# Assessing Polyseme Sense Similarity through Co-predication Acceptability and Contextualised Embedding Distance

**Janosch Haber** and **Massimo Poesio**
Queen Mary University of London
{j.haber|m.poesio}@qmul.ac.uk

## Abstract

Co-predication is one of the most frequently used linguistic tests to tell apart shifts in polysemic sense from changes in homonymic meaning. It is increasingly coming under criticism as evidence is accumulating that it tends to mis-classify specific cases of polysemic sense alteration as homonymy. In this paper, we collect empirical data to investigate these accusations. We asses how co-predication acceptability relates to explicit ratings of polyseme word sense similarity, and how well either measure can be predicted through the distance between target words' contextualised word embeddings. We find that sense similarity appears to be a major contributor in determining co-predication acceptability, but that co-predication judgements tend to rate less similar sense interpretations as being as unacceptable as homonym pairs, effectively misclassifying these instances. The tested contextualised word embeddings fail to predict word sense similarity consistently, but the similarities between BERT embeddings show a significant correlation with co-predication ratings. We take this finding as evidence that BERT embeddings might be better representations of context than encodings of word meaning.

## 1 Introduction

Polysemy is a form of lexical ambiguity which occupies a unique middle ground between *monosemy* –word forms with exactly one interpretation– and *homonymy* –word forms associated with two or more completely unrelated interpretations. Unlike monosemes, polysemes can evoke different interpretations, but unlike homonyms, polysemic sense interpretations are thought to be closely related to each other (Lyons, 1977). It is commonly assumed that most words in natural language are in fact polysemous to some degree (Falkum and Vicente, 2015), and the question whether there in fact are any proper monosemes has been the source of ongoing debate (see for example Jackendoff, 1989; Fodor, 1998). Homonyms have been a driving factor in developing contextualised language models (e.g. Peters et al., 2018; Devlin et al., 2018; Radford et al., 2019) in order to account for the different unrelated meanings some words can evoke in different contexts:

a. The match burned my fingers.
b. The match ended without a winner.

Comparing these uses of the word *match* to the various closely related interpretations of canonical polyseme *school* illuminates the conceptual difference between the two phenomena of lexical ambiguity:[1]

a. The school [building] is on fire.
b. The school [rules] has prohibited wearing hats in the classroom.
c. I have talked to the school [director, staff] about it already.
d. The school [participants] went for a visit to the cathedral.

Although the distinction is clear in theory, distinguishing monosemy, polysemy and homonymy in practice proves exceedingly difficult: At what point are interpretation nuances pronounced enough to speak of two different word senses? Is the coercion of word sense a manifestation of polysemic sense alteration or a context effect on a monosemic word form? Do word senses related through metaphor qualify as polysemes or are their interpretations a form or homonymic ambiguity? Traditionally, co-predication tests are used to provide a linguistic means to answer these questions and attempt a



[1] Examples taken from Ortega-Andrés and Vicente (2019)



*Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 114–124
Barcelona, Spain (Online), December 12–13, 2020

classification of word sense interpretations into one of the three categories. In co-predication tests, two interpretations of a word form are simultaneously invoked by the context. If this renders a felicitous construction (see Example 1), the two interpretations are considered to evoke the same sense or meaning of the word; if the reading is infelicitous (Example 2) they are considered to be derived from two different word meanings.

(1) The newspaper wasn't very interesting, so she folded it and put it away. [content/object]

(2) # The match burned my fingers but ended without a winner.

Based on a range of experiments finding that homonyms seem to be processed differently than polysemes (Frazier and Rayner, 1990; Rodd et al., 2002; Klepousniotou et al., 2008, 2012), the prevailing understanding of co-predication is that it is rendered felicitous if the different sense interpretations are activated simultaneously and can be shifted between without additional processing costs. Co-predication is thought to lead to infelicitous sentences if the different activations are not activated automatically, and cognitive effort is involved in updating the assumed meaning of a word. These hypotheses informed a number of linguistic models to define different mental representations of homonymic meaning and polysemic sense, respectively. The Generative Lexicon (Pustejovsky, 1991; Asher and Pustejovsky, 2006; Asher, 2011) for example postulates individual lexicon entries for different interpretations of a homonym, while all sense interpretations of a polyseme are represented by a single under-specified entry and therefore do not require any processing cost for sense switching. Recently, a growing body of work however came to challenge a unified, under-specified representation of polysemic sense (see Klepousniotou, 2002; Pylkkänen et al., 2006; Frisson, 2015). Klepousniotou et al. (2012) for example report that their experiments indicate that the processing of irregular polysemes resembles homonymic meaning alterations more than the sense alterations in regular polysemes, while an ongoing series of co-predication studies (e.g. Antunes and Chaves, 2003; Traxler et al., 2005; Schumacher, 2013; Filip and Sutton, 2017; Zobel, 2017; Sutton and Filip, 2018) show that not all polysemic senses can be co-predicated either, and that the co-predication of some polysemic interpretations can lead to infelicitous and zeugmatic expressions:[2]

a. # The newspaper fired its editor in chief and got wet from the rain. [publisher/publication]

b. # They took the door off its hinges and walked through it. [object/opening]

A recent model of polyseme sense clustering proposed by Ortega-Andrés and Vicente (2019) tries to explain why certain polyseme senses lead to infelicitous co-predication by suggesting that polyseme senses might be grouped based on their similarity. According to their grouping, closely related senses are thought to form co-activation packages that remain active for a while, allowing for cost-free sense shifting and therefore felicitous co-predication. Distantly related sense interpretations on the other hand would not co-activate and therefore require cognitive effort to be changed, much like homonymic meaning alterations.

