COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 48 (2024) e13416

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13416

Probing the Representational Structure of Regular Polysemy via Sense Analogy Questions: Insights from Contextual Word Vectors

Jiangtian Li,[a] Blair C. Armstrong[b]

[a] *Department of Psychology, University of Toronto*
[b] *Department of Psychology and Department of Computer Science, University of Toronto BCBL, Basque Center on Cognition, Brain, and Language*

Received 14 September 2023; received in revised form 2 January 2024; accepted 5 February 2024

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 distributional semantic model, here exemplified by bidirectional encoder representations from transformers (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) similar to that which maps between LOBSTER (as an animal) to LOBSTER (as a meat)?” We did so using the LRcos model, which combines a logistic regression classifier of different categories (e.g., animal vs. meat) with a measure of cosine similarity. We found that (a) the model was sensitive to the shared structure within a given regular relationship; (b) the shared structure varies across different regular relationships (e.g., animal/meat vs. location/organization), potentially reflective of a “regularity continuum;” (c) some high-order latent structure is shared across different regular relationships, suggestive of a similar latent structure across different types of relationships; and (d) there is a lack of evidence for the aforementioned effects being explained by meaning overlap. Lastly, we found that both components of the LRcos model made important contributions to accurate responding and that a variation of this method could yield an accuracy boost of 10% in answering sense analogy questions. These findings enrich previous theoretical work on regular polysemy with a computationally explicit

Correspondence should be sent to Jiangtian Li, Department of Psychology, University of Toronto Scarborough, 1265 Military Trail, SW427A, Toronto, Ontario, M1C 1A4, Canada. E-mail: jiangtianli91@gmail.com

theory and methods, and provide evidence for an important organizational principle for the mental lexicon and the broader conceptual knowledge system.

Keywords: Regular polysemy; Semantic ambiguity; Word analogy; Word sense analogy; Distributional semantic model; Lexical semantics; BERT model

1. Introduction

Most words are semantically ambiguous and denote different meanings in different contexts (Rodd, Gaskell, & Marslen-Wilson, 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, which has been the subject of extensive linguistic, computational, and psycholinguistic study, is in terms of the relatedness between their meanings, and how the relatedness of an ambiguous words meanings are represented in the mental lexicon. Initially, researchers made 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) (Azuma & Van Orden, 1997; Jastrzembski, 1981; Rodd et al., 2002). As the field has progressed, the field has further differentiated the degree and nature of the relationship between the meanings of polysemes, which are the most numerous type of ambiguous word.

Relatedness of meaning can manifest in different ways in the representational structure within our mental lexicon. This can potentially further distinguish between different types of polysemes. For example, relatedness can be represented in terms of featural overlap. Thus, different types of polysemes can differ in terms of the amount of overlap among the representations of their meanings (Klepousniotou, Titone, & Romero, 2008). For instance, the meanings of CHICKEN, which refers to an animal or its meat, may be highly 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, relatedness of meaning may be represented not in terms of featural overlap but through multiple words sharing the same relationship between each sense of a word. 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. How the representations of regular polysemes are structured to reflect this common relationship between meanings is the subject of our inquiry. A number of different regular relationships have been attested across different languages (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) and how the conceptual system categorizes and generalizes similar relationships between different concepts (Lakoff, 1987).

The relatively recent development of language models that can produce context-sensitive representations of word meaning offers a valuable way of developing computationally explicit theories of the representational structure underlying regular polysemy based on the patterns of word co-occurrences observed in natural language. These computational models will be unpacked in more detail later, but in a nutshell, they can produce high-dimensional vectors for individual word senses based on how those word senses are used in different contexts. That is, they do not produce "static" representations of a word that reflect how that word is used across all contexts, but rather produce "contextual" word vectors that reflect what a word means in a specific context. Each dimension of those vectors can be thought of as coding for a (latent) semantic feature associated with a word, and several studies have established that such computational vectors can also predict neural representations of meaning (e.g., Ettinger, Feldman, Resnik, & Phillips, 2016; Søgaard, 2016), with considerable recent research focusing on improving the strength of these predictions (e.g., Chersoni, Santus, Huang, & Lenci, 2021; Sassenhagen & Fiebach, 2020; Schrimpf et al., 2021). Thus, the relationships between the vectors from the computational models can be analyzed to understand how the meanings of regular polysemy relate to one another—that is, to understand the representational structure of regular polysemy—and the results of this work can connect to a range of computational, behavioral, and neuroimaging research.

In what follows, we first discuss the specific literature that motivates our computational work on regular polysemy. We then examine the shared representational structure of regular polysemes as reflected in a distributional semantic model (DSM), here exemplified by the bidirectional encoder representations from transformers (BERT) model (Devlin, Chang, Lee, & Toutanova, 2018). This approach is distinct from many previous studies in psycholinguistics (e.g., Frazier & Rayner, 1990; Frisson, 2015; Rabagliati, Pylkkänen, & Marcus, 2013; Srinivasan & Snedeker, 2011; Zhu, 2021) and from theoretical linguistics work (e.g., Copestake & Briscoe, 1995; Nunberg, 1995; Pustejovsky, 2005) on regular polysemy, which has relied on verbal theorizing instead of developing an explicit computational account of how the different senses of a regular polyseme are represented and how those representations relate to one another. Previous research (e.g., Mandera, Keuleers, & Brysbaert, 2017) has shown that the vectors produced by DSMs can capture key aspects of linguistic cognition. BERT, as a contextualized distributional semantic model, produces different vectors for the same word in different contexts (i.e., contextual word vectors). In this study, we utilized the contextual vectors from the BERT model to examine the regular meaning variations of regular polysemes.

More specifically, our methods were 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 using BERT, we report the findings of our analyses based on the vector representations of each sense for 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 attention 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 representational structure in a DSM reflect the shared structure of a given type of regular polysemy? (Yes.)
2. Are there variations in the degree of regularity across different regular relationships? If so, this could indicate that regularity is a graded, continuous construct (i.e., a "regularity continuum") and is not a dichotomous construct (i.e., polysemes are either regular or irregular). (Yes.)
3. Is there any higher order latent structure shared across the different types of regular polysemy, suggestive of similar underlying pressure in the emergence of each type? (Yes.)
4. Can the degree of regularity be fully explained by the degree to which the semantic representations denoting each of the regular meanings overlap? (No.)

2. Prior work within theoretical and experimental linguistics

The fields of theoretical linguistics, experimental psycholinguistics, and computational modeling have each provided important insights that inform our current understanding of regular polysemy. To begin, we review relevant prior work from theoretical linguistics and experimental psycholinguistics.

The notion of regular structures that are shared across sets of polysemes was first proposed in theoretical linguistics 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 proposal has been further refined by Pustejovsky (2005), Nunberg (1995), and Copestake and Briscoe (1995) into a well-defined area of research, including extensive descriptions of various types of regular polysemy and hypotheses regarding how the meanings from two categories are related to one another.

The experimental psycholinguistics literature has sought to evaluate whether words that are regular polysemes exhibit different processing and learning effects relative to other types of ambiguity. For example, several studies that used online reading, offline rating, or some combination of these methods (Fishbein & Harris, 2014; Frazier & Rayner, 1990; Frisson, 2015; Frisson & Frazier, 2005; Rabagliati et al., 2013) have reported processing differences between regular polysemes and irregular polysemes or homonyms. Similarly, developmental studies using artificial language learning tasks have reported differences in how children learn and extend the meanings of regular polysemy when this type of ambiguity is compared to other types of ambiguity such as homonymy (Srinivasan, Al-Mughairy, Foushee, & Barner, 2017; Srinivasan, Berner, & Rabagliati, 2019; Srinivasan & Snedeker, 2011, 2014; Zhu, 2021). Collectively, this work indicates that regular polysemy is not only an abstract construct from theoretical linguistics but also taps into an important aspect of how the human mental lexicon is learned, represented, and processed.

Notwithstanding the major value of the aforementioned contributions, this work is limited by the lack of an explicit computational formalization of regular polysemy. For example, when studying real words sampled from natural language, researchers manually selected both the types of regular polysemes to study and the specific examples of these types. This serves to validate basic intuitions about regular polysemy, but does not shed direct light on what aspects of the regularity are represented by humans, or can be learned from patterns of word co-occurrence in natural text. An analogous limitation exists in the context of designing artificial regular polysemes. 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, Lambon Ralph, & Rogers, 2013), the frequency with which each meaning is used (e.g., Rice, Beekhuizen, Dubrovsky, Stevenson, & Armstrong, 2019), and the relatedness among a word's meanings (e.g., DeLong, Trott, & Kutas, 2022). It also motivates our review of computational models related to regular polysemy, discussed next.

3. Prior work with distributional semantic models

Given that our current work, as well as much prior work from the field of computational modeling, draws heavily from DSMs, we first provide a general overview of this approach to modeling and then review several specific computational investigations related to regular polysemy.

DSMs derive meaning from the statistical relationships between words in large corpora, formalizing the intuition that a word's meaning can be defined in terms of the company that it keeps (Firth, 1957). A key advantage of these types of models is that they can be implemented in relatively computationally efficient terms, allowing them to be trained on large volumes of text and thus develop rich, nuanced representations of word meanings. This type of model had previously been dismissed as unsuitable for capturing the essence of human meaning representations, as advocated by referential theories of meaning (Lewis, 1970; Putnam, 1975), because they lack grounding in the real world (e.g., they lack vision, touch, and the ability to interact with the environment). However, a growing body of evidence has suggested that despite a lack of grounding, which would inevitably be part of a complete model of human meaning representations, DSMs can be used to solve a range of linguistic and cognitive tasks and thus appear to be tapping into key aspects of human meaning representation (Mandera et al., 2017). Most critically for our work, these models do appear able to capture one key aspect of meaning: the conceptual relationship between words as proposed by conceptual role semantics (Block, 1987; Harman, 1982; Piantadosi & Hill, 2022). Conceptual role semantics posits that the meaning of a word is encapsulated by its relationship to other words that are conceptually related to it. For example, the meaning of "cat" is defined by conceptual relationships such as that "being a cat" entails "being an animal," "being carnivorous," "valuing its territory," and so on. Previous research has demonstrated that DSMs, including both those that we refer to as "classic" approaches that modeled every word's meaning as a single static vector, such as latent semantic analysis (Landauer & Dumais, 1997) and word2vec (Mikolov, Yih, & Zweig, 2013), as well as more recent approaches that generate context-sensitive representations of a word's meaning, such as BERT (Devlin et al., 2018), effectively capture this kind of conceptual relationship between words. This is illustrated through the use of these models to complete word analogy tasks (Mikolov et al., 2013), as well as the use of these models to compare multiple words along a conceptual dimension, for example, comparing different animals in terms of whether they are large or dangerous (Grand, Blank, Pereira, & Fedorenko, 2022). With regard to regular polysemy, understanding the different meanings of CHICKEN requires knowledge that one sense is more conceptually related to food while the other sense is more related to animals. This aligns with the "conceptual role" aspect of linguistic meaning that DSMs have proven able to capture in past work in this general vein. On another related front, a number of computational models (e.g., Chersoni et al., 2021; Ettinger et al., 2016; Sassenhagen & Fiebach, 2020; Schrimpf et al., 2021; Søgaard, 2016), including BERT, have been used to predict significant variation in the neural activity elicited when processing language. This suggests that these models—while clearly not comprehensive accounts of semantic cognition—have broad relevance and capture core relationships in diverse behavioral and neuroimaging data that are of interest within the cognitive sciences. As such, explorations of such models can help inform our understanding of research questions such as the key questions that we previously outlined as motivating our work. Or, expressed pragmatically, although these models are not perfect (as no model of cognition is perfect), prior research suggests that they are sufficient approximations of a range of relevant language abilities to help inform our understanding of how regular polysemy is represented. These models may also help generate novel predictions that can guide future experimental research and theoretical refinements.

Further honing in on regular polysemy, complementing ongoing work in experimental psycholinguistics, prior modeling work has suggested that regular polysemy can be better understood in explicit computational terms with the help of DSMs. For example, based on Corelex definitions, Boleda, Padó, and Utt (2012) used distributional vectors of monosemous words (e.g, robin or pork) to derive sense classes representations for predetermined classes, such as ANIMAL or MEAT. They then used these abstract sense classes to identify regular polysemes (e.g., lamb) belonging to both of the sense classes. They found that their computational approach performed better than two baselines based on either random or frequency-ranked combinations of lemmas from each of the categories to form artificial regular polyseme controls. This work clearly establishes that there is more to regular polysemy than simply having two associated interpretations across classes or having highly frequent interpretations.

On another related front, and one that bears some conceptual similarity to our own approach, Lopukhina and Lopukhin (2016) used a modified version of the word2vec model, which originally only represented a single distributional vector for a given word regardless of the number of meanings associated with it, to represent different distributional vectors for different senses associated with a word. Here, the number of distinct senses that could be associated with a word was determined via a parameter that varied the grain size of the difference between senses. The specific senses learned for a given word were, however, learned in an unsupervised fashion. They then used several examples of a type of regular polysemy as "anchor" words to define a type of regular relationship by example and evaluated whether this relationship could be applied to identify other "target" polysemes that share this relationship. This approach performed well above chance (but considerably below perfect accuracy) in inferring whether a given polyseme belonged to a given regularity type, as outlined in prior research (Apresjan, 1974). However, only four to seven examples of each type of regular polysemy were used in their evaluations, so the robustness of these findings to broader sets of example words is an open empirical question.

