Polysemy through the lens of psycholinguistic variables: a dataset and an evaluation of static and contextualized language models

Andrea Bruera
Cognition and Plasticity
Max Planck Institute for
Human Cognitive and
Brain Sciences
Leipzig, Germany
bruera@cbs.mpg.de

Farbod Zamani
Department of Computing
Goldsmiths
University of London
United Kingdom

Massimo Poesio
CogSci Research Group
School of Electronic Engineering
and Computer Science
Queen Mary
University of London
United Kingdom

Abstract

Polysemes are words that can have different senses depending on the context of utterance: for instance, 'newspaper' can refer to an organization (as in 'manage the newspaper') or to an object (as in 'open the newspaper'). Contrary to a large body of evidence coming from psycholinguistics, polysemy has been traditionally modelled in NLP by assuming that each sense should be given a separate representation in a lexicon (e.g. WordNet). This led to the current situation, where datasets used to evaluate the ability of computational models of semantics miss crucial details about the representation of polysemes, thus limiting the amount of evidence that can be gained from their use.

In this paper we propose a framework to approach polysemy as a continuous variation in psycholinguistic properties of a word in context. This approach accommodates different sense interpretations, without postulating clear-cut jumps between senses. First we describe a publicly available English dataset that we collected, where polysemes in context (verb-noun phrases) are annotated for their concreteness and body sensory strength. Then, we evaluate static and contextualized language models in their ability to predict the ratings of each polyseme in context, as well as in their ability to capture the distinction among senses, revealing and characterizing in an interpretable way the models' flaws.

1 Introduction

The meaning of individual words taken in isolation can look unambiguous. Take for instance the word *book*. If encountered on its own, it evokes the image of an object made of sheets of paper bound together. However, when put in context, such as in the phrase 'explain the book', it clearly does not refer to that same concrete object - rather, it denotes its immaterial, abstract content. A word like *book* is called a polyseme (Falkum and Benito, 2015; Vicente and Falkum, 2017; Haber and Poesio, 2023). Polysemes are easily understood when contrasted with monosemes (words with only one possible interpretation, like *leaf*) and homonyms (words that can that can take two completely unrelated interpretations, like *bat*): polysemes can take different interpretations - also called **senses** - which are related among them and that follow patterns that also apply to other words (so-called **regular polysemy**; Apresjan, 1974). In the case of *book*, for instance, the pattern is an alternation between a concrete object and an abstract meaning, which also characterizes other words like *newspaper* or *painting*.

In computational linguistics and Natural Language Processing (**NLP**), a large body of work has looked at polysemy. Mainly, the aim is that of finding out to what extent the distinctions between different senses can be captured by current models - either with a theoretical focus (Erk and Padó, 2010; Boleda et al., 2012; Del Tredici and Bel, 2015; Lopukhina and Lopukhin, 2016; Garí Soler and Apidianaki, 2021; Haber and Poesio, 2021; Li and Armstrong, 2023) or in applied tasks (word sense disambiguation Navigli, 2009; Bevilacqua et al., 2021; Loureiro et al., 2021 and induction Agirre and Soroa, 2007; Manandhar et al., 2010; Lau et al., 2012; Eyal et al., 2022). However, as pointed out in McCarthy et al. (2016); Haber and Poesio (2023), a fundamental conceptual limitation has characterized approaches to polysemy in NLP so far. Namely, they have (almost) exclusively assumed a traditional view of polysemy, the so-called **sense enumeration view** (Katz and Fodor, 1963), which has been shown to afford only limited explanatory power. According to this theory, each sense of a polysemous word like *book* should be given a separate, dedicated representation – like the meanings of distinct words like *leaf* and *curtain*. This is the way in which knowledge graphs like WordNet (Miller, 1995) or BabelNet (Navigli and Ponzetto, 2012), the resources that are most typ-

*Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 35 - 48

ically used as the golden standard for polysemy in NLP, are structured: for *book*, we find multiple entries - e.g. <*noun.communication*> and <*noun.artifact*>. However, this view is challenged from a large body of work in cognitive psychology and psycholinguistics. Experimental approaches have rather proposed the so-called **one representation view** of polysemous nouns: different senses are not assumed not to be represented differently, but just to be different aspects or facets of the same semantic representation (among others, Klepousniotou, 2002; Rodd et al., 2004; Schumacher, 2013; see Falkum and Benito, 2015; Haber and Poesio, 2023 for comprehensive reviews).

As a reflection of this theoretical gap, the datasets typically used for the evaluation of computational models of language at capturing polysemy are built according the sense enumeration view. Lack of diverse evaluation approaches not only leaves a large amount of potential evidence untapped, but also obscures important insights that could emerge by taking a different perspective.

We concur with McCarthy et al. (2016); Haber and Poesio (2023) that, to investigate in depth the ability of current computational models of semantic to capture polysemy, it is necessary to go beyond the sense enumeration view. To this aim, we propose to take a hybrid approach. We break down regularized patterns of polysemy – from the sense enumeration view – in terms of psycholinguistic variables like concreteness – inspired by the one representation view. In this framework, the variation happening when varying the interpretation of *book* from <*noun.artifact*> to <*noun.communication*> can be captured by observing that the second is interpreted as a less concrete entity – which can be further characterized as a reduction in manipulability (touch) and readability (sight), possibly accompanied by an increase in its audibility (hearing). We build on previous work showing how hard distinctions between senses emerge from (and are contained by) complex representations of words (Pustejovsky, 1991; Cruse, 1995; Ortega-Andrés and Vicente, 2019). What we add is an explicit specification (i.e. in terms of psycholinguistic variables) of how sense alternations in polysemy take place. From previous approaches in NLP that rely on similarities in latent vector spaces (Boleda et al., 2012; McCarthy et al., 2016; Haber and Poesio, 2021), we retain the notion of using continuous measures of similarity/distance – i.e. a 'soft' approach to senses: however, while dimensions of language are not interpretable from a cognitive point of view, ours are. Importantly, this framework has been previously successfully applied to model how the brain processes fine-grained lexical meaning variations (Bruera et al., 2023). Since our framework revolves around cognitively motivated semantic features, it aims at fostering research connecting computational and cognitive models of language – with the broader goal of allowing to gain insights on how similar the two are, which is a fundamental open question in the field (Antonello and Huth, 2023; Beinborn and Hollenstein, 2023; Golan et al., 2023; Kanwisher et al., 2023).

