Brighter than Gold: Figurative Language in User Generated Comparisons

Vlad Niculae and **Cristian Danescu-Niculescu-Mizil**
MPI-SWS
Cornell University
vniculae@mpi-sws.org, cristian@mpi-sws.org

Abstract

Comparisons are common linguistic devices used to indicate the likeness of two things. Often, this likeness is not meant in the literal sense—for example, "I slept like a log" does not imply that logs actually sleep. In this paper we propose a computational study of figurative comparisons, or *similes*. Our starting point is a new large dataset of comparisons extracted from product reviews and annotated for figurativeness. We use this dataset to characterize figurative language in naturally occurring comparisons and reveal linguistic patterns indicative of this phenomenon. We operationalize these insights and apply them to a new task with high relevance to text understanding: distinguishing between figurative and literal comparisons. Finally, we apply this framework to explore the social context in which figurative language is produced, showing that similes are more likely to accompany opinions showing extreme sentiment, and that they are uncommon in reviews deemed helpful.

1 Introduction

> *In argument similes are like songs in love; they describe much, but prove nothing.*
>
> — Franz Kafka

Comparisons are fundamental linguistic devices that express the likeness of two things—be it entities, concepts or ideas. Given that their working principle is to emphasize the relation between the shared properties of two arguments (Bredin, 1998), comparisons can synthesize important semantic knowledge.

Often, comparisons are not meant to be understood literally. Figurative comparisons are an important figure of speech called *simile*. Consider the following two examples paraphrased from Amazon product reviews:

(1) Sterling is much cheaper than gold.

(2) Her voice makes this song shine brighter than gold.

In (1) the comparison draws on the relation between the price property shared by the two metals, *sterling* and *gold*. While (2) also draws on a common property (*brightness*), the polysemantic use (vocal timbre vs. light reflection) makes the comparison figurative.

Importantly, there is no general rule separating literal from figurative comparisons. More generally, the distinction between figurative and literal language is blurred and subjective (Hanks, 2006). Multiple criteria for delimiting the two have been proposed in the linguistic and philosophical literature—for a comprehensive review, see Shutova (2010)—but they are not without exceptions, and are often hard to operationalize in a computational framework. When considering the specific case of comparisons, such criteria cannot be directly applied.

Recently, the simile has received increasing attention from linguists and lexicographers (Moon, 2008; Moon, 2011; Hanks, 2013) as it became clearer that similes need to be treated separately from metaphors since they operate on fundamentally different principles (Bethlehem, 1996). Metaphors are linguistically simple structures hiding a complex mapping between two domains, through which many properties are transferred. For example the conceptual metaphor of *life as a journey* can be instantiated in many particular ways: *being at a fork in the road*, *reaching the end of the line* (Lakoff and Johnson, 1980). In contrast, the semantic context of similes tends to be very shallow, transferring a single property (Hanks, 2013). Their more explicit syntactic structure allows, in exchange, for more lexical creativity. As Hanks (2013) puts it, similes "tend to license all

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2008–2018,
October 25-29, 2014, Doha, Qatar.

sorts of logical mayhem." Moreover, the overlap between the expressive range of similes and metaphors is now known to be only partial: there are similes that cannot be rephrased as metaphors, and the other way around (Israel et al., 2004). This suggests that figurativeness in similes should be modeled differently than in metaphors. To further underline the necessity of a computational model for similes, we give the first estimate of their frequency in the wild: over 30% of comparisons are figurative.[1] We also confirm that a state of the art metaphor detection system performs poorly when applied directly to the task of detecting similes.

In this work we propose a computational study of figurative language in comparisons. To this end, we build the first large collection of naturally occurring comparisons with figurativeness annotation, which we make publicly available. Using this resource we explore the linguistic patterns that characterize similes, and group them in two conceptually distinctive classes. The first class contains cues that are agnostic of the context in which the comparison appears (domain-agnostic cues). For example, we find that the higher the semantic similarity between the two arguments, the less likely it is for the comparison to be figurative—in the examples above, *sterling* is semantically very similar to *gold*, both being metals, but *song* and *gold* are semantically dissimilar. The second type of cues are domain-specific, drawing on the intuition that the domain in which a comparison is used is a factor in determining its figurativeness. We find, for instance, that the less specific a comparison is to the domain in which it appears, the more likely it is to be used in a figurative sense (e.g., in example (2), *gold* is very unexpected in the musical domain).

We successfully exploit these insights in a new prediction task relevant to text understanding: discriminating figurative comparisons from literal ones. Encouraged by the high accuracy of our system—which is within 10% of that obtained by human annotators—we automatically extend the figurativeness labels to 80,000 comparisons occurring in product reviews. This enables us to conduct a fine-grained analysis of how comparison usage interacts with their social context, opening up a research direction with applications in sentiment analysis and opinion mining. In particular we find that figurative comparisons are more likely to accompany reviews showing extreme sentiment, and that they are uncommon in opinions deemed as being helpful. To the best of our knowledge, this is the first time figurative language is tied to the social context in which it appears.