More recently, Floyd, Dalawella, Goldberg, Lew-Williams, and Griffiths (2021) examined a related issue: why multiple senses are colexified as a single regular polyseme. They tested two distinct principles of colexification: were the senses related through a regular rule or were they closely related based on overall similarity. To test these distinctions using a computational model, they first devised a novel task for inferring the likelihood of colexifying (i.e., using the same word) two concepts in a foreign language (e.g., the concepts of LEG as a noun and FOOT as a noun). Critically, the specific concept pairs being rated were sampled from among a set that included the most frequently colexified meanings across languages, excluding examples that exist in English, the language used in the study. They then used three different computational models, one in which colexification was entirely rule-based, defined in terms of 32 different meaning extension rules that the authors expected to be productive and general across languages (e.g., an Animal-for-Meat rule, a Part-for-Whole rule, and a Cause-and-Effect rule), one in which colexification was entirely similarity based, as determined by the similarity of the embeddings for each concept in the word2vec model (Mikolov et al., 2013) and Sentence-BERT variation of the BERT model (Reimers & Gurevych, 2019), and one hybrid model in which both rules and similarity codetermined colexification. Although in isolation the similarity-based model outperformed the rule-based model, it was the hybrid model that produced the highest overall levels of correct colexifications. These results indicate that there is more to regular polysemy than simply the similarity between the two related meanings. In a sense, a portion of our work can be viewed as building upon this work to identify how such rules (or rule-like constructs) could be derived and studied using DSMs, and how these rule-like constructs might relate to meaning similarity (i.e., featural overlap).

Collectively, the aforementioned computational investigations of regular polysemy provide an important initial demonstration that there is shared structure across regular polysemes. However, it still leaves much unanswered in terms of exactly how regular polysemy manifests in DSMs, and by proxy, how humans represent regular polysemy, insofar as the link between DSMs and human language representations observed in studies of several other aspects of language continues to hold true in this context. For example, in the aforementioned work, there was no examination of the similarities and differences within and between individual regularity types (e.g, how similar is the transformation between the Animal and Meat senses of CHICKEN and SALMON; how similar is the transformation between Animal and Meat senses overall to that between Part and Whole senses overall). Thus, there are still clearly a number of major unanswered questions. For example, are all types of regularity the same, or is there graded variation of the regularity of the relationships exhibited across regularity types (i.e., is regularity a dichotomous construct such that polysemes are either regular or irregular, or is there a graded and continuous "regularity continuum" that maps between extreme cases of regularity and irregularity)? If similarity is not the sole determinant of regular polysemy, how are rules (or rule-like structures) represented, and is this representation present in DSMs (as exemplified here by BERT)? Our work is a major extension of this prior work, employing a novel extension of more robust methods for examining word analogies to answer several additional theoretical questions, as outlined above.

4. Theoretical approach

Our approach to probing the relationships between regular polysemes was inspired by related work on word analogies. This work (e.g., Drozd, Gladkova, & Matsuoka, 2016; Mikolov et al., 2013; Turney, Littman, Bigham, & Shnayder, 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, prior work has examined how models can fill in a missing word in an analogy such as:

$$\_\_\_\_\_\_ : \mathit{QUEEN} :: \mathit{MAN}{:}\mathit{WOMAN} \tag{2}$$

This is equivalent to asking a human to identify "which word is to QUEEN as MAN is to WOMAN." Much work has succeeded in answering such analogies by identifying the relationship between each of the words in the representational space, such as by subtracting the semantic vector for WOMAN from that of QUEEN, and adding the vector for MAN.

Our work extends the aforementioned 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:

$$\_\_\_\_\_\_ : \mathit{CHICKEN}_{Meat} :: \mathit{SALMON}_{Animal} : \mathit{SALMON}_{Meat} \tag{3}$$

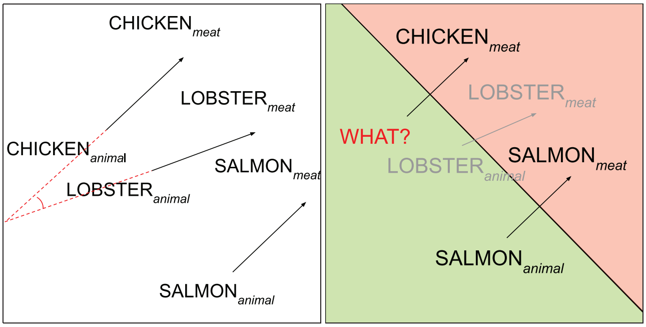
This is equivalent to answering the question "which word sense is to CHICKEN$_{Meat}$ as SALMON$_{Meat}$ is to SALMON$_{Animal}$?" In our analyses, we first answer sense analogy questions that involve exemplars from within a given type of regular polysemy (e.g., Animal/Meat) and then compare these results with those from control conditions, being 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 within a given regularity type 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 first requirement can be addressed in a relatively straightforward manner. However, the two other points warrant some additional discussion, which we present, in turn, below.

4.1. Selecting an appropriate distributional semantic model

In our work, we chose to employ the BERT model. Several considerations motivated our use of this model. BERT, as a contextual DSM, offers several advantages for the task at hand relative to other models (e.g., classic "static" models that produce context-invariant representations of word meaning). First, this model generates different semantic vectors for the same word in different contexts and has been extensively used to study semantic phenomena (Rogers, Kovaleva, & Rumshisky, 2021). Second, unlike the GPT-style of transformer architecture (Brown et al., 2020; Radford et al., 2019), the semantic vectors in BERT are created by leveraging information from both the context that precedes and follows a target word. The subsequent context may be important for activating the correct interpretations of ambiguous words in at least some circumstances, as examined in the psycholinguistic literature, because it can enable the appropriate meaning of an ambiguous word to be identified in cases where the preceding context does not specify the correct interpretaton (Frazier & Rayner, 1990; Vitello, Warren, Devlin, & Rodd, 2014). Although in many cases, words, and presumably their meanings, are predictable strictly from prior context, regressions to review previous text in natural reading suggest that both preceding and subsequent context are at least sometimes important for the correct interpretation of the ambiguous words. Additionally, some of the prior data by Alonso, Pedersen, and Bel (2013) that we built upon, as well as additional data we annotated ourselves, consists of sentences that were annotated in isolation from broader (preceding or following) context—to a first approximation, the annotated data typically consist of a single sentence or a small number of very short sentences. Thus, using a model that can leverage the full constraints available both before and after a target word should more closely align with the information available to human annotators. Next, BERT is one of the largest models that is open-sourced and that can conceivably be trained by researchers without access to extensive specialized computational equipment, allowing us to probe into the hidden layers, or retrain the entire model in future work. This is not currently an option with more recent close-sourced large language models (LLMs) such as GPT3 and GPT4 (Brown et al., 2020; OpenAI, 2023), and even if these models were open-sourced, most academic researchers would not have access to the computational resources required to retrain many LLMs. The BERT model is also the most cited model in the BERT family (which also includes RoBERTa [Liu et al., 2019], ALBERT [Lan et al., 2020], and DistilBERT [Sanh, Debut, Chaumond, & Wolf, 2020]). This means that using BERT will be most relevant for connecting with a wide group of researchers. Finally, a number of studies have established a significant relationship between the representations of BERT (and DSMs more broadly) and various neural measures of the representation of meaning (e.g., Chersoni et al., 2021; Ettinger et al., 2016; Sassenhagen & Fiebach, 2020; Schrimpf et al., 2021; Søgaard, 2016). Although this statistical relationship is not perfect (i.e., it is far from explaining 100% of the variance in neural data), the presence of such a relationship indicates that an analysis of regular polysemy in BERT has the potential to generate predictions that can guide neuroimaging research related to regular polysemy and connect with the neuroimaging research on semantic ambiguity more broadly (e.g., Klepousniotou, Pike, Steinhauer, & Gracco, 2012; MacGregor, Bouwsema, & Klepousniotou, 2015; Rodd, Davis, & Johnsrude, 2005; Vitello et al., 2014; Yurchenko, Lopukhina, & Dragoy, 2020).

Although we could envision that some detailed aspects of our findings are shaped by the particular implementational details of the BERT model, we nevertheless predict that our overall insights into regular polysemy should be reasonably robust and generalize to other similar models, as has been observed in other contexts using distributional vectors. This prediction is motivated by the qualitatively similar results that have been obtained using different models when studying other aspects of polysemy (e.g., BERT vs. ELMo in Trott & Bergen, 2023; word2vec vs. Sentence-BERT in Floyd et al., 2021), as well as the prior success in using a variety of models to successfully study other aspects of regular polysemy (e.g., Lopukhina &



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

Fig. 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 (a) the likelihood that a vector is a member of the animal class (green area) as opposed to the meat class (the red area) and (b) how close it is to CHICKEN$_{Meat}$.

Lopukhin, 2016). Of course, this prediction is fundamentally an empirical question which we leave for future work.

4.2. Answering sense analogy questions with a distributional semantics model

Two main approaches to operationalizing how a model could be used to simulate answering sense analogy questions were considered for our work, each of which is reviewed in turn.

4.2.0.1. 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 Fig. 1A). An analysis of *parallelity* assesses the similarity of the directions of the vectors mapping between two senses for a pair of words, with the notion that more similar (i.e., more parallel) mapping vectors are reflective of a more regular mapping between senses. For example, Fig. 1A illustrates how there is a low angle (high parallelity) between the vectors that map between the animal and meat interpretations of both CHICKEN and LOBSTER. Despite the intuitiveness of this method, recent work has identified several major drawbacks with analyses based on parallelity (Drozd et al., 2016). Most critically, idiosyncratic variation among the senses of individual polysemes may make it more difficult to observe the overall systematicity across all polysemes within a type. This led us to employ a more sophisticated and robust method, described next. Nevertheless, we replicated most of our key findings using the parallelity method and found that the overall results correlated strongly with our more sophisticated methods (see Supporting Information). We take this as evidence that our main findings do not critically depend on a specific analytical method.

4.2.0.2. Assessing sense classes and similarity: The second more analytically sophisticated and quantitatively sensitive approach to answering sense analogy questions was inspired by recent work from Drozd et al. (2016). This approach adapts the sense analogy question in 3: “which word sense is to CHICKEN$_{Meat}$ as is SALMON$_{Animal}$ to SALMON$_{Meat}$?” into a new closely related two-part question in (4a, 4b). This question essentially asks “which word sense belongs to the same sense class as SALMON$_{Animal}$ and is also similar to CHICKEN$_{Meat}$?” The first part of this question asks about the sense class of the target word sense (4a), and the second part asks about the sense similarity of the target word sense (4b).

______ belongs to the sense class of SALMON$_{Animal}$ (4a)

______ is ALSO similar to CHICKEN$_{Meat}$ (4b)

We operationalized the answer to the sense analogy question using the logistic regression (LR) times cosine approach (LRcos) illustrated in Fig. 1B, with LR providing the answer to the first part of the question and cosine similarity providing the answer for the second part. With this approach, we can generate a score for every sense as a potential answer to the sense analogy question by multiplying the scores from each of the two parts. A higher score is given to senses that are both (a) more likely to belong to the same class as SALMON$_{Animal}$ (LR probability) and (b) more similar to CHICKEN$_{\text{Meat}}$ (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$_{\text{Animal}}$ was selected) or incorrect (0).

Critically for our purposes, if there is no consistent relationship among the senses of the sense class, it will not be possible to form an accurate classification model using LR. This should impair the overall accuracy of the method in answering sense analogy questions. Furthermore, if two senses of a word are not similar, senses of other words may be more similar and be selected incorrectly as the answer to a sense analogy question.

Furthermore, by training the LR model on the polysemes from one regularity type and applying the model to classify the polysemes of another type (e.g., train the model on the Container/Content classes and then use this trained model to classify senses from the Animal/Meat classes), we can examine the existence of higher order latent regularity shared across different types of regular polysemy. We refer to this as a cross-type analysis. This analysis is equivalent to asking sense analogy questions such as “which word sense is to CHICKEN$_{Meat}$ as CANADA$_{Location}$ is to CANADA$_{Organization}$?”

5. Methods

To implement the theoretical approach described above, we needed datasets of regular polysemes and control words, methods to derive sense vectors from them, and a computational implementation of LRcos. We describe each of these components of our methods in turn. All code and data necessary to replicate our work are available at: https://osf.io/bctp4/?view_only=b981b9a66b5046459a0201e14f0a5d26

Table 1
Summary of the datasets used in our study

	A2013		B2009		Raw Total		Cleaned Total		Interrater Reliability	
	$n(w)$	$n(s)$	$n(w)$	$n(s)$	$n(w)$	$n(s)$	$n(w)$	$n(s)$	Percentage	κ
A/M	55	9	10	164	55	38	27	73	0.93	0.86
C/C	17	29	10	94	17	79	12	91	0.80	0.57
L/O	11	45	10	85	11	111	10	119	0.92	0.41
A/I	5	100	10	98	10	135	10	126	0.78	0.52
P/R	13	38	10	87	13	96	13	95	0.75	0.51

Note. A/M = Animal/Meat, C/C = Container/Content, L/O = Location/Organization, A/I = Artifact/Information, P/R = Process/Result. A2013 = Data from Alonso et al. (2013). B2009 = Annotated data derived from Brysbaert and New (2009). Cleaned total refers to the number of words and sentences used in the analyses after excluding words that were not part of the BERT base vocabulary and plurals. *n(w)* is the number of words. *n(s)* is the average number of sentences for each word. Percentage is the percentage of agreement between raters. κ is Cohen's Kappa.

5.1. *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 loved 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., Copestake & Briscoe, 1995; Dölling, 2020; Pustejovsky, 2005; Rabagliati, Marcus, & Pylkkänen, 2011; Srinivasan & Rabagliati, 2015). Furthermore, Alonso et al. (2013) provide an excellent starting source for sense-annotated data.

5.2. *Target polysemes and annotated data*

Our initial set of regular polysemes was sourced from the English dataset reported by Alonso et al. (2013). We thus started with between 5 and 55 polysemes in each regularity type (see Column A2013 in Table 1). For the one type with fewer than 10 polysemes (Artifact/Information), we manually added additional polysemes so that it also contained 10 items. We focused only on the singular form of each polyseme and avoided a small number of polysemes in the original set whose singular versus plural forms could introduce additional ambiguity (e.g., *glasses* can denote several drinking containers or spectacles).

