Probing the Representational Structure of Regular Polysemy in a Contextual Word Embedding Model via Sense Analogy Questions

Jiangtian Li (jiangtian.li@utoronto.ca)
Department of Psychology, University of Toronto Scarborough, 1265 Military Trail
Toronto, Ontario, M1C 1A4, Canada

Blair C. Armstrong (blair.armstrong@utoronto.ca)
Department of Psychology, University of Toronto Scarborough, 1265 Military Trail
Toronto, Ontario, M1C 1A4, Canada

Abstract

Regular polysemes are sets of ambiguous words that all share the same relationship between their meanings, such as CHICKEN and LOBSTER both referring to an animal or its meat. To probe how a context embedding model, here exemplified by BERT, represents regular polysemy, we analyzed whether its embeddings support answering sense analogy questions similar to "is the mapping between CHICKEN (as an animal) and CHICKEN (as a meat) the same as that which maps between LOBSTER (as an animal) to LOBSTER (as a meat)?" We found that (1) the model was sensitive to the shared structure within a regularity type; (2) the shared structure varies across regularity types, potentially reflective of a "regularity continuum;" (3) some high-order latent structure may be shared across regularity types, suggestive of a similar latent structure across types; and (4) there is equivocal evidence that the aforementioned effects are explained by meaning overlap.

Keywords: regular polysemy; semantic ambiguity; word analogy; contextual word embeddings; lexical semantics; BERT model

Introduction

Most words are semantically ambiguous and denote different meanings in different contexts (Rodd et al., 2002). As such, understanding how ambiguous words are represented and processed is an absolutely essential component of any theory of word or discourse comprehension (Rodd, 2020). Semantic ambiguity is not a monolithic phenomenon, however. One key way in which ambiguous words vary that has been the subject of extensive linguistic, computational, and psycholinguistic study is the relatedness between their meanings. Researchers typically make a broad delineation between homonyms, which have unrelated meanings (e.g., BAT refers to an animal or to baseball equipment), and polysemes, which have related meanings (e.g., POWER can refer to political authority or to electrical energy). Additional research has further differentiated among the polysemes. For example, polysemes can differ in terms of the relatedness among their meanings (Klepousniotou et al., 2008), which may be attributable, at least in part, to featural overlap. For instance, the meanings of CHICKEN, which refers to an animal or its meat, may be more related because they both can denote the same basic body parts (e.g., wing, thigh, leg, etc.) whereas other polysemes have fewer overlapping features, such as STAR, which refers to a celestial body or an actor.

On another distinct but potentially related front, sets of polysemes may be related to one another because they embody the same latent relationship among their meanings. For example, CHICKEN, LOBSTER, and SALMON all denote both an animal and its meat. This is referred to as *regular* polysemy, which can be contrasted against *irregular* polysemy, as exemplified by STAR, which has a more idiosyncratic relationship between its meanings. These same regular relationships exist across different languages as well (Srinivasan & Rabagliati, 2015). As such, the cross-word and cross-language structures among regular polysemes make them an ideal tool for drawing inferences regarding how meanings are organized (Lakoff & Johnson, 1980), how new meanings are extended during learning (Srinivasan & Snedeker, 2011), and how the conceptual system categorizes and generalizes similar relationships between different concepts (Lakoff, 1987).

We examine the underlying shared structure of regular polysemes as reflected in contextual word embeddings, here exemplified by the BERT model (Devlin et al., 2018). We chose this model for our initial exploration reported here because it is based on the principle of distributional semantics (Landauer & Dumais, 1997) and captures several aspects of how humans represent ambiguous words (Trott & Bergen, 2023). Our investigations are inspired by the "reason by analogy" logic developed to study word analogies using distributional semantic vectors, such as inferring that *Queen* is the appropriate completion for "_____ is to King as Woman is to Man." After first providing some additional background on regular polysemy and how we implemented "reason by analogy" logic to study sense analogies, we report the findings of our analyses based on the sense embeddings for words from five different types of regular polysemy derived from annotated texts. Our first goal was to confirm our intuitions that methods previously applied to study word analogies could be extended to study the structure present in regular polysemy. We then turned our atten-

tion to three other goals focused on aspects of regularity that have been discussed in prior work but have not, to our knowledge, been studied in explicit, computational terms. Our goals can be summarized in four questions (whose answers are foreshadowed in parentheses):

