A Grounded Typology of Word Classes

Coleman Haley[A] **Sharon Goldwater**[A] **Edoardo Ponti**[A][🖼]
[A]University of Edinburgh [🖼]University of Cambridge
{coleman.haley,sgwater,eponti}@ed.ac.uk

Abstract

We propose a grounded approach to meaning in language typology. We treat data from perceptual modalities, such as images, as a language-agnostic representation of meaning. Hence, we can quantify the function–form relationship between images and captions across languages. Inspired by information theory, we define "groundedness", an empirical measure of contextual semantic contentfulness (formulated as a difference in surprisal) which can be computed with multilingual multimodal language models. As a proof of concept, we apply this measure to the typology of word classes. Our measure captures the contentfulness asymmetry between functional (grammatical) and lexical (content) classes across languages, but contradicts the view that functional classes do not convey content. Moreover, we find universal trends in the hierarchy of groundedness (e.g., nouns > adjectives > verbs), and show that our measure partly correlates with psycholinguistic concreteness norms in English. We release a dataset of groundedness scores for 30 languages. Our results suggest that the grounded typology approach can provide quantitative evidence about semantic function in language.

1 Introduction

Within linguistics, *typology* is the subfield focused on the study of patterns and variation across the world's languages (Croft, 2002, pp. 1–2). To identify such patterns, linguists must carefully identify phenomena of interest within languages, and then align them with one another. For example, vowels exist in a continuous acoustic and perceptual space, without clear boundaries between them. To define vowel categories and align systems across languages, linguists rely largely on acoustic properties of the speech signal—reducing the problem to a physically grounded, empirical one (Liljencrants et al., 1972; Cotterell and Eisner, 2017).

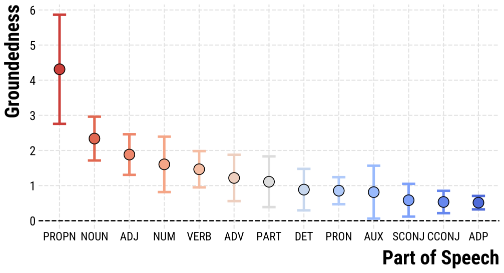


Figure 1: Mean and standard deviation of per-language mutual information estimates between word class and image. Across 30 languages, we see clear and consistent tendencies about which parts of speech are more "grounded", corresponding to a distinction between lexical and functional classes.

While empirically grounding language form (surface structure like vowels) is typically straightforward, language is not just a formal system, but also a functional one. Many questions within typology relate to the relationship between form and *meaning*, especially in domains like morphology and syntax. Typically, typologists manually identify semantic/functional roles such as "subject", and "causative" and study their expression across languages (Haspelmath, 2010; Greenberg, 1966). Unlike with many definitions based on form, definitions based on meaning are left up to subjective discretion, leading to debates which reduce to the definition of particular terms cross-linguistically (Haspelmath, 2007, 2012; Plank, 1994).

Instead, we propose a "grounded" approach to typology, which (under certain assumptions), allows the quantification and cross-linguistic comparison of language function and semantics across languages. By looking at sentences produced as captions of the same image across languages, we can use the image as an evidence-based, language-agnostic representation of the shared semantics underlying these utterances, similar to the evidence-based acoustic signal in the study of vowel spaces.

In this work, we specifically focus on semantic

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 10380–10399
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

contentfulness—how semantically informative a given word token is. We introduce a way to empirically quantify contentfulness, *groundedness*, which relies on vision-and-language models. Groundedness measures how much less surprising a word is when we know the perceptual stimuli (i.e., the image) it describes. This *surprisal difference* between the surprisal of the word token in an image captioning model versus its surprisal in a language model is an estimate of the pointwise mutual information: the greater this difference (LM > captioning), the more *grounded* the word is in that context.

As a case study, we apply this measure to the study of the typology of word classes ("parts of speech"). Literature from cognitive, pyscho- and neurolinguistics all point to contentfulness being an organizing factor in word class processing and even formation and structure: low-content (functional) word classes have many different properties from high-content (lexical) classes (Dubé et al., 2014; Bird et al., 2003; Chiarello et al., 1999). Yet, there has been no cross-linguistic study of the relationship between contentfulness and word class.