Starting from this theoretical approach, in the current work we present two main contributions. First, we describe how we created an original dataset of examples of lexical polysemy. For each polyseme, the dataset provides ratings provided by human subjects in terms of concreteness and of sensory strength (with separate ratings for sight, hearing, touch, smell, taste) for phrases where the different senses are evoked. Our dataset is carefully crafted by controlling for psycholinguistic variables, with the aim of allowing its use both for *in silico* and cognitive experiments.

Secondly, we evaluate static and contextualized language models on their ability to predict the ratings provided by humans and to distinguish among different senses of polysemous words. We hypothesized that contextualized language models would consistently outperform static language models. Our results confirm our prediction, but they also show that there is large room for improvement in overall accuracy for contextualized language models too - indicating that polysemy is still a challenging semantic phenomenon for language models to capture.

We publish the dataset together with the code[1].

2 Data

2.1 Overview of the dataset

We select a set of 25 polysemic nouns admitting both an abstract and a concrete interpretation. Then, for each noun we select two verbs that, when combined with the noun in a verb-noun phrase, give rise to an abstract (e.g. 'explain the book', 'describe the picture', 'know the medicine') interpretation and two that evoke a concrete (e.g. 'open the book', 'carry the picture', 'swallow the medicine') read-

[1]they can be found at this link: https://osf.io/nfcuq/?view_only=9c7137bc88d543dbaaa17225cbfdef34

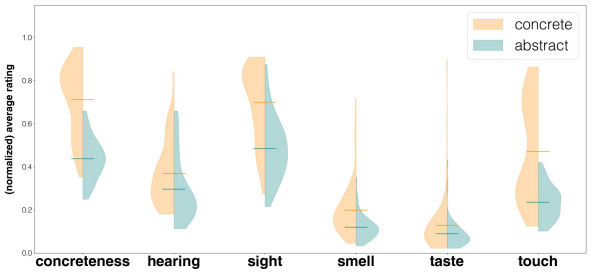

Figure 1: **Distribution of concreteness and sensory strength ratings for the 100 verb-noun polysemic phrases**. Ratings (y axis) are normalized in the range 0-1. As shown by the averages (horizontal coloured lines), concrete phrases show higher concreteness and stronger involvement of all types of sensory information.

ing of the noun. This is the process of so-called 'sense coercion' (Pustejovsky, 1991; Lauwers and Willems, 2011) or 'sense selection', where verbs make the interpretation of the noun go towards one sense or the other (Klepousniotou, 2002). In this way, phrases are equally divided into two mirrored sets of abstract and concrete senses.

Finally, we collect a set of psycholinguistic ratings for all of the nouns within each phrase. We collect ratings for concreteness – the most relevant cognitive dimension –and for the five body senses, since sensory strength can better characterize variation in meaning than simple concreteness (Lynott et al., 2020).

The main aim of this dataset is to fill a gap in existing resources that can be used to evaluate NLP models with respect to polysemy. Our hope is also to foster further research along these lines, with a strong focus on cognitive evaluation of computational models of semantics (Beinborn and Hollenstein, 2023). Therefore, we wanted our stimuli selection to be valid for further testing involving the collection of behavioural and brain data. In such studies, it is fundamental to control for experimental confounds which are not relevant for NLP models, but play an important role in human cognition. Such confounds can be related to non-semantic, low-level sensory properties of the stimuli (Hauk and Pulvermüller, 2004; Laszlo and Federmeier, 2014; Dufau et al., 2015) or, within semantics, to emotional processing (Kuperman et al., 2014; Hinojosa et al., 2020).

In the following we will describe the stimuli selection procedure in detail. A visualization of the distributions of the ratings, directly comparing abstract and concrete senses, is displayed in Figure 1.

2.2 Stimuli selection

2.2.1 Nouns

We selected the set of 25 polysemous nouns to be used among the polysemes annotated in CoreLex (Buitelaar, 1998). CoreLex is an annotation made on top of WordNet (Miller, 1995) specifically created for polysemy. In CoreLex, a number of polysemous nouns from WordNet are annotated according to their polysemy pattern - e.g. annotating with the same label all words that behave similarly to 'book'. For our purpose, the advantage of the annotation provided by CoreLex is that it allows to automatically isolate cases of polysemy where an alternation of a concrete and an abstract sense is present (cf. Boleda et al., 2012).

To extract the nouns, we therefore first looked at the types of nouns present in CoreLex (e.g. 'art'='artifact' or 'com'='informational content'; so-called 'Corelex basic types'). We annotated them according to whether they referred to 'concrete', 'abstract' or 'other' entities (where 'art'='concrete', 'com'='abstract'). From this list, we moved to the list of the polysemy classes ('CoreLex classes'), retaining only the classes where an alternation of an abstract and a concrete sense was present (e.g. a CoreLex class like 'cae', where both a 'art' and a 'com' sense are found). Finally, we chose our candidate nouns by taking the nouns which were annotated in CoreLex as instances of the selected polysemous classes - like 'book', which is a case of the CoreLex class 'cae'.

In parallel, we computed word (lemma) frequencies for the selected polysemous nouns from UKWaC (Baroni et al., 2009), a corpus reflecting general internet language use which has been validated as a corpus for psycholinguistic studies in previous work (Mandera et al., 2017). Since most words occurred with very low frequencies in the corpus, we selected as our candidate polysemes only the top 10% most frequent nouns . Among those, we tried to minimize variance in word length, so as to minimize this possible confounding factor which has a strong impact on cognitive processing (Hauk and Pulvermüller, 2004) . Given that word concreteness correlates negatively with word length (Reilly et al., 2017), we had to strike a balance, avoiding short (whose majority would be concrete) and long (overwhelmingly abstract) words. Therefore, we chose as a criterion to consider nouns between six and nine letters in length. This left us with 571 candidate polysemous nouns.

