Modeling rules and similarity in colexification

Sammy Floyd (sfloyd@princeton.edu)
Kavindya Dalawella (kavindya@princeton.edu)
Adele E. Goldberg (adele@princeton.edu)
Casey Lew-Williams (caseylw@princeton.edu)
Department of Psychology
Princeton University, Princeton, NJ, USA, 08540
Thomas L. Griffiths (tomg@princeton.edu)
Departments of Psychology and Computer Science
Princeton University, Princeton, NJ, USA, 08540

Abstract

Colexification, or the expression of multiple concepts by the same word, is ubiquitous in language. Colexifications may appear rule-like, as when an artifact is used for an activity (*repair the shower/take a shower*), or similarity-based (*child* refers to both "young person" and "descendant"). We investigate whether these two modes of generalization (rules and similarity) reflect how people structure new meanings. We propose computational models based on rules, similarity, and a hybrid of the two, and correlate model predictions to human behavior—in a novel task, participants generalized labels across colexified meanings. We found that a model using similarity correlated much better with human behavior than rules, and that the similarity model was significantly outperformed by a hybrid model of the two mechanisms. However, the difference between similarity and hybrid was modest, suggesting that a framework which combines rules and similarity largely relies on similarity-based generalization to characterize human expectations about colexification.

Keywords: colexification; semantics; Bayesian modeling; polysemy; generalization; natural language processing

Introduction

English speakers can use the name of a species to refer to the animal's meat (she fed the *fish*/she ate *fish*). But speakers also face uncertainty: the meat of a *cow* is instead called *steak*. If offered the meat of a zebra at a restaurant, it is unclear whether an English speaker should apply a label for a similar concept, referring to it as *steak*, or follow the animal-for-meat rule, calling the meat *zebra*. Though this example is novel, a plurality of familiar concepts are **colexified**, sharing a word label with at least one other concept (Zipf, 1945). Like the multiple meanings of *fish* and *shower*, most colexified words name related or **polysemous** meanings (Bréal, 1897; Navigli & Ponzetto, 2012; Durkin & Manning, 1989; Schmitt & McCarthy, 1997; Rodd, Gaskell, & Marslen-Wilson, 2002), while a relatively low proportion of colexified meanings are unrelated or **homonymous** (Dautriche, 2015); for instance, flying bats and baseball bats share only a common label. What underlies the relationships between multiple meanings in the minds of language users?

Though polysemous meanings are related, they do not generally lend themselves to a single definition—instead, they require learners to somehow form a complex category or relationship from input in their language (Wittgenstein, 1953). A quick examination of cross-linguistic differences confirms that there are a multitude of cases of polysemy that are not universal (Fillmore & Langėndoen, 1971; François, 2008; Rzymski et al., 2020). For example, English's *straight*, French's *droit*, and Persian's *rost* can all be used to mean both "rectilinear" and "honest" (e.g., *straight talk*). This semantic extension from "rectilinear" to "honest" is motivated by a general metaphorical relationship between speaking directly and speaking honestly, but not all languages use the same word to mean both "straight" and "honest" (e.g. Spanish), so this is something that English, French, and Persian speakers must learn. English additionally uses *straight* to mean, "undiluted" and "heterosexual", while neither French nor Spanish do (François, 2008). This idiosyncrasy across languages has been used as evidence that individuals must in fact learn the extensions of their language as a set of conventions, and that new senses in a language may be learned on the basis of any semantic relationship between concepts (Murphy, 2004; Lehrer, 1990).