To summarize, the main contributions of this work are as follows:

- it introduces the first large dataset of comparisons with figurativeness annotations (Section 3);
- it unveils new linguistic patterns characterizing figurative comparisons (Section 4);
- it introduces the task of distinguishing figurative from literal comparisons (Section 5);
- it establishes the relation between figurative language and the social context in which it appears (Section 6).

2 Further Related Work

Corpus studies on figurative language in comparisons are scarce, and none directly address the distinction between figurative and literal comparisons. Roncero et al. (2006) observed, by searching the web for several stereotypical comparisons (e.g., *education is like a stairway*), that similes are more likely to be accompanied by explanations than equivalent metaphors (e.g., *education is a stairway*). Related to figurativeness is irony, which Veale (2012a) finds to often be lexically marked. By using a similar insight to filter out ironic comparisons, and by assuming that the rest are literal, Veale and Hao (2008) learn stereotypical knowledge about the world from frequently compared terms. A similar process has been applied to both English and Chinese by Li et al. (2012), thereby encouraging the idea that the trope behaves similarly in different languages. A related system is the *Jigsaw Bard* (Veale and Hao, 2011), a thesaurus driven by figurative conventional similes extracted from the Google N-grams. This system aims to build and generate canned expressions by using items frequently associated with the simile pattern above. An extension of the principles of the *Jigsaw Bard* is found in *Thesaurus Rex* (Veale and Li, 2013), a data-driven partition of words into ad-hoc categories. *Thesaurus Rex* is constructed using simple comparison and hypernym patterns

[1] This estimate is based on the set of noun-noun comparisons with non-identical arguments collected for this study from Amazon.com product reviews.

and is able to provide weighted lists of categories for given words.

In text understanding systems, literal comparisons are used to detect analogies between related geographical places (Lofi et al., 2014). Tandon et al. (2014) use relative comparative patterns (e.g., *X is heavier than Y*) to enrich a common-sense knowledge base. Jindal and Liu (2006) extract graded comparisons from various sources, with the objective of mining consumer opinion about products. They note that identifying objective vs. subjective comparisons—related to literality—is an important future direction. Given that many comparisons are figurative, a system that discriminates literal from figurative comparisons is essential for such text understanding and information retrieval systems.

The vast majority of previous work on figurative language focused on metaphor detection. Tsvetkov et al. (2014a) propose a cross-lingual system based on word-level conceptual features and they evaluate it on Subject-Verb-Object triples and Adjective-Noun pairs. Their features include and extend the idea of abstractness used by Turney et al. (2011) for Adjective-Noun metaphors. Hovy et al. (2013) contribute an unrestricted metaphor corpus and propose a method based on tree kernels. Bridging the gap between metaphor identification and interpretation, Shutova and Sun (2013) proposed an unsupervised system to learn source-target domain mappings. The system fits conceptual metaphor theory (Lakoff and Johnson, 1980) well, at the cost of not being able to tackle figurative language in general, and similes in particular, as similes do not map entire domains to one another. Since similes operate on fundamentally different principles than metaphors, our work proposes a computational approach tailored specifically for comparisons.

3 Background and Data

3.1 Structure of a comparison

Unlike metaphors, which are generally unrestricted, comparisons are more structured but also more lexically and semantically varied. This enables a more structured computational representation of which we take advantage. The constituents of a comparison according to Hanks (2012) are:

- the TOPIC, sometimes called tenor: it is usually a noun phrase and acts as logical subject;
- the VEHICLE: it is the object of the comparison and is also usually a noun phrase;
- the shared PROPERTY or ground: it expresses what the two entities have in common—it can be explicit but is often implicit, left for the reader to infer;
- the EVENT (eventuality or state): usually a verb, it sets the frame for the observation of the common property;
- the COMPARATOR: commonly a preposition (*like*) or part of an adjectival phrase (*better than*), it is the trigger word or phrase that marks the presence of a comparison.

The literal example (1) would be segmented as:

> [Sterling /TOPIC] [is /EVENT] much [cheaper /PROPERTY] [than /COMPARATOR] [gold /VEHICLE]

3.2 Annotation

People resort to comparisons often when making descriptions, as they are a powerful way of expressing properties by example. For this reason we collect a dataset of user-generated comparisons in Amazon product reviews (McAuley and Leskovec, 2013), where users have to be descriptive and precise, but also to express personal opinion. We supplement the data with a smaller set of comparisons from WaCky and WaCkypedia (Baroni et al., 2009) to cover more genres. In preliminary work, we experimented with dependency parse tree patterns for extracting comparisons and labeling their parts (Niculae, 2013). We use the same approach, but with an improved set of patterns, to extract comparisons with the COMPARATORS *like*, *as* and *than*.[2] We keep only the matches where the TOPIC and the VEHICLE are nouns, and the PROPERTY, if present, is an adjective, which is the typical case. Also, the head words of the constituents are constrained to occur in the distributional resources used (Baroni and Lenci, 2010; Faruqui and Dyer, 2014).[3]

[2] We process the review corpus with part-of-speech tagging using the IRC model for *TweetNLP* (Owoputi et al., 2013; Forsyth and Martell, 2007) and dependency parsing using the *TurboParser* standard model (Martins et al., 2010).