2.2.2 Verbs

Having thus reduced the set of polysemous nouns, we moved on to select the verbs to be used to create the phrases. We applied a procedure inspired by recent work on predicting concreteness from distributional semantics models (Bhaskar et al., 2017). First, we took the 40000 concreteness ratings for English words from (Brysbaert et al., 2014). Then, we filtered this list, considering only words whose most common POS was that of verb. To do so we used a corpus-based measure of POS prevalence provided by the same authors (Brysbaert et al., 2012). Then, to find verbs eliciting the concrete senses of the polysemes, we took the 1000 most concrete verbs; for the abstract senses, we took the 1000 least concrete verbs. We decided, here again, to reduce the variance in word length for the verbs. However, we kept a wider variance range (4-8 letters, extremes included), considering that we could balance length when choosing the final phrases. After this selection step, the number of concrete verbs was 811, and of abstract verbs 571 (incidentally, the same number of nouns retained from CoreLex).

2.2.3 Verb-noun phrases

Then we looked for the selected verbs' frequencies of co-occurrence with the polysemous nouns within the UKWaC corpus. The aim was that of obtaining a measure of the frequency of occurrence of each of the potential verb-noun phrases, so as to balance them for frequency across abstract and concrete senses. To do so, we exploited the POS annotation provided by UKWaC. We adapted the procedure already validated by Bruera et al. (2023) to extract verb-noun phrase mentions from corpora to be used with language models. We thus considered as relevant verb-noun co-occurrences (i.e. mentions of phrases) only cases where the (lemmatized) verb preceded the (lemmatized) polysemous noun, within a window of three words to the right (to be able to consider cases such as "open an old book", where the linear distance in words between the verb and the noun is three). Then, for each polyseme, we retained the 100 abstract and 100 concrete verbs that co-occurred the most with it. Finally, we proceeded to manually select the twenty-five nouns for which we could find clear cases of sense selection for two verbs and two nouns, thus obtaining the final set of 100 stimuli. We adjusted iteratively our choices so the resulting phrases did not differ statistically across abstract and concrete senses along relevant psycholinguistic variables. As statistical tests we used t-tests; reported p-values are not corrected for multiple comparisons - corrected p-values would be even more conservative. All differences among concrete and abstract phrases are not statistically significant. Since the nouns were the same in both conditions (abstract and concrete), for most variables it was enough to look at the verbs - the main exceptions being phrase frequency ($p = 0.952$) and phrase length ($p = 0.79$). Regarding verbs, we checked that no difference in valence ($p = 0.298$), arousal ($p = 0.103$), dominance ($p = 0.769$) was statistically significant, using the norms provided by (Warriner et al., 2013). Additionally, difference in frequency for concrete and abstract verbs is also not significant ($p = 0.0687$). By contrast, statistically significant differences between verbs emerge, as required by design, in concreteness ($p < 0.0001$).

2.3 Concreteness and sensory strength ratings

Given the 100 phrases selected following the procedure reported above, we then collected from 25 human volunteers ratings for concreteness and sensory strength in all of the five body senses. Sensory strength norms capture more precisely what drives the sense alternation in terms of semantic variables (e.g. the case of book can be explained in terms of variation in sight and touch, but no taste is involved). Participants were recruited among the communities of the authors' university departments, which are located in the same anglophone country. We did not require participants to be native speakers of English. Twenty-five (25) subjects, between 18 and 40 years of age, took part as volunteers to the rating experiment after giving their written consent. In the rating experiment, subjects were presented one by one with all of the 100 phrases, and asked to rate on a Likert scale from 1 to 5 how concrete the polysemous noun in that context was, as well as its so called sensory strength (Lynott et al., 2020). Before starting the experiment, participants were provided with an explanation for each variable, taken from previous rating experiments (Scott et al., 2019; Lynott et al., 2020) and with an example.

The distributions of the resulting ratings are reported in Figure 1. As it can be seen, the largest difference between distributions for concrete/abstract senses is found for concreteness, sight and touch (in all cases $p < 0.0001$), followed by hearing ($p = 0.0163$). The difference is also statistically

significant for smell ($p = 0.00012$) and close to significance for taste ($p = 0.083$), however the ratings for the nouns are in both cases always low (averages after normalization: $abstract_{smell} = 0.12, concrete_{smell} = 0.198, abstract_{taste} = 0.088, concrete_{taste} = 0.128$).

We further compute the reliability of the scores provided by the raters. As a measure of inter-rater reliability we use the mean intra-class correlation (**ICC**, Shrout and Fleiss, 1979), which can take a value between 0 (random agreement) and 1 (perfect agreement). This is the recommended choice for cases like ours where multiple raters provide a single non-nominal score for the same set of items (Hallgren, 2012). We treat subjects as random effects, thus we report what is referred to as type 2 ICC, with 25 subjects – in the terminology of Shrout and Fleiss (1979), $ICC(2, k = 25)$. When aggregating all types of scores together (i.e. concreteness and all sensory modalities), $ICC = 0.945$, indicating excellent agreement (the lower threshold for excellence, according to the guidelines of Cicchetti, 1994; Hallgren, 2012, is $ICC > 0.75$). This confirms that the measurements contained in our dataset are reliable. To understand whether reliability is affected by each of the sensory modalities, we further compute the corresponding separate ICC scores. We find that reliability is highest for concreteness ($ICC_{concreteness} = 0.924$), touch ($ICC_{touch} = 0.913$) and sight ($ICC_{sight} = 0.895$. ICCs are slightly lower, but still indicate excellent agreement, for taste $ICC_{taste} = 0.87$, hearing ($ICC_{hearing} = 0.82$) and smell ($ICC_{smell} = 0.789$).

