A Continuous Approach to Metaphorically Motivated Regular Polysemy in Language Models

Anna Temerko and **Marcos Garcia** and **Pablo Gamallo**
CiTIUS – Research Center in Intelligent Technologies
Universidade de Santiago de Compostela
{a.temerko,marcos.garcia.gonzalez,pablo.gamallo}@usc.gal

Abstract

Linguistic accounts show that a word's polysemy structure is largely governed by systematic sense alternations that form overarching patterns across the vocabulary. While psycholinguistic studies confirm the psychological validity of regularity in human language processing, in the research on large language models (LLMs) this phenomenon remains largely unaddressed. Revealing models' sensitivity to systematic sense alternations of polysemous words can give us a better understanding of how LLMs process ambiguity and to what extent they emulate representations in the human mind. For this, we employ the measures of surprisal and semantic similarity as proxies of human judgment on the acceptability of novel senses. We focus on two aspects that have not received much attention previously – metaphorically motivated patterns and the continuous nature of regularity. We find evidence that surprisal from language models represents regularity of polysemic extensions in a human-like way, discriminating between different types of senses and varying regularity degrees, and overall strongly correlating with human acceptability scores.

1 Introduction

Polysemy, a linguistic phenomenon whereby a word is associated with multiple related senses, is fundamental to language. As most lexical words are polysemes to varying degrees (Zipf, 1945; Durkin and Manning, 1989; Haber and Poesio, 2024), this form of ambiguity remains a challenge for NLP. However, recent studies show that current language models (LMs) based on Transformers are able to reveal the degree of a word's polysemy, meaningfully cluster word senses, distinguish homonymy from polysemy or perform superior word sense disambiguation (see Garí Soler and Apidianaki, 2021; Li and Joanisse, 2021; Nair et al., 2020; Wiedemann et al., 2019 for each of the above).

We focus on the topic that received less attention in LM research – the regularity dimension of polysemy and its continuous nature. The definition and scope of regular polysemy vary depending on the linguistic theory. The widely cited definition has been proposed by Apresjan (1974, p. 16) and states that "Polysemy of the word A with the meanings a_i and a_j is called regular if [...] there exists at least one other word B with the meanings b_i and b_j, which are semantically distinguished from each other in exactly the same way as a_i and a_j [...]". Pustejovsky's (1991) approach, also frequently adopted, frames regular polysemy as an ability of words that belong to one semantic type to act as members of another, behaving predictably, unlike irregular (accidental) polysemes.

To exemplify, the word *star* shows irregularity in its sense structure referring to a celestial body or a highly celebrated, famous person. To our knowledge, such meaning extension is not shared with any other lexical units of English. In contrast, regular polysemy forms patterns of meaning structure across the vocabulary. Some of the widely used examples of such patterns are ANIMAL - MEAT pattern (instantiated by *chicken* or *salmon*) or CONTAINER - CONTENT (e.g. *cup*, *glass*), as exemplified with sentences (1) and (2) below.

(1) a. We keep our chicken in the backyard.
b. Chicken is a great source of protein.

(2) a. Glass is made of liquid sand.
b. He drank the whole glass.

These are an instance of metonymy – a sense extension device that is based on contiguity (association, referential co-existence) of two concepts. The theoretical approaches mentioned above largely attribute regular polysemy to this figure, and so do the researchers in computational linguistics who adopt these theories (see Section 2 for their overview).

There is, however, another cognitive tool that structures polysemy – metaphor. Unlike metonymy, it is based on analogy, or referential disjunction (Lombard et al., 2023). Regular polysemy by metaphor can be exemplified by such polysemes as *antenna* (insect's organ, signal transmission device) or *leg* (limb, table support) instantiating the pattern BODY PART - OBJECT PART. The two figures are based on different cognitive mechanisms, have different processing profiles in our brain (Klepousniotou et al., 2012), but, as recent psycholinguistic studies show, they equally govern polysemous sense extensions (Lombard et al., 2023, 2024).

Another important aspect of regular polysemy is its continuous nature. In a recent study, Lombard et al. (2024) introduce a method to extract regular polysemes (including metaphors) from WordNet and suggest metrics to measure the degree of regularity of the patterns they are governed by (Table 1). Their findings are in contrast with the widely applied categorical approach to polysemy, where a sense extension of a polyseme is labeled in a binary way, i.e. as either regular or irregular.

Here we adopt this continuous view, aligning with recent work that argues that word meaning, polysemy regularity, and productivity form a continuum rather than discrete representations (Trott and Bergen, 2023; Li, 2024). To the best of our knowledge, no experimental design has previously targeted the graded aspect of regularity in LLMs, although researchers have noted that some patterns seemed more regular or productive than others (Li and Armstrong, 2024). We also contribute by focusing on metaphorically motivated regular polysemy. Only a handful of works in computational linguistics include regular metaphor in their experiments, and even less in the experiments with LMs in particular.

In order to investigate the effect of graded regularity on models' representation of metaphorically motivated polysemes, we rely on datasets compiled for psycholinguistic studies on human polysemy processing in French and English (Lombard et al., 2023, 2024). The datasets feature semantic neologisms – novel senses of existing words created using polysemy patterns of varying regularity degrees. These are compared against attested, existing polysemes and nonsensical derivations (refer to Table 2 for the examples). Human acceptability assessment confirmed the psychological validity of graded regularity for human processing: the more regular the polysemy pattern, the more acceptable its novel senses. Using surprisal and semantic similarity measures, we aim to find out how closely language model processing of semantic neologisms aligns with human processing, and whether the degree of regularity plays a role in it. With this in mind, we outline the following research questions:

RQ1. Which of the two measures (surprisal or semantic similarity) would be a better proxy for human behaviour in our task? As discussed in Methods section (§3), both proved to have psycholinguistic predictive power, despite operating at different levels of language structure.

RQ2. Are the results consistent across model types and sizes? Oh and Schuler (2022) show, e.g., that larger models do not necessarily deliver more human-like linguistic representations.

RQ3. Do models distinguish between the novel senses based on existing regular polysemy patterns and the senses created using the patterns that do not exist? To match human behaviour, models should be able to discriminate between these groups.

RQ4. Are LMs sensitive to the varying degrees of regularity of polysemy patterns? If their processing matches human ratings, we should expect the models to be less surprised by neologisms from highly regular patterns and vice versa.

RQ5. What type of regularity metrics (as defined in Table 1) are models more sensitive to: count-based or consistency-based? Do word frequency and word length play a role, and how does this compare with data from human evaluators?

In the case of LLMs, evaluating novel senses allows us to test their ability to generalize beyond previously seen material and avoid data contamination. Additionally, on a higher level, we can assess their sensitivity to the polysemy patterns abstracted from concrete, previously seen words.

Our results show that LLMs could discriminate between different sense types and regularity gradations in a human-like way, and overall correlated well with human sense plausibility judgment.

In the following sections we will briefly discuss the existing work on regular polysemy (§2), justify our methodology (§3), present the experiments (§4) and discuss their results (§5).

2 Related Work

Aside from the theoretical frameworks cited in the Introduction (§1), regular polysemy is studied in several areas dealing with language processing.

Psycholinguistics. In psycholinguistics, regular

polysemy is addressed in the discussion about the meaning representation in human mind and the nature of restrictions that govern polysemy patterns in language. Many authors defend hybrid approaches to these problems. Rabagliati and Snedeker (2013) suggest that irregular senses are stored separately, while senses that follow regular patterns form core meanings. Analyzing co-predication acceptability and sense similarity of polysemes and homonyms, Haber and Poesio (2020) suggest that senses form groups according to their similarity (in line with Ortega-Andrés and Vicente, 2019), and reject the idea of a fully underspecified representation. In contrast, Vicente (2024) analyses regular and irregular polysemy along several dimensions and defends the one-representation hypothesis.

In the discussion on whether linguistic conventions or an underlying conceptual structure restrict polysemy patterns, Srinivasan and Rabagliati (2015) propose the "conventions-constrained-by-concepts" model. Their study across 15 languages suggests that while the conceptual structure governs the patterns, the language-specific conventions define senses that instantiate them. A hybrid approach is also supported by the investigations in language learning: Zhu (2021) studies how preschoolers acquire regular metonymies, highlighting their ability to quickly grasp semantic generalizations without extensive prior exposure. Children rely on an early-emerging conceptual structure, although at later stages linguistic generalizations also play a crucial role in word learning.

