Word Sense Distance in Human Similarity Judgements and Contextualised Word Embeddings

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

Abstract

Homonymy is often used to showcase one of the advantages of context-sensitive word embedding techniques such as ELMo and BERT. In this paper we want to shift the focus to the related but less exhaustively explored phenomenon of polysemy, where a word expresses various distinct but related senses in different contexts. Specifically, we aim to i) investigate a recent model of polyseme sense clustering proposed by Ortega-Andrés and Vicente (2019) through analysing empirical evidence of word sense grouping in human similarity judgements, ii) extend the evaluation of context-sensitive word embedding systems by examining whether they encode differences in word sense similarity and iii) compare the word sense similarities of both methods to assess their correlation and gain some intuition as to how well contextualised word embeddings could be used as surrogate word sense similarity judgements in linguistic experiments.

1 Introduction

Homonymy, the linguistic phenomenon of a word taking on a different meaning based on its context, such as *match* in (1), is often used to showcase one of the advantages of context-sensitive word embedding techniques such as ELMo (Peters et al., 2018), BERT (Devlin et al., 2018) and GPT-2 (Radford et al., 2019) over their traditional word-vector counterparts such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), which are unable to encode context-dependent meaning.

(1) a. The match burned my fingers.
b. The match ended without a winner.

In this paper we want to shift the focus to the related but less exhaustively explored phenomenon of polysemy. We speak of polysemy when a word takes on different distinct but related senses given its context (Lyons, 1977), such as *school* in the various contexts of (2).[1]

(2) 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.

Specifically, we aim to investigate a recent model of polyseme sense clustering proposed by Ortega-Andrés and Vicente (2019), suggesting that similarity differences in polysemic senses could lead to a grouping in their representation in the Generative Lexicon (Pustejovsky, 1991), addressing and attempting an explanation for processing differences observed within the seemingly homogeneous group of polysemes.

Through a range of surveys we collect word sense similarity judgements for a set of polysemes to provide empirical data for an investigation of word sense clustering as proposed by Ortega-Andrés and Vicente. We then aim to extend the linguistic evaluation of context-sensitive word embeddings by examining whether their contextualised encodings of polysemes show signs of word sense grouping, and whether these groupings correlate with the patterns observed in the human judgements. If this is the case, contextualised word embeddings could be used as surrogate word sense indicators in linguistic experiments.

1.1 Processing of Polysemes

While on a first glance homonymy and polysemy seem to be two closely related phenomena, poly-

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

Proceedings of the Probability and Meaning Conference (PaM 2020), pages 128–145
October 14-15 2020 ©2020 Association for Computational Linguistics

semy should not be viewed as a simple extension of homonymic ambiguity: While the interpretation of a homonym requires the selection of one and only one specific meaning, polysemes have been found to activate multiple sense interpretations simultaneously and in many cases accommodate for sense shifting without additional processing cost. Frazier and Rayner (1990) for example showed that late disambiguating contexts can cause processing difficulties for homonyms but not so for polysemes. This observation led them to postulate a fully specified mental representation for homonymic meaning (i.e. one entry per meaning), but an un- or under-specified representation of polysemic sense. Studies like Klepousniotou (2002); Pylkkänen et al. (2006) and Klepousniotou et al. (2012) later revisited this experiment with the support of MEG and EEG readings, observing significant priming effects in homonyms but not so for polysemes. This led them, too, to postulate a principled processing difference in the interpretation of homonyms and polysemes.

A second case for a systematic difference between homonymy and polysemy has been made using so-called co-predication tests. In co-predication, two different meanings or senses of a word are simultaneously invoked by the context. In the case of homonymy, co-predication will always result in an infelicitous sentence, like for example in (3). For polysemous words on the other hand, co-predication with different senses seems to be felicitous in principle (e.g. example (4)).

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

(4) Lunch was delicious but took forever. [food/meal]

1.2 Representation of Polysemes

A variety of linguistic models, including the Generative Lexicon (Pustejovsky, 1991) and Type Theory with Records (TTR, e.g. Cooper and Ginzburg, 2015), have been proposed to accommodate the observed processing differences between homonyms and polysemes. Specifically, Gotham (2014) proposed methods for addressing co-predication, quantification and individuation of polysemic senses in TTR, and Asher and Pustejovsky (2006) and Asher (2011, 2015) augmented the Generative Lexicon model by proposing that the various senses of a polyseme are represented by so-called dot-objects, complex objects that distinguish the different aspects, facets and types of polysemic sense interpretations, arguing that a word's context selects for the appropriate sense from within that representation.

Opposing a unified, under-specified representation of polysemic sense, a growing body of work however also collected a range of observations indicating that there might be significant and potentially systematic differences between various polysemic interpretations as well. Dating back to at least Apresjan (1974), for example, stems the idea that polysemes should be sub-divided into two types, *regular* (or *systematic*), and *irregular* polysemy, based on whether a polyseme's set of interpretations is idiosyncratic or shared among a group of similar words (also see Falkum (2015)). Supporting this principled split, Klepousniotou et al. (2012) report that their experiments indicate that regular polysemes might be represented differently than their irregular counterparts, arguing that in their processing, irregular polysemes more resemble homonymic meaning alterations than the sense alterations in regular polysemes. Furthering this discussion, Dölling (Forthcoming) recently collected a fine-grained distinction of 19 different patterns of polysemic sense alteration within the set of systematic polysemes, begging the question whether even regular polysemes form a homogeneous group and share a common representation, or whether these, too, require a more structured distinction than previously assumed.

Other evidence comes from an ongoing series of co-predication studies (Antunes and Chaves (2003); Traxler et al. (2005); Zobel (2017), and Filip and Sutton (2017); Sutton and Filip (2018); Schumacher (2013) for observations and models specifically concerning content/container alterations), showing 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, too (see example (5)).[2]

(5) 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/aperture]

[2] Examples from Cruse (1995)

$$\text{School:}\begin{bmatrix} \text{Formal: Institution = x} \\ \text{Telos: educating (e, x, y, z)} \\ \text{Participants:}\begin{bmatrix}\text{Students = y}\\ \text{Teachers = z}\end{bmatrix} \\ \text{Social Realization:}\begin{bmatrix}\text{Organization (x):}\begin{bmatrix}\text{Rules}\\ \text{Staff}\end{bmatrix}\\ \text{Representation (x):}\begin{bmatrix}\text{Director}\\ \text{Student Rep.}\\ \text{Sport team}\\ \dots\end{bmatrix}\end{bmatrix} \\ \text{Temp. Realization:}\begin{bmatrix}\text{Process (x) = e:}\begin{bmatrix}\text{Temp. Org.}\\ \text{Timetable}\\ \text{Academic course}\end{bmatrix}\end{bmatrix} \\ \text{Phys. Realization:}\begin{bmatrix}\text{Building (x):}\begin{bmatrix}\text{Outside}\\ \text{Inside}\\ \text{Occupants = participants (y, z)}\end{bmatrix}\end{bmatrix}\end{bmatrix}$$

Figure 1: Knowledge structure for polyseme *school* proposed by Ortega-Andrés and Vicente (2019). Figure replicated from ibid., page 5.

To account for processing differences among polysemic senses, Ortega-Andrés and Vicente (2019) recently proposed an extension to Asher and Pustejovsky's model by postulating a hierarchical representation of polysemic sense that groups target word senses based on their similarity, creating so-called activation packages. Sense shifting is assumed to be automatic and free of processing costs within the under-specified representation of an activation package, but will lead to processing difficulties and infelicitous co-predication when moving outside of it. Figure 1 shows a proposed ordering of the hierarchical structure and resulting co-activation packages for polyseme *school* according to this model.