The difficulty in assessing this hypothesis is the unavailability of a ready reference of contextualised word sense similarity for polysemes. To mitigate this, we collected human annotated data on a number of different measures of polysemic sense similarity to empirically investigate the correlation between sense similarity ratings and co-predication acceptability judgements. Specifically, we use crowdsourcing to collect i) graded co-predication acceptability judgements, ii) explicit (meta-linguistic) word sense similarity judgements, iii) word class similarity ratings, and iv) determine the similarity in a target word's contextualised embeddings derived from different models. If word sense similarity indeed governs co-activation and therefore co-predication acceptability, we expect similarity judgements to be a strong predictor for acceptability judgements. Conversely, if co-predication acceptability is a representative test of the mental processing of lexically ambiguous items, we expect acceptability judgements to be a strong predictor of similarity judgements and reliably tell apart homonyms from polysemes.

We find that sense similarity appears to be a major contributor in determining co-predication acceptability, but that co-predication judgements tend to rate less similar sense interpretations equally as unacceptable as homonym pairs, effectively misclassifying these instances. We therefore argue

[2]Examples from Cruse (1995)

that these findings provide both, a) support for a more hierarchical representation of polysemic sense based on sense similarity, and, b) an additional, empirically founded argument against co-predication as a prevailing test for distinguishing polysemy and homonymy. Finally, the tested contextualised word embeddings fail to predict word sense similarity consistently, but the similarities between BERT embeddings show a significant correlation with co-predication ratings.

## 2 Method

In order to evaluate both, i) the hypothesis that polysemic senses might form groupings based on their similarity, and ii) the prevalence of co-predication as a linguistic test for the distinction between homonymy and polysemy, we collect three human annotated measures of word sense similarity together with five word sense similarity proxies derived from computational methods. We investigate how well these different metrics distinguish homonyms from polysemes, and to what degree they can predict one another. In order to achieve a fair comparison of the different measures, we defined a fixed set of target words, sense interpretations and contexts to be used in all experiments.

### 2.1 Samples

Since at least Apresjan (1974), polysemes are generally considered to be either regular or irregular, depending on whether or not their sense patterns are shared with other word forms. Irregular polysemes often demonstrate a metaphorical connection between the different interpretations of their senses that does not carry over for other uses (see Example 3), regular, or systematic polysemes on the other hand exhibit the same interpretation patterns across a number of word forms (Example 4). See Moldovan (2019) for a recent in-depth discussion of this distinction.

(3) **[cold]** I got a cold/#hot/#tired after getting caught in the rain last week.
The librarian gave me a cold/#hot/#tired stare when my phone rang.

(4) **[liquid-for-container]** He took a sip and put his beer/coffee/juice/gin/soup... back on the kitchen table.

With growing evidence that irregular polysemes might be processed differently than their regular counterparts (e.g. Klepousniotou et al., 2012), we decided to focus on regular polysemic nouns for this study. Regular polysemes can be more clearly distinguished from homonyms, maximising the impact of our findings if metrics fail to classify them correctly. With their canonical division of sense interpretations, they also allow for a clear separation of different sense interpretations, making it easier to generate contexts that unequivocally evoke the different senses. We selected ten of the systematic polysemy types compiled in Dölling (Forthcoming), with target expressions having between two and four clearly distinct but related senses, and picked one of the most frequently used expressions representing each class from his compilation.

To create sample contexts invoking the different interpretations, we followed a custom template designed to guarantee that samples could be used individually to collect graded word sense judgements, class ratings and context embedding similarity, but could also be combined into a co-predication structure without invalidating acceptability due to repetitions or temporal or logical mis-matches. Following this template, samples were created such that i) the ambiguous target expression is the subject of the sentence, ii) the context is kept as short as possible, and iii) the context invokes a certain sense as clearly as possible without mentioning that sense explicitly.[3] Besides creating clear sample sentences for our human participants, these guidelines also minimise the impact of syntactic features and compounding context effects for contextualised models, which are shown to significantly impact embbedings (see e.g. Wiedemann et al., 2019) and cloud the accessibility of meaning representations.

Two sample contexts were created for every sense interpretation of the ten polysemes, resulting in a total of 54 sentences. As an example, consider the six sample sentences of polyseme *newspaper*, generated for its three senses (1) *organisation/institution*, (2) *physical object* and (3) *information/data*:

1a The newspaper fired its editor in chief.,
1b The newspaper was sued for defamation.
2a The newspaper lies on the kitchen table.,
2b The newspaper got wet from the rain.
3a The newspaper wasn't very interesting.,
3b The newspaper is rather satirical today.

Besides the polyseme samples, we created an ad-

[3] As in "The school is an old building." for sense *building*. See Haber and Poesio (2020) for more details.

ditional two samples sets. The first set is made up of 15 common homonyms, with two sentences invoking their two most dominant senses each. While our focus is on polysemes, comparing ratings for the homonym samples to ratings assigned to polyseme pairs, we will be able to test the different similarity measures' performance in predicting whether an ambiguous target pair is polysemic or homonymic. The second set contains 15 pairs of synonyms meant to be used as quality control and to calibrate the rating scale. All sample sentences were rated to be acceptable by annotators recruited from Amazon Mechanical Turk (AMT)[4] in a validation experiment.