Using our groundedness measure to quantify semantic contentfulness, we can estimate the mutual information of a word class with a caption's meaning (image). We find our measure largely rediscovers the distinction between lexical and functional word classes across 30 languages. Further, though it correlates only weakly with psycholinguistic norms for imageability and concreteness in English, it provides an intuitive ranking (noun > adjectives > verbs) across languages. On the other hand, it contradicts the view of adpositions as a "semi-lexical" class (Corver and Riemsdijk, 2001) and suggests grammatical word classes do carry some semantic content. These results thus partly validate and partly falsify received wisdom about word class contentfulness. They suggest the utility of this measure as a general tool for studying contentfulness in linguistics, and of taking a grounded approach to typological problems. We release the model used to estimate our measure and a dataset of groundedness values in 30 languages.[1]

2 Background

An excellent example of the relevance of the relationship between semantic function and linguistic form to typology is *word classes*. Within a particular language, there are typically groups of words unified by the (formal) contexts in which they can appear. Further, this distribution of words is not arbitrary, but unified by a particular semantic prototype. For example, in English, nouns are a class of words which prototypically denote physical objects or things and can follow words like "*the*", "*this*", and "*that*". However, not all languages have words like "*the*", and so an equivalent formal–structural criterion cannot be given (Haspelmath, 2012). On the other hand, semantic criteria are not sufficient to describe these classes: most languages can express prototypical verb or adjective meanings with the syntactic distribution of a noun.

The elusiveness of a cross-linguistic definition for word classes leads to many debates about particular languages "having" or "not having" a distinction between (e.g.) nouns and verbs on the basis of a mix of formal and semantic criteria (cf. Kaufman, 2009; Hsieh, 2019; Richards, 2009; Weber, 1983; Floyd, 2011). Here, we investigate word classes as operationalized in a framework where there is a fixed set of *universally applicable* word classes, as set out in the Universal Dependencies project (de Marneffe et al., 2021).While this is problematic in general, our aim is not to claim that the assignment of word classes is precisely correct, but rather to empirically and quantitatively investigate the functional/semantic dimension of this common operationalisation of word class. In future work, we aim to investigate the relationship between these measures and non-prototypical parts of speech.

2.1 Contentfulness and word class

In this work, we focus on the related distinction between lexical/contentful word classes (e.g. nouns, verbs, and adjectives) and functional/grammatical word classes. Functional word classes are typically closed-class, meaning they do not admit new members and typically do not exhibit rich productive morphology; they tend to express highly grammatical and abstract meanings. Lexical classes are typically open class, productively admitting new members, and their meanings tend to be more concrete and contentful (Corver and Riemsdijk, 2001).

Complications about these generalized categories and tendencies abound, however. For example, in some languages like Jaminjung, prototypically lexical categories like verbs are closed class (Schultze-Berndt, 2000; Pawley, 2006). Further, both the abstraction and semantic contentfulness of particular members of a given word class can be quite variable. For example, a noun like "*factor*"

[1] `https://osf.io/bdhna/`

has a highly abstract meaning, while the meaning of the preposition "*to*" is intuitively more abstract than the preposition "*above*", despite belonging to the same, "abstract" grammatical word class. Further, over time words can change in both their contentfulness and even word class through processes like grammaticalization (Bisang, 2017).

Nevertheless, the complex relationship between contentfulness and word class remains unexplored through a cross-linguistic empirical lens—perhaps due to the difficulties of measuring such properties.

2.2 Measuring contentfulness

The relationship between contentfulness and word class has not been explored cross-linguistically; however, a significant literature within the language sciences has investigated related concepts.