In summary, a number of recent observations, hypotheses and models concerning polysemy point towards a continuum of sense (or meaning) similarity between truly polysemous expressions (or, in Pinkal's terms, p-type ambiguity) and homonymic ambiguity (or h-type ambiguity, see Pinkal (1985) and Poesio (Forthcoming)) where some senses might be more closely related to one another than others. In this paper we aim to provide additional empirical data for investigating this claim by i) collecting graded word sense similarity judgements to assess the notion of word sense grouping as a driving factor in determining the representation of polysemic sense and accounting for differences in their processing costs and co-predication acceptability. In addition, we ii) investigate if clustering the representations of polysemes as generated by contextualised word-embedding techniques such as ELMo and BERT develops a word sense grouping and whether this grouping correlates with that derived from the collected sense similarity judgements. If this is the case, contextualised word embeddings could be used as surrogate word sense indicators in linguistic experiments.

2 Method

In order to generate the clearest results possible for investigating potential distances between different polysemic sense interpretations, we decided to use custom samples instead of resorting to corpus samples in this study. By creating the samples ourselves, we can construct contexts that invoke a certain polysemic sense as clearly as possible, and we can create sentence pairs that combine any of the different interpretations in order to have annotators judge their similarity directly. In addition, a preliminary investigation of context-sensitive polyseme representations obtained from ELMo revealed that factors such as i) the position of the target expression in the sentence, ii) its syntactic function and iii) the overall sentence length all significantly influence the resulting embedding and might overshadow the differences in encoding stemming from interpretation differences.[3] Designing custom samples helps us to control for these factors.

[3] See Appendix A

2.1 Samples

As target expressions for our samples we decided to focus on regular polysemes, as they are more likely to produce the clearest results possible due to their canonical division of sense interpretations. 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. We then created a sample set for each of the ten polysemes, containing two sample sentences for each of the target expression's interpretations.[4] The samples were created such that i) the 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.[5] As an example, consider the six sample sentences for 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.

All sample sentences were rated to be acceptable by annotators recruited from Amazon Mechanical Turk (AMT)[6] in a validation experiment. Individual sample sentences were then combined into pairs invoking all possible combinations of sense interpretations (i.e. creating nine sentence pairs for *newspaper*) and distributed over books so that no target expression appears twice in any book. In total, we generated 67 target pairs and distributed them over 15 books. We then followed Lau et al. (2014) by adding one of 15 sentence pairs containing homonyms and one of 15 sentence pairs containing synonyms to each book to create test items for spotting spammers, and further filled the books with random combinations of filler sentence pairs in order to disguise the focus on polysemes and present objectively low similarity items to calibrate the annotator's ratings.

2.2 Human Judgements

We used AMT to collect word sense similarity judgements by highlighting (polysemic) target expressions in the sentence pairs and asking 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.[7] The submitted slider positions are translated to a similarity score between 0 and 100 and stored in combination with a workers unique ID. To improve judgement quality, we required workers to have obtained a US high school degree and reached the "AMT Master" qualification.[8] Workers were paid 0.35 USD for every completed book.

We collected 20 judgements for each book. A total of 65 individual workers contributed to the study, with HITs taking an average of 133.4 seconds (median of 90.0). Through filtering out any books where the homonym sentence pair or a filler pair was given a similarity score higher than 60, or where the synonym sentence pair was rated lower than 50, we removed a total of 51 books and obtained an average of 16.6 judgements per sentence pair (min = 13).

2.3 Word Embeddings

Models of polysemy have previously been proposed in distributional semantics (see for exanple 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. ELMo was developed explicitly to capture a target word's context, processing sentences with a two-layered, bi-directional LSTM and computing the weighted sum of their hidden states depending on the task at hand to create contextualised word embeddings. BERT on the other hand is a language model that borrows and stacks the encoder architecture of the Transformer (Vaswani et al., 2017), an attention mecha-

[4] See Appendix B for details.

[5] As in "The school is an old building." for sense *building*

[6] `https://www.mturk.com/`

[7] See Figure 7 for a screenshot of the AMT HIT interface.

[8] 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`

nism for learning the contextual relations between words, and adds a masking technique that allows for processing sentences in a non-directional fashion with minimised interference among the layers. While BERT's output, either an array of embeddings or a single pooled one, is normally fed to a further model to process a language-based task, our aim is to see whether it is able to capture any differences in polysemic sense and use its outputs directly.

For the ELMo analysis we used a pretrained model available on TensorFlow Hub[9] 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 (and consequently the ELMo output layer that combines them. See Appendix A, and Ethayarajh (2019)). For the investigation of BERT's embeddings we used the output of a pretrained cased model from the same repository[10] with 12 layers, a hidden state size of 768 and 12 attention heads. We extract i) sub-word vectors before pooling, ii) use the pooled sentence vector or iii) the embedding of the special [CLS] token.

3 Results

3.1 Similarity Differences in Judgements

As a first step, we calculated the overall means of word similarity judgements for all polyseme, homonym, synonym and filler sentence pairs in the dataset to determine any principled differences among these groups. The polyseme sentence pairs obtained a mean similarity rating of 87.12 (std=20.92), synonym sentence pairs a mean of 92.38 (std=10.35), homonym pairs a mean of 3.76 (std=8.37) and filler sentence pairs a mean of 2.71 (std=7.19). We then used Student's T-Test to compare the distributions of judgements; The polyseme and synonym distributions each are significantly different from all other distributions ($p<0.05$). This means that annotators rated synonyms to be overall more similar to each other than different uses of polysemes - a first indicator that word sense interpretations might not be perceived as carrying identical meaning.

Polyseme	**Same-sense**		**Cross-sense**	
	mean	std	mean	std
Newspaper (3)	99.17	2.36	77.71	30.08
Hemingway (2)	96.26	16.64	85.64	24.71
War and Peace (3)	99.55	2.65	91.78	22.73
Lunch (2)	96.15	11.98	80.35	24.51
Door (2)	99.33	2.27	95.88	9.73
DVD (3)	95.56	12.34	88.12	20.58
School (4)	96.30	8.57	88.08	22.97
Wine (2)	99.85	0.50	92.30	17.25
Glass (2)	70.39	35.02	65.03	38.02
Construction (2)	86.49	21.65	59.93	33.44

Table 1: Polysemic target expression (number of senses), and means and standard deviations of the same-sense and cross-sense samples' pairwise similarity ratings.

Because the ten different polysemous target expressions used in this study each represent a different type of regular polysemy, we next split the collected judgements based on their target expression and calculated the mean sense similarity judgements for same-sense and cross-sense sentence pairs. Table 1 displays these numbers, showing that same-sense means are consistently higher than the cross-sense ones, and except for *glass* and *construction* range above 95 (i.e. higher than the synonym mean). This means that barring these two outliers, the generated same-sense pairs were rated as invoking an almost identical interpretation of the polysemic target expression. The average similarity of cross-sense pairs often ranges between 80 and 90, showing a high similarity still, but indicating that not all cross-sense pairs seem to be perceived as invoking the same sense.

Turning to a more qualitative analysis of the results obtained for each individual polyseme, we investigated the similarity ratings obtained for sentence pairs containing a specific target expression to assess whether the collected data provides any evidence for sense clustering as proposed by Ortega-Andrés and Vicente (2019). Since it is difficult to collapse results over the different types of polysemes tested, we here exemplify our analyses through a summary of the observations concerning polyseme *newspaper* and draw parallels to other test items where possible. As mentioned above, the polyseme *newspaper* was taken to invoke one of three distinct but related senses; (1) *organisation/institution*, (2) *physical object* and (3) *information/data*, and creating all combinations of senses generates the

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

[10] https://tfhub.dev/google/bert_cased_L-12_H-768_A-12/1

following nine sentence pairs:[11]

11 organisation/organisation
22 physical/physical
33 information/information

12 organisation/physical
21 physical/organisation

13 organisation/information
31 information/organisation

23 physical/information
32 information/physical

Figure 2 shows the mean word similarity judgements for these nine sentence pairs. The three same-sense pairs 11, 22, and 33 (red) receive mean similarity ratings close to 100, showing that in these cases annotators indeed perceive the target word contexts to invoke exactly the same sense in both sample sentences. This effect can be observed for all tested polysemes except for *glass*, where one of the same-sense pairs does not actually seem to elicit the same sense (rated at a similarity of 48) and a same-sense pair for *construction* which only received a similarity score of 82 (being higher still than the cross-sense pairs). Returning to *newspaper*, all six cross-sense pairs receive lower ratings than the same-sense pais: Both, the *organisation/physical* sentence pairs 12 and 21 (yellow), and the *organisation/information* sentence pairs 13 and 31 (green) receive significantly lower similarity ratings than the same-sense pairs. The similarity ratings for the *physical/information* pairs 23 and 32, (blue) are ranging between 90 and 100, being significantly higher than the ratings for pairs 12, 21, 13, but significantly lower than same sense-sense pair 22. This indicates that at least between the *organisation* and *physical* sense interpretation there seems to be a notable difference in meaning, while the *information* readings are judged to be relatively similar to either - however not to a level that same-sense sample pairs are similar to each other. We see a similar but less pronounced effect for the tested polysemes with two senses, where cross-sense pairs are rated as being less similar than the same-sense pairs, as well as in some of the senses of target expressions with three or four interpretations, with significant differences between the *building* and *administration* and *institution* senses of polyseme *school*.