Nonetheless, past research has looked for general rules or patterns which might be used to predict colexifications. On these accounts, multiple meanings are learned and represented when a general rule or pattern is applied to an existing meaning (Fauconnier, 1994; Nunberg, 1979), and these patterns are predicted to reappear across languages (Copestake & Briscoe, 1995; Lakoff & Johnson, 2008; Ostler & Atkins, 1991; Pustejovsky, 1998). Cross-linguistic work tested 14 languages for instances of 27 English patterns of polysemy (such as "animal-for-meat") (Srinivasan & Rabagliati, 2015). Multilingual English speakers were presented with with a single example of a rule and asked to come up with further examples in their non-English language. In every language surveyed, there was at least some vocabulary which followed the English extension rules tested. However, there was also considerable variation in which particular vocabulary items the rule could apply to. For example, while English can be said to have a "material-for-artifact" relationship (e.g. the material glass, a drinking glass), other languages may apply this relationship to other materials instead. Another study found that English speakers' acceptance of artificial, English-based colexifications was better predicted by rules than similarity (Rabagliati, Marcus, & Pylkkänen, 2010), but did not test the two mechanisms in tandem. And, as both of these studies investigated how speakers familiar with English extensions rated English (or English-like) colexifications, they could have relied on non-rule structures learned from experience, it is not clear if these participants would have learned naturalistic colexifications via rules.

Other evidence suggests a role for semantic similarity in colexification (Wittgenstein, 1953). One historical study predicted meaning change across time in English, finding that similarity played a strong role in predicting which sense a word acquired (Ramiro, Srinivasan, Malt, & Xu, 2018). This work used a nearest-neighbor chaining algorithm over Word2Vec embeddings to predict the emergence of new, polysemous meanings for an existing form at later timepoints. However, this work did not contrast results with a rule-based account, and predicted historical change for English rather than directly testing human performance with colexifications. This is relevant because historical change may face pressures from many other sources such as environment, geography, colonialism and other language contact (Thompson, Roberts, & Lupyan, 2020; Gordon, 2004; Wiseman, 2015; Kirby et al., 2016), making it difficult to use historical change to draw conclusions about the role of similarity in the minds of language users. Similar work has focused on the importance of semantic relatedness in colexification across languages (Xu, Duong, Malt, Jiang, & Srinivasan, 2020; Youn et al., 2016). These approaches predict attested word extensions across many languages, finding that more similar concepts are more likely to share a label. We build on this work by investigating rule-based polysemy, which was not previously tested, and exploring novel word learning. Rather than predicting crosslinguistic variation, we ask a distinct question: When presented with novel word meanings, to what extent do learners rely on expectations from productive, rule-based extensions vs. semantic similarity?

Because language is rife with colexifications between concepts, the present work aims to determine the role of two candidate mechanisms in the human mind. We test whether rule-based generalizations predict how learners expect concepts to colexify, e.g. a rule would predict that learners would expect the concepts "foot" and "leg" are likely to colexify because they follow a part-for-whole pattern. We also investigate the contrasting proposal: that similarity predicts expectations about colexification (e.g. the concepts "float" and "swim" are likely to colexify because their semantic similarity is high). We suggest a third possibility: that humans simultaneously use *both* rules and similarity in structuring word meanings. Past work has used Bayesian modeling to compare rule and similarity-based learning in the domain of numerical concepts (Tenenbaum, 2000). The Bayesian modeling approach is well-suited to investigating these questions, as it offers a mechanism-agnostic approach to formalizing multiple kinds of hypotheses, such as numerical and semantic similarity, as well as rule or pattern-based structures. In this vein, we propose and evaluate three Bayesian models of colexification: one based in rules gathered from extant literature, one based in similarity as measured through word embeddings, and a hybrid rule-and-similarity model which generates predictions from a mixture of the two lone models. Based on past literature on concept learning, we predict that the hybrid model will correlate best with human expectations for colexifications in natural language.

To test this prediction, we asked adult, English-speaking participants to rate the probability of pairs of concepts being expressed by the same label in another language. The concept pairs were selected to be commonly attested colexifications in a database of over 3K languages, but novel to our English-speaking participants. We then correlated the human ratings to the three models' predictions. Critically, the participants' and models' ratings depend on the prior distribution of both rule-based and similarity-based colexifications in their experience.Therefore, the task was preceded by a warm-up phase, in which participants were exposed to a set of frequent colexifications from outside their language (randomly sampled without replacement from the full set used in the main task), so that the participants' prior expectations could be aligned with the models' priors.