3 Models

A fine-grained semantic phenomenon like polysemy has proven particularly challenging to capture for language models. Older approaches (so-called **static** language models; Bommasani et al., 2020), were particularly unsuited to face its subtleties (Camacho-Collados and Pilehvar, 2018). Static language models learn fixed semantic representations for words, abstracted from specific contexts of usage. This made it hard to successfully model meaning of words in context - and consequently context-dependent phenomena such as polysemy (Schütze, 1998; Yaghoobzadeh and Schütze, 2016). The more recent language models, called **contextualized** language models (Rogers et al., 2021; Min et al., 2023)), should be in principle better equipped to face the challenge of polysemy. They are trained to create semantic representations of words which are context-specific. When focusing broadly on NLP tasks requiring to consider contextual semantic knowledge (e.g. natural language generation, inference, relation classification), contexualized models are clearly able to reach impressive performance, outperforming static models (Lenci et al., 2022). However, when zooming in through the lens of extremely specific semantic knowledge such as polysemy, synonymy, hypernymy and categorization, the picture changes: contextualized models appear to capture such phenomena only to a modest extent, leaving much room for improvement (Ravichander et al., 2020; Haber and Poesio, 2021; Lenci et al., 2022; Haber and Poesio, 2023).

To provide a better picture with regards to this, we use four models, including both static and contextualized language models (Lenci et al., 2022). In the following we will briefly describe each model, and how the vectors for the polysemous nouns in context were extracted from each one of them. In Appendix A we report an analysis measuring how similar the representations are across the models: the phrases that compose our dataset make notable differences emerge across different types of models, converging with our prediction and sense discrimination results (see Sections 5.1, 3, 4).

3.1 Baseline: count-based model

As a baseline model, we use a so-called count model, following previous work on using distributional models predicting concreteness ratings (Bhaskar et al., 2017). We used the same window size used for fasttext (Bojanowski et al., 2017) - therefore we counted word co-occurrences within a sliding window of ten words (five on the left and five on the right of the target word). As training corpus we used UKWaC. To reduce computational effort, we tried to keep vector dimensionality low by reducing the vocabulary size as done in Bhaskar et al. (2017); Charbonnier and Wartena (2019). Therefore, we reduced the vocabulary to the top 20% most frequent words that appeared in the concreteness norms of (Brysbaert et al., 2014), which makes vectors have 5220 dimensions. As is commonplace in the literature, we transform the raw co-occurrence counts using Pointwise-Mutual Information - therefore the model will be referred to as **count-pmi** (Levy et al., 2015).

We modelled the meaning of the polysemous

noun in the phrase by following the procedure validated in Bruera et al. (2023). It consists of adapting the noun's representation to the context by averaging it with the representation for the verb. Averaging was chosen because, despite its simplicity, it has been shown to be a strong baseline to compose the meaning of words both in NLP and in cognitive neuroscience (Dinu et al., 2013; Wu et al., 2022). We first extracted the pre-trained vector representations for each verb and noun present in the set of stimuli. Then, each phrase's vector representation was obtained by averaging the vectors for the verb and the noun.

3.2 fasttext

As a static model, we chose **fasttext**, using the pre-trained version for English, which is publicly available (Bojanowski et al., 2017; Grave et al., 2018). This version was trained on a combination of Common Crawl and Wikipedia and has 300-dimensional vectors. We extract word vectors for all nouns and verbs and create a phrase-specific representation for each noun as described for count-pmi.

3.3 ConceptNet Numberbatch

As discussed above, senses for polysemous are annotated explicitly in graph-based resources like WordNet. In recent years, ways to integrate graph- and vector- based approaches to semantic representation have been devised. To evaluate how the explicit knowledge about senses encoded in graph-based models can help language models, we used ConceptNet Numberbatch (in the following, **numberbatch**; Speer et al., 2017). Numberbatch is a widely used model that combines distributional and graph-based information: it brings together semantic knowledge from ConceptNet, a graph-based resource that includes WordNet annotations, and two word embeddings models (word2vec (Mikolov et al., 2013) and Glove (Pennington et al., 2014)) using the retrofitting procedure (Faruqui and Dyer, 2015). Recently, its performance has been shown to be superior to distributional-only models in modelling cognitive data (Turton et al., 2020; Alacam et al., 2022; Yang et al., 2024). We compose word vectors for the phrase using the same methodology as count-pmi and fasttext; the resulting phrase vectors have 300 dimensions.

3.4 XGLM

As a contextualized language model, we used XGLM, a recently proposed multilingual model (Lin et al., 2021). Since contextualized models are specialized for representation of language in context, and given previous results (Haber and Poesio, 2021; Bruera et al., 2023), we expect that XGLM should in principle provide the best performance at capturing polysemy. XGLM can beat a similarly-sized GPT-3, a monolingual model, at a number of NLP tasks – arguably thanks to the cross-linguistic transfer of semantic information (Lin et al., 2021). Also, it is publicly available and it has been already used in previous experiments with cognitive datasets (De Varda and Marelli, 2023). We experiment with different model sizes (as reported in the Section 5.3) and for the main comparisons we report results using the best layer (7) for the best-performing model, **XGLM-1.7B**.

To extract vectors for the phrases, we use HuggingFace's Transformers library (Wolf et al., 2020). We employed 'representation pooling', a methodology for creating 'static' representations in contextualized language models that was validated in (Bommasani et al., 2020; Vulić et al., 2020; Apidianaki, 2022) for NLP tasks and in (Bruera and Poesio, 2022, 2023; Bruera et al., 2023) for brain data. In our implementation, first we collected from UKWaC all the sentences containing each one of the selected phrases. To do so, we used the procedure described above for counting the frequencies of verb-noun co-occurrences during stimuli selection. Then, we used XGLM to encode all the sentences separately. Having done so, we extracted the hidden layers of the deep neural network, considering the tokens corresponding to the words contained in the phrase. We followed Bruera et al. (2023), where authors found that the best results with a causal language model like XGLM are obtained when considering all of the phrase tokens + 1, thus capturing both the meaning of the verb and the noun. In Section 5.3 we report results using different sizes of XGLM and all the layers. For the analyses reported in Sections 5.1 and 5.2 we use the layer and the model with the best performances (XGLM 1.7B, layer 7). For each mention of the phrase, we averaged vectors across layers and tokens. In this way, we could obtain a single contextualized vector for each phrase mention. Finally, we averaged, for each phrase, ten randomly sampled mention vectors, following (Vulić et al.,

2020). This allowed us to obtain one single vector capturing reliably the information encoded in XGLM for each phrase.

