Coin it up: Generalization of creative constructions in the wild

Julia Watson
Department of Computer Science
University of Toronto
(jwatson@cs.toronto.edu)

Farhan Samir
Department of Computer Science
University of Toronto
(fsamir@cs.toronto.edu)

Suzanne Stevenson
Department of Computer Science
University of Toronto
(suzanne@cs.toronto.edu)

Barend Beekhuizen
Department of Language Studies
University of Toronto, Mississauga
(barend.beekhuizen@utoronto.ca)

Abstract

Language is inherently flexible: people continually generalize over observed data to produce creative linguistic expressions. This process is constrained by a wide range of factors, whose interaction is not fully understood. We present a novel study of the creative use of verb constructions "in the wild", in a very large social media corpus. Our first experiment confirms on this large-scale data the important interaction of category variability and item similarity within creative extensions in actual language use. Our second experiment confirms the novel hypothesis that low-frequency exemplars may play a role in generalization by signaling the area of semantic space where creative coinages occur.

Keywords: linguistic generalization; verb constructions; denominalization

Introduction

Language is flexible, enabling people to creatively express new ideas or nuances of meaning. But such creativity is not unconstrained; for example, it seems very natural to generalize from phrases like *wrap up* and *snuggle up* to use *burrito up* as a verb in (1), but it is less natural to use *bookshelf* as a verb in (2):

1. He slept **burritoed up** all night.
2. She **bookshelfed up** the novel.

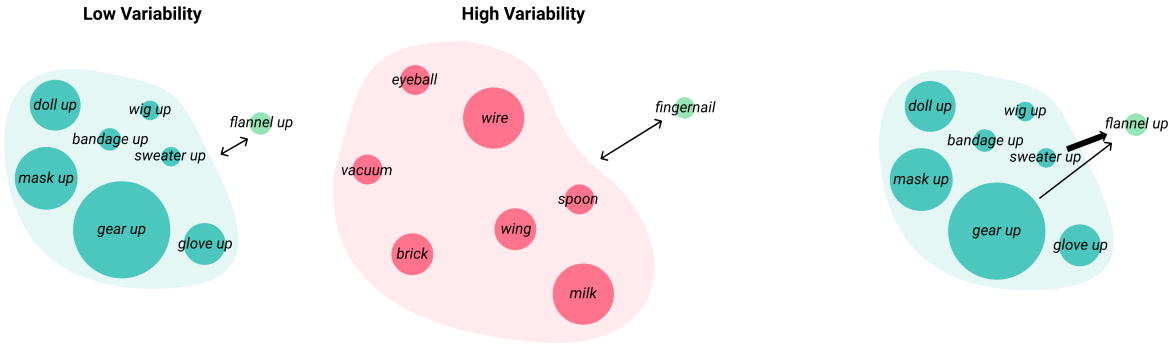
A key issue in the cognitive science of language is understanding which aspects of linguistic experience drive generalizability and which creative extensions are more likely.

We investigate these issues with a novel study on language use "in the wild". Generalization happens at all levels of language; here we focus on verb constructions – a rich domain that involves complex distributional factors and semantic constraints on the (potentially novel) verbs that can occur in them (e.g., Goldberg, 2006). Complementing previous work on construction generalization using artificial language experiments (e.g., Suttle & Goldberg, 2011), small case studies (e.g., Bybee & Eddington, 2006), or more formal corpora (e.g., Perek, 2016), here we analyze data from a very large-scale, informal, and interactive online discussion platform. Moreover, we focus on denominalization: the use of a noun as a verb, as in example (1). This linguistic process is a common source of creative usages (e.g., Clark & Clark, 1979; Yu et al., 2020), enabling us to identify hundreds of one-off examples of denominal constructions in our extracted data. To our knowledge, this is the first large-scale analysis of creative generalization of verb constructions in social media data.

