Digital Scholarship in the Humanities, 2024, **39**, 1075–1091
https://doi.org/10.1093/llc/fqae050
Advance access publication 16 September 2024
Full Paper

Measuring the semantic headedness of English blends with token-based semantic vector space modeling: a corpus-based study

Qingnan Meng [1], **Martin Hilpert** [2,*]

[1]School of Foreign Languages, Dalian Maritime University, Dalian, 116026, China
[2]Department of English, Université de Neuchâtel, Neuchâtel, 2000, Switzerland
*Corresponding author. Department of English, Université de Neuchâtel, Neuchâtel, Switzerland. E-mail: martin.hilpert@unine.ch

Abstract

This article analyzes the semantic headedness of English blends with distributional semantics methods. The semantic head of a blend is the source word that transfers its semantic information to the blend as a whole. For example, a *sitcom* is a kind of *comedy*. But is *FedEx* a kind of *express*, and is *wi-fi* a kind of *fidelity*? We use corpus data and token-based semantic vector space modeling in order to address these questions. Specifically, we investigate whether Plag's ternary division of endocentric, exocentric, and coordinative compounds based on semantic headedness can also be applied to English blends, and whether the general tendency of semantic right-headedness can be observed for all three subtypes. We analyze a dataset of fifty-five blends and their respective source words, using data from the Corpus of Contemporary American English and the English Web Corpus 2021. We measure the degree of semantic similarity between each blend and its two source words. The results show that for most endocentric blends, the hypothesis of semantic right-headedness holds true. At the same time, exocentric blends and coordinative blends are shown to behave differently. We conclude that Plag's classification offers a useful point of departure for the semantic analysis of blends and that distributional semantics methods can provide new insights into their semantic behavior.

Keywords: English blends; semantic headedness; token-based semantic vector space modeling; multinomial logistic regression analysis.

1. Introduction

Blending in English is a word formation process with a long history. The earliest blends, also known as portmanteau words, amalgams, or coalesced words, can be dated back to Old English manuscripts from the seventh century (Thurner 1993: viii). Since the late 19th century, blending has become a highly frequent word-formation process, and in Present-Day English blending remains very productive (Lehrer 1996: 360). However, in many linguistic frameworks, blends are seen as peripheral to English word formation, or of no importance to morphological theory, and they are consequently confined to "extra-grammatical morphology" (Dressler 2000: 5; Mattiello 2013: 111). In Marchand (1969: 451), blending is only mentioned in passing and is claimed to "have no grammatical, but a stylistic status." What is more, existing definitions of blending are quite heterogeneous, depending on whether both clipping and overlap are involved, how many source words are included, which parts of the source words are shortened, whether we are dealing with phonological blending or orthographic blending, etc. Be that as it may, there are some widely agreed prototypical characteristics, which are summarized by Grlj (2022: 86–87):

> [Blending] involves the shortening of at least one source word, but frequently the source words display some overlap. The most typical pattern of blending is the front part of the first source word and the last part of the second source word.

As to the classification of blends, many studies exclusively focus on structural properties such as categorial composition, morphosyntactic headedness, and phonological properties (Algeo 1977; Lehrer 1996; Kelly 1998). In more recent research, more attention has been paid to the semantic relationships between blends and their source words. According to Bauer, Lieber, and Plag (2013: 485), "blends generally are interpreted in the same way that compounds are, though not necessarily in the same proportions." In their classification of

compounds, two orthogonal distinctions are proposed: "endocentric" versus "exocentric" in terms of semantic headedness, and "argumental" versus "non-argumental" on the basis of the argument structure of the two components. In an argumental compound, one element is interpreted as a grammatical relation (e.g. subject, object) of the respective other. For example, in *club member*, which refers to "a member of the club," the first element is the prepositional object of the second element. For the semantic classification of blends, however, Bauer, Lieber, and Plag (2013) only mention the "argumental" versus "non-argumental" distinction. The latter is claimed to be the more common type, with a further division into "attributive" (also termed "determinative," "telescopic" elsewhere) and "coordinative" blends, depending on whether the blend as a whole is a hyponym of the second blended element. For example, blends such as *daycation*, *beersicle*, or *carbage* belong to the attributive type, since they are a subtype of *vacation*, *popsicle*, and *garbage*, respectively. For coordinative blends, a further distinction of "appositive" versus "compromise" is made, depending on whether the blend retains the semantic features of both source words (e.g. *actorvist—actor/activist*) or is only a hybrid of the two (e.g. *tigon—tiger/lion*). These distinctions are captured in Fig. 1.

Another semantic distinction in compounds that cross-cuts with "argumental" versus "non-argumental" is "endocentric" versus "exocentric." This pair of terms was first introduced by Bloomfield (1933: 235) in order to describe the relation between a compound and its component words. In Bauer, Lieber, and Plag's (2013) discussion of non-argumental compounds, it is stated explicitly that attributive compounds can be either endocentric or exocentric, as are the coordinative ones. For exocentric coordinative compounds, three relationships are further distinguished (Bauer, Lieber, and Plag 2013: 481): conjunctive (e.g. a *father–daughter* dance), translative (e.g. *Arab–Israel* conflict), and disjunctive (e.g. *pass–fail* test). Besides, it is also argued that exocentric attributive compounds such as *egghead* can be taken as regular endocentric compounds with a metonymic or metaphoric reading of the head. Bauer, Lieber, and Plag (2013) do not offer a parallel description for non-argumental blends, save for a passing note that for some of the blends that have coordinative interpretations, "an endocentric reading is also possible" (2013: 483). More recent discussions have provided further insights into the semantics of English blends. For instance, Beliaeva (2014) conducted a multifactorial analysis examining the interaction between the form and meaning of blends, which allowed her to pinpoint phonological and structural differences between blends and clipping compounds. Beliaeva (2019) further provided a nuanced classification of blends, distinguishing between "associative" blends (which include synonymic, paradigmatic, hybrid, and jumble blends) and "syntagmatic" blends. Renner (2015) proposed a prototype-based typology on the basis of increasing degrees of playfulness inherent in lexical blends. Finally, Tarasova and Beliaeva (2020) conducted a comparative experimental study focusing on the semantic relationship between constituents of English determinative blends and N + N subordinative compounds. Building on these studies, which have significantly advanced our understanding of English blends, this study aims to offer a new perspective on English blends by adopting a quantitative methodology that harnesses corpus data.

Our theoretical point of departure is the question of how the headedness of blends can be analyzed. Based on Bauer, Lieber, and Plag (2013), Plag (2018) has proposed a ternary classification for English compounds in terms of semantic headedness, namely "endocentric," "exocentric," and "coordinative." A compound is endocentric (Sanskrit "tatpuruṣa") if its semantic head is one of its components. If the semantic head of a compound is not one of its components, then it is exocentric (Sanskrit "bahuvrihi"). This type seems

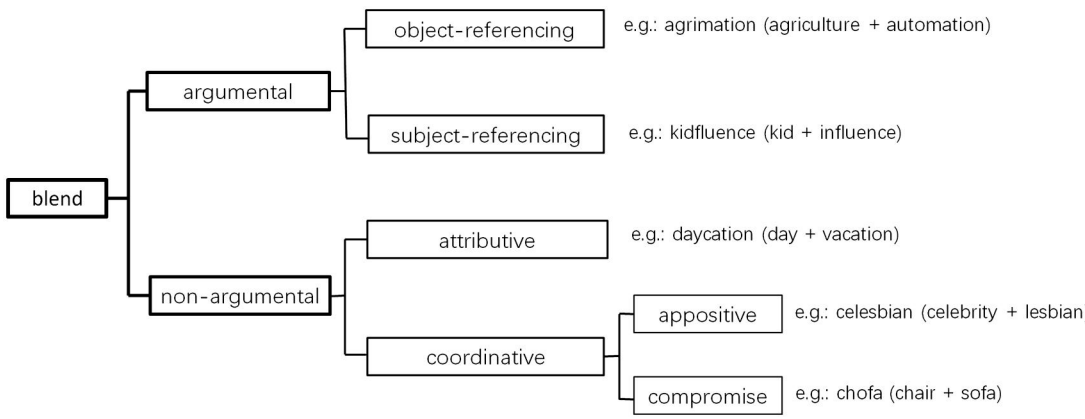

Figure 1. Semantic classification of English blends (adapted from Bauer, Lieber, and Plag, 2013: 484).

to be restricted to human beings or higher animals (Plag 2018: 138). If both components contribute equally to the meaning of a compound, then it is coordinative (Sanskrit "dvandva").[1] In terms of English blends, however, Plag (2018: 120) only makes a brief comment to the effect that "semantically, blends behave like compounds," and a similar claim can be found in Renner (2019: 38), namely "it is expected that, like compounds, subordinate blends should be canonically right-headed in English." However, neither of the two studies makes explicit how the notions of endocentricity and exocentricity should be operationalized. Bat-El (2006) employs the terms "endocentric blends" and "exocentric blends" in her discussions, but if we apply Plag's (2018) ternary division, examples such as *smog* and *brunch* that Bat-El classifies as exocentric are in fact coordinative.[2] This raises the question of whether there really are exocentric blends in English. In existing studies, only a few have been identified as such. For example, in Mattiello (2013: 124–125), three exocentric blends are given: *Frutopia* (< fruit + Utopia), *helilift* (< helicopter + lift), and *fortran* (< formula + translation). Tomić (2019) presents another six examples: *bromance* (< brother + romance), *Koreegro* (< Korea + negro), *mangina* (< man + vagina), *neek* (< nerd + geek), *schmiddy* (< schooner + middy) and *yestergay* (< yesterday + gay). Apart from these, the blends *bionic* (< biological + electronic), *FedEx* (< federal + express), and *pixel* (< pix + element) are another three possible candidates. For *bionic*, its two source words do not belong to the same semantic domain, unlike most coordinative blends. For the other two, they do not exhibit a subordinate relationship, at least not in their usual senses or as typical collocations. However, using a strict definition of exocentric blends, we may argue that most of these examples are still semantically compositional and transparent, and some of them are merely proper names or dialectal variants. To what extent the semantic head of a blend can be taken as "outside" its two component parts is an empirical question. Thus, the main goal of this study is to explore whether Plag's (2018) ternary semantic classification for compounds is applicable to English blends as well, and whether the category of exocentric blend can be clearly defined in English morphology.

