METAPHOR AND SYMBOL
https://doi.org/10.1080/10926488.2026.2627594

Routledge
Taylor & Francis Group

Two Context Effects on Metaphor Usage. A Computational, Corpus-Based Approach

Barend Beekhuizen and Claudia Raihert

University of Toronto Mississauga

ABSTRACT

This paper studies the relation between metaphor use and its linguistic and extralinguistic contexts. We propose a novel computational method, using token-level distributional semantic models and cluster analysis, to identify different metaphorical "senses" of a lexical item (e.g. *cascade* as 'a sequence of events, each causing the next' versus 'a diagonally flowing array of matter') and validate it qualitatively on a large corpus of American English fiction and newspaper texts (approximately 100 million words each). Next, we use the identified sense clusters to study how metaphor use varies as a function of extralinguistic context, in this case: genre. In a regression analysis, we model the genre bias of the sense clusters on the basis of three factors motivated by accounts of the interaction between metaphor and genre, namely sense concreteness, fixedness, and valence. Studying the metaphorical senses of 45 lexical items from the source domains of landscape and weather, we find that newspaper-biased senses involve more abstract target domains than fiction-biased ones. These results show that understanding the relation between linguistic contexts and metaphorical senses can support insight in usage variation across extralinguistic contexts, and, more broadly, that computational methods allow for novel ways to study the context-dependent nature of metaphor use.

Introduction

The same lexical item can have multiple, rather distinct, metaphorical meanings (Deignan, 1999; Hanks, 2004). Consider the following two examples (from the Corpus of Contemporary American English; Davies, 2008), with the metaphorical token underlined:

(1) The Osp-A protein in a vaccine could start a similar immunological cascade and cause arthritis months or even years after inoculation.
(2) So now Della's beautiful hair fell about her, rippling and shining like a cascade of brown waters.

In the former, the cascade metaphorizes a stepwise, causal relation between a protein and arthritis, whereas in the latter, any sense of causality is absent, and perceptual properties of source domain ('ripple', 'shine') are drawn upon to characterize someone's hair. How do recipients know which metaphorical meaning is intended, and how do senders formulate their messages to ensure the recipients arrive at the right interpretation? In the present study, we consider two components of an answer to this question: the linguistic context of the metaphorically used word, and its extra-linguistic context in the form of the type of text (e.g., news, fiction, also referred to as the *genre*) that the

CONTACT Barend Beekhuizen barendbeekhuizen@gmail.com Department of Language Studies, University of Toronto Mississauga, Mississauga, Canada

metaphor occurs in. To this end, we present a generalizable and fully automated computational model, and apply it to metaphorical usages of 98 lexical items in a corpus of approximately 200 million words.

As we know from corpus linguistics (Firth, 1957; Hanks, 2013; Sinclair, 1996), distinct linguistic-contextual patterns of the same lexical item tend to reflect distinct lexical semantic interpretations. For instance, the verb "to bother" has a different meaning in the expression *Don't bother* ('don't spend any effort') as opposed to the transitive usage in *It bothers me* ('it annoys me'), and language users distinguish between these meanings based on the linguistic environment of the verb. Following Hanks (2004), we argue that this property holds for distinct metaphorical interpretations as well. Here, we use computational techniques to test whether recurrent patterns of linguistic context allow us to automatically identify distinct metaphorical meanings associated with the same lexical item.

These methods moreover allow us to investigate how different kinds of metaphor are distributed across different text types. For instance, is it the case that the metaphorical use of *cascade* exemplified in (2) is found in both fiction and newspapers at a similar rate? In line with research on the interaction between metaphor and genre (e.g., Cameron & Maslen, 2010; Deignan et al., 2013; Goatly, 1997, 2011), we expect that the kinds of metaphor that are thought of as appropriate and effective in a given genre will depend on the communicative goals of the genre and its associated stylistic conventions and practices. Our second goal is hence to consider how genre modulates the frequency of occurrence of different metaphorical senses of a word. We investigate several communicative factors that have been associated with newspapers and novels, and demonstrate quantitatively that one of these factors, namely the imagistic quality of metaphors, indeed presents a reliable association with text type.

As such, our research hopes to contribute to a discourse-oriented perspective on metaphor by providing computational methods to study the situatedness or context-dependence of metaphor usage in novel ways. While rooted in this tradition, our paper also asks a fundamentally cognitive question, namely: what does a language user need to keep track of in order to be a functioning member of a speech community – that is: a question regarding the nature of the information that language users draw on when using metaphor.

Metaphor, practice, and usage

A lexicological perspective on metaphor use

Our first goal is to provide evidence for the position that lexical items have different, conventional, metaphorical senses, and that those senses can be fruitfully studied through the linguistic contexts the lexical items occur in. Issues of metaphorical polysemy have mostly come up in discussions about the relative contribution of conceptual metaphors on the one hand (i.e., general mappings between conceptual domains that productively generate novel expressions, as suggested by Conceptual Metaphor Theory; Lakoff & Johnson, 1980) and lexical factors, in particular: conventionality (Deignan, 1999; Svanlund, 2007) on the other. Deignan (1999; see also Deignan, 2005) uses the fact that metaphorical usages occur in distinct sets of linguistic environments to identify challenges for metaphor researchers, such as the characterization of a metaphor as active or "dead," and to raise issues with the Invariance Principle of Conceptual Metaphor Theory (Lakoff, 1990, 1993), that is: the expectation that a conceptual metaphor allows for the free, unconstrained, generation of novel instantiations of a source-target domain mapping.

Deignan illustrates this point with temperature terms: while instances of *My geography's not all that hot* are prevalent in usage, examples of other temperature adjectives (e.g., *My geography is not all that cool/cold/warm*) with similar target domain meanings are notably absent from text corpora. This means that the Invariance Principle overpredicts metaphorical usage (see Svanlund, 2007 for a similar argument). An attempt to bring these facts in line with the Invariance Principle comes from Sullivan (2013), who identifies differences in the lexical meanings of the source domain words that may explain the differences in mappings to target domains, thus motivating the variation in terms of properties of the source domain meanings of the lexical items. Gibbs (2017, pp. 120–123) presents a balanced review

of this discussion, concluding that, while such lexical variation does not challenge the Invariance Principle per se, it does indicate that the generative principles of Conceptual Metaphor Theory do not operate in an unconstrained way. Regardless of the consequences for the Invariance Principle, what is relevant for our purposes is the insight that the lexical items *hot, warm*, *cool* and *cold* all have their own (conventional) sets of metaphorical senses, whose analogous metaphorical senses are not necessarily (conventionally) present in the other members of the same lexical field. This means that metaphorical senses are associated with lexical items (and their non-metaphorical senses) at least in part through convention, rather than being exclusively licensed by more general conceptual metaphors, a claim that is not very controversial (cf. Gibbs, 2017, p. 123) but whose extent and explanatory power has been only considered in a limited way. Our paper contributes a generalizable method for studying such patterns of polysemy, which can in turn further contribute to the kinds of theoretical discussions as laid out above.

Along similar lines, Hanks (2004) inquires into the link between metaphorical polyfunctionality and syntagmatic contexts, that is: the (grammatically) adjacent linguistic context of a metaphorically used word. In particular, he considers adjacent lexical environments known in the corpus-linguistic tradition as "collocations," syntactic environments known as "colligations" and the co-presence of lexical items in the wider contextual environment known as "semantic prosody" (cf. Sinclair, 1991, pp. 70–75). The study of these environments, Hanks proposes, allows us to identify (1) whether the lexical item is used metaphorically, and (2) if so, what metaphorical contribution to the meaning of the utterance can be derived from it. When organizations *weather a storm*, we know that *storm* is likely used metaphorically, and that the *storm* pertains to negatively impacting actions' (that are 'withstood', or: *weathered*), whereas if someone is *greeted by a storm of applause*, we can identify *storm* as being metaphorically used to refer to "a large quantity." Different syntagmatic contexts are thus associated with different aspects of the meaning of a lexical item, and this association may become conventionalized over time, as in the case of *weather a storm*, where the metaphor user can no longer cancel the inference that *storm* alludes to negatively impacting actions.

Thus, as Hanks (2004) argues, contextual properties are informative of lexical senses. Metaphor, then, is a special case of lexical usage that has been elevated to a *sui generis*, presumably due to its status as a rhetorical form, but that, from a lexicologer's perspective, is arguably just another case of pragmatically informed lexical usage (see Sperber & Wilson, 2008). In Hanks' theory of lexical norms and exploitations, conventionalized metaphors are part of the norm of usage of lexical items. On the other hand, active metaphorizing involves exploiting such norms (by placing them in unexpected syntagmatic environments) to arrive at a non-conventional message.

Metaphorical variation and genre

The second point of this paper is that the usage patterns of metaphorical senses are not only explained by reference to linguistic contexts (collocations, colligations, and semantic prosodies) but also to extra-linguistic contexts. Specifically, we draw on research into the relation between metaphor and genre (or *text type*) to consider how extra-linguistic context affects the metaphorical potential of different source domains.

We find an early extensive treatment of the relation between metaphor and genre in Goatly (1997, 2011). Goatly argues that metaphors fulfill different communicative functions: metaphorical language can be deployed to fill lexical gaps or to explain and model a poorly understood phenomena, but also to express emotional attitudes, to decorate, disguise and exaggerate, or to enhance memory of or foreground particular contents. Goatly subsequently finds these functions to be non-uniformly distributed across different genres, and suggests that the greater association of certain functions of metaphor use with certain genres is due to the properties of the genre. He analyzes these properties in terms of Halliday and Hasan's (1985) Register theory. The latter presents three dimensions along which situated language use varies: the Field (the nature of the social action taking place), Tenor (the participants in the action), and Mode (the role language plays). Goatly then observes through a corpus

study that the metaphorical functions are distributed across genres in ways that follow from the Field, Tenor and Modes of the genres.

The detailed monograph of Deignan et al. (2013) also deals with the notions of metaphor (and other figurative language) and genre. Like Goatly, they draw on Haliday & Hasan's (1985) Register theory, complemented with Swales' (Swales, 1990) model of genre which emphasizes a different set of dimensions (viz. the discourse community, purpose, and rhetorical structure of the genre). These dimensions are then applied to a series of case studies analyzing metaphor use in particular genres (e.g., figurative language used by daycare workers). These case studies, together with many more that can be found in the literature (e.g., (Caballero, 2003, Caballero, 2017; Dorst, 2015; Fludernik, 2019; Kearns, 1987, Porto & Romano, 2013; Semino, 2011; Semino et al., 2013), present a fertile ground for hypothesizing that genre does indeed interact with *how* and *to what ends* metaphorical language is deployed.

Research goals

The two lines of research we presented have primarily been developed through (mostly qualitative) case studies involving individual lexical items or small sets of lexical items. To further develop our understanding of the relation between metaphor use, and linguistic and extra-linguistic context, we propose a computation-driven corpus study. This methodology will allow us to test previous observations about metaphorical variation at a larger scale than previous studies and draw conclusions based on complementary qualitative and quantitative evidence.

We start off from Hanks' (2004) insight that metaphorical senses are lexical norms of lexical items, and that they can be distinguished from other senses through the consideration of the syntagmatic context. We build on the traditional corpus-based methodology of studying collocations and colligations by proposing an automated procedure for identifying groups of (syntagmatically similar) usages of a particular lexical item, on the basis of distributional semantic representations of the usages and validating it qualitatively.

Next, we show how instances of these metaphorical senses are not distributed uniformly across the genres of national newspapers and popular fiction in American English. We argue that the distribution of metaphorical senses across these genres can be explained to a substantial extent in terms of properties of the genres. As such, we provide a large-scale validation of the more qualitatively supported finding that metaphor and genre interact in structured ways. However, in line with the findings from the first part of the paper, we will argue that there also needs to remain space in the explanatory model for more arbitrary, lexical, associations.