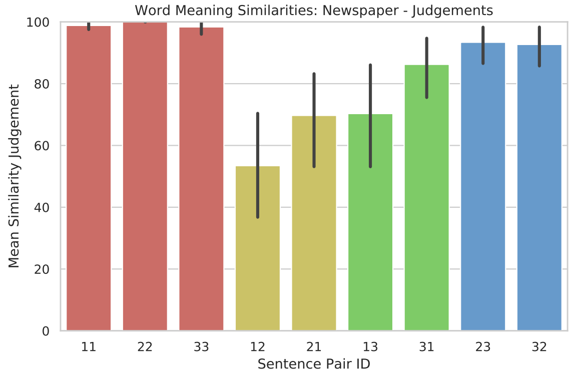


Figure 2: Similarity judgements for sentence pairs containing the polyseme *newspaper*. The two numbers in the sentence pair IDs indicate the combination of senses. The first three bars thus indicate same-sense pairs, the other three groups the different variations of cross-sense samples. The full set of similarity judgements graphs can be found in Appendix D.

Returning to the *newspaper* samples, a second point of interest are the notable though non-significant differences in similarity ratings for sentence pairs 12 and 21, and 13 and 31, respectively. Since these sentence pairs were created to invoke the same pair of (cross-sense) interpretations, it is noteworthy that their ratings differ so much. This difference can be the result of two factors: i) the sentence pairs contain different sample sentences, which within the same sense interpretation could evoke interpretation differences, and ii) the order of presentation for the two sentence pairs is different, and presentation order is known to induce biases and affect acceptability in co-predication studies. To control for the latter, we repeated our experiments with the same set of samples, but inverting the presentation order within the sentence pairs. Based on an average of ten judgements, only one of the 67 sentence pairs' similarity ratings changed significantly, indicating that the observed difference in similarity ratings is not an effect of presentation order, but indeed due to subtle interpretation differences in the contexts used to elicit a certain sense.

3.2 Correlation with Embedding Techniques

Observing noticeable differences in the word sense similarity ratings between some of the sample sentences invoking different interpretations of a polyseme - and in some cases even within sentence pairs that were designed to invoke the same

[11] see Appendix G for the full list of sample sentences.

	Newsp.	Hemingw.	W&P	Lunch	Door	DVD	School	Wine	Glass	Constr.
BERT WE	0.383	0.692	**0.235**	**0.899**	0.079	0.409	0.259	0.459	-0.739	**0.623**
BERT SE	0.591	**0.999***	-0.159	0.316	**0.449**	0.355	0.092	0.458	-0.973*	-0.115
BERT CLS	0.317	0.960*	0.017	0.152	-0.202	**0.517**	0.084	0.216	-0.933	-0.492
ELMo WE	**0.919***	0.916	-0.310	-0.278	0.018	-0.167	**0.332**	0.442	-0.666	0.648
Word2Vec SE	0.576	0.126	0.089	-0.923	0.177	0.361	-0.310	**0.795**	-0.614	0.117

Table 2: Correlations between human sense similarity judgements and the similarities in the representations derived from different contextualised word embedding techniques as measured with Pearson's r. Highest correlating model output in bold font, significant correlations ($p<0.05$) starred.

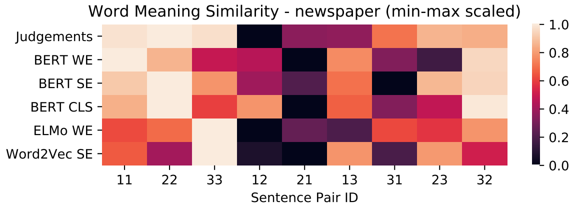


Figure 3: Comparison of word sense similarity ratings based on annotator judgements and ELMo and BERT context-sensitive word embeddings, min-max normalised to amplify the visibility of effects. Brighter indicates higher similarity.

sense - we proceeded to investigate whether the contextualised embeddings of polysemes generated by ELMo and BERT also exhibit word sense similarity differences. To this end we used the raw embeddings as returned by the models and calculate their similarity based on cosine. If a target expression contains multiple words (or sub-word tokens in the case of BERT), we average the embeddings of all parts. In addition to ELMo and BERT word embeddings for the target expressions alone, we also consider BERT's pooled sentence embedding, the embedding of the special [CLS] token, and a sentence embedding by Word2Vec created by averaging all word embeddings in the sentence. Table 2 displays the correlations between the human sense similarity ratings and the cosine similarities of the target expressions (or sentences) given these different embedding techniques. With only a fraction of the correlations being significant,[12] none of the embedding techniques appears to capture the similarity patters observed in the human judgements consistently, with each of the methods achieving the strongest correlation for one or two of the target expressions, but also showing instances of negative or no correlation for some samples.

Moving to a more qualitative analysis of the contextualised embeddings, we created heat maps to display the similarity patterns for the different polysemic expressions tested. The resulting heat map for *newspaper* is shown in Figure 3, displaying on a more accessible level the difference in correlation between the human judgements and contextualised embeddings.[13] While in some cases the cosine similarities between the contextualised embeddings seem to reflect the human judgements - especially so for sense interpretations rated to be highly similar (e.g. 11, 22 and 32) or dissimilar (12, 21) - overall the differences in embeddings do not consistently resemble the human judgements. This observation is replicated throughout the ten polysemes tested in this study, with some of the 2-sense samples also exhibiting more consistent patterns.

While the similarities between contextualised embeddings do not consistently match the collected sense similarity ratings, the patterns in their embeddings indicate that they do differentiate between the different contexts. We further investigated this intuition by applying a non-linear function to reduce the dimensionality of 15 different word embeddings for polyseme *newspaper* produced by ELMo using t-SNE (van der Maaten and Hinton, 2008) and visualising the result in the two-dimensional scatter plot displayed in Figure 4. The samples for this experiment were created to invoke the polyseme's three senses *organisation* - red (1-5), *physical* - yellow (6-10), and *information* - green (11-15).[14] And while no clear grouping into different sense clusters seems to emerge, we do observe a similar pattern to that found in Figure 2, namely that the *physical* interpretations seem to be more similar to the *information* senses,

[12] Note that the compared similarity vectors are of length 4-16 only

[13] The heat maps for the full set of tested polysemes can be found in Appendix E.

[14] See Appendix F for the list of samples.