Table 1: Examples of hypotheses from English as well as our task (non-English colexifications).

Hypothesis type	Example
Animal-for-meat (rule)	***feed a chicken, eat chicken*** [English]
Metaphor (rule)	***wear a crown, the Crown*** [English] ***voice, word*** [other languages, task item]
Cause and Effect (rule)	***move quickly, fast*** [other languages, task item] ***drop, fall*** [other languages, task item]
Part-for-whole (rule)	***walk down Wall Street, Wall Street panicked*** [English] ***hand, arm*** [other languages, task item]
Similarity	***old (aged), old (expired)*** [English] ***float, swim*** [other languages, task item] ***learn, study*** [other languages, task item] ***grandson, nephew*** [other languages, task item]

Experiment

Method

Participants Adults (*n* = 60) were recruited from Amazon's Mechanical Turk via CloudResearch (Litman, Robinson, & Abberbock, 2017) and compensated $1.20 to perform 100 ratings, preceded by the priors warm-up phase which exposed them to 70 other colexifications. 100% of participants who completed the task were approved for payment and no data were excluded.

Stimuli We obtained natural language colexifications from the Database of Cross-Linguistic Colexifications (CLICS), an inventory of 2,919 concepts 3,156 languages across 20 different language families (Rzymski et al., 2020). Because our intent was to test how English speakers structured novel colexifications, we extracted the 300 most frequently colexified meanings across languages, excluding any which also occurred in English using the CLICS colexification set for English. However, the English colexification set did not span the full list of concepts, so a coder reviewed and removed an additional 55 colexifications from the top 300 which were identified as attested in English, leaving 245 pairs of novel colexified concept pairs. From this set, we randomly selected 70 colexifications to use as exposure items in the warm-up phase which was run to match the participants' prior experience to the model. The remaining 175 were used for the main task. In order to estimate how participants rated both attested and non-attested colexifications, an additional 175 pairs were created by scrambling the 175 attested.

In order to determine which type of hypothesis the colexified pairs corresponded to, two coders were trained to recognize rule-based extensions based on past literature (Lakoff & Johnson, 2008; Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002) and coded each pair of colexified concepts used in the experiment. Whenever a pair did not correspond to one of the 32 rules, it was categorized as a similarity-based colexification.

Procedure The study consisted of two phases. First, in the priors exposure phase, participants were shown a randomly-selected set of 70 colexifications in order to align their experience in the task with the basis for the models' priors. Participants were shown two concepts from a colexification (such as "leg" and "foot") and told that a foreign language uses the same word to express both concepts. On the next page, they were prompted with one of the concepts from the colexification they were just exposed to ("foot") and asked to select the corresponding concept ("leg") from among two distractors and could not proceed to the next exposure until they correctly identified the target concept.

In the main phase of the task, participants were exposed to a randomly-selected 100 items, 1 per page, and prompted to rate the item's probability, e.g. "How likely do you think it is that a foreign language would use the same word to express the concept LEG(noun) and the concept FOOT(noun)?" Participants then selected a value from 0 to 100 on a slider, and this rating was compared with model predictions for each item.

Models

We compare three types of Bayesian models. First, we propose a rule-based model which uses productive patterns of colexification which have been identified in past research (see Experiment). Next, we tested a similarity model with two methods for measuring similarity: cosine similarity between the two colexified concepts' word embeddings from Sentence-BERT, a state-of-the-art transformer model of language (Reimers & Gurevych, 2019), and cosine similarity between the concepts' Word2Vec embeddings (Mikolov, Chen, Corrado, & Dean, 2013), which relies on a much simpler algorithm and less training. This was done in order to determine that the similarity models' performance was consistent, independent of the quality of the word embeddings. Finally, we present a hybrid model which represents a mixture of both rule and similarity models, again separately testing this model using both measures of similarity. We used each model to generate a prediction for the probability that two meanings would be colexified, and this was compared to human predictions for the same pair of meanings from the Experiment.