1. Does the representation in a contextual word embedding model reflect the shared structure of a specific type of regular polysemy? (Yes.)
2. Is the degree of regularity for each type different from each other, reflective of a graded "regularity continuum?" (Yes.)
3. Is there any higher-order latent structure shared across different types of regular polysemy (e.g., Animal/Meat and Location/Organization), suggestive of similar underlying pressure in the emergence of each type? (Yes.)
4. Can the degree of regularity be explained by the degree to which the semantic representations denoting each of the regular meanings overlap? (Equivocal.)

Prior Work

Regular structures that are shared across sets of polysemes were first described by Apresjan (1974), who also outlined several types of regular polysemy (e.g., COOK can refer to an action or the agent of the action). This initial theoretical distinction was further fleshed out in psycholinguistic experiments that identified processing differences (e.g., Fishbein & Harris, 2014; Frazier & Rayner, 1990) and learnability differences (e.g., Srinivasan & Snedeker, 2011; Zhu, 2021) between regular and irregular polysemes. However, the polysemes used in these studies were manually identified by the researchers, without any independent evidence for what aspects of the regularity are represented by humans or can be extracted from patterns of word co-occurrence in natural text. This contrasts with the formal quantification of other aspects of semantic ambiguity such as how much a word's meaning varies across contexts (e.g., Hoffman et al., 2013), the frequency with which each meaning is used (e.g., Rice et al., 2019), and the relatedness amongst a word's meanings (e.g., DeLong et al., 2022).

Additionally, several computational investigations have focused on regular polysemy using distributional semantic vectors derived from word co-occurrence in natural text. For example, Boleda et al. (2012) used CoreLex defined sense labels to examine regular polysemy, or the lack thereof, in words that were members of one class (e.g., SALMON, SHEEP as animals) and potentially members of a related class (e.g., SALMON but not SHEEP as meat). Lopukhina and Lopukhin (2016) used a sense-aware skip-gram model to induce sense vectors in a more unsupervised manner and infer whether a polyseme belonged to a given regularity type, but did not examine whether there is distinct or shared structure across types. Collectively, this work provides an important initial demonstration that there is shared structure across regular polysemes, but still leaves much unanswered. Our work is a major extension of this prior work, leveraging recent developments in deep learning based contextual word embeddings (Devlin et al., 2018) and more robust methods for examining word analogies (Drozd et al., 2016) to answer several additional theoretical questions, as outlined above, regarding how regular polysemy is represented.

Theoretical Approach

Our approach to probing the relationships between regular polysemes was inspired by related work on word analogies. This work (e.g., Drozd et al., 2016; Mikolov et al., 2013; Turney et al., 2003) has examined how distributional semantic vectors can be used to complete analogies of the form a is to a* as b is to b*, denoted as:

$$a : a^* :: b : b^* \tag{1}$$

For example, this work has examined how models can fill in a missing word in an analogy such as:

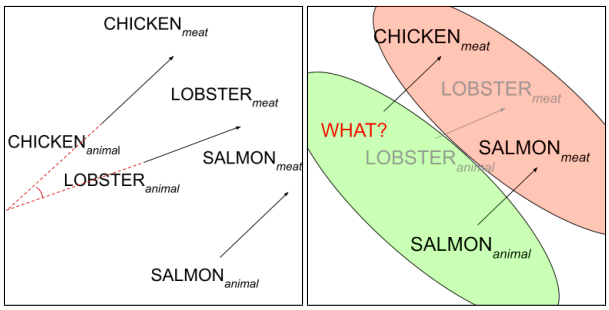
$$\text{-------} : QUEEN :: MAN : WOMAN \tag{2}$$