A wide range of factors has been proposed to influence linguistic generalization, including both semantic and distributional factors (lexical statistics of various kinds) (e.g., Baayen & Lieber, 1991; Goldberg, 2006; Barðdal, 2008; Suttle & Goldberg, 2011; Perek, 2016; Barak & Goldberg, 2017; Pierrehumbert & Granell, 2018). Some of these factors have to do with how generalizable a construction is. For example, it has been proposed that constructions that are more variable (have a diverse set of exemplars) generate more novel coinages (e.g., Barðdal, 2008; Perek, 2016) Other factors in creative language use involve the "fit" of novel coinages with the construction they extend. For example, *burrito up*'s acceptability may arise from its shared semantic properties with existing exemplars such as *wrap up* and *snuggle up*. The key question we are concerned with is: how does the generalizability of a construction affect how we assess the fit of a new, creative coinage? In answering this question, we focus on the semantic properties of constructions: the variability of attested exemplars (generalizability), and the similarity of novel coinages to attested exemplars (fit).

Suttle & Goldberg (2011) showed in an artificial language learning experiment that the less variable a construction is, the more people are attuned to semantic similarity. We explore this here by comparing two denominal verb constructions. As Fig. 1a illustrates, the Suttle & Goldberg (2011) results suggest that people will be more stringent in extending a "low variability" construction, only permitting coinages that are very similar to attested exemplars, but will extend a "high variability" construction more freely, attending less to the similarity of a novel coinage. In our first experiment, we test this hypothesis on our constructions in the wild, finding support for the interaction of variability and similarity in actual creative language use.

In our second experiment, we look at which exemplars matter more in assessing fit, and whether this is affected by the variability of a construction. Work on morphology has shown that a high proportion of low-frequency examples (specifically, singletons) is a signal to language users that a construction is productive (Baayen & Lieber, 1991; Pierrehumbert & Granell, 2018). Building on this insight, we developed two novel hypotheses on how low-frequency exemplars impact the process of generalization – beyond simply signaling that generalization has occurred. First, we propose that for novel coinages, semantic similarity with low-frequency



(a) Do constructions with different levels of variability have different "fit" functions?

(b) Are high or low frequency exemplars more important in "fit" functions?

Figure 1: Key Questions

exemplars is more important than similarity with higher-frequency exemplars – i.e., low-frequency exemplars indicate not only **that** a construction generalizes, but **where** in semantic space the generalization is happening (see Fig. 1b). Second, we expect that the affinity to low-frequency exemplars will be stronger for low variability constructions: for these, speakers may feel limited to extending only in the pockets of semantic space where others have done so. We find support for the more general of these hypotheses but not the second one. To our knowledge, this is the first time that the semantic properties of low-frequency exemplars have been suggested to have a significant impact on generalization – a finding that contrasts with the emphasis on high-frequency exemplars from earlier studies on verb constructions (e.g., Bybee & Eddington, 2006; Casenhiser & Goldberg, 2005).

Case Study: Denominal Verbs and VPCs

We focus our case study on denominalization, a highly productive creative process in English and other languages (e.g., Clark & Clark, 1979; Yu et al., 2020). We first present two denominal verb constructions that we hypothesized would have differing levels of generalizability (measured as variability), enabling us to compare them on the properties of interest. We then describe how we extracted a sample of instances of these constructions from an online discussion platform.

Denominal Verb Constructions

The denominalizations we consider are uses of a noun as the verb in a verb-argument construction, without explicit derivational morphology marking the denominalization. Such denominals have been proposed to communicate an event of which a salient participant is expressed by the denominalized noun (Clark & Clark, 1979). These denominalizations can highlight various eventive participants, such as instruments (e.g., *fingernail the card* to mean "mark the card using a fingernail") or locations (e.g., *casket the remains* to mean "put the remains into the casket"). The first construction we consider, which we refer to simply as Denominal Verbs (DVs), is the broad range of uses of a "bare" denominal (in contrast to the second construction described below). The semantics of DVs can be quite varied; e.g., even within instrument-derived DVs, novel examples in our dataset range from ***fingernail*** *the edge of an ace for later* to *you can* ***whataboutism*** *the left all you want*.