Rule-only model

Each colexification used in the experiment was coded for which hypothesis type (rule or similarity) it corresponded to, and in the case of a rule-type colexification, which particular rule hypothesis (see Table 1). Model predictions (the product of the prior and likelihood for each data point) were then compared with z-scored human ratings from the Experiment.

Formally, the hypothesis space $\mathcal{H} = \{h_1, h_2, \ldots, h_N\}$, consisted of N rules, each represented as the set of valid colexifications under the rule (determined by trained coders; see Experiment). Rules were constructed from from $N = 32$ extensions believed to be productive and general across languages in past work (see Table 1 for examples).

The prior probability $p(h_i)$ was defined as the proportion of colexifications that adhered to each rule h_i from a priors dataset. We obtained this using the same subset of 70 items which were randomly selected for the priors exposure task in the Experiment, which were coded for which hypothesis best characterized the relationship between their two concepts (calculating the relative frequency of each rule and normalizing their probability). In the hybrid model, a hierarchical prior was used for rule hypotheses by calculating the relative frequency of rule-type hypotheses to similarity hypotheses in this set. The likelihood function was binary, based on whether or not the example followed a rule and given by:

$$P(C|h_i) = \begin{cases} 1 & \text{if } C \text{ is consistent with } h_i \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where a colexification C is an unordered pair $C = C_1, C_2$ of meanings.

Given this prior and likelihood, the probability that a colexification would occur under any rule is given by:

$$p(C|R) = \sum_{i=1}^{n} p(C|h_i)p(h_i). \tag{2}$$

In the rule-only model, $p(C|R) = 1$ (as the hypothesis space consists of rules only), but it captures the degree that a colexification is supported by rules in the hybrid model.

Similarity-only model

The similarity-based model consists of a single hypothesis that expresses the probability that two concepts would be colexified based on their semantic similarity, and its prior was therefore set to 1. The likelihood of a colexification under the similarity rule is given by:

$$P(C|S) = \frac{1}{1 + \mathrm{e}^{-(\beta_0 + \beta_1 \mathrm{Sim}(C_1, C_2))}}. \quad (3)$$

This likelihood consists of a softmaxed linear model over similarity between concepts. The Sim function was computed via cosine similarity between the embeddings of C_1 and C_2, obtained using Sentence-BERT pretrained embeddings (Reimers & Gurevych, 2019) in one version of the model and Word2Vec pretrained embeddings in another (Mikolov et al., 2013). The final softmaxed linear model was obtained by fitting a logistic regression fit to the full set of colexifications used in the task via Ordinary Least Squares.

Hybrid rule-and-similarity model

Our main model combines predictions from both similarity and rule models described above, using a mixture given by:

$$p(C) = p(C|R)p(R) + P(C|S)p(S) \quad (4)$$

where $p(R)$ and $p(S)$ were scaled so that $p(R) + p(S) = 1$ using the distribution from the stimuli presented in the prior exposure phase (see Experiment), leading to $p(R) = 0.57$ and $p(S) = 0.43$.

Results

We correlated model predictions for each pair of concepts with z-scored human predictions for that item. Though all models showed some evidence of predicting human behavior (see Table 2), the hybrid rule-and-similarity model consistently and significantly outperformed the next-best performing similarity-only model in both versions ($z = 3.19$, $p = 0.001$ with S-BERT, z = 2.68, p = 0.01 with Word2Vec). Our prediction, that the hybrid rule-and-similarity model would best correlate with human judgments, was borne out in the data. However, panels C and E of figure 1 show quite a few items which are outside of the hybrid model's predictions (panels C and E). For example, the colexification between "dye" and "paint", which is attested in many languages, was given quite high ratings from the participants, but less high predictions from the similarity model. And, even though the two concepts follow a productive rule ("same function"), the frequency of this rule in the prior estimation data set was quite low, meaning that the combinations of low similarity and low rule-based hierarchical prior generated a much lower prediction from the model than the probability assigned by participants.