Prior work has succeeded in completing such analogies by identifying the relationship between each of the words in the representational space such as subtracting the semantic vector for WOMAN from that of QUEEN, and adding the vector for MAN. Our work extends this approach to *sense* (as opposed to *word*) analogies. We first derive separate representations for each word sense (e.g., CHICKEN as an animal, hereafter denoted as CHICKEN_{animal}). we then complete analogies in the form of:

$$\text{-------} : CHICKEN_{meat} :: SALMON_{animal} : SALMON_{meat} \tag{3}$$

In our analyses, we first complete sense analogies within one type of regular polysemy (e.g., Animal/Meat) and then compare these results with those from control conditions comprised of polysemes or homonyms that do not share the same regularity. We can thus assess whether there is additional structure shared by regular polysemes and answer our four key questions.

There are several requirements for implementing our approach, including the need for sense-annotated data; a computational model that generates representations of each sense from these data; and a method for computing the answers to sense analogy questions. The majority of these requirements can be addressed in a straightforward manner. However, the final point warrants additional consideration.



(a) Analysis of Parallelity (b) Analysis of Sense Classes and Similarity

Figure 1: Illustration of two analytical methods in a simplified 2-D semantic space. (a) Analysis of Parallelity involves measuring the angle between each pair of vectors that maps between the two senses of a polyseme. Low angles denote more parallel, and thus potentially more analogous, mappings. Only the angle between CHICKEN and LOBSTER is shown. (b) Analysis of Sense Classes and Similarity answers questions of the type _____ : $CHICKEN_{meat}$ :: $SALMON_{animal}$: $SALMON_{meat}$ in two steps. First, it involves assessing (1) the likelihood that a vector is a member of the animal class (green circle) as opposed to the meat class (the red circle) and (2) how close it is to $CHICKEN_{meat}$.

Answering Sense Analogy Questions

Assessing Parallelity. Arguably the simplest method for answering sense analogy questions is to analyze the geometric relationship between the senses associated with two polysemes and their respective interpretations (see Figure 1a). An analysis of *parallelity* assesses how similar the directions of the vectors mapping two senses for each word, with the notion that more similar (i.e., more parallel) mapping vectors are reflective of a more regular mapping between senses, as in Figure 1a. Despite the intuitiveness of this method, it has some major drawbacks (Drozd et al., 2016). Most critically, idiosyncratic variation among the individual polysemes makes it more difficult to observe the systematicity across all words within a type. This led us to employ a more sophisticated and robust method, described next. Nevertheless, we replicated all of our key findings using the parallelity method and found the same qualitative pattern of results, the only deviations being a few instances where comparisons were numerically but not statistically different. This is to be expected from a less sensitive method. We took this as evidence that our findings do not critically depend on a specific analytical method.

Assessing Sense Classes and Similarity. The second more analytically sophisticated and quantitatively sensitive approach to answering sense analogy questions was inspired by recent work by Drozd et al. (2016). This approach adapts the sense analogy question in Equation 3 into "which word sense belongs to the same sense class as $SALMON_{animal}$ and is similar to $CHICKEN_{meat}$?" We operationalized the question using the logistic regression plus cosine approach (LRCos) illustrated in Figure 1b. With this approach, we can generate scores for every sense of a type as an answer to the question. A higher score is given to senses that are both (a) more likely to belong to the same class as $SALMON_{animal}$ (logistic regression probability) and (b) that are closer to $chicken_{MEAT}$ (cosine similarity). The sense with the highest score is considered to be the answer to the sense analogy question, and can be scored as either correct (1, e.g., if $CHICKEN_{animal}$ was selected) or incorrect (0).