Data, extraction methods, and a first exploration

The corpus and item selection

We used two segments of the Corpus of Contemporary American English (COCA; Davies, 2008), one comprising newspaper texts and the other popular fiction texts (approximately 100 million words each, written between 1990 and 2012). To narrow down the comparison between the two genres, we extracted nouns denoting landscape and weather concepts. We chose these two semantic domains because they involve concrete concepts that have a substantial expected metaphorical propensity owing to their experiential "basicness" (e.g., a *mountain* of papers, a *storm* of protests). While there are other semantic domains that meet this criterion (e.g., bodily concepts), we restricted ourselves to two domains as a feasible starting point for the manual validation of the method.

We first determined a set of lexical items involving landscape and weather concepts by extracting all instances of these two semantic domains in WordNet, a large lexical database for English that organizes words in terms of their synonymic and taxonomic relations (Fellbaum, 1998; Miller, 1995). For example, *river* is a hyponym, or a subcategory, of *body of water*, its hypernym. The

hypernymic senses that were manually selected for landscape descriptors were 'body of water' (sense 1), 'geological formation' (sense 1), 'land' (senses 2 and 4), and 'biome' (sense 1); and for weather phenomena, 'weather' (sense 1) alone. The lexical items that were hyponyms of these senses were extracted automatically. We omitted proper names (e.g., *Skagerrak*) and compounds (e.g., *international waters*). Words with a frequency below 100 in our combined corpus (e.g., *divot*, $N = 74$) were discarded, as well as words whose first dictionary entry involved a non-metaphorical sense from a different domain (e.g., *gut* having 'bowels, entrails' as sense 1, and 'a narrow passage' as sense 5).[1] This selection process resulted in 72 lexical items for landscape concepts, and 26 lexical items for weather concepts (see Table A1 in Appendix 1 for the full list). We extracted all sentences that involved instances of the 98 lexical items from the news and fiction COCA samples ($N = 399{,}927$).

Automated extraction methodology

To separate metaphorical usages from non-metaphorical ones at this scale, manual filtering would take prohibitively much time. Instead, we automatically identified which tokens were metaphorical using a computational metaphor detection model, MelBERT (Choi et al., 2021), built on the pre-trained contextualized language model BERT, which represents instances of words as numerical vectors (see below belowfor a more extensive introduction). MelBERT calculates the probability of a word being metaphorical based on the difference between (i) the meaning of the word in context and its generalized meaning, and (ii) the meaning of the word in context and the overall meaning of the (rest of the) sentence. Hence, MelBERT's predictions correspond to how "typical" a given context is for the target word. When trained and tested on the VU Amsterdam Metaphor Corpus (VUAMC) (Krennmayr & Steen, 2017), MelBERT is reported to perform better than other models (see Ge et al., 2023, for an overview): given a portion of the VUAMC that it has not seen before, MelBERT will accurately predict whether a word is metaphorical 77% of the time and catch 81% of the annotated metaphors (Choi et al., 2021). Notably, this means that while MelBERT is a somewhat blunt instrument (only picking up on general contextual atypicality as a predictor of metaphoricity), it achieves satisfactory performance on the test data. However, we note here that it is not guaranteed for computational models to perform well when tested on a different dataset than the one they were trained on. Hence, while MelBERT's performance on the VUAMC is satisfying, it is useful to assess the model's accuracy on our current dataset. Similarly, the use of contextual atypicality as a predictor may lead to recurrent false positives in cases where the context is more generally atypical, or creative, as we will see below.

Exploring this method with our data, we found three groups of cases that, upon inspection, are rarely metaphorical, but were nonetheless labeled as metaphorical by MelBERT. First, there are lexicalized compounds where the erstwhile compositional meaning typically did not involve metaphor but does reflect a contextually uncommon usage of the word, hence making MelBERT classify it as a metaphor (e.g. *pit bull* or *iceberg lettuce*, where, etymologically, the compounds express a specific dog associated with a literal pit for bullfighting, and lettuce associated with transport on metonymic icebergs). Second, we found proper names labeled as metaphorical (e.g. *Alabama*, *Gulf Shores*), and, third, rich descriptions where the target word was not metaphorical either (*rushing with mimic roar over the flat meadows beside it*). To improve accuracy, we implemented three heuristics to exclude these cases. We automatically labeled as non-metaphorical instances where the source domain noun was modifying another noun (e.g., *pit bull*). We also excluded capitalized nouns, except when following a punctuation sign. Finally, we excluded cases where the source domain noun (e.g., *pit*) was in a prepositional phrase (e.g., *He's a yard from the edge of the pit now*), unless it was modified by another noun or PP (e.g., *the bad feeling churning in the pit of her stomach*). This heuristic was more particularly targeted at the landscape terms in our data set, given they tended to indicate (literal)

[1] Using the Merriam-Webster (n.d.) for the individual words, and going by the assumption that the Merriam-Webster entries are mostly organized by usage frequency, see https://www.merriam-webster.com/grammar/how-to-use-the-dictionary (visited April 25, 2024).

Table 1. Number of metaphorical and non-metaphorical tokens (as classified by MelBERT) in the two COCA subcorpora (newspaper, fiction) for the two metaphorical source domains (landscape, weather).

	Newspaper		Fiction	
	Landscape	Weather	Landscape	Weather
non-metaphorical	76,679	16,126	116,449	46,968
metaphorical	3,875	3,332	6,605	5,181

locations following a preposition. Further examples and a more detailed rationale is presented in Appendix 2.

When evaluated on a manually labeled test set of 105 metaphorical and 105 non-metaphorical tokens, the combination of MelBERT and the heuristics led to an accuracy of metaphor identification, or *precision* (0.64) and completeness or *recall* (0.95), which we believe is sufficient for our procedure (though a substantial amount of false positives remains – which, as we will see, tend to be grouped together in the clustering procedure). The code implementing the extraction method, together with the evaluation of the automatic extraction procedure are given in the Supplementary Materials. An overview of the extraction statistics can be found below in Table 1.

Finding groupings of metaphorical usages

Our first research goal is to determine if we can study metaphorical polysemy in an automated, and therefore generalizable, way by looking at patterns of linguistic contexts. In this section, we introduce and validate a clustering method that uses such contextual information to find groupings of contextually similar tokens.

We take up two very similar landscape terms, *cascade* and *waterfall*, and demonstrate the syntagmatic properties of their clusters that plausibly underlie their closeness in their numerical representation (see below), and discuss more qualitatively how clusters correspond to metaphorical senses. An analysis of the metaphorical senses of two lexical items with very similar non-metaphorical senses further allows us to assess the insight (Deignan, 1999; Hanks, 2004), that metaphorical mappings are *linguistic* phenomena, governed not only by conceptual-metaphorical potential, but equally by lexical conventions. An overview of the clusters alongside examples for all the other lexical items is given in the Supplementary Materials.

Cluster analysis over token-level distributional semantic representations

There is a small literature on computational approaches to metaphor interpretation in context, that is: automated approaches to determining or providing a helpful approximation of what the intended sense of a metaphor is (see Reid & Katz, 2018 for a recent overview). Here, we use a computational approach based on similar principles to the models presented in Reid and Katz (2018), but leveraging the power of contemporary distributional semantic models, which are trained on billions of words of text. More precisely, we use a clustering approach similar to Chronis and Erk (2020) to identify coherent groups of metaphorical usages of each of the lexical items. Figure 1 schematically represents the procedure. First, for every token of a word identified to be metaphorical, we use the pre-trained parameters of BERT (Devlin et al., 2018, using the bert-base-uncased model of the transformers library: Wolf et al., 2020) to retrieve the (768-dimensional) vector representing the token on the final hidden layer of the model (henceforth: the contextual vector of the token). Steps (i) and (ii) in Figure 1 depict this step: a sentence is tokenized (made into BERT-compliant tokens, e.g., splitting *bankruptcies* into *bankrupt* and *##cies*) and each of the tokens is subsequently predicted on the basis of all the other tokens in the sentence, through a layered neural network capturing the interactions between all tokens in predicting each target token. The resulting representations on each of the layers can next be used as representations of each word

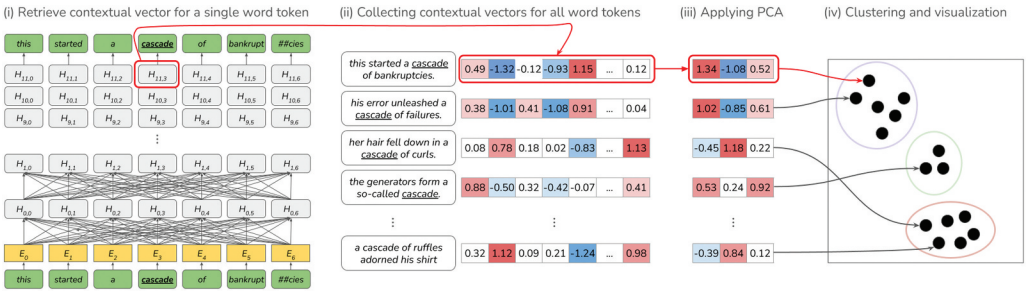

Figure 1. Schematic representation of the contextual vector extraction, dimensionality reduction, and clustering. The bottom layers of the BERT network in panel (i) indicate the complete connectedness of all nodes on one layer to all nodes on the next; this holds true of the upper three layers shown as well, but the arrows were omitted to reduce visual clutter.

token in context. Here, we use the final hidden layer for the target metaphorical token (the circled $H_{11,3}$ in the figure), as it is thought to represent the contextualized lexical semantic information about that token (cf. Chronis & Erk, 2020; Devlin et al., 2018; Ethayarajh, 2019).

An important property of these representations is that word tokens with similar contexts have similar vectors, which allows us to group together contextually-similar tokens of a lemma. Figure 1, step (ii) illustrates this: usages of *cascade* with more similar contexts (e.g., abstract events in the *of*-phrase and causal verbs like *start* and *unleash* in the first two examples) have more similar vectors (witness the color coding of the vectors in panel (ii) of Figure 1). The mechanical reason for this is that in similar contexts, certain sets of words are expected to occur that are not expected to occur in other contexts (e.g., *chain* and *onslaught* for the first two examples; *garland* and *blanket* in examples 3 and 5), which in turn leads to different contextual vectors in the training procedure. Given that contextual similarity can, per the Distributional Hypothesis (Firth, 1957; Lenci & Sahlgren, 2023), be thought of as a proxy of conceptual similarity, these groupings are assumed to represent conceptually-similar metaphorical usages of a lexical item. Indeed, Chronis and Erk (2020) deploy a similar method, and argue for its cognitive validity by showing that the inference of such clusters leads to a superior fit (compared to models without sense-like clusters) for word-level similarity judgments. Here, we present an approach similar to Chronis and Erk (2020), that is: by clustering over the contextualized vectors, but suggest two methodological improvements.

First, we apply dimensionality reduction, here: Principal Component Analysis (PCA; Hotelling, 1933) to the set of contextual vectors of a metaphorically used lexical item, capturing the main patterns of variance in the data in some substantially lower number (step (iii) in Figure 1). We did so for two reasons: first, clustering in high-dimensional spaces is fraught due to the curse of dimensionality (i.e., in high-dimensional spaces, everything tends to be far from everything, and as such distance becomes less meaningful). Second, these are all token-level representations of the same lexical items and as such much of the 768-dimensional vector space is redundant (see e.g., the analysis of the variance in these representations in Ethayarajh, 2019). Across lexical items, the PCA turned up between 0 and 4 components with an explained variance ratio of 5% or more, and lexical items with fewer than 2 usable components were excluded.

Second, we use a clustering algorithm that does not require us to set a specific number of clusters to be extracted, since we do not know what the right number of senses is, and the inference of meaningful senses is critical to the lexicological component of this work. Specifically, we used Bayesian Gaussian Mixture Model (GMM, as implemented in sklearn; Pedregosa et al., 2011). We used the default parameters except for setting the maximal number of iterations to 500 and the number of seed clusters to 20, with both settings leading to more stable and interpretable clusters. Figure 1, step (iv) shows how the dimensionality-reduced token-level contextual vectors are grouped together, visualizing them spatially in a *t*-SNE representation (van der Maaten & Hinton, 2008) that groups together in two dimensions higher-dimensional vectors that are similar to each other. The cluster analysis led to the

grouping of all the tokens for each lexical item into a set of clusters (between 1 and 11, and modally 4, clusters per lexical item). Given that the cluster analysis leads to hard-to-interpret results on lexical items with low numbers of tokens, we further omitted lexical items with fewer than 50 metaphorical tokens ($N = 51$). This left us with 45 total nouns (36 landscape nouns and 9 weather nouns). As our computational approach involves some design choices, which furthermore differ from a similar approach proposed by Chronis and Erk (2020), a supplemental analysis of variants of the clustering approach is reported in Appendix 3.