Denominal verbs also commonly occur in conjunction with a particle, such as *up* or *out*, as in *I'm all* ***hydroed up*** *on good old h2o* and *should I go with one color or* ***rainbow*** *it* ***up****?*. When a noun combines with a particle in this way we refer to this as a Denominal Verb Particle Construction (DVPC). DVPCs are more semantically constrained than DVs, since each particle "selects for" a verb (noun) that has appropriate properties to combine with it. For example, with the particle *up*, DVPCs (like VPCs generally) often communicate a sense of increasing or becoming more positive (Lindner, 1982), such that examples like *I decided to* ***friend*** *him* ***up*** are more semantically acceptable than *I decided to* ***enemy*** *him* ***up***. Due to such semantic considerations, we expect that DVPCs are less variable than DVs. We consider DVPCs with a particular particle (*up*) in order to have a specific lexically-anchored construction, in contrast to the general construction of DVs. DVPCs with *up* are highly productive in our corpus.

Creating the EXEMPLAR and COINAGE Datasets

We collected examples of DVs and DVPCs from Reddit, a social media site that contains an abundance of creative language use. Specifically, we collected comment data from 11 subreddits: `AmItheAsshole`, `AskReddit`, `explainlikeimfive`, `IAmA`, `legaladvice`, `mac`, `malefashionadvice`, `movies`, `tifu`, `relationships`, and `unpopularopinion`. We selected these subreddits because we anticipated they would have high amounts of personal narratives, which we expected to be a good source for creative language use. We determined which words were acting as verbs or VPCs using a rule-based method that took syntactic parses as input. A verb usage counted as denominal if its lemma occurred at least twice as often as a noun than as a verb. Using this approach, we extracted a total

of $1,100,144$ DV tokens and $76,195$ DVPC tokens.[1]

To model how a construction is generalized to create novel coinages, we need two sets of types for each construction: a set of examples that speakers are exposed to (which we refer to as the EXEMPLAR set), and a set of novel coinages (which we refer to as the COINAGE set). One approach would be to treat all the one-off types (hapax legomena) as COINAGES, and take types that occur two or more times as known EXEMPLARS. A problem with this approach is that there are also one-off usages among the examples speakers are exposed to, so eliminating them from the EXEMPLAR sets does not yield data that realistically represents the full set of usages that speakers have likely seen. To allow inclusion of one-off examples in the EXEMPLAR sets, we partition the data into two parts, using one part to extract the EXEMPLAR types, potentially including one-offs, and the other as a pool for extracting hapax legomena, which we treat as COINAGES.

Because we can't know for certain which examples are novel coinages (i.e., usages where the author extends a construction in a way they haven't heard previously) and which are simply low frequency usages, we make the necessary simplification of treating most one-off examples as novel COINAGES. Pierrehumbert & Granell (2018) took a similar approach, treating sufficiently low frequency items – below 0.01 per 1 million words – as novel. We estimate that items in our COINAGE sets are of comparably low frequency. The COINAGE sets we create are full of examples like *to buzzword it up* ('to use a lot of buzzwords'), *to warrior up* ('to act like a warrior'), *to bystander a situation* ('to act as a bystander'), and *to yoda something* ('to make something sound like Yoda said it'), which seem novel to us.

To be able to compare DVs and DVPCs, we create EXEMPLAR sets of similar sizes by selecting an appropriate number of tokens from each construction that yields comparable numbers of types after filtering (manually removing examples that are false positives, such as parsing errors). That is, we use a random sample of $1,500$ of the $1,100,144$ DV tokens and $7,619$ of the $76,195$ DVPC tokens as our EXEMPLAR sets, which amount to 354 DV types and 301 DVPC types after manual filtering. We then extract samples of one-off types from the remaining tokens to form the COINAGE sets, similarly aiming for comparable numbers of types after filtering. Specifically, we obtain 175 DV COINAGE types and 205 DVPC COINAGE types after filtering, from the remaining $1,098,644$ DV tokens and $68,576$ DVPC tokens.

As semantic representations for our types, we used word2vec pretrained on Google News articles (Mikolov et al., 2013). Specifically, we used normalized embeddings of the verb lemmas. There are some limitations associated with using these embeddings. First, they are trained on news data, rather than social media data. We used the original word2vec embeddings trained on GoogleNews because they have been evaluated for their efficacy across a multitude of tasks and experiments in both natural language processing and psychology (e.g., Schnabel et al., 2015; Hollis & Westbury, 2016). Future work should validate the analyses presented here using embeddings trained on social media data. Pretrained static embeddings also do not capture regularities in how denominalization affects meaning. In future work, we hope to use vector representations that take denominalization into account (e.g., Yu et al., 2020). Lastly, vector space models are not entirely representative of judgments of similarity (Nematzadeh et al., 2017; Tversky, 1977), so it is important to replicate our findings with other operationalizations of semantic similarity.