One important point to note about our target polysemes was that they were all included as part of the base BERT vocabulary.[1] This point is important because BERT uses the WordPiece tokenization algorithm (Wu et al., 2016) to decompose words that are not included in the base vocabulary (i.e., low-frequency words) into subword units, which are often morpheme-like (e.g., -ion, -s). For example, the word "representationalism," which does not appear in the vocabulary, would be split into three subcomponents that are part of the base vocabulary: "representation," "#alis#," "#m." However, words that are in the base vocabulary are not subject to tokenization and are each represented as separate, independent inputs to the model (e.g., "construct" and "construction" are each represented separately and independently because they are in the base vocabulary).

Had all of our stimuli not been included in the base vocabulary, the tokenizer would have been necessary to process those polysemes and this could have had a direct impact on some of our results. For instance, many of the words in the Process/Result type end in "-ion," so if the words in that category had been subject to tokenization, it could have been the case that our regularity types were confounded to varying degrees because of how BERT was forming and integrating the representation of the subword components produced by the tokenizer. However, we avoided this major complication by using words in the base vocabulary. We originally made this choice not to avoid the issue of tokenization per se, but because we wanted to focus on words of a sufficient frequency[2] so that BERT would have enough data to develop representations of each sense usage that could be sensitive to the potentially subtle effects of regularity, as well as to increase the likelihood of identifying sufficient annotated examples of their uses in our labeled data (discussed below).

Notwithstanding the fact that we avoided the invocation of the tokenizer for our target polysemes, it is possible that there has been an indirect effect of tokenization on our results. For example, consider the hypothetical case that target polyseme "building" was in our vocabulary but "buildings" (or any other variation of the target polsyeme) was not. In this case, our target polyseme would not invoke the tokenizer but these other variations of that target polyseme would (e.g., decomposition "buildings" into "building" and "#s"). This would, in turn, lead BERT to adjust the weights related to "building" when training on "buildings" because of how that item was tokenized. Insofar as such decompositions are correlated with the presence of distinct but related meanings (e.g., countable nouns from "#s," use of an object to complete an action from "#ed"), this could have indirectly impacted our results. However, we consider it highly unlikely that such indirect effects are substantial drivers of any of the results that we observed. This is because the BERT base vocabulary includes the most frequent words in language, so the impact of such indirect effects is necessarily a rare occurrence that should not lead to major changes in the representations formed by BERT for each word. Confirming this prediction, however, is necessarily an empirical question that we consider outside the scope of the present work that focuses on the representation of regular polysemy because of the substantial theoretical and methodological issues that must be tackled (e.g., should the base vocabulary be expanded to reduce the need for tokenization, but at the expense of substantially increased computational costs? Should data related to words that are not in the base vocabulary be discarded? If yes, should only that word be omitted, or should surrounding context also be discarded? If surrounding context should be discarded, how much?).

Next, our initial source for annotated data was the set of sense-annotated sentences from Alonso et al. (2013). Alonso and colleagues sampled 500 sentences for each type from the American National Corpus (Ide & Macleod, 2001) and annotated them via Amazon Mechanical Turk. To increase the size of this initial sample, we supplemented this set of annotated data with our own annotations of lines of dialog taken from the Brysbaert and New (2009) subtitle database, for the top 10 most frequent words in each regularity type of the Alonso dataset. Specifically, we aimed to collect at least 100 sense-annotated lines of dialog for the most frequent polysemes (see Column B2009 in Table 1). For some of the polysemes, we exhausted all available lines in the subtitle database, limiting the total counts slightly below this level. The senses evoked in each line of dialog were then annotated by one of eight research assistants or by the first author. The annotators were tasked with indicating if the word denoted either one of the target senses (e.g., for CHICKEN, either an animal or a meat sense) or some other senses. Note that in both the data taken from Alonso et al. and our own annotated data, raters only had access to partial contextual information available from a sentence or a line of dialog, not the full context in which this text occurred (e.g., a full book, paragraph, or movie).

By combining these two sources of data, we were able to obtain a relatively large sample of polysemes and annotated sentences. Our expectation was the performance would increase as a function of large sample sizes, mainly because each sense, and by proxy, each class of senses (e.g., Animals, Meats) would be estimated in a more precise way. However, combining data in this way could also raise concerns about the introduction of other potential confounds. In supplemental analyses that parallel the main analyses reported in the results section, we confirmed that we obtained qualitatively similar results by analyzing the annotated data from each source separately and that overall performance increased when these two sources were merged, indicating that the sources themselves did not critically determine our results (see Supporting Information). Similarly, to rule out the possibility that our key results were impacted by the number of polysemes and/or the number of annotated senses associated with each type, we downsampled our data to match all regularity types on these measures and again obtained similar results (see Supporting Information). On this basis, our main results section focuses on analyses relating to the entire cleaned dataset since more data simply seems to have boosted overall accuracy, as expected.

Raters were also tasked with indicating if they were certain or uncertain about their rating as basic assay of rating confidence. We discarded all the lines for which the raters could not confidently identify a single specific annotation prior to conducting our main analysis (28 % of the data), although we retained them temporarily for computing interater reliability, described next.

5.2.1. *Interrater reliability*

We had 25% of the lines of subtitles dialog in our own annotated set reannotated by a different rater to gauge interrater agreement. After excluding ratings for which either rater was uncertain about which meaning was evoked (29% of the interrater data), and ratings that either rater indicated interpretations other than the two target senses (18% of the interrater data), we calculated the interrater reliability for each type. This was computed as the average of the interrater reliability scores computed across all pairs of different raters that contributed ratings for that type, weighted by the number of sentences rated by each pair of raters.[3] A summary of the agreement levels is presented in Fig. 2. This figure presents the degree to which raters were certain (bar labels prefixed with “C”) or one or both were uncertain (bar labels prefixed with “U”) and whether they indicated that the first sense (a) or the second

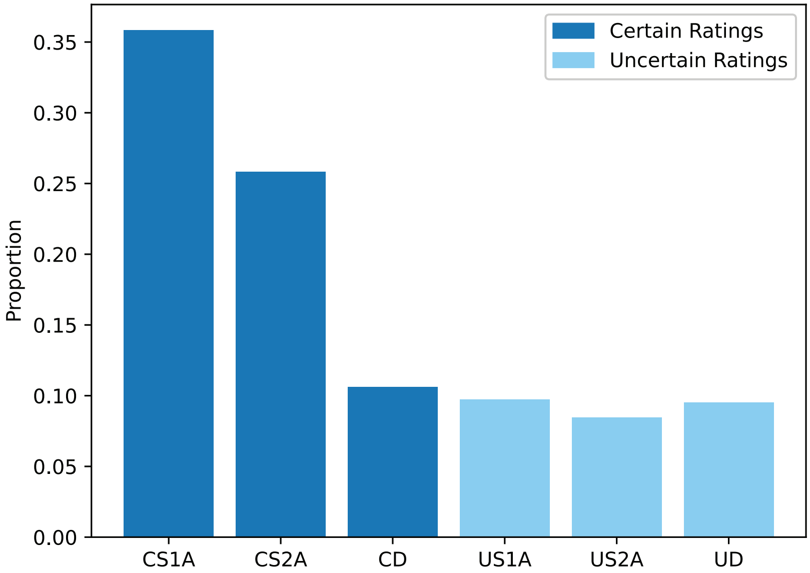


Fig. 2. Distribution of rater agreement and disagreement. CS1A = certain for sense1 agreement (Animal, Container, Location, Artifact, Process); CS2A = certain agreement for sense2 (Meat, Content, Organization, Information, Result); CD is certain disagreement (e.g., rater 1 indicated sense1, whereas rater 2 indicated sense2); US1A = uncertain agreement on sense1; US2A = uncertain agreement on sense2; UD is uncertain disagreement.

sense (b) of the polyseme was denoted by context. Here, the first sense was always defined as the first class listed when we enumerated the types of regular polysemy (e.g., Container in Container/Content), a decision that we will return to in more detail in our evaluation of shared cross-type structure. Cases of agreement were coded with the “A” suffix, whereas cases of disagreement were coded with the “D” suffix in the bar labels. As is clear from this figure, there was certain agreement for the majority of the lines of rated dialog and cases of disagreement were somewhat rare (only approximately 10% of the data when looking at certain disagreement, and approximately 20% of the data when collapsing both certain and uncertain disagreement). The ratio of agreement to disagreement did vary as expected as a function of certainty, with relatively more cases of disagreement when one or both of the raters was uncertain. The data from this figure also suggest that the first sense, as defined above, is more frequent than the second sense.

We also report two measures of interrater reliability for each category in Table 1: the percent of ratings in agreement and Cohen’s Kappa (Cohen, 1960). As summarized in the figure and table, we observed moderate levels of agreement across all types of regular polysemy, although some types were associated with higher overall levels of agreement than others. This moderate overall agreement was expected because the two senses of regular polysemy are very related hence more difficult to separate as compared to homonyms (e.g., Rice et al., 2019 obtained interrater reliability scores in excess of 90% for homonyms in the same Brysbaert & New, 2009 data). These results are also consistent with the levels of interrater reliability obtained in other earlier studies of sense annotation agreement for similar regular polyseme types, and for regular polysemy more broadly (Alonso et al., 2013; Markert & Nissim, 2002; Navarro, Marcos, & Abad, 2005; Véronis, 1998). Although we consider these levels of agreement more than sufficient for our present purposes, based on our experience here, we speculate that future work could improve overall agreement levels. It is also possible that some lines of dialog are consistent with both senses but different raters each confidently chose only one of them, which decreased the interrater reliability. We will return to discuss these points in the discussion section.

We also included two additional annotated datasets as control conditions. The first set consisted of homonyms and their corresponding annotated sentences in the same subtitles database reported by Rice et al. (2019). Our second set of control items consisted of samples of polysemes and their corresponding annotated sentences from Evans and Yuan (2017), who re-annotated all polysemes in the MASC corpus (Passonneau, Baker, Fellbaum, & Ide, 2012) and SemCor datasets (Mihalcea, 1998) with the New Oxford American Dictionary. Critically, this control set of polysemes (a) comprised a mixture of both regular and irregular polysemes, and (b) although it did include some regular polysemes, they were sampled at random across the population regularity types in natural language and were not grouped based on specific types as in our regularity types of primary interest.

5.3. *Deriving sense vectors*

For each regular polyseme, we derived a sense vector that corresponded to each of the two target senses of the polyseme. This was done by providing every sense-annotated sentence corresponding with the senses of interest 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 (i.e., contextual word vector) of this word for a given sentence (see Jawahar et al., 2019). We then took the average of the vectors associated with a given sense to derive the sense vector. The same method was used to derive representations for the control items.

5.4. *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. We first formulated all the sense analogy questions that could be asked for every polyseme of a given type in the form of (4a, 4b) (e.g., "which word sense belongs to the same sense class as SALMON_{Animal} and is similar to CHICKEN_{Meat}?" for the Animal/Meat type). Formally, for each sense analogy question $\_ : a^* :: b : b^*$, we transformed it into the form: $\_ \in cat(b) \wedge \_$ *is similar to* a^*. We then calculated a score for each candidate word sense as an answer for a given sense analogy question following the steps below.

1. We quantified the probability that this sense S is a member of the sense class of b, for example, $P(S \in cat(S^{animal}_{SALMON}))$, with a LR 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^* and this sense vector S using cosine similarity, for example, $cos(S, S^{meat}_{CHICKEN})$;
3. We multiplied the values obtained from the LR and cosine similarity to yield a score for this sense, for example, $P(S \in cat(S^{animal}_{SALMON})) \cdot \cos(S, S^{meat}_{CHICKEN})$.

Following these steps, a score was obtained for each sense vector as a candidate answer for a given sense analogy question. The sense vector with the highest score was the answer to the analogy question given by the model (Eq. 5), and was classified as either correct or incorrect. We averaged the classification accuracies for all sense analogy questions within each regularity type. Our assumption was that greater underlying regularity would yield higher accuracy.

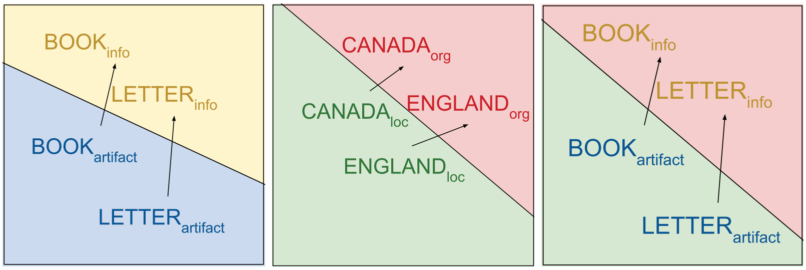
$$\arg\max_{S} P(S \in cat(S^{animal}_{SALMON})) \cdot cos(S, S^{meat}_{CHICKEN}) \tag{5}$$

We also computed equivalent analyses for each of our control conditions (random polysemes, homonyms, and cross-type polysemes, described below).

5.4.0.1. Random polyseme and homonym controls: The control conditions served to establish a baseline regarding the capacity to answer sense analogy questions for words that did not share a specific underlying regularity. For the random polyseme and homonym controls, we randomly selected an equal number of words in the respective datasets, matching the average count for a regular polysemy type, so that any differences we observed were not due to the amount of data included in each condition. For each word, we randomly selected two meanings/senses and randomly mapped them onto two sense classes for the control LRcos analyses, which were otherwise identical to those described above. This whole procedure was repeated 5,000 times and the results were averaged to ensure that the results obtained via random samples were stable.

Homonyms are defined as having unrelated meanings, so we expected that sampling random sets of homonyms and arbitrarily assigning each of their meanings to one of two classes as a control "type" will lead to very poor accuracy using the LRcos method. We also expected random polysemes to be associated with lower performance as compared to our analyses of regular types. However, insofar as there is some latent relationships shared across types (as might be reflected in the LR component of the model) and/or reflected in the relatedness between senses (as reflected in the cos component of the model), we might expect higher overall performance in this control condition as compared to the homonym controls.