### 2.2 Graded Co-predication Acceptability

Traditionally, co-predication acceptability is one of the most frequently used linguistic tests for distinguishing homonyms from polysemes. Acceptability usually is determined through introspection, classifying a sentence invoking two different interpretations of the same word form as either acceptable or not. When assessed through annotator judgements, co-predication acceptability however appears to be a graded measure (Lau et al., 2014). We therefore decided to collect empirical data on graded annotator judgements, asking participants to rate the acceptability of co-predication structures combining different pairings of target word samples through conjunction reduction (Zwicky and Sadock, 1975). As an example, the previously shown *newspaper* contexts 1a and 1b where combined into co-predication sample 1ab for data collection:

1ab The newspaper fired its editor in chief and was sued for defamation.

Co-predication samples were generated for all combinations of sense interpretations, resulting in four samples for polysemes with two senses, nine for polysemes with three senses, and 16 for those with four, and a grand total of 75. We manually inspected the co-predication structures for any inconsistencies that might have emerged through the conjunction, and corrected issues with the least invasive measures possible. The samples were then distributed over 15 questionnaires so that no target expression appeared twice in any questionnaire. We added one of the homonym and synonym validation samples to each questionnaire, and filled all questionnaires to a total of ten items with co-predication structures generated from random sentence pairs to obfuscate the focus on polysemes. Item order was then randomised per questionnaire.

We used AMT to collect graded co-predication acceptability judgements by asking workers to rate a given sentence using a slider labelled with "The sentence is absolutely unacceptable" on the left hand side and "The sentence is absolutely acceptable" on the right. The submitted slider positions were translated to a 100-point acceptability score ranging between 0 and 1, and stored in combination with a worker's unique ID. To improve judgement quality, we required workers to have obtained a US high school degree and reached the "AMT Master" qualification.[5] Workers were paid 0.35 USD for every completed questionnaire.

We collected between 20 and 40 judgements for each item. A total of 43 individual workers contributed to the study, with HITs taking an average of 146 seconds (median of 93). Through filtering out any submissions that rated at least two filler samples higher than 0.66 or the synonym sample lower than 0.33,[6] we excluded a total of 44 judgements. The resulting dataset features an average of 28 judgements per item.

### 2.3 Graded Word Sense Similarity

As a first measure of sense similarity, we collected graded annotator judgements explicitly rating the similarity of word sense interpretations as invoked by different pairings of sample sentences. In contrast to co-predication judgements, these pairwise similarity ratings are less influenced by factors like sentence order and compound consistency, but do provide a meta-linguistic signal rather than the more ecological acceptability rating derived from co-predication. Still, if word sense similarity is the driving factor in determining the mental representation of polysemic sense, we should find a strong correlation between these judgements and the previously measured co-predication judgements.

We collected word sense similarity judgements using our custom polyseme sample set, this time combining samples into sentence pairs invoking

[4] https://www.mturk.com/

[5] According to AMT's website, "[T]hese Workers have consistently demonstrated a high degree of success in performing a wide range of HITs across a large number of Requesters," https://www.mturk.com/worker/help

[6] Note that in co-predication the synonymity effect is lost as only one subject noun phrase remains in the conjunction.

different combinations of sense interpretations instead of joining them into a single co-predication structure. The same method as in the first experiment was used for distributing test items over questionnaires, with the distinction that now homonym, synonym and filler samples were presented as sentence pairs rather than co-predication structures as well. We highlighted target expressions in bold font and asked workers to rate the highlighted expressions using a slider labelled with "The highlighted words have a completely different meaning" on the left hand side and "The highlighted words have completely the same meaning" on the right. Qualification requirements and payment remained identical.

We collected 20 judgements for each questionnaire. 65 individual workers in total contributed to the study, with HITs taking an average of 133 seconds (median = 90). Applying the same filtering as with the co-predication samples, we removed 9 submissions and retained at least 18 judgements per item.

### 2.4 Word Sense Class Ratings

As a second judgement of word sense similarity, we collected categorical sense class labels. If the determining factor in whether or not word senses can be co-predicated is not specifically their distance, but whether or not both interpretations refer to the same type or class of object, the agreement in assigned sense class should be a good predictor of co-predication acceptability - and valid proxy of word sense similarity.

To collect sense class labels, AMT Workers were presented with individual sample sentences together with a list of 16 sense class labels. Class labels were derived from the descriptions of the ten polyseme's different interpretations as used in Dölling (Forthcoming) and included an "other" category label. We used the same set of polyseme samples as before, with target expression highlighted like in the second experiment. Designed to validate the other two annotation metrics, we did not include any homonym, synonym or filler items in this experiment. Workers were asked to classify the highlighted target expression by selecting all applicable labels. Submissions were stored in 16-dimensional multi-hot vectors indicating the selection of labels together with the worker's ID. We kept the same worker qualification requirements and payment regime as before and collected 15 labels for each item, incidentally provided by exactly 15 individual workers, i.e. each individual worker completed all 15 questionnaires. HIT's took an average of 178 seconds (median of 107). Classification results were not filtered, but averaged per item in order to create word sense class vectors. Pairwise sense class similarity was then calculated through the cosine between the different combinations of sense interpretations, i.e. the overlap in their averaged multi-class assignments.