Critically for our purposes, if there is no consistent relationship among the senses that share the same sense class, it will not be possible to form an accurate classification model using logistic regression. This should impair the overall accuracy of the method in answering sense analogy questions. Thus, we expect better performance for regular polysemes within a regularity type than in control conditions consisting of homonyms or randomly selected polysemes. Furthermore, by training the logistic regression on the polysemes of one regularity type and applying the model to classify the polysemes of another type (e.g. train the model on Animal/Meat and classify senses in Container/Content), we can examine the existence of higher-order latent structure across different types of regular polysemy. We refer to this as a cross-type control.

Methods

Types of Regular Polysemy

We focused on five major types of regular polysemy adapted from Alonso et al. (2013):

- Animal/Meat: "The CHICKEN flew" versus "the delicious CHICKEN"
- Container/Content: "The red BOX" versus "I hated the whole BOX"
- Location/Organization: "ENGLAND is far" versus "ENGLAND instituted reforms"
- Artifact/Information: "The BOOK fell" versus "the suspenseful BOOK"
- Process/Result: "The BUILDING took months to finish" versus "the BUILDING is sturdy"

We chose these five types because of their history of wide use in the field (e.g., Dölling, 2020; Rabagliati et al., 2011; Srinivasan & Rabagliati, 2015). Furthermore, Alonso et al. (2013) provides an excellent starting source for sense-annotated data.

Target Polysemes and Annotated Data

Our initial set of regular polysemes was taken from from the English dataset reported by Alonso et al. (2013). We excluded a small number of polysemes which did not, according to Alonso and colleagues, have both regular senses annotated. We thus started with between 5 and 55 polysemes in each regularity type (see Table 1). For the type with less than 10 polysemes (Container/Content), we manually added additional polysemes so that it also contained 10 items.

Similarly, our initial source for annotated data was the sense-annotated sentences from Alonso et al. (2013). For the top 10 most frequent words in each regularity type of the Alonso dataset, we further supplemented their annotated data with our own annotations for lines of dialog taken from the Brysbaert and New (2009) subtitles database. Specifically, we extracted 100 lines of dialog for each polyseme, or for polysemes for which there were fewer than 100 lines, all available lines. The senses evoked in each of these lines were then annotated by three research assistants. We discarded all lines for which the raters could not consistently identify a single specific annotation. This was the case for 30% of ratings, primarily because individual lines often were not constrained to evoke a single meaning, such as "Go get the chicken." The combined polysemes and annotated text formed our raw data set. This data set was further cleaned to remove polysemes that did not occur in the BERT base vocabulary (e.g., yak, prawn, quail), or for which only one sense appeared in the annotated data. A summary of the data appears in Table 1.

We also included two additional annotated data sets as control conditions. The first set consisted of homonyms and their corresponding 100 annotated sentences in the same subtitles database reported by Rice et al. (2019). Homonyms are defined as having unrelated meanings, so we expect them to exhibit very poor accuracy using the LRCos method. Our second set of control items consisted of samples of polysemes and their corresponding annotated sentences from Evans and Yuan (2017), which used sentences from the SemCor (Mihalcea, 1998) and MASC corpora (Passonneau et al., 2012). This set was comprised of a mixture of both regular and irregular polysemes, although the regular polysemes were sampled at random from across the population of regularity types. The number of items in each control condition was matched to the average number of regular polysemes in each regularity type, as was the average number of annotated sentences/lines of dialog associated with each item.

Deriving Sense Vectors

For each regular polyseme we derived a sense vector that corresponded to each interpretation of the polyseme. This was done by providing each sense-annotated sentence corresponding with a given sense as input to the BERT base model (Devlin et al., 2018). We computed the average vector from the last four 768-dimensional layers of the model to produce the contextual representation of this word in this sentence (see Jawahar et al., 2019). The vector for this sense was then computed as the grand average of the vectors from all sentences annotated with this sense. The same method was used to derive representations for the control items.