While theoretical linguistics has focused on a distinction between content and function words, psycholinguistics has focused on semantic dimensions like imageability, concreteness, and strength of perceptual experience. Measures of these dimensions have relied on subjective, decontextualized human judgements, but nevertheless predict processing differences between word classes, such as asymmetries in the processing of nouns and verbs in certain aphasias (Bird et al., 2003; Dubé et al., 2014; Lin et al., 2022). Because we operationalize meaning as images, notions such as imageability seem especially related to our groundedness measure. However, as discussed in Section 5.4, these concepts differ from our measure in that informativity is not a major factor in their definition. For example, while both "zebra" and "woman" are highly concrete nouns, the former has higher groundedness on average, because although both are often strongly associated with an image, "zebra" is more informative/surprising, especially if the image is unavailable—thus, the image adds more information in that case.

As shown by the prior example, our measure is also closely related to another concept widely studied in computational psycholinguistics: *surprisal.* Like our groundedness measure, surprisal has an intuitive link to contentfulness from an information theoretic perspective, and has been extensively studied in relation to processing difficulty (Hale, 2001; Levy, 2008; Smith and Levy, 2013; Wilcox et al., 2023; Staub, Forthcoming). However, surprisal entangles formal and functional information in language. As such, cross-linguistic comparisons based on surprisal are challenging, since form is language specific (Park et al., 2021). We aim to focus on information due to language *function*, separated from form. Surprisal must also encode grammatical uncertainty (alternative ways of expressing the same meaning like "knight" and "cavalier"), as opposed to surprisal due only to what meanings are being expressed. Our image captioning model quantifies how many bits of information remain after the meaning is known. Our measure then quantifies how much of the LM surprisal is explained by the meaning (image).

3 Method

In this section, we define a token's *groundedness*, and show how we can use this to estimate the mutual information between parts of speech and representations of meaning. Let the set of word types in a language be $\mathcal{W}$. We assume a model of the data generation process where given a meaning m, a sentence is constructed by iteratively sampling a word $w_t \in \mathcal{W}$ conditioned on m and previous words $\mathbf{w}_{<t}$. As mentioned previously, the groundedness of a token is given by its pointwise mutual information (PMI) with the meaning.

$$\mathrm{PMI}(w_t; m \mid \mathbf{w}_{<t}) = \log \frac{p(w_t \mid m, \mathbf{w}_{<t})}{p(w_t \mid \mathbf{w}_{<t})} \quad (1)$$

As we cannot access the true meaning m, we must approximate it with a proxy. A good proxy for m should be language-neutral, and will make estimating the probabilities in Equation 1 straightforward across languages. In this work, we focus on *images* as a language-neutral representation of meaning. Images capture rich, language-independent information about the world state described by an image, and have proved useful as a method for aligning meanings across languages (Rajendran et al., 2016; Gella et al., 2017; Mohammadshahi et al., 2019; Wu et al., 2022). Further, a major strength of images as a meaning representation is that estimating both quantities in Equation 1 becomes straightforward with neural models: $p_{\boldsymbol{\phi}}(w_t \,|\, m, \mathbf{w}_{<t})$ corresponds to the probability of the token under an image captioning model, while $p_{\boldsymbol{\theta}}(w_t|\mathbf{w}_{<t})$ corresponds to its probability under a language model.

Using images as a representation of meaning does have some implications for our approach. For instance, verbs, which usually denote events and are more temporally unstable (Givon, 1984) than other parts of speech, may be less grounded than with a different meaning representation, such as videos. Further, the language of image captions is somewhat restricted in terms of grammatical structure and lexical items, making the analysis of long-tail phenomena or highly abstract language challenging (Ferraro et al., 2015; Alikhani and Stone, 2019). Future work could use our framework to explore other meaning representations, such as symbolic models or videos (though doing so involves overcoming further dataset and modeling challenges). Still, the language-neutral nature and rich information content of images allows us to study groundedness for a wide range of words, languages, and linguistic contexts.

Noting that a model's surprisal is negative log probability, we can view groundedness as a *difference in surprisal*, corresponding to how much more expected the token is under the grounded model than under the textual model. As such, the PMI should rarely take on negative values—because the captioning model has more information (both image and text) than the language model (text only). However, some tokens, such as those that are highly grammatical or structural, should be close to 0.