The resulting dataset containing all three types of human annotations is publicly available.[7]

### 2.5 Word Embedding Similarity

Because the three previously described measures of word sense similarity are based on costly human-annotated labels, we were also interested in investigating how well sense similarity estimates derived from computational models would correlate with these metrics. Models of polysemy have previously been proposed in distributional semantics (see for example Boleda et al., 2012), but for the most part, such models found limited application in computational linguistics. With the recent development of context-sensitive models of word embeddings such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018), the field however obtained a new tool to capture polysemic sense alterations, leading to a demonstrated improvement in various NLP systems. While ELMo was developed explicitly to capture a target word's context, BERT is a language model based on the encoder architecture of the Transformer model (Vaswani et al., 2017), an attention mechanism for learning the contextual relations between words. While BERT's output is usually fed to a downstream model, our aim is to see whether it is able to capture differences in word sense by using its outputs directly.

To obtain ELMo embeddings we used a pretrained model available on TensorFlow Hub[8] and extracted target word vectors from the LSTM's second layer hidden state, which has previously been shown to encode more semantic information than the character-level first layer or the LSTM's first layer (Ethayarajh, 2019; Haber and Poesio, 2020). For the investigation of BERT's embeddings we used the output of a pretrained cased model as provided by Huggingface[9] with 12 layers, a hidden

[7] https://github.com/dali-ambiguity/Word-Sense-Dataset-v1

[8] https://tfhub.dev/google/ELMo/3

[9] https://huggingface.co/transformers/

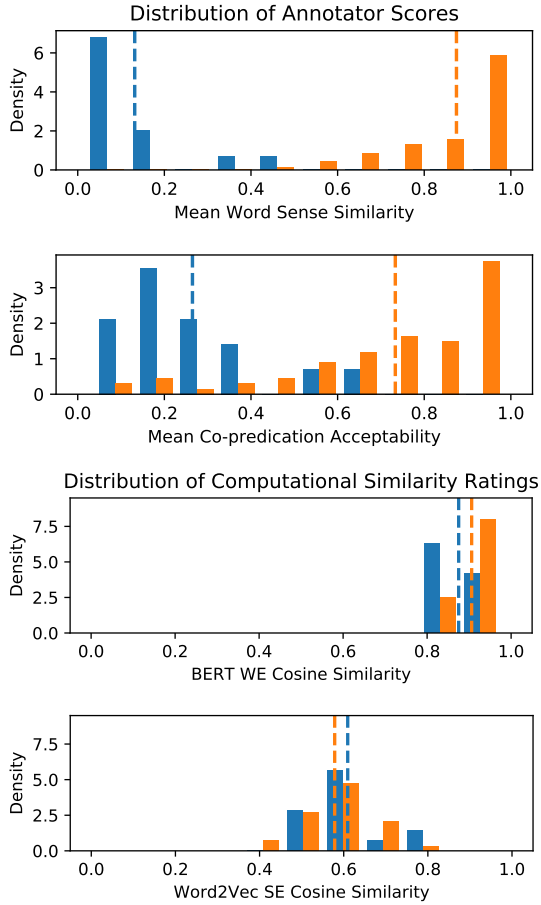


Figure 1: Distribution of human annotation ratings and computational similarity ratings for homonymic (blue) and polysemic (orange) sentence pairs, together with their means.

state size of 768 and 12 attention heads. We i) extracted and averaged sub-word vectors before pooling, ii) extracted the embedding of the [CLS] token, and iii) used the pooled sentence embedding. Lastly, we also determined a primitive contextualised sentence embedding by averaging over the sentence's token embeddings as derived from Word2Vec (Mikolov et al., 2013) pretrained on the Google News Dataset.[10]

## 3 Results

We report the collected data in four steps: Firstly, we inspect to what degree the different metrics and combinations thereof can predict whether a pair of target sense interpretations is polysemic or homonymic. We then investigate the correlation between the three collected annotation metrics, and

model_doc/bert.html

[10] https://code.google.com/archive/p/word2vec/

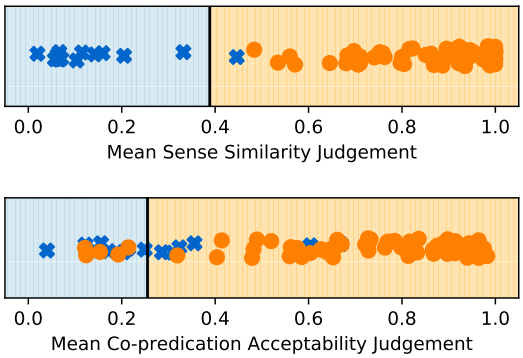


Figure 2: Classification of homonym (blue) and polyseme (orange) sample pairs based on pairwise similarity annotations and co-predication acceptability judgements.

report how well the computational measures predict the human annotations. Finally, we move to a more qualitative analysis, investigating in more detail the distribution of ratings over the different sense interpretations of a polyseme.