Mental processing of ambiguous words is affected by the degree of relatedness of meanings in memory. This is demonstrated by Brocher (2016; 2018), who report increased processing effort associated with disambiguation of unrelated meanings.

Computational Linguistics. In this field, regular polysemy is addressed in a variety of works, such as Boleda et al. (2012a,b); Lopukhina and Lopukhin (2016), who model systematic polysemy, or Del Tredici and Bel (2015), exploring the representations of polysemous and monosemous words in static word embeddings. A number of researchers propose methods of sense annotation for regular polysemy (Nimb and Pedersen, 2000; Freihat et al., 2013; Martinez Alonso, 2013), while other authors use WordNet to automatically extract regular polysemes (Peters and Peters, 2000; Barque and Chaumartin, 2009; Lombard et al., 2024). Interestingly, the latter authors recognize metaphoric extensions as types of systematic polysemy patterns, in contrast to most of the previously mentioned studies. Peters and Peters (2000) depart from an assumption that metaphoric alternations are irregular, but after applying their extraction method, "stumble upon" the instances of metaphoric sense extensions that can only be described as regular. Only a few more works mentioned in this section fully recognize that regular polysemy by metaphor is possible: Nimb and Pedersen, 2000; Freihat et al., 2013; Lopukhina and Lopukhin, 2016; Lombard et al., 2023 and Lombard et al., 2024.

Language models. Regarding regular polysemy and neural language models, Haber and Poesio (2021) test BERT's ability to predict human assessment of sense similarity degree. They report that $\text{BERT}_{\text{LARGE}}$ captures distinctions between polysemic, homonymic and same-sense samples in a human-like way. BERT delivers sensible results in sense clustering, suggesting that this model is sensitive to polysemy patterns. Sørensen et al. (2023) explore BERT sense clustering as a guidance tool for annotation of systematic polysemy in lexical resources. Similarly to Haber and Poesio (2021), they got mixed results but see potential: for one of the patterns, BERT discovered a sense that the authors overlooked when creating the dataset. Finally, Li and Armstrong (2024) use sense analogy questions to investigate how regular polysemy is represented in BERT embeddings. The authors observe that the pattern of BERT's sense similarity score distribution reflects differences not only in the processing of regular polysemes and irregular/homonymous controls, but also of distinct polysemy patterns. They also note on the scalar nature of regularity, an observation that contributes to Li's (2024) comprehensive approach to polysemy as continuous in its sense individuation, regularity, and productivity.

The present paper adopts the recent insights about the graded nature and metaphoric motivation of regular patterns and incorporates them in the experimental design.

3 Materials and Methods

3.1 Data

To answer our research questions, we evaluated two datasets compiled by Lombard et al. (2023) and Lombard et al. (2024)[1]. Both data sets were created for psycholinguistic experiments investigating the effect of graded regularity on the human

Metr.	Definition	Formula
R1	Number of words having SENSE_1 and SENSE_2 in a given pattern.	$R_1 = N_{S2}$
R2	Ratio of R1 and the number of words with SENSE_1, whether or not they have SENSE_2.	$R_2 = \frac{N_{S2}}{N_{S1}}$
R3	R1 weighted by the log-frequency of occurrence of the word.	$R_3 = \sum_{w=1}^{N_{S2}} \log(f_w)$
R4	R2 weighted by the log-frequency of occurrence of the word	$R_4 = \frac{\sum_{w=1}^{N_{S2}} \log(f_w)}{\sum_{w=1}^{N_{S1}} \log(f_w)}$

Table 1: Regularity metrics as proposed by (Lombard et al., 2024, pp. 4–5). While R1 and R3 capture the number of pattern instantiations, R2 and R4 reflect the consistency with which words having a base sense (SENSE_1) also have a derived sense (SENSE_2) within a pattern.

Type	Pattern	Example	W.	S.
new	ANIMAL - ARTIFACT	My sister cleaned the **porcupine** of the brush.	35	70
	ANIMAL - PERSON	The chessplayer is always a cruel **spider** with his opponents.		
	ARTIFACT - MESSAGE	A mean **spear** slipped through her lips in an angry tone.		
	BODY PART - OBJECT PART	We can see the **knee** of the chair getting damaged.		
	NATURAL EVENT - HAPPENING	There was a huge **tornado** of claps at the final of the challenge.		
	PERSON - ANIMAL	Some zoos are trying to protect the **doctor** from extinction.		
	PHYS. PROP. - PSYCHOL. PROP.	She said that the **density** of the project was an issue.		
illegal		My brother painted the **curry** of the controller in blue.	40	80
existing		My dog chewed the **tongue** of my new shoes	40	40
all			115	**190**

Table 2: Sentence examples of each sense type, labeled in the original dataset as *new*, *illegal*, and *existing*. *New* senses include 7 polysemy patterns (5 words per pattern). *Illegal* and *existing* senses are not annotated with patterns in the original dataset. The column *W.* lists the number of words per sense type, while *S.* – the number of sentences.

perception of neology. The more recent study is in English and focuses solely on regular metaphor, whereas the earlier one is in French and involves both metaphor and metonymy. Since the present research focuses on metaphoric polysemy, we only evaluate the part of the French dataset containing metaphors. Having removed the metonymies, we were left with only 42 sentences to evaluate, which limited our ability to derive meaningful results for French (see §4.2 and §5). The English dataset comprises 190 sentences.

The stimuli. The datasets contain sentences with target words of three types:

1. Semantic neology: words used in a novel, unattested sense. The derived metaphoric sense, together with the base sense, represent a polysemy pattern that a given word has never developed, unlike other words from its semantic field. To exemplify, the word *knee* represents a pattern BODY PART-OBJECT PART and is used in the sentence *We can see the knee of the chair getting damaged*. For comparison, some of the words that actually developed both senses are *leg, heart, artery, vein, antenna, wing, head, skeleton, brow, tongue* etc.

2. Non-sensical derivation: semantic neologisms that follow a non-existent pattern in each language. For instance, *curry* in *My brother painted the curry of the controller in blue* represents an unattested pattern FOOD-OBJECT PART.

3. Existing polysemy: words used in an attested sense of a valid, existing polysemy pattern. For example, *tongue* in *My dog chewed the tongue of my new shoes* is used in an attested sense of an OBJECT PART. An overview of the English dataset with sentence examples is presented in Table 2.

The dataset is annotated with human acceptability scores, regularity degree of polysemy patterns, word frequency and word length.

Human acceptability rating. Human acceptability scores are derived from the initial psycholinguistic experiment. They reflect how plausible the annotators found each sentence on a scale from 'no sense at all' (0) to 'completely acceptable' (100).

Regularity. Each target word is annotated with a score reflecting the degree of regularity of a polysemy pattern it instantiates. For the two languages,

this metric has been calculated using different procedures. For English, the authors developed an automatic extraction technique using WordNet and proposed several formulas to calculate the regularity degree of a pattern based on the extracted data. These regularity metrics are summarized in Table 1. For French, the authors relied on the judgment of experts in French lexicology to assess the degree of regularity for each pattern. The methodological differences in the compilation of both datasets seem to affect our results, which will be discussed in more detail in Section 4.2.

3.2 Methods

To answer our research questions, we explore two common methods in NLP and computational psycholinguistics – surprisal and semantic similarity from large language models.

Surprisal. Surprisal is the negative log-probability of a token given its immediate context. Surprisal theory (Hale, 2001; Levy, 2008) assumes that the processing difficulty of the word is based on its predictability. This information-theoretic measure is typically used in the studies on human reading, where it proved to predict reading times and, consequently, cognitive processing difficulty in multiple languages (for recent work, see de Varda and Marelli, 2022; Nair and Resnik, 2023; Wilcox et al., 2023; Xu et al., 2023). It is also used to assess the models' ability to predict linguistic acceptability (grammaticality) of sentences (Noh et al., 2024). In our study, we use surprisal from language models as a proxy of human acceptability judgment of novel word senses: we assume that higher surprisal values assigned to a target word by an LM correspond to lower acceptability scores obtained from human evaluation.

Semantic relatedness. Semantic similarity between a word and its context is used along with surprisal to predict reading times, assess processing difficulty and explain brain activity during language processing (Leal et al., 2021; Salicchi et al., 2021; Kun et al., 2023). Specifically, we apply the cosine similarity between the vector of the target word and the vector of the sentence obtained by mean-pooling. Additionally, since a few rogue dimensions often dominate similarity measures in transformer models (Timkey and van Schijndel, 2021), we compare the original and normalized vectors (z-scoring) to assess their impact. We also use Spearman's ρ as a similarity metric, another technique suggested by Timkey and van Schijndel (2021) and replicated by Lyu et al. (2023) and Salicchi et al. (2023).