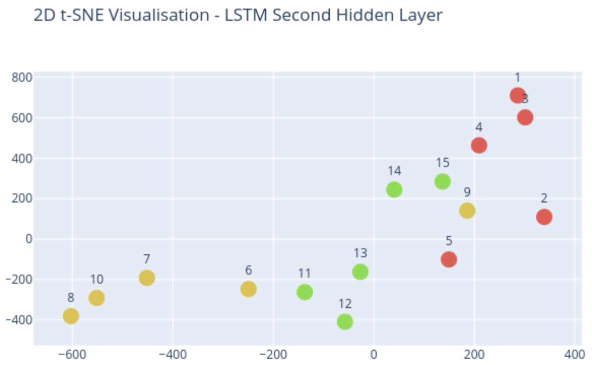


Figure 4: t-SNE scatter plot of the reduced ELMo embeddings for 15 instances of polyseme *newspaper* presented with disambiguating contexts for its three senses: *organisation* - red, *physical* - yellow, and *information* - green. Sample sentences in Appendix F.

which are occupying the space between *physical* and *organisation* readings. Note however that due to the working of t-SNE's dimensionality reduction algorithm, results can change between iterations, and the observed pattern is not always visible. This means that from this test alone it is unclear whether polysemic sense is indeed encoded in the ELMo embeddings.

4 Conclusion

While our results are difficult to collapse as they survey different types polysemes, there are some overarching conclusions to be drawn from the data collected in this study. First of all, we provide empirical evidence that readers are indeed sensitive to differences between polysemic word senses. Polysemic target expressions in contexts designed to invoke the same sense interpretation are consistently rated as highly similar, while similarity ratings of cross-sense pairs receive ratings on a spectrum ranging from highly similar to significantly dissimilar. It thus seems that some sense interpretations are perceived to be more similar to each other than others, providing support for a similarity-based grouping of word senses like the one as proposed by Ortega-Andrés and Vicente (2019). In some cases, distances in sense interpretations correspond to intuitive groupings of senses, but the collected judgements also reveal a notion of gradedness that usually is not assumed to be present in canonical samples (see e.g. Lau et al. (2014)). Given these observations, we see merit in exploring a more structured representation of polysemic sense, since a fully underspecified, single-entry approach would be insufficient to fully account for them. Having investigated only one target expression for a small set of systematic polysemes, we acknowledge that more empirical research is needed to investigate potential patterns within polysemy types or in the much larger set of irregular polysemes in order to determine whether there are any systematic effects - or whether each and every polysemic expression requires its own idiosyncratic representation structure. The graded word sense judgements obtained through our data collection however also indicate to us that a categorical approach to word sense disambiguation (WSD) such as implied by a number of recent models (e.g. Levine et al. (2019); Wiedemann et al. (2019); Blevins and Zettlemoyer (2020)) might be geared more towards the distinction of homonymic ambiguity and fall short of capturing the full spectrum of polysemic sense alterations. Current approaches focusing on graded word sense similarity and word sense shifting (see for example Armendariz et al. (2019)) on the other hand might produce new insights in mapping out the intricacies of polysemic word sense interpretation and representation.

Concerning the encoding of polysemic sense in the contextualised word embeddings of ELMo and BERT, we do observe differences in the representation of polysemic expressions invoking different sense interpretations which could indicate the encoding of context-specific information, but similarities between word embeddings do not consistently correlate with the collected human similarity judgements. While the raw embeddings thus cannot directly be used to distinguish polysemic senses to the same degree as human judgements do, they still could contain word sense information that requires non-linear functions in order to be accessed. Failing to provide conclusive answers in this respect, we hope that future work will help to determine to what extend - and how exactly - polysemic sense is represented in contextualised embeddings to shed more light into the black box processes that improve so many NLP systems.

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.

Carlos Santos Armendariz, Matthew Purver, Matej Ulčar, Senja Pollak, Nikola Ljubešić, Marko Robnik-Šikonja, Mark Granroth-Wilding, and Kristiina Vaik. 2019. Cosimlex: A resource for evaluating graded word similarity in context.

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

Nicholas Asher. 2015. Types, meanings and coercions in lexical semantics. *Lingua*, 157:66–82.

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

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss-informed biencoders.

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.

Robin Cooper and Jonathan Ginzburg. 2015. *Type Theory with Records for Natural Language Semantics*, chapter 12. John Wiley & Sons, Ltd.

Alan D. Cruse. 1995. Polysemy and related phenomena from a cognitive linguistic viewpoint. In Patrick Saint-Dizier and EvelynEditors 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. 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.

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

Matthew Graham Haigh Gotham. 2014. *Copredication, quantification and individuation*. Ph.D. thesis, UCL (University College London).

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*.

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)*.

Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2019. Sensebert: Driving some sense into bert.

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

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

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*.

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

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*.

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.

Manfred Pinkal. 1985. *Logik Und Lexikon: Die Semantik des Unbestimmten*. De Gruyter.

Massimo Poesio. Forthcoming. Ambiguity. In Daniel Gutzmann, Lisa Matthewson, Cécile Meier, Hotze Rullmann, and Thomas Ede Zimmermann, editors, *The Blackwell Companion to Semantics*. Wiley.

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.

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, Rihana S. Williams, and Martin J. 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.

Appendices

A Analysis of ELMo Embeddings

In order to determine what factors would have to be taken into account when generating test samples for analysing word sense differences through their embedding, we ran a series of preliminary experiments comparing the embeddings of target words in different context settings. We conducted these experiments using ELMo embeddings obtained by accessing specific target words from sentence embeddings created based on these different context conditions. Using canonical co-predication examples at first, we quickly realised that the position and function of the target word within the sentence has a significant effect on the resulting embedding and thus potentially overshadows any effects caused by sense shifting. To control for this, we next created a set of sample sentences that fixed the position and function of the target word and generated four levels of context: 1) the absolute minimal context to invoke a certain sense, 2) compact context, 3) extensive but descriptive context, 4) extensive, natural context. Using polyseme *newspaper*, with senses a) *physical*, b) *information*, and c) *organisation* we generated the following samples according to these guidelines:

1a The newspaper is folded.
1b The newspaper is boring.
1c The newspaper is famous.
2a The newspaper is lying on the table.
2b The newspaper is listing job openings.
2c The newspaper is struggling financially.
3a The newspaper is made up of 40 sheets of thin, recycled paper, has three columns of text and only a few colour images.
3b The newspaper contains reports on national and international incidents, the daily weather report and sports results.
3c The newspaper fired its editor in chief after her new business strategy caused the company to lose important partners.
4a The newspaper got wet from the sprinklers because the paper boy hadn't thrown it far enough to reach the front porch.
4b The newspaper wasn't very interesting but got the local obituaries and job offers which were read by almost everyone.
4c The newspaper was attacked over its populist coverage of the recent events surrounding the general election in May.

We then calculated the cosine similarities between the embeddings of the target word *newspaper* for all sentence pairs using the LSTM's first layer's hidden state, the LSTM's second layer's hidden state and the ELMo output. See Figure 5 for results. The embeddings of sample sentences 1-6 seem to form a cluster of high similarity compared to the rest of the pairwise comparisons in all of the embedding layers. It thus seems that the extensive context of samples 7-12 causes the target word embeddings to be noticeably different from those of the short context samples. As we aim to analyse the differences between the different senses of a target word and solely need context to invoke these different senses, we propose to keep the context in the test samples for our experiments as short and descriptive as possible.

To determine which output layer provides the most sensitivity to word sense, we calculated the similarity of each sense cluster's mean to the other cluster mean vectors to establish the overall distances between embedding vectors of different senses, i.e. the amount of variance in the outputs. We propose that if this variance is higher, the embeddings are easier to differentiate and different senses therefore might be identified more easily. The result of this experiment is shown in Figure 6, revealing the the second layer's hidden state exhibits the largest differences among the three sense cluster means. We therefore decided to use this embedding layer output for our experiments.