We also found that the similarity-based model correlated fairly well with human judgments (panels B and D). As in the hybrid model, there is still a cluster of attested colexification items which neither the model or humans think are likely, suggesting that even though similarity offers a gradient prediction, the similarity likelihood function assigns very low probability to less similar items.

In the top right-hand corner of panels B and D, predictions from the similarity-based model shows that there is a rather large cluster of colexifications with high similarity ratings. In comparison to the hybrid model (panels C and E), this suggests that the hybrid model may have benefitted from the hyperprior scaling down the similarity-only model's high predictions, introducing more variation and increasing the correlation with human predictions.

The rule model showed the weakest correlation with human performance (panel A). One possible reason for this is that, because each rule represents a categorical relationship, a model based in rules is ill-suited to capture the variation in participants' finer-grained predictions for each item. A second explanation we considered was that participants may not have been exposed to examples of each of the rules during the priors exposure warm-up, since they were only exposed to 70 colexifications. However, every rule type which appeared in the main task did appear in the prior estimation data, so participants would have had experience with colexifications from each rule. Nonetheless, it is possible that the frequency of these examples was not high enough in the prior phase to sufficiently reflect the frequency of that rule type in the main task.

Table 2: Pearson's correlations between human and model predictions

Model type	Correlation
Rule-only model	0.45
Similarity-only model (w/Sentence-BERT embeddings)	0.72
Similarity-only model (w/Word2Vec embeddings)	0.75
Hybrid rule-and-similarity model (w/Sentence-BERT embeddings)	0.75
Hybrid rule-and-similarity model (w/Word2Vec embeddings)	0.79

As described in the Models section, we ran both the hybrid and similarity-only models using Word2Vec embeddings as well as Sentence-BERT. This was done to test the possibility that the similarity model's results were not dependent on the state-of-the-art transformer architecture or extensive training used to create the S-BERT embeddings. However, we found the same pattern of results with Word2Vec as with S-BERT (see Table 2). In fact, the correlations between humans and the hybrid rule-and-similarity model as well as the similarity-only model were higher for Word2Vec, suggesting that the superior performance of models which use similarity are not entirely dependent on the sophistication of the similarity measure. It is, however, not intuitive that embeddings from a simpler model with less training such as Word2Vec