5.4.0.2. Cross-type analysis: The cross-type analysis was implemented to probe for higher order regularity shared across different regularity types. In other words, it aimed to examine how analogous sense mappings in one type might relate to those in another type. We conducted this analysis because the theoretical linguistics literature suggests, under various different labels, that there is an abstract relationship that helps shape the emergence of polysemous senses across all types, such as concrete to abstract sense extension (e.g., Copestake & Briscoe, 1995; Nunberg, 1995; Pustejovsky, 2005). If such accounts are correct



(a) *Target regularity type* (b) *Source regularity type* (c) *Using the source regularity type to classify the target regularity type*

Fig. 3. Simplified depiction of the cross-type analysis using one source type (Location/Organization) as the classification base for the target type (Article/Information). This allows the model to answer questions of the form what is to BOOK$_{Information}$ as CANADA$_{Location}$ is to CANADA$_{Organization}$? Panel (A) Target regularity type (Artifact/Information) and the logistic regression model trained on this type separating the two classes (blue vs. yellow). (B) Analogous panel for a different type of regularity (Location/Organization) and the classes (green/red) identified by logistic regression for that type. (C) Using the logistic regression from the source regularity in (B) to answer sense analogy questions from the target regularity type in (A).

and there is some type of latent regularity between the base and extended senses of a polyseme that is shared across regularity types, we should be able to train the LR model on one regularity type (e.g., Location/Organization) and use that model to successfully classify another regularity type (e.g., Artifact/Information) above the level expected by chance, as established by our other control conditions. Put concretely, the cross-type condition involved asking sense analogy questions such as “what is to BOOK$_{Information}$ as CANADA$_{Location}$ is to CANADA$_{Organization}$?”

To test the cross-type regularity of a given type (e.g., Artifact/Information) (Fig. 3A), we applied the LRcos method mentioned above with the following modification: For a target type of regular polysemy (e.g,. Artifact/Information), we first trained separate LR models on each of the other regularity types (e.g., Animal/Meat, Container/Content, Location/Organization, Process/Result). We then used these models as the basis for classifying items from the target type (e.g., Artifact/Information) in four separate LRcos models (one of which is exemplified in Fig. 3B). The cross-type accuracy for the given type is the average accuracy of all four LRcos models. In additional analyses, we also provide a more detailed breakdown of how the classifier trained on each type of regular polysemy contributed to this average accuracy. The process of training on one classifier and testing on another classifier is illustrated in Fig. 3C. The intuition tested here is that if the LRcos model can still achieve a reasonable accuracy above baseline, it suggests that there is some latent structure shared across types.

In the cross-type analyses, the first class in the classifier was always the first class noted for each regularity type (e.g., Artifact in Artifact/Information; Location in Location/Organization). This approach was taken because it aligns the base sense for a given regularity type with the base sense from another regularity type, as defined in past theoretical linguistics work (e.g., Copestake & Briscoe, 1995; Nunberg, 1995; Pustejovsky, 2005). In a supplemental analysis, we reversed the order of classes in the classifier when mapping to another regularity type, and observed a decrease in performance, as predicted by this intuition (see Supporting Information).

6. Results

Before turning to the detailed quantitative results as they relate to our four key questions, we first present a visualization of the space which we used to probe sense analogy questions. Fig. 4 displays the sense vectors for each type of regularity after the 768-dimensional representational space employed by BERT had been reduced to two dimensions using principal component analysis (PCA). In this figure, the two senses of each regular polyseme are linked with a line, with the dotted end of the line corresponding to the base sense, as described above. Although PCA is designed to identify dimensions that best explain variance in the distribution of senses in the overall semantic space, which is not identical to identifying the optimal plane for delineating between two sense classes as in the LR component of the LRcos model, several broad patterns of structure are clearly present in these data that are worthy of note. First, the senses from the different regularity types are clearly separated into different clusters, with only minimal overlap between the Artifact/Information and Process/Result types. This indicates that answering sense analogy questions for a given type requires models that are sensitive to relatively fine-grained distinctions within different parts of the representational space so as to distinguish between closely related senses. It also indicates that our cross-type analyses will involve inferences made across much larger portions of the representational space, for instance, when using Location/Organization relationships to make inferences about Artifact/Information relations. Second, within each regularity type cluster, the vectors mapping between the base sense and the extended sense generally point in the same direction. This suggests that there is some shared transformation that relates each base sense to its extended sense. (We have further investigated the geometric structure between pairs of senses in our parallelity analysis, presented in Supporting Information, which repeats our key analyses related to each question of interest using this alternative geometric method for answering sense analogy questions). Of course, these vectors are not all perfectly parallel, which is to be expected for several reasons: First, there are obviously idiosyncratic features associated with each sense of a polyseme that will shape the overall direction of the vector. Second, as alluded to earlier, the dimensionality reduction produced through PCA here does not identify the optimal sets of dimensions for each individual cluster, which we would expect to vary somewhat even if there is some shared latent structure across regularity types. Thus, the fact that there is some overall structure within each type bodes well for analyses using the LRcos method, as we would expect that separate LRs conducted for each regularity type will identify better dimensions for delineating between the classes than are reflected in the PCA representation. Third, there are typically one or two classes that have broadly similar directions. This might suggest that our cross-type analyses, which involve using the LR classifiers from one type to classify the senses of another type, could be shaped by the different patterns

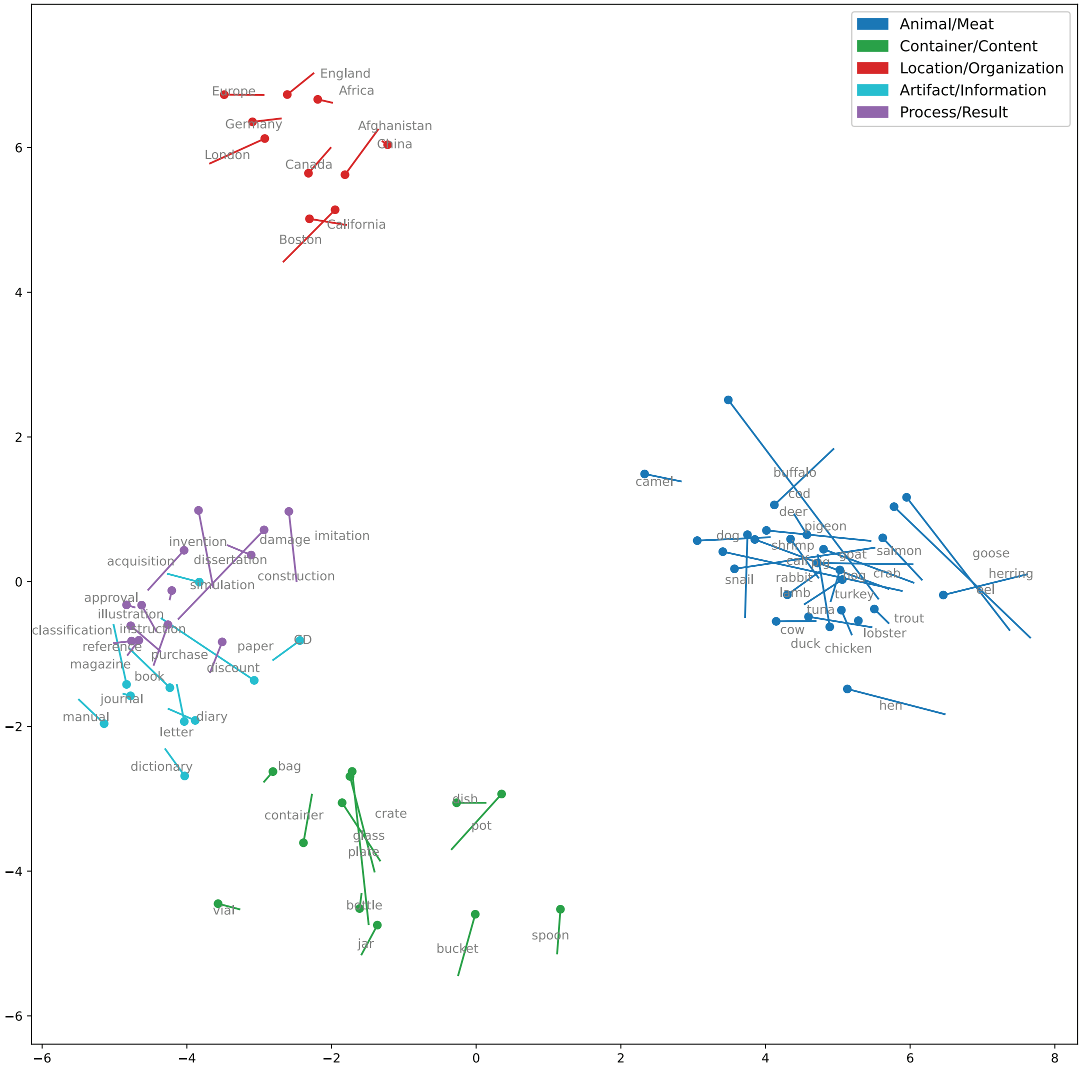


Fig. 4. Plot of the sense vectors in the BERT representational space after being reduced to the first two principal components through PCA analysis. Each line represents two senses of a polyseme. The dotted end of the line represents the base sense (Animal, Container, Location, Artifact, Process) of the polyseme. The other end represents the extended sense of the polyseme.

of similarity between pairs of regularity types, although this remained to be tested using LR given the aforementioned potential limitations of the PCA representation.

With these insights in mind regarding the structure that exists among regularity types, we now turn to our main results and how they relate to each of our key questions. Our key results are presented in Fig. 5 and are unpacked in more detail in relation to each key question, in

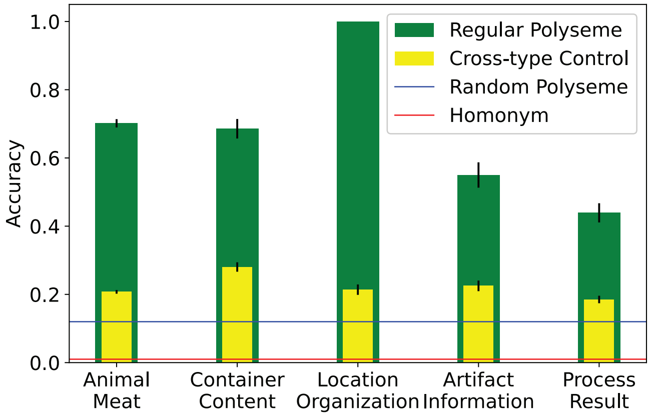


Fig. 5. Mean accuracy for each regular polysemy type and for the control conditions. Error bars correspond to the standard error for all bars in the plot. For the random polyseme and homonym controls, the average performance across 5000 samples is plotted.

turn. For each of the key questions, we controlled for multiple comparisons in our statistical analyses using the Holm correction (Holm, 1979).

6.1. Question 1: Does the representation in a distributional semantic model reflect the shared structure of a given type of regular polysemy?

To answer our first question regarding whether the representations in a DSM reflects the shared structure of a specific type of regular polysemy, we compared the accuracy obtained using the LRcos method within the five regularity types against the random polyseme and homonym controls with chi-square tests of goodness-of-fit. Since the LRcos model leverages the underlying regularity to answer sense analogy questions, a higher accuracy for the regular type was expected relative to the controls. All of these comparisons were statistically significant, as summarized in Table 2 and Fig. 5. This indicates that YES, the representations in the model are sensitive to the shared regularity structure within each regularity type.

6.2. Question 2: Are there variations in the degree of regularity across different regular relationships, reflective of a graded "regularity continuum"?

Traditionally, regular polysemy is considered to be a discrete, dichotomous category within polysemy, such that a polyseme is either regular, or is not. However, an alternative perspective suggests that regularity might exist on a continuum, allowing for varying degrees of regularity for a given polyseme rather as opposed to a dichotomous distinction. To answer this question, we first conducted a chi-square test of homogeneity on the five types. Since the LRcos model relies on regularity within a given type to answer sense analogy questions, its accuracy reflects the degree of regularity in a given regularity type. The result was significant ($\chi^2(4) = 183.58$, $p < .001$), which indicated that these five types were not homogeneous

Table 2
Summary of the mean accuracies and statistical comparisons from the main analyses and the cross-type analyses

	Main Analysis					Cross-type Analysis				
	Accuracy	versus RP		versus H		Accuracy	versus RP		vs. Main Analysis	
		$\chi^2(1)$	p	$\chi^2(1)$	p		$\chi^2(1)$	p	$\chi^2(1)$	p
A/M	0.70	4497	**<.001**	67,827	**<.001**	0.21	1288	**<.001**	407	**<.001**
C/C	0.69	800	**<.001**	12,172	**<.001**	0.28	149	**<.001**	257	**<.001**
L/O	1.00	1320	**<.001**	17,820	**<.001**	0.21	378	**<.001**	60	**<.001**
A/I	0.55	315	**<.001**	5302	**<.001**	0.23	72	**<.001**	75	**<.001**
P/R	0.44	301	**<.001**	5803	**<.001**	0.19	88	**<.001**	50	**<.001**
RP	0.12	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
H	0.01	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

Note. A/M = Animal/Meat, C/C = Container/Content, L/O = Location/Organization, A/I = Artifact/Information, P/R = Process/Result, RP = random polyseme control, and H = homonym control. Significant *p*-values after the Holm correction are presented in boldface.

Table 3
Summary of the statistical tests comparing the accuracy for each regularity type against the accuracy for every other regularity type

group1	group2	$\chi^2(1)$	*p*
A/M	C/C	0.199	.656
A/M	L/O	71.512	**<.001**
A/M	A/I	16.273	**<.001**
A/M	P/R	76.541	**<.001**
C/C	L/O	67.549	**<.001**
C/C	A/I	7.877	**.005**
C/C	P/R	34.149	**<.001**
L/O	A/I	101.951	**<.001**
L/O	P/R	154.259	**<.001**
A/I	P/R	5.189	**.023**

Note. Significant *p*-values after the Holm correction are presented in boldface.

in terms of their overall accuracy levels as determined via the LRcos method. We then compared the five regular polysemy types against one another. The results of these comparisons are reported in Table 3. All of the comparisons except between Animal/Meat and Container/Content were statistically significant. The large number of significant differences between the types, 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.