[3] Due to the strong tendency of comparisons with the same TOPIC and VEHICLE to be trivially literal in the WaCky examples, we filtered out such examples from the Amazon product reviews. We also filtered proper nouns using a capitalization heuristic.

We proceed to validate and annotate for figurativeness a random sample of the comparisons extracted using the automated process described above. The annotation is performed using crowdsourcing on the Amazon Mechanical Turk platform, in two steps. First, the annotators are asked to determine whether a displayed sentence is indeed a comparison between the highlighted words (TOPIC and VEHICLE). Sentences qualified by two out of three annotators as comparisons are used in the second round, where the task is to rate how metaphorical a comparison is. We use a scale of 1 to 4 following Turney et al. (2011), and then binarize to consider scores of 1–2 as literal and 3–4 as figurative. Finally, in this work we only consider comparisons where all three annotators agree on this binary notion of figurativeness. For both tasks, we provide guidelines mostly in the form of examples and intuition, motivated on one hand by the annotators not having specialized knowledge, and on the other hand by the observation that the literal-figurative distinction is subjective. All annotators have the *master worker* qualification, reside in the U.S. and completed a linguistic background questionnaire that verifies their experience with English. In both tasks, control sentences with confidently known labels are used to filter low quality answers; in addition, we test annotators with a simple paraphrasing task shown to be effective for eliciting and verifying linguistic attention (Munro et al., 2010). Both tasks seem relatively difficult for humans, with inter-annotator agreement given by Fleiss' k of 0.48 for the comparison identification task and of 0.54 for the figurativeness annotation after binarization. This is comparable to 0.57 reported by Hovy et al. (2013) for general metaphor labeling. We show some statistics about the collected data in Table 1. Overall, this is a costly process: out of 2400 automatically extracted comparison candidates, about 60% were deemed by the annotators to be actual comparisons and only 12% end up being selected confidently enough as figurative comparisons.

Our dataset of human-filtered comparisons, with the scores given by the three annotators, is made publicly available to encourage further work.[4] This also includes about 400 comparisons where the annotators do not agree perfectly on binary figurativeness. Such cases can be interesting to other analyses, even if we don't consider them in our experiments. It is worth noting that the existing corpora annotated for metaphor cannot be directly used to study comparisons. For example, in TroFi (Birke and Sarkar, 2006), a corpus of 6436 sentences annotated for figurativeness, we only find 42 noun-noun comparisons with sentence-level (thus noisy) figurativeness labels.

Domain	fig.	lit.	% fig.
Books	177	313	36%
Music	45	68	40%
Electronics	23	105	18%
Jewelery	9	126	7%
WaCky	19	79	19%
Total	273	609	31%

Table 1: Figurativeness annotation results. Only comparisons where all three annotators agree are considered.

4 Linguistic Insights

We now proceed to exploring the linguistic patterns that discriminate figurative from literal comparisons. We consider two broad classes of cues, which we discuss next.

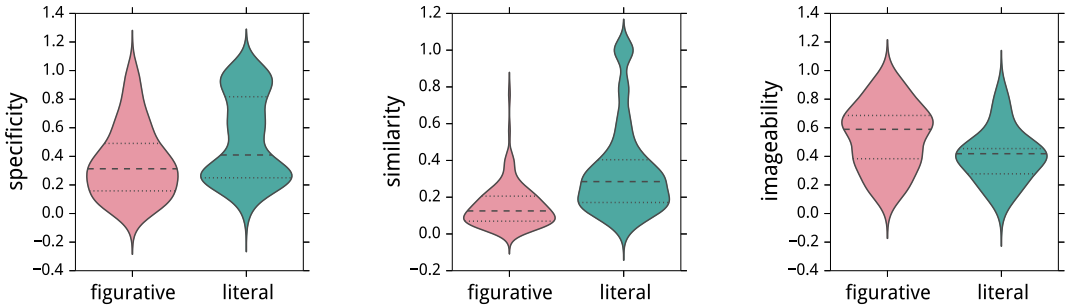
4.1 Domain-specific cues

Figurative language is often used for striking effects, and comparisons are used to describe new things in terms of something given (Hanks, 2013). Since the norms that define what is surprising and what is well-known vary across domains, we expect that such contextual information should play an important role in figurative language detection. This is a previously unexplored dimension of figurative language, and Amazon product reviews offer a convenient testbed for this intuition since category information is provided.

Specificity To estimate whether a comparison can be considered striking in a particular domain—whether it references images or ideas that are unexpected in its context—we employ a simple measure of word *specificity* with respect to a domain: the ratio of the word frequency within the domain and the word frequency in all domains being considered.[5] It should be noted that specificity is not purely a function of the word, but

[4] http://vene.ro/figurative-comparisons/

[5] We measure specificity for the VEHICLE, PROPERTY and EVENT.