Cascade

To validate the clustering, we first consider the lexical item *cascade*, for which we have 483 metaphorical tokens, each represented as a 768-dimensional vector. Running the GMM, we find five clusters. Figure 2 shows a dimensionality-reduced representation of the tokens, with the genres and clusters indicated. The full sets of sentences grouped in each cluster can be found in Supplementary Materials. Two properties of particular interest turned out to be the nature of postnominal *of*-phrases, which are frequently present and denote the metaphorical target domain (cf. Steen et al., 2010), and the syntagmatic embedding of the noun phrase containing *cascade* (e.g., the kind of predicate it is in a syntactic relationship with), as these dimensions of variation can point to differences in meaning between clusters. Table 2 presents our analysis of the five clusters, alongside several examples.

Cluster 1 is characterized by the frequent occurrence of postnominal *of*-phrases (present in 81% of the tokens) which contain predominantly nouns denoting natural or perceptual phenomena that occur in stream-like sequences, such as *rivulets*, *lights* and *sparks* (as in example (b)), but also sounds that can be construed in that way (*thumps*, *thunder*, *boos*) and objects making sounds when moving in a cascade-like fashion (*rocks*, *glass and crockery and books*). The noun *cascade* is found to be modified by adjectives denoting auditory properties such as *soft, roaring, thundering, snarling* and descriptors of the consistency such as *vitreous, steady, twisting*. The metaphorical cascade is often in the subject or manner adjunct of a motion verb (28/68 cases; e.g., *flood*, *come off*).

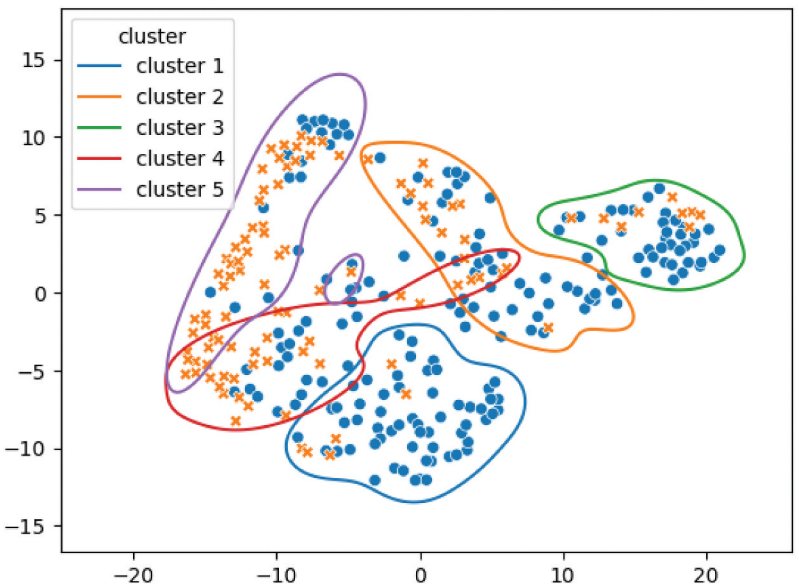

Figure 2. t-SNE projection (van der Maaten & Hinton, 2008) of BERT vectors of *cascade* with clusters indicated as lines. Dots represent tokens from the fiction subcorpus; crosses tokens from the newspaper subcorpus.

Table 2. Clusters for the lexical item *cascade* with characterizations and examples.

#	N	% Fiction	Target domain (partitives)	Typical syntagms	Examples
1	68	91	Perceptual phenomena (*sparks*, *thumps*, *boos*, *crockery*, *books*)	Motion verbs (*flood*, *erupt*), adjectives like *roaring, steady, twisting*	(a) Finally the mantels lit and a cascade of light flooded the garage. (b) They had driven a few kilometers now, and as I watched the cascade of sparks coming off the front of the truck [. . .]
2	53	73	Similar to Cluster 1; fewer partitives	Motion verbs (*swirl, tumble*), color adjectives	(a) A broad red cascade of carpet led up the stairway (b) Clara [saw] a cascade of dust swirling by her desk (c) She whirled around, mystified, and the lilacs tumbled from her arms in a lavender cascade.
3	48	81	hair, clothing, and jewelry	Motion verbs (*splash, sweep*), *with*-adjuncts	(a) one of whom had a remarkably level head tucked beneath that familiar cascade of dark curls. (b) her hair was long and brown and splashed in an unruly cascade about her face and neck (c) . . . a cool wedge haircut with the sides shaved and a cascade of dangly earrings on one ear.
4	53	45	Abstract event nouns (*bankruptcies*, *embellishments*, *injustice*, *disclosures*)	Causal verbs (*cause*, *launch*, *set off*, *unleash*)	(a) The rest of the film, incredibly densely packed with a cascade of images – documentary, fictional, speculative – seemed to hurtle forward (b) Even with it zipped, she feared that the wallet would blurt into view in some way that she couldn't control, unleashing a cascade of horrors: arrest, shame, poverty, death.
5	66	34	Abstract event nouns, mechanical artifacts (*centrifuges*)	Causal verbs	(a) But routine screening . . . [has] the potential to do harm . . . , potentially setting off a cascade of unnecessary events like ultrasounds, needle biopsies in the neck, operations to remove the thyroid and complications . . . (b) the Osp-A protein in a vaccine could start a similar immunological cascade and cause arthritis months or even years after inoculation (c) . . . to link 164 spinning centrifuges in what nuclear experts call a cascade.

Cluster 2, next, has a lower incidence of postnominal *of*-phrases (51% of the tokens), but where they occur, they are similar to those in Cluster 1 (denoting natural phenomena and sounds). What stands out is that *cascade* is with some frequency (13/53 cases) modified by color adjectives, as in examples (a) and (c), or more broadly adjectives denoting visual properties (*sparkling, eye-catching, Technicolor, flashing*). Like Cluster 1, this cluster also frequently features *cascade* as the subject or manner adjunct of motion verbs (17/53 cases).

Cluster 3 revolves overwhelmingly around the target domain of hair (34/48 cases; instantiated in nouns like *hair, curls, dreadlocks, braids, extensions*) and, secondarily, clothing and jewelry (*ruffles, ribbons, lace, earrings*). These nouns are furthermore frequently modified with adjectives describing visual properties such as color (*auburn, black, red*) and shape and texture (*navette-cut*, *loose*, *carefully tousled*), and similarly occur with motion predicates, though frequently as *with*-adjuncts (9/48 cases), as in example (c). Given the coherence of the examples in this cluster, it arguably illustrates a conventionalized metaphorical use of *cascade*.

Turning to their meaning, these three clusters draw on imagistic or sensory properties of the source domain *cascade*. They are clear cases of image metaphors (Lakoff, 1987; Lakoff & Turner, 2009), where a salient perceptual correlation between the source and target domain is drawn upon, without implying a fuller domain-to-domain mapping. This means that the target domains of these clusters are also overwhelmingly concrete (*lights*, *sparks*, *carpet*, *dust*, and *curls*), and likened to a cascade because of sensory (visual, auditory, textural) properties explicated as attributive adjectives as well as their moving appearance as – as suggested by the extensive collocation with motion verbs, which are often to be interpreted as *fictive* motion (i.e., an imagined event of motion while the actual event is one

of stasis; Talmy, 2000, pp. 100–101). As we will see below, it is the imagistic nature of these metaphors that sets these clusters apart from the remaining two clusters.

In Cluster 4, all instances have postnominal *of*-phrases. These are mostly (abstract) event nouns such as *bankruptcies*, *embellishments*, *injustice*, and *disclosures*. Here, the most frequent type of syntagmatic relation of the noun phrase is not as the subject of manner of motion verbs, but rather as the direct object of a verb denoting creation or causation (*cause, launch, set off, unleash*) – verbs known for taking negatively valenced direct objects (Louw, 1993).

Similarly for Cluster 5, the role as direct object of a verb of creation or causation is the dominant usage (17/66 cases, as in examples (a-b)), with roles as subject of causal verbs being another frequent category (10/66 cases; e.g., *cause, create, develop, play a role*). These cases frequently display *of*-phrases, with target domains that, impressionistically speaking, involve more cases of medical processes than Cluster 4. However, this cluster also contains a set of cases where *cascade* is used without much modification, in the sense of 'a sequence of mechanical implements or artifacts', as in example (c).

In contrast with the first three clusters, the instances in these clusters do not seem to (primarily) evoke the imagistic properties of the cascade, but rather the fact that cascades are differentiable (stepwise) waterways that move with force, with one part affecting the next (through gravity and the force of the water). What supports this analysis, are the use of abstract event nouns and (bio-)mechanistic systems as target domains, as well as the prevalence of verbs of causation. Broader contextual cues like *hurtle forward* (example (a) of Cluster 4), *couldn't control* (example (b) of Cluster 4) add to a semantic prosody (Louw, 1993) in which the metaphorization of a phenomenon as a *cascade* invites inferences of 'one event/entity causally affecting the next (in a way that is hard to stop)'.

Waterfall

We now turn to a near-synonymous lexical item, *waterfall*. The clusters are introduced in Table 3.

Table 3. Clusters for the lexical item *waterfall* with characterizations and examples.

#	N	% Fiction	Target domain (partitives)	Typical syntagms	Examples
1	38	89	Perceptual phenomena (*bottles*, *noise*, *trills*) including liquids (*milk*, *glaze, drink*)	Similes (*like*), manner adjuncts of motion verbs	(a) … plaster and brick fell to the floor in a waterfall of noise … (b) … flowers were pouring out of the place like a waterfall of bold blue irises. (c) … sent my nights reeling into a waterfall of drink.
2	45	88	Perceptual phenomena	More elaborate similes ("like Ving P a waterfall of N")	(a) It was like standing under a pounding waterfall, … (b) the brief tropical showers can be like living inside a waterfall, (c) There was a roaring in his ears like a distant waterfall …
3	11	100	Few partitives, though domain (hair) is contextually expressed	Similes (*like*)	(a) … hair pouring forward over her face like a blonde waterfall … (b) leaving her hair loose as a waterfall down her back
4	12	100	Hair: many partitives	Variable. Notably few motion verbs.	(a) The handsome young man … represented the best London society had to offer – an exquisite plum tailcoat, a high-tied waterfall of white about his neck, … (b) Rosita … straightened the driver's cap over her waterfall of dark curly hair (c) [her hair was] shoulder-length now, not the rich waterfall to her waist …
5	21	47	Liquids + "rest"	Variable	(a) Inside the … dining room, with a waterfall spilling down the wall (b) BP's waterfall of cash has changed people's lives profoundly. (c) … without the threat of falling over an unseen waterfall of a canceled World Series.

Cluster 1 ($N = 38$, 89% fiction usage) predominantly comprises *of*-phrases (31/38) followed by a concrete target domain (27/31). Among these target domains, we frequently find liquids (e.g., *milk, glaze, blood*) and sounds (e.g., *trills*). In terms of their syntagmatic embedding, the tokens in Cluster 1 (13/38) show a fairly even distribution of grammatical roles (see examples (a–c)).

Cluster 2 ($N = 45$; 88% fiction usage) is similar to Cluster 1 with respect to the concreteness of its target domains, but it is syntagmatically different. Like Cluster 1, it presents metaphors about sound, objects, and liquids, including less metaphorical instances of the target domain 'water' (e.g. *rain, showers, water*). However, syntagmatically, Cluster 2 consists exclusively of similative constructions. Among these similes, some are longer metaphorical comparisons (16/45), where *waterfall* is not directly mapped onto a target domain, but rather forms part of a larger source domain – such as, in (1), the experience of standing near a waterfall. These longer comparisons tend to pick on more phenomenological, abstract, and human-centered properties, such as the weight of the water and the fear that can be associated with waterfalls. In this way, Cluster 2 also differs from the previous Cluster 1, where target domains were more consistently concrete.