	A2013		B2009		Raw total		Cleaned total	
	$n(w)$	$n(s)$	$n(w)$	$n(s)$	$n(w)$	$n(s)$	$n(w)$	$n(s)$
A/M	55	9	10	64	55	20	26	34
C/C	5	29	10	45	17	56	12	55
L/O	17	45	10	77	11	115	10	111
A/I	11	100	10	63	10	113	10	90
P/R	13	38	10	56	13	81	13	71

Table 1: Rows are Animal/Meat, Container/Content, Location/Organization, Artifact/Information, and Process/Result. A2013 = Data from Alonso et al. (2013). B2009 = Annotated data derived from Brysbaert and New (2009). *n(w)* is the number of words. *n(s)* is the average number of sentences for each word.

Analysis of Sense Classes and Similarity

We used the LRCos method described by Drozd et al. (2016) to compute sense analogies for each type of regular polysemy and for our various control conditions. Given a regularity type, we first formulated all the sense analogy questions that could be asked for all the polysemes of this type (e.g., "which word sense belongs to the same sense class as SALMON_{animal} and is similar to $\text{CHICKEN}_{\text{meat}}$?"). For each sense analogy question, (1) we quantified the probability that each sense is a member of the sense class of b (in this example, the animal class) with a logistic regression model, which was trained on all the sense vectors not in the sense analogy question (e.g., in the aforementioned case, the CHICKEN and SALMON senses would have been excluded from the training set); (2) we computed the similarity between a^* (e.g, $\text{CHICKEN}_{\text{meat}}$) and each sense vector with cosine similarity; (3) we multiplied the probability obtained from the logistic regression and cosine similarity to yield a score for each sense. The sense vector with the highest score was the answer to the analogy question given by the model, and was classified as either correct or incorrect. We averaged the classification accuracies within this regularity type. Our assumption was that greater underlying regularity would yield higher overall accuracy.

We also computed equivalent analyses for each of our control conditions (random polysemes, homonyms, and cross-type polysemes). For the random polyseme and homonym controls, to ensure that a given random sample of control items did not distort our results, we re-

peated this sampling process 5000 times and averaged the results. The cross-type control was used to probe how much latent regularity was shared across regularity types. To answer sense analogy questions for a given cross-type control, we trained the logistic regression model on each of the other regularity types and used that model as the basis for classifying items from the given type. For example, to answer sense analogy questions related to Animal/Meat, we used the models trained on each of four other types. The first class in the classifier was always the first class noted for each regularity type (e.g., Animal in Animal/Meat; Location in Location/Organization). We did this because the theoretical linguistics literature suggests, under various different labels, that there is one base sense whose meaning is extended/transferred to form the secondary sense (e.g., Copestake & Briscoe, 1995; Nunberg, 1995; Pustejovsky, 2005). If true, switching the base with the extended sense in cross-type controls might further impair the accuracy, because we would be aligning the base sense from one regularity type derived when training the classifier with the extended sense from the other regularity type at test. In a supplemental analysis, we reversed the order of base and extended sense when training versus testing the classifier and observed decreases in performance in all but one of the regularity types, as predicted by this intuition.

Results

Our key results are presented in Figure 2. We review how they bear on each of our four research questions, in turn. We controlled for multiple comparisons using Bonferonni-corrected p-values. Except as noted below, we used χ^2 goodness of fit tests to compare the conditions given the binomial nature of the accuracy data.

(1) To answer our first question regarding whether the representation in a contextual word embedding model reflects the shared structure of a specific type of regular polysemy, we compared the five types of regular polysemy against the random polyseme and homonym controls. All of these comparisons were statistically significant, indicating that YES, the model's representations reflect the shared structure within a regularity type.