As an additional control analysis, to test for a potential confound due to the different numbers of polysemes included in each regularity type in determining LRcos accuracy, we tested for the correlation between these two measures. This correlation was low and non-significant, ($r(3) = -0.04$, $p = .95$), which indicated that the number of polysemes per type did not drive our results.

6.3. Question 3: Is there any higher order latent structure shared across the different types of regular polysemy?

To answer our third question regarding whether there is any higher order latent regularity shared across different types of regular polysemy, we initially conducted two tests. First, we compared each cross-type accuracy against the random polyseme control with chi-square tests of goodness-of-fit to see if there is shared regularity contributing to the regularity of each different type. The assumption here is that if we can successfully employ the classifier from one regularity type to support classifications in another regularity type above the levels established by the random polyseme control, this would indicate that there is shared structure within a regularity type. We found that all five cross-type accuracies were associated with significantly higher accuracy than the random polysemy control, as summarized in Table 2 and Fig. 5. This suggests that YES, there is regularity shared across regularity types. Second, We compared each regularity type with its cross-type analog with chi-square tests of independence to see if accuracy was significantly higher when the classifier for a given regularity type was employed instead of the cross-type classifier. We found that accuracy was higher when the classifier for a given type was used as opposed to the cross-type classifier, as also summarized in Table 2 and Fig. 5. Considered together, these two findings suggest that there is indeed some shared structure across regularity types, but there is also structure unique to each regularity type.

Having established that there is indeed some shared structure across regularity types, we delved deeper into the nature of this shared regularity. In the cross-type data analyzed thus far, we averaged the accuracy of the four separate cross-type LRcos analyses conducted for a given regularity type, wherein we used each of the other types as the basis for classifying the target type. However, focusing only on the average cross-type performance could mask a more nuanced understanding of the nature of the shared relationship between types. One possibility is that there is a rather homogeneous level of latent structure shared across all types, which would be reflected in similar levels of accuracy if we examined the performance of each individual classifier prior to producing the average cross-type accuracy. Another possibility is that accuracy is higher for only a subset of the classifiers, which could indicate that there are multiple types of latent structure that are shared across some, but not all types. Put differently, although our initial analysis established that there is shared structure across types, these additional analyses aim to determine whether this reflects a single latent structure or multiple latent structures (for a parallel debate on these types of considerations, see the discussion of a single general intelligence factor versus seven primary mental abilities as underlying performance in a range of tasks, as presented by Spearman, 1904 and Thurstone, 1938, respectively).

To probe the aforementioned issue, instead of averaging the cross-type accuracies for the different source types for the LR model, we examined accuracy for all pairs of source types for the LR models and target regularity types. For a given pair of types—Animal/Meat and Location/Organization, for example—we computed the average of two accuracy results: the results of using Location/Organization as the base for classifying Animal or Meat, and of using Animal/Meat as the base for classifying Location or Organization. This average reflects

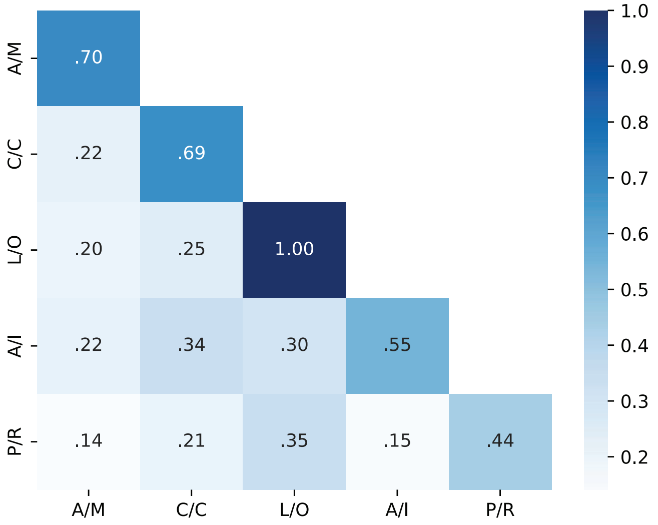


Fig. 6. Accuracy obtained by using each regularity type as the basis for classifying every other regularity type. A/M = Animal/Meat, C/C = Container/Content, L/O = Location/Organization, A/I = Artifact/Information, P/R = Process/Result. The diagonal corresponds to the within-type classifications. All cells below the diagonal correspond to the individual cross-type comparisons.

the shared regularity that is being used to infer all relevant sense classes in Animal/Meat and Location/Organization. The full results are plotted in the heatmap in Fig. 6.

From the results, we can see that the accuracy obtained using different individual types as the basis for the cross-type controls was non-identical across the cells, ranging from 0.14 to 0.35. To better understand the nature of these differences, we conducted several follow-up tests. First, we compared each specific cross-type accuracy against that of the random polyseme control with chi-square tests of goodness-of-fit to see if there is shared regularity above the random polyseme control. The results indicated that all specific cross-type comparisons were significantly above the random polyseme control except when using Animal/Meat to classify Process/Result and vice versa, and using Artifact/Information to classify Process/Result, and vice versa. (See Table A1 in the Appendix.) This indicated that there is some latent structure shared across all of the types. Second, we ran a chi-square test of homogeneity on the 10 cross-type conditions to determine if the accuracy levels obtained in some cross-types were significantly different from those in other cross-types. The result was significant ($\chi^2(9) = 179.1$, $p < .001$). We therefore followed up this test by comparing each cross-type against the others. The detailed results of these comparisons are presented in Table A2 in the Appendix. The results revealed that although many of the comparisons produced similar levels of accuracy, there were also many significant differences across types. These results indicate that although there may be some higher order latent structure shared across all types, as reflected by the non-significant differences between many cross-type accuracy comparisons, there may also be more nuanced substructures shared across some but not all regularity types. For example, overall accuracy was non-significantly different for the following pairs of cross-type comparisons: Container/Content versus Artifact/Info (0.34), Artifact/Info versus Location/Organization (0.30), and Process/Result versus Location/Organization (0.35). The accuracy levels of these three comparisons were also significantly higher than those in many of the other cells.

Taken together, these results are consistent with some general latent structure shared across types, for instance, the transformation from the base sense to the extended sense posited by Nunberg (1995), Copestake and Briscoe (1995), and Pustejovsky (2005), as well as some substructure common to some but not all types, for instance, the Concrete/Abstract distinction posited by Lakoff and Johnson (1980).

6.4. Question 4: Can the degree of regularity be fully explained by the degree to which the semantic representations denoting each of the regular meanings overlap?

Finally, we probed whether variations in regularity could be explained in terms of the meaning overlap among a word's senses. For instance, is it possible that higher degrees of regularity are observed when the distance between the two sense classes is small? To answer this question, we first computed the average meaning overlap between the two classes in each regularity type by averaging the cosine similarity between the two senses of each word. We then correlated these results with the accuracy data from the LRcos analysis for the five regularity types. There was no significant correlation between these two measures ($r(3) = -0.292$, $p = 0.63$), which indicates that there is no evidence that semantic overlap explains a significant amount of the degree of regularity in our data, as reflected in the accuracy data from the LRcos model. However, we acknowledge that the lack of a significant effect in this instance may be shaped by two distinct but potentially related aspects of our experimental methods. First, it could be the case that our regularity types all have relatively high levels of overall regularity, in which case the restricted range of regularity values near ceiling could impede the detection of a significant effect. Second, analyzing only five regularity types may have reduced the power of this inference —a correlation of -0.29 in a larger dataset could indicate that some (but not all) of our results could be explained by meaning overlap.[4] Based on the results at hand then, we take our data as indicative that our regularity effects cannot simply be reduced to effects of meaning overlap, but we do not rule out the possibility that there is some relationship between these constructs. We are hopeful that the work reported here will serve as a basis for expanded analyses on this front in the future.

6.5. Investigating how the LRcos method works

As described above, our LRcos model performed well in detecting structure among regular polysemes, although it did not achieve perfect accuracy. This good-but-not-perfect performance motivated us to further investigate the operation of the LRcos model. The goal here was to understand how the LR and cos components contribute to determining overall accuracy, as well as to explore whether a variation of this approach can be created that would yield even higher performance.

First, to better understand the operation of the LRcos model, particularly with respect to errors, we examined the incorrect answers it produced to our sense analogy questions. In so doing, we inferred that the model could produce an error in two different ways. For the first type of error, the model misidentified the target word entirely. For instance, given the sense analogy question, "What is to TROUT$_{Animal}$ as CHICKEN$_{Animal}$ is to CHICKEN$_{Meat}$?," the LRcos model suggested SALMON$_{Meat}$—a completely incorrect word, although from within the correct class. This type of error suggests that it is difficult in some circumstances to make fine-grained distinctions between the highly similar representations of words within a sense class (e.g., the MEAT senses of different fish), but the model is able to identify the correct sense class. Put differently, these results suggest that the LR classifier was supporting correct classification, but the model could not use word sense similarity information from the cos term to identify the correct word. For the second type of error, the model answered with an incorrect sense of the correct target word. For instance, for the analogy question "what is to IMITATION$_{Result}$ as DAMAGE$_{Process}$ is to DAMAGE$_{Process}$," the model answered IMITATION$_{Result}$, an incorrect sense, and a repetition of one of the probe words in the question. This second type of error highlights a subtle but important implementational aspect of the LRcos model: when searching for answers to a sense analogy question: $\_ : a^* :: b : b^*$, the question terms a^*, b, and b^* were *not* excluded when scoring candidate answers. This made the task more challenging than if we had simply eliminated the items presented in question as viable responses. However, not eliminating these items would allow the LRcos method to rely strongly on the similarity between a and a^* to give the correct answer (Linzen, 2016; Rogers, Drozd, & Li, 2017). Excluding a^*, b, and b^* as candidate answers prevented the model from giving the wrong answer a^*, despite this wrong answer being the "best" answer otherwise. Given our aim here was to measure regularity, as reflected in a consistent change in the relative location of the two senses of a regular polyseme within the semantic space of a distributional model, we allowed the items presented in the question to be included in the question as a more stringent and accurate test of the model's performance. This ensures that regularity, not simply similarity, supports answering sense analogy questions for regular polysemes. In sum, when considering how the model made such an error in light of this important implementational detail, it is apparent that the LR component failed to identify the correct sense class. However, it is not so simple as to attribute this error solely to the LR component because it is also possible that the similarity between the first and second senses of the word (as reflected by the cos term) was low in the case of these types of errors, so even correct classification of the class could still have yielded an error, similar to the first type of error described above.

Building upon the basic insights produced from our error analyses, given that the LRcos model involves combining two distinct components—the probability derived from LR and the similarity derived from the cosine measure (cos)—we investigated how each of these components was able to contribute to correctly answering sense analogy questions. As an initial step on this front, we examined whether each of these components captured different aspects of regularity. To do so, we correlated the probability from the LR term and the similarity from the cos term, and observed only a very low correlation ($r(97, 463) = -.08$, $p < .001$). This low correlation indicated that these two components captured distinct aspects of the sense analogy question.

Next, we conducted a grid search for the weightings of *LR*, *cos*, and the multiplicative term $LR \cdot cos$. Our goal was to investigate how varying the weightings of these terms will influence accuracy, including the effect of ablating each term. In this way, we aimed to ascertain not only the relative contribution of each component of LRcos in accurately answering sense analogy questions, but also to determine the potential accuracy improvement with different weightings of the LRcos components if we included both the additive effects of LR and cos, in addition to their interaction. In addition to shedding basic light onto the relative contributions of each of these terms, we were also motivated to conduct this analysis by the original work on LRcos reported by Drozd et al. (2016), who noted that their formulation, while effective at answering analogy questions, was not necessarily formally optimal. Thus, a variation on their approach could yield additional performance improvements that could facilitate the study of (sense) analogy questions. In the grid search, we incorporated both the sum and product of LR and cos in scoring each candidate: $\alpha \cdot LR + \beta \cdot cos + \gamma \cdot LR \cdot cos$, with α, β, and γ being the adjustable parameters. The range of parameters was $\alpha, \beta, \gamma \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. The full results of the grid search are presented in Fig. 7.

First, we found that models that only included the LR term or the cos term performed very poorly, with accuracy levels of 0.07 and 0.00, respectively. This indicates that both components are necessary to obtain acceptable levels of performance. Next, we found that using only $LR \cdot cos$ with the weights for the two additive terms set to zero, which was what was used in the analyses conducted to answer our four key questions above, produced essentially identical performance to a model with equally weighted *LR* and *cos* components $LR + cos$ (0.68 vs. 0.67, respectively). However, selecting the optimal weightings of the *LR* and *cos* terms in the $LR + cos$ model improved performance for this additive model to 0.79 (with $0.4 \cdot LR + 1 \cdot cos$). Thus, an additive as opposed to a multiplicative combination of these terms could be used to further improve the performance if optimal weightings of each term are used. Furthermore, if we included the optimal weightings of all three terms, performance further increased to 0.81 (with $0 \cdot LR + 0.6 \cdot cos + 0.4 \cdot LR \cdot cos$, or, effectively a model that includes both an additive contribution from cos as well as the interaction between LR and cos). If we take a step back from the optimal weightings of the model, it is clear that even if the exact optimal parameters are not used, there exists a large range of weightings that produce quite strong overall results, suggesting that there is robust value in considering the contributions from each of these three terms in optimizing the model used to answer sense analogy questions.