(a) VEHICLE specificity. (b) TOPIC-VEHICLE similarity. (c) Imageability of the PROPERTY.

Figure 1: Distribution of some of the features we use, across literal and figurative comparisons in the test set. The profile of the plot is a kernel density estimation of the distribution, and the markers indicate the median and the first and third quartiles.

of the word and the context in which it appears. A comparison in the *music* domain that involves melodies is not surprising:

> But the title song really feels like a pretty bland vocal **melody** [...]

But the same word can play a very different role in another context, for example, *book* reviews:

> Her books are like sweet **melodies** that flow through your head.

Indeed, the word *melody* has a specificity of 96% in the *music* domain and only of 3% in the *books* domain.

An analysis on the labeled data confirms that literal comparisons do indeed tend to have more domain-specific VEHICLES (Mann-Whitney U test, $p < 0.01$) than figurative ones. Furthermore, the distribution of specificity across both types of comparisons, as shown in Figure 1a, has the appearance of a mixture model of general and specific words. Figurative comparison VEHICLES largely exhibit only the general component of the mixture.[6]

Domain label An analysis of the annotation results reveals that the percentage of comparisons that are figurative differs widely across domains, as indicated in the last column in Table 1. This suggests that simply knowing the domain of a text can serve to adjust some prior expectation about figurative language presence and therefore improve detection. We test this hypothesis using a Z-test comparing all Amazon categories. With the exception of *books* and *music* reviews, that have similar ratios, all other pairs of categories show significantly different proportions of figurative comparisons ($p < 0.01$).

4.2 Domain-agnostic cues

Linguistic studies of figurative language suggest that there is a fundamental generic notion of figurativeness. We attempt to capture this notion in the context of comparisons using syntactic and semantic information.

Topic-Vehicle similarity The default role of literal comparisons is to assert similarity of things. Therefore, we expect that a high semantic similarity between the TOPIC and the VEHICLE of a comparison is a sign of literal usage, as we previously hypothesized in preliminary work (Niculae, 2013). To test this hypothesis, we compute TOPIC-VEHICLE similarity using Distributional Memory (Baroni and Lenci, 2010), a freely available distributional semantics resource that captures word relationships through grammatical role co-occurrence.

By applying this measure to our data, we find that there is indeed an important difference between the distributions of TOPIC-VEHICLE similarity in figurative and literal comparisons (shown in Figure 1b); the means of the two distributions are significantly different (Mann-Whitney $p < 0.01$).

Metaphor-inspired features We also seek to understand to what extent insights provided by computational work on metaphor detection can be

[6]The mass around 0.25 in Figure 1a is largely explained by generic words such as *thing, others, nothing, average* and barely specific words like *veil, reputation, dream, garbage*.

	more concrete	less concrete
more imageable	*cinnamon, kiss*	*devil, happiness*
less imageable	*casque, pugilist*	*aspect, however*

Table 2: Examples of words with high and low concreteness and imageability scores from the MRC Psycholinguistic Database.

applied in the context of comparisons. To that end we consider features shown to provide state of the art performance in the task of metaphor detection (Tsvetkov et al., 2014a): abstractness, imageability and supersenses.

Abstractness and imageability features are derived from the MRC Psycholinguistic Database (Coltheart, 1981), a dictionary based on manually annotated datasets of psycholinguistic norms. Imageability is the property of a word to arouse a mental image, be it in the form of a mental picture, sound or any other sense. Concreteness is defined as "any word that refers to objects, materials or persons," while abstractness, at the other end of the spectrum, is represented by words that cannot be usually experienced by the senses (Paivio et al., 1968). Table 2 shows a few examples of words with high and low concreteness and imageability scores. Supersenses are a very coarse form of meaning representation. Tsvetkov et al. (2014a) used WordNet (Miller, 1995) semantic classes for nouns and verbs, for example *noun.body*, *noun.animal*, *verb.consumption*, or *verb.motion*. For adjectives, Tsvetkov et al. (2014b) developed and made available a novel classification in the same spirit.[7] We compute abstractness, imageability and supersenses for the TOPIC, VEHICLE, EVENT, and PROPERTY.[8] We concatenate these features with the raw vector representations of the constituents, following Tsvetkov et al. (2014a).

We find that such features relate to figurative comparisons in a meaningful way. For example, out of all comparisons with explicit properties, figurative comparisons tend to have properties that are more imageable (Mann-Whitney $p < 0.01$), as illustrated by Figure 1c. This is in agreement with Hanks (2005), who observed that similes are characterized by their appeal to sensory imagination.

Definiteness We introduce another simple but effective syntactic cue that relates to concreteness: the presence of a definite article versus an indefinite one (or none at all). We search for the indefinite articles *a* and *an* and the definite article *the* in each component of a comparison.

We find that similes tend to have indefinite articles in the VEHICLE more often and definite articles less often (Mann-Whitney $p < 0.01$). In particular, 59% of comparisons where the VEHICLE has a indefinite article are figurative, as opposed to 13% of the comparisons where VEHICLE has a definite article.