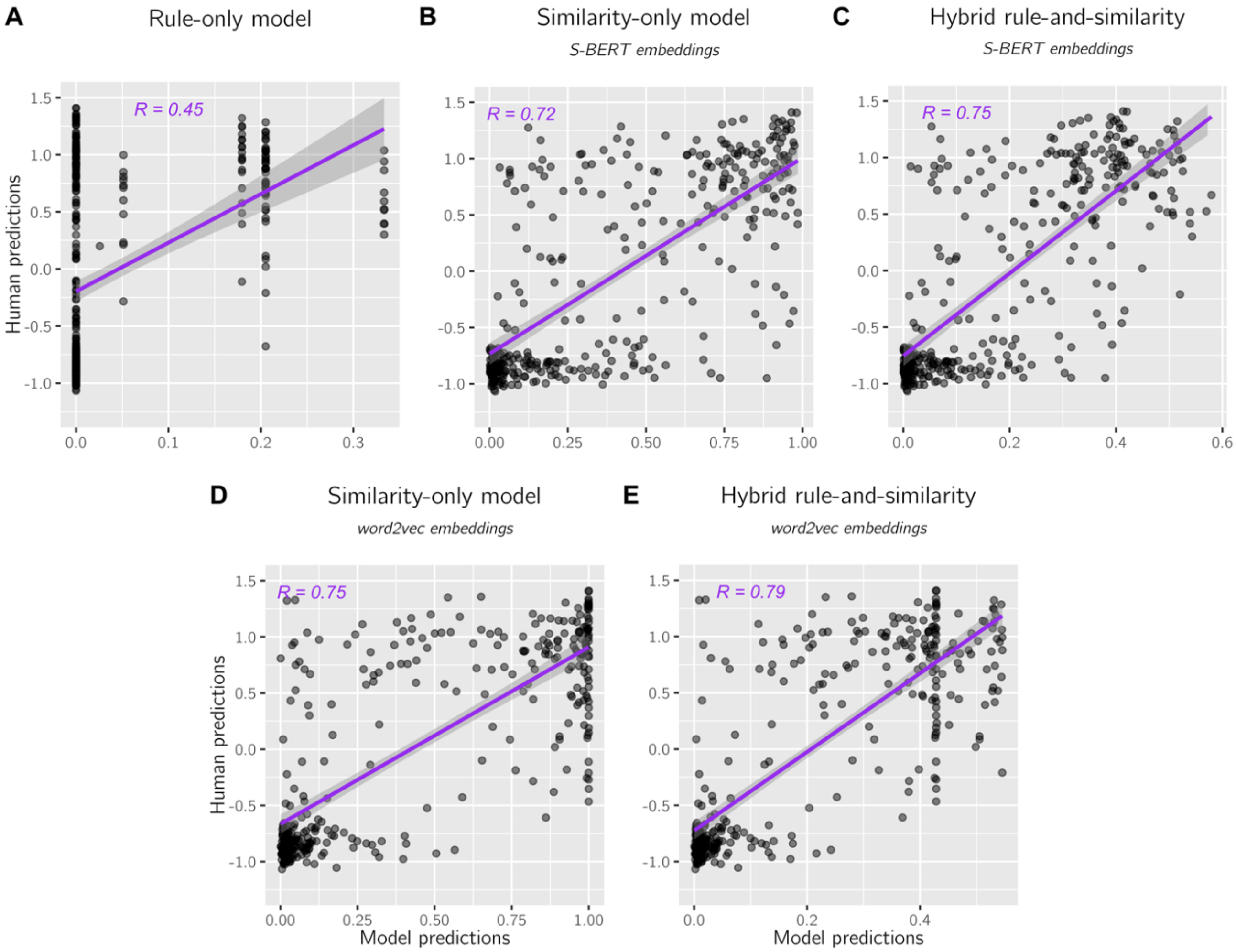

Figure 1: A: rule-only model weakly predicts variation in human responses. B: similarity-only model using Sentence-BERT embeddings performs noticeably better. C: a mixture of rules and similarity using Sentence-BERT embeddings correlates even better than similarity alone with the same embeddings. D & E show the same similarity-only and hybrid models, instead using Word2Vec embeddings, which had the same pattern of results (hybrid model $>$ similarity-only).

should deliver superior results. One possible explanation is that these pre-trained vectors were used to create embeddings for the short concepts described in the CLICS database, which often did not exceed a single word in length. And while Sentence-BERT was trained to build representations for entire sentences, Word2Vec was trained to build representations for words, which are overwhelmingly the semantic unit in our task, rather than phrases or sentences.

Discussion

The present study investigated mechanisms which underlie the human ability to share labels across meanings. Specifically, we formalized a label-generalization task as Bayesian inference over hypotheses based in semantic similarity, hypotheses based in rules, and a mixture of the two. We collected behavioral data from 60 English-speaking participants, asking them to rate the probability of novel colexifications based in attested language data. We found that the hybrid model, which combined predictions from both rule-based and similarity-based hypotheses, outperformed each lone model. Additionally, though the hybrid model's correlation with human predictions was significantly higher than the similarity-only model, the similarity-only model's correlation was almost as high. The rule-based model correlated worst with human judgments.

To represent each concept in the similarity model, we used word embeddings, which are models trained on large text corpora, and then calculated distance between the colexified concepts in this semantic space. The ability of these corpus-based semantic representations to capture human predictions might suggest that at least one of the possible structures underlying colexifications can be learned. Of course, the sheer amount of training required in these embeddings far exceeds what individual learners experience. Critically, however, we tested our models with two different kinds of word embeddings, and found that our similarity and hybrid models correlate just as well when trained on significantly less data with a much simpler algorithm.