### 3.1 Predicting Ambiguity Types

The top two graphs in Figure 1 show the distribution of human annotations for homonymic (blue) and polysemic (orange) target words based on their explicit word sense similarity ratings or co-predication acceptability, respectively. Both annotation measures clearly separate the modes of the distributions, but while co-predication acceptability judgements for the tested polyseme pairs occupy the entire rating scale, explicit word sense similarity ratings only span the upper half (the lowest score is 0.48). Conversely, co-predication acceptability ratings for homonym pairs reach up to 0.67, while the highest-scoring homonym pair only reaches a similarity score of 0.44. This impacts the distribution means, which are closer to each other in the co-predication metric than in the similarity scores. The computational approaches to rating word sense similarities overall return relatively high scores for both, homonym and polyseme pairs, often only occupying the top 20% of the scale. As a result, the means of their distributions are significantly closer, as exemplified by the distributions of BERT word embedding similarity ratings for polyseme and homonym pairs in the third graph of Figure 1. The primitive Word2Vec sentence embeddings lastly assign a higher mean similarity score to homonym pairs than to polysemes (last graph).

Because co-predication acceptability judge-

| Combination | | Correlation | | Ordinary Least Squares (OLS) Regression Analysis | | | | | | Prediction | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **First Measure** | **Second Measure** | **r** | **p** | **Coef.** | **$R^2$** | **F-stat.** | **Prob.** | **Omnib.** | **Prob.** | **MSE** | **$R^2$** |
| **Similarity** | **Acceptability** | 0.529 | 2.08E-06 | 0.910 | 0.280 | 26.855 | 2.08E-06 | 3.756 | 0.153 | 0.040 | 0.208 |
| **Similarity** | **Classification** | 0.539 | 1.21E-06 | 1.091 | 0.291 | 28.320 | 1.21E-06 | 6.587 | 0.037 | 0.057 | 0.162 |
| **Acceptability** | **Similarity** | 0.529 | 2.08E-06 | 0.308 | 0.280 | 26.855 | 2.08E-06 | 22.297 | 0.000 | 0.014 | 0.149 |
| **Acceptability** | **Classification** | 0.563 | 3.21E-07 | 0.662 | 0.317 | 32.015 | 3.21E-07 | 11.321 | 0.003 | 0.050 | 0.301 |
| **Classification** | **Similarity** | 0.539 | 1.21E-06 | 0.267 | 0.291 | 28.320 | 1.21E-06 | 29.957 | 0.000 | 0.014 | 0.175 |
| **Classification** | **Acceptability** | 0.563 | 3.21E-07 | 0.479 | 0.317 | 32.015 | 3.21E-07 | 6.101 | 0.047 | 0.037 | 0.258 |
| **BERT WE** | **Similarity** | 0.211 | 0.077 | 0.762 | 0.045 | 3.226 | 0.077 | 14.001 | 0.001 | 0.018 | -0.214 |
| **BERT WE** | **Acceptability** | 0.482 | 0.000 | 2.991 | 0.233 | 20.936 | 0.000 | 21.974 | 0.000 | 0.041 | 0.204 |
| **BERT WE** | **Classification** | 0.221 | 0.064 | 1.614 | 0.049 | 3.553 | 0.064 | 15.446 | 0.000 | 0.069 | -0.007 |
| **BERT CLS** | **Similarity** | -0.038 | 0.756 | -0.390 | 0.001 | 0.097 | 0.756 | 12.775 | 0.002 | 0.019 | -0.298 |
| **BERT CLS** | **Acceptability** | 0.271 | 0.023 | 4.832 | 0.073 | 5.448 | 0.023 | 13.459 | 0.001 | 0.049 | 0.033 |
| **BERT CLS** | **Classification** | 0.051 | 0.672 | 1.075 | 0.003 | 0.181 | 0.672 | 17.604 | 0.000 | 0.073 | -0.051 |
| **BERT SE** | **Similarity** | -0.007 | 0.955 | -0.067 | 0.000 | 0.003 | 0.955 | 13.383 | 0.001 | 0.020 | -0.322 |
| **BERT SE** | **Acceptability** | 0.011 | 0.929 | 0.181 | 0.000 | 0.008 | 0.929 | 14.479 | 0.001 | 0.058 | -0.162 |
| **BERT SE** | **Classification** | -0.016 | 0.895 | -0.317 | 0.000 | 0.018 | 0.895 | 17.751 | 0.000 | 0.073 | -0.067 |
| **ELMo WE** | **Similarity** | 0.295 | 0.012 | 1.191 | 0.087 | 6.600 | 0.012 | 10.325 | 0.006 | 0.018 | -0.188 |
| **ELMo WE** | **Acceptability** | 0.178 | 0.138 | 1.233 | 0.032 | 2.257 | 0.138 | 13.644 | 0.001 | 0.051 | -0.015 |
| **ELMo WE** | **Classification** | 0.323 | 0.006 | 2.630 | 0.104 | 8.022 | 0.006 | 14.382 | 0.001 | 0.065 | 0.063 |
| **Word2Vec SE** | **Similarity** | 0.053 | 0.662 | 0.085 | 0.003 | 0.193 | 0.662 | 13.484 | 0.001 | 0.020 | -0.305 |
| **Word2Vec SE** | **Acceptability** | 0.245 | 0.039 | 0.681 | 0.060 | 4.423 | 0.039 | 16.732 | 0.000 | 0.051 | -0.006 |
| **Word2Vec SE** | **Classification** | 0.249 | 0.036 | 0.813 | 0.062 | 4.555 | 0.036 | 15.828 | 0.000 | 0.070 | -0.026 |