The two smaller Clusters 3 and 4 ($N = 11$resp. $N = 12$; both 100% fiction), display a very large degree of coherence in their target domains. In Cluster 3, 73% of metaphors are about 'hair' – with other examples targeting *fingers* and *roses* – and in Cluster 4, 100% of tokens are about 'hair'. However, these clusters differ from each other in their syntagmatic patterns as well as in the creativity of their metaphors. Cluster 3 consists exclusively of similes, whereas Cluster 4 is characterized by a majority of *of*-phrases following the verb (10/12), with the target of the metaphor (e.g. 'hair') not always being explicitly mentioned in the *of*-phrase, but rather implied (3/10), often via a color term, as in (a). In the two examples that do not contain an *of*-phrase, the target of the metaphor is altogether left implicit, for instance in (c). These stylistic choices contribute to lending the cluster greater creativity.

Finally, Cluster 5 ($N = 21$; 47% fiction) is more diverse and comprises some literal uses (3/21), as well as less metaphorical uses targeting 'water' (6/21), and instances where *waterfall* is used as a lexical gap-filler for water installations that resemble a waterfall (4/21), such as (a). The remaining tokens are more disparate, some of which instantiate more genuine metaphorical usages, like (b), or metaphors about 'money,' like (c). These tokens are also more syntagmatically diverse. For our purposes, we treat this as a leftover cluster, which is less interpretable, and which may also be bound to happen given the diversity of linguistic usages.

Discussion

In short, we found that our clustering method picks up on syntagmatic similarities, such as the presence of pre-nominal modifiers, post-nominal *of*-phrases, relations to the verbal predicate, and the nature of the verbal predicate. These syntagmatic similarities correspond to similarities in the level of concreteness or abstractness of the target domains found in each cluster, and overall make it possible to analyze clusters in terms of different, interpretable senses. Some leftover clusters remain, for instance Cluster 5 for *waterfall*, that are nonetheless conveniently grouped as such.

We also find that multiple clusters present overlap in terms of their metaphorical meaning: the first three clusters of *cascade*, for instance, all involve image-metaphoric mappings of visual and textural properties of cascades. The third cluster, targeting 'hair', can be argued to differ on the basis of its narrower target domain (a possible case of sense narrowing; Darmesteter, 1887). The first two clusters, however, seem to differ mostly on formal, rather than semantic or conceptual grounds: it is the greater presence of partitives in the former that likely led the clustering algorithm to determine that there were two, rather than one, clusters here. As such, it is important to keep in mind that the cluster analysis does not provide one-to-one, but rather many-to-one mappings between clusters and interpretable senses. Nonetheless, having two slightly distinct collocational patterns may over time lead to two more distinguishable metaphorical senses, as the two patterns lexicalize and each obtain different inferential, and then conventionally anticipated, content (see, for a similar argument about "tracing lineages" of metaphors, Svanlund, 2007).

In short, while the cluster analysis is by no means a substitute for a quantitative corpus-based, or qualitative text-oriented analysis, it constitutes a useful starting point for such analysis, as the clusters clearly display coherence in the types of cases they group together. For our purposes, this means that we can use the approach for our subsequent large-scale quantitative analysis of variation across corpora.

Now turning to our comparison between *waterfall* and *cascade*, we see that, as predicted by the Invariance Principle, some metaphorical senses of these two near-synonymous lexical items do overlap. It is worth noting, however, that even when they do so, there are differences in the syntagmatic patterns of the metaphors, that is, differences in the conventional way in which these overlapping senses are used for the two lexical items. For instance, both *waterfall* and *cascade* presented clusters targeting 'hair', but *waterfall* displayed a cluster where the target domain ("hair") was exclusively the subject of similes (e.g., *hair fell like a waterfall*), whereas no such cluster was found for *cascade*. This means that even where the source-to-target mapping is equally available for both lexical items, the conventionalized ways of using the mapping may differ across lexical items.

However, not all metaphorical senses found for *waterfall* and *cascade* in our clustering analysis were shared by both lexical items. Some were exclusive to just one lexical item, such as the "causation" metaphors for *cascade* (Cluster 5) or the "phenomenological" meaning of *waterfall* (Cluster 2; e.g., *like standing under a waterfall*). Similarly, *waterfall* had a sizable set of cases, in Clusters 1 and 2, where the target domain was a liquid or even more narrowly: water. In such cases, the distance between the source and target domain is small, straddling the line between metaphor and other figures of speech, such as hyperbole (e.g., a large amount of a liquid) or broadening (e.g., from a naturally formed downward flowing body of water to any downward flowing body of water). Critically for our discussion, these cases occur with some frequency for *waterfall* but less so for *cascade*.

Such differences in conventional senses seem to remain unexplained if we take metaphoric behavior to just be a productive mapping between domains. We would expect *waterfall* to also be used to mean 'hard-to-stop causal sequence' with some frequency, and *cascade* be used in expressions such as *it felt like sitting at the bottom of a thundering cascade*, but such cases are vanishingly rare. As such, our data supports Deignan's (1999) and Svanlund's (2007) arguments, namely that *lexical convention* – that is: which metaphorical senses are conventional parts of our lexical knowledge – at the very least guides metaphorical usage alongside more general and productive domain-to-domain mappings as proposed by Conceptual Metaphor Theory. Of course, further analysis may reveal that differences in the properties of the source domains (e.g., *waterfall* versus *cascade*) explain their variable mapping to target domains (along the lines of Sullivan's, 2013 analysis). However, such an argument would have to show that (1) the variation in source domain conceptualization is itself not a matter of convention, and (2) that different syntagmatic patterns associated with the same mapping (e.g., 'vertically falling body of water' to 'hair') are somehow predicted by the conceptualization of the source domain rather than by conventionality as arising through usage. We take this question to be an important one for future work.

Barring such an account, we consider both contrasts discussed above ("different availability of senses" and "same senses – different syntagms") as pointing to the indispensability of the linguistic dimension of metaphorical expression, namely: the organization of lexical items in terms of norms and violations drawing on those norms (Hanks, 2004, 2013), and processes of conventionalization and entrenchment (Svanlund, 2007 for metaphor; Schmid, 2020 for a more general framework). Our data shows that lexical convention does indeed play a (seemingly substantial) role in shaping metaphorical usage, and that, while metaphor is undoubtedly also a product of non-linguistic conceptual mappings, it is also intrinsically a linguistic, lexical, phenomenon.

Understanding variation in metaphor usage

The previous section demonstrated that there is a close connection between recurrent linguistic contexts and inferrable metaphorical senses. As such, syntagmatic contexts can be considered as

one component of a model that tries to explain (quantitatively and theoretically) patterns of metaphor use. Here, we consider a second component, namely: genre. We will develop the argument that to understand the differing pattern of metaphor usage across genres, we need to consider how the recruitment of metaphoric language (i.e., the choice to speak or write metaphorically) supports the communicative goals and stylistic practices of different genres.

Motivating the analysis at the level of usage clusters

Crucially, the clusters established with the methods presented above should be thought of as the ideal starting point for such an analysis. Given that each cluster of usages (more or less) represents a metaphorical sense, we expect it to be the clusters that are the primary locus of variation across genres (rather than lexical items, or even entire source domains), as the utility of particular metaphorical meanings might vary across genres. Individual lexical items are indeed expected to have clusters whose biases are opposite – some clusters occurring more in newspaper, others in fiction. This section first considers the empirical adequacy of this motivation.

We formalize the bias (or: numerical overrepresentation) of a metaphorical usage type (i.e., a cluster) as the proportion of tokens in that usage type that come from the fiction subcorpus. For the analysis in this Section, we quantize the genre bias of usage types into three bins depending on the proportion of that usage type that came from the fiction corpus: if one-third or less, we call the cluster *newspaper-biased*; if two-thirds or more, we call it *fiction-biased*; and otherwise, we call it *unbiased*. Using this measure, we can then characterize each lexical item by the set of biases of its metaphorical usage types. *Cascade*, for instance, has both fiction-biased and unbiased clusters. As such, not all metaphorical usage types of *cascade* behave alike in their bias to either genre.

We indeed find such heterogeneity across the board: Table 4 presents all lexical items, grouped by source domain (columns) and by the type of biases their clusters display (rows). We find heterogeneity in over 90% (41 out of 45) of the lexical items from the two source domains. Furthermore, 40% of all lexical items (18 out of 45) have both newspaper-biased and fiction-biased clusters, and thus display substantial internal variation in how they are used.

These results suggest that variation in genre association is indeed found at the level of metaphorical usage types and that it is that level that is a more fruitful starting point for analysis.

Predicting genre variation on a cluster level

Having established that clusters of usages are the right level to analyze between-genre variation, we can now ask what factors influence the greater prevalence of a metaphorical sense in one genre than the other. Here, we draw on Goatly (1997, 2011) observation that news and fiction mostly differ in their Field, that is: the nature of the social action taking place (cf. Halliday & Hasan, 1985). The genre of news is more explicitly informational, with its Field involving "publishing (selling) newspapers, giving/receiving information and forming opinion about recent events of, or in the, public interest;

Table 4. Classification of lemmas based on the genre bias of their respective clusters.

Contains:	N	'Landscape' examples	N	'Weather' examples
Balanced + fiction	14	Beach, cascade, cavern, desert, forest, fountain, lake, mountain, quicksand, river, sea, swamp, trench, waterfall	5	Fog, frost, snow, sunshine, wind
All	12	Bay, cliff, gulf, hill, island, land, pond, precipice, puddle, slope, stream, valley	5	Breeze, flurry, hail, torrent, wave
Balanced + newspaper	4	Gackwater, iceberg, plateau, quagmire	0	
Fiction	1	Cave	2	Gust, rain
Fiction + newspaper	0		1	Deluge
Balanced	1	Morass	0	

entertainment" (Goatly, 2011, p. 315), with a "critical bias towards content over form, the word over the image, and what a paper says as opposed to how it says it" (Mussell, 2014, p. 20). The Field of fiction, conversely, involves "publishing books; entertainment, aesthetics, literature; creation of a fictional world and fictional characters, which more or less reflect society and psychology – thereby exploring themes of interpersonal and social significance; exciting emotions of identification, sympathy or antipathy towards characters; and of suspense and curiosity in the reader" (Goatly, 2011, p. 316), and it is a genre where authors stand out from each other based, in part, on the language they use. We believe this, concretely, leads to three predictions.

First, we expect metaphors in fiction to have more concrete target domains than metaphors in newspaper text. This is in line with Goatly's finding that metaphors in fiction frequently fulfill the rhetorical goal of helping readers "reconceptualize" familiar states of affairs in a novel way (e.g., a light beam as a *cascade of light*), whereas they are not found to serve this function in newspapers (Goatly, 2011, p. 151). In Conceptual Metaphor Theory (CMT), these examples of "reconceptualization" broadly correspond to what scholars call image (or poetic) metaphors (Lakoff, 1987; Lakoff & Turner, 2009), where one perceptual source domain (e.g., *cascade*) is used to qualify another perceptual target domain (e.g., *light*). Metaphors in newspapers, on the other hand, primarily serve the function to supplement literal language where it is found lacking, for instance, to describe abstract concepts or events like a series of bankruptcies causing each other in *a cascade of bankruptcies*. Given this, we expect to find more examples of conceptual metaphor in newspapers where the target domain is abstract and made more tangible (i.e., cognitively "accessible") via the concrete source domain (Lakoff & Johnson, 1980). While the association between concreteness and metaphor type is not perfect, it is strong enough to use concreteness to operationalize the expectation that fiction uses more *image* metaphors, and news more *conceptual* metaphors. In particular, for every sentence in a cluster, we gather all lexical context words occurring in a five-word window around the metaphorical word.[2] We then determine, per context word, the maximal sensori-motor value from the Lancaster Sensory-motor database (Lynott et al. 2020), which encodes concreteness on various sensory-motor dimensions. We then average these, first per sentence, and then for the whole cluster. This gives us an index (**concreteness**) of how concrete the contexts in a metaphor cluster are. Implementational details for the computation of all the independent variables can be found in the Supplementary Materials.