The remainder of this article is structured as follows. Section 2 first critically reviews Plag's (2018) formal classification of English blends, and then provides a working definition of semantic headedness, the coverage of which is in line with but also beyond the traditional hyponymic definition of endo/exocentricity. Section 3 explains how a token-based semantic vector space is created with data extracted from two large corpora. Section 4 first discusses the applicability of the semantic right-headedness hypothesis within the context of English blends, and then presents visualizations that allow analysis for each subtype of blends, with a special focus on two counterexamples to the more general tendencies. Section 5 concludes the article, summarizes the main findings, and suggests new directions for further corpus-based morphological research.

2. Plag's Classification of English Blends

In Plag's (2018: 120–122) formal discussion on English blends (see Table 1), two subtypes are distinguished, namely "clipped compounds" (or "clipping compounds" as in Marchand 1969: 445) and "proper blends." Type-1 blends are shortened from the first part of each source word in existing compounds (AB + CD → AC), so they are also termed "AC blends." For Type-2 blends, usually there is no corresponding full form, and they are composed of the first part of source word 1 and the final part of source word 2, hence "AD blends" (AB + CD → AD), where either B or C can be null, as in *guesstimate*.[3] According to Bauer, Lieber, and Plag (2013: 459), the first type only accounts for 5 per cent of all the blends in English.

When we take a closer look at the classifications above, there are a few questions that remain to be further elaborated. First of all, this concerns the presence of the full forms in the language from which the blend is, or has been, derived. For *sci-fi* and *sitcom*, it is clear that their full forms exist as well, but for *modem*, it is debatable whether it is appropriate to call it a clipped compound, given that speakers of English no longer use the full form "modulator (and) demodulator." Apart from the items listed above, how can we deal with cases in which the full forms are only theoretically possible but are seldom observed in authentic language use, such as *wi-fi* (< wireless + fidelity)? Should we set a threshold value for the minimum number of hits in a large corpus to justify the existence of its full form? Besides, how can we connect this binary formal distinction to the ternary semantic classification which was originally intended for the analysis of compounds and their semantic headedness? If Type-1 corresponds to endocentric blends and Type-2 corresponds to coordinative blends, does this mean that English has no exocentric blends at all?[4]

Given the problems mentioned above, this study explores whether Plag's (2018) ternary semantic classification for compounds based on semantic headedness can be applied to English blends as well. Specifically, we investigate whether endocentric, exocentric, and coordinative blends display different distributional patterns of semantic overlap with the two source words.

Table 1. Plag's (2018: 121) classification of English blends

Type 1 clipped compounds (AC blends)	Type 2 proper blends (AD blends)
science + fiction → sci-fi	boat + hotel → boatel
hydrogen + railway → hydrail	boom + hoist → boost
modulator + demodulator → modem	breakfast + lunch → brunch
situation + comedy → sitcom	channel + tunnel → chunnel
	compressor + expander → compander
	goat + sheep → geep
	guess + estimate → guesstimate
	sheep + goat → shoat
	smoke + fog → smog
	Spanish + English → Spanglish
	stagnation + inflation → stagflation

In previous studies, semantic headedness as an indicator of endo/exocentricity is usually defined by using a test of hyponymy between the source words and the blend itself in a narrow sense. While this often yields useful results, tests of this kind are not foolproof. Renner (2019: 37) has pointed out that sometimes only select conceptual features are inherited from the source word. For example, *affluenza* (< affluence + influenza) is not literally (a form of) influenza but rather an inheritance of the latter's lexical–conceptual feature of "malaise." Only in this broad sense can *affluenza* be seen as endocentric.[5] Another issue in the traditional definition is the problem of polysemy: when a blend has several senses, only the earliest, original sense is retained for its semantic classification (p. 37). But this may be inconsistent with the actual language use, as the original sense may not necessarily be the most frequent one in present-day English. On the other hand, even for the least ambiguous blends, with satisfactory contextual cues, alternative interpretations are still possible, as is shown in English compounds. Tarasova and Beliaeva (2020: 24) offer the example of a *police dog*, which can mean "a dog IN the police," "a dog that IS police," or "a dog that the police HAS," amongst other interpretations.

To avoid such problems, this study employs a quantitative corpus-based approach in order to model blend meaning as a highly plausible combination of all the possible readings, which is meant to uncover the prototypical meaning. Unlike traditional approaches which deal with the semantic head of endocentric blends or compounds only, we offer a working definition of semantic headedness not only as an approximation to the narrow hyponymic definition of endocentricity, but also as a tentative explanation of the sequence of components in coordinative and exocentric blends. Specifically, we rely on distributional semantic methods (cf also Günther and Marelli 2022 on semantic aspects of compounds and how they can be measured using distributional semantic methods). For the purpose of this study, we use token-based semantic vector space modeling as a tool to display the distributional behavior of each English blend we select. The semantic (dis)similarity between the blend and its source words is operationalized by the distance between their medoids. If the medoid of the blend is closer to a certain source word, then that source word is taken as the semantic head. Our hypothesis is that for endocentric blends, the semantic head is usually source word 2, whereas coordinative blends will be semantically double-headed, so that the blend and its two source words are equally distant in their respective meanings. For exocentric blends, we suppose that it is unpredictable which source word is semantically closer to the blend, since the semantic heads of exocentric blends are claimed to be outside the blend. Our approach will be described in detail in the next section.

3. Data and methodology

Drawing on Plag (2013, 2018), Mattiello (2013), and other publications on word formation, we make a selection of fifty-five blends (see the Appendix Table A.1 for a full list) that form the basis for our study. There are two main reasons why we did not select equal subsets of endocentric, exocentric, and coordinative blends. First, as is argued in Section 1, it is very difficult to find exocentric blends in previous studies, and even those labeled as exocentric blends are, in fact, coordinative according to Plag's (2018) ternary division. Secondly, our data-driven approach mainly aims to reconsider the semantic categories assigned to each blend in the existing literature based on their semantic headedness. Therefore, it is not crucial for the number of blends in the three subsets to be equal, since we are not dealing with the overall distribution of the three subtypes of blends. Following Plag's (2018) approach, only blends with two source words are included (i.e. excluding blends such as *turducken* < turkey + duck + chicken), but unlike Plag (2018), we also include blends such as *newscast* and *docudrama*, where the entire source word 1 or 2 is retained.[6] We analyze the fifty-five blends on the basis of corpus data. Our main source of data for this study is the 2014 release of the off-line version of the Corpus of Contemporary American English (COCA; Davies 2008), which contains 440 million word tokens. We use a context span of ten words to the left and right to retrieve the concordance lines for the fifty-five blends and their respective two source words. If a given blend or its source words are extremely low in frequency which is, as a rule of thumb, below 900 hits in the COCA, the fifty-two billion words English Web Corpus 2021 (enTenTen21, "TenTen

Corpus" for short below; Suchomel 2020) is resorted to as a supplement to ensure that after eliminating those semantically less informative concordance lines, there are still around 200 hits left for the subsequent step of plotting. The TenTen Corpus can be accessed through the Sketch Engine (Jakubíček *et al.*, 2013; Kilgarriff *et al.*, 2014).

In order to analyze the semantic relationship between blends and their two source words, a token-based semantic vector space model was constructed. The idea of semantic vector space modeling can be traced back to Harris's (1954: 156) "difference of meaning correlates with the difference of distribution" and Firth's (1957: 11) oft-cited aphorism "you shall know a word by the company it keeps."[7] A modern interpretation of this theoretical framework is "the statistical semantics hypothesis" (Turney and Pantel 2010), which is operationalized by semantic vector space modeling. This technique provides a means to explore lexical semantic structure in large corpora (Heylen *et al.*, 2015). In this approach, both the identification of relevant context features and the detection of patterns in the co-occurring features are achieved through the analysis of co-occurrence frequencies in corpus data, which thus provides an empirical basis for the analysis of meaning in corpus data.

Two levels of semantic vector space models can be distinguished. Type-based models differ from token-based models, depending on whether we want to model a word form by its surrounding context words (i.e. first-order collocates) or the context words themselves by means of their immediate collocates (i.e. second-order collocates). The token-based model, as a methodological extension of the type-based model, is a more recent proposal which was originally developed in computational linguistics in the mid-1990s (cf Schütze 1992), and since Heylen, Speelman, and Geeraerts (2012), it gradually became popular in lexical semantic studies, either to separate near-synonyms from an onomasiological perspective or to explore the semantic structure of a certain lexical item from a semasiological perspective. With a token-based model, researchers can empirically assess the context-dependent variability that is inherent in the usage of blends, since different usage events are represented in their own right. In other words, the issue of polysemy is directly reflected in differences between the semantic vectors. To be more precise, for each concordance line, the situated meaning of a blend or its source word which is barely ambiguous in a given context is represented by a unique vector. Each of these vectors corresponds to a unique datapoint in the overall distribution, so that the relative frequencies of the different senses can be displayed by the relative positions and density differences of the datapoints that are visualized. Another advantage of token-based models is that both situated meanings and abstractions from those meanings can be reflected at the same time. The prototypical semantic structure of a blend or its source word thus naturally emerges from the relative weights of different senses which represent generalizations across contexts.