5 Prediction Task

We now turn to the task of predicting whether a comparison is figurative or literal. Not only does this task allow us to assess and compare the efficiency of the linguistic cues we discussed, but it is also highly relevant in the context of natural language understanding systems.

We conduct a logistic regression analysis, and compare the efficiency of the features derived from our analysis to a **bag of words** baseline. In addition to the features inspired by the previously described linguistic insights, we also try to computationally capture the lexical usage patterns of comparisons using a version of bag of words adapted to the comparison structure. In this **slotted bag of words** system, features correspond to occurrence of words within constituents (e.g., *bright* $\in$ PROPERTY).

We perform a stratified split of our comparison dataset into equal train and test sets (each set containing 408 comparisons, out of which 134 are figurative),[9] and use a 5-fold stratified cross validation over the training set to choose the optimal value for the logistic regression regularization parameter and the type of regularization (ℓ_1 or ℓ_2) for each feature set.[10]

[7] Following Tsvetkov et al. (2014a) we train a classifier to predict these features from a vector space representation of a word. We use the same cross-lingually optimized representation from Faruqui and Dyer (2014) and a simpler classifier, a logistic regression, which we find to perform as well as the random forests used in Tsvetkov et al. (2014a). We treat supersense prediction as a multi-label problem and apply a one-versus-all transformation, effectively learning a linear classifier for each supersense.

[8] If the PROPERTY is implicit, all corresponding features are set to zero. An extra binary feature indicates whether the PROPERTY is explicit or implicit.

[9] The entire analysis described in Section 4 is only conducted on the training set. Also, in order to ensure that we are assessing the performance of the classifier on unseen comparisons, we discard from our dataset all those with the same TOPIC and VEHICLE pair.

[10] We use the logistic regression implementation of `liblinear` (Fan et al., 2008) wrapped by the `scikit-learn` library (Pedregosa et al., 2011).

Model	# features	Acc.	P	R	F_1	AUC
Bag of words	1970	0.79	0.63	0.84	0.72	0.87
Slotted bag of words	1840	0.80	0.64	0.90	0.75	0.89
Domain-agnostic cues	357	0.81	0.70	0.74	0.72	0.90
only metaphor inspired	345	0.75	0.60	0.72	0.65	0.84
Domain-specific cues	8	0.69	0.51	0.81	0.63	0.76
All linguistic insight cues	365	0.86	0.76	0.83	0.79	0.92
Full	2202	0.88	0.80	0.84	0.82	0.94
Human	-	0.96	0.92	0.96	0.94	-

Table 3: Classification performance on the test set for the different sets of features we considered; human performance is shown for reference.

Classifier performance The performance on the classification task is summarized in Table 3. We note that the **bag of words** baseline is remarkably strong, because of common idiomatic similes that can be captured through keywords. Our **full** system (which relies on our linguistically inspired cues discussed in Section 4 in addition to slotted bag of words) significantly outperforms the bag of words baseline and the slotted bag of words system in terms of accuracy, F_1 score and AUC ($p < 0.05$),[11] suggesting that linguistic insights complement idiomatic simile matching. Importantly, a system using only our **linguistic insight cues** also significantly improves over the baseline in terms of accuracy and AUC and it is not significantly different from the full system in terms of performance, in spite of having about an order of magnitude fewer features. It is also worth noting that the **domain-specific cues** play an important role in bringing the performance to this level by capturing a different aspect of what it means for a comparison to be figurative.

The features used by the state of the art metaphor detection system of Tsvetkov et al. (2014a), adapted to the comparison structure, perform poorly by themselves and do not improve significantly over the baseline. This is consistent with the theoretical motivation that figurativeness in comparisons requires special computational treatment, as discussed in Section 1. Furthermore, the linguistic insight features not only significantly outperform the metaphor inspired features ($p < 0.05$), but are also better at exploiting larger amounts of data, as shown in Figure 2.

[11] All statistical significance results in this paragraph are obtained from 5000 bootstrap samples.

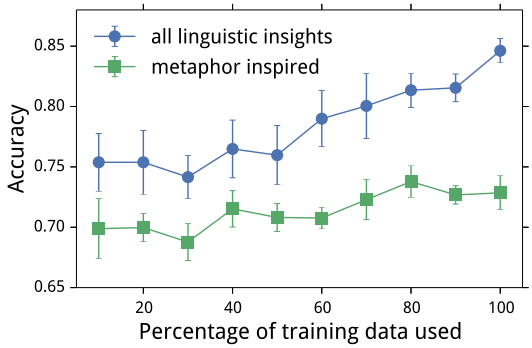


Figure 2: Learning curves. Each point is obtained by fitting a model on 10 random subsets of the training set. Error bars show 95% confidence intervals.

Comparison to human performance To gauge how well humans would perform at the classification task on the actual test data, we perform another Amazon Mechanical Turk evaluation on 140 examples from the test set. For the evaluation, we use majority voting between the three annotators,[12] and compare to the agreed labels in the dataset. Estimated human accuracy is 96%, placing our full system within 10% of human accuracy.