Of course, given our grid search represents only a single case of post-hoc analyses, we remain cautious in making strong overall recommendations for future work because these performance improvements might hinge on the specific data or question that we studied. For instance, we could imagine a scenario where the items being compared are distributed differently and lead to different sets of optimal parameters and relative differences in the performance across models with additive and/or interactive terms. Thus, we used the established LRcos model to answer our four key questions. However, we are excited by the possibility that substantial performance improvements of 10% or more could be obtained by optimizing the model that is used to answer sense analogy questions, as well as by how probing the model's errors and optimal parameter weightings could shed additional insight into how the

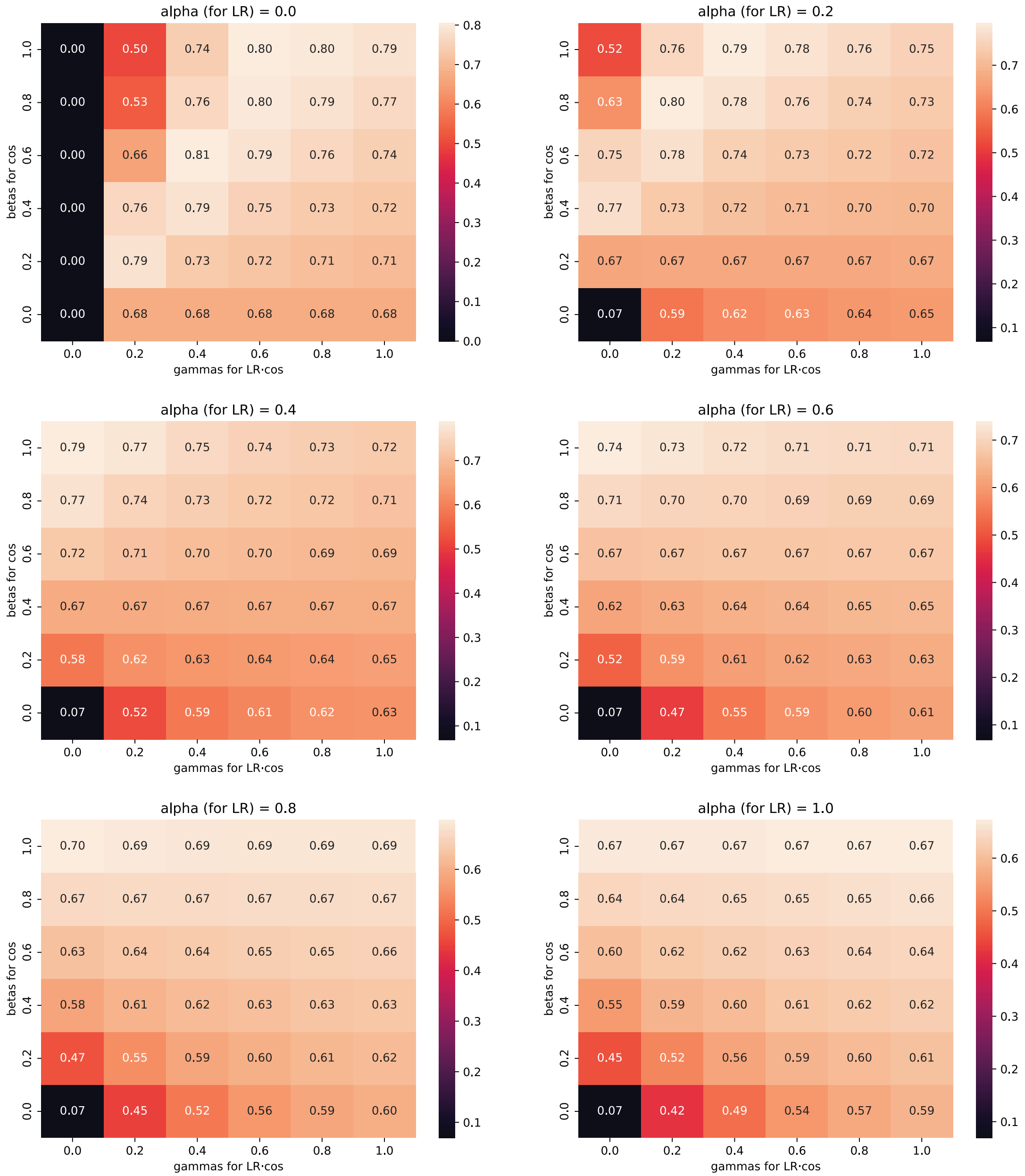


Fig. 7. Overall accuracy in answering sense analogy questions for different weighting combinations of the LR, cos, and LRcos components of a model based on $LR + cos + LR \cdot cos$.

model actually answers such questions. We expect that our initial analyses on this front will provide some a priori predictions for future work in this vein.

7. Discussion

Most words are polysemous and have related but distinct senses. Many polysemes can further be classified as regular polysemes because they share the same overall representational structure of their senses (e.g., Animal/Meat). In our work, we aimed to further our collective interdisciplinary understanding of regular polysemy with the help of a DSM, here exemplified by BERT, which can provide us with quantitative measures of the regularity of a given type of regular polysemy and the intricate relationships between different types of regular polysemy, complementing and refining prior researcher intuitions and verbal theorizing. More specifically, we examined the representational structure of regular polysemy in BERT via sense analogy questions, a method that we adapted from the word analogy research in the field computational linguistics.

We answered four main research questions with respect to the representational structure of regular polysemy. First, is there significant shared structure across regular polysemes sharing the same regularity type? 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 learning a new regular meaning for an existing word (Brochhagen, Boleda, Gualdoni, & Xu, 2023; Rabagliati et al., 2011; Srinivasan & Rabagliati, 2015; Srinivasan & Snedeker, 2011). For example, upon learning that a novel word denotes a container, an individual could infer that this word could also be used to denote its contents.

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 either 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 false dichotomy was revised along a relatedness continuum that delineates between homonyms and polysemes (e.g., Armstrong & Plaut, 2016; Klepousniotou et al., 2008; Rodd et al., 2002), we argue for an additional decomposition of polysemy across a regularity continuum.

Third, we asked whether there is any higher 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 overall. Furthermore, we observed that there may be additional subtypes of even more related structure that are common to only a subset of regularity types. This suggests that although there may be some latent structure shared across different types when mapping from a base sense to an additional sense, there may also be multiple additional structures that are shared to varying degrees across different regularity types. Exactly what this latent structure might be will require additional investigation. Speculatively, based on the cases we examined, one type of structure that could underlie some of our effects is the Concrete/Abstract structure proposed by Lakoff and 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, Malt, & Srinivasan, 2017). However, this structure does not appear to fit with all of the categories that we examined, such as the Animal/Meat type, for which both senses appear to be highly concrete. Conceivably, the more abstract structure shared between this category and the other categories could be attributable to a distinction between a base sense (Animal) and an extended sense (Meat), although such an explanation currently suffers from the fact that it is not clear how to determine, a priori, which sense is the base sense (e.g., is it the class with the greatest number of exemplars? The first class learned? Some other possibility?).

Fourth, we tested whether the systematic differences between regularity types could be explained by meaning overlap. On this point, our analyses did not provide evidence that meaning overlap recapitulates the regularity effects we obtained through our LRcos analyses at the level of different types. This is in contrast with the meaning overlap at the item level, where for each sense analogy question the meaning overlap, as reflected by the cos term on the LRcos model, did contribute to the performance of LRcos model, which is in line with the conclusion in Floyd et al. (2021) that similarity (in terms of meaning overlap) is the major determiner of colexification of two senses. Nevertheless, more data are clearly needed to probe this issue in a more comprehensive way. For theoretical reasons, however, we expect that although these factors may ultimately be found to be at least partially related to one another, each will continue to make unique contributions to the structure of the representational space and one will not ultimately be reduced to the other. For instance, a polyseme could have two very closely related meanings that are related in a very idiosyncratic way, making the relationship between regularity and relatedness imperfect at best. This result, along with the other results summarized above and the previously noted history of using models such as BERT as approximations of the human mental lexicon, suggests that the organization of the mental lexicon and the broader conceptual system is driven not only by semantic overlap. Rather, there are additional pressures such as the consistency with which many words share similar relationships among their meanings.

Lastly, we examined how the LRcos model works to better understand what contributed to the model's capacity to detect regularity in regular polysemy. According to previous research, the LRcos model performs better in detecting word analogies than traditional parallelity methods (Drozd et al., 2016). The primary advantage of LRcos is that it evaluates the entire set of analogous items instead of just comparing individual items. This feature aligns perfectly with our objective of quantifying the regularity of each type overall. We used grid search to determine that although neither LR nor cos on their own produce accuracies above floor, their additive combination produces similar accuracy to the multiplicative LRcos model that we employed in our work. This suggests that the abstract intuition of combining a similarity measure with a classification measure is what is critical to answering sense analogy questions, not the specific LRcos implementation (for additional discussion, see Drozd et al., 2016). Furthermore, we observed post hoc that an optimal weighting of the LR, cos, and LRcos terms could yield a further 10% increase in accuracy. If this post-hoc finding stands up to additional scrutiny, it could offer an important methodological advance for testing analogy questions, including sense analogy questions, particularly if tapping the lower end of the regularity continuum where performance would be expected to decrease.

Of course, the empirical basis for the inferences we have drawn, as outlined above, hinges to some degree on the five regularity types that we analyzed. These types were selected because they were the subjects of extensive prior study and in this respect are arguably the most informative for developing an integrated understanding of regular polysemy from various theoretical and experimental angles (e.g., Copestake & Briscoe, 1995; Dölling, 2020; Pustejovsky, 2005; Rabagliati et al., 2011; Srinivasan & Rabagliati, 2015). However, a potential concern from this selection is that prior research has focused on a non-random sample of regular polysemes and regular types. For instance, this work may have focused on types that have the most consistent underlying regularity structure, and/or which might be driven most strongly by an abstract/concrete relationship between sense classes. Understanding 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 many words that have a sense associated with each of these classes. Recent work by Yu and Xu (2023) employed a similar approach. In that work, the authors inferred the extensibility of a new sense from the old senses of a given word by employing the chaining algorithms previously outlined by Lakoff (1987) as a likelihood function. With this function, they calculated the posterior probability of a new sense being an extension of the older senses of a word. Analogously, our methods could utilize learned regularity structures as a likelihood function to calculate how extensible a new sense is from old senses of a word, essentially instantiating another aspect of Lakoff's chaining theory in an explicit computational model, which can then be the subject of experimental validation. In so doing, we could move toward a more complete computational analysis and mechanistic theory of the relationships between polysemes in language overall, assuming that the representations from models like BERT are sufficient approximations of representations in the human mental lexicon so as to generate informative insights. For reasons outlined earlier, we view the current literature as supporting this assumption, notwithstanding the imperfect match between the models and human data, and we expect that newer computational models with further improve on this front (as discussed below). However, even if this assumption ultimately proves to be false, there is still much value to be gained from this type of work: For example, it provides an additional novel approach for evaluating the cognitive plausibility of the models, and mismatches between human and model performance can feed back to give clues regarding how to further improve the models. Furthermore, if the model can accurately discover new regular polysemes and types of regular polysemy, even if the mechanism used to do so is cognitively implausible, this would nevertheless be very valuable as a methodological tool. For instance, it would allow researchers to develop improved samples of regular polysemes rather than rely upon small sets of manually selected regular polysemes that may be recycled across studies. Such recycling of items has previously been found to inflate the magnitude of experimental effects and hinder accurate inferences regarding the magnitude and significance of a number of linguistic effects (Forster, 2000).

Having highlighted the empirical contributions of our work in answering our key research questions and its implication for broader field including conceptual system and computational methodology, it is also worth discussing several issues that we encountered in conducting this work that are particularly salient in the study of sense analogies questions and regular polysemy. We elaborate these issues below and discuss how these issues could be ameliorated in future work.

7.1. Computational methodologies

Our results indicate that BERT can capture the conceptual relationships between the senses of regular polysemes, even within subtitle datasets that inherently emphasize visual elements. This finding suggests that this model, and, we assume, the class of models that it exemplifies, succeeds to some degree at least in capturing the conceptual role aspect of meaning even if it lacks other important aspects of meaning representation, such as referential grounding (i.e., embodiment).

Despite the well above chance performance of the LRcos model, most types of regular polysemy still only achieved below ceiling accuracy scores even after our parameterization of an improved variation of the technique using grid search. Apart from the intrinsic difference in regularity among each type, there are a couple of potential reasons why LRcos might struggle to detect sufficient regularity within a certain type. First, the boundary between two senses of ambiguous words could be nonlinear so that the LR in the model might not be fully suited to making these types of delineations. Second, there is a potential void in the BERT semantic space between some classes of senses of a word wherein this part of the space does not correspond to any intelligible sense, as investigated by Karidi, Zhou, Schneider, Abend, and Srikumar (2021). This might also decrease the performance of the LR in LRcos. Lastly, a single polyseme token could sometimes take on two senses simultaneously, a phenomenon referred to as "copredication." For example, in the sentence "Judy's dissertation is still thought-provoking although yellowed with age" (Cruse, 1986), the regular polyseme "dissertation," an Artifact/Information type, is used in both artifact and information senses. We did not include instructions to explicitly label such cases in our annotation procedure. A partial inspection of our annotated subtitles database indicated that such sentences were rare, but we expect that flagging copredication in the future may yield further performance increases.

7.2. Distributional semantic models, regular polysemy, and human cognition

Of course, for the work that we have conducted to be maximally relevant to the cognitive sciences, the computational model that we employed, as well as the general class of such DSMs, must have a history of being relevant to human cognition. To briefly reiterate some of the points raised in the introduction: we chose BERT for our examination of the representational structure of regular polysemy due to its large size and proven track record in simulating a range of aspects of human cognition (Rogers et al., 2021), as well as its capability to produce contextual word vectors by attending to both the preceding and subsequent contexts of a given target word. However, choosing BERT does not preclude the generalizability of our results to other DSMs, as previous research has reached similar conclusions using a variety of DSMs (Floyd et al., 2021; Lopukhina & Lopukhin, 2016; Trott & Bergen, 2023). More importantly, probing DSMs is a well-established method for examining human language cognition, not only in terms of linguistic behavior, such as response times in lexical decision or naming tasks (Mandera et al., 2017), and eye-tracking-based reading times (Heilbron, van Haren, Hagoort, & de Lange, 2023; Pimentel, Meister, Wilcox, Levy, & Cotterell, 2023), but also in internal cognitive/neural representations aligned with EEG (Ettinger, Feldman, Resnik, & Philips, 2016) and fMRI data (e.g., Schrimpf et al., 2021). We therefore interpret our work as shedding light on the internal cognitive representations of regular polysemy, engaging with work in theoretical linguistics, and as being suitable for potential extension to help understand the human neural representation of regular polysemy.