In the process of model construction, many practical decisions need to be made, such as the vector size and window size, feature selection, similarity metric, weighting scheme, stopwords, and cut-off values (Kiela and Clark 2014). Additionally, in practice, we have to strike a balance between computational resources and the performance of the model. Drawing on Hilpert and Correia Saavedra (2020) and Hilpert, Correia Saavedra, and Rains (2023), this study uses the following steps in order to create token-based semantic vector spaces.[8]

1) A type-based semantic vector space is created using a frequency list of 20,000 elements (i.e. word tokens) from the British National Corpus (BNC; Davies, 2004). Excluding the top 200 most frequent elements and other unwanted items such as single letters, numbers, primary verbs, and punctuation marks, the vector space is reduced to a matrix with 19,429 rows and 19,429 columns. The cells in the matrix are filled with raw co-occurrence frequency values between each vocabulary item and its context items in a context span of two elements to the left and right. The frequency values are transformed with Pointwise Mutual Information (PMI). Rows and columns that do not show at least one PMI value that exceeds the threshold value of PMI = 5.5 are excluded. This reduces the matrix to 12,621 rows and 12,619 columns, with only the most informative rows and columns retained (for full details on this process, see Hilpert and Correia Saavedra, 2020).
2) For each blend and its two source words, concordance lines with a context span of ten words to the left and right, respectively are extracted from the off-line version of COCA (2014 release). The TenTen corpus is used as a supplement for those infrequent blends or source words in order to avoid data sparseness in the subsequent analytic steps. If the token frequency is over 10,000 in the TenTen corpus, only the first 10,000 concordance lines of the blend and/or its source word(s) are extracted. If the token frequency for any of the three elements is still below 200, then the blend is excluded from our analysis.
3) The concordance lines extracted from the TenTen corpus are tagged with TreeTagger (Schmid 1995) using the BNC C5 CLAWS tagset, and the

results are stored in a three-column plain text file: word, lemma, and part-of-speech (wlp) tags. Then all these tagged words are scanned into R (R Core Team 2023) and concatenated to construct an R workspace. To make it consistent with the tagged version of BNC used for type-based vector space prepared by Hilpert and Correia Saavedra (2020), all the words and tags are further converted to the C5 tagset in the format of the tagged BNC.

4) In each concordance line, all the context items that are not represented in the type-based BNC data frame are removed. We only retain cleaned-up concordance lines that contain at least four elements that can be expressed by column vectors of PMI values in the BNC vector space. The PMI values of these column vectors are summed up and averaged, so that each concordance line is now represented by a single vector of PMI values.
5) To ensure that the blend and its two source words are evenly represented in the visualization step (i.e. the MDS plot in step 7), a threshold value of 200 is adopted for the number of cleaned-up concordance lines for each of the three elements. If the minimum number is smaller than 200, then the actual number was used as the reference. Next, all the concordance line vectors are combined to form a single data frame with the same context items as row variables. This new data frame is the so-called token-based semantic vector space.
6) A cosine-based distance matrix is created to display the (dis)similarities between different concordance lines. To display the results in a 2D space, a classical metric multidimensional scaling (MDS) technique is adopted with cmdscale() in R. The x and y values represent the coordinates for each cleaned-up concordance line in the 2D space. The closer the two datapoints are in the graph, the more similar are the meanings of the two concordance lines they represent.
7) In order to further display how well the model can distinguish between the blend and its two source words, a multinomial logistic regression analysis is conducted with two predictors from the first two dimensions in the MDS plot. Its classification accuracy is calculated based on a 3×3 confusion matrix. In addition, for a clear display of the semantic relations between the three elements, their corresponding medoids[9] with 90 per cent confidence ellipses are added to the graph, and then the Euclidean distances between medoids are calculated, so that the semantic headedness for each blend can be assessed. As a rule of thumb, we assign 1.5 as a cut-off point. If the distance between the blend and source word 1 (d1) is at least 1.5 times the distance between the blend and source word 2 (d2) or the other way around, we would say that it is a typical endocentric blend. If the ratio of d1 to d2 is between 0.67 and 1.5, then it is taken as a typical coordinative blend. For exocentric blends, the relative position of the three ellipses is a more important criterion: the smaller the degree of overlap, the more typical its degree of exocentricity.

To justify the research methods and procedures aforementioned, we now include an independent example of the English compound *goldfinch* which is traditionally classified as endocentric, since according to the restrictive hypoymic definition, a *goldfinch* is a kind of *finch*, but not a hyponym of *gold*. Its graphical representation is displayed in Fig. 2. The values in the two dimensions represent the first two principal coordinates calculated through metric MDS, which aims to maximally preserve the relative distances between original datapoints on the basis of an underlying high-dimensional vector space. The semantic centers for *goldfinch* and its two source words *gold* and *finch* are operationalized by the medoids (marked by triangles) of 200 concordance lines. A 90 per cent confidence ellipse is plotted around the medoids. As is shown in the graph, the distance between *goldfinch* and *gold* (0.216) is 1.6 times between *goldfinch* and *finch* (0.135). The classification accuracy of the multinomial regression model is 62.67 per cent, which is not ideal but still much higher than the baseline accuracy. Consequently, the distributional feature of *goldfinch* and its two source words is in line with its semantic classification of endocentric blend with the canonical right-headedness.

4. Results and discussion

This section presents findings that are obtained from the visualization of token-based vector spaces. We first test the hypothesis of semantic right-headedness, and then discuss three typical examples of endocentric, exocentric, and coordinative blends, respectively. Next, we move on to some cases where the graphical configurations are inconsistent with Plag's (2018) semantic categories. In the end, we develop some new ideas that are inspired by the graphs and that allow us to rethink the current semantic classification of English blends.

4.1 Semantic headedness in English compounds and blends

In Plag's (2018: 140–141) discussion of compounding, it is argued that "English compounds are generally

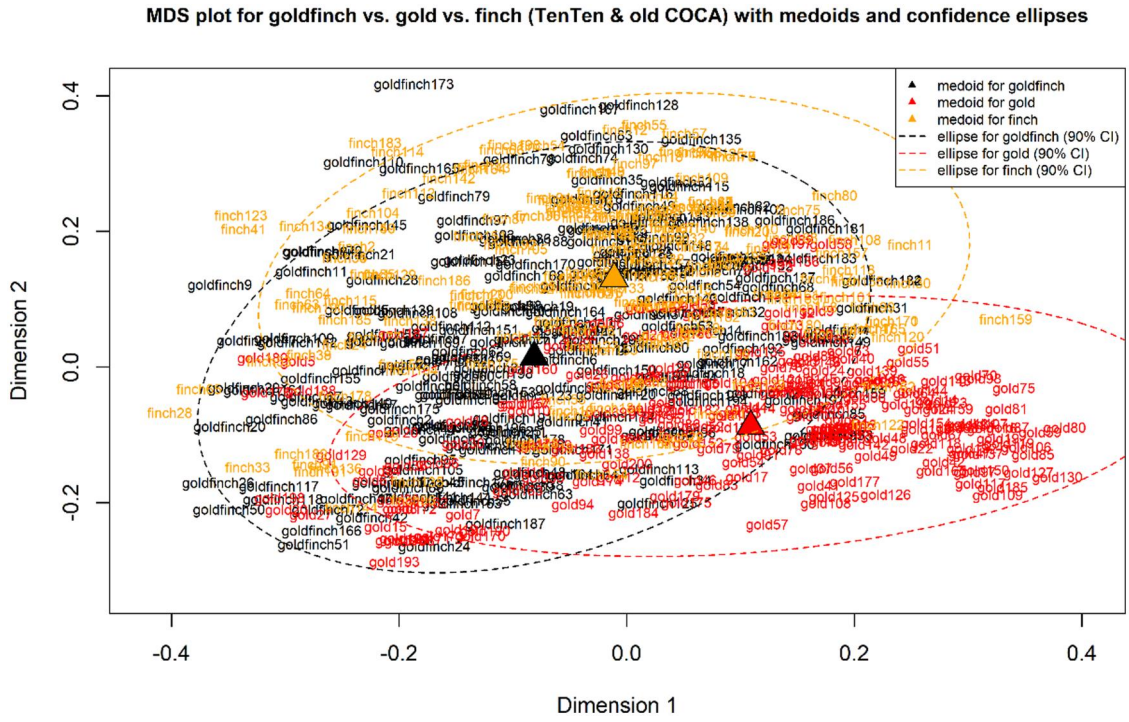

Figure 2. MDS plot for a prototypical endocentric compound *goldfinch*.

headed, and the head is always the right-hand member." Coordinative compounds, however, have two heads, so whether they are perhaps equally right-headed as the other two types is an open question. Plag (2018: 140) argues that right-headedness at least holds true in terms of morphosyntactic headedness—"plural marking occurs only on the right-hand member," as illustrated in *poet-translators* (**poets-translator*, **poets-translators*).[10] Starting from this point, we test the hypothesis of semantic right-headedness in English blends. Specifically, for endocentric blends, the fact that the semantic head is inside the blend naturally presupposes right-headedness. We therefore hypothesize that endocentric blends should display a smaller semantic distance between the medoid of the blend and source word 2, with a ratio of $d1/d2 \geq 1.5$. For exocentric and coordinative blends, we would expect a relatively larger distance between the medoid of the blend and source word 2, and the ratio of d1/d2 is expected to be between 0.67 and 1.5 (see graphs in Section 4.2 below for examples and further explanation).