(2) To answer our second question regarding whether polysemes associated with different types of regular polysemy share different amounts of structure, we compared the five regular polysemy types against one another. These comparisons were statistically significant, with the exception of Animal/Meat vs. Container/Content, and Process/Result vs. Artifact/Information. The large number of significant differences between the categories, as well as their distribution across a broad range of accuracy values, suggests that YES, regularity varies as a graded, continuous construct, and is not a monolithic construct wherein all types of regular polysemy are equal.

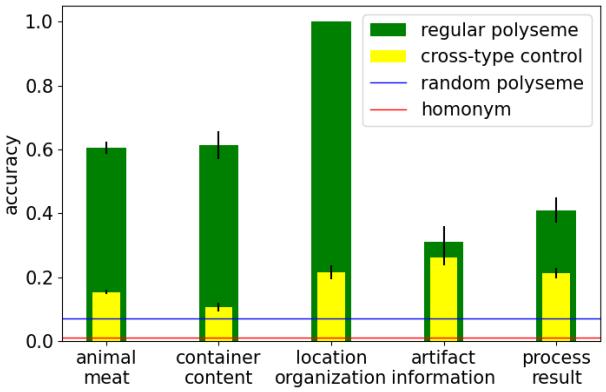


Figure 2: Mean accuracy and standard errors for regular each regular polysemy type and the control conditions.

(3) To answer our third question regarding whether there is any higher-order latent structure shared across different types of regular polysemy, we conducted two tests. First, we compared the cross-type regular polysemy controls against the random polyseme and homonym controls to see if there was shared regularity across different types. We found all five cross-type controls were associated with significantly higher accuracy than the other control conditions. This suggests that YES, there is some regularity shared across regularity types. Second, We compared each regularity type with its cross-type control to see if each regularity type had additional distinct structure. We found that accuracy was higher for four out of five regularity types than their cross-type controls (the exception being Artifact/Information). Collectively, this suggests that there is some shared latent structure across regularity types, but also unique structure associated with each type.

(4) Finally, we probed whether the variations in regularity could be explained by meaning overlap among a word's senses. To answer this question, we first computed the average meaning overlap between the two classes in each regularity type using cosine similarity. We then correlated these results with the accuracy data from the LRCos analysis for the five regularity types. We obtained a moderately strong correlation, but it did not reach statistical significance ($r(3) = -0.34$, $p = .58$). Analogous Bayesian analyses yielded intermediate Bayes factors within the 0.3-3 range, indicating that our data are EQUIVOCAL and do not provide a strong basis for making strong claims either way on this issue.

Additionally, to test for a potential confound due to the different number of words included in each regularity type, we correlated the number of polysemes in each type with the accuracy data. This correlation was very low and non-significant, ($r(3) = -0.02$, $p = 0.96$); analogous Bayes factor in the 0.5-2 range. Although we can-

not make strong claims either way based on these results, we do have some limited evidence that the number of polysemes per type did not drive our results.

Discussion

Most words are polysemous and have related but distinct senses. Many polysemes can further be classified as regular polysemes because multiple polysemes share the same overall regularity between their senses (e.g., Animal/Meat). We investigated if and how regular polysemy structure manifests in a contextual word embedding model via sense analogy questions. In particular, we answered four main research questions. First, is there significant shared structure across regular polysemes sharing the same regularity types? Our analyses indicated that this is clearly the case. The existence of this shared representational structure is important because it indicates that the structure of the mental lexicon could, in principle, serve as the basis for facilitating learning new regular polysemes, or learning a new regular meaning for an existing word (Rabagliati et al., 2011; Srinivasan & Rabagliati, 2015; Srinivasan, 2011).

Second, we investigated whether the degree of regularity varied across different types of regular polysemy. We observed substantial variability across regularity types, suggestive of a "regularity continuum." More generally, this finding suggests that classifying polysemes as regular or irregular is a false dichotomy. Rather, our results are more consistent with a graded transition from polysemes adhering closely to a particular underlying regularity to those with more idiosyncratic mappings between their senses. This parallels the older semantic ambiguity literature, which originally focused on a (false) dichotomy between ambiguous and unambiguous words, treating ambiguity as a monolithic category and collapsing homonyms and polysemes together (see Rodd et al., 2002, for discussion). Just as that dichotomy was further decomposed into polysemes and homonyms (e.g., Klepousniotou et al., 2008; Rodd et al., 2002), we argue for a a further decomposition across a regularity continuum.