In reading experiments, low similarity between a word and its context is associated with increased human reading difficulty. In our study, we expect to associate low similarity rating from LMs with low human acceptability of semantic neologisms.

As shown by Salicchi et al. (2023), both surprisal and semantic relatedness equally contribute to the prediction of reading difficulty, despite operating at different levels of language structure. While surprisal operates at the syntagmatic level and reflects how predictable the word is from its context, semantic relatedness reflects coherence of a word with its context modeling paradigmatic dimension. Both surprisal and semantic relatedness proved to predict brain activity during language comprehension and are associated with signals from distinct brain areas (Frank and Willems, 2017; Michaelov et al., 2023; Salicchi and Hsu, 2025).

3.3 Models

We used a set of masked language models and compared them with an autoregressive Llama.

For English, we use the monolingual BERT as well as RoBERTa. For French, we took the BERT-based FlauBERT, and the RoBERTa-based CamemBERTv2. We also evaluate multilingual models on both languages: mBERT and XLM-RoBERTa.

Surprisal experiments typically use unidirectional decoder models (e.g., GPT), as they rely only on left-context to emulate human reading, avoiding access to future words. In our case, the experimental settings of the initial psycholinguistic study entail the choice of a masked model: the evaluators were first presented with the context on both sides before seeing the full sentence. We still include an autoregressive LM to compare the results and challenge our assumption about masked language modeling being more suitable for our task. For this, we chose Llama 3.1 8B and Llama 3.2 3B, which we oppose to BERT as more recent and significantly larger multilingual models that include English and French. For all models, weights were taken off HuggingFace. Additional information on these models is presented in Table 5 of Appendix A.

4 Experiments and Results

4.1 Experiments

We feed the sentences into each of the language models and compute[2] the surprisal and semantic relatedness scores as described in Methods section (§3). For this, we use the `minicons` library provided by Misra (2022). For bi-directional models, we rely on the 'pseudo-log-likelihood' proposed by Kauf and Ivanova (2023), which takes into account multi-token and out-of-vocabulary words.[3]

We then compute the Spearman correlation between the human acceptability scores and each of the measures (target word surprisal and the similarity between the target word and its context).

4.2 Results

Models	Default	K & I	PF
$\text{BERT}_{\text{BASE}}$	**-0.65**	-0.63	-0.61
$\text{BERT}_{\text{LARGE}}$	**-0.68**	-0.65	-0.64
$\text{RoBERTa}_{\text{BASE}}$	-0.67	**-0.76**	-0.68
$\text{RoBERTa}_{\text{LARGE}}$	-0.72	**-0.78**	-0.70
$\text{XLM-RoBERTa}_{\text{BASE}}$	-0.38	**-0.56**	-0.39
$\text{XLM-RoBERTa}_{\text{LARGE}}$	-0.44	**-0.63**	-0.43
$\text{mBERT}_{\text{BASE}}$	-0.26	**-0.47**	**-0.47**
Llama 3.1 8B	**-0.65**	-	-
Llama 3.2 3B	**-0.65**	-	-

Table 3: Results of the surprisal experiment in English. The column *Default* reports results obtained from the default implementation of `minicons` (Misra, 2022), the column *K & I* reports the results from the method by Kauf and Ivanova (2023), and *PF* – from the PsychFormers application (Michaelov et al., 2023). All results are statistically significant (p<.05). Bold formatting points to the strongest correlation achieved by each model.

Surprisal. Across models, we observe moderate to strong correlation with human judgment. As expected, models correlate negatively, showing that more acceptable senses elicit lower surprisal.

Among masked models, the strongest correlation was achieved by $\text{RoBERTa}_{\text{LARGE}}$ at -0.78, p<.001. It is followed by $\text{BERT}_{\text{LARGE}}$ showing moderate negative correlation of -0.68, p<.001. Multilingual models demonstrated poorer results with correlation coefficients of -0.63 for $\text{XLM-RoBERTa}_{\text{LARGE}}$ (p<.001), as well as -0.47 for mBERT (p<.001). As mentioned previously, the method of Kauf and Ivanova (2023) yielded the best results, except for the BERT models which performed slightly better using the standard metric. See Table 3 for a complete overview of the different models and metrics.

As for autoregressive models, Llama 3.1 8B and Llama 3.2 3B achieved correlation of -0.65 (p<.001 for both), yielding the best results among the multilingual models but exhibiting a lower correlation than the smaller monolingual encoders.

In French, none of the models gave statistically significant correlation at the word level. We attribute this to the much smaller dataset size (42 sentences). However, we could still obtain usable results by changing the experimental settings: we checked correlation of sentence-wise surprisal with human judgment (obtained by sum and mean) and received statistically significant results for $\text{XLM-RoBERTa}_{\text{LARGE}}$, at -0.32, p=.039 (sum). We compared this result with the sentence surprisal of the English version from XLM-RoBERTa, and curiously, for English, this was the only model that showed stronger correlation when computing sentence surprisal instead of the target word surprisal (-0.65 vs. -0.63, p<.001 in both cases). Table 6 in Appendix C presents all scores obtained from the sentence-wise correlation experiment. Additionally, it reports correlation of sentence surprisal with the acceptability of polysemy patterns, where $\text{XLM-RoBERTa}_{\text{LARGE}}$ achieved moderate significant correlation.

Semantic relatedness. The results of the experiment with semantic relatedness are more difficult to summarize, as the data does not allow to discern clear trends. In different models, the highest correlation was achieved across varying layers, model sizes and normalization approaches. Moreover, some models show positive correlation with human judgment, while others correlate negatively. This is not expected, as usually we assume a better word/context coherence to elicit higher acceptability scores. Tables 7 to 10 of Appendix D offer a full overview of the correlation scores distribution within several selected models: masked RoBERTa and FlauBERT for English and French, as well as a significantly bigger multilingual autoregressive Llama 3.1 8B. Here, we will only highlight the best results achieved by the models to give an idea of how inconsistent they are across experimental settings.

The strongest correlation was reached by Llama

[2]The information on GPU use and computation time is reported in Appendix B.

[3]We also tested the standard scoring based on Salazar et al., 2020, and that of PsychoFormers (Michaelov and Bergen, 2022), obtaining generally lower results, as shown in Table 3.

3.1 8B (32 layers) in the layer 4 using Spearman's ρ instead of cosine, the correlation being positive (0.66, p<.001). RoBERTa$_{\text{BASE}}$ (12 layers) follows with coefficient of -0.58, p<.001 in the ninth layer without applying any normalization techniques. Finally, BERT$_{\text{BASE}}$ (12 layers) achieved the correlation of 0.52 in the last layer when applying Spearman's ρ instead of the cosine (p<.001). Multilingual models score 0.5 and below, their best achieved correlation coefficients being scattered across different experimental settings.

For French, the scores lie in the same range, but with the strongest correlation achieved by a smaller Llama 3.2 3B (-0.53, p<.001, in the last 28$^{\text{th}}$ layer, non-normalized).

Lyu et al. (2023) report similar outcome of their study of lexical stylistic features in language models: although normalization generally improves the results (especially for the multilingual models), it is hard to single out the best technique for all models and experimental settings. As for Salicchi et al. (2023), they do not notice any effect of BERT's embedding anisotropy on reading times prediction.

Overall, semantic relatedness results show no clear interpretable trend across models and settings.

5 Discussion

In this section, we will address the research questions presented in the Introduction (§1).

RQ1. Regarding the choice of measure (surprisal or semantic relatedness), the results suggest that surprisal is preferable. Not only because it achieved strong correlation (-0.78 for RoBERTa$_{\text{LARGE}}$ in surprisal vs. -0.61 for Llama 3.2 3B in similarity setting), but also because it is consistent, more interpretable, and easier to obtain. While we can confirm the assumption that surprisal is in an inverse relationship with sense plausibility as assessed by humans, the semantic similarity scores correlated both negatively and positively depending on the model, its layer and the embedding normalization technique (see Tables in Appendix D). Finding the most suitable configuration thus demands running a considerable number of trials.