In this work, we study the groundedness of *word classes*. Drawing inspiration from functionalist typology, we treat a word class $\mathcal{C}_i$ as a label selected by a linguist for a word in its context. We make an assumption that this label is independent of our meaning representation given a word's context, allowing us to define the following joint distribution:

$$p(\mathcal{C}_i, m \mid \mathbf{w}_{<t}) = \sum_{w_t \in \mathcal{W}} \big[p(\mathcal{C}_i \mid w_t, \mathbf{w}_{<t}) p(w_t, m \mid \mathbf{w}_{<t}) \big]. \quad (2)$$

We can then formulate the mutual information between a word class and meaning as the expected value of the PMI between each token labeled with that class, and the token's associated image:

$$I[\mathcal{C}_i; m | \mathbf{w}_{<t}] = \mathop{\mathbb{E}}_{p(\mathcal{C}_i, m, \mathbf{w}_{<t})} \left[\log \frac{p(w_t | \mathbf{w}_{<t}, m)}{p(w_t | \mathbf{w}_{<t}))} \right]. \quad (3)$$

Given our factorization of the joint, we can perform a Monte Carlo estimation of the expectation by simply averaging groundedness over all the tokens tagged with $\mathcal{C}_i$ in the data $\mathcal{D}$:

$$\hat{I}[\mathcal{C}_i; m \mid \mathbf{w}_{<t}] = \sum_{(m, \mathbf{w}_{<t}) \in \mathcal{D}} \frac{\mathbb{1}_{\mathcal{C}_{w_t} = \mathcal{C}_i} \log \frac{p_{\phi}(w_t | \mathbf{w}_{<t}, m)}{p_{\theta}(w_t | \mathbf{w}_{<t})}}{\sum_{w_t \in \mathcal{D}} \mathbb{1}_{\mathcal{C}_{w_t} = \mathcal{C}_i}} \quad (4)$$

where $\mathbb{1}_{\mathcal{C}_{w_t} = \mathcal{C}_i}$ is 1 when a token's class is C_i and 0 otherwise. We note that our groundedness measure and our mutual information estimates are conditional on *linguistic context*. As such, words which are very grounded in one context could be hardly grounded in another, due to disambiguating information in the preceding context. Some information about m will be generally conveyed by $\mathbf{w}_{<t}$; however, our mutual information estimates are aggregated over all contexts in which a word class occurs, and on average this contribution is small.

Model	Gemma PT	PaliGemma CT	COCO-35L FT
Img. Cap.	**A**	🖼**A**	🖼**A**
LM	**A**	🖼**A**	**A**

Table 1: We match the data points on which the language model and image captioning model were trained. The three datasets are the Gemma pre-training mixture (PT), PaliGemma multimodal data for continued training (CT), and COCO image–caption pairs for fine-tuning (FT). Symbols indicate whether models are trained on text data (**A**) or on multimodal data (🖼**A**).

4 Experimental setup

Captioning model $p_{\phi}(w_t \mid \mathbf{w}_{<t}, m)$ As our image captioning model, we use the recently released PaliGemma model (Beyer et al., 2024). This model is by far the state-of-the-art among publicly available multilingual image captioning models. PaliGemma consists of an image encoder, initialized from the SigLIP-So400m model (Zhai et al., 2023), and a transformer decoder language model, initialized from the Gemma-2B language model (Gemma, 2024). A linear projection maps from the image encoder space to a sequence of 256 tokens in the language model's embedding space. The whole system is then trained on a mix of vision-and-language datasets, including the unreleased WebLI dataset with 10 billion image-caption pairs in 109 languages (Chen et al., 2023), and the CC3M-35L dataset consisting of 3 million image-caption pairs in each of 35 languages (Thapliyal et al., 2022).

While PaliGemma is a general-purpose vision-and-language model, it is designed to be fine-tuned on and applied to individual tasks. As such, we use the open-source `paligemma-3b-ft-coco35-224` checkpoint for multilingual captioning, which has been fine-tuned on COCO-35L.