Turning to our results, a complete list of the 55 blends, their source words, semantic distances, and semantic headedness is provided in the Appendix Table A.1. The results indicate that among the 55 blends we analyzed, 34 are right-headed in terms of the operationalization that we set out above. If we take a closer look at the three categories, we see that the behavior of endocentric, exocentric, and coordinative blends is not uniform. For endocentric blends, twenty out of twenty-seven are right-headed. The seven counterexamples are *breathalyzer*, *chocoholic*, *shopaholic*, *bromance*, *docudrama*, *wi-fi*, and *netizen*. On the one hand, this means that the semantic head and the morphosyntactic head may not necessarily be the same. For example, we may argue that a *netizen* is a kind of citizen, since source word 1 *Internet* syntactically modifies source word 2 *citizen*, but whether the semantic class of *netizen* is inherited from *citizen* is an open-ended question, since the identity of a netizen is quite different from traditional citizens: they are usually anonymous, and can be actively involved in any online discussion with no fixed affiliation to a particular community or country. So *Internet* would be the more salient semantic feature for *netizen*, as is reflected in the semantic overlapping of the MDS plot. The same goes for *bromance*, since the non-sexual relationship between males is definitely not a kind of romance in the traditional sense, and that is perhaps why Tomić (2019) labels it as exocentric. On the other hand, it may also be due to the fact that the second source word is polysemous, and that the sense we want to explore is downplayed by other irrelevant senses. For example, a *breathalyzer* is "a device used by the police to measure the amount of alcohol in a driver's breath"

according to the *Oxford Advanced Learner's Dictionary*, 10th edition (Oxford University Press 2020). Thus, it is a piece of equipment, rather than a human being who analyzes.

For coordinative blends, the results are quite heterogeneous—eleven out of twenty-three blends are semantically left-headed in our dataset (i.e. *guestimate*, *modem*, *vog*, *smog*, *boost*, *stagflation*, *happenstance*, *brunch*, *frenemy*, and *transceiver*) according to our working definition of semantic headedness, namely the blend is semantically closer to its left constituent than its right constituent. This is actually in line with Kelly's (1998) "prototypical first" principle for the sequence of coordinative blends. But Kelly's example of *spork*, on the contrary, is right-headed based on our plot, which means other factors may play a more important role in determining the sequence. For example, according to the constraint of initial consonant complexity (cf Renner 2014), blends should retain the more complex onset of the two source words. In our examples of left-headed blends, other factors mentioned in Kelly (1998) such as length (shorter first, as in *guestimate*, *happenstance*), frequency (more frequent first, as in *boost*), and pragmatics (temporal order, as in *brunch*) may serve as an explanation. Additionally, it should be noted that for some of our examples, the distances between the medoid of the blend and that of its two source words are quite close (ten out of twenty-three blends have a ratio of d1/d2 between 0.67 and 1.5), and whether they are right-headed is difficult to determine on the basis of eyeballing our visualizations (shown below). For exocentric blends, since we do not have enough a large number of different types, it is hard to draw any general conclusion. Among the five types that we analyzed, *fortran*, *FedEx* and *pixel* are right-headed, whereas *bionic* and *helilift* are left-headed. To the naked eye, however, most of them seem to be double-headed, in that the distances between the medoid of each blend and its two source words are almost the same, and this is further supported by their ratios of d1/d2 ranging from 0.79 to 1.22. Thus, whether exocentric blends are semantically right-headed is left in doubt as well, and we may safely conclude that the general tendency of semantic right-headedness is only confirmed for most endocentric blends.

4.2 Prototypical members for endocentric, exocentric, and coordinative blends

Is there evidence to say that endocentric, exocentric, and coordinative blends display different patterns of semantic overlap with their two source words, or is it more adequate to maintain the traditional binary distinction of "endocentric" versus "exocentric" blends? In this section, we choose one prototypical member for each semantic category and explain in detail how their distributional behavior can be analyzed with token-based vector spaces.

Taking *sitcom* as a prototypical example of endocentric blends, its graphical representation is displayed in Fig. 3. The semantic centers for a *sitcom* and its two source words are operationalized by the medoids of 200 concordance lines. As is shown in the graph, the distance between *sitcom* and *comedy* (0.060) is almost 1/6 of that between *sitcom* and *situation* (0.338). This is further validated by the degree of semantic overlap between the confidence ellipses—an almost complete overlap between *sitcom* and *comedy* illustrates the blend's semantic right-headedness. The classification accuracy of the multinomial regression model is 66.32 per cent, which means that the two dimensions in the MDS plot can distinguish between *situation* and the other two elements to a large extent, but less so between *sitcom* and *comedy*.

For *FedEx* (case-insensitive), which we argue may be a qualified candidate for an exocentric blend, its 90 per cent confidence ellipse is almost entirely outside the two corresponding ellipses for *Federal* and *Express*. This is consistent with what we would expect from an exocentric blend, whose semantic head is supposed to be outside the two source words. Since *express* is polysemous and can be tagged differently in different contexts, we only retrieve nominal *express* from the COCA, in conformity with its part-of-speech in the full compound form *Federal Express*. Similarly, we exclude the verbal use of *FedEx* as in "I was dismayed that Sprint Customer Service didn't offer to *fedex* me a free or cheap phone overnight" in the TenTen corpus (2021 release). For naturally occurring language data, a clear separation of the three clusters is impossible, as we can examine in the datapoint express114, which enters the 90 per cent confidence ellipse of *FedEx*:

> (1) … for replacement lenses if no new prescription was needed. Lens **Express** has encountered plenty of resistance from the traditional distribution channel. (COCA, 1994, mag)

Here the word *Express* is part of the proper name *Lens Express*, a direct-mail-order company that sells contact lenses that can be ordered over the phone, and that explains why it is distributionally similar to *FedEx*.

The classification accuracy of the multinomial logistic regression model with two dimensions in the MDS plot (Fig. 3) as predictors is 86.5 per cent, which is much higher than the accuracy for *sitcom*. That means the model can tease apart the three elements quite effectively, especially between *FedEx* and the other two.

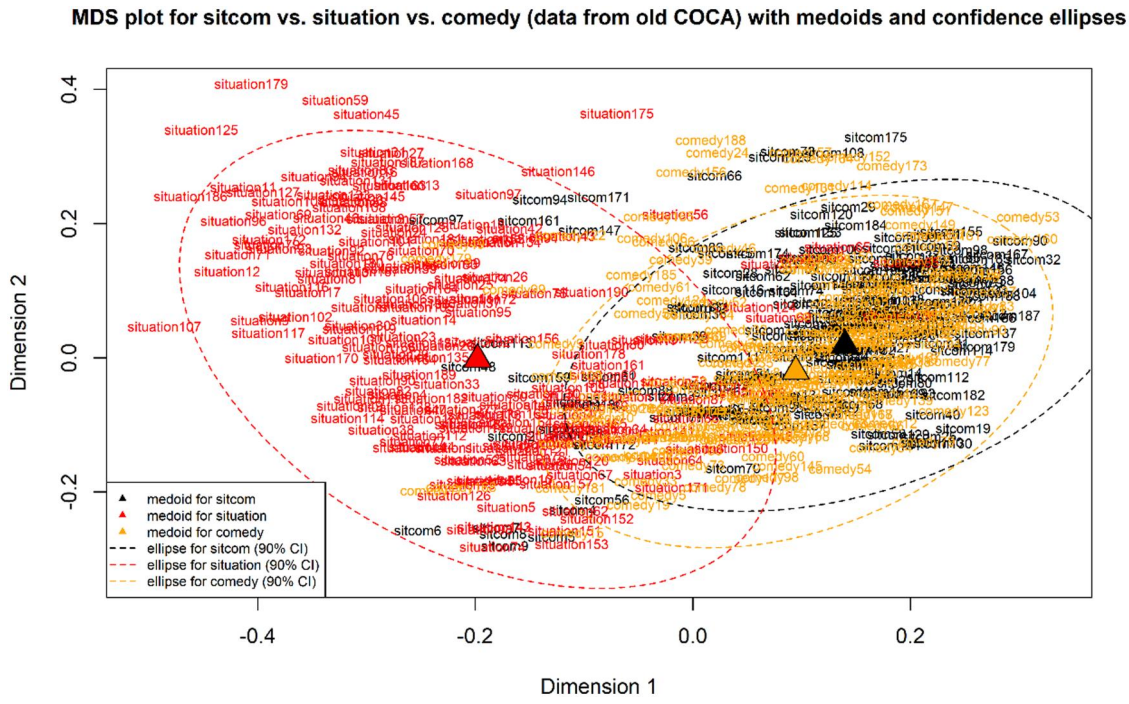

Figure 3. MDS plot for a prototypical endocentric blend *sitcom*.