Table 1 shows the number of types, number of types having a word2vec representation, and an example from each dataset.

Differing levels of variability

We selected our two constructions, DVs and DVPCs, on the intuition that the former is more inherently generalizable than the latter, so that we could investigate how factors of fit might interact with generalizability. Variability is a measure of generalizability that refers to the spread of exemplars of a construction in semantic space. We follow Suttle & Goldberg (2011) in assuming that the sub-clustering of exemplars of a construction is crucial to capturing this property, but we differ in using word embeddings and clustering methods to compute variability in semantic space.[2]

Because variability is a measure over existing instances of a construction, we perform this analysis using the EXEMPLAR datasets for DVs and DVPCs. The goal is to assess how tightly clustered the EXEMPLARS are for each construction. Since the correct number of clusters for each dataset is unknown, we use a k-means clustering algorithm, trying all values of $k = [1..20]$. To compute the variability of a particular clustering, we do the following: We take the prototype (centroid) vector of each of the k clusters and calculate the average (cosine) similarity between each exemplar and the prototype of its cluster. This average similarity indicates how closely exemplars are clustered together, and thus expresses the inverse of semantic variability for that clustering. We used bootstrapping (1000 samples with replacement for each construction) to enable us to construct confidence intervals over these values. The results for each value of k, for each construction, are shown in Fig. 2.[3]

We find that DVs have higher variability (lower intra-cluster similarity) than DVPCs for all values of k considered, supporting our hypothesis that DVs are an inherently more generalizable construction than DVPCs.[4] To give an intuition

[1]The $76,195$ DVPC tokens are a subset of around 1.1 million VPC tokens that we scraped. Because these started as separate projects, we applied the denominal filter during extraction for DVs and after extraction for DVPCs.

[2]We also go beyond a qualitative assessment of semantic spread of constructions using word embeddings, as in Perek (2016).

[3]These results are not specific to the k-means clustering algorithm: We find the same pattern of results with hierarchical clustering.

[4]Note that the significant difference in variability between the constructions is not driven purely by one-off exemplar types. (By one-off exemplar types, we refer to items in the EXEMPLAR sets

Table 1: Number of types, number found in word2vec, and an example, for each dataset.

Construction	Dataset	# types	# in w2v	Example
DV	EXEMPLAR	354	354	But serious question, I'm not **trolling** or joking. I'm honestly asking.
DV	COINAGE	175	164	You can **whataboutism** the left all you want.
DVPC	EXEMPLAR	301	301	It is found in all sorts of consumer products that **foam up**.
DVPC	COINAGE	205	201	I'm gonna have to **karen** it **up** later today ... I gotta self quarantine.

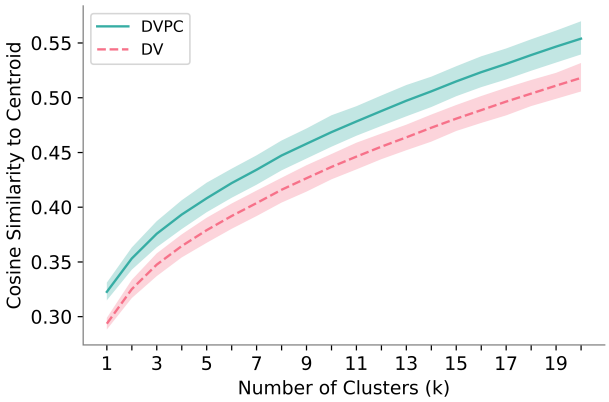


Figure 2: Variability analysis by construction. Error bars indicate 95% confidence intervals.

for this pattern, Fig. 1a shows a selection of items from a low variability DVPC cluster (containing *gear up*, *sweater up* and other examples related to clothing) and a high variability DV cluster (containing items that range from *vacuum* to *milk*).