Second, we expect metaphors in newspaper texts to be more fixed in their form than metaphors in fiction. The "attention to content over form" might steer the journalist to draw more on well-established metaphorical expressions, whereas the demand on the literary writer to be creative with more or less well-established metaphorical senses (to "evoke curiosity," for instance) might lead to a lower degree of formal fixedness. Goatly (2011, p. 323) further argues that "serious newspapers, with their informative rather than entertainment properties, generally discourage metaphorical processing." As there is no standard way of quantifying fixedness, we approximate this construct by measuring how diverse the contexts are that a word can be found in. To compute an index of **fixedness**, we gathered the set of all the lemmas of dependent words in a dependency parse, as well as the lemma of the head word, for each metaphorical token (e.g. *beneath* as the head and *that*, *familiar* and *curls* as dependents in the COCA example (3) below). Finally, we calculated the pairwise Jaccard similarity between the sets of all tokens in a cluster. The fixedness index ranges from 0 (all tokens occur in completely unique contexts) to 1 (all usages in a cluster display identical contexts). We expect that, on average, the fixedness is higher for newspaper-biased clusters than for fiction-biased clusters.

[2]This operationalization is based on the assumption that the target domain of the metaphor is indeed mentioned in the context. To validate the measure, we derived a similar measure in which only the concreteness score of any dependent noun of the metaphorical word was used, thus only finding cases like *assumptions* in *quagmire of assumptions*, or *revenue* in *revenue stream*. Those are virtually exclusively used as descriptors of the target domain and as such allow us to provide an index of the concreteness with a high certainty that we're actually picking up on the target domain. However, they don't occur with every metaphorical noun and as such lead to the absence of an index of concreteness for some clusters. Nonetheless, a concreteness score per cluster extracted this way has a strong correlation with the window-based concreteness score per cluster for clusters where both could be extracted ($r = .47$; $p < .001$).

(3) . . . a remarkably level head tucked beneath that familiar cascade of dark curls.

Third, we expect metaphors in newspaper text to be more negatively valenced than metaphors in fiction. Newspapers (in the United States, as in most places) are commercial products which need to attract a paying audience. This goal can be expected to affect the choices of topics discussed: negative topics might be more prevalent as they draw more attention (cf. Soroka & McAdams, 2015; Soroka et al., 2019), and metaphors – which draw attention to themselves by virtue of extending the (literal) meaning of a word – could serve as ideal means to present such negative topics. This resonates with Goatly's (1997, p. 164) function of "enhancing memorability, foregrounding, and informativeness." To operationalize this expectation, we applied an off-the shelf sentiment analysis tool (the sentiment-analysis model from the transformers library) to each token of each cluster to determine its valence (positive or negative). This model uses the contextualized vector representations of a sentence to predict its sentiment. We defined the **valence** index per cluster to be the proportion of positive tokens in a cluster.

As clusters are inferred on the basis of the BERT vectors, and given that the BERT vectors represent aspects of collocational structure, we have to be concerned with the possibility that any correlation between genre-bias and the three indices was confounded by certain formal properties of the clusters. In particular, we incorporated two factors as control variables. First, the presence of *metaphor flags*, or: MFlags (the proportion of tokens in a cluster having *as* or *like* heading the noun phrase containing the metaphor) and target-domain markers, or TDMs (nominal expressions of the target domain that are syntagmatically related to the metaphorical expression as pre- or postnominal modifiers, e.g., *dust* in a *cascade* of dust, again counting their proportion in a cluster), in conforming with the MIPVU procedure (Steen et al., 2010). MFlags and TDMs, as identified using the SpaCy library (Honnibal & Montani, 2017), are distributed unevenly across genres: 5% of newspaper tokens contain an MFlag, versus 14% of fiction tokens, and 65% of newspaper tokens display a TDM vs. 54% of fiction tokens. As such, it is sensible to control for them in a statistical experiment: after all, if more concrete senses also tend to have more MFlags (as is the case: they correlate substantially – Pearson's $r = .50$, $p < .001$), the correlation with the dependent variable of genre bias is confounded.

Before applying a multivariate model to these factors, we first removed outliers on each of the five independent and one dependent variables by omitting any datapoint that had a value of greater than three times the Median Absolute Deviation from the median on any of the six variables (see Leys et al., 2013 for the motivation for using this measure over the more well-known use of *N* Mean Standard Deviations over or under the Mean). Subsequently, the values for all five factors were *z*-transformed so that the magnitudes of their effects could be compared. We use the measure of genre-bias introduced above as the dependent variable in a beta regression, the appropriate method of analysis for proportional data (Geissinger et al., 2022): any factor with a positive coefficient will indicate that it predicts a cluster to be biased toward fiction, and any negative coefficient will indicate a bias toward newspaper text. The regression analysis was applied to 131 unique clusters across 44 lexical items, using the Python statsmodels library (Seabold & Perktold, 2010).[3]

Results and analysis

Results of the regression analysis are presented in Table 5. The only factor showing a significant effect is the concreteness of the context, predicting a positive relation to the proportion of fiction. The other two factors, fixedness and valence, were not found to be significant predictors.

To interpret these results, we refer back to the fact that fiction texts might use more poetic or image metaphors, where the target described by the metaphor is concrete (e.g. a *cascade* of red carpet), whereas newspapers might use metaphors more often in combination with abstract concepts (e.g., social phenomena, economics) to render them more cognitively accessible, that is: as conceptual

[3]As outlier removal reduced the size of the dataset by about a third, a supplemental analysis without outlier removal, confirming the same statistical patterns, is reported in Appendix 4.

Table 5. Results from beta regression predicting cluster bias on the basis of fixedness, concreteness, and valence, alongside the control variables of proportion MFlags and proportion TDMs ($N = 131$ items).

	Coefficient	st. err.	Z	*p*
Intercept	0.309	0.084	3.666	< .001
Concreteness	0.502	0.089	5.620	< .001
Fixedness	0.113	0.087	1.297	0.194
Valence	−0.078	0.085	0.992	0.357
Proportion MFlags	0.126	0.089	1.422	0.155
Proportion TDMs	0.084	0.086	0.974	0.330

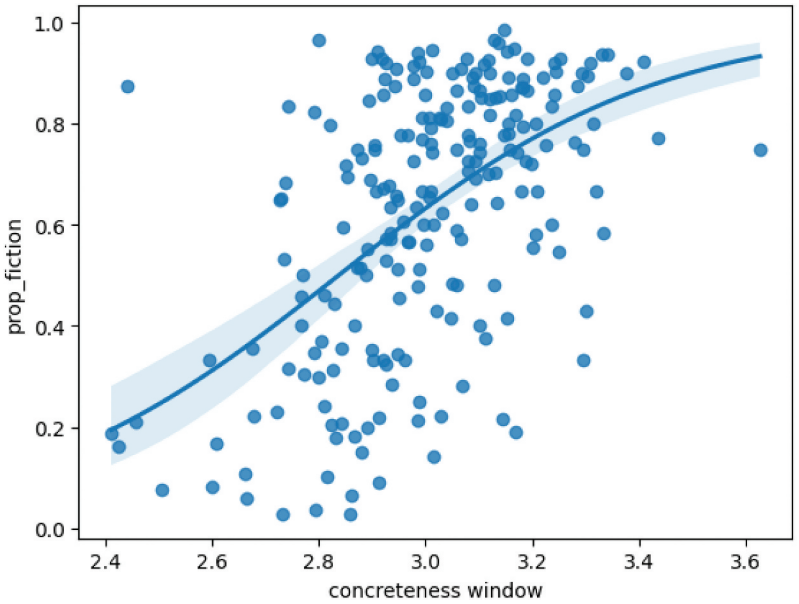

Figure 3. Correlation of concreteness with the proportion fiction per cluster.

metaphors. Figure 3 shows this correlation ($r = .54$, $p < .001$). Note that there is a double dissociation: concrete phenomena are described in newspapers too, but do not generally warrant metaphorical language use as much, while abstract phenomena are described in fiction too, but they, in turn, do not warrant metaphorical language use there either. In other words: the function of metaphorizing is different in the two genres.

Of course, there are individual cases that go against this patterning. Many of them are artifacts of the estimation method: Cluster 3 of *pond* ($N = 19$; 14% fiction) has a high concreteness score despite being newspaper-biased, but this is primarily due to it involving the fixed expression *(small) fish in a (big) pond*, where the presence of the word *fish*, being a rather concrete entity, inflates the concreteness score of the clusters. Some, furthermore, illustrate the heterogeneity of the genres, which we have conveniently glossed over so far: Cluster 10 of *puddle* ($N = 21$; 19% fiction), for instance, involves highly concrete target domains, but most examples here seem to be drawn from food reviews in newspapers (*served in a puddle of lightly sweetened soy sauce*; *chilled mascarpone cream in a puddle of warm chocolate espresso*; *rare meat and a little puddle of sweet brown sauce*), where the target domain is already a liquid, and hence concrete, just not one that is normally considered to occur in puddles (e.g., unlike mud and rain). Food reviews might be a genre in its own right, whose norms might sit closer to fiction in some ways than to news reporting. Interestingly, the identification and inspection of outliers in

a statistical analysis allows us to detect such variation within the (corpus-delineated) genres and make sense of them.

We finally observe that the lack of a result on the other two dimensions might be an effect of our operationalization (though fixedness is trending, albeit against the expected direction). Each of these factors could be the subject of a paper-length study into how best to operationalize them as a quantitative index, and we would like to encourage future research in that direction.

Conclusions

Recapitulating

In this paper, we set out to develop an account of metaphor usage in terms of its linguistic and extralinguistic contexts. The non-metaphorical meaning of a lexical item may be used metaphorically in different ways: *cascade* could be metaphorized as a 'causal chain of events' or a '(seemingly) diagonally moving object or mass'. We developed a computational method that was applied at unprecedented scale to extract metaphors from a large corpus (COCA; Davies, 2008) and group them on the basis of their syntagmatic environments. We qualitatively investigated two lexical items, *cascade* and *waterfall*, in order to validate the method. Using this method, we further developed the argument, first proposed by Deignan (1999), Hanks (2004), and Svanlund (2007), that different metaphorical meanings of the same lexical item are each associated with their own syntagmatic environments: their collocations, colligations, and semantic prosodies tend to differ. This means that the basis of the interpretation is larger than the individual word; aspects of the context may (more or less conventionally) be necessary as components of the metaphorical interpretation.

This analysis of metaphorical polysemy allowed us to provide further evidence for Deignan's (1999) proposed upper bound on the Invariance Principle (Lakoff, 1990, 1993): words denoting highly similar concepts in the source domain do not always have the same distribution of metaphorical mappings to various target domains. Our clustering method facilitates the analysis that lets us pinpoint such differences. More positively, such analyses underscore that metaphor usage to a large extent consists of lexical norms governing contextual usages, which are somewhat conventional: many recurrent sets of metaphorical usages in particular contexts are not quite dead metaphors yet, but indeed perhaps losing some of their metaphorical strength (see Svanlund, 2007) due to their conventionality and therefore lesser reliance on activating the metaphorical mapping.