4 Evaluation

Having obtained the vectors for each verb-noun phrase, we measure to what extent it is possible to learn to predict the ratings obtained from human subjects. We use a cross-validated procedure, with a Ridge regression model (α is cross-validated within the train set among $0.01, 0.1, 1, 10, 100, 1000$). We employ a linear model, an efficient choice given the low number of data points (100; Lin et al., 2023). For cross-validation, we use Monte Carlo Cross-Validation (Kim, 2009) - which entails randomly sampling train and test sets many times (in our case, 20), in order to obtain a reliable average statistics. For the evaluation, we use two measures, explained below.

Correlation The first one simply measures the average Pearson correlation between predicted and real values, averaged across all 20 randomized train-test splits (proportion: 80% train - 20% test). This is the metric typically used in similar studies using language models to predict psycholinguistic variables (Bhaskar et al., 2017; Charbonnier and Wartena, 2019; Chersoni et al., 2020).

Sense discrimination The second measure, by contrast, is directly aimed at testing the ability of each language model to distinguish among different senses. It was originally introduced in cognitive neuroscience, to quantify how well a model could distinguish between two brain images referring to two different concepts (Mitchell et al., 2008; Pereira et al., 2018).

It works in the following way. First, as in Bruera et al. (2023), we consider each word and its two senses as a separate test set – consisting of two phrases for each sense. Suppose they are named $a = phr1_{sense1}, b = phr2_{sense1}, p = phr1_{sense2}, q = phr2_{sense2}$. At test time, the desired semantic variable for the four test items is predicted (e.g. for concreteness $\hat{a}_{conc}, \hat{b}_{conc}, \hat{p}_{conc}, \hat{q}_{conc}$). The predicted ratings are then used to quantify, with a binary accuracy metric, how well the model can distinguish between different senses. All possible pairs of phrases belonging to two different senses are taken (i.e. $\{a, p\}$, $\{b, p\}$, $\{a, q\}$, $\{b, q\}$). Intuitively, given a pair (e.g. $\{a, p\}$) we measure if the prediction $\hat{a}_{conc}$ is closer to the real value for its corresponding sense a_{conc} than it is to the other sense p_{conc}; and *vice versa*. If this is the case, then $accuracy = 1$ because the distinction between the two senses has been correctly captured; else, $accuracy = 0$.

More formally, $accuracy = 1$ if $abs(a_{conc} - \hat{a_{conc}}) + abs(p_{conc} - \hat{p_{conc}}) < abs(a_{conc} - \hat{p_{conc}}) + abs(p_{conc} - \hat{a_{conc}})$; else $accuracy = 0$. This evaluation is repeated for all combinations of phrases for the two senses of each word, then averaged; the final evaluation is the average of the scores for all the test sets. This procedure is repeated for all the semantic variables; overall results refer to their average. Since it is a binary accuracy measure, chance performance is at 0.5.

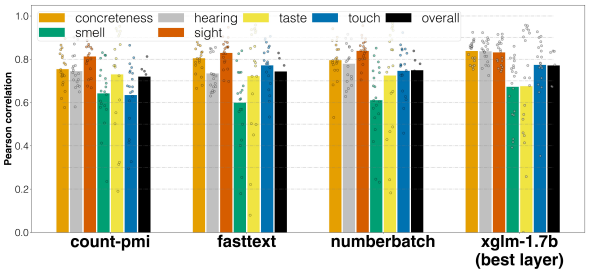

Figure 2: **Pearson correlation between predicted and true variables for each model**. We plot each cross-validation split as a separate scatter point. XGLM consistently provides the best correlation scores across all variables.

5 Results and discussion

5.1 Correlation analysis

In Figure 2 we report the average Pearson correlation between predicted and real ratings. XGLM (best performing layer and version: layer 7 of XGLM-1.7B; see Section 5.3) provides the best performance in all variables except taste ($XGLM_{sight} = 0.839, XGLM_{touch} = 0.774, XGLM_{hearing} = 0.837, XGLM_{smell} = 0.672$; best performance in taste by Conceptnet Numberbatch $numberbatch_{taste} = 0.725$). Overall low performance in taste and smell can be explained by the fact that, as shown in Figure 1, these two sensory variables had the smallest variance overall, and tended to cluster around low values – thus making it difficult to differentiate among values for different phrases.

Despite the superiority of XGLM, however, differences between different models are surprisingly small ($XGLM_{overall} = 0.771, count - pmi_{overall} = 0.72, fasttext_{overall} =$

$0.743, numberbatch_{overall} = 0.749$). This suggests that simpler, more efficient approaches can capture information about polysemy. Importantly, this concurs with the results of Lenci et al. (2022) in showing that even count-based models often can outperform much more complex ones at fine-grained semantic tasks.

The performance of our models are largely comparable to those obtained when predicting single-word semantic variables. For concreteness, Charbonnier and Wartena (2019) report scores for fasttext oscillating among 0.85 and 0.9, depending on the dataset; here fasttext is at 0.804 (the best performance is afforded by XGLM at 0.838). For sensory strength, Chersoni et al. (2020) report overall lower Spearman correlation for fasttext (average across body senses: 0.596) than us (body sensory average for fasttext: 0.731; top performance by XGLM at 0.758). We assume that such differences are due to the fact that our dataset is much smaller than those used for single-words evaluations, that range in the tens of thousands of words, and possibly to the different correlation metrics used (Spearman vs Pearson correlation).