Table 1: Correlations between the three different metrics of word sense similarity based on annotation judgements, and correlation between computational proxies of word sense similarity as compared to the human judgements. The first set of columns displays pairwise correlation based on Pearson's r, the second set shows the key statistics obtained from their OLS regression, and the third set contains the mean regression scores based on 5-fold cross validation.

ments show a higher overlap between the distributions of homonym and polyseme ratings than the similarity ratings, we expect similarity to be a stronger predictor in classifying target pairs as either homonyms or polysemes. To validate this intuition, we classified items through a support vector machine (SVM) with linear kernel under five-fold cross-validation. As our dataset is skewed towards polysemy samples, baseline performance is an accuracy of 0.825, achieved by assigning all samples to the polyseme class. Both classification based on similarity ratings and co-predication ratings outperform this baseline, with an accuracy of 0.988 for similarity ratings, and 0.895 for co-predication ratings, respectively. Figure 2 shows the optimal decision boundary between homonym samples (blue) and polyseme pairs (orange) calculated for the two annotation metrics. The higher overlap in homonym and polyseme ratings indeed prevents a clear delineation between the two ambiguity types. None of the computational metrics manages to outperform the baseline, and consistently apply max-class labels. Neither combining the two human annotated metrics, nor combining any of the computational metrics improves their respective classification performance over the best individual score.

### 3.2 Relation Between Different Annotations of Sense Similarity

In order to establish a measure of correlation between the three human annotation metrics, we consider all six combinations of metrics and i) calculate their Pearson's r, ii) perform an ordinary least squares (OLS) regression, and iii) calculate the mean squared error (MSE) of OLS predictions under five-fold cross validation. The results of these calculations are displayed in Table 1, and visualised in Figure 3. We find a moderate but significant correlation between the three human annotation metrics. Similarity judgements and co-predication acceptability judgements show the lowest correlation in the set (Pearson's r of 0.529), while acceptability judgements and categorical class similarity achieve the highest correlation (Pearson's r of 0.563). These results indicate that categorical class boundaries between referent interpretations might have a more direct influence on whether two different senses can felicitously be co-predicated than their graded similarity score. The correlation graphs in Figure 3 again display the coverage of judgements obtained for the three human annotation metrics, indicating that class similarity ratings, like co-predication acceptability, span over the full scale, while similarity judgements only cover the

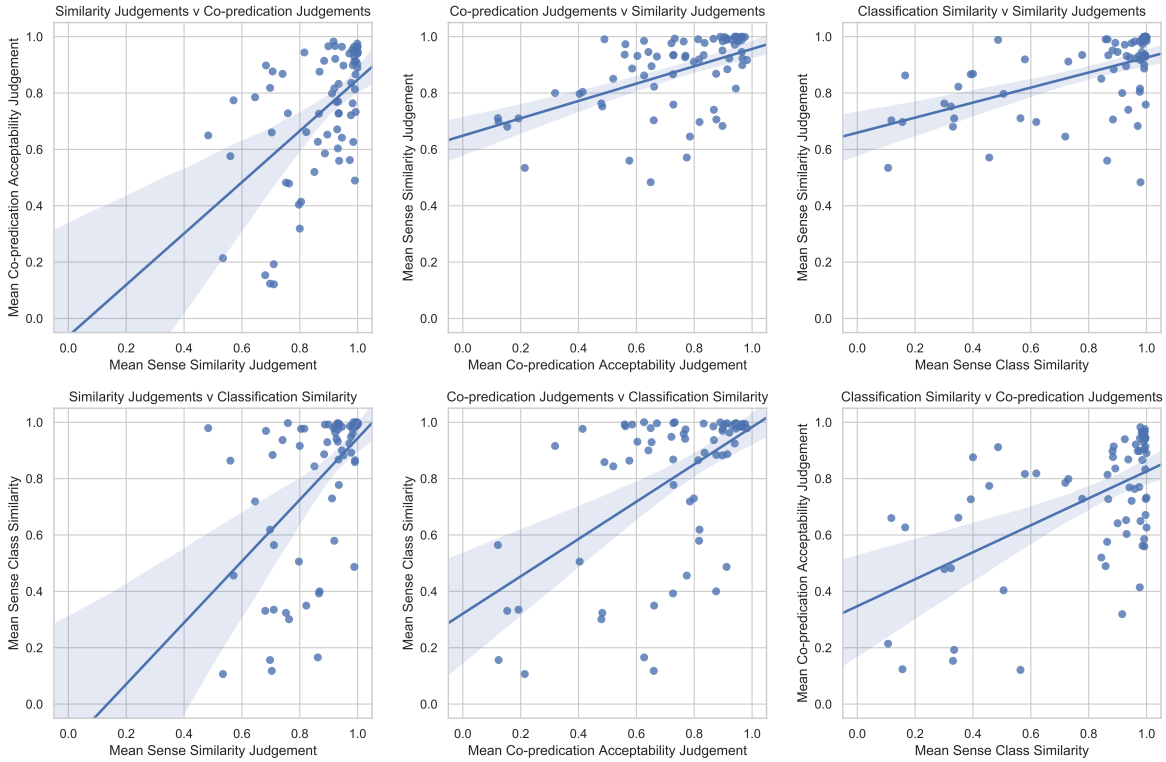


Figure 3: Correlations between polysemic target word pairs based on the three collected judgements of word sense similarity, together with their best linear fit.

top half. Here however this means that predicting co-predication ratings from similarity scores is more difficult than the inverse, leading to a higher error rate in the prediction of low-similarity items, and an overall higher mean squared error (MSE; 0.014 to 0.04). The same holds for predicting similarity class labels from similarity judgements, which is more difficult than predicting similarity judgements based on class similarity.