With these clusters of usages for 45 lexical items, we considered a second aspect of a model of metaphor usage, namely the extralinguistic context (i.e., the text type or genre) that a metaphor occurs in. We first established that the clusters of metaphorical usages differed in their frequency across genres. We furthermore expected a cluster's frequency bias to either genre to be predictable from properties of the genres, in particular the Field – the social activity taking place in the genre; cf. Halliday and Hasan (1985), as previously made relevant for metaphor in Goatly (1997, 2011) and Deignan et al. (2013). We predicted, based on this observation, that metaphors in news might be more *conceptual* (i.e., targeting abstract concepts for which literal language might be lacking), more fixed in their form (i.e., not trying to use innovative language), and more negatively valent, since newspapers (in the U.S. context, at least, and likely beyond) are also commercial products that need to draw in audiences with shock value (Soroka & McAdams, 2015). Conversely, we predicted that metaphors in fiction might be more *imagistic* (i.e., reconceptualizing concrete target domains), more creative, and less negatively valent. Operationalizing these factors in a way that allowed for a large-scale study of their effects, and controlling for two extraneous factors (explicit indicators of metaphoricity, i.e., MFlags, and linguistic expressions of the target domain in partitive *of*-phrases), we found that only the concreteness of the target domain (i.e., whether a metaphor is conceptual or imagistic) was a significant predictor of genre bias, with metaphorical senses biased toward fiction being more often imagistic, and metaphorical senses biased toward news being more often conceptual. We take this to reflect the different function that metaphor fulfills in the two genres: the attention-grabbing, creative reconceptualizing functions of language in fiction license

more imagistic metaphors, while the necessity to express abstract, complex phenomena in newspapers makes conceptual metaphors more useful.

In short, our evidence suggests that computational modeling can be a useful starting point for studying metaphorical polysemy and variation at scale, and that the linguistic and broader communicative contexts in which metaphorical language is used (e.g., text type or genre) play an important role in shaping the metaphorical potential of a source domain. Moreover, our findings lend support to the argument that metaphor is not only a conceptual, but also a linguistic phenomenon that is strongly rooted in usage events (Deignan, 1999): language users need to know how (in what linguistic and extra-linguistic contexts) the words *cascade* and *waterfall* are commonly used metaphorically in a speech community, in ways that do not directly fall out from the properties of the source domains, to show communicative competence. In language usage, metaphor is mediated by lexical convention and the linguistic activities at hand through and through.

Taken together, the picture that emerges is one in which metaphorical language use is best understood through the interaction of various components, as a "complex dynamic system" perhaps (cf. Cameron & Deignan, 2006; Gibbs & Cameron, 2008), where (emergent) properties of higher-order systems, such as genres, affect the likelihood with which novel metaphorical usages of lexical items are coined and spread through a community. Such an account explains both why genre-level similarities across lexical items exist, and why we observe differences between minimal pairs such as *cascade* and *waterfall.*

Vantage points and vistas

By means of a conclusion, we would like to lay out several implications of our analysis.

First, a central instrument in our analysis was the use of computational methods. We used such methods to extract metaphors at a large scale, with a reasonable reliability, to group together similar usages, and to estimate indices of how concrete, negatively valenced, and conventional groupings of metaphorical usages were. We believe that the use of cluster analysis to group together distributional-semantically similar metaphorical tokens of a lexical item is a generalizable method that could support the analysis of other corpus-oriented metaphor researchers. The method can be applied to any corpus data in a language for which contextual semantic representations exist or can be trained, and allows researchers to find groupings of similar usages that tend to reflect homogenous metaphorical patterns, for instance in involving similar kinds of target domains, similar conceptual properties of the source domain drawn upon, and similar linguistic expressions, such as verbal idioms. We leveraged these clusters to study variation between genres, but these groupings can equally well be used to study other kinds of contrasts (different authors, different languages, different political positions) and to, more generally, obtain insight in how the same source concept is metaphorically used in different ways.

Our analysis considers variation as primarily rooted in lexico-pragmatic convention (following Deignan, 1999; Hanks, 2004) and discourse-pragmatic practice (i.e., genre, following Deignan et al., 2013; Goatly, 1997, 2011). This paradigm suggests interesting future steps. Regarding embeddedness in the linguistic context, we could ask how more creative or active metaphors function – are they just the unconstrained expression of novel or existing conceptual metaphors, or does active metaphorizing still happen against the backdrop of lexical norms as well, through purposeful violations of other lexical norms (in Hanks' 2013 framework), or otherwise? Second, the interaction with discourse-pragmatic practice similarly raises further questions. Is figurative language somehow specifically normatively governed in practices (somewhat similar to Goddard's, 2004 pragmatic scripts for studying crosslinguistic variation in figurative language use), or do the patterns we observe fall out from more general normative patterns of language use (as e.g., Leezenberg, 2007 argues for metaphor over historical time)?

Our characterization also bears on the question how metaphorical language functions on the individual cognitive level, a topic we did not address in this paper. Nonetheless, our characterization of the situatedness of metaphorical language use in linguistic and extra-linguistic context opens

research avenues for considering how these sources of information are implemented in the human mind. For instance, we can wonder: do metaphors incur an additional processing cost when used in extra-linguistic contexts in which they are not licensed? Do they pick up on the normative dimension of such extra-linguistic associations (i.e., by judging it as inappropriate or ineffectual)? Moreover, it cautions against drawing all too strong conclusions from experimentation on decontextualized metaphor usage.

All in all, our paper hopes to have contributed to the understanding of the relation between linguistic and extralinguistic contexts and metaphorical senses. Importantly, the former insight feeds the latter: having clusters of usage types allows us to research between-genre variation in a novel way, thus demonstrating the potential of computational methods to study the context-dependent nature of metaphor use.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by an NSERC Discovery Grant Word Meaning in Action (RGPIN-2019-06917) to Barend Beekhuizen.

ORCID

Barend Beekhuizen http://orcid.org/0000-0003-1275-2974

Data availability statement

All code used to generate the results, supplementary materials, and a sample of the data can be found on an OSF repository: https://osf.io/yq3rf/?view_only=6a3739330fad4e98aaa17a1f8d280788.

References

Caballero, R. (2003). Metaphor and genre: The presence and role of metaphor in the building review. *Applied Linguistics*, *24*(2), 145–167. https://doi.org/10.1093/applin/24.2.145

Caballero, R. (2017). Genre and metaphor: Use and variation across usage events. In E. Semino & Z. Demjén (Eds.), *The Routledge handbook of metaphor and language* (pp. 211–223). Routledge.

Cameron, L., & Deignan, A. (2006). The emergence of metaphor in discourse. *Applied Linguistics*, *27*(4), 671–690. https://doi.org/10.1093/applin/aml032

Cameron, L., & Maslen, R. (2010). *Metaphor analysis*. Equinox.

Chacón, J. E., & Rastrojo, A. I. (2023). Minimum adjusted rand index for two clusterings of a given size. *Advances in Data Analysis and Classification*, *17*(1), 125–133. https://doi.org/10.1007/s11634-022-00491-w

Choi, M., Lee, S., Choi, E., Park, H., Lee, J., Lee, D., & Lee, J. (2021). Melbert: Metaphor detection via contextualized late interaction using metaphorical identification theories. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1763–1773).

Chronis, G., & Erk, K. (2020). When is a bishop not like a rook? When it's like a rabbi! Multi-prototype BERT embeddings for estimating semantic relationships. *Proceedings of the 24th Conference on Computational Natural Language Learning (CoNLL 2020)* (pp. 227–244). Association for Computational Linguistics.

Darmesteter, A. (1887). *La vie des mots*. Delagrave.

Davies, M. (2008). The corpus of contemporary American English. www.english-corpora.org/coca/

Deignan, A. (1999). Corpus-based research into metaphor. In G. Low & L. Cameron (Eds.), *Researching and applying metaphor (pp* (pp. 177–200). Cambridge University Press.

Deignan, A. (2005). *Metaphor and corpus linguistics*. John Benjamins.

Deignan, A., Littlemore, J., & Semino, E. (2013). *Figurative language, genre and register*. Cambridge University Press.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.). *Proceedings of the 2019 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171–4186). Association for Computational Linguistics, Minneapolis, Minnesota.
Dorst, A. G. (2015). More or different metaphors in fiction? A quantitative cross-register comparison. *Language and Literature*, *24*(1), 3–22. https://doi.org/10.1177/0963947014560486
Ethayarajh, K. (2019). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (EMNLP-IJCNLP) (pp. 55–65). Association for Computational Linguistics.
Fellbaum, C. (1998). A semantic network of English: The mother of all WordNets. *Computers and the Humanities*, *32* (2–3), 209–220. https://doi.org/10.1023/A:1001181927857
Firth, J. R. (1957). A synopsis of linguistic theory, 1930–1955. *Studies in Linguistic Analysis* (pp. 1–32). Blackwell.
Fludernik, M. (2019). *Metaphors of confinement: The prison in fact, fiction, and fantasy*. Oxford University Press.
Ge, M., Mao, R., & Cambria, E. (2023). A survey on computational metaphor processing techniques: From identification, interpretation, generation to application. *Artificial Intelligence Review*, *56*(Suppl 2), 1829–1895. https://doi.org/10.1007/s10462-023-10564-7
Geissinger, E. A., Khoo, C. L., Richmond, I. C., Faulkner, S. J., & Schneider, D. C. (2022). A case for beta regression in the natural sciences. *Ecosphere*, *13*(2), e3940. https://doi.org/10.1002/ecs2.3940
Gibbs, R. W., Jr. (2017). *Metaphor wars: Conceptual metaphors in human life*. Cambridge University Press.
Gibbs, R. W., Jr., & Cameron, L. (2008). The social-cognitive dynamics of metaphor performance. *Cognitive Systems Research*, *9*(1), 64–75. https://doi.org/10.1016/j.cogsys.2007.06.008
Goatly, A. (1997). *The language of metaphors* (First ed.). Routledge.
Goatly, A. (2011). *The language of metaphors* (Second ed.). Routledge.
Goddard, C. (2004). The ethnopragmatics and semantics of 'active metaphor'. *Journal of Pragmatics*, *36*(7), 1211–1230. https://doi.org/10.1016/j.pragma.2003.10.011
Halliday, M. A. K., & Hasan, R. (1985). *Language, context and text: A social semiotic perspective*. Deakin University Press.
Hanks, P. (2004). The syntagmatics of metaphor and idiom. *International Journal of Lexicography*, *17*(3), 245–274. https://doi.org/10.1093/ijl/17.3.245
Hanks, P. (2013). *Lexical analysis: Norms and exploitations*. MIT Press.
Honnibal, M., & Montani, I. (2017). Spacy. Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.
Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *25*(6), 417–441. https://doi.org/10.1037/h0071325
Kearns, M. S. (1987). *Metaphors of mind in fiction and psychology*. University Press of Kentucky.
Krennmayr, T., & Steen, G. (2017). VU Amsterdam metaphor corpus. In N. Ide & J. Pustejovsky (Eds.), *Handbook of Linguistic Annotation* (pp. 1053–1071). Springer Verlag. https://doi.org/10.1007/978-94-024-0881-2_39
Lakoff, G. (1987). Image metaphors. *Metaphor and Symbol*, *2*(3), 219–222. https://doi.org/10.1207/s15327868ms0203_4
Lakoff, G. (1990). The invariance hypothesis: Is abstract reason based on image-schemas? *Cognitive Linguistics*, *1*(1), 39–74. https://doi.org/10.1515/cogl.1990.1.1.39
Lakoff, G. (1993). How metaphor structures dreams: The theory of conceptual metaphor applied to dream analysis. *Dreaming*, *3*(2), 77. https://doi.org/10.1037/h0094373
Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press.
Lakoff, G., & Turner, M. (2009). *More than cool reason: A field guide to poetic metaphor*. University of Chicago press.
Leezenberg, M. (2007). Metaphor and metalanguage. *Baltic International Yearbook of Cognition, Logic and Communication*, *3*(1), 1–24. https://doi.org/10.4148/biyclc.v3i0.24
Lenci, A., & Sahlgren, M. (2023). *Distributional semantics*. Cambridge University Press.
Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4), 763–766. https://doi.org/10.1016/j.jesp.2013.03.013
Louw, B. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker (Ed.), *Text and technology: In honour of John Sinclair* (Vol. 157, p. 176). John Benjamins Publishing.
Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020). The lancaster sensorimotor norms: Multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, *52*(3), 1271–1291. https://doi.org/10.3758/s13428-019-01316-z
Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, *9*(11), 2579–2605.
Merriam-Webster. (n.d.). Merriam-webster.com dictionary. Retrieved April 25, 2024, from, https://www.merriam-webster.com/dictionary/<word>forallwords.
Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41. https://doi.org/10.1145/219717.219748
Mussell, J. E. P. (2014). Elemental forms: The newspaper as popular genre in the nineteenth century. *Media History*, *20* (1), 4–20. https://doi.org/10.1080/13688804.2014.880264

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Porto, M. D., & Romano, M. (2013). Newspaper metaphors: Reusing metaphors across media genres. *Metaphor and Symbol*, *28*(1), 60–73. https://doi.org/10.1080/10926488.2013.744572

Reid, J. N., & Katz, A. N. (2018). Vector space applications in metaphor comprehension. *Metaphor and Symbol*, *33*(4), 280–294. https://doi.org/10.1080/10926488.2018.1549840

Schmid, H.-J. (2020). *The dynamics of the linguistic system: Usage, conventionalization, and entrenchment*. Oxford University Press.