Language model $p_{\theta}(w_t \mid \mathbf{w}_{<t})$ Our aim is to use a language model as similar to our captioning

model $p_\phi(w_t|\mathbf{w}_{<t}, m)$ as possible. This is critical to getting good (P)MI estimates, which relies on estimating a difference in surprisal between the two models. If the language model is not adapted to the image captioning domain, it may under-estimate the probability of particular words, leading to an over-estimation of mutual information. We therefore aim to *match* the training data between the language model and image captioning model, such that they see the same set of captions.

To do so, we initialize our language model with the weights from the pretrained PaliGemma model `paligemma-3b-pt-224`. However, out of the box, the decoder behaves degenerately when no image is provided, so we need to adapt the model to not expect image information and to match the training data of the captioning model. To do so, we fine-tune the language model on the *captions only* from the COCO-35L dataset. In this way, we ensure the models have observed the same data during training and are adapted to the same domain, and are thus maximally comparable. Table 1 summarizes the data matching between the two models. Further implementational and POS tagging details are in Appendix A.

Evaluation Datasets We also need multilingual image captioning datasets for evaluation which are not observed during training. For this, we measure groundedness on three separate datasets, each with its own strengths and weaknesses. First, we use **Crossmodal-3600**. This dataset includes captions for 3,600 images across a range of cultures, manually captioned by fluent speakers of 36 typologically diverse languages. However, it is relatively small per language compared to other datasets. Further, the independence of the captions means that there is greater diversity in what aspects of an image are being described across languages (Liu et al., 2021; Ye et al., 2024; Berger and Ponti, 2024).

Our second dataset, the validation set of **COCO-35L**, addresses several of these issues. It is larger, with 5 captions each for 5000 images and 35 languages,[2] yielding 25,000 captions per language. Further, the captions are machine translations of each other, ensuring more comparable semantic content across languages (Beekhuizen et al., 2017) at the expense of centering the perspective of English speakers and machine translation issues.

Finally, we consider **Multi30K**. This dataset comprises 30,000 images captioned 5 times each in English, with a single caption per image manually translated into French, German, Czech, and Arabic. This dataset is therefore large on the individual language level, but with limited language coverage. It has the comparability of being translated and the trustworthiness of human translation, but may still be vulnerable to translationese. By looking at all three of these datasets for similar generalizations about the relationship between groundedness and part of speech, we obtain a picture that is robust to the weaknesses of the individual datasets.

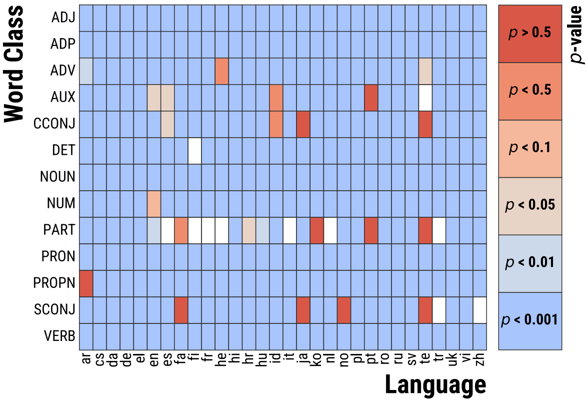


Figure 2: Heatmap of mutual information estimates across parts of speech in thirty languages. Cells show the statistical significance of a word class's groundedness (MI > 0). Unattested classes are white. Some functional classes display non-significant levels of groundedness in several languages, while lexical classes dominantly show highly significant grounding.