Third, we asked whether there is any high-order latent structure shared across different types of regular polysemy. In what is arguably our most surprising finding, we found that this was the case. This suggests that there is a latent structure shared across different types when mapping from a base sense to an additional sense. Exactly what this latent structure might be will require additional investigation. Speculatively, this could be the Concrete/Abstract structure proposed by (Lakoff & Johnson, 1980). Essentially, this structure maps concrete concepts onto more abstract constructs, such as an artifact (e.g., a heavy BOOK) onto information (e.g., an interesting BOOK). This same structure has also been found to underlie at least some types of diachronic meaning extension for ambiguous words (Xu et al., 2017).

Of course, the empirical basis for our claim at present hinges on the five regularity types that we analyzed. These types were selected because they were the subject of extensive prior study. However, a potential concern from this selection is that prior research has focused on a non-random sample of regular polysemes (e.g., those that have the most consistent underlying representational structure, which might follow the abstract/concrete relationship). Probing this potential limitation further is, of course, an empirical question, and one that traditionally has been challenging because of the resource intensiveness of annotating data and the reliance upon intuitions regarding what regularity types exist and what words are associated with these types. This latter point may be particularly challenging to probe for what might be called "somewhat regular" regularity types that fall part way down the regularity continuum. However, in our view, an extension of our methods may offer a way forward on this front by allowing for the unsupervised detection of regular polysemy types that can be the subject of subsequent targeted analyses. This would involve first clustering annotated senses together into sense classes, and then examining for pairs of sense classes that have relatively consistent mappings between words that have a sense associated with each of these classes.

Finally, we tested whether the systematic differences between regularity types could be explained by meaning overlap. On this point, our analyses were equivocal. More data is clearly needed to answer this question in a compelling way. For theoretical reasons, however, we expect that although these factors will ultimately be found to be related, one will not ultimately be reduceable to the other. For instance, a polyseme could have two very closely related meanings, but that are related in a very idiosyncratic way, making the relationship between regularity and relatedness imperfect at best.

Conclusion

Taken together, our findings advance the understanding of the representational structure of regular polysemy by shedding light on how the relationships between senses are represented, how regularity is likely a continuous factor, and how similar latent pressures may drive the formation of several types of regular polysemy. Doing so in computationally explicit terms has also offered a platform for future extensions of this work, such as the unsupervised discovery of other types of regular polysemy. Given that most words in languages are polysemous, we expect these findings to help drive forward our understanding of an important facet of word and discourse comprehension, which may also be relevant to how we categorize and generalize knowledge more broadly.

Acknowledgments

This work was supported by NSERC Grant RGPIN-2017- 06310 to Blair Armstrong. The authors are grateful to Barend Beekhuizen, Julia Watson, and Yang Xu for their feedback on an a draft of this work.

References

Alonso, H. M., Pedersen, B. S., & Bel, N. (2013). Annotation of regular polysemy and underspecification. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 725–730.

Apresjan, J. D. (1974). Regular Polysemy. *Linguistics*, *12*(142), 5–32.

Boleda, G., Padó, S., & Utt, J. (2012). Regular polysemy: A distributional model. * *SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, 151–160.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990.

Copestake, A., & Briscoe, T. (1995). Semi-productive Polysemy and Sense Extension. *Journal of Semantics*, *12*(1), 15–67.

DeLong, K. A., Trott, S., & Kutas, M. (2022). Offline dominance and zeugmatic similarity normings of variably ambiguous words assessed against a neural language model (BERT). *Behavior Research Methods*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*.