B Sample Creation

Instead of creating sentence pairs by combining every sample sentence for a given polyseme with every other one, we decided to create two pairs for every cross-sense sentence condition only and just one pair for every same-sense condition. This was done by combining the selected first sample sentence for every sense (a) with the selected second sample sentence for every sense (b). For a polyseme with two senses, this results in the four sentence pairs

1a - 1b (ID=11)	2a - 2b (ID=21)
1a - 2b (ID=12)	2a - 1b (ID=22)

By analogy, a polyseme with three distinct senses generates nine samples:

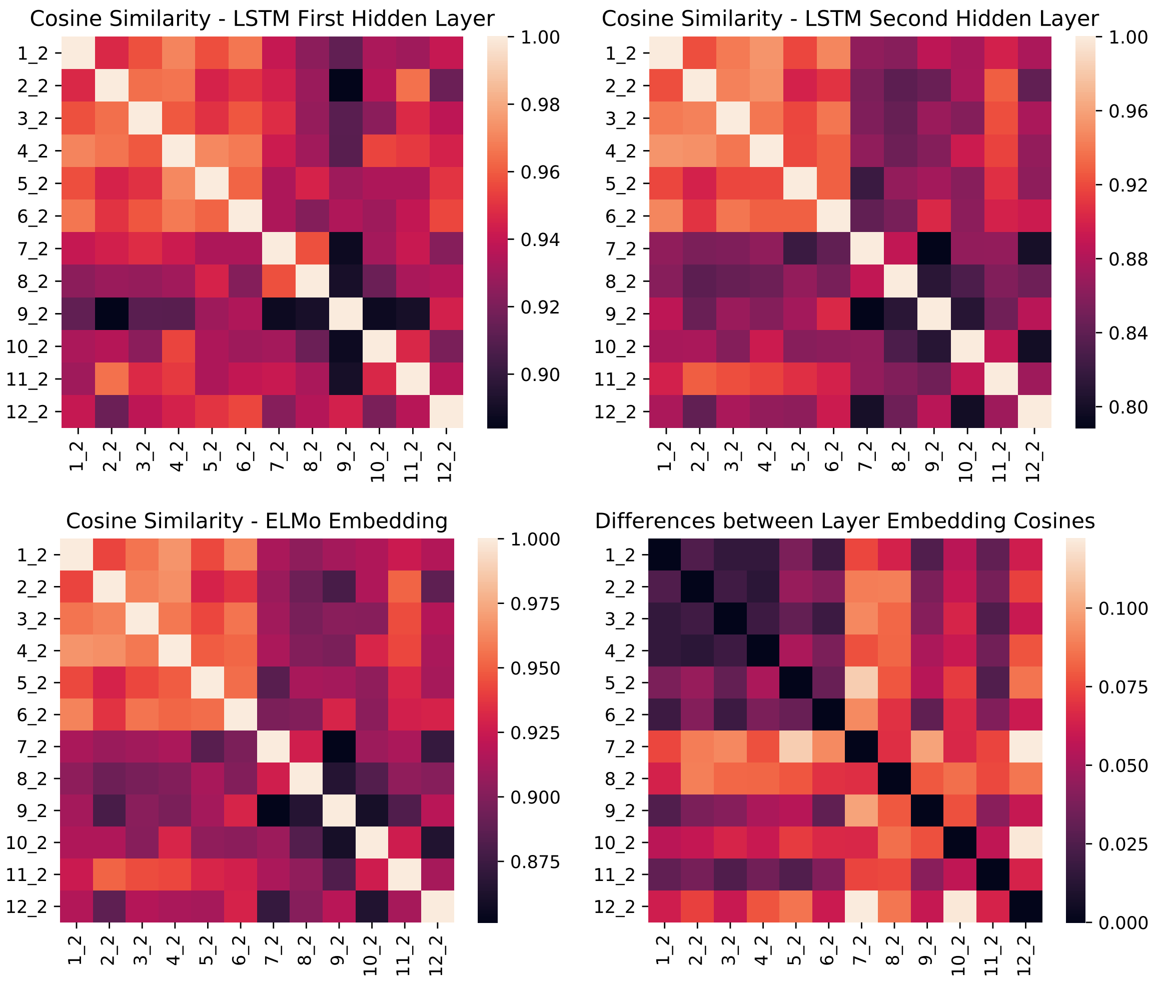


Figure 5: Heat maps of the pairwise cosine similarity of target word embeddings using a given ELMo layer, and a heat map of the differences in cosine similarity between the first and second LSTM layers' hidden state representation

1a - 1b (ID=11)
1a - 2b (ID=12)
1a - 3b (ID=13)
2a - 2b (ID=22)
2a - 1b (ID=21)

2a - 3b (ID=23)
3a - 3b (ID=33)
3a - 1b (ID=31)
3a - 2b (ID=48)

We leave it to the reader to apply this system to generate the 16 pairs for polysemes with four senses. Note that this procedure creates cross-sense pairs with each of the two senses being the first one in the pair once.

C AMT Interface

A screenshot of the AMT user interface can be found in Figure 7.

D Word Sense Similarity Graphs

Graphs of the word sense similarity judgements for the ten regular polysemes tested can be found in Figure 8.

E Comparison of Similarity Ratings

Graphs of the correlation between human word sense similarity judgements and ELMo and BERT embeddings for the ten polysemes tested in this study can be found in Figure 9.

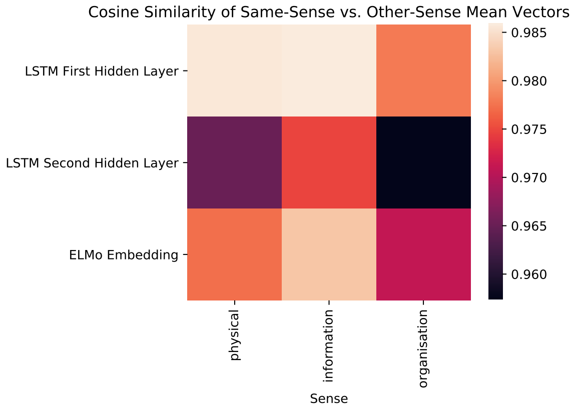
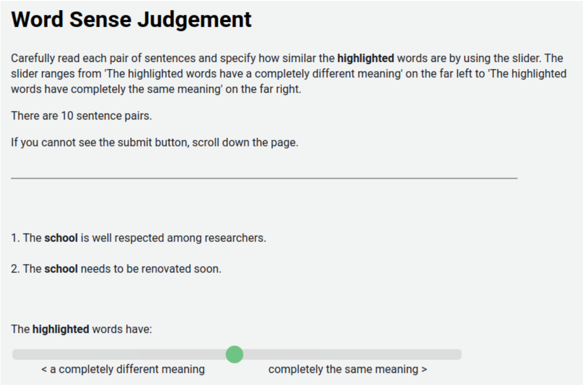


Figure 6: Cosine similarities of sense cluster means to the other senses' means - a measure of overall sense embedding differences in the ELMo layers.



Word Sense Judgement

Carefully read each pair of sentences and specify how similar the **highlighted** words are by using the slider. The slider ranges from 'The highlighted words have a completely different meaning' on the far left to 'The highlighted words have completely the same meaning' on the far right.

There are 10 sentence pairs.

If you cannot see the submit button, scroll down the page.

1. The **school** is well respected among researchers.

2. The **school** needs to be renovated soon.

The **highlighted** words have:

< a completely different meaning
completely the same meaning >

Figure 7: Screenshot of the Amazon Mechanical Turk (AMT) user interface designed to collect the word sense similarity judgements.

F Sense Clustering Samples

The newspaper fired its editor in chief.
The newspaper struggles financially.
The newspaper hired a new designer.
The newspaper was sued for defamation.
The newspaper has around 150 employees.
The newspaper has large coffee stains.
The newspaper lies on the kitchen table.
The newspaper got wet from the rain.
The newspaper weighs less than yesterday.
The newspaper fell behind the counter.
The newspaper contains advertisements.
The newspaper listed all affected stores.
The newspaper has a large sports section.
The newspaper wasn't very interesting.
The newspaper is rather satirical today.