Variability and similarity

In this section, we test whether the level of a construction's generalizability affects how people assess the fit of a coinage. Most existing work has looked at these factors – of level of generalizability and semantic fit – independently: more variable constructions are likely to be extended more readily (e.g., Perek, 2016), while acceptability of a coinage is modulated by its similarity to existing instances (e.g., Bybee & Eddington, 2006; Perek, 2016).[5] Suttle & Goldberg (2011) go further to provide insight into the interaction of these factors: in a comprehension experiment on an artificial language, they found that the similarity of a coinage to existing exemplars is most relevant to acceptability judgments when the construction has low variability.

Here, we similarly investigate the interaction of variability and similarity, but ask instead if this effect holds if we look at the production of novel coinages of actual constructions, as

that are sufficiently low frequency that they would have been considered COINAGES had they been assigned to the COINAGE partition). Excluding such low frequency examples from the EXEMPLAR sets gives the same pattern of results.

[5]Though see Barðdal (2008) for a discussion of the relationship between variability and similarity, among other factors in generalization.

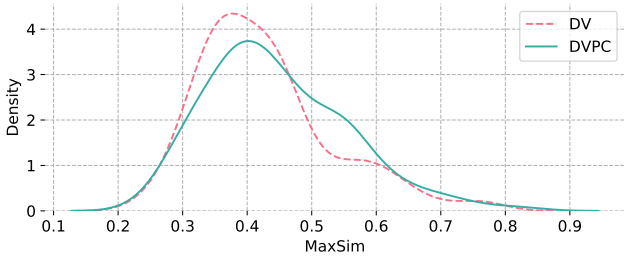


Figure 3: Kernel density estimates for MAXSIM for the DVPC and DV constructions.

found "in the wild". Our two constructions form an ideal test case for this question: with DVs found to be more variable than DVPCs, we hypothesize that DV COINAGES are on average less similar to DV EXEMPLARS than DVPC COINAGES are to DVPC EXEMPLARS. That is, we expect similarity to attested exemplars to be of less importance in the generalization of DVs than of DVPCs. We present two complementary ways of studying this generalizability and fit interaction.

Variability and Maximally Similar Exemplars

Our first analysis explores whether variability influences the level of (cosine) similarity found between a coinage and the exemplar that it is maximally similar to (denoted MAXSIM, cf. Suttle & Goldberg, 2011). MAXSIM provides insight into the degree of semantic deviation of a construction's coinages: the lower the similarity of a coinage to the most similar exemplar, the more that coinage deviates from what a speaker has already seen. When comparing the aggregated MAXSIM scores across our two constructions, we predict the more variable construction (the DVs) to have lower scores than the less variable construction (the DVPCs).

To give an intuition for this pattern, we show a selection of exemplars alongside a novel coinage from each construction in Fig. 1a. The novel DVPC coinage *flannel up* is very similar to its nearest DVPC exemplar *sweater up* as well as with other nearby exemplars, all related to clothing (*gear up*, *sweater up*, *mask up*, etc). In contrast, the novel DV coinage *fingernail* is not as strongly related to its nearest DV exemplar *spoon*, and even less so to other nearby items (which range from *vacuum* to *milk*).

Fig. 3 shows that the more variable construction (DVs; $N = 164$) indeed has significantly lower MAXSIM scores than the less variable construction (DVPCs; $N = 201$) (Mann-Whitney $U = 3.72$, $p < 0.005$ 1-tailed). This result

holds when we consider the average similarity not to just 1, but to the $k = [2..5]$ most similar exemplars as well. These results support the hypothesis that greater generalizability (as indicated by our measure of variability) allows for novel coinages to be less similar to the set of attested exemplars.

Variability and Coinage Fit

MAXSIM provides insight into the relation between variability and semantic similarity (as measures of generalizability and fit, respectively), but it abstracts away from the mechanisms of category extension assumed to underlie creative uses of a construction. Here, we approach the generalizability-fit interaction through models of semantic category extension (e.g. Ramiro et al., 2018), which assess the goodness of fit of an item to an existing category.