Dölling, J. (2020). Systematic Polysemy. In *The Wiley Blackwell Companion to Semantics* (pp. 1–27). American Cancer Society.

Drozd, A., Gladkova, A., & Matsuoka, S. (2016). Word Embeddings, Analogies, and Machine Learning: Beyond king - man + woman = queen. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 3519–3530.

Evans, C., & Yuan, D. (2017). *A Large Corpus for Supervised Word-Sense Disambiguation*. Google AI Blog. Retrieved 2019, from http : / / ai . googleblog.com/2017/01/a-large-corpus-for-supervised-word.html

Fishbein, J., & Harris, J. (2014). Making sense of Kafka: Structural biases induce early sense commitment for metonyms. *J. Mem. Lang.*, *76*, 94–112.

Frazier, L., & Rayner, K. (1990). Taking on semantic commitments: Processing multiple meanings vs. multiple senses. *Journal of Memory and Language*, *29*(2), 181–200.

Hoffman, P., Lambon Ralph, M. A., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior research methods*, *45*, 718–730.

Jawahar, G., Sagot, B., Seddah, D., Unicomb, S., Iñiguez, G., Karsai, M., Léo, Y., Karsai, M., Sarraute, C., & Fleury, É. (2019). What does BERT learn about the structure of language? *57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy*.

Klepousniotou, E., Titone, D., & Romero, C. (2008). Making sense of word senses: The comprehension of polysemy depends on sense overlap. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(6), 1534.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. University of Chicago Press.

Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By*. University of Chicago Press.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211.

Lopukhina, A., & Lopukhin, K. (2016). Regular polysemy: From sense vectors to sense patterns. *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex - V)*, 19–23.

Mihalcea, R. (1998). Semcor semantically tagged corpus. *Unpublished manuscript*.

Mikolov, T., Yih, W.-t., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 746–751.

Nunberg, G. (1995). Transfers of Meaning. *Journal of Semantics*, *12*(2), 109–132.

Passonneau, R. J., Baker, C., Fellbaum, C., & Ide, N. (2012). The MASC word sense sentence corpus. *Proceedings of LREC*.

Pustejovsky, J. (2005). A survey of dot objects. *Author's weblog*.

Rabagliati, H., Marcus, G. F., & Pylkkänen, L. (2011). Rules, Radical Pragmatics and Restrictions on Regular Polysemy. *Journal of Semantics*, *28*(4), 485–512.

Rice, C. A., Beekhuizen, B., Dubrovsky, V., Stevenson, S., & Armstrong, B. C. (2019). A comparison of homonym meaning frequency estimates derived from movie and television subtitles, free association, and explicit ratings. *Behavior Research Methods*, *51*(3), 1399–1425.

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.

Rodd, J. (2020). Settling Into Semantic Space: An Ambiguity-Focused Account of Word-Meaning Access. *Perspectives on Psychological Science*, *15*(2), 411–427.

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

Srinivasan, M., & Snedeker, J. (2011). Judging a book by its cover and its contents: The representation of polysemous and homophonous meanings in four-year-old children. *Cognitive Psychology*, *62*(4), 245–272.

Srinivasan, M. (2011). Flexibility in language and thought. *Dissertation Abstracts International: Section B: The Sciences and Engineering*, *72*(4-B), 2469–2469.

Trott, S., & Bergen, B. (2023). Word meaning is both categorical and continuous. *Psychological Review*.

Turney, P. D., Littman, M. L., Bigham, J., & Shnayder, V. (2003). *Combining independent modules to solve multiple-choice synonym and analogy problems*.

Xu, Y., Malt, B., & Srinivasan, M. (2017). Evolution of word meanings through metaphorical mapping: Systematicity over the past millennium. *Cognitive Psychology*, *96*, 41–53.

Zhu, R. (2021). Preschoolers' acquisition of producer-product metonymy. *Cognitive Development*, *59*.