Feature analysis The predictive analysis we perform allows us to investigate to what extent the features inspired by our linguistic insights have discriminative power, and whether they actually cover different aspects of figurativeness.

[12] Majority voting helps account for the noise inherent to crowdsourced annotation, which is less accurate than professional annotation. Taking the less optimistic invididual turker answers, human performance is on the same level as our full system.

Feature	Coef.	Example where the feature is positively activated
TOPIC-VEHICLE similarity	−11.3	the older *man* was wiser and stronger than the *boy*
VEHICLE specificity	−5.8	the cord is more durable than the *adapter* [Electronics]
VEHICLE imageability	4.9	the explanations are as clear as *mud*
VEHICLE communication supersense	−4.6	the book reads like six short *articles*
VEHICLE indefiniteness	4.0	his fame drew foreigners to him like *a* magnet
life ∈ VEHICLE	7.1	the hero is truly larger than *life*: godlike, yet flawed
picture ∈ VEHICLE	−6.0	the necklace looks just like the *picture*
other ∈ VEHICLE	−5.9	this one is just as nice as the *other*
others ∈ VEHICLE	−5.5	some songs are more memorable than *others*
crap ∈ VEHICLE	4.7	the headphones sounded like *crap*

Table 4: Top 5 linguistic insight features (top) and slotted bag of words features (bottom) in the full model and their logistic regression coefficients. A positive coefficient means the feature indicates figurativeness.

Table 4 shows the best linguistic insight and slotted bag of words features selected by the full model. The strongest feature by far is the semantic similarity between the TOPIC and the VEHICLE. By itself, this feature gets 70% accuracy and 61% F_1 score.

The rest of the top features involve mostly the VEHICLE. This suggests that the VEHICLE is the most informative element of a comparison when it comes to figurativeness. Features involving other constituents also get selected, but with slightly lower weights, not making it to the top.

VEHICLE specificity is one of the strongest features, with positive values indicating literal comparisons. This confirms our intuition that domain information is important to discriminate figurative from literal language.

Of the adapted metaphor features, the noun communication supersense and the imageability of the VEHICLE make it to the top. Nouns with low communication rating occurring in the training set include *puddles, arrangements, carbohydrates* while nouns with high communication rating include *languages* and *subjects*.

Presence of an indefinite article in the VEHICLE is a strong indicator of figurativeness. By themselves, the definiteness and indefiniteness features perform quite well, attaining 78% accuracy and 67% F_1 score.

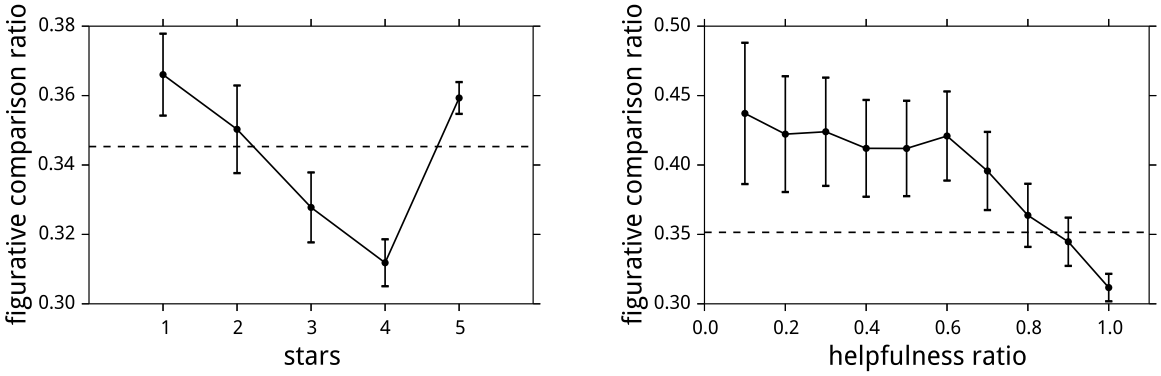
The salient bag of words features correspond to specific types of comparisons. The words *other* and *others* in the VEHICLE indicate comparisons between the same kind of arguments, for example *some songs are more memorable than others,* and these are likely to be literal. The word *picture* is specific to the review setting, as products are accompanied by photos, and for certain kinds of products, the resemblance of the product with the image is an important factor for potential buyers.[13] The bag of words systems are furthermore able to learn idiomatic comparisons by identifying common figurative VEHICLES such as *life* and *crap*, corresponding to fixed expressions such as *larger than life*.

Error analysis Many of the errors made by our full system involve indirect semantic mechanisms such as metonymy. For example, the false positive *the typeface was larger than most books* really means *larger than the typefaces found in most books*, but without the implicit expansion the meaning can appear figurative. A similar kind of ellipsis makes the example *a lot [of songs] are even better than sugar* be wrongly classified as literal. Another source of error is polysemy. Examples like *the rejuvelac formula is about 10 times better than yogurt* are misclassified because of the multiple meanings of the word *formula*, one being closely related to yogurt and food, but the more common ones being general and abstract, suggesting figurativeness.