For *brunch* as a typical coordinative blend, its two source words are co-hyponyms of *meal*, and the three elements are paradigmatically related, so we would expect their 90 per cent confidence ellipses to overlap to a large extent, and this is in fact attested in Fig. 4 below. According to Bat-El (2006: 67), *brunch* has two different senses: it means either "lunch with some characteristics of breakfast" or "a mixture of breakfast and lunch." If this is true, then the medoid for *brunch* should be closer to that for *lunch*, since in the first sense *brunch* is semantically closer to *lunch* and in the second *brunch* is supposed to be roughly equidistant from *breakfast* and *lunch*. But as can be observed in Fig. 5, the medoids between *brunch* and *breakfast* are relatively closer (0.052), compared with those between *brunch* and *lunch* (0.104). So we would argue that the first sense in Bat-El (2006) is not its prototypical meaning, and that *brunch* is more frequently used to refer to a meal that is served midway between the time when breakfast and lunch are customarily eaten, probably temporally proximal to breakfast. This interpretation also corresponds to the earliest attestation in 1896 in the OED, which indicates that the combination-meal, when nearer the usual breakfast hour, is "brunch," and when nearer luncheon, is "blunch" (Renner, 2015: 130). In addition, though intuitively *lunch* is a more prototypical meal than *breakfast*, as is attested in idiomatic expressions such as "go to lunch" or "let's have lunch" which has become a fixed expression for "see you again; goodbye," yet for some groups of speakers such as construction workers *breakfast* may be the more prototypical meal, since they have to be out early and thus may think of breakfast as a meal and lunch as a snack.[11] This may serve as another possible explanation for the slight semantic left-headedness of *brunch* in our MDS plot, apart from the sampling error of the data. An alternative explanation for this current order may also be due to the constraint of initial consonant complexity for blends (cf Renner 2014), as the onset of the first syllable in *breakfast* is more complex than that of *lunch*.

The classification accuracy of the corresponding multinomial logistic regression model is 46.5 per cent, only slightly above the baseline accuracy of 33.33 per cent. That means the two dimensions in the MDS plot really cannot effectively distinguish the three elements, and this is further evidence of the blend's coordinative nature. Of course, in coordinative blends, the two source words can also be near-synonyms or near-antonyms or even from different semantic domains, but the general feature of the graph, namely an equal degree of distributional overlap between the blend and its two source words, will remain roughly the same.

Judging from the three MDS plots above, we can conclude that for prototypical endocentric, exocentric, and coordinative blends, their distributional behaviors are distinct from one another, and based on all the fifty-five blends in our dataset, the order of average classification

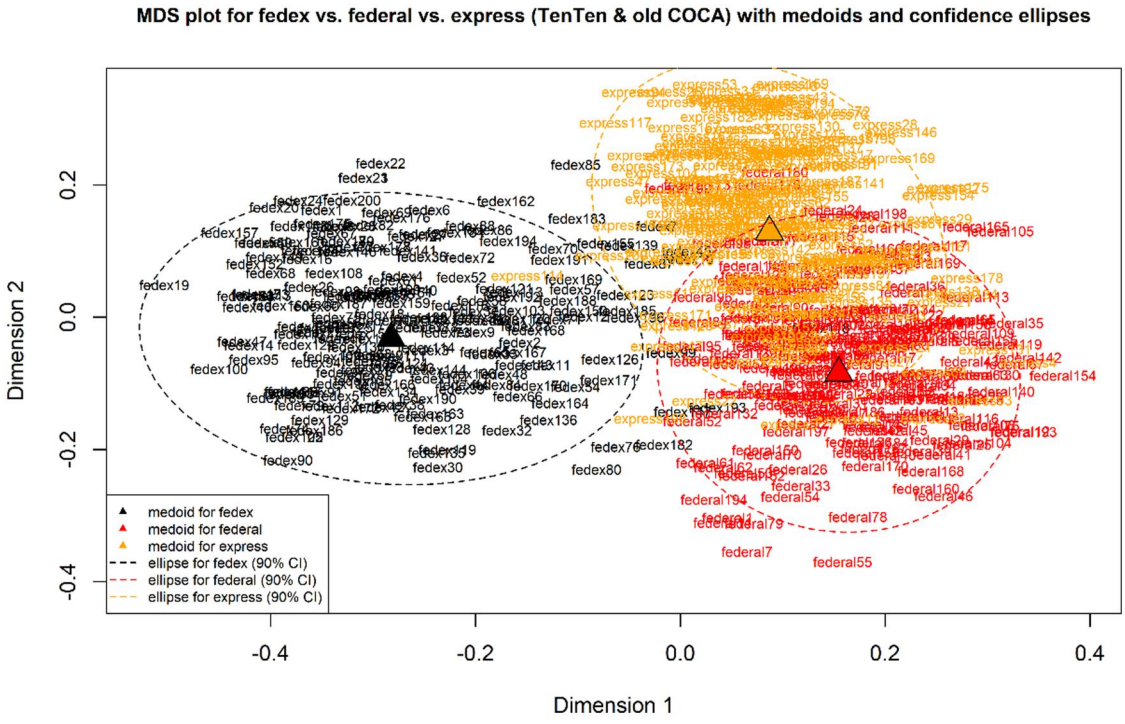

Figure 4. MDS plot for a prototypical exocentric blend *FedEx*.

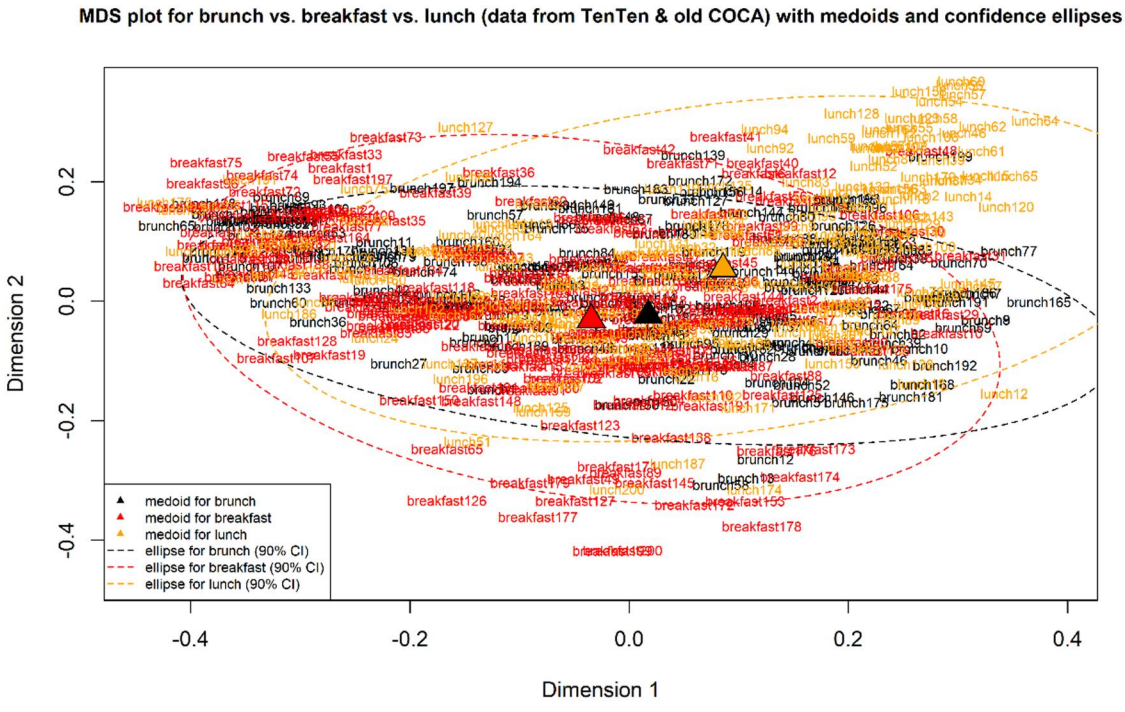

Figure 5. MDS plot for a prototypical coordinative blend *brunch*.

accuracy for each semantic category in the multinomial logistic regression model is "coordinative (55.59%) < endocentric (56.92%) < exocentric (69.89%)" in ascending order.

4.3 Inconsistencies between linguists' intuitions and graphical representations

We admit that the conclusions above only reflect probabilistic tendencies, and there is no clear-cut boundary between two adjacent semantic categories. In other words, some coordinative blends may exhibit a relatively high classification accuracy in a multinomial logistic regression model, while for certain endocentric blends, the classification accuracy may even be below the average value of that for coordinative blends. Thus, it is better to take Plag's (2018) three semantic categories as a gradient. Additionally, for some blends, the semantic headedness may also not be consistent with our expectations and those of other linguists who have analyzed blends. In this section, we analyze two such counter-examples to explain in detail how this can come about.

First, we examine *wi-fi*, a commonly used blend in our daily life. According to the etymological information in the third edition of (online) The Oxford English Dictionary (OED),[12] it is originally from *wireless* (adj.) + *-fi*, and the second element is arbitrarily chosen, probably for euphony and in analogy with the second syllable of *hi-fi* (< high + fidelity). Later, it was reinterpreted as a shortening of *wireless fidelity*. Its first OED quotation is from 1999, much later than that for *hi-fi* (from 1935). We plot it against the two components in its supposed full form *wireless fidelity*, and since the capitalized *Fidelity* is usually part of proper names such as *Fidelity Fund* or *Fidelity Investments*, we limit our search query to the non-capitalized form *fidelity* exclusively. There are still some issues to be further discussed in the process. It should be noted that in the COCA, all the instances of *wireless* are tagged as singular common nouns (nn1) which used to be a common term for "radio" in British English. In addition, even if we exclude all instances of *Fidelity* used as proper names, still most of the concordance lines do not concern the sense of "high quality." As is shown in its 90 per cent confidence ellipse in Fig. 5, *fidelity* has a very wide range of semantic coverage. We randomly check three data points around its medoid, and find that their meanings still differ to a large extent. In fidelity161, fidelity120, and fidelity158, their respective meanings are "loyalty," "honesty," and "accuracy," as we can see from the corresponding concordance lines in Fig. 6.