RQ2. In surprisal setting, masked LMs performed better, confirming our assumption that masked model scoring with its access to the bi-directional context would be more suitable for our task. Previous research has repeatedly shown that larger model size delivers a poorer prediction of processing difficulty (Oh and Schuler, 2022; Salicchi et al., 2023; Liu et al., 2024; Shain et al., 2024). In contrast, in our experiments, large varieties of the same models always performed above the base ones (see Table 3). Interestingly, much larger Llama 3.1 8B and 3.2 3B did not outperform masked monolingual BERT and RoBERTa (330M and 355M respectively for large varieties). We attribute this to differences in model architecture, although it requires further investigation. The Limitations section (Appendix E) elaborates more on this issue. In the case of semantic similarity, results are not consistent enough to draw conclusions on this topic, as explained in the Results section.

We further analyse the results to test the models' sensitivity to such features as sense types, regularity degrees, word frequency and word length.

RQ3. We run a series of tests to confirm whether the models discriminate between the senses derived using the existing and non-existing patterns, as well as to see if they are sensitive to the varying pattern regularity degrees. We took our best-performing masked model RoBERTa$_{\text{LARGE}}$ and an autoregressive Llama 3.1 8B, for comparison. For French, we picked the same Llama model and FlauBERT$_{\text{LARGE}}$. A Mann-Whitney U test on two independent samples for the two sense types (two-sided, p<0.05) shows that the difference is significant for RoBERTa, Llama and human evaluators. They could distinguish between all three groups of senses (Figure 1). For French, FlauBERT and Llama 3.1 8B did not yield significancy (see Figure 4 in Appendix E for score distribution).

RQ4. In the same way, we established that language models were sensitive to the degrees of regularity of the polysemy patterns the senses instantiated, although not as fine-grained as humans: while the Mann-Whitney U test shows significancy in the difference between low, medium and high regularity of patterns for humans, the models only discriminate between high/low and medium/low groups (Figure 2). Again, neither Llama nor FlauBERT reached statistical significance in French. The distribution plot can be found in Appendix E, Figure 5.

RQ5. We also establish whether there is a relation between the model scoring and such factors as the degree of polysemy pattern regularity, word frequency and word length. The latter two factors contribute to the cognitive processing load in humans since less frequent and longer words require more time to process (Pollatsek et al., 2008). Figure 3 illustrates the pattern of correlation (Pearson) between four regularity metrics and the measures

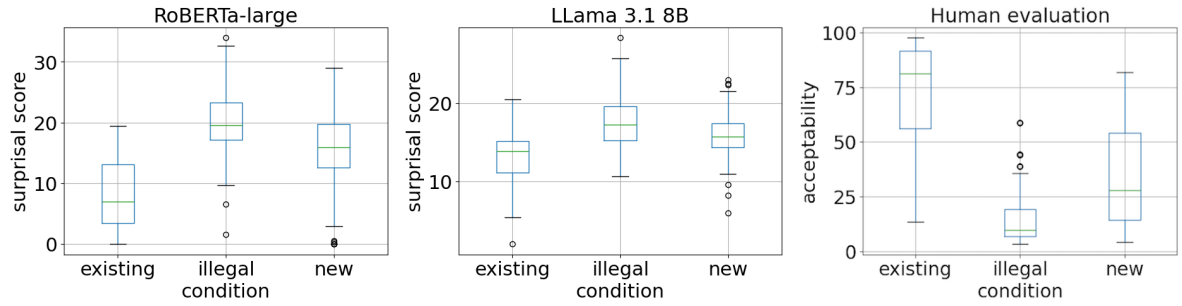

Figure 1: Distribution of English surprisal scores by condition labeled in the original dataset as *new*, *illegal* and *existing*. These correspond to the groups (1), (2) and (3) respectively, as described in the Section 3.1. High model surprisal is expected to correspond to the low acceptability scores in the human rating.

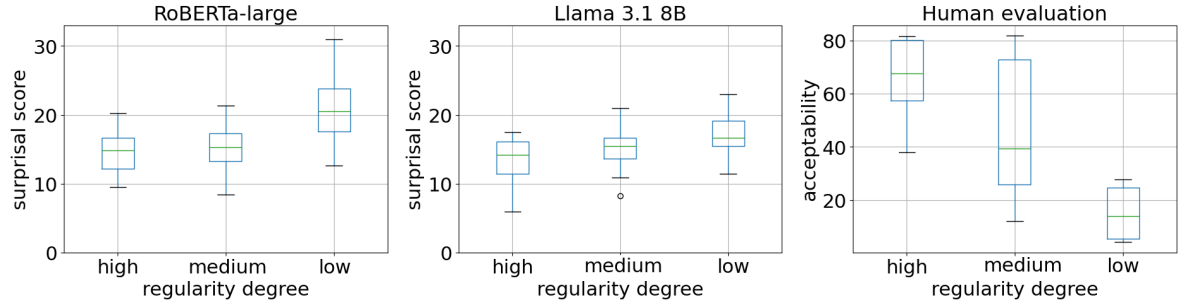

Figure 2: Distribution of surprisal scores by regularity degrees labeled as *high*, *medium* and *low* for the English data. High model surprisal is expected to correspond to the low acceptability scores in the human rating.

of semantic similarity and surprisal from RoBERTa and Llama, as compared to human evaluators. As described in Table 1, we consider a count-based metric R1, consistency-based metric R2 and two metrics that weight them by the log-frequency of the occurrence of the word – R3 and R4. The measure of surprisal is more aligned with the human judgment correlated with regularity, Llama showing almost identical coefficients for most metrics. The same as for human evaluators, for both language models, the regularity metrics that reflect how consistently words instantiate a polysemy pattern appeared to be more relevant than the sheer number of words having the SENSE_1 and the SENSE_2. Weighting R1 and R2 by word frequency generally did not improve the correlation coefficients (except for RoBERTa in R4 where it gains one point). Again, the correlation scores for the measure of semantic relatedness are generally low, with apparent preference for consistency-based and frequency-weighted metrics. All correlation coefficients are listed in Table 11, Appendix E.

As for the effect of the word frequency and word length, the models generally show a low correlation, although it is higher than the one computed with human scoring. The exceptions are RoBERTa in the similarity setting and FlauBERT in the surprisal setting relying on these features more and correlating moderately (see Table 4 for correlation scores and Figure 6 in Appendix E for visualization).

English		
Models	**W. freq.**	**W. length**
$\text{RoBERTa}_{\text{LARGE}}$ SURP	-0.21*	0.24*
EN Llama 3.1 8B SURP	-0.27*	0.16*
$\text{RoBERTa}_{\text{LARGE}}$ SIMIL	-0.34*	0.33*
Llama 3.1. 8B SIMIL	0.07	0.1
Human acceptability rating	0.04*	-0.09*
French		
$\text{FlauBERT}_{\text{LARGE}}$ SURP	0.38*	0.41*
Llama 3.1 8B SURP	0.00	0.29
$\text{FlauBERT}_{\text{LARGE}}$ SIMIL	0.06	-0.07
Llama 3.1 8B SIMIL	-0.16	0.16
Human acceptability rating	-0.15	0.17

Table 4: Correlation (Pearson) of word length and word frequency with model scoring and human evaluation. Asterisk (*) indicates a statistically significant correlation (p<.05).

6 Conclusions

In this paper, we investigated the effect of the graded regularity of polysemy patterns on the pro-

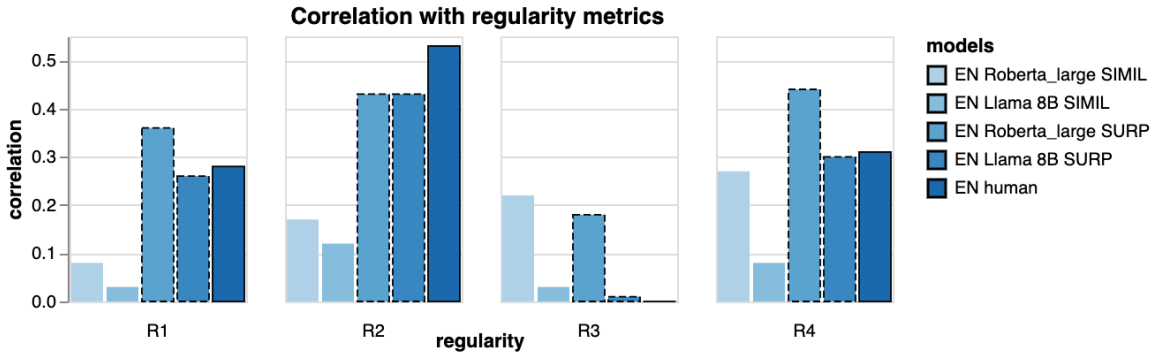

Figure 3: Pearson's *r* for regularity metrics as described in Section 5, in absolute numbers. The dashed line marks the measure of surprisal from RoBERTa and Llama, while word/context similarity remains unmarked.

cessing of novel metaphorical word senses by large language models. Using surprisal and semantic relatedness as proxies, we found evidence that models represent regularity of polysemy extensions in a human-like way. Especially surprisal proved to adequately model sense plausibility, showing a strong correlation with human judgment. Among models, RoBERTa delivered the best results. Furthermore, the distributions of model scores suggest sensitivity to different types of sense extensions and regularity degrees. Similarly to humans, LLMs could discriminate between attested polysemes, novel senses derived from regular polysemy patterns and nonsensical derivations. They were, however, less responsive to the gradations in regularity, only differentiating very regular and weakly regular patterns. These observations allow us to better understand how LLMs model lexical ambiguity and to what extent such factors as regularity, continuity and sense relatedness affect model representations.