Specifically, we treat the novel COINAGES for each construction as extensions of a semantic category, and use a category extension model, trained on the EXEMPLARS, to estimate how likely the novel COINAGES are as extensions of the construction. We then assess the fit of each coinage as an extension of the construction (COINFIT for short) by ranking the coinage's likelihood score to the likelihood scores of a set of *hypothetical coinages*: a set of 7085 nouns from our corpus that have not been attested in either construction in our sample. The higher the rank of the coinage, the worse the fit of the coinage to the model (relative to the fit of hypothetical coinages to the model). We take higher COINFIT scores to reflect a greater permissiveness of the construction to allow more dissimilar novel coinages. Following our hypothesis that more variable constructions allow for more dissimilar coinages, we predict DVs to have higher COINFIT scores than DVPCs.

Different category extension models determine the likelihood of a novel coinage in different ways. To generalize over these different approaches, we employ the same set of models used in (Ramiro et al., 2018): a prototype model, an exemplar model, and k nearest-neighbor chaining models for $k = 1 \ldots 5$, which are inspired by theoretical frameworks of categorization (e.g. Rosch, 1975; Lakoff, 1987). Results for both constructions given all models are shown in Fig. 4. Here, we see that the DVs, the more variable construction, have significantly higher COINFIT scores than the DVPCs (164 DVs, 201 DVPCs; Mann-Whitney U tests for each model: all $p < .001$). This means that, in line with our hypothesis, novel COINAGES of the DVs on average have a lower fit with the attested EXEMPLARS than the DVPCs (as estimated through all the models).

As Fig. 4 shows, many hypothetical extensions are at least as likely as the COINAGES. This is to be expected, as the COINAGES are best thought of as only a small sample of all likely extensions of a construction. However, a relevant difference between the two constructions is that the average COINFIT of the DVs does not differ from a random ranking of the DV COINAGES, where we expect a coinage to rank on average halfway down the list (the dotted line in Fig. 4). Whereas DVPCs fall significantly below this line,

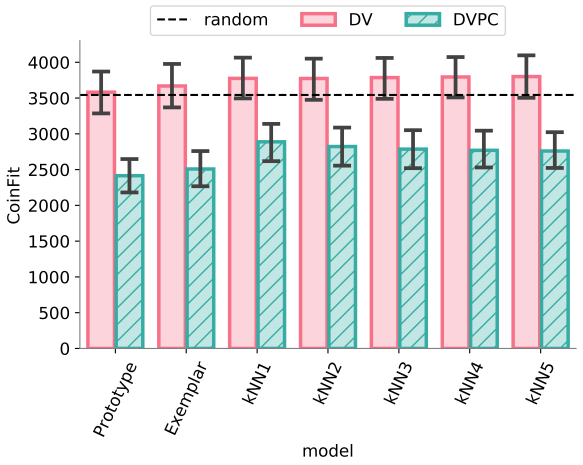


Figure 4: Average COINFIT values for both constructions. Lower values indicate that similarity is a better predictor of whether an item is a coinage. Error bars are bootstrapped 95% confidence intervals.

DV COINAGES are ranked at chance. This further supports our claim that the lower variability of DVPCs highly restricts the possibilities of category extension, compared to the fairly weak restrictions found for DVs.

Discussion

Taken together, our analyses provide novel evidence, on a dataset of actual language usage, for the interaction between variability and similarity found in artificial language experiments by Suttle & Goldberg (2011). Our first approach (MAXSIM), inspired by the findings of Suttle & Goldberg (2011), measures how similar novel coinages are to the exemplar most similar to them. In line with our hypothesis, we find that the COINAGES from the (more variable) DV construction are less similar to the EXEMPLARS closest to them, than the DVPC COINAGES are to their most-similar EXEMPLARS.

In our second approach, we study the interaction in a new way, through the lens of category extension models. Here we estimate how likely such models are to extend a construction to each example in our COINAGE datasets, compared to nouns not attested as denominals in our corpus. Using this COINFIT measure, we observe that DVPC COINAGES have significantly lower rank than DV COINAGES, whose ranks do not differ from random. These results again show that the less variable construction more strongly prioritizes coinages that are similar to its exemplars.