Turning our approach on its head, our results show that it is possible to automatically obtain reliable concreteness and sensory ratings for phrases (an approach that has been recently advocated especially for low resource languages; Turton et al., 2020; Grand et al., 2022; Wang et al., 2023), and use those to *induce* word senses. In other words, our methodology can be used to automatically find in corpora contexts of use where the same polysemous word is used in different senses. This would also allow for an automated large scale expansion of the current dataset .

5.2 Sense discrimination analysis

While correlation scores provide a general evaluation of prediction performance, we separately assess the ability of the four models at discriminating among different senses of polysemous words using the dedicated pairwise evaluation (see above). We also run statistical significance t-tests against the chance baseline of 0.5. Results are reported in Figure 3. XGLM performs better overall ($XGLM_{overall} = 0.672, p = 0.0001; XGLM_{concreteness} = 0.88, p < 0.0001; XGLM_{hearing} = 0.62, p = 0.093; XGLM_{smell} = 0.61, p = 0.156; XGLM_{taste} = 0.35, p = 0.99$), as hypothesized. ConceptNet Numberbatch affords the best results only for sight ($numberbatch_{sight} = 0.81, p = 0.0002; XGLM_{sight} = 0.8, p = 0.0004$. The performance of the contextualized model is always better at capturing polysemy than both purely distributional models (count-pmi and fasttext), confirming previous reports (Haber and Poesio, 2021; Bruera et al., 2023). XGLM can also (in most cases) outperform ConceptNet Numberbatch, which incorporates hand-coded information about senses. This suggests that such fine-grained semantic knowledge can be alternatively captured by looking at linguistic contexts – i.e. at language in use. However, the fact that all models perform significantly above chance for the same variables, the small magnitude of the differences among models, and the rather low average performance taken together suggest that polysemy is still hard to capture.

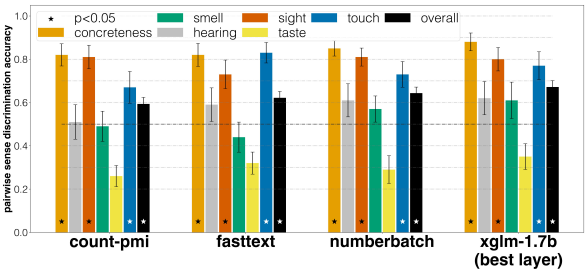

Figure 3: **Sense discrimination scores for each model, using all semantic variables**. Error bars indicate the standard error of the mean across test splits. Overall indicate that the sense discrimination task is challenging for all models.

5.3 In-depth evaluation of XGLM on sense discrimination

In Figure 4 we report the layer-by-layer results for the XGLM family of models (1.7B, 4.5B, 7.5B parameters). We plot overall performance – i.e. the average across all variables. In accordance with previous results on lexical information encoded in contextualized models, performance is better in earlier layers (Bommasani et al., 2020). A relatively small model (1.7B) can provide the best results overall, outperforming both static and larger-sized variants in almost all layers. This converges with previous results casting doubts over the need of ever-larger language models when it comes to modelling human cognition (Oh and Schuler (2023) for reading times, De Varda and Marelli (2023) for eye-tracking, Bruera et al. (2023) for fMRI; cf. Rogers et al., 2021).

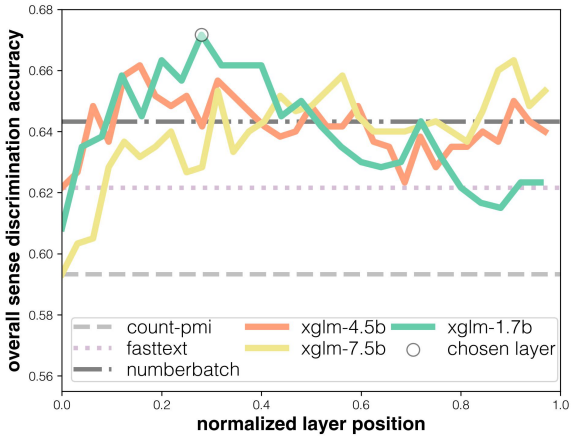

Figure 4: **Overall sense discrimination scores for a number of contextualized models, across all layers**. Overall, all versions of XGLM perform better in the first half of the layers. We indicate with a circle the layer used for the analyses reported above.

6 Limitations and future directions

The main limitation of our study is the size of the dataset, and the fact that we focus on only one case of regular polysemy. Future work could expand this dataset by considering more, and more specific types of polysemy that can be modelled within a similar framework – cases like *chicken* where another variable, taste, can explain sense alternations (animal vs taste; Boleda et al., 2012).

Another interesting direction could be investigating to what extent language models and human cognition align while processing these polysemes (e.g. using brain data; cf. Bruera et al., 2023).

References

Eneko Agirre and Aitor Soroa. 2007. Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the fourth international workshop on semantic evaluations (semeval-2007)*, pages 7–12.

Özge Alacam, Simeon Schüz, Martin Wegrzyn, Johanna Kißler, and Sina Zarrieß. 2022. Exploring semantic spaces for detecting clustering and switching in verbal fluency. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 178–191, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Richard Antonello and Alexander Huth. 2023. Predictive coding or just feature discovery? an alternative account of why language models fit brain data. *Neurobiology of Language*, pages 1–16.

Marianna Apidianaki. 2022. From word types to tokens and back: A survey of approaches to word meaning representation and interpretation. *Computational Linguistics*, pages 1–60.

JD Apresjan. 1974. Regular polysemy. *Linguistics*, 12(142):5–32.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43:209–226.

Lisa Beinborn and Nora Hollenstein. 2023. *Cognitive Plausibility in Natural Language Processing*. Springer Nature.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conference on Artificial Intelligence, Inc.

Sai Abishek Bhaskar, Maximilian Köper, Sabine Schulte Im Walde, and Diego Frassinelli. 2017. Exploring multi-modal text+ image models to distinguish between abstract and concrete nouns. In *Proceedings of the IWCS workshop on Foundations of Situated and Multimodal Communication*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Gemma Boleda, Sebastian Padó, and Jason Utt. 2012. Regular polysemy: A distributional model. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 151–160.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting pretrained contextualized representations via reductions to static embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781.