Acknowledgments

This work was funded by MCIU/AEI/10.13039/501100011033 (grants with references PID 2021-128811OA-I00, PRE2022-102762, CNS2024-154902, and TED 2021-130295B-C33, the latter also funded by "European Union Next Generation EU/PRTR"), by the Galician Government (ERDF 2024-2027: Call ED431G 2023/04, and ED431F 2021/01), and by a Ramón y Cajal grant (RYC2019-028473-I).

References

Wissam Antoun, Francis Kulumba, Rian Touchent, Éric de la Clergerie, Benoît Sagot, and Djamé Seddah. 2024. Camembert 2.0: A smarter french language model aged to perfection. *Preprint*, arXiv:2411.08868.

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

Lucie Barque and François-Régis Chaumartin. 2009. Regular polysemy in WordNet. *Journal for language technology and computational linguistics*, 24(2):5–18.

Gemma Boleda, Sebastian Padó, and Jason Utt. 2012a. 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, Montréal, Canada. Association for Computational Linguistics.

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

Andreas Brocher, Stephani Foraker, and Jean-Pierre Koenig. 2016. Processing of irregular polysemes in sentence reading. *Journal of experimental psychology. Learning, memory, and cognition*, 42 11:1798–1813.

Andreas Brocher, Jean-Pierre Koenig, Gail Mauner, and Stephani Foraker. 2018. About sharing and commitment: the retrieval of biased and balanced irregular polysemes. *Language, Cognition and Neuroscience*, 33:443 – 466.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Andrea de Varda and Marco Marelli. 2022. The effects of surprisal across languages: Results from native and non-native reading. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP*

2022, pages 138–144, Online only. Association for Computational Linguistics.

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, Denver, Colorado. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kevin Durkin and Jocelyn Manning. 1989. Polysemy and the subjective lexicon: Semantic relatedness and the salience of intraword senses. *Journal of Psycholinguistic Research*, 18:577–612.

S. Frank and Roel M. Willems. 2017. Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Language, Cognition and Neuroscience*, 32:1192 – 1203.

Abed Alhakim Freihat, Fausto Giunchiglia, and Biswanath Dutta. 2013. Regular polysemy in wordnet and pattern based approach. *International Journal On Advances in Intelligent Systems*, 6:199–212.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic,

Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Janosch Haber and Massimo Poesio. 2020. Assessing polyseme sense similarity through co-predication acceptability and contextualised embedding distance. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 114–124, Barcelona, Spain (Online). Association for Computational Linguistics.

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, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Janosch Haber and Massimo Poesio. 2024. Polysemy—Evidence from linguistics, behavioral science, and contextualized language models. *Computational Linguistics*, 50(1):351–417.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Carina Kauf and Anna Ivanova. 2023. A better way to do masked language model scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 925–935, Toronto, Canada. Association for Computational Linguistics.

Ekaterini Klepousniotou, G. Bruce Pike, Karsten Steinhauer, and Vincent L. Gracco. 2012. Not all ambiguous words are created equal: An eeg investigation of homonymy and polysemy. *Brain and Language*, 123:11–21.

Sun Kun, Qiuying Wang, and Xiaofei Lu. 2023. An interpretable measure of semantic similarity for predicting eye movements in reading. *Psychonomic Bulletin & Review*, 30:1227 – 1242.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2020. Flaubert: Unsupervised language model pre-training for french. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Sidney Leal, Edresson Casanova, Gustavo Paetzold, and Sandra Aluísio. 2021. Evaluating semantic similarity methods to build semantic predictability norms of reading data. In *Text, Speech, and Dialogue*, pages 35–47, Cham. Springer International Publishing.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Jiangtian Li. 2024. Semantic minimalism and the continuous nature of polysemy. *Mind & Language*, 39(5):680–705.

Jiangtian Li and Blair C Armstrong. 2024. Probing the representational structure of regular polysemy via sense analogy questions: Insights from contextual word vectors. *Cognitive Science*, 48(3):e13416.

Jiangtian Li and Marc Joanisse. 2021. Word senses as clusters of meaning modulations: A computational model of polysemy. *Cognitive Science*, 45.

Tong Liu, Iza Škrjanec, and Vera Demberg. 2024. Temperature-scaling surprisal estimates improve fit to human reading times – but does it do so for the "right reasons"? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9598–9619, Bangkok, Thailand. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Alizée Lombard, Richard Huyghe, Lucie Barque, and Doriane Gras. 2023. Regular polysemy and novel word-sense identification. *The Mental Lexicon*, 18(1):94–119.

Alizée Lombard, Anastasia Ulicheva, Maria Korochkina, and Kathy Rastle. 2024. The regularity of polysemy patterns in the mind: Computational and experimental data. *Glossa Psycholinguistics*, 3(1).

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.

Qing Lyu, Marianna Apidianaki, and Chris Callison-burch. 2023. Representation of lexical stylistic features in language models' embedding space. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 370–387, Toronto, Canada. Association for Computational Linguistics.

Hector Martinez Alonso. 2013. *Annotation of Regular Polysemy: An empirical assessment of the underspecified sense*. Ph.D. thesis, Universitat Pompeu Fabra and University of Copenhagen, Barcelona, Spain.

James A. Michaelov, Megan D. Bardolph, Cyma K. Van Petten, Benjamin K. Bergen, and Seana Coulson. 2023. Strong prediction: Language model surprisal explains multiple n400 effects. *Neurobiology of Language*, 5:107 – 135.

James A. Michaelov and Benjamin K. Bergen. 2022. Do language models make human-like predictions about the coreferents of Italian anaphoric zero pronouns? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1–14, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Kanishka Misra. 2022. minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

Sathvik Nair and Philip Resnik. 2023. Words, subwords, and morphemes: What really matters in the surprisal-reading time relationship? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11251–11260, Singapore. Association for Computational Linguistics.

Sathvik Nair, Mahesh Srinivasan, and Stephan Meylan. 2020. Contextualized word embeddings encode aspects of human-like word sense knowledge. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 129–141, Online. Association for Computational Linguistics.

Sanni Nimb and Bolette Sanford Pedersen. 2000. Treating metaphoric senses in a danish computational lexicon –different cases of regular polysemy. In *Proceedings of the 9th EURALEX International Congress*, pages 679–691, Stuttgart, Germany. Institut für Maschinelle Sprachverarbeitung.

Kangsan Noh, Eunjeong Oh, and Sanghoun Song. 2024. Testing language models' syntactic sensitivity to grammatical constraints: a case study of wanna contraction. *Frontiers in Communication*, 9.

Byung-Doh Oh and William Schuler. 2022. 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).

Wim Peters and Ivonne Peters. 2000. Lexicalised systematic polysemy in WordNet. In *Proceedings of the*

Second International Conference on Language Resources and Evaluation (LREC'00), Athens, Greece. European Language Resources Association (ELRA).

Alexander Pollatsek, Barbara Jean Juhasz, Erik D. Reichle, Debra Machacek, and Keith Rayner. 2008. Immediate and delayed effects of word frequency and word length on eye movements in reading: a reversed delayed effect of word length. *Journal of experimental psychology. Human perception and performance*, 34 3:726–50.

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

Hugh Rabagliati and Jesse Snedeker. 2013. The truth about chickens and bats: Ambiguity avoidance distinguishes types of polysemy. *Psychological science*, 24(7):1354–1360.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Lavinia Salicchi, Emmanuele Chersoni, and Alessandro Lenci. 2023. A study on surprisal and semantic relatedness for eye-tracking data prediction. *Frontiers in Psychology*, 14.