### 3.3 Relation between Computational Estimates and Human Judgements

The bottom part of Table 1 displays the results of predicting human judgements of polyseme sense similarity based on the different computational proxies. Only seven of the pairwise correlations are significant, and only the correlation between BERT contextualised word embeddings and co-predication acceptability ratings approaches a moderate degree (Pearson's r of 0.48). We argue that it was to be expected that the correlation between the similarity of BERT's contextualised embeddings and co-predication acceptability should be higher than between BERT scores and explicit similarity ratings, as BERT does not specifically capture the sense of a target word, but rather the diversity and type of context it appears in. This way it is easier to predict whether a combined context as created by co-predication is natural to occur (and therefore more felicitous) than to directly predict the targets' sense similarity. Other notable significant pairs are ELMo word embeddings and classification similarity (Pearson's r of 0.32), ELMo and similarity ratings (r = 0.3), as well as BERT classification token similarity and co-predication acceptability (r = 0.27), indicating that BERT and ELMo might capture slightly different facets of word sense - but, as indicate above - not in such a way that combining them would improve their performance in predicting the ambiguity type of a target word pair.

### 3.4 Qualitative Analysis

While the correlation between explicit similarity judgements and co-predication acceptability is imperfect, our analysis reveals that judgements are more similar towards the upper end of the rating scale than at the lower end. To investigate this observation in more detail, we here analyse polyseme *newspaper*, which provides two samples to the low-similarity cluster. As mentioned before, in our experiments we assume that *newspaper* has three distinct but related sense interpretations: (1)

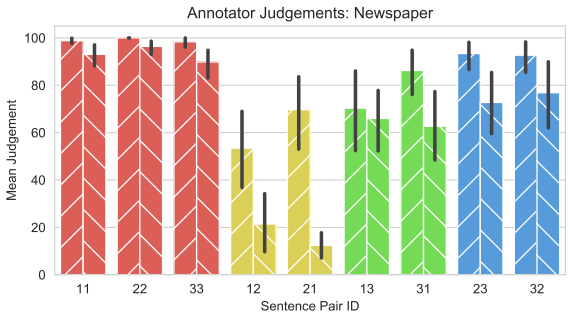


Figure 4: Mean similarity ratings (left, ascending hatch) and co-predication acceptability ratings (right, descending hatch) for the nine sense interpretation pairs of polyseme *newspaper*. The first three bars represent same-sense pairs, the other three groups the different combinations of cross-sense readings, respectively.

*organisation/institution*, (2) *physical object*, and (3) *information/data*. Figure 4 shows the mean similarity and acceptability ratings for the nine combinations of sense interpretations: The first three bars represent same-sense pairs 11, 22 and 33, the other three groups the different combinations of cross-sense pairs. The figure reveals that the three same-sense pairs receive equally high similarity and acceptability ratings, but while similarity ratings show a gradual decrease in scores assigned to cross-sense pairs, the co-predication acceptability scores are only gradual for more similar cross-sense pairs, and drop significantly for less similar ones. These results indicate that similarity ratings appear to be a more nuanced, continuous measure than co-predication acceptability, which can assigns extremely low scores for readings deemed to be infelicitous. A more detailed investigation of the grouping of polyseme senses and its implications for the hypothesis of hierarchical sense representation can be found in Haber and Poesio (2020).

## 4 Conclusion

The data collected in this study allows for a number of observations about the role of word sense similarity in the processing of homonyms and polysemes. On the one hand, graded co-predication acceptability ratings are shown to be less able to tell apart samples of homonymic and polysemic sense pairs than explicit sense similarity ratings. This supports the growing collection of studies indicating that co-predication might not be as suited a tool to distinguish different types of lexical ambiguity as traditionally assumed. On the other hand, the collected judgements of word sense similarity indicate that polyseme sense pairs mis-classified by co-predication acceptability are overall less similar to each other than other sense pairs, and significantly so than same-sense interpretations. This to some degree vindicates co-predication as a linguistic test, suggesting that rather than distinguishing homonyms form polysemes per se, it might be a coarse indication of the underlying word sense similarity.

Our results also provide support for recent hypotheses suggesting that polyseme representation in the mental lexicon cannot be fully underspecified. During data collection, annotators rated some polysemic sense interpretations to be significantly less similar to each other than other sense pairs, and even rated some of the polyseme cross-sense co-predication samples as unacceptable. This indicates that the interpretations of polysemic words might be grouped based on their similarity, and only grouped interpretations are available for cost-free sense shifting and felicitous co-predication. Because only a single target word per type of systematic polysemy was tested here, we cannot ascertain whether sense groupings are idiosyncratic or systematic across target words of a certain polysemy type. Data for an analysis of this question can however easily be obtained by repeating our experiments with a larger set of target words. In a similar vain, we also recommend an in-depth analysis of irregular or metaphorical polysemes, which were omitted in this data collection effort.