5 Results

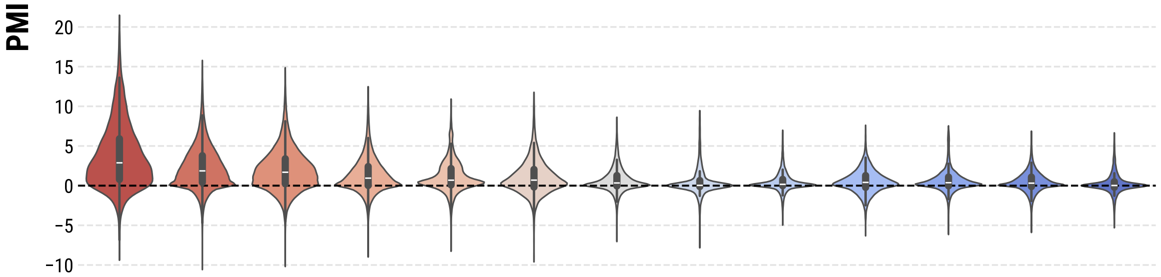
The following sections quantitatively investigate the trends in our groundedness measure across languages and word classes. We begin by examining which word classes exhibit significant groundedness (Section 5.1), followed by an analysis of cross-linguistic trends and their consistency (5.2 and 5.3). Finally, we relate our findings to contentfulness-related psycholinguistic norms (5.4).

5.1 Which word classes are grounded?

We first investigate the evidence for groundedness in each word classs—that is, for each part of speech, we ask whether its estimated mutual information with the image is significantly greater than zero.

To compute significance levels, we use a one-sample permutation test. Taking the set of PMIs for a part of speech (POS) in a language, we sample up to 500 PMIs at a time from all datasets and

[2] Crossmodal-3600 and COCO-35L cover the same languages with the exception of Quechua.



POS	PROPN	NOUN	ADJ	VERB	NUM	ADV	PRON	PART	AUX	CCONJ	DET	ADP	SCONJ
Mean	3.82	2.46	2.02	1.67	1.47	1.23	0.77	0.76	0.67	0.56	0.54	0.53	0.48
Rank	1	2	3	4	5	6	7	7	8	9	10	11	12

Figure 3: Word token level distributions of the groundedness measure (PMI) across all languages and datasets, grouped by part of speech (word class). We also report the estimated marginal mean and ranking of each word class. Colors are based on the ranking of classes, rather than their average PMIs. Overall, the distribution and estimated ranking of word classes strongly suggest our groundedness measure quantitatively captures the distinction between lexical and functional classes.

randomly permute their signs (assign + or - with equal probability to each PMI value), then average these values to produce a new estimate of mutual information (MI). We repeat this process to produce 10^5 permuted estimates. By measuring how often our estimate based on the observed data is greater than the permuted estimate, we obtain the p-value,[3] i.e., the probability that our observations would have occurred under the null hypothesis of MI = 0.

Results are shown in Figure 2. Overall, the results suggest most or all word classes contribute some information about the image they describe—in line with theories in linguistics that emphasize the lexical aspects of categories which are traditionally considered functional (Corver and Riemsdijk, 2001; Bisang, 2017). Interestingly, subordinating and coordinating conjunctions do not consistently reject the null hypothesis, suggesting there is little evidence the image is informative for how many clauses a speaker uses to describe an image.

5.2 Which word classes are more grounded?

We hypothesize that the cross-linguistically consistent trends in word class groundedness correspond to a cline which is a continuous analogue of the lexical–functional word class distinction. To isolate the contribution of word class identity to mutual information cross-linguistically, we compute estimated marginal means (EMMs) for each word class's groundedness,[4] and perform a post-hoc pairwise comparison test of the means.[5] The results of this analysis are displayed in Figure 3. All pairwise comparisons except between pronouns and particles are statistically significant, leading to a near total ranking of word classes. We find that lexical word classes (Proper nouns, nouns, adjectives, verbs, numbers, and adverbs) have higher groundedness than functional word classes (particles, auxiliaries, conjunctions, determiners, and adpositions), with pronouns ranking together with particles at the upper end of the functional categories. The ranking corroborates ideas from cognitive linguistics which place nouns, adjectives, and verbs along a lexical–functional continuum, with nouns > adjectives > verbs (Rauhut, 2023). On the other hand, it does not neatly align with ideas in linguistic theory about adpositions as a semi-lexical class (Corver and Riemsdijk, 2001), which suggest they should behave more like other lexical classes compared to functional classes. Instead we see similar or greater mutual information for other functional classes, suggesting they could be more meaning-bearing than traditionally viewed.