This is one of the limitations of token-based vector space modeling—the senses presented in the graph may not be what we are interested in, and the senses we intend to explore may be marginalized due to their low frequency or because of the indistinctiveness of their context words. Even if we manually remove instances of *fidelity* with irrelevant senses beforehand, we still believe that this endocentric blend is a counterexample to the hypothesis of semantic right-headedness, since the second source word *fidelity* is not easily recognizable from the blend even by contemporary native English speakers, and it only serves the purpose of filling in the syllable for rhyming motivation, as is speculated in the OED entry. Although *fidelity* is the morphosyntactic head of the full-form *wireless fidelity* since it is pre-modified by an adjective, the semantically more salient element for *wi-fi* should be *wireless*. In addition, this semantic left-headedness can also be explained by the fact that *wi-fi* uses radio waves, whereas the notion of *fidelity* is strongly tied to sound reproduction such as "a stereo *hi-fi*," "a *hi-fi* video." This indicates that the semantic head and morphosyntactic head of a blend may not necessarily align with each other.

Next, we examine *helilift* (< helicopter + lift), which has been suggested as a possible candidate for an exocentric blend. According to Bryant (1974: 175), *helilift* means "a group transported by helicopter," as in "Up in Quang Tin province, near Danang, a *helilift* of South Vietnamese paras, hoping to provoke a big battle, made contact with the Communists in a slough of serried hills, scuffled briefly but bloodily, then withdrew to regroup." (cited from *Time* 27, 9 April 1965). Thus, Mattiello (2013: 125) takes it as an exocentric blend, since it is neither "a helicopter" nor "a lift." However, if we closely examine the concordance lines in the TenTen corpus, we may easily find that its meaning has already changed. Apart from the capitalized *Heli-Lift* which is the name of an international incorporation founded in 1994, its new meaning—"lifting with a helicopter"—is quite compositional and transparent, as is shown in the three clear examples below. So its earlier meaning of "individuals being transported by a helicopter" is now extended to refer to the event itself—a typical part-for-whole metonymy. According to the OED, *helilift* can also be used as a verb meaning "to transport by helicopter," and *heli-* has already become a combining form, as in words like *helibus* (a helicopter with accommodation for a large number of passengers), *heliport* (a landing place for helicopters), *helidrome*, *helipad*, *helipod*, etc.

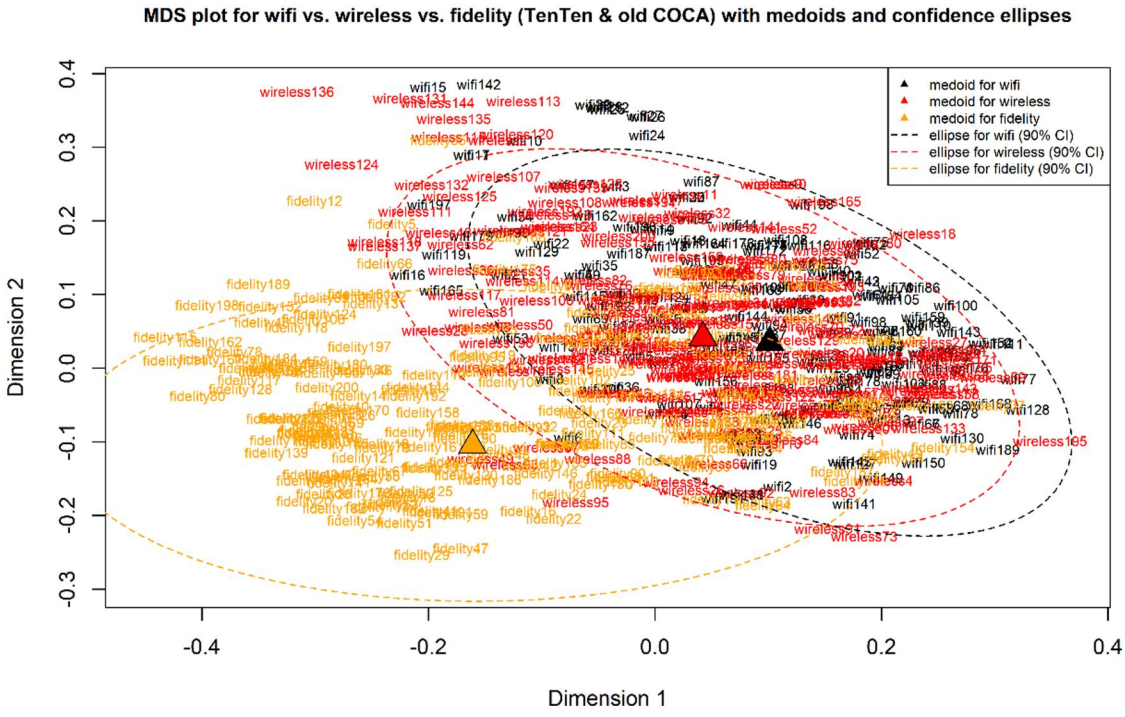

Figure 6. MDS plot for *wi-fi*.

	Full left context	Key word	Full right context
(2) fidelity161	A Supreme Court justice must meet the highest standard of legal excellence, while serving with humility and	fidelity	to our founding promise of equal justice under the law.
(3) fidelity120	… register with the Bar as a settlement agent under CRESPA, including the obtaining of the required surety and	fidelity	bonds, the completed registration form, and registration fee were not sent to the Bar.
(4) fidelity158	David Barnard's story of King David belongs to this genre. With	fidelity	to the English text and respect for current scholarship, he has written an extremely readable and very believable book.

(5) Because the 2nd Battalion, 5th Cavalry stealthily closed in the battlefield by foot instead of by **heli-lift**.

(6) … volunteers came forward and some of them worked independently of Living Oceans to collect and prepare debris for our **heli-lifts**.

(7) … the mission had to be aborted—and he was evacuated by **heli-lift** somewhere in the mid-Pacific.

Judging from the graphical features in Fig. 7, it would be more appropriate to classify *helilift* as a left-headed endocentric blend, since the semantic distance between *helilift* and *helicopter* is relatively closer. Of course, the "elevator" meaning of *-lift* may bias the results to some extent, but due to its compositionality in semantics, we still believe it is not a proper candidate for exocentric blend.

In sum, the blends whose graphical representations are different from the semantic categories we intuitively assign based on Plag's (2018) classification framework may be either due to the limitations of the vector space modeling itself or due to our incomplete understanding of their semantic structures.

4. Concluding remarks

This study has addressed the question of whether Plag's (2018) ternary division for English compounds

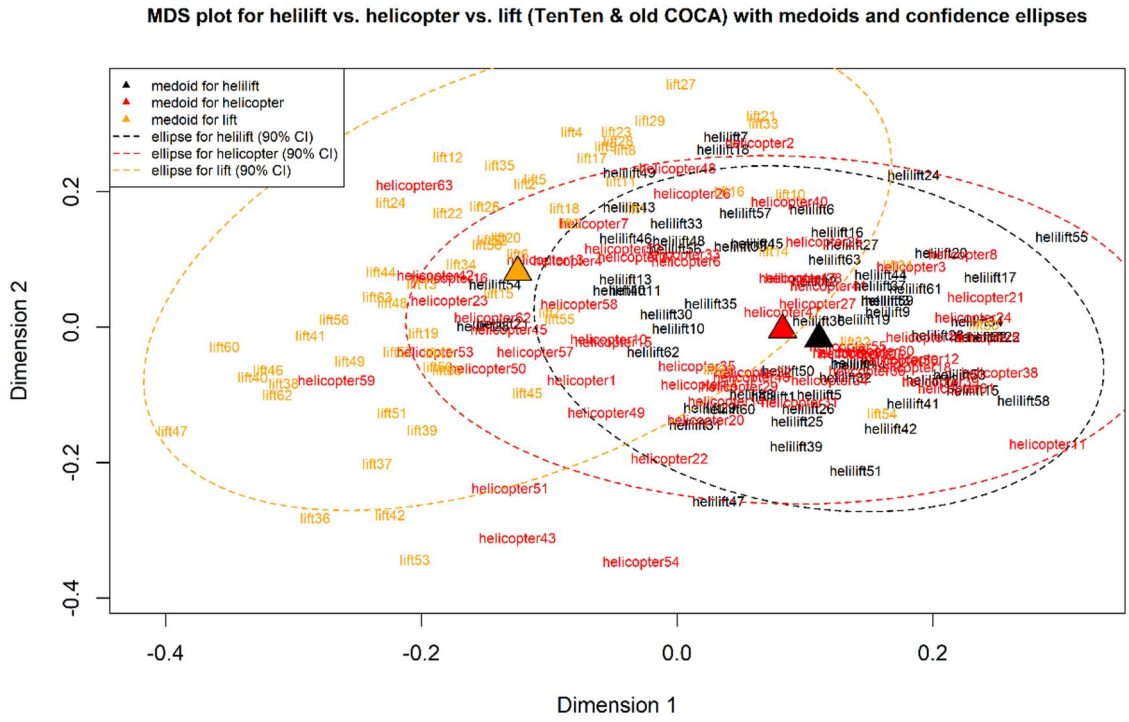

Figure 7. MDS plot for *helilift*.

based on semantic headedness can also be applied to English blends. Taking a distributional semantics approach, we operationalized the notion of semantic headedness in terms of distances between medoids in the distribution of a token-based semantic vector space. Our general conclusion is that token-based semantic vector space modeling can provide new insights into the semantic classification of English blends. For prototypical members of endocentric, exocentric, and coordinative blends, their distributional patterns in the corresponding MDS plots are quite different, not only in terms of the relative positions of the medoids and the degree of semantic overlapping in their 90 per cent confidence ellipses, but also in terms of the classification accuracy of the multinomial logistic regression model with two MDS dimensions as predictors. In addition, the hypothesis of semantic right-headedness only holds true for most endocentric blends.