G Full Sample List

Using the procedure described in 2 and Appendix B, the following sentence pairs were created:

Newspaper

11: newspaper: organisation/organisation,
The newspaper fired its editor in chief.,
The newspaper was sued for defamation.
12: newspaper: organisation/physical,
The newspaper fired its editor in chief.,
The newspaper got wet from the rain.
13: newspaper: organisation/information,
The newspaper fired its editor in chief.,
The newspaper is rather satirical today.
21: newspaper: physical/organisation,
The newspaper lies on the kitchen table.,
The newspaper was sued for defamation.
22: newspaper: physical/physical,
The newspaper lies on the kitchen table.,
The newspaper got wet from the rain.
23: newspaper: physical/information,
The newspaper lies on the kitchen table.,
The newspaper is rather satirical today.
31: newspaper: information/organisation,
The newspaper wasn't very interesting.,
The newspaper was sued for defamation.
32: newspaper: information/physical,
The newspaper wasn't very interesting.,
The newspaper got wet from the rain.
33: newspaper: information/information,
The newspaper wasn't very interesting.,
The newspaper is rather satirical today.

Hemingway

11: Hemingway: person/person,
Hemingway was born in Illinois.,
Hemingway won a Nobel prize.
12: Hemingway: person/work,
Hemingway was born in Illinois.,
Hemingway is not suitable for children.
21: Hemingway: work/person,
Hemingway is still widely read today.,
Hemingway won a Nobel prize.
22: Hemingway: work/work,
Hemingway is still widely read today.,
Hemingway is not suitable for children.

War and Peace

11: War and Peace: work/work,
War and Peace was finally published in 1869.,
War and Peace won a range of international awards.
12: War and Peace: work/content,
War and Peace was finally published in 1869.,
War and Peace describes a number of historic battles.
13: War and Peace: work/physical,
War and Peace was finally published in 1869.,
War and Peace is bound in black embossed leather.
21: War and Peace: content/work,
War and Peace chronicles the period of 1805 to 1820.,
War and Peace won a range of international awards.
22: War and Peace: content/content,
War and Peace chronicles the period of 1805 to 1820.,
War and Peace describes a number of historic battles.
23: War and Peace: content/physical,
War and Peace chronicles the period of 1805 to 1820.,
War and Peace is bound in black embossed leather.
31: War and Peace: physical/work,
War and Peace gathers dust on the top shelf.,
War and Peace won a range of international awards.

32: War and Peace: physical/content,
War and Peace gathers dust on the top shelf.,
War and Peace describes a number of historic battles.
33: War and Peace: physical/physical,
War and Peace gathers dust on the top shelf.,
War and Peace is bound in black embossed leather.

Lunch
11: lunch: food/food,
Lunch was exceptionally delicious today.,
Lunch got cold while we waited for someone.
12: lunch: food/event,
Lunch was exceptionally delicious today.,
Lunch is great for socialising and networking.
21: lunch: event/food,
Lunch took more than an hour yesterday.,
Lunch got cold while we waited for someone.
22: lunch: event/event,
Lunch took more than an hour yesterday.,
Lunch is great for socialising and networking.

Door
11: door: physical/physical,
The door was turned into a table top.,
The door splintered when they hit it.
12: door: physical/aperture,
The door was turned into a table top.,
The door connects the two rooms.
21: door: aperture/physical,
The door leads to a long hallway.,
The door splintered when they hit it.
22: door: aperture/aperture,
The door leads to a long hallway.,
The door connects the two rooms.

DVD
11: DVD: physical/physical,
The DVD has some scratches but looks fine.,
The DVD got stuck in the player yesterday.
12: DVD: physical/content,
The DVD has some scratches but looks fine.,
The DVD wasn't very entertaining somehow.
13: DVD: physical/medium,
The DVD has some scratches but looks fine.,
The DVD has won the battle against VHR.
21: DVD: content/physical,
The DVD is a low resolution home movie.,
The DVD got stuck in the player yesterday.
22: DVD: content/content,
The DVD is a low resolution home movie.,
The DVD wasn't very entertaining somehow.
23: DVD: content/medium,
The DVD is a low resolution home movie.,
The DVD has won the battle against VHR.
31: DVD: medium/physical,
The DVD will be replaced by BluRay soon.,
The DVD got stuck in the player yesterday.
32: DVD: medium/content,
The DVD will be replaced by BluRay soon.,
The DVD wasn't very entertaining somehow.
33: DVD: medium/medium,
The DVD will be replaced by BluRay soon.,
The DVD has won the battle against VHR.

School
11: school: building/building,
The school was painted during the holidays.,
The school needs to be renovated soon.
12: school: building/administration,
The school was painted during the holidays.,
The school informed parents about this year's events.
13: school: building/institution,
The school was painted during the holidays.,
The school recently got a more modern website.
14: school: building/students,
The school was painted during the holidays.,
The school went on a field trip last summer.
21: school: administration/building,
The school requires students to wear a uniform.,
The school needs to be renovated soon.
22: school: administration/administration,
The school requires students to wear a uniform.,
The school informed parents about this year's events.
23: school: administration/institution,
The school requires students to wear a uniform.,
The school recently got a more modern website.
24: school: administration/students,
The school requires students to wear a uniform.,
The school went on a field trip last summer.
31: school: institution/building,
The school is well respected among researchers.,
The school needs to be renovated soon.
32: school: institution/administration,
The school is well respected among researchers.,
The school informed parents about this year's events.
33: school: institution/institution,
The school is well respected among researchers.,
The school recently got a more modern website.
34: school: institution/students,
The school is well respected among researchers.,
The school went on a field trip last summer.
41: school: students/building,
The school developed an important algebraic proof.,
The school needs to be renovated soon.
42: school: students/administration,
The school developed an important algebraic proof.,
The school informed parents about this year's events.
43: school: students/institution,
The school developed an important algebraic proof.,
The school recently got a more modern website.
44: school: students/students,
The school developed an important algebraic proof.,
The school went on a field trip last summer.

Wine
11: wine: container/container,
The wine lay in a padded wooden box.,
The wine is a little dusty from storage.
12: wine: container/content,
The wine lay in a padded wooden box.,
The wine tastes great with fish.
21: wine: content/container,
The wine had a beautiful red tint.,
The wine is a little dusty from storage.
22: wine: content/content,
The wine had a beautiful red tint.,
The wine tastes great with fish.

Glass
11: glass: container/container,
The glass broke when she dropped it.,
The glass fits about 200 ml of liquid.
12: glass: container/content,
The glass broke when she dropped it.,
The glass was absolutely refreshing.
21: glass: content/container,
The glass had a thick layer of foam.,
The glass fits about 200 ml of liquid.
22: glass: content/content,
The glass had a thick layer of foam.,

The glass was absolutely refreshing.

Construction

11: construction: process/process,
The construction took far longer than expected.,
The construction will begin in early September.
12: construction: process/product,
The construction took far longer than expected.,
The construction is larger than most in the city.
21: construction: product/process,
The construction has a solid steel frame.,
The construction will begin in early September.
22: construction: product/product,
The construction has a solid steel frame.,
The construction is larger than most in the city.

Homonyms

0: Homonym: bat,
The bat came in through the open window.,
The bat broke when he hit the fence with it.
1: Homonym: match,
The match burned my fingers.,
The match ended without a winner.
2: Homonym: club,
The club only admits women older than 50.,
The club felt very heavy and unwieldy.
3: Homonym: bank,
The bank was washed out by the current.,
The bank increased the interest rate.
4: Homonym: mole,
The mole dug tunnels all throughout the garden.,
The mole needs to be removed as it is cancerous.
5: Homonym: pitcher,
The pitcher threw a number of perfect curveballs.,
The pitcher broke when the waiter dropped it.
6: Homonym: rocket,
The rocket left the atmosphere at 2AM tonight.,
The rocket was bitter taste and ruined the pizza.
7: Homonym: tank,
The tank could easily fit 500 litres of water.,
The tank could easily shoot further than 3 miles.
8: Homonym: watch,
The watch slipped off his hand while he was swimming.,
The watch reported troop movements on the south border.
9: Homonym: yard,
The yard equals exactly three feet.,
The yard is just over 10 feet wide.
10: Homonym: stall,
The stall barely fit the large bull.,
The stall didn't have any toilet paper.
11: Homonym: spring,
The spring in the garden feeds the little pond with fresh water.,
The spring in the ballpen lets you open it with a simple click.
12: Homonym: mine,
The mine had to close after the accident.,
The mine could be defused by an expert.
13: Homonym: order,
The order welcomed the new members.,
The order was shipped two weeks late.
14: Homonym: jumper,
The jumper broke a long-standing record.,
The jumper didn't really fit her that well.