A possible explanation for the relatively poor performance of the rule-only model is that our hypothesized rules could not adequately characterize the colexifications tested, especially because the patterns we included were deliberately selected for their novelty to English speakers. However, past research has found that these rules are attested, in some form, in each of the 14 languages they tested (Srinivasan & Rabagliati, 2015). Secondly, we expanded our rule set by adding 5 even more broad and inclusive rule types from additional literature, such as "metaphor", "cause-and-effect" and "same function" (Lakoff & Johnson, 2008). Moreover, participants were exposed to at least one instance of each rule type that appeared in the main task, during the priors exposure phase. Therefore, we do not think the best explanation for the rule model's low performance is that our candidate rules were insufficient.

Our results leave several interesting questions unaddressed. It is still possible that novel rule hypotheses could instead be learned from the data, rather than simply gathered from past literature: future work should examine possible systematicity in colexifications outside of English using human coders and/or machine learning techniques. Additionally, the present work does not include a systematic analysis of errors of the model, which would improve our understanding of why the rule approach was so unsuccessful. And, while we focus on modeling people's expectations about colexification, past work has investigated attested colexifications across languages and history (Xu et al., 2020; Ramiro et al., 2018). Future work can combine these approaches by exploring how human learning expectations do (or do not) contribute to the conventionalization of label extensions. Further, recent work has shown that learners as young as 2 years old are familiar with the language-specific polysemous colexifications from their environment (Floyd, Goldberg, & Lew-Williams, 2020), and children can rapidly learn and retain networks of meaning which do not follow a single pattern or rule (Floyd & Goldberg, 2020). This ability to flexibly form relationships across meanings is key, as individuals with a reduced propensity to generalize across items have been shown to be challenged by polysemous word learning (Floyd, Jeppsen, & Goldberg, 2020). Therefore, future work should investigate how young children learn to flexibly generalize using rules and similarity, and how they come to combine generalizations based on both kinds of mechanisms.

Together, we show the first quantitative modeling evidence that human learners use expectations about colexifications based both in semantic similarity *and* in rules (such as the semantic similarity between "cow" and "bison" and part-for-whole rule relationship between the concepts "leg" and "foot"). This is consistent with similar findings from concept learning, which show that human predictions can reflect a mixture of both kinds of extensions (Tenenbaum, 2000). We build on earlier work in semantics by testing more than one simultaneous account, going beyond cases already known to participants, and generating quantifiably-testable predictions which we compare with human performance. Our results can shed light on previous, seemingly contrasting findings, which have shown evidence for rules (Srinivasan & Rabagliati, 2015), evidence for rules over similarity (Rabagliati et al., 2010), and as evidence for similarity-based chaining over other category structures (Ramiro et al., 2018). By using a modeling framework that can simultaneously represent multiple mechanisms, we show that a combination predicts human behavior better than either mechanism alone.

References

Bréal, M. (1897). Une science nouvelle: la sémantique. *Revue des Deux Mondes (1829-1971)*, *141*(4), 807–836.

Copestake, A., & Briscoe, T. (1995). Semi-productive polysemy and sense extension. *Journal of semantics*, *12*(1), 15–67.

Dautriche, I. (2015). *Weaving an ambiguous lexicon* (Unpublished doctoral dissertation). Sorbonne Paris Cité.

Durkin, K., & Manning, J. (1989). Polysemy and the subjective lexicon: Semantic relatedness and the salience of intraword senses. *Journal of Psycholinguistic Research*, *18*(6), 577–612.

Fauconnier, G. (1994). *Mental spaces: Aspects of meaning construction in natural language*. Cambridge University Press.

Fillmore, C. J., & Langèndoen, D. T. (1971). Studies in linguistic semantics.