6 Social Correlates

The advantage of studying comparisons situated in a social context is that we can understand how their usage interacts with internal and external human factors. An internal factor is the sentiment of

[13] This feature is highly correlated with the domain: it appears 25 times in the training set, 24 of which in the *jewelery* domain and once in *book* reviews.



(a) Figurative comparisons are more likely to be found in reviews with strongly polarized sentiment.

(b) Helpful reviews are less likely to contain figurative comparisons.

Figure 3: Interaction between figurative language and social context aspects. Error bars show 95% confidence intervals. The dashed horizontal line marks the average proportion of figurative comparisons. In Figure 3b the average proportion is different because we only consider reviews rated by at least 10 readers.

the user towards the reviewed product, indicated by the star rating of the review. An external factor present in the data is how helpful the review is perceived by other users. In this section we analyze how these factors interact with figurative language in comparisons.

To gain insight about fine grained interactions with human factors at larger scale, we use our classifier to find over 80,000 figurative and literal comparisons from the same four categories. The trends we reveal also hold significantly on the manually annotated data.

Sentiment While it was previously noted that similes often transmit strong affect (Hanks, 2005; Veale, 2012a; Veale, 2012b), the connection between figurativeness and sentiment was never empirically validated. The setting of product reviews is convenient for investigating this issue, since the star ratings associated with the reviews can be used as sentiment labels. We find that comparisons are indeed significantly more likely to be figurative when the users express strong opinions, i.e., in one-star or five-star reviews (Mann-Whitney $p < 0.02$ on the manually annotated data). Figure 3a shows how the proportion of figurative comparisons varies with the polarity of the review.

Helpfulness It is also interesting to understand to what extent figurative language relates to the external perception of the content in which it appears. We find that comparisons in helpful reviews[14] are less likely to be figurative. Figure 3b shows a near-constant high ratio of figurative comparisons among unhelpful and average reviews; as helpfulness increases, figurative comparisons become less frequent. We further validate that this effect is not a confound of the distribution of helpfulness ratings across reviews of different polarity by controlling for the star rating: given a fixed star rating, the proportion of figurative comparisons is still lower in helpful (helpfulness over 50%) than in unhelpful (helpfulness under 50%) reviews; this difference is significant (Mann-Whitney $p < 0.01$) for all classes of ratings except one-star. The size of the manually annotated data does not allow for star rating stratification, but the overall difference is statistically significant (Mann-Whitney $p < 0.01$). This result encourages further experimentation to determine whether there is a causal link between the use of figurative language in user generated content and its external perception.

7 Conclusions and Future Work

This work proposes a computational study of figurative language in comparisons. Starting from a new dataset of naturally occurring comparisons with figurativeness annotation (which we make publicly available) we explore linguistic patterns that are indicative of similes. We show that these

[14] In order to have reliable helpfulness scores, we only consider reviews that have been rated by at least by ten readers.

insights can be successfully operationalized in a new prediction task: distinguishing literal from figurative comparisons. Our system reaches accuracy that is within 10% of human performance, and is outperforming a state of the art metaphor detection system, thus confirming the need for a computational approach tailored specifically to comparisons. While we take a data-driven approach, our annotated dataset can be useful for more theoretical studies of the kinds of comparisons and similes people use.

We discover that domain knowledge is an important factor in identifying similes. This suggests that future work on automatic detection of figurative language should consider contextual parameters such as the topic and community where the content appears.

Furthermore, we are the first to tie figurative language to the social context in which it is produced and show its relation to internal and external human factors such as opinion sentiment and helpfulness. Future investigation into the causal effects of these interactions could lead to a better understanding of the role of figurative language in persuasion and rhetorics.

In our work, we consider common noun TOPICS and VEHICLES and adjectival PROPERTIES. This is the most typical case, but supporting other parts of speech—such as proper nouns, pronouns, and adverbs—can make a difference in many applications. Capturing compositional interaction between the parts of the comparison could lead to more flexible models that give less weight to the VEHICLE.

This study is also the first to estimate how prevalent similes are in the wild, and reports that about one third of the comparisons we consider are figurative. This is suggestive of the need to build systems that can properly process figurative comparisons in order to correctly harness the semantic information encapsulated in comparisons.

Acknowledgements

We would like to thank Yulia Tsvetkov for constructive discussion about figurative language and about her and her co-authors' work. We are grateful for the suggestions of Patrick Hanks, Constantin Orăsan, Sylviane Cardey, Izabella Thomas, Ekaterina Shutova, Tony Veale, and Niket Tandon. We extend our gratitude to Julian McAuley for preparing and sharing the Amazon review dataset. We are thankful to the anonymous reviewers, whose comments were *like a breath of fresh air*. We acknowledge the help of the Amazon Mechanical Turk annotators and of the MPI-SWS students involved in pilot experiments.