Lavinia Salicchi and Yu-Yin Hsu. 2025. Not every metric is equal: Cognitive models for predicting n400 and p600 components during reading comprehension. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3648–3654, Abu Dhabi, UAE. Association for Computational Linguistics.

Lavinia Salicchi, Alessandro Lenci, and Emmanuele Chersoni. 2021. Looking for a role for word embeddings in eye-tracking features prediction: Does semantic similarity help? In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 87–92, Groningen, The Netherlands (online). Association for Computational Linguistics.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences of the United States of America*, 121.

Nathalie Sørensen, Sanni Nimb, and Bolette Sandford Pedersen. 2023. How do we treat systematic polysemy in wordnets and similar resources?–using human intuition and contextualized embeddings as guidance. In *Proceedings of the 12th Global Wordnet Conference*, pages 117–126.

Mahesh Srinivasan and Hugh Rabagliati. 2015. How concepts and conventions structure the lexicon: Cross-linguistic evidence from polysemy. *Lingua*, 157:124–152.

William Timkey and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4527–4546, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sean Trott and Benjamin Bergen. 2023. Word meaning is both categorical and continuous. *Psychological Review*, 130(5):1239.

Agustín Vicente. 2024. Polysemies and the one representation hypothesis. *The Mental Lexicon*.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers*, pages 161–170, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

Ethan G. Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the Predictions of Surprisal Theory in 11 Languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Weijie Xu, Jason Chon, Tianran Liu, and Richard Futrell. 2023. The linearity of the effect of surprisal on reading times across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15711–15721, Singapore. Association for Computational Linguistics.

Rebecca Zhu. 2021. Preschoolers' acquisition of producer-product metonymy. *Cognitive Development*, 59:101075.

George Kingsley Zipf. 1945. The meaning-frequency relationship of words. *The Journal of general psychology*, 33:251–6.

A Model Information

Table 5 offers a model description in terms of size and languages, HuggingFace names, as well as paper references.

Model	Reference	Layers	Lang.	Size
google-bert/bert-base-uncased	Devlin et al., 2019	12	EN	110M
google-bert/bert-large-uncased	Devlin et al., 2019	24	EN	330M
FacebookAI/roberta-base	Liu et al., 2019	12	EN	125M
FacebookAI/roberta-large	Liu et al., 2019	24	EN	355M
almanach/camembertv2-base	Antoun et al., 2024	12	FR	112M
FlauBERT/flaubert_base_cased	Le et al., 2020	12	FR	138M
FlauBERT/flaubert_large_cased	Le et al., 2020	24	FR	373M
google-bert/bert-base-multilingual-uncased	Devlin et al., 2019	12	multi	179M
FacebookAI/xlm-roberta-base	Conneau et al., 2019	12	multi	279M
FacebookAI/xlm-roberta-large	Conneau et al., 2019	24	multi	560M
meta-llama/Llama-3.1-8B	Grattafiori et al., 2024	32	multi	8B
meta-llama/Llama-3.2-3B	Grattafiori et al., 2024	28	multi	3B

Table 5: Models used in this experiment can be obtained from https://huggingface.co.

B GPU Use

We ran the experiments with Llama 3.1 8B and LLama 3.2 3B on an Nvidia Ampere A100 80GB GPU node. We estimate the overall execution time to be 200 hours.

C Sentence-wise Surprisal

Sentence surprisal for EN models, word acceptability			Sentence surprisal for FR models, word acceptability			Sent. surprisal for FR models, pattern acceptability	
Model	**Default**	**K & I**	**Model**	**Default**	**K & I**	**Default**	**K & I**
$\text{BERT}_{\text{BASE}}$	-0.46*	-0.54*	$\text{CamemBERTv2}_{\text{BASE}}$	0.02	-0.06	-0.2	-0.11
$\text{BERT}_{\text{LARGE}}$	-0.46*	-0.55*	$\text{FlauBERT}_{\text{BASE}}$	-0.18	-	-0.11	-
$\text{RoBERTa}_{\text{BASE}}$	-0.57*	**-0.63***	$\text{FlauBERT}_{\text{LARGE}}$	-0.17	-	-0.06	-
$\text{RoBERTa}_{\text{LARGE}}$	-0.54*	-0.61*	$\text{XLM-RoBERTa}_{\text{BASE}}$	-0.24	-0.15	-0.3	0.01
$\text{XLM-RoBERTa}_{\text{BASE}}$	-0.36*	-0.54*	$\text{XLM-RoBERTa}_{\text{LARGE}}$	**-0.32***	-0.23	**-0.52***	-0.21
$\text{XLM-RoBERTa}_{\text{LARGE}}$	-0.42*	-0.65*	$\text{mBERT}_{\text{BASE}}$	0.02	0.02	-0.12	-0.02
$\text{mBERT}_{\text{BASE}}$	-0.16	-0.41*	Llama 3.2 3B	-0.2	-	-0.23	-
Llama 3.2 3B	-0.62*	-	Llama 3.1 8B	-0.19	-	-0.27	-
Llama 3.1 8B	-0.62*	-					

Table 6: Spearman correlaiton coefficients for sentence surprisal and human acceptability scores in English and French. For French, we additionally include the correlation with human acceptability as per polysemy pattern (averaged over acceptability of each word instance of a pattern). Asterisk (*) indicates a statistically significant result at $p < .05$.

D Semantic Similarity vs. Human Acceptability Judgment

We show mixed results obtained from the experiments with semantic similarity for selected models in Tables 7 to 10. Results are presented for RoBERTa, FlauBERT and Llama 3.1 8B covering both languages.

L.	Cosine	Cosine norm.	Spearman
RoBERTa$_{\text{BASE}}$			
0	-0.2, p=.034	0.08, p=.422	0.0, p=.964
1	-0.18, p=.058	0.27, p=.003	0.2, p=.031
2	-0.27, p=.003	0.37, p<.001	0.25, p=.006
3	-0.28, p=.002	0.44, p<.001	0.34, p<.001
4	-0.31, p<.001	**0.56, p<.001**	0.38, p<.001
5	-0.32, p<.001	0.54, p<.001	**0.42, p<.001**
6	-0.39, p<.001	0.42, p<.001	0.32, p<.001
7	-0.52, p<.001	0.25, p=.008	0.15, p=.105
8	-0.55, p<.001	0.1, p=.301	0.06, p=.494
9	**-0.58, p<.001**	-0.14, p=.129	-0.14, p=.144
10	-0.38, p<.001	-0.26, p=.006	-0.27, p=.003
11	-0.15, p=.102	-0.33, p<.001	-0.26, p=.005
12	-0.19, p=.04	-0.3, p=.001	-0.23, p=.013
RoBERTa$_{\text{LARGE}}$			
0	-0.02, p=.808	-0.0, p=.959	-0.05, p=.6
1	-0.15, p=.121	0.15, p=.104	0.0, p=.966
2	-0.32, p<.001	0.27, p=.004	-0.07, p=.485
3	-0.36, p<.001	0.29, p=.001	0.0, p=.974
4	-0.29, p=.002	0.32, p<.001	0.22, p=.017
5	-0.14, p=.131	0.36, p<.001	0.17, p=.062
6	-0.16, p=.086	0.43, p<.001	0.2, p=.034
7	-0.21, p=.022	0.48, p<.001	0.33, p<.001
8	-0.2, p=.029	**0.5, p<.001**	0.31, p<.001
9	-0.26, p=.005	0.47, p<.001	0.29, p=.002
10	-0.17, p=.078	0.41, p<.001	0.16, p=.083
11	0.01, p=.923	0.36, p<.001	0.09, p=.325
12	-0.25, p=.007	0.42, p<.001	0.14, p=.123
13	-0.17, p=.078	0.32, p<.001	0.06, p=.508
14	-0.21, p=.022	0.18, p=.048	-0.0, p=.99
15	-0.23, p=.016	0.11, p=.238	-0.09, p=.316
16	-0.23, p=.014	0.13, p=.166	-0.06, p=.549
17	-0.21, p=.023	0.11, p=.254	-0.08, p=.376
18	-0.2, p=.029	-0.05, p=.589	-0.2, p=.033
19	-0.37, p<.001	-0.18, p=.059	-0.26, p=.006
20	**-0.48, p<.001**	-0.29, p=.002	-0.18, p=.057
21	-0.48, p<.001	-0.36, p<.001	-0.12, p=.188
22	-0.34, p<.001	-0.46, p<.001	-0.2, p=.028
23	-0.0, p=.974	-0.49, p<.001	**-0.34, p<.001**
24	-0.4, p<.001	-0.26, p=.005	-0.21, p=.025

Table 7: Layerwise correlation of human acceptability scores with semantic relatedness of a word with its context, as computed by RoBERTa$_{\text{BASE}}$ and RoBERTa$_{\text{LARGE}}$ models.