This latter avenue is also a compelling direction for future research, given that existing studies involving neural representations have primarily examined the relatedness between different meanings of polysemes in terms of meaning overlap (Klepousniotou et al., 2012; MacGregor et al., 2015; Yurchenko et al., 2020). These studies often conflate various types of regular polysemy into a single category (e.g., metonymic or metaphorical polysemy), while overlooking the nuances in the direction of meaning extension across different types, among many other facets of meaning representation that our computational work has highlighted. Linking our computational findings with a coordinated neuroimaging research program could therefore serve as an excellent platform for evaluating the novel predictions arising from our work, and for guiding that experimental research agenda. As but a few examples of what may be possible on this front: Could the featural (voxel) overlap from fMRI measures of the two senses of regular polysemes from different types (dis)confirm our computational finding that degree of regularity is not reducible to amount of featural overlap? Could the voxel patterns of different types of regular polysemy indicate some high-order latent structures shared across these different types, confirming our finding based on computational methodologies?

There are also other potential avenues for future improvements that build upon the work we have reported here and connect this work with other important advances within the computational modeling literature. First, multimodal models trained on both text and images might provide richer and more comprehensive semantic representations (e.g., Betker et al., 2023; Ramesh et al., 2021; Xu et al., 2015), particularly for capturing concrete features that relate to vision, which forms a key part of our semantic knowledge (Cree & McRae, 2003). Second, recent advancements in LLMs (OpenAI, 2023; Touvron et al., 2023), trained on trillions of tokens, have significantly improved performance across diverse benchmarks. Future work incorporating open-sourced LLMs (e.g., Touvron et al., 2023) could enable a more sensitive assay of the effects of regular polysemy and of more nuanced differences between types than is possible with the BERT model. Indeed, it may be the case that replicating our analyses using the representations derived from one of these larger models would substantially boost accuracy, and this, in turn, could modulate other aspects of our results. For example, although we do not expect that degree of regularity could be fully distilled down to amount of featural overlap between senses across types, a larger model, perhaps coupled with an analysis of a larger set of regularity types, might reveal a statistically significant (although far from perfect) relationship between featural overlap and degree of regularity. However, for reasons outlined earlier in our paper, the challenges associated with training these models at scale pose important limitations on how these models can be probed to understand exactly how they operate, even setting aside other issues such as the implausibly extensive experience with text that the models experience during training. Thus, work on this front will have to deal with additional considerations that are less of an issue with our work. Third, as described in the methods section, although we consider it unlikely that the use of a word tokenizer to process words not in the BERT base vocabulary has substantially shaped our results, an explicit investigation of the role of tokenization in shaping the representation of polysemy would be useful in confirming this prediction. It may also reveal ways in which tokenization of unfamiliar words could offer complementary insight into how new or rarely encountered word meanings can be inferred based on the representation of the meaning of their constituent subwords, and how this type of inference may relate to regular polysemy (e.g., in terms of how new regularity types emerge, or how a new word may have its senses extended).

Taken together, despite the imperfect alignment between BERT (including other language models based at least in part on co-occurrence patterns in natural text) and human cognition, we still view this work as tapping into important aspects of the human representation of polysemy. These models are clearly developing sensitivity to facets of meaning representation based on word co-occurrence statistics much as humans appear to do, even if the detailed processes used to do so by the models and by humans are non-identical. Going forward, we expect that these models can be useful in guiding thinking and in developing targeted, testable predictions for empirical evaluation, which is a key reason to develop computational models, even if they are “wrong” in some respects.

7.3. *Splitting senses*

A splitting theory of polysemy (Katz & Fodor, 1963; Pustejovsky, 1998) suggests that the meanings of a polyseme can be split into separate sense categories. According to this theory, a key task for the language system is therefore to uncover these categories and outline their relationships. Our work also leverages this splitting assumption based on how it leverages the LR component of the LRcos model to answer sense analogy questions. However, this theory may not be accurate due to the challenges in separating senses, the possibility of a word carrying multiple senses simultaneously, and recent computational and psycholinguistic evidence (Li & Joanisse, 2021; Trott & Bergen, 2023) supporting the graded and continuous representation of specific polysemous senses. For these reasons, we expect that the more nuanced views advocated by this recent work will underlie a full account of regular polysemy.

Why then did we develop our work based on sense splitting? In our view, this was simply a pragmatic methodological decision, and not one intended to reflect a strong theoretical claim about the organization of senses. What was critical, in our view, was the notion that there is a consistent transformation between the two senses of polysemes from a given type. This transformation can, in principle, be estimated using either discrete splits or graded distributions of senses. Employing a method based on discretization greatly simplifies the approach, much as it does in other areas of semantic ambiguity research (e.g., delineating relatedness of meaning into the categories of unrelated homonyms and related polysemes, despite the fact that some polysemes may have more related meanings than others, as reflected by a relatedness continuum; Armstrong & Plaut, 2016). Having validated the basic premises of the approach, we expect that more complex and sophisticated methods may be leveraged to better capture the nuances between a polysemes meanings, as reflected in the aforementioned work.

7.4. Interrater reliability

In our study, we supplemented a prior source of sense annotated sentences with additional examples of usages of regular polysemes in lines of dialog extracted from movies and television subtitles. Generally speaking, these lines of dialog typically correspond to either one or a small number of short sentences. That is, they typically provide more context than just one or two adjacent words around the target word, but provide far less context than would be available in a full paragraph or more of text. In our annotations, the mean percentage of interrater agreement was 85% and mean Cohen's kappa value was 0.70. These results clearly indicate agreement levels well above chance, but also far from perfect agreement. To confirm that this was not an anomalous finding related to our specific norming methods or set of raters, we examined the interrater agreement levels reported in prior work on regular polysemy and found the levels of agreement to be quite consistent across studies (Alonso et al., 2013; Markert & Nissim, 2002; Navarro et al., 2005; Véronis, 1998). Adding another relevant comparison point, Rice et al. (2019) computed an interrater reliability measure for annotating homonyms, as opposed to polysemes using the exact same subtitles dataset used in our study. They observed slightly higher levels of overall agreement (92%). Taken together, these results indicate that raters have greater difficulty in agreeing on an annotation for the more closely related senses of polysemes than they do for homonyms for the amount of context we included in our norms.

This is perhaps not unsurprising for the following reasons. First, previous work has suggested that the different senses of a polyseme are closely related within a neural network's internal representational space, with somewhat fuzzy boundaries between some senses (Li & Joanisse, 2021). Given the long history of using similar neural networks to make inferences about human cognition, this work suggests that humans should also be challenged in providing appropriate annotations for polysemes in many contexts. Second, copredication, even if it happens infrequently, could further erode model performance, although our relatively high overall levels of accuracy and our informal inspection of our annotated data suggest that the effects of this factor are likely small. Finally, and perhaps most importantly, our experience in annotating data might suggest that the context we provided (on average, about 13 words) was insufficient to make confident annotations between the related senses of a polyseme in many cases. Increasing the context window in future annotations may therefore be worthwhile even if it does increase the resource intensiveness of the task.

8. 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, how similar latent pressures may drive the formation of several types of regular polysemy, whether semantic overlap fully explains regularity, and how the LRcos methodology detects the regularity in sense analogies. Doing so in computationally explicit terms has offered a platform for future extensions of this work, such as the unsupervised discovery of other types of regular polysemy, and has also offered targeted directions for future research, such as the expanded study of the regularity continuum using a broader range of regularity types and interdisciplinary methods. 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 can also inform 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 the members of the Language, Computation, and Cognition group at the University of Toronto for their feedback on this project.

Open Research Badges

This article has earned Open Data and Open Materials badges. Data and materials are available at https://osf.io/bctp4/?view_only=b981b9a66b5046459a0201e14f0a5d26.

Notes

[1] We thank the helpful comment of an anonymous reviewer for highlighting the importance of discussing the BERT tokenizer in our work.

[2] For example, the Animal/Meat polysemes that were excluded included low-frequency words such as *yak, prawn*, and *quail*. These items also had very few—if any—examples of one of their two senses in our corpora.

[3] For practical reasons, it was not possible to have every rater to provide equal numbers of annotations for each polyseme/polyseme type, hence the weighted average approach.

[4] To rule out the possibility that any single one of our types was an outlier and a significant relationship was not observed as a result, we conducted an additional sanity check.

We first left out the data from one of our types and reran the main analysis on the four remaining types. We then imputed the expected accuracy for the withheld type and reran the regression including the four "real" data points and the imputed datapoint. By definition, such imputation must produce either the same or a higher correlation to that observed with the four real data points. However, when we ran this analysis, we always failed to observe a significant correlation. For example, we failed to detect a significant correlation when we withheld the real data from the Process/Result type and imputed the data for that type and obtained a correlation of $r(3) = -.289$, $p = .64$. This category shared a large number of words with the same morphological structure (words ending in "-ion;" although as described in the methods section, our polysemes are not subject to tokenization into subwords), and was also associated with the lowest overall accuracy. This supplemental analysis confirms that in this aspect, the Process/Result type (or indeed, any idiosyncratic property associated with any one type) is not responsible for the non-significant relationship between sense overlap and meaning regularity.

References

Alonso, H., Pedersen, B., & Bel, N. (2013). Annotation of regular polysemy and underspecification. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Vol. 2, pp. 725–730).

Apresjan, J. U. D. (1974). Regular polysemy. *Linguistics*, *12*(142), 5–32. https://doi.org/10.1515/ling.1974.12.142.5

Armstrong, B. C., & Plaut, D. C. (2016). Disparate semantic ambiguity effects from semantic processing dynamics rather than qualitative task differences. *Language, Cognition and Neuroscience*, *31*(7), 940–966. https://doi.org/10.1080/23273798.2016.1171366

Azuma, T., & Van Orden, G. C. (1997). Why SAFE is better than FAST: The relatedness of a word's meanings affects lexical decision times. *Journal of Memory and Language*, *36*(4), 484–504. https://doi.org/10.1006/jmla.1997.2502

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., … Ramesh, A. (2023). Improving image generation with better captions. https://cdn.openai.com/papers/dall-e-3.pdf

Block, N. (1987). Advertisement for a semantics for psychology. *Midwest Studies in Philosophy*, *10*(1), 615–678. https://doi.org/10.1111/j.1475-4975.1987.tb00558.x

Boleda, G., Padó, S., & Utt, J. (2012). Regular polysemy: A distributional model. In ** 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)* (pp. 151–160).

Brochhagen, T., Boleda, G., Gualdoni, E., & Xu, Y. (2023). From language development to language evolution: A unified view of human lexical creativity. *Science*, *381*(6656), 431–436. https://doi.org/10.1126/science.ade7981

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language models are few-shot learners*. arXiv: 2005.14165. http://arxiv.org/abs/2005.14165

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. https://doi.org/10.3758/BRM.41.4.977

Chersoni, E., Santus, E., Huang, C.-R., & Lenci, A. (2021). Decoding word embeddings with brain-based semantic features. *Computational Linguistics*, *47*(3), 663–698. https://doi.org/10.1162/coli_a_00412

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Copestake, A., & Briscoe, T. (1995). Semi-productive polysemy and sense extension. *Journal of Semantics*, *12*(1), 15–67. https://doi.org/10.1093/jos/12.1.15

Cree, G. S., & McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of Experimental Psychology: General*, *132*(2), 163–201. https://doi.org/10.1037/0096-3445.132.2.163

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge: Cambridge University Press.

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*, *55*(4), 1537–1557. https://doi.org/10.3758/s13428-022-01869-6

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv: http://arxiv.org/abs/1810.04805

Dölling, J. (2020). Systematic polysemy. In D. Gutzmann, L. Matthewson, C. Meier, H. Rullmann, & T. E. Zimmerman (Eds.), *The Wiley Blackwell companion to semantics* (pp. 1–27). Hoboken, NJ: Wiley-Blackwel. http://onlinelibrary.wiley.com/doi/abs/10.1002/9781118788516.sem099

Drozd, A., Gladkova, A., & Matsuoka, S. (2016). Word embeddings, analogies, and machine learning: Beyond king - man + woman = queen. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 3519–3530). https://www.aclweb.org/anthology/C16-1332

Ettinger, A., Feldman, N., Resnik, P., & Phillips, C. (2016). Modeling N400 amplitude using vector space models of word representation. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 1445–1450).

Evans, C., & Yuan, D. (2017). *A large corpus for supervised word-sense disambiguation*. Google Research Blog. http://ai.googleblog.com/2017/01/a-large-corpus-for-supervised-word.html

Firth, J. (1957). A synopsis of linguistic theory, 1930-1955. In *Studies in linguistic analysis* (pp. 1–31). Special Volume of the Philological Society. Oxford: Blackwell.

Fishbein, J., & Harris, J. (2014). Making sense of Kafka: Structural biases induce early sense commitment for metonyms. *Journal of Memory and Language*, *76*, 94–112. https://doi.org/10.1016/j.jml.2014.06.005

Floyd, S., Dalawella, K., Goldberg, A., Lew-Williams, C., & Griffiths, T. (2021). Modeling rules and similarity in colexification. In *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society* (pp. 1830–1836).