These findings provide further evidence for the variability-similarity interaction from two angles: both an isolated measure of similarity and a more integrated model of category extension show that the more variable construction is more permissive of novel coinages that have a comparably low semantic similarity. Importantly, our evidence for this interaction is found for a pair of actual constructions (vs. artificial ones; e.g., Suttle & Goldberg, 2011), is based on informal language data (vs. data from more formal registers e.g.,

Perek, 2016), and considers production data (vs. comprehension/acceptability data; e.g., Bybee & Eddington, 2006). Our results thus lend important convergent evidence to the previous studies that we build on.

Frequency and similarity

It has been argued that high-frequency exemplars "serve as the basis for the production of novel expressions" (Bybee & Eddington, 2006), and that such exemplars have a privileged role in the acquisition of grammatical constructions (Casenhiser & Goldberg, 2005). On the other hand, the ratio of hapaxes has been proposed to indicate a construction's productivity, because one-off examples signal to people that a construction can be extended (e.g., Baayen & Lieber, 1991; Pierrehumbert & Granell, 2018; Perek, 2018). Here, we propose that one-off examples also signal *how* a construction can be extended: namely, such examples indicate the areas in semantic space that are open for extension.

Based on this insight, we present the novel hypothesis that it is the low frequency exemplars that are attended to and that form the basis for the generalization of a construction to novel coinages. Formally, we expect the nearest neighbors of the COINAGES – the EXEMPLARS most similar to them, which we'll refer to collectively as the "nearest neighbor EXEMPLARS" – to sit on the low end of the frequency distributions over all EXEMPLARS. To be precise, we expect the frequency of the nearest neighbor EXEMPLARS to be lower than the frequency of the average EXEMPLAR. Fig. 1b illustrates this idea: the coinage *flannel up* is closer to the low frequency example *sweater up* than to higher frequency items like *gear up*. We further expect this effect to be more pronounced for DVPCs, the less generalizable construction: if a construction is not very generalizable, and similarity to existing exemplars is very important, speakers may feel more comfortable extending it in creative ways in the area of semantic space where they have clear evidence that other speakers have done so.

For each construction, we evaluated the influence of its low frequency members by computing the average (log-transformed) relative frequency[6] of the nearest neighbor EXEMPLARS, and comparing it, using a bootstrap test (Efron & Tibshirani, 1994), to the frequency of the larger population of EXEMPLARS. The bootstrap test involved drawing 10K samples (with replacement, from the full set of EXEMPLARS) of the same size as our set of nearest neighbor EXEMPLARS, and measuring if the average frequency of the nearest neighbors was lower than the lower end of a 95% confidence interval. (Having a directed hypothesis, we ran a one-tailed test.)

Fig. 5 shows that the mean frequency of the nearest neighbor EXEMPLARS is substantially lower than that of the average equal-sized sample drawn from the broader set of EXEMPLARS, and below the 95% CI. This supports our hypothesis that it is the lower-frequency exemplars that have a privileged role in the generalization of grammatical constructions.

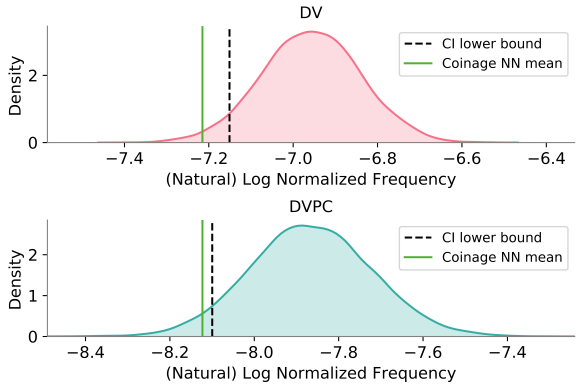


Figure 5: Distribution of bootstrapped mean frequencies of EXEMPLAR samples (density curves) and the mean frequency of those EXEMPLARS that are the nearest neighbor of an item in the COINAGE set (vertical solid green line).

Discussion

Our results are compatible with accounts of productivity that emphasize the role of one-off examples in generalization, but go beyond those accounts by suggesting that the semantic properties of low-frequency exemplars are in fact used in the process of generalizing a construction to a novel coinage. This result contrasts with previous work that emphasized the role of high frequency exemplars in generalization (Bybee & Eddington, 2006), and more research will be needed to understand the complex relationship between exemplar frequency and generalization.