L.	Cosine	Cosine norm.	Spearman
FlauBERT$_{\text{BASE}}$			
0	-0.01, p=.937	0.15, p=.356	-0.05, p=.764
1	0.05, p=.763	0.19, p=.22	0.08, p=.623
2	0.03, p=.848	0.19, p=.221	0.05, p=.773
3	0.02, p=.901	0.22, p=.17	0.04, p=.796
4	0.02, p=.882	0.24, p=.126	0.07, p=.677
5	0.04, p=.825	0.24, p=.125	0.07, p=.673
6	0.01, p=.935	0.21, p=.187	0.05, p=.737
7	0.05, p=.736	0.2, p=.205	0.04, p=.793
8	0.06, p=.696	0.21, p=.181	0.1, p=.51
9	0.11, p=.491	0.15, p=.331	0.09, p=.562
10	0.09, p=.55	0.15, p=.347	0.09, p=.578
11	0.13, p=.404	0.21, p=.182	0.13, p=.401
12	**0.31, p=.049**	**0.35, p=.023**	**0.33, p=.031**
FlauBERT$_{\text{LARGE}}$			
0	-0.05, p=.765	0.17, p=.293	-0.06, p=.719
1	-0.08, p=.628	0.18, p=.241	0.04, p=.81
2	-0.11, p=.471	0.14, p=.375	-0.07, p=.647
3	-0.15, p=.352	0.09, p=.553	-0.05, p=.76
4	-0.32, p=.041	0.04, p=.79	-0.15, p=.359
5	-0.34, p=.026	0.1, p=.513	-0.12, p=.464
6	-0.35, p=.022	0.2, p=.207	-0.08, p=.634
7	-0.31, p=.045	0.4, p=.01	-0.01, p=.944
8	-0.32, p=.036	0.45, p=.003	0.09, p=.577
9	-0.41, p=.008	0.45, p=.002	0.06, p=.684
10	-0.39, p=.011	0.45, p=.003	0.12, p=.436
11	-0.4, p=.009	0.45, p=.003	0.16, p=.316
12	-0.38, p=.014	0.46, p=.002	0.15, p=.329
13	-0.38, p=.014	**0.5, p<.001**	0.14, p=.369
14	**-0.42, p=.006**	0.42, p=.005	0.12, p=.43
15	-0.39, p=.01	0.37, p=.016	0.13, p=.42
16	-0.41, p=.007	0.38, p=.014	0.16, p=.324
17	-0.35, p=.023	0.42, p=.006	0.19, p=.232
18	-0.31, p=.044	0.43, p=.005	0.14, p=.38
19	-0.35, p=.023	0.41, p=.008	0.15, p=.332
20	-0.32, p=.036	0.39, p=.011	0.13, p=.414
21	-0.24, p=.121	0.38, p=.013	0.12, p=.438
22	-0.23, p=.136	0.39, p=.01	0.13, p=.395
23	-0.21, p=.18	0.4, p=.01	0.11, p=.489
24	-0.25, p=.106	0.44, p=.003	0.15, p=.336

Table 8: Layerwise correlation of human acceptability scores with semantic relatedness of a word with its context, as computed by FlauBERT$_{\text{BASE}}$ and FlauBERT$_{\text{LARGE}}$.

Llama 3.1 8B (on English)			
L.	**Cosine**	**Cosine norm.**	**Spearman**
0	-0.01, p=.886	-0.16, p=.095	0.07, p=.464
1	0.02, p=.814	-0.01, p=.938	0.13, p=.164
2	-0.03, p=.749	0.19, p=.039	0.39, p<.001
3	-0.13, p=.157	0.34, p<.001	0.49, p<.001
4	0.06, p=.522	**0.54, p<.001**	**0.66, p<.001**
5	0.07, p=.449	0.5, p<.001	0.61, p<.001
6	0.13, p=.164	0.41, p<.001	0.51, p<.001
7	0.21, p=.026	0.43, p<.001	0.54, p<.001
8	0.28, p=.002	0.43, p<.001	0.54, p<.001
9	0.42, p<.001	0.4, p<.001	0.5, p<.001
10	0.42, p<.001	0.41, p<.001	0.52, p<.001
11	0.48, p<.001	0.4, p<.001	0.48, p<.001
12	0.51, p<.001	0.44, p<.001	0.51, p<.001
13	0.58, p<.001	0.48, p<.001	0.54, p<.001
14	**0.58, p<.001**	0.49, p<.001	0.54, p<.001
15	0.54, p<.001	0.48, p<.001	0.5, p<.001
16	0.52, p<.001	0.43, p<.001	0.47, p<.001
17	0.47, p<.001	0.42, p<.001	0.49, p<.001
18	0.38, p<.001	0.38, p<.001	0.45, p<.001
19	0.36, p<.001	0.38, p<.001	0.46, p<.001
20	0.32, p<.001	0.36, p<.001	0.47, p<.001
21	0.31, p<.001	0.35, p<.001	0.47, p<.001
22	0.29, p=.002	0.26, p=.005	0.41, p<.001
23	0.25, p=.007	0.19, p=.041	0.35, p<.001
24	0.22, p=.016	0.16, p=.08	0.33, p<.001
25	0.2, p=.03	0.13, p=.152	0.3, p=.001
26	0.19, p=.042	0.11, p=.222	0.29, p=.001
27	0.21, p=.022	0.1, p=.28	0.27, p=.003
28	0.15, p=.109	0.09, p=.328	0.29, p=.002
29	0.12, p=.194	0.07, p=.486	0.26, p=.005
30	0.11, p=.228	0.05, p=.563	0.25, p=.007
31	0.04, p=.648	0.06, p=.54	0.21, p=.023
32	-0.26, p=.004	0.07, p=.428	-0.2, p=.032

Table 9: Layerwise correlation of human acceptability scores with semantic relatedness of a word with its context, as computed by Llama 3.1 8B, on English.

Llama 3.1 8B (on French)			
L.	**Cosine**	**Cosine norm.**	**Spearman**
0	-0.04, p=.812	0.12, p=.478	-0.02, p=.885
1	0.04, p=.816	0.05, p=.779	-0.01, p=.943
2	-0.13, p=.444	0.29, p=.076	0.03, p=.867
3	-0.23, p=.17	0.06, p=.74	-0.01, p=.947
4	-0.29, p=.073	0.11, p=.525	0.08, p=.651
5	-0.21, p=.214	0.31, p=.059	0.16, p=.348
6	-0.06, p=.741	0.4, p=.013	0.19, p=.263
7	0.09, p=.596	**0.41, p=.01**	0.2, p=.228
8	-0.06, p=.703	0.37, p=.022	0.2, p=.239
9	-0.03, p=.859	0.3, p=.07	0.19, p=.263
10	0.06, p=.723	0.27, p=.099	0.18, p=.27
11	0.17, p=.31	0.3, p=.066	0.2, p=.234
12	0.12, p=.455	0.26, p=.121	0.14, p=.388
13	0.06, p=.725	0.25, p=.128	0.2, p=.238
14	0.12, p=.492	0.24, p=.141	0.2, p=.226
15	0.09, p=.608	0.15, p=.385	0.12, p=.472
16	0.04, p=.805	0.09, p=.609	0.01, p=.94
17	-0.01, p=.968	0.12, p=.478	0.11, p=.518
18	-0.09, p=.591	0.1, p=.569	0.14, p=.411
19	-0.14, p=.392	0.09, p=.605	0.13, p=.431
20	-0.22, p=.185	0.1, p=.555	0.1, p=.55
21	-0.23, p=.156	0.09, p=.581	0.04, p=.811
22	-0.24, p=.14	0.1, p=.541	0.02, p=.895
23	-0.25, p=.133	0.07, p=.655	-0.0, p=.996
24	-0.27, p=.107	0.15, p=.364	0.0, p=.98
25	-0.27, p=.097	0.1, p=.537	0.04, p=.806
26	-0.3, p=.065	0.17, p=.316	0.07, p=.695
27	-0.24, p=.139	0.16, p=.345	0.09, p=.611
28	-0.26, p=.117	0.16, p=.34	0.1, p=.559
29	-0.26, p=.113	0.17, p=.316	0.11, p=.524
30	-0.33, p=.04	0.14, p=.393	0.02, p=.888
31	-0.29, p=.074	0.1, p=.549	-0.1, p=.541
32	**-0.33, p=.046**	0.09, p=.609	-0.21, p=.21

Table 10: Layerwise correlation of human acceptability scores with semantic relatedness of a word with its context, as computed by Llama 3.1 8B, on French.