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. *Proceedings of the 9th Python in Science Conference*, Austin, TX (pp. 57–61). https://doi.org/10.25080/Majora-92bf1922-011

Semino, E. (2011). The adaptation of metaphors across genres. *Review of Cognitive Linguistics*, *9*(1), 130–152. https://doi.org/10.1075/rcl.9.1.07sem

Semino, E., Deignan, A., & Littlemore, J. (2013). Metaphor, genre, and recontextualization. *Metaphor and Symbol*, *28*(1), 41–59. https://doi.org/10.1080/10926488.2013.742842

Sinclair, J. (1991). *Corpus, concordance, collocation*. OUP.

Sinclair, J. (1996). The search for units of meaning. *Textus*, *9*(1), 75–106.

Soroka, S., Fournier, P., & Nir, L. (2019). Cross-national evidence of a negativity bias in psychophysiological reactions to news. *Proceedings of the National Academy of Sciences*, *116*(38), 18888–18892. https://doi.org/10.1073/pnas.1908369116

Soroka, S., & McAdams, S. (2015). News, politics, and negativity. *Political Communication*, *32*(1), 1–22. https://doi.org/10.1080/10584609.2014.881942

Sperber, D., & Wilson, D. (2008). A deflationary account of metaphors. In R. W. Gibbs Jr. (Ed.), *The Cambridge handbook of metaphor and thought* (pp. 84–105). Cambridge University Press.

Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A., Krennmayr, T., & Pasma, T. (2010). *A method for linguistic metaphor identification: From MIP to MIPVU*. John Benjamins Publishing.

Sullivan, K. (2013). *Frames and constructions in metaphoric language*. John Benjamins Publishing Company.

Svanlund, J. (2007). Metaphor and convention. *Cognitive Linguistics*, *18*(1), 47–89. https://doi.org/10.1515/COG.2007.003

Swales, J. M. (1990). *Genre analysis*. Cambridge University Press.

Talmy, L. (2000). *Toward a cognitive semantics. Volume, 1: Concept structuring systems*. The MIT Press.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., & Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45).

Appendices

Appendix 1. List of lexical items used

Table A1. All lexical items with frequency statistics and extraction information.

Lexical item	Frequency	% Metaphorical	N clusters
Archipelago	157	1.3	–
Backwater	212	40.1	4
Basin	1111	2.4	–
Bay	4454	9.1	7
Beach	10,163	0.6	5
Bog	309	6.1	–
Breeze	4278	7.4	6
Canyon	1834	1.5	–
Cascade	483	59.6	5
Cave	3862	6.2	3
Cavern	791	7.0	5
Cliff	2854	4.5	4
Coast	4706	0.5	–
Continent	1939	0.5	–
Crater	973	2.2	–
Cyclone	268	4.5	–
Deluge	221	38.5	2
Desert	5303	2.3	3
Downhill	57	1.8	–
Downpour	374	1.9	–
Draught	128	13.3	–
Dune	907	0.1	–
Fjord	88	0	–
Flurry	948	87.4	2
Fog	2558	24.9	9
Foothill	639	0	–
Forest	9452	2.4	4
Fountain	2171	8.2	6
Frost	756	8.1	6
Gale	349	11.5	–
Geyser	198	18.7	–
Glacier	681	3.1	–
Gorge	530	0.4	–
Gulf	396	31.8	4
Gust	908	34.3	7
Hail	455	38.5	6
Highland	166	0	–
Hill	10,930	0.6	5
Hillside	1446	0.6	–
Iceberg	457	39.4	4
Island	10,776	1.3	4
Lagoon	507	1.0	–
Lake	7782	1.8	3
Land	23,223	3.2	6
Landfill	669	1.3	–
Lowland	134	0	–
Mainland	884	0	–
Marsh	668	0.9	–
Marshland	122	0	–
Meadow	1635	1.0	–
Mire	94	45.7	–
Monsoon	258	1.2	–
Moor	194	0.5	–
Morass	140	85	1
Mount	613	4.6	–
Mountain	15,174	3.1	3
Murk	134	20.1	–

(*Continued*)

Table A1. (Continued).

Lexical item	Frequency	% Metaphorical	N clusters
Ocean	6485	3.5	–
Overcast	41	0	–
Pasture	1632	1.3	–
Peninsula	591	0.3	–
Pit	3673	29.5	–
Plateau	472	26.9	4
Pond	3417	4.8	3
Prairie	668	1.6	–
Precipice	249	24.9	5
Precipitation	120	0	–
Promontory	58	0	–
Puddle	1339	19	7
Quagmire	180	61.7	5
Quicksand	142	40.8	2
Rain	13,026	2.1	4
Rainfall	310	1.0	–
Rainstorm	253	5.1	–
Ravine	597	0.5	–
Reef	675	1.8	–
River	15,132	2.8	4
Sea	14,203	7.6	5
Shore	4276	0.8	–
Sleet	253	2	–
Slope	3316	5.4	4
Slough	100	8	–
Snow	11,230	1.8	4
Snowfall	331	2.1	–
Steppe	201	0.5	–
Strait	298	1	–
Stream	5741	37	9
Sunshine	1342	6.1	5
Swamp	1506	5.7	3
Torrent	644	66	4
Trench	1301	5.5	3
Valley	4618	1.2	11
Volcano	1002	4.9	–
Warming	1233	0.5	–
Waterfall	1012	12.5	5
Wave	12,283	32.8	8
Wetland	840	0	–
Wind	18,906	5	8

Appendix 2. Heuristics for automatic metaphor detection

In this Appendix, we describe the additional heuristics that we implemented to exclude classes of cases that were classified as "metaphorical" by MelBERT, but that we found to be non-metaphorical upon manual inspection. We give examples of the non-metaphorical cases that were successfully excluded in this way. The code for automatically implementing each heuristic can be found in the Supplementary Materials; grammatical categories and relations were determined using the SpaCy library (Honnibal & Montani, 2017), using the en_core_web_sm pipeline. Here, relevant segments of examples are underlined, and the detected metaphor is given in bold.

Heuristic 1: A type of compound

We excluded cases of noun compounds where the target word (e.g., *iceberg*) was modifying another noun (e.g., *iceberg lettuce*). In our manual inspection of MelBERT's output, we found that these cases were often erroneously flagged as metaphorical. Examples (1–4) illustrate compounds that were detected to be metaphorical: the first three are literal usages (though sometimes using different senses of the target nouns, as in example (2)) and should thus be excluded. We found that these types of examples were substantially more frequent than compounds where the first noun *was*

metaphorical, as in (4). Removing all compounds thus improves the Precision (i.e., accuracy) substantially while only somewhat hurting the model's Recall (i.e., completeness).

Furthermore, there were compounds removed this way that were non-compositional, or at least: opaque to many language users, but that are nonetheless not metaphorical. For instance *pit bull* originally referred to a specific dog associated with a (literal) pit for bullfighting; *iceberg lettuce* in (5) was associated with transport on metonymic icebergs (i.e., ice). We moreover note that in examples (6–7) below, *pit bull* is used metaphorically, but crucially, these metaphors do not derive their meaning from the basic meaning of *pit* but rather from the compound *pit bull*.

Examples of excluded cases (non-metaphorical except (4))

(1) Fort Belvoir in southern Fairfax County and the FBI 's relocation of its Northern Virginia field office from Tysons Corner to Manassas. Local officials and planners have criticized those moves.
(2) Like a bank shot, on a pool table: you kill your wife – and destroy her lover. With one bullet.
(3) The definition of cave man: "One who acts in a rough primitive manner."
(4) Soon Ponytail was peering in at Harris on one of those fake freeze frames Harris would trust in any movie from that moment on.
(5) The proportions of iceberg lettuce, tomato, onion, mustard, mayo and pickles are preternaturally right with the good griddled beef.
(6) The 34-year-old champ of show-and-tell has not only "metamorphosed from Marilyn Monroe to a pit bull," as Rosenbaum puts it, but also inspired a generation of women.
(7) He calls Tyson a "pit bull" for his fouls against Holyfield.

Heuristic 2: Proper names

We excluded cases where the target word was capitalized, such as (8–10). We found that MelBERT tended to categorize instances of proper names as metaphorical, perhaps because they are semantically more opaque and involve contexts that do not typically occur with the more basic (literal) meaning of a target word. However, if the capitalization occurred following a punctuation, MelBERT's verdict was not overruled.

Examples of excluded cases (non-metaphorical):

(1) But Stewart cautioned that she knows little about the Peninsula market and urged Anagnostou to have a study done to determine what people want [. . .].
(2) < p > Shari Mosier, Carol Stream < p > Cold facts on Barbie < p > It's not my goal to defend the Chicago [. . .].
(3) Price per square foot: $150 # MEADOWS ON THE PARKWAY, 4800 Baseline Road, Boulder # Sale price: $30.8 million.

Heuristic 3: Locative PP

We found that for our data set (consisting mostly of nouns that denote landscape features: $N = 72$, as well as nouns denoting weather, though fewer: $N = 26$), prepositional phrases tended to denote (literal) locations. However, when the context was somewhat creative (e.g., including a metaphor elsewhere, like *wildcatted* in (7)) or when the context featured transitions between semantically divergent topics (e.g., *stream* and *blood* in (10)), MelBERT tended to classify the target word as metaphorical. We therefore excluded cases where the target word was embedded in a prepositional phrase (typically denoting a location), with some notable exceptions to this rule listed below.

Examples of excluded cases (non-metaphorical)

(1) He wildcatted down in Mexico and set fire to a sludge pit next to the rig, then the wind blew the flames through a dry field.
(2) [She] waited for Leroy. The beauty of the birds as they rose above the meadow was majestic. A ring-necked pheasant in flight was one of the wonders of nature.
(3) [. . .] hiking near the ranch to look for Fremont pictographs in a cave. There are quite a few extended hikes in the area.
(4) Then, kneeling by the stream, he washed his hands clean of blood and grime.
(5) Before us, a short dive and a half-dozen breaststrokes from the shore, three massive whirlpools churned, mutating to whirlwinds as they rose, humming kazoo-like [. . .].

However, if the target word: (i) was further modified by another complement (specifically: a prepositional phrase with *of*, adjective, or noun, e.g., (16–18)), (ii) contained a possessive marker, or (iii) was used in a sentence containing a comparative word (e.g., *like* in (19)), then MelBERT's metaphorical verdict was not overruled.

Examples of included cases (metaphorical)

(1) Hard as she tried, she could not ignore the bad feeling churning in the pit of her stomach.
(2) [. . .] she turned slightly, staring dramatically through glossy streams of hair.
(3) explained a tall, sandy-haired executive well on the downslope to forty, who was waiting for Nick on the second-floor landing.
(4) A timeline was like a river cutting through a ravine, a gorge.

Appendix 3. Model variant analysis for clustering

In this Appendix, we take up possible alternative ways of approaching the clustering step of the method, as introduced in 'Cluster analysis over token-level distributional semantic representations'. Concretely, we consider three variations on the approach we report on in the paper:

(1) Instead of the final (11th) layer of the token representation in BERT, we can use the 7th layer, which Chronis & Erk (2020) have argued to model lexical semantic similarity optimally.
(2) In our main analysis, we dimensionality-reduce the BERT vectors, for reasons spelled out in 'Cluster analysis over token-level distributional semantic representations'. Chronis & Erk do not do so, and here we consider the effects of dimensionality-reducing the contextualized vectors vs. not doing so.
(3) We use Bayesian Gaussian Mixture Models to cluster the vectors; Chronis & Erk (2020) use *k*-Means, tuning the setting for *K* (the number of clusters). Here, we consider *k*-Means with values for *k* ranging from 1 to 10.