For some blends, we found the semantic categories suggested by MDS graphs and those proposed by previous morphological experts are inconsistent due to the polysemy of the element(s). Inevitably, this is one of the limitations of distributional semantics methods: some senses may not necessarily be what we are interested in, and other senses that we are interested in may be underrepresented in the graph, since the context items of the blend or its source word(s) may be too infrequent or indistinctive to be retained.

Apart from that, the token-based vector space modeling also prompts us to reconsider the existing semantic categories for certain blends. For example, are *motel* and *boatel* indeed so different as to be classified into two distinct categories? Instead of a categorical approach, a prototype-based gradient approach may be a better solution to the semantic classification. However, it should be noted that whether the categories suggested by distributional semantics methods indeed have a psychological reality (cf Lehrer 1996) still awaits cross-validation through psycholinguistic experiments.

In conclusion, Plag's (2018) ternary division offers a useful framework for the semantic analysis of blends, and distributional semantics methods can provide new insights into their semantic behavior, thereby contributing toward extending morphological theory on the basis of quantitative corpus-based methods. We hope that our discussion can inspire further research on English word formation along these lines, such as whether the semantic right-headedness hypothesis is applicable to exocentric and coordinative compounds as well, whether the average classification accuracy values in logistic regression analysis are comparable

for each subtype of blends and compounds, and whether this token-based vector space modeling technique can be further extended to study the directionality of conversion with diachronic corpus data. To explore these guestimates, we need more workaholics.

Author contributions

Qingnan Meng (Data curation, Funding acquisition, Investigation, Resources, Visualization, Writing—original draft) and Martin Hilpert (Methodology, Project administration, Software, Supervision, Writing—review & editing)

Funding

This work was supported by the China Scholarship Council (Grant No. 202206575004) and the Fundamental Research Funds for the Central Universities in China [3132023329] ("Quantitative Linguistic Studies from the Perspective of Machine Learning," PI Qingnan Meng).

Notes

1. It should be noted that according to Renner (2019: 36), the last type is controversial in the morphological literature, since some scholars claim that such compounds have no semantic head. In this study, we mainly follow Plag's (2018) statement to view such equivalents in English blends as double-headed.
2. Other similar terms include "coordinate," "copulative," "associative," and "paradigmatic."
3. We appreciate the reviewers' comments highlighting that this is only a general tendency for AC blends, and there are exceptions such as *moped* (<motor + pedal), *parsec* (< parallax + second), and *rawin* (< radar + wind), the latter two mentioned in Renner (2023).
4. In an earlier version, Plag (2003) tries to combine both formal and semantic features in the classification of blends, depending on whether the corresponding full form exists and whether the two source words are in paradigmatic relations. Thus *boatel* and *motel* are classified differently, since "a *boatel* is both a boat and a hotel" (2003: 122), but a *motel* is not a motor. But this fails to take into account the polysemy of *boatel* which can mean either "a hotel reached by boat" or "a hotel on a boat" (Lehrer 2007: 123), let alone the legitimacy of the full form *motor hotel* in present-day English. It is perhaps due to the fact that sometimes it is hard to judge whether the full form of a blend is completely impossible that in Plag (2018) the less controversial terms AC-blends and AD-blends are adopted instead.
5. In the discussion part of this present research, we still stick to the narrow hyponymic definition to treat such blends as exocentric.
6. It should be noted that according to Adams (1973: 188), the distinction between affixes, compound-elements and splinters is not always clear-cut. For example, the element *-scope* in *biographoscope* is more of a splinter than in *diamondscope*, since the source word *microscope* underlies its formation. Consequently, *biographoscope* is more of a blend, whereas *diamondscope* is more of a neoclassical compound. Following this claim, we tend to perceive *newscast* and *docudrama* as blends, since *-cast* and *docu-* readily evoke their corresponding source words and are thus splinters still on their way to become combining forms.
7. It should be noted that Geeraerts (2017) argues for at least five differences between Harris's distributionalism and the contemporary distributional semantics approach, so it would be misleading to refer to them as the founding fathers of distributional semantics. But we believe his pursuit of methodological precision on a firm empirical basis at least indirectly inspired the current corpus-based distributional approaches.
8. The R codes that are used to create token-based vector spaces and for visualization are available from the authors upon request.
9. Medoid is the most centrally located object of each cluster, with minimum sum of distances to other points. In comparison with the mean, it is more robust to the outlier and can better represent the cluster center.
10. Even for morphosyntactic headedness, we can still find some exceptions, such as *chiefs of staff*, *fathers-in-law*, *passers-by*.
11. This explanation was offered by Elizabeth Closs Traugott in a personal email correspondence with the first author on Feb 9, 2024. But Traugott also reminds that which meal is the prototype may be culturally dependent.
12. See https://www.oed.com/.

References

Adams, V. (1973) *An Introduction to Modern English Word-formation*. London: Longman.

Algeo, J. (1977) 'Blends, a Structural and Systemic View', *American Speech*, **52**: 47–64.

Bat-El, O. (2006) 'Blend', in K. Brown (ed.) *Encyclopedia of Language and Linguistics* (Vol. 2), pp. 66–70. Oxford: Pergamon.

Bauer, L., Lieber, R., and Plag, I. (2013) *The Oxford Reference Guide to English Morphology*. Oxford: Oxford University Press.

Beliaeva, N. (2014) 'A Study of English Blends: From Structure to Meaning and Back Again', *Word Structure*, **7**: 29–54.

Beliaeva, N. (2019) 'Blending in Morphology', *The Oxford Research Encyclopedia of Linguistics (online)*. Oxford: Oxford University Press. Retrieved 15 July 2024, from https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-511.

Bloomfield, L. (1933) *Language*. New York, NY: Holt, Rinehart and Winston.

Bryant, M. (1974) 'Blends are Increasing', *American Speech*, **49**: 163–84.

Davies, M. (2004) 'BYU-BNC. Based on the British National Corpus from Oxford University Press', https://www.english-corpora.org/bnc, accessed 7 Feb. 2024.

Davies, M. (2008) 'The Corpus of Contemporary American English (COCA): 400+ Million Words', https://www.english-corpora.org/coca, accessed 7 Feb. 2024.

Dressler, W. (2000) 'Extragrammatical vs. Marginal Morphology', in U. Doleschal, and A. Thornton (eds.) *Extragrammatical and Marginal Morphology*, pp. 1–10. Munich: Lincom Europa.

Firth, J. R. (1957) 'A Synopsis of Linguistic Theory, 1930–1955', in J. R. Firth. (ed.) *Studies in Linguistic Analysis*, pp. 1–32. Oxford: Blackwell.

Geeraerts, D. (2017). Distributionalism, old and new. In A., Makarova, S. M., Dickey, and D., Divjak (eds), *Each Venture a New Beginning: Studies in Honor of Laura A. Janda*. Bloomington: Slavica, pp. 29–38.

Grlj, T. (2022) 'Blending as a Word-formation Process: A Comparative Analysis of Blends in English and French', *Journal for Foreign Languages*, **14**: 85–106.

Günther, F., and Marelli, M. (2022) 'Patterns in CAOSS: Distributed Representations Predict Variation in Relational

Interpretations for Familiar and Novel Compound Words', *Cognitive Psychology*, **134**: 101471.

Harris, Z. (1954) 'Distributional Structure', *Word*,**10**: 146–62.

Heylen, K., Speelman, D., and Geeraerts, D. (2012) 'Looking at word meaning. An interactive visualization of semantic vector spaces for Dutch synsets', *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pp. 16–24. Avignon: Association for Computational Linguistics.

Heylen, K. *et al.* (2015) 'Monitoring Polysemy: Word Space Models as a Tool for Large-scale Lexical Semantic Analysis', *Lingua*, **157**: 153–72.

Hilpert, M., and Correia Saavedra, D. (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**: 393–424.

Hilpert, M., Correia Saavedra, D., and Rains, J. (2023) 'Meaning Differences between English Clippings and their Source Words: A Corpus-based Study', *ICAME Journal*, **47**: 19–37.

Jakubíček, M. *et al.* (2013) 'The TenTen corpus family', *Proceedings of the 7th International Conference on Corpus Linguistics*, pp. 125–7. Lancaster: UCREL.

Kelly, M. (1998) 'To 'Brunch' or to 'Brench': Some Aspects of Blend Structure', *Linguistics*, **36**: 579–90.

Kiela, D., and Clark, S. (2014) 'A systematic study of semantic vector space model parameters', *Proceedings of EACL 2014, Second Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, pp. 21–30. Gothenburg, Sweden: Association for Computational Linguistics.