Vlad Niculae was supported in part by the European Commission, Education & Training, Erasmus Mundus: EMMC 2008-0083, Erasmus Mundus Masters in NLP & HLT.

References

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Louise Shabat Bethlehem. 1996. Simile and figurative language. *Poetics Today*, 17(2):203–240.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *Proceedings of EACL*.

Hugh Bredin. 1998. Comparisons and similes. *Lingua*, 105(1):67–78.

Max Coltheart. 1981. The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, 33(4):497–505.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of EACL*.

Eric N Forsyth and Craig H Martell. 2007. Lexical and discourse analysis of online chat dialog. In *Proceedings of ICSC*.

Patrick Hanks. 2005. Similes and sets: The English preposition 'like'. In R. Blatná and V. Petkevic, editors, *Languages and Linguistics: Festschrift for Fr. Cermak*. Charles University, Prague.

Patrick Hanks. 2006. Metaphoricity is gradable. *Trends in Linguistic Studies and Monographs*, 171:17.

Patrick Hanks. 2012. The roles and structure of comparisons, similes, and metaphors in natural language (an analogical system). Presented at the Stockholm Metaphor Festival.

Patrick Hanks. 2013. *Lexical Analysis: Norms and Exploitations*. MIT Press.

Dirk Hovy, Shashank Srivastava, Sujay Kumar Jauhar, Mrinmaya Sachan, Kartik Goyal, Huiying Li, Whitney Sanders, and Eduard Hovy. 2013. Identifying metaphorical word use with tree kernels. In *Proceedings of the NAACL Workshop on Metaphors for NLP*.

M. Israel, J.R. Harding, and V. Tobin. 2004. On simile. *Language, Culture, and Mind. CSLI Publications*.

Nitin Jindal and Bing Liu. 2006. Identifying comparative sentences in text documents. In *Proceedings of SIGIR*.

George Lakoff and Mark Johnson. 1980. *Metaphors we live by*. University of Chicago Press.

Bin Li, Jiajun Chen, and Yingjie Zhang. 2012. Web based collection and comparison of cognitive properties in English and Chinese. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*.

Christoph Lofi, Christian Nieke, and Nigel Collier. 2014. Discriminating rhetorical analogies in social media. In *Proceedings of EACL*.

André FT Martins, Noah A Smith, Eric P Xing, Pedro MQ Aguiar, and Mário AT Figueiredo. 2010. Turbo parsers: Dependency parsing by approximate variational inference. In *Proceedings of EMNLP*.

Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: Understanding rating dimensions with review text. In *Proceedings of RecSys*.

George A Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Rosamund Moon. 2008. Conventionalized as-similes in English: A problem case. *International Journal of Corpus Linguistics*, 13(1):3–37.

Rosamund Moon. 2011. Simile and dissimilarity. *Journal of Literary Semantics*, 40(2):133–157.

Robert Munro, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen, and Harry Tily. 2010. Crowdsourcing and language studies: The new generation of linguistic data. In *Proceedings of the NAACL Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Vlad Niculae. 2013. Comparison pattern matching and creative simile recognition. In *Proceedings of the Joint Symposium on Semantic Processing*.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL-HLT*.

Allan Paivio, John C Yuille, and Stephen A Madigan. 1968. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology*, 76(1p2):1.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Carlos Roncero, John M Kennedy, and Ron Smyth. 2006. Similes on the internet have explanations. *Psychonomic Bulletin & Review*, 13(1):74–77.

Ekaterina Shutova and Lin Sun. 2013. Unsupervised metaphor identification using hierarchical graph factorization clustering. In *Proceedings of NAACL-HLT*.

Ekaterina Shutova. 2010. Models of metaphor in NLP. In *Proceedings of ACL*.

Niket Tandon, Gerard de Melo, and Gerhard Weikum. 2014. Smarter than you think: Acquiring comparative commonsense knowledge from the web. In *Proceedings of AAAI*.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014a. Metaphor detection with cross-lingual model transfer. In *Proceedings of ACL*.

Yulia Tsvetkov, Nathan Schneider, Dirk Hovy, Archna Bhatia, Manaal Faruqui, and Chris Dyer. 2014b. Augmenting English adjective senses with supersenses. In *Proceedings of LREC*.

Peter D Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of EMNLP*.

Tony Veale and Yanfen Hao. 2008. A context-sensitive framework for lexical ontologies. *Knowledge Engineering Review*, 23(1):101–115.

Tony Veale and Yanfen Hao. 2011. Exploiting readymades in linguistic creativity: A system demonstration of the Jigsaw Bard. In *Proceedings of ACL (System Demonstrations)*.

Tony Veale and Guofu Li. 2013. Creating similarity: Lateral thinking for vertical similarity judgments. In *Proceedings of ACL*.

Tony Veale. 2012a. A computational exploration of creative similes. *Metaphor in Use: Context, Culture, and Communication*, 38:329.

Tony Veale. 2012b. A context-sensitive, multi-faceted model of lexico-conceptual affect. In *Proceedings of ACL*.