Forster, K. I. (2000). The potential for experimenter bias effects in word recognition experiments. *Memory & Cognition*, *28*(7), 1109–1115. https://doi.org/10.3758/BF03211812

Frazier, L., & Rayner, K. (1990). Taking on semantic commitments: Processing multiple meanings vs. multiple senses. *Journal of Memory and Language*, *29*(2), 181–200. https://doi.org/10.1016/0749-596X(90)90071-7

Frisson, S. (2015). About bound and scary books: The processing of book polysemies. *Lingua*, *157*, 17–35. https://doi.org/10.1016/j.lingua.2014.07.017

Frisson, S., & Frazier, L. (2005). Carving up word meaning: Portioning and grinding. *Journal of Memory and Language*, *53*(2), 277–291. https://doi.org/10.1016/j.jml.2005.03.004

Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E. (2022). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature Human Behaviour*, *6*(7), 975–987. https://doi.org/10.1038/s41562-022-01316-8

Harman, G. (1982). Conceptual role semantics. *Notre Dame Journal of Formal Logic*, *23*(2), 242–256.

Heilbron, M., van Haren, J., Hagoort, P., & de Lange, F. P. (2023). Lexical processing strongly affects reading times but not skipping during natural reading. *Open Mind*, *7*, 757–783. https://doi.org/10.1162/opmi_a_00099

Hoffman, P., Lambon Ralph, M., & Rogers, T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, *45*(3), 718–730. https://doi.org/10.3758/s13428-012-0278-x

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65–70.

Ide, N., & Macleod, C. (2001). The American National Corpus: A standardized resource of American English. *Proceedings of Corpus Linguistics*, *3*, 1–7.

Jastrzembski, J. E. (1981). Multiple meanings, number of related meanings, frequency of occurrence, and the lexicon. *Cognitive Psychology*, *13*(2), 278–305.

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? In *57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy*.

Karidi, T., Zhou, Y., Schneider, N., Abend, O., & Srikumar, V. (2021). Putting words in BERT's mouth: Navigating contextualized vector spaces with pseudowords. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 10300–10313). https://doi.org/10.18653/v1/2021.emnlp-main.806

Katz, J. J., & Fodor, J. A. (1963). The structure of a semantic theory. *Language*, *39*(2), 170–210.

Klepousniotou, E., Pike, G. B., Steinhauer, K., & Gracco, V. (2012). Not all ambiguous words are created equal: An EEG investigation of homonymy and polysemy. *Brain and Language*, *123*(1), 11–21. https://doi.org/10.1016/j.bandl.2012.06.007

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), 1534–1543.

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

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). *ALBERT: A lite BERT for self-supervised learning of language representations*. arXiv: 1909.11942 [cs]. http://arxiv.org/abs/1909.11942

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), 211–240.

Lewis, D. (1970). General semantics. *Synthese*, *22*(1), 18–67. http://www.springerlink.com/index/V084851U826907J7.pdf

Li, J., & Joanisse, M. F. (2021). Word senses as clusters of meaning modulations: A computational model of polysemy. *Cognitive Science*, *45*(4), e12955. https://doi.org/10.1111/cogs.12955

Linzen, T. (2016). Issues in evaluating semantic spaces using word analogies. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP* (pp. 13–18).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach*. arXiv: 1907.11692 [cs]. http://arxiv.org/abs/1907.11692

Lopukhina, A., & Lopukhin, K. (2016). Regular polysemy: From sense vectors to sense patterns. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex - V)* (pp. 19–23). https://www.aclweb.org/anthology/W16-5303

MacGregor, L. J., Bouwsema, J., & Klepousniotou, E. (2015). Sustained meaning activation for polysemous but not homonymous words: Evidence from EEG. *Neuropsychologia*, *68*, 126–138. https://doi.org/10.1016/j.neuropsychologia.2015.01.008

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78.

Markert, K., & Nissim, M. (2002). Towards a corpus annotated for metonymies: The case of location names. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*.

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

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

Navarro, B., Marcos, R., & Abad, P. (2005). Semantic annotation and inter-annotation agreement in Cast3LB corpus. In *Proceedings of the 4th Workshop on Treebanks and Linguistic Theories* (pp. 125–135).

Nunberg, G. (1995). Transfers of meaning. *Journal of Semantics*, *12*(2), 109–132. https://doi.org/10.1093/jos/12.2.109

OpenAI. (2023). *GPT-4 Technical Report*. arXiv: 2303.08774 [cs]. https://doi.org/10.48550/arXiv.2303.08774

Passonneau, R. J., Baker, C., Fellbaum, C., & Ide, N. (2012). The MASC word sense sentence corpus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*.

Piantadosi, S. T., & Hill, F. (2022). Meaning without reference in large language models. In *NeurIPS 2022 Workshop on Neuro Causal and Symbolic AI (nCSI)*.

Pimentel, T., Meister, C., Wilcox, E. G., Levy, R., & Cotterell, R. (2023). *On the effect of anticipation on reading times*. arXiv: 2211.14301 [cs]. https://doi.org/10.48550/arXiv.2211.14301

Pustejovsky, J. (1998). *The generative lexicon*. Cambridge, MA: MIT Press.

Pustejovsky, J. (2005). *A survey of dot objects*. Unpublished manuscript, Brandeis University, Waltham.

Putnam, H. (1975). The meaning of "meaning". In K. Gunderson (Ed.), *Minnesota studies in the philosophy of science* (Vol. 7, pp. 131–193). Minneapolis: University of Minnesota Press.

Rabagliati, H., Marcus, G. F., & Pylkkänen, L. (2011). Rules, radical pragmatics and restrictions on regular polysemy. *Journal of Semantics*, *28*(4), 485–512. https://doi.org/10.1093/jos/ffr005

Rabagliati, H., Pylkkänen, L., & Marcus, G. (2013). Top-down influence in young children's linguistic ambiguity resolution. *Developmental Psychology*, *49*(6), 1076–1089. https://doi.org/10.1037/a0026918

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). *Zero-shot text-to-image generation*. http://arxiv.org/abs/2102.12092

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence embeddings using siamese BERT-networks*. arXiv: 1908.10084 [cs]. https://doi.org/10.48550/arXiv.1908.10084

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. https://doi.org/10.3758/s13428-018-1107-7

Rodd, J. (2020). Settling into semantic space: An ambiguity-focused account of word-meaning access. *Perspectives on Psychological Science*, *15*(2), 411–427. https://doi.org/10.1177/1745691619885860

Rodd, J., Davis, M., & Johnsrude, I. (2005). The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity. *Cerebral Cortex*, *15*(8), 1261–1269. https://doi.org/10.1093/cercor/bhi009

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. https://doi.org/10.1006/jmla.2001.2810

Rogers, A., Drozd, A., & Li, B. (2017). The (too many) problems of analogical reasoning with word vectors. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)* (pp. 135–148). https://doi.org/10.18653/v1/S17-1017

Rogers, A., Kovaleva, O., & Rumshisky, A. (2021). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, *8*, 842–866. https://doi.org/10.1162/tacl_a_00349

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter*. arXiv: [cs]. http://arxiv.org/abs/1910.01108

Sassenhagen, J., & Fiebach, C. J. (2020). Traces of meaning itself: Encoding distributional word vectors in brain activity. *Neurobiology of Language*, *1*(1), 54–76. https://doi.org/10.1162/nol_a_00003

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45). https://doi.org/10.1073/pnas.2105646118

Søgaard, A. (2016). Evaluating word embeddings with fMRI and eye-tracking. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP* (pp. 116–121). https://aclanthology.org/W16-2521.pdf

Spearman, C. (1904). General ability, objectively determined and measured. *American Journal of Psychology*, *15*, 201–293.

Srinivasan, M., Al-Mughairy, S., Foushee, R., & Barner, D. (2017). Learning language from within: Children use semantic generalizations to infer word meanings. *Cognition*, *159*, 11–24. https://doi.org/10.1016/j.cognition.2016.10.019

Srinivasan, M., Berner, C., & Rabagliati, H. (2019). Children use polysemy to structure new word meanings. *Journal of Experimental Psychology: General*, *148*(5), 926–942. https://doi.org/10.1037/xge0000454

Srinivasan, M., & Rabagliati, H. (2015). How concepts and conventions structure the lexicon: Cross-linguistic evidence from polysemy. *Lingua*, *157*, 124–152. https://doi.org/10.1016/j.lingua.2014.12.004

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. https://doi.org/10.1016/j.cogpsych.2011.03.002

Srinivasan, M., & Snedeker, J. (2014). Polysemy and the taxonomic constraint: Children's representation of words that label Multiple Kinds. *Language Learning and Development*, *10*(2), 97–128. https://doi.org/10.1080/15475441.2013.820121

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs*, *1*, ix + 121.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (2023). *Llama 2: Open foundation and fine-tuned chat models*. arXiv: 2307.09288 [cs]. https://doi.org/10.48550/arXiv.2307.09288

Trott, S., & Bergen, B. (2023). Word meaning is both categorical and continuous. *Psychological Review*, *130*(5), 1239–1261. https://doi.org/10.1037/rev0000420

Turney, P. D., Littman, M. L., Bigham, J., & Shnayder, V. (2003). Combining independent modules to solve multiple-choice synonym and analogy problems. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing, 2003* (pp. 482–489).

Vе́ronis, J. (1998). A study of polysemy judgements and inter-annotator agreement. In *Programme and Advanced Papers of the Senseval Workshop* (pp. 2–4).

Vitello, S., Warren, J., Devlin, J., & Rodd, J. (2014). Roles of frontal and temporal regions in reinterpreting semantically ambiguous sentences. *Frontiers in Human Neuroscience*, *8*, 530. https://doi.org/10.3389/fnhum.2014.00530

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., … Dean, J. (2016). *Google's neural machine translation system: Bridging the gap between human and machine translation*. arXiv: 1609.08144 [cs]. http://arxiv.org/abs/1609.08144

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning* (pp. 2048–2057).

Xu, Y., Malt, B., & Srinivasan, M. (2017). Evolution of word meanings through metaphorical mapping: Systematicity over the past millennium. *Cognitive Psychology*, *96*, 41–53. https://doi.org/10.1016/j.cogpsych.2017.05.005

Yu, L., & Xu, Y. (2023). Word sense extension. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 3281–3294). https://aclanthology.org/2023.acl-long.184

Yurchenko, A., Lopukhina, A., & Dragoy, O. (2020). Metaphor is between metonymy and homonymy: Evidence from event-related potentials. *Frontiers in Psychology*, *11*, 2113. https://doi.org/10.3389/fpsyg.2020.02113

Zhu, R. (2021). Preschoolers' acquisition of producer-product metonymy. *Cognitive Development*, *59*, 101075. https://doi.org/10.1016/j.cogdev.2021.101075

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supporting Information

Appendix: Cross-type specific analysis tables

Table A1
Summary of the mean accuracies and statistical comparisons from the pairwise cross-type comparisons

		versus RP	
	Accuracy	χ^2	*p*
A/M C/C	0.22	152.40	**<.001**
A/M L/O	0.20	107.22	**<.001**
A/M A/I	0.22	137.98	**<.001**
A/M P/R	0.14	7.18	.01
C/C L/O	0.25	68.62	**<.001**
C/C A/I	0.34	207.86	**<.001**
C/C P/R	0.21	44.25	**<.001**
L/O A/I	0.30	110.45	**<.001**
L/O P/R	0.35	236.98	**<.001**
A/I P/R	0.15	4.90	.03

Note. A/M = Animal/Meat, C/C = Container/Content, L/O = Location/Organization, A/I = Artifact/Information, P/R = Process/Result. Significant *p*-values after the Holm correction are presented in boldface.

Table A2
Summary of the statistical tests comparing the accuracy for each cross-type comparison against the accuracy for every other cross-type comparison

group1	group2	$\chi^2(1)$	p
A/M C/C	A/M L/O	0.83	.36
A/M C/C	A/M A/I	0.01	.91
A/M C/C	A/M P/R	33.77	**<.001**
A/M C/C	C/C L/O	1.59	.21
A/M C/C	C/C A/I	28.59	**<.001**
A/M C/C	C/C P/R	0.12	.72
A/M C/C	L/O A/I	10.63	**<.001**
A/M C/C	L/O P/R	32.40	**<.001**
A/M C/C	A/I P/R	9.75	**<.001**
A/M L/O	A/M A/I	0.55	.46
A/M L/O	A/M P/R	22.94	**<.001**
A/M L/O	C/C L/O	3.60	.06
A/M L/O	C/C A/I	35.90	**<.001**
A/M L/O	C/C P/R	0.05	.82
A/M L/O	L/O A/I	14.92	**<.001**
A/M L/O	L/O P/R	40.37	**<.001**
A/M L/O	A/I P/R	6.23	.01

(*Continued*)

Table A2
(Continued)

group1	group2	$\chi^2(1)$	p
A/M A/I	A/M P/R	31.20	**<.001**
A/M A/I	C/C L/O	1.85	.17
A/M A/I	C/C A/I	29.40	**<.001**
A/M A/I	C/C P/R	0.05	.82
A/M A/I	L/O A/I	11.19	**<.001**
A/M A/I	L/O P/R	33.25	**<.001**
A/M A/I	A/I P/R	9.03	**<.001**
A/M P/R	C/C L/O	28.68	**<.001**
A/M P/R	C/C A/I	94.51	**<.001**
A/M P/R	C/C P/R	14.91	**<.001**
A/M P/R	L/O A/I	52.52	**<.001**
A/M P/R	L/O P/R	104.02	**<.001**
A/M P/R	A/I P/R	0.32	.57
C/C L/O	C/C A/I	9.10	**<.001**
C/C L/O	C/C P/R	1.82	.18
C/C L/O	L/O A/I	2.49	.12
C/C L/O	L/O P/R	10.18	**<.001**
C/C L/O	A/I P/R	12.77	**<.001**
C/C A/I	C/C P/R	21.71	**<.001**
C/C A/I	L/O A/I	1.44	.23
C/C A/I	L/O P/R	0.00	.97
C/C A/I	A/I P/R	44.79	**<.001**
C/C P/R	L/O A/I	9.21	**<.001**
C/C P/R	L/O P/R	23.88	**<.001**
C/C P/R	A/I P/R	5.50	.02
L/O A/I	L/O P/R	1.76	.18
L/O A/I	A/I P/R	25.97	**<.001**
L/O P/R	A/I P/R	48.02	**<.001**

Note. Significant *p*-values after the Holm correction are presented in boldface.