Kilgarriff, A. *et al.* (2014) 'The Sketch Engine: Ten years on', *Lexicography: Journal of ASIALEX*, **1**: 7–36.

Lehrer, A. (1996) 'Identifying and Interpreting Blends: An Experimental Approach', *Cognitive Linguistics*, **7**: 359–90.

Lehrer, A. (2007) 'Blendalicious', in J. Munat (ed.) *Lexical Creativity, Texts and Contexts*, pp. 115–33. Amsterdam, the Netherlands: John Benjamins.

Marchand, H. (1969) *The Categories and Types of Present-day English Word-formation: A Synchronic-diachronic Approach*. München: C.H. Becksche Verlagsbuchhandlung.

Mattiello, E. (2013) *Extra-grammatical Morphology in English: Abbreviations, Blends, Reduplicatives, and Related Phenomena*. Berlin, Germany: Walter de Gruyter.

Oxford University Press. (2020) *Oxford Advanced Learner's Dictionary*, 10th edn. Oxford: Oxford University Press.

Plag, I. (2003) *Word-Formation in English*. Cambridge: Cambridge University Press.

Plag, I. (2018) *Word-Formation in English*, 2nd edn. Cambridge: Cambridge University Press.

R Core Team. (2023) *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. https://www.r-project.org (accessed 7 February 2024).

Renner, V. (2014) 'A Study of Element Ordering in English Coordinate Lexical Items', *English Studies*, **95**: 441–58.

Renner, V. (2015) 'Lexical Blending as Wordplay', in A. Zirker, and E. Winter-Froemel (eds.) *Wordplay and Metalinguistic/Metadiscursive Reflection: Authors, Contexts, Techniques, and Meta-Reflection*, pp. 119–34. Berlin, Germany: De Gruyter.

Renner, V. (2019). French and English lexical blends in contrast. *Languages in Contrast*, **19**(1): 27–47.

Renner, V. (2023) 'Blending', in P. Ackema, S. Bendjaballah, E. Bonet, and A. Fábregas (eds.) *The Wiley Blackwell Companion to Morphology*, pp. 1–19. New York, NY: JohnWiley & Sons.

Schmid, H. (1999) Improvements in Part-of-Speech Tagging with an application to German. *In: Armstrong, S., Church, K., Isabelle, P., Manzi, S., Tzoukermann, E., Yarowsky, D. (eds) Natural Language Processing Using Very Large Corpora. Text, Speech and Language Technology*, Vol. 11, pp. 13–25. Dordrecht: pringer.

Schütze, H. (1992) 'Dimensions of meaning', *Proceedings of the 1992 ACM/IEEE Conference on Supercomputing (Supercomputing '92)*, pp. 787–96. Washington, DC: IEEE Computer Society Press.

Suchomel, V. (2020) 'Better Web Corpora for Corpus Linguistics and NLP', PhD dissertation, Masaryk University, Brno, Czechia.

Tarasova, E., and Beliaeva, N. (2020) On twittizens and city residents: Experimental study of semantic relations in English compounds and blends. *The Mental Lexicon*, **15**: 21–41.

Thurner, D. (1993) *Portmanteau Dictionary: Blend Words in the English Language, Including Trademarks and Brand Names*. Jefferson, NC: McFarland.

Tomić, G. (2019) 'Headedness in Contemporary English Slang Blends', *Lexis: Journal in English Lexicology*, **14**: 1–24.

Turney, P. D., and Pantel, P. (2010) 'From Frequency to Meaning: Vector Space Models of Semantics', *Journal of Artificial Intelligence Research*, **37**: 141–88.

Appendix Table A.1. Dataset used in this study for semantic vector space modeling

Semantic type by narrow hyponymic definition	Blend	Classification accuracy (%)	Right-headedness	Source word 1 (sw1)	Source word 2 (sw2)	Distance 1 (blend-sw1)	Distance 2 (blend-sw2)	d1/d2
Coordinative	moped	70.17	no	motor	pedal	0.192551	0.26077	0.738394
coordinative	chunnel	66.53	yes	channel	tunnel	0.224427	0.185698	1.208559
coordinative	guestimate	66.17	no	guess	estimate	0.155736	0.181615	0.85751
coordinative	boatel	65.27	yes	boat	hotel	0.278854	0.12003	2.3232
coordinative	demopublican	63.84	yes	democratic	republican	0.2718	0.151387	1.795401
coordinative	modem	62.33	no	modulator	demodulator	0.28326	0.300566	0.942421
coordinative	airtel	62	yes	air	hotel	0.187549	0.170398	1.100652
coordinative	frenemy	59	no	friend	enemy	0.042845	0.257686	0.166267
coordinative	vog	57.67	no	volcanic	fog	0.128122	0.237268	0.539989
coordinative	compander	57.52	yes	compressor	expander	0.241633	0.026198	9.223349
coordinative	smog	56.5	no	smoke	fog	0.075692	0.257315	0.294161
coordinative	transceiver	55.67	no	transmitter	receiver	0.172946	0.182866	0.945748
coordinative	edutainment	55.33	yes	education	entertainment	0.151245	0.096758	1.563126
coordinative	shoat	54.78	yes	sheep	goat	0.15139	0.099689	1.518619
coordinative	boost	54	no	boom	hoist	0.063928	0.229134	0.278998
coordinative	geep	51.99	yes	goat	sheep	0.174626	0.147091	1.18719
coordinative	stagflation	48.33	no	stagnation	inflation	0.045482	0.095892	0.474308
coordinative	happenstance	47.21	no	happen	circumstance	0.05715	0.125353	0.455918
coordinative	himbo	47.2	yes	him	bimbo	0.122942	0.082999	1.481242
coordinative	adorkable	42.14	yes	adorable	dorky	0.1026953	0.0785544	1.307314
coordinative	brunch	46.5	no	breakfast	lunch	0.052466	0.103726	0.505812
coordinative	Spanglish	46	yes	Spanish	english	0.1426822	0.0617837	2.3093810
coordinative	spork	42.33	yes	spoon	fork	0.135954	0.09804	1.386722
endocentric	breathalyzer	80.67	no	breath	analyzer	0.321235	0.359086	0.894591
endocentric	motel	75	yes	motor	hotel	0.349837	0.226703	1.543151
endocentric	sitcom	66.32	yes	situation	comedy	0.338467	0.059912	5.649437
endocentric	floatel	64.56	yes	float	hotel	0.355242	0.11161	3.182881
endocentric	flexitarian	62.67	yes	flexible	vegetarian	0.352979	0.121716	2.90003
endocentric	wi-fi	60.83	no	wireless	fidelity	0.059372	0.297101	0.199838
endocentric	netizen	60.83	no	network	citizen	0.135747	0.194902	0.696488
endocentric	infomercial	60.83	yes	information	commercial	0.203343	0.094236	2.157803
endocentric	sci-fi	60.06	yes	science	fiction	0.260606	0.151817	1.716574
endocentric	workaholic	61.5	yes	work	alcoholic	0.206931	0.116893	1.770265
endocentric	newscast	58.67	yes	news	broadcast	0.224491	0.208754	1.075385
endocentric	mockumentary	57	yes	mock	documentary	0.189431	0.099032	1.912830
endocentric	chocoholic	56.83	no	chocolate	alcoholic	0.028194	0.348691	0.080855
endocentric	webinar	54.83	yes	web	seminar	0.110825	0.083373	1.329262
endocentric	blog	67.83	yes	web	log	0.339294	0.295581	1.147885
endocentric	glamping	53.95	yes	glamour	camping	0.124435	0.089925	1.383769
endocentric	podcast	53.33	yes	ipod	broadcast	0.187967	0.017082	11.00413
endocentric	telethon	52.67	yes	television	marathon	0.148665	0.039985	3.718048
endocentric	telecast	52.67	yes	television	broadcast	0.206889	0.144724	1.429547
endocentric	shopaholic	50.17	no	shop	alcoholic	0.038384	0.143791	0.266943
endocentric	beefburger	49.83	yes	beef	hamburger	0.109059	0.10859	1.004326
endocentric	bromance	48.67	no	brother	romance	0.079087	0.121222	0.652417
endocentric	staycation	47.83	yes	stay	vacation	0.12734	0.037163	3.42649
endocentric	medicare	47	yes	medical	care	0.139084	0.029277	4.750586

(continued)

Appendix Table A.1. (continued)

Semantic type by narrow hyponymic definition	Blend	Classification accuracy (%)	Right-headedness	Source word 1 (sw1)	Source word 2 (sw2)	Distance 1 (blend-sw1)	Distance 2 (blend-sw2)	d1/d2
endocentric	tragicomedy	46	yes	tragic	comedy	0.085518	0.059039	1.448499
endocentric	cheeseburger	44.5	yes	cheese	hamburger	0.116036	0.087984	1.318833
endocentric	docudrama	41.67	no	documentary	drama	0.014285	0.116207	0.122927
exocentric	FedEx	86.5	yes	federal	express	0.440052	0.402791	1.092505
exocentric	bionic	75.67	no	biological	electronic	0.270763	0.273859	0.988694
exocentric	fortran	69.67	yes	formula	translation	0.301701	0.265727	1.135383
exocentric	pixel	69	yes	pix	element	0.340640	0.188505	1.807062
exocentric	helilift	53.97	no	helicopter	lift	0.031273	0.255085	0.122597

Digital Scholarship in the Humanities, 2024, 39, 1075–1091
https://doi.org/10.1093/llc/fqae050
Full Paper