Synonyms

0: Synonym: answer/reply,
The answer came after more than a month.,
The reply arrived within a couple of minutes.
1: Synonym: street/road,
The street leads to a small town in the mountains.,
The road ends at a beautiful hut made from wood.
2: Synonym: world/planet,
The world is heating up because of CO2 emissions.,
The planet is heading towards a serious climate crisis.
3: Synonym: computer/PC,
The computer suddenly turned off.,
The PC needs to be replaced soon.
4: Synonym: problem/issue,
The problem was solved by replacing a cable.,
The issue couldn't be resolved without tools.
5: Synonym: capability/ability,
The capability of modern computers is astonishing.,
The ability to read and write is crucially important.
6: Synonym: area/space,
The area was roped off by the police.,
The space was littered with rubbish.
7: Synonym: audience/crowd,
The audience was very quiet during the concert.,
The crowd was cheering on the football team.
8: Synonym: note/memo,
The note on the fridge read "clean me!".,
The memo simply said "Meeting at 1PM".
9: Synonym: advice/tip,
The advice wasn't very good.,
The tip helped to fix the TV. 10: Synonym: photo/image,
The photo was of a picturesque lake.,
The image shows a red muscle car.
11: Synonym: building/structure,
The building burned down last week.,
The structure collapsed years ago.
12: Synonym: company/organisation,
The company had to find a new office building.,
The organisation expanded to Eastern Europe.
13: Synonym: plank/board,
The plank was torn out of the floor.,
The board covered up a crack in the wall.
14: Synonym: sea/ocean,
The sea was much colder than the beach.,
The ocean looked beautiful in the sunset.

Fillers

11: Filler: bottle/The Guardian,
The bottle fell off the kitchen counter.,
The Guardian contains advertisements.
12: Filler: War of the Worlds/The Guardian,
War of the Worlds is made up of five consecutive parts.,
The Guardian is rather satirical today.
13: Filler: Dickens/university,
Dickens was born in Portsmouth.,
The university was closed during the holidays.
14: Filler: CD/Dinner,
The CD broke when I accidentally sat on it.,
Dinner got cold while we waited for someone.
15: Filler: Dickens/CD,
Dickens didn't really grip me.,
The CD sparked discussions about copyright laws.
16: Filler: War of the Worlds/The Guardian,
War of the Worlds is made up of five consecutive parts.,
The Guardian wasn't very interesting.
17: Filler: Dinner/War of the Worlds,
Dinner was moved to 7:00 PM earlier today.,
War of the Worlds only took a few months to be completed.
18: Filler: The Guardian/beer,
The Guardian has around 150 employees.,
The beer is a little dusty from storage.
19: Filler: The Guardian/War of the Worlds,
The Guardian hired a new designer.,
War of the Worlds is an expensive,

signed first edition.
20: Filler: The Guardian/War of the Worlds,
The Guardian contains advertisements.,
War of the Worlds won a range of international awards.
21: Filler: War of the Worlds/beer,
War of the Worlds just wouldn't fit on the new shelves.,
The beer had a hand-drawn label.
22: Filler: The Guardian/War of the Worlds,
The Guardian fired its editor in chief.,
War of the Worlds only took a few months to be completed.
23: Filler: War of the Worlds/record,
War of the Worlds is an expensive,
signed first edition.,
The record contained times and dates.
24: Filler: The Guardian/bottle,
The Guardian fired its editor in chief.,
The bottle had a hand-drawn label.
25: Filler: Dickens/The Guardian,
Dickens is full of satire and caricature.,
The Guardian listed all affected stores.
26: Filler: The Guardian/Dinner,
The Guardian wasn't very interesting.,
Dinner was moved to 7:00 PM earlier today.
27: Filler: The Guardian/War of the Worlds,
The Guardian has around 150 employees.,
War of the Worlds was first published in 1898.
28: Filler: The Guardian/university,
The Guardian was sued for defamation.,
The university was closed during the holidays.
29: Filler: Dickens/milk,
Dickens advocates Children's rights.,
The milk had a red cow on the label.
30: Filler: picture/War of the Worlds,
The picture was propped up on the mantelpiece.,
War of the Worlds was used to weigh down the mail.
31: Filler: Dickens/The Guardian,
Dickens grew up very poor.,
The Guardian hired a new designer.
32: Filler: Dinner/War of the Worlds,
Dinner was exceptionally delicious today.,
War of the Worlds describes an alien attack on Earth.
33: Filler: War of the Worlds/Dinner,
War of the Worlds only took a few months to be completed.,
Dinner was held in a restaurant in London.
34: Filler: Dinner/War of the Worlds,
Dinner was so spicy that it made me cry.,
War of the Worlds is an expensive,
signed first edition.
35: Filler: The Guardian/War of the Worlds,
The Guardian has around 150 employees.,
War of the Worlds is made up of five consecutive parts.
36: Filler: The Guardian/War of the Worlds,
The Guardian listed all affected stores.,
War of the Worlds is bound in black embossed leather.
37: Filler: hatch/Dinner,
The hatch leads to a long tunnel.,
Dinner is great for socialising and networking.
38: Filler: War of the Worlds/bottle,
War of the Worlds was used to weigh down the mail.,
The bottle had a hand-drawn label.
39: Filler: The Guardian/university,
The Guardian struggles financially.,
The university went on a field trip last summer.
40: Filler: War of the Worlds/Dinner,
War of the Worlds was first published in 1898.,
Dinner was exceptionally delicious today.
41: Filler: Dinner/The Guardian,
Dinner was hastily devoured before the meeting.,
The Guardian is rather satirical today.
42: Filler: The Guardian/Dickens,
The Guardian hired a new designer.,
Dickens is about social equality.
43: Filler: university/War of the Worlds,
The university recently got a more modern website.,
War of the Worlds won a range of international awards.
44: Filler: The Guardian/Dickens,
The Guardian contains advertisements.,
Dickens didn't really grip me.
48: Filler: The Guardian/Dickens,
The Guardian struggles financially.,
Dickens didn't really grip me.
46: Filler: beer/The Guardian,
The beer has a rich golden tint.,
The Guardian wasn't very interesting.
47: Filler: The Guardian/War of the Worlds,
The Guardian wasn't very interesting.,
War of the Worlds was adapted as a movie multiple times.
48: Filler: The Guardian/Dickens,
The Guardian fired its editor in chief.,
Dickens advocates Children's rights.
49: Filler: milk/Dinner,
The milk tastes a little bitter today.,
Dinner got cold while we waited for someone.
50: Filler: bottle/War of the Worlds,
The bottle fell off the kitchen counter.,
War of the Worlds is an expensive,
signed first edition.
51: Filler: The Guardian/War of the Worlds,
The Guardian contains advertisements.,
War of the Worlds describes an alien attack on Earth.
52: Filler: The Guardian/beer,
The Guardian contains advertisements.,
The beer lay in a padded wooden box.
53: Filler: War of the Worlds/bottle,
War of the Worlds gathers dust on the top shelf.,
The bottle has a modern screw-on cap.
54: Filler: War of the Worlds/The Guardian,
War of the Worlds describes an alien attack on Earth.,
The Guardian listed all affected stores.
55: Filler: War of the Worlds/The Guardian,
War of the Worlds is bound in black embossed leather.,
The Guardian struggles financially.
56: Filler: picture/The Guardian,
The picture stands on the living room table.,
The Guardian was sued for defamation.
57: Filler: War of the Worlds/Dinner,
War of the Worlds is an expensive,
signed first edition.,
Dinner was so spicy that it made me cry.
58: Filler: War of the Worlds/picture,
War of the Worlds is made up of five consecutive parts.,
The picture was glued into a photo album.
59: Filler: Dinner/bottle,
Dinner was moved to 7:00 PM earlier today.,
The bottle lay in a padded wooden box.