Andrea Bruera and Massimo Poesio. 2022. Exploring the representations of individual entities in the brain combining eeg and distributional semantics. *Frontiers in Artificial Intelligence*, 5:796793.

Andrea Bruera and Massimo Poesio. 2023. Family lexicon: using language models to encode memories of personally familiar and famous people and places in the brain. *bioRxiv*, pages 2023–08.

Andrea Bruera, Yuan Tao, Andrew Anderson, Derya Çokal, Janosch Haber, and Massimo Poesio. 2023. Modeling brain representations of words' concreteness in context using gpt-2 and human ratings. *Cognitive Science*, 47(12):e13388.

Marc Brysbaert, Boris New, and Emmanuel Keuleers. 2012. Adding part-of-speech information to the subtlex-us word frequencies. *Behavior research methods*, 44:991–997.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46:904–911.

Peter Paul Buitelaar. 1998. *CoreLex: systematic polysemy and underspecification*. Brandeis University.

Jose Camacho-Collados and Mohammad Taher Pilehvar. 2018. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, 63:743–788.

Jean Charbonnier and Christian Wartena. 2019. Predicting word concreteness and imagery. In *Proceedings of the 13th International Conference on Computational Semantics-Long Papers*, pages 176–187.

Emmanuele Chersoni, Rong Xiang, Qin Lu, and Chu-Ren Huang. 2020. Automatic learning of modality exclusivity norms with crosslingual word embeddings. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 32–38.

Domenic V Cicchetti. 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological assessment*, 6(4):284.

D. A. Cruse. 1995. Polysemy and related phenomena from a cognitive linguistic viewpoint. In *Computational Lexical Semantics*, Studies in Natural Language Processing, page 33–49. Cambridge University Press.

Andrea De Varda and Marco Marelli. 2023. Scaling in cognitive modelling: A multilingual approach to human reading times. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 139–149.

Marco Del Tredici and Núria Bel. 2015. A word-embedding-based sense index for regular polysemy representation. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 70–78.

Georgiana Dinu, Marco Baroni, et al. 2013. General estimation and evaluation of compositional distributional semantic models. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 50–58.

Stéphane Dufau, Jonathan Grainger, Katherine J Midgley, and Phillip J Holcomb. 2015. A thousand words are worth a picture: Snapshots of printed-word processing in an event-related potential megastudy. *Psychological science*, 26(12):1887–1897.

Katrin Erk and Sebastian Padó. 2010. Exemplar-based models for word meaning in context. In *Proceedings of the acl 2010 conference short papers*, pages 92–97.

Matan Eyal, Shoval Sadde, Hillel Taub-Tabib, and Yoav Goldberg. 2022. Large scale substitution-based word sense induction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4738–4752, Dublin, Ireland. Association for Computational Linguistics.

Ingrid Lossius Falkum and Agustín Vicente Benito. 2015. Polysemy: current perspectives and approaches. *Lingua: International review of general linguistics*, (157):1–16.

Manaal Faruqui and Chris Dyer. 2015. Non-distributional word vector representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 464–469.

Aina Garí Soler and Marianna Apidianaki. 2021. Let's play mono-poly: Bert can reveal words' polysemy level and partitionability into senses. *Transactions of the Association for Computational Linguistics*, 9:825–844.

Tal Golan, Matthew Siegelman, Nikolaus Kriegeskorte, and Christopher Baldassano. 2023. Testing the limits of natural language models for predicting human language judgements. *Nature Machine Intelligence*, 5(9):952–964.

Gabriel Grand, Idan Asher Blank, Francisco Pereira, and Evelina Fedorenko. 2022. Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature human behaviour*, 6(7):975–987.

Édouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomáš Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Janosch Haber and Massimo Poesio. 2021. Patterns of polysemy and homonymy in contextualised language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2663–2676.

Janosch Haber and Massimo Poesio. 2023. Polysemy-evidence from linguistics, behavioural science and contextualised language models. *Computational Linguistics*, pages 1–67.

Kevin A Hallgren. 2012. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1):23.

Olaf Hauk and Friedemann Pulvermüller. 2004. Effects of word length and frequency on the human event-related potential. *Clinical Neurophysiology*, 115(5):1090–1103.

José A Hinojosa, Eva M Moreno, and Pilar Ferré. 2020. Affective neurolinguistics: towards a framework for reconciling language and emotion. *Language, Cognition and Neuroscience*, 35(7):813–839.

Nancy Kanwisher, Meenakshi Khosla, and Katharina Dobs. 2023. Using artificial neural networks to ask 'why'questions of minds and brains. *Trends in Neurosciences*, 46(3):240–254.

Jerrold J Katz and Jerry A Fodor. 1963. The structure of a semantic theory. *language*, 39(2):170–210.

Ji-Hyun Kim. 2009. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational statistics & data analysis*, 53(11):3735–3745.

Ekaterini Klepousniotou. 2002. The processing of lexical ambiguity: Homonymy and polysemy in the mental lexicon. *Brain and language*, 81(1-3):205–223.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, page 4.

Victor Kuperman, Zachary Estes, Marc Brysbaert, and Amy Beth Warriner. 2014. Emotion and language: valence and arousal affect word recognition. *Journal of Experimental Psychology: General*, 143(3):1065.

Sarah Laszlo and Kara D Federmeier. 2014. Never seem to find the time: evaluating the physiological time course of visual word recognition with regression analysis of single-item event-related potentials. *Language, Cognition and Neuroscience*, 29(5):642–661.

Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. 2012. Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 591–601.

Peter Lauwers and Dominique Willems. 2011. Coercion: Definition and challenges, current approaches, and new trends. *Linguistics*, 49(6):1219–1235.