This leads to a total of 2×2×11 = 44 models considered. We will look at two possible ways the models may differ: first, the resulting clusters, and second the predictivity of the clustering over its genre dominance (i.e., the analysis of 'Understanding variation in metaphor usage').

First, considering the similarity of the resulting clusters. The similarity of any two clustering solutions to the same data can be measured with the Adjusted Rand Index. Given two clustering solutions, *x* and *y*, the Rand Index *r* is defined as $(a+b)/(a+b+c+d)$, where:

- *a* is the number of pairs of items that are part of the same cluster in *x* and part of the same cluster in *y*
- *b* is the number of pairs of items that are part of different clusters in *x* and part of different clusters in *y*
- *c* is the number of pairs of items that are part of the same cluster in *x* and part of different clusters in *y*
- *d* is the number of pairs of items that are part of different clusters in *x* and part of the same cluster in *y*

In other words, it is the proportion of pairs of items in a dataset for which the two clustering solutions agree on whether they belong to the same or different clusters.

The Adjusted Rand Index (ARI), then, is the Rand Index, but corrected for chance. This can be done in various ways, but here we follow the standard implementation in the Python library sklearn based on Chacón and Rastrojo (2023). Values of 0 represent at-chance similarity, 1 complete similarity, with increasing values in between indicating greater beyond-chance similarity of two clustering solutions.

Figure A4 presents all the pairwise ARI values for the 44 unique models, averaged across all the lemmas for each pairwise comparison. The shorthand notation is: "<BERT layer>_<PCA?>_<clustering algorithm>"; so 7_True_kM@10 means: layer 7, use of PCA is True and clustering algorithm is *k*Means. With an average pairwise ARI of 0.194, we see that the similarity of clustering solutions is on average above chance, even though substantial variation remains.

Crucially, however, that variation is not due to the use of BERT layer 7 vs. layer 11, or the use of PCA or not, or the use of a particular clustering algorithm (*k*-means vs. Bayesian Gaussian Mixture Models) or the setting of *k*. In fact, for each of these contrasts, the clustering similarity between models with different values on the contrasting model properties is consistently greater than between models with the same values, suggesting that these contrasts do not constitute the main drivers of clustering variability across models. Concretely: the average pairwise ARI for the 22 models using BERT layer 7 is 0.212 vs. 0.191 for the models using layer 11, but the pairwise ARI between all 22 layer-7 models and all 22 layer-11 models (i.e., 22×22 = 484 pairwise comparisons) is 0.284. Looking at the use of PCA, we find ARI = 0.199 among models using PCA, ARI = 0.175 among models not using PCA, and ARI = 0.309 for model comparisons across the two groups. Finally, for model choice, we find ARI = 0.277 among the Bayesian Gaussian Mixture Models vs. ARI = 0.173 among the *k*-Means models, and ARI = 0.302 across the two groups.

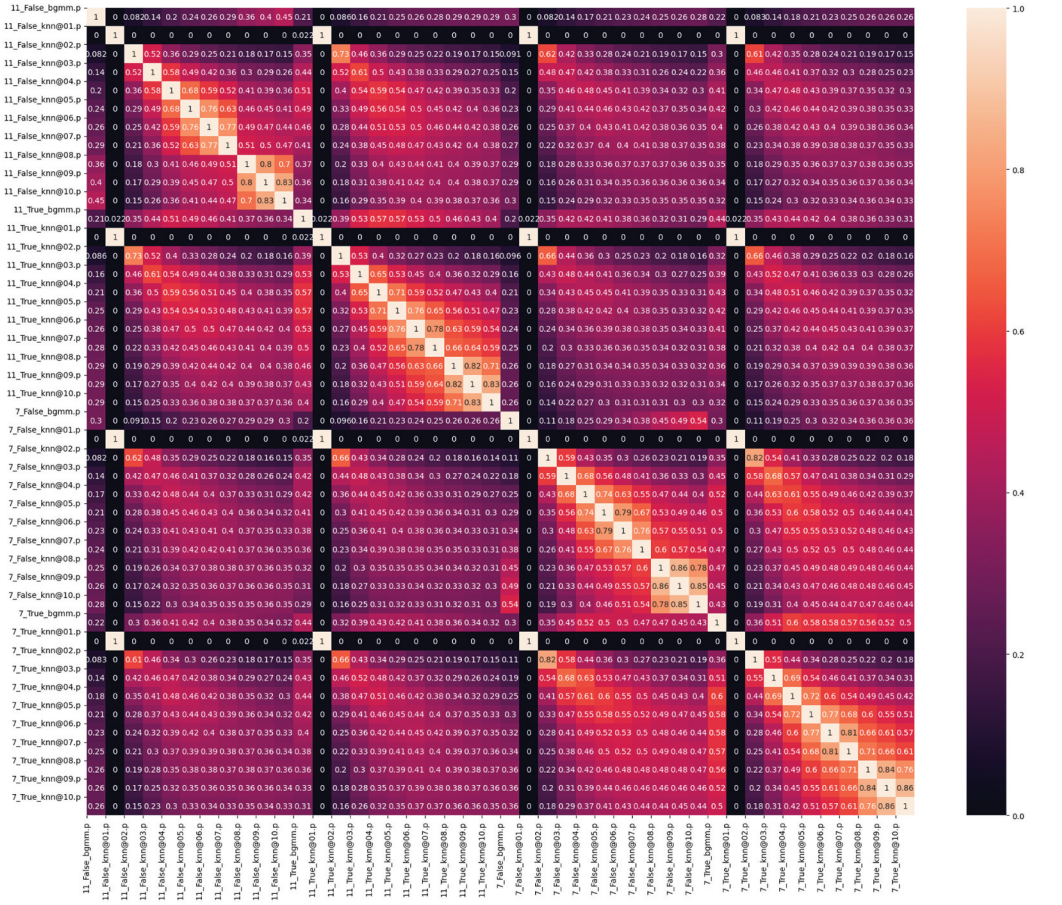

Figure A4. Similarity of clustering solutions (Adjusted Rand scores) between the 44 models.

Further analysis could assess if the lexicological sensibility of the clusters varies along with these contrasting model parameters (i.e., determining which model comes up with the most sensible clusters), but as this requires large-scale data annotation for sense assignment, we defer to future work to do so.

Next, we look at the performance of the 44 models on the regression analysis reported in 'Predicting genre variation on a cluster level'. For each model, we follow the preprocessing steps described there, namely outlier removal and using *z*-scores of the data. Figure A5 presents a summary view of the resulting 44 statistical models, with the level of significance being indicated by the color intensity (the darker, the more strongly significant, with absolute values of 3 meaning $p < .001$, values of 2 meaning $p < .01$, absolute values of 1 meaning $p < .05$, and values of 0 meaning non significant). All effects across all models were positive.

This representation allows us to see in one glance that the observed positive effect of concreteness on the proportion of fiction is robust across all models, reaching significance values of at the highest $p < .01$ across all models. Any other of the dependent variables display substantially more variable patterns of significance across the models, and in general are seldomly significant. Missing out on them is therefore not a quirk of the reported model in comparison to a broad set of model variants.

Taken together, the results discussed here suggest that while the clustering may vary across models and model settings, this variation is not primarily governed by specific model properties such as the use of particular representations, dimensionality reduction, or the specific clustering algorithm, and the association with concreteness holds across the model variants.

Appendix 4 :Predicting genre variation without outlier removal

As the removal of outliers in 'Predicting genre variation on a cluster level' substantially reduced the size of the dataset (from $N = 220$ clusters to $N = 131$), it would be worthwhile knowing if the findings from 'Predicting genre variation on a cluster level', viz. that only concreteness is a significant positive predictor of the

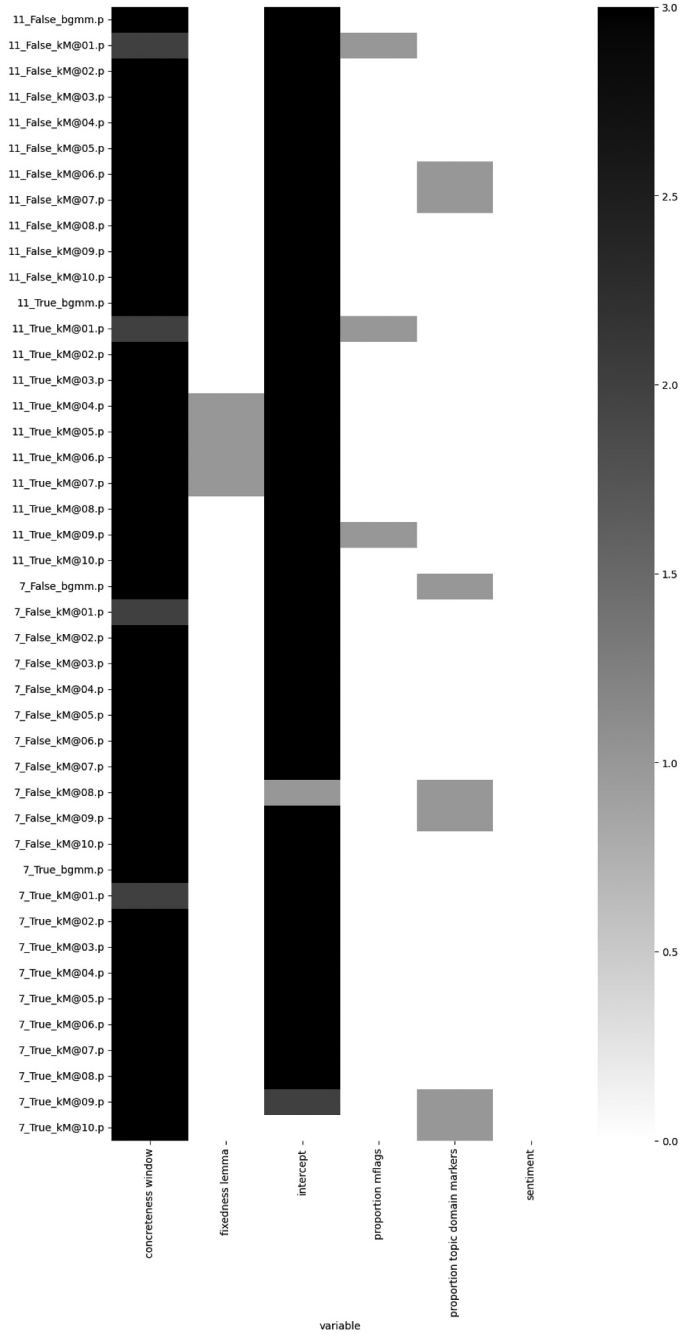

Figure A5. Regression results for the 44 models, where the (positive) effect of each independent variable on the dependent variable is given as an integer between 0 and 3, where the value is the standard level of significance of the *p*-value (3 = $p < .001$, 2 = $p < .01$, 1 = $p < .05$, 0 = n.S.).

proportion of tokens in a cluster that belong to the fiction subcorpus, hold on the larger dataset. Here, we present a supplemental analysis in which we run the same analysis except the outlier removal step (i.e., including a *z*-transform of the independent variables) led to the results in Table A6. When we compare the pattern of significance and the direction of the effects to those in Table 5 (i.e., the model applied to outlier-removed data), we see that the same pattern of effects is present, underscoring the robustness of the finding: the outliers do not behave different from the non-outliers in such a way that the pattern of results is affected.

Table A6. Supplemental results from beta regression predicting cluster bias on the basis of fixedness, concreteness, and valence, alongside the control variables of proportion MFlags and proportion TDMs without outlier removal ($N = 216$ items).

	Coefficient	st. err.	Z	*p*
Intercept	0.496	0.063	7.906	< .001
Concreteness	0.571	0.078	7.327	< .001
Fixedness	0.110	0.066	1.650	0.099
Valence	0.001	0.064	0.018	0.985
Prop. MFlags	0.118	0.079	1.500	0.134
Prop. TDMs	0.094	0.065	1.439	0.150