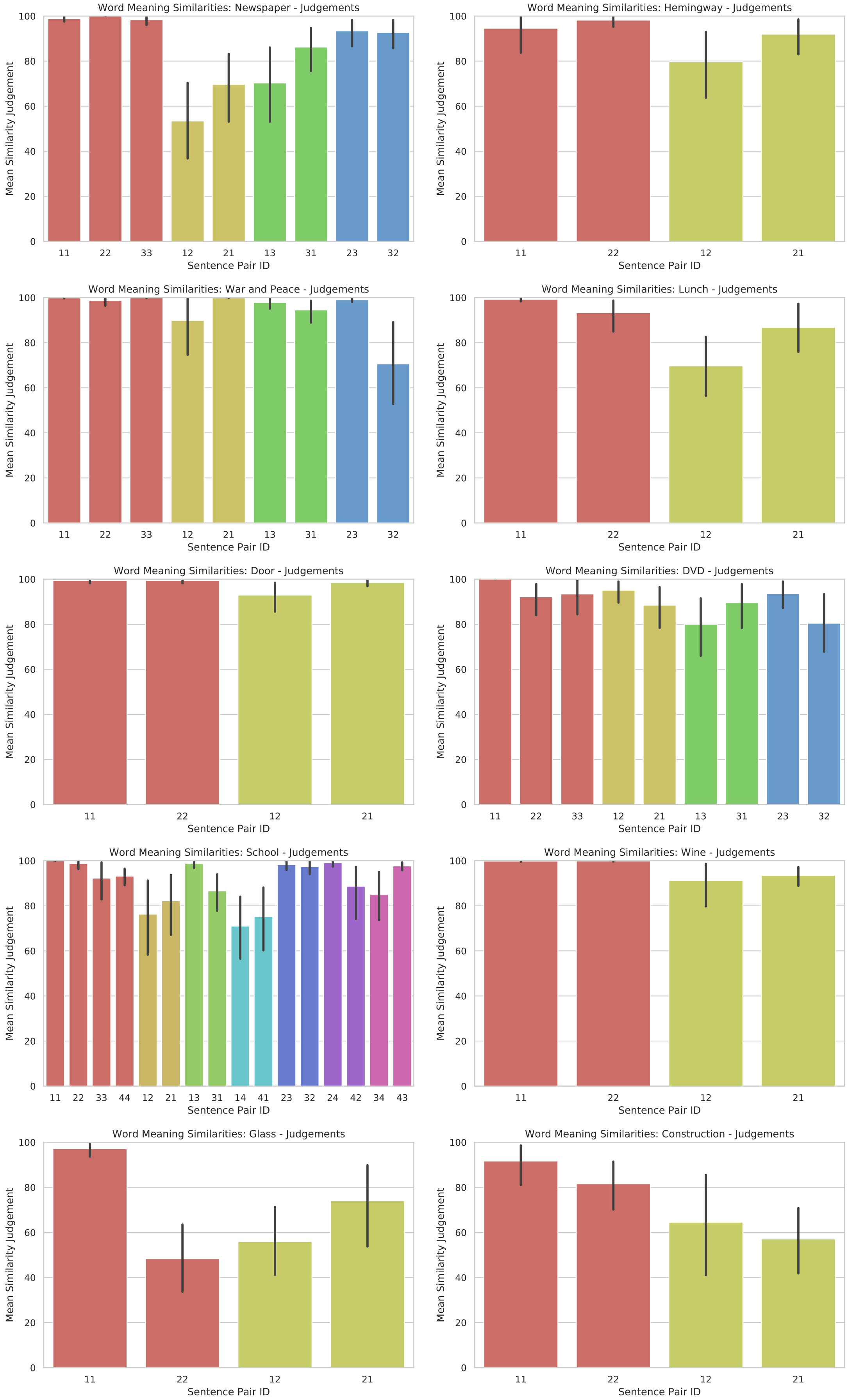


Figure 8: Word sense similarity judgements for the ten tested types of regular polysemy.

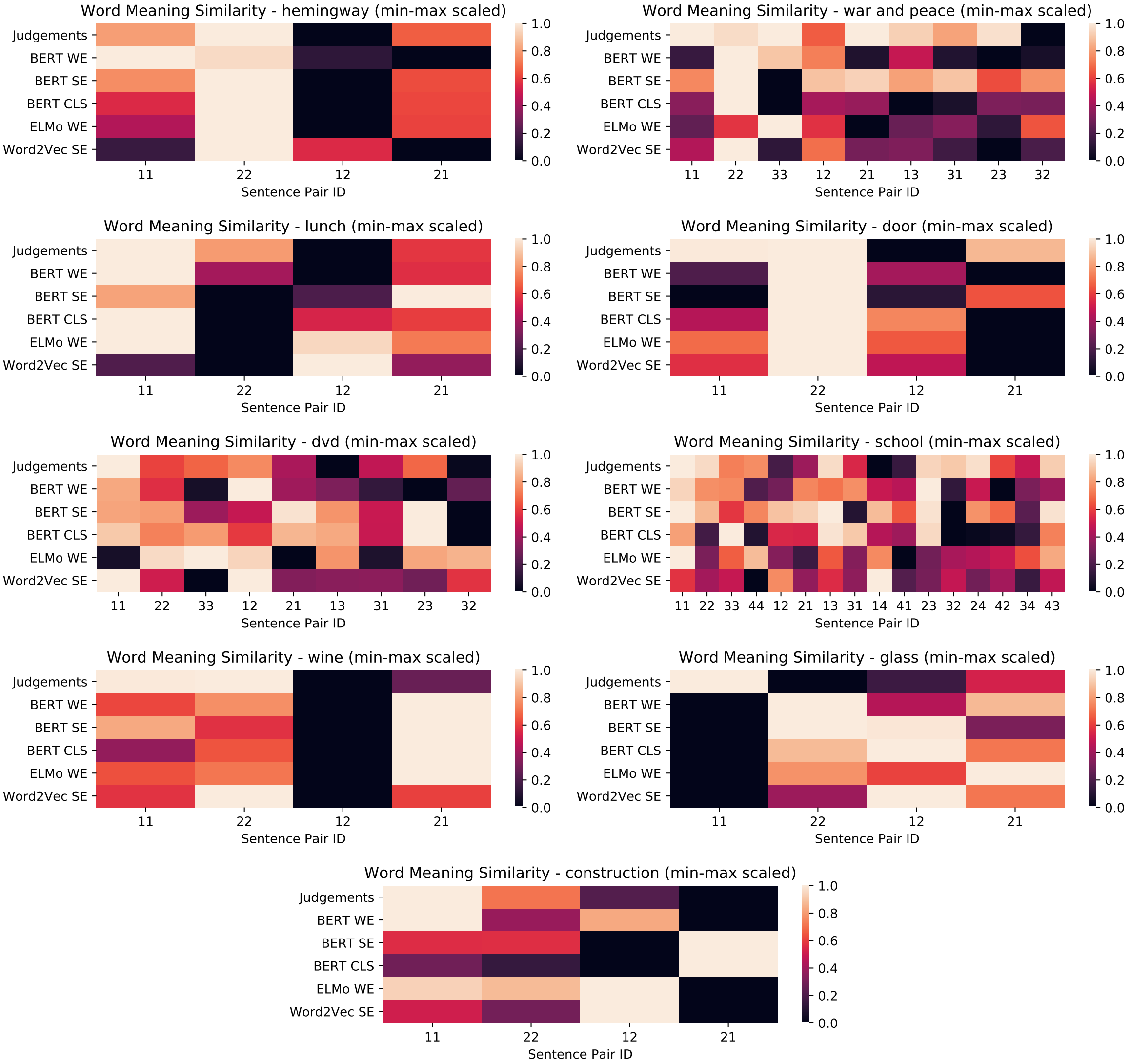


Figure 9: Correlations between human word sense similarity judgements and ELMo and BERT embeddings for the regular polysemes tested in this study.