Floyd, S., & Goldberg, A. E. (2020). Children make use of relationships across meanings in word learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Floyd, S., Goldberg, A. E., & Lew-Williams, C. (2020). Toddlers recognize multiple meanings of polysemous words.

Floyd, S., Jeppsen, C., & Goldberg, A. E. (2020). Brief report: Children on the autism spectrum are challenged by complex word meanings. *Journal of Autism and Developmental Disorders*, 1–7.

François, A. (2008). Semantic maps and the typology of colexification. *From polysemy to semantic change: Towards a typology of lexical semantic associations*, *106*, 163.

Gordon, P. (2004). Numerical cognition without words: Evidence from amazonia. *Science*, *306*(5695), 496–499.

Kirby, K. R., Gray, R. D., Greenhill, S. J., Jordan, F. M., Gomes-Ng, S., Bibiko, H.-J., … others (2016). D-place: A global database of cultural, linguistic and environmental diversity. *PloS one*, *11*(7), e0158391.

Lakoff, G., & Johnson, M. (2008). *Metaphors we live by*. University of Chicago press.

Lehrer, A. (1990). Polysemy, conventionality, and the structure of the lexicon. *Cognitive linguistics*, *1*(2), 207–246.

Litman, L., Robinson, J., & Abberbock, T. (2017). Turkprime. com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior research methods*, *49*(2), 433–442.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Murphy, G. (2004). *The big book of concepts*. MIT press.

Navigli, R., & Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, *193*, 217–250.

Nunberg, G. (1979). The non-uniqueness of semantic solutions: Polysemy. *Linguistics and philosophy*, *3*(2), 143–184.

Ostler, N., & Atkins, B. T. S. (1991). Predictable meaning shift: Some linguistic properties of lexical implication rules. In *Workshop of siglex (special interest group within acl on the lexicon)* (pp. 87–100).

Pustejovsky, J. (1998). *The generative lexicon*. MIT press.

Rabagliati, H., Marcus, G. F., & Pylkkänen, L. (2010). Shifting senses in lexical semantic development. *Cognition*, *117*(1), 17–37.

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

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Rodd, J., Gaskell, G., & Marslen-Wilson, W. (2002). Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language*, *46*(2), 245–266.

Rzymski, C., Tresoldi, T., Greenhill, S. J., Wu, M.-S., Schweikhard, N. E., Koptjevskaja-Tamm, M., … others (2020). The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific data*, *7*(1), 1–12.

Schmitt, N., & McCarthy, M. (1997). *Vocabulary: Description, acquisition and pedagogy*. Cambridge university press.

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological science*, *13*(1), 13–19.

Srinivasan, M., & Rabagliati, H. (2015). How concepts and conventions structure the lexicon: Cross-linguistic evidence from polysemy. *Lingua*, *157*, 124–152.

Tenenbaum, J. B. (2000). Rules and similarity in concept learning. *Advances in neural information processing systems*, *12*, 59–65.

Thompson, B., Roberts, S. G., & Lupyan, G. (2020). Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour*, *4*(10), 1029–1038.

Wiseman, R. (2015). Interpreting ancient social organization: conceptual metaphors and image schemas. *Time and Mind*, *8*(2), 159–190.

Wittgenstein, L. (1953). *Philosophical investigations*. John Wiley & Sons.

Xu, Y., Duong, K., Malt, B. C., Jiang, S., & Srinivasan, M. (2020). Conceptual relations predict colexification across languages. *Cognition*, *201*, 104280. doi: https://doi.org/10.1016/j.cognition.2020.104280

Youn, H., Sutton, L., Smith, E., Moore, C., Wilkins, J. F., Maddieson, I., … Bhattacharya, T. (2016). On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences*, *113*(7), 1766–1771. Retrieved from `https://www.pnas.org/content/113/7/1766` doi: 10.1073/pnas.1520752113

Zipf, G. K. (1945). The meaning-frequency relationship of words. *The Journal of general psychology*, *33*(2), 251–256.