Interestingly, we find similar results for both constructions. While variability may influence how similar a novel coinage needs to be to existing exemplars, variability does not seem to modulate which exemplars play a central role in generalization. Regardless of construction variability, speakers seem to prefer to generalize constructions by creating coinages that share semantic properties with low frequency examplars, perhaps because those are creative usages themselves.

Conclusions

Creative extensions of a construction are not unconstrained – some extensions are a better "fit" for a construction than others. We explored how this assessment of "fit" may be affected by a construction's inherent generalizability. For our exploration, we used data from an informal online discussion platform, where we find an abundance of creative language use. We see this as a key contribution of our work: to our knowledge, generalization of creative constructions has yet to be explored using such informal corpora. Using this data, we showed that a construction that is normally used to express a semantically limited set of meanings (less generalizable) will be extended in ways that are similarly limited. Conversely, a construction that has been used to express semantically diverse meanings will generate semantically diverse coinages. Regardless of a construction's inherent generaliz-

[6]We computed relative frequency for each type by dividing that type's frequency by the number of tokens in the EXEMPLAR sample for the associated construction.

ability, however, we found novel creative expressions emerge near other low-frequency expressions in semantic space, suggesting that creative (low-frequency) expressions inspire similarly creative expressions. This exploratory work is a first step towards understanding the processes underlying creative language use in the wild. Further research is needed to understand the complex interplay of factors that influence creative generalization.

References

Baayen, H., & Lieber, R. (1991). Productivity and English derivation: A corpus-based study. *Linguistics*, *29*(5), 801–844.

Barak, L., & Goldberg, A. E. (2017). Modeling the partial productivity of constructions. In *2017 AAAI spring symposium* (pp. 131–138).

Barðdal, J. (2008). *Productivity: Evidence from case and argument structure in Icelandic*. John Benjamins.

Bybee, J., & Eddington, D. (2006). A usage-based approach to Spanish verbs of 'becoming'. *Language*, 323–355.

Casenhiser, D., & Goldberg, A. E. (2005). Fast mapping between a phrasal form and meaning. *Developmental science*, *8*(6), 500–508.

Clark, E. V., & Clark, H. H. (1979). When nouns surface as verbs. *Language*, 767–811.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press.

Hollis, G., & Westbury, C. (2016). The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic bulletin & review*, *23*(6), 1744–1756.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. U. of Chicago Press.

Lindner, S. J. (1982). *A lexico-semantic analysis of English verb particle constructions with out and up.* Indiana University Linguistics Club.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Proceedings NeurIPS*, *26*, 3111–3119.

Nematzadeh, A., Meylan, S. C., & Griffiths, T. L. (2017). Evaluating vector-space models of word representation, or, the unreasonable effectiveness of counting words near other words. In *Cogsci.*

Perek, F. (2016). Using distributional semantics to study syntactic productivity in diachrony: A case study. *Linguistics*, *54*(1), 149–188.

Perek, F. (2018). Recent change in the productivity and schematicity of the way-construction: A distributional semantic analysis. *Corpus Linguistics and Linguistic Theory*, *14*(1), 65–97.

Pierrehumbert, J., & Granell, R. (2018). On hapax legomena and morphological productivity. In *Proceedings of the fifteenth workshop on computational research in phonetics, phonology, and morphology* (pp. 125–130).

Ramiro, C., Srinivasan, M., Malt, B. C., & Xu, Y. (2018). Algorithms in the historical emergence of word senses. *PNAS*, *115*(10), 2323–2328.

Rosch, E. (1975). Cognitive representations of semantic categories. *JEP: Gen*, *104*(3), 192.

Schnabel, T., Labutov, I., Mimno, D., & Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. In *Proceedings EMNLP* (pp. 298–307).

Suttle, L., & Goldberg, A. E. (2011). The partial productivity of constructions as induction. *Linguistics*, *49*(6), 1237–1269.

Tversky, A. (1977). Features of similarity. *Psychological review*, *84*(4), 327.

Yu, L., El Sanyoura, L., & Xu, Y. (2020). How nouns surface as verbs: Inference and generation in word class conversion. In *Proceedings CogSci.*