Alessandro Lenci, Magnus Sahlgren, Patrick Jeuniaux, Amaru Cuba Gyllensten, and Martina Miliani. 2022. A comparative evaluation and analysis of three generations of distributional semantic models. *Language resources and evaluation*, 56(4):1269–1313.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the association for computational linguistics*, 3:211–225.

Jiangtian Li and Blair C Armstrong. 2023. Probing the representational structure of regular polysemy in a contextual word embedding model via sense analogy questions. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

Yu-Chen Lin, Si-An Chen, Jie-Jyun Liu, and Chih-Jen Lin. 2023. Linear classifier: An often-forgotten baseline for text classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1876–1888, Toronto, Canada. Association for Computational Linguistics.

Anastasiya Lopukhina and Konstantin Lopukhin. 2016. Regular polysemy: from sense vectors to sense patterns. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex-V)*, pages 19–23.

Daniel Loureiro, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. Analysis and evaluation of language models for word sense disambiguation. *Computational Linguistics*, 47(2):387–443.

Dermot Lynott, Louise Connell, Marc Brysbaert, James Brand, and James Carney. 2020. The lancaster sensorimotor norms: multidimensional measures of perceptual and action strength for 40,000 english words. *Behavior Research Methods*, 52:1271–1291.

Suresh Manandhar, Ioannis Klapaftis, Dmitriy Dligach, and Sameer Pradhan. 2010. Semeval-2010 task 14: Word sense induction &disambiguation. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 63–68.

Paweł Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2017. Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, 92:57–78.

Diana McCarthy, Marianna Apidianaki, and Katrin Erk. 2016. Word sense clustering and clusterability. *Computational Linguistics*, 42(2):245–275.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023.

Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.

Tom M Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, Robert A Mason, and Marcel Adam Just. 2008. Predicting human brain activity associated with the meanings of nouns. *science*, 320(5880):1191–1195.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, 193:217–250.

Byung-Doh Oh and William Schuler. 2023. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

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 *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, 9(1):963.

J Pustejovsky. 1991. The generative lexicon. *Computational Linguistics*, 17(4):409–441.

Abhilasha Ravichander, Eduard Hovy, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2020. On the systematicity of probing contextualized word representations: The case of hypernymy in bert. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 88–102.

Jamie Reilly, Jinyi Hung, and Chris Westbury. 2017. Non-arbitrariness in mapping word form to meaning: Cross-linguistic formal markers of word concreteness. *Cognitive Science*, 41(4):1071–1089.

Jennifer M Rodd, M Gareth Gaskell, and William D Marslen-Wilson. 2004. Modelling the effects of semantic ambiguity in word recognition. *Cognitive science*, 28(1):89–104.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

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

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational linguistics*, 24(1):97–123.

Graham G Scott, Anne Keitel, Marc Becirspahic, Bo Yao, and Sara C Sereno. 2019. The glasgow norms: Ratings of 5,500 words on nine scales. *Behavior research methods*, 51:1258–1270.

Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Jacob Turton, David Vinson, and Robert Smith. 2020. Extrapolating binder style word embeddings to new words. In *Proceedings of the second workshop on linguistic and neurocognitive resources*, pages 1–8.

Agustín Vicente and Ingrid L Falkum. 2017. Polysemy. In *Oxford research encyclopedia of linguistics*.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240.

Shaonan Wang, Yunhao Zhang, Weiting Shi, Guangyao Zhang, Jiajun Zhang, Nan Lin, and Chengqing Zong. 2023. A large dataset of semantic ratings and its computational extension. *Scientific Data*, 10(1):106.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45:1191–1207.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Meng-Huan Wu, Andrew J Anderson, Robert A Jacobs, and Rajeev DS Raizada. 2022. Analogy-related information can be accessed by simple addition and subtraction of fmri activation patterns, without participants performing any analogy task. *Neurobiology of Language*, 3(1):1–17.

Yadollah Yaghoobzadeh and Hinrich Schütze. 2016. Intrinsic subspace evaluation of word embedding representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 236–246.

Yang Yang, Luan Li, Simon de Deyne, Bing Li, Jing Wang, and Qing Cai. 2024. Unraveling lexical semantics in the brain: Comparing internal, external, and hybrid language models. *Human Brain Mapping*, 45(1):e26546.

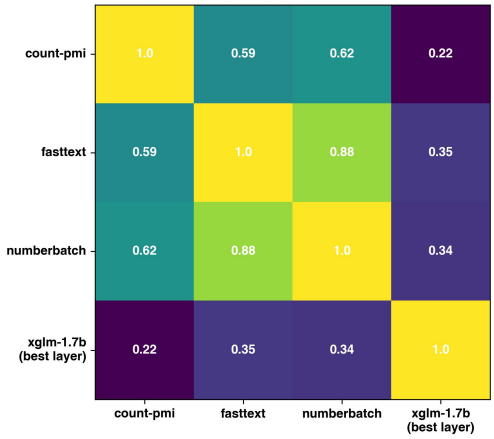

Figure 5: **Pairwise similarities as measured by Representational Similarity Analysis among models**. The scores reported in white are Pearson correlation scores, indicating a clear distinction between static and contextualized models.

A Appendix A: Representational Similarity Analysis of the models' representations

In order to gain some insights into how the models used in our work relate to each other, in Figure 5 we report a visualization of the similarity of the semantic representations across all pairs of models. We carry out the comparisons using the Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008) framework. RSA measures how similar two quantitative ways of representing the same stimuli are by looking at the similarity between the vectors of all pairwise similarities between individual representations in the space. We follow the traditional implementation and we measure similarity with Pearson correlation. As we can see, as it can be expected, static models are rather similar among each other ($corr_{count-pmi,\ fasttext} = 0.59$, $corr_{count-pmi,\ numberbatch} = 0.62$, $corr_{fasttext,\ numberbatch} = 0.88$), while the contextualized model has a different way of representing the phrases ($corr_{XGLM-7.5B,\ count-pmi} = 0.22$, $corr_{XGLM-7.5B,\ fasttext} = 0.35$, $corr_{XGLM-7.5B,\ numberbatch} = 0.34$).