Lastly, investigating the suitability of contextualised language models as proxies for human word sense similarity judgements, we find that the tested contextualised embeddings fail to predict word sense similarity consistently, but that the similarities between BERT embeddings show a significant correlation with co-predication acceptability ratings. We take this finding as evidence that BERT might create better encodings of complex contexts than encodings of actual word meaning, as it seems to perform well in determining whether contexts can be felicitously combined without consistently determining the similarity of word senses from these contexts first. We strongly encourage further research into determining the exact lexical semantic information available in BERT encodings in order to shed more light on this issue.

## Acknowledgements

The work presented in this paper was supported by the DALI project, ERC Grant 695662. The authors would like to thank Derya Çokal and Andrea Bruera for their input, and the anonymous reviewers for their feedback.

## References

Sandra Antunes and Rui Pedro Chaves. 2003. On the Licensing Conditions of Co-Predication. In *Proceedings of the 2nd International Workshop on Generative Approaches to the Lexicon*.

Juri D. Apresjan. 1974. Regular polysemy. *Linguistics*, 12:5–32.

Nicholas Asher. 2011. *Lexical Meaning in Context: A Web of Words*. Cambridge University Press.

Nicholas Asher and James Pustejovsky. 2006. A type composition logic for generative lexicon. *Journal of Cognitive Science*, 6(1).

Gemma Boleda, Sabine Schulte im Walde, and Toni Badia. 2012. Modeling regular polysemy: A study on the semantic classification of catalan adjectives. *Computational Linguistics*, 38(3):575–616.

Alan D. Cruse. 1995. Polysemy and related phenomena from a cognitive linguistic viewpoint. In Patrick Saint-Dizier and Evelyn Viegas, editors, *Computational Lexical Semantics*, Studies in Natural Language Processing, page 33–49. Cambridge University Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Johannes Dölling. Forthcoming. Systematic Polysemy. In Daniel Gutzmann, Lisa Matthewson, Cécile Meier, Hotze Rullmann, and Thomas Ede Zimmermann, editors, *The Blackwell Companion to Semantics*. Wiley.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Ingrid Lossius Falkum and Augustin Vicente. 2015. Polysemy: Current perspectives and approaches. *Lingua*, 157:1–16.

Hana Filip and Peter Sutton. 2017. Singular count NPs in measure constructions. In *Semantics and Linguistic Theory*, volume 27, pages 340–357.

Jerry A. Fodor. 1998. *Concepts: Where Cognitive Science Went Wrong*. Oxford University Press.

Lyn Frazier and Keith Rayner. 1990. Taking on semantic commitments: Processing multiple meanings vs. multiple senses. *Journal of Memory and Language*.

Steven Frisson. 2015. About bound and scary books: The processing of book polysemies. *Lingua*, 157:17 – 35. Polysemy: Current Perspectives and Approaches.

Janosch Haber and Massimo Poesio. 2020. Word sense distance in human similarity judgements and contextualised word embeddings. In *Proceedings of the Probability and Meaning Conference (PaM 2020)*, pages 128–145, Gothenburg. Association for Computational Linguistics.

Ray Jackendoff. 1989. What is a concept, that a person may grasp it?1. *Mind & Language*, 4(1-2):68–102.

Ekaterini Klepousniotou. 2002. The Processing of Lexical Ambiguity: Homonymy and Polysemy in the Mental Lexicon. *Brain and Language*, 81(1-3):205–223.

Ekaterini Klepousniotou, G. Bruce Pike, Karsten Steinhauer, and Vincent Gracco. 2012. Not all ambiguous words are created equal: An EEG investigation of homonymy and polysemy. *Brain and Language*.

Ekaterini Klepousniotou, Debra Titone, and Carolina Romero. 2008. Making sense of word senses: The comprehension of polysemy depends on sense overlap.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2014. Measuring Gradience in Speakers' Grammaticality Judgements. *Proceedings of the 36th Annual Meeting of the Cognitive Science Society (CogSci 2014)*.

John Lyons. 1977. *Semantics*, volume 2. Cambridge University Press.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*.

Andrei Moldovan. 2019. Descriptions and tests for polysemy. *Axiomathes*, pages 1–21.

Marina Ortega-Andrés and Agustín Vicente. 2019. Polysemy and co-predication. *Glossa: a journal of general linguistics*, 4(1).

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *CoRR*, abs/1802.05365.

James Pustejovsky. 1991. The Generative Lexicon. *Comput. Linguist.*, 17(4):409–441.

Liina Pylkkänen, Rodolfo Llinás, and Gregory L. Murphy. 2006. The representation of polysemy: Meg evidence. *Journal of cognitive neuroscience*, 18(1):97–109.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

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

Petra Schumacher. 2013. When combinatorial processing results in reconceptualization: toward a new approach of compositionality. *Frontiers in Psychology*, 4:677.

Peter R. Sutton and Hana Filip. 2018. Counting Construcions and Coercion: Container, Portion and Measure Interpretations. *Oslo Studies in Language*, 10(2).

Matthew J. Traxler, Brian McElree, and Martin J. Williams, Rihana S .and Pickering. 2005. Context effects in coercion: Evidence from eye movements. *Journal of Memory and Language*, 53(1):1–25.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does bert make any sense? interpretable word sense disambiguation with contextualized embeddings.

Sarah Zobel. 2017. The sensitivity of natural language to the distinction between class nouns and role nouns. In *Semantics and Linguistic Theory*, volume 27, pages 438–458.

Arnold M. Zwicky and Jerrold M. Sadock. 1975. Ambiguity tests and how to fail them. In *Syntax and Semantics volume 4*, pages 1–36. Brill.