E Analysis of Model Scoring

Figure 4 demonstrates the surprisal score distribution in French for three types of senses as described in Section 3.1. The data on human evaluation is absent from this figure, since it was not obtained by the authors of the initial psycholinguistic experiment (Lombard et al., 2023).

Figure 5 shows surprisal score distribution in French as per regularity degree, along with human acceptability judgment.

Table 11 reveals the relationships of the models with the degree of pattern regularity, while Figure 4 shows correlation with word frequency and word length, as compared to the human evaluators.

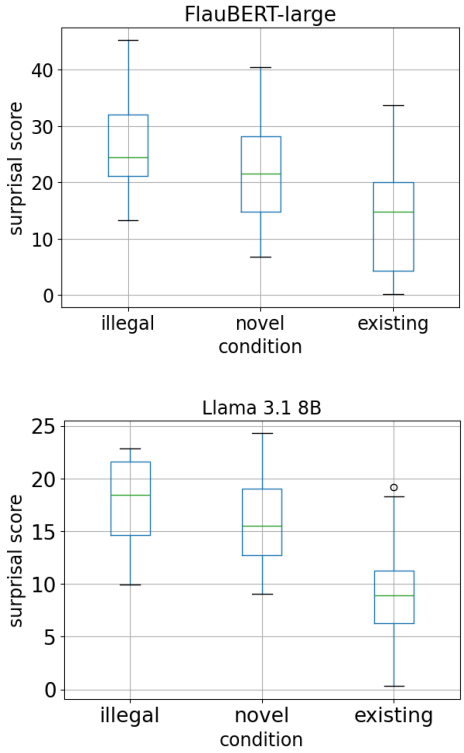

Figure 4: Distribution of French surprisal scores by condition labeled in the original dataset as *new*, *illegal* and *existing*. These correspond to the groups (1), (2) and (3) described in the Section 3.1

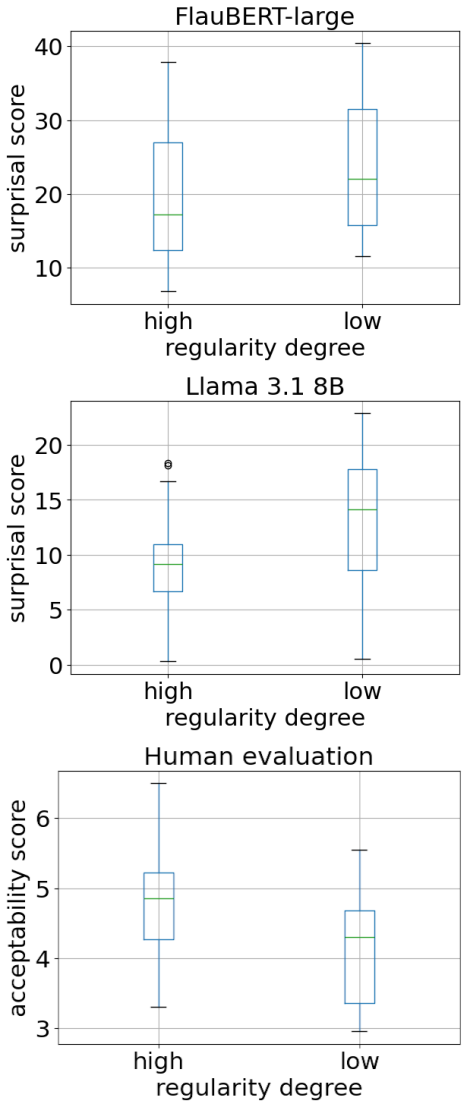

Figure 5: Distribution of surprisal scores by regularity degrees labeled as *high* and *low* for the French data.

Models	R1	R2	R3	R4
RoBERTa$_{\text{LARGE}}$ SURP	-0.36*	-0.43*	-0.18	**-0.44***
Llama 3.1 8B SURP	-0.26*	**-0.43***	-0.01	-0.3*
RoBERTa$_{\text{LARGE}}$ SIMIL	-0.08	-0.17	-0.22	**-0.27***
Llama 3.1 8B SIMIL	0.03	-0.12	-0.03	-0.08
Human acceptability rating	0.28*	**0.53***	0.00	0.31*

Table 11: Correlation (Pearson) between the regularity metrics and scores from surprisal, similarity and human evaluation. The regularity metrics are defined in Table 1. Asterisk (*) indicates a statistically significant correlation ($p < .05$). Since the high human acceptability is associated with low model surprisal, the negative correlation with model-derived metrics is expected.

Limitations

Languages. In our experiments, we mainly focus on English data. The size of the dataset in French often did not allow us to make meaningful comparisons or confirm the validity of results received from English. As shown by Srinivasan and Rabagliati (2015), polysemy patterns and their degree of regularity overlap only partially across languages. Hence, studying the phenomenon of continuous regular polysemy in one language has limited generalization potential. More languages from diverse language families need to be involved in such investigations.

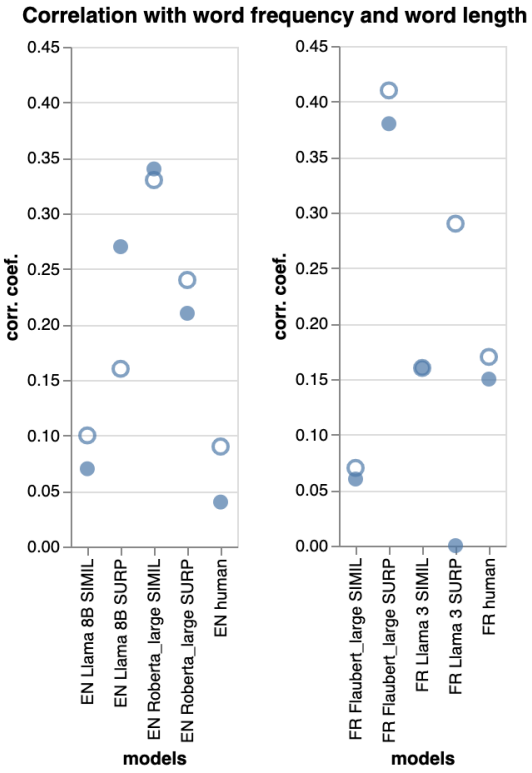

Figure 6: Correlation (Pearson) of word frequency and word length with the measures of semantic relatedness, surprisal and human rating. The empty circles represent word length, while the filled circles represent word frequency.

Model size. Regarding the models represented in our study, it would be interesting to compare masked BERT-based varieties with an autoregressive model that actually matches their size (such as GPT-2). Certain differences in polysemy processing were difficult to interpret since both varying architectures and wide-ranging model sizes were involved.

Model architecture. It is possible that the weaker correlation with human judgment achieved by the autoregressive language models is due to the initial psycholinguistic experiment design that favors bidirectional language models over unidirectional ones. Precisely, the authors first introduce the participants to a sentence with the target word masked. Afterwards, the participants see the full sentence and are asked to judge its plausibility on a scale from 1 to 100 by moving the cursor. This way, the human evaluators have a chance to consider the context on both sides before deciding on plausibility. In the surprisal setting, having access to the left context only, the autoregressive Llama does not mimic this behavior and does not align as closely with human estimation. Another argument in favor of this idea is that the subsequent analysis of the correlation of surprisal scores with the different regularity metrics showed that Llama delivers identical results to human evaluators for almost all regularity metrics. At the same time, RoBERTa, while generally following the correlation pattern, does not align as closely (see Figure 3 and Table 11, Appendix E). This suggests that our experimental design might poorly accommodate this model type, and we need other means of establishing whether the sensitivity to scalar regularity in models of different architectures varies.

Methodology. In order to understand the effect of such variables as word length, word frequency, and pattern regularity degree on the model scoring, we conducted correlation analyses and were able to gain insights from them. However, a deeper understanding of each variable's contribution to the results requires a regression study, which we plan to conduct as part of future work.