5.3 How consistent is word class groundedness across languages?

We quantify the strength of the association between groundedness and word class on two levels: language-level MI estimates (Figure 1), and token-level PMI (Figure 3). The first level quantifies how consistent languages are in the groundedness of word classes, while the second level quantifies how much word class drives the groundedness of individual tokens. In both cases, we use ANOVA to estimate the amount of the variance in groundedness explained by word class.

[3] We use the Benjamini and Yekutieli (2001) corrections.

[4] Averaged over values of language and dataset.

[5] Using Šidák corrections; significance threshold = 0.01.

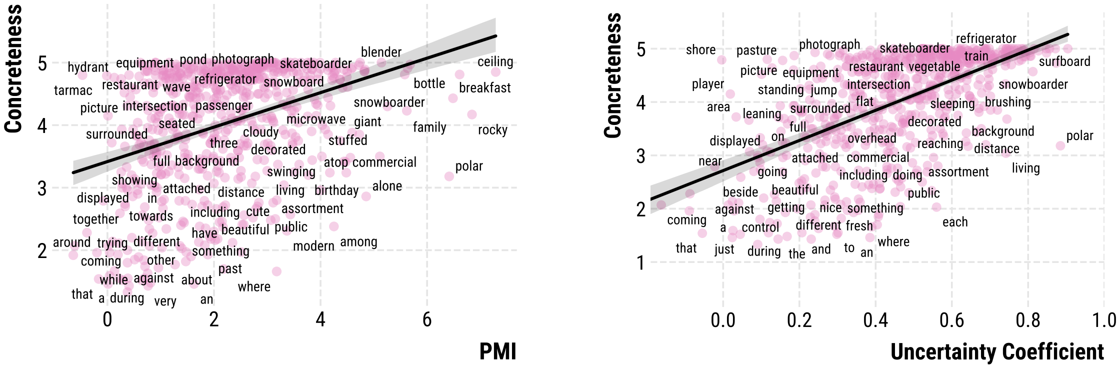


Figure 4: Correlation between human concreteness ratings and type-level groundedness (PMI; left, $\rho = 0.368$) or uncertainty coefficent (right, $\rho = 0.609$): i.e., the average ratio between LM surprisal and captioning model surprisal.

MI estimates For the language-level MI estimates in Figure 1, we consider the separate effects of language, dataset, and POS on groundedness. Because the meanings (images) are matched across languages, this allows us to estimate and control for some languages having consistently larger or smaller MI estimates (due to language-specific variation in our neural estimators). We find significant effects of all 3 factors, but they differ dramatically in how much variation they explain. The effect of dataset is extremely small, explaining 0.5% of the observed variance ($F_{3,816} = 5.71$, $p < 0.01$). Language identity has a larger effect, explaining 8.2% of the variance ($F_{29,789} = 6.42$, $p < 0.001$). However, word class dominates, explaining most of the total variance (57.3%, $F_{12,806} = 775$, $p < 0.001$), and 62.8% of the remaining variance after controlling for variance due to dataset and language. Altogether, these factors explain 65.6% of the variance, leaving the remaining variance to cross-linguistic differences in the MI of specific parts of speech.

PMI distributions We also investigate how much variation in the full distribution of contextual groundedness estimates (PMIs) is explained by word class (shown in Figure 3). Within a POS, groundedness is expected to vary substantially: for example, some (concrete, visually distinct) nouns have much higher PMI with the image than others, and tokens of the same word type also have different groundedness (e.g. "lot" referring to a location vs. "lot" as a quantity expression) Therefore, we expect word class to explain much less variance than in the overall MI estimates. Language, dataset, and their interaction account for 2.4% of the total variation in PMIs across the three datasets ($F_{64,10^7} = 4727$, $p < 0.001$). Word class accounts for 12.0% of the total variation ($F_{12,10^7} = 123583$, $p < 0.001$). Additionally, the interaction between word class and language (cross-linguistic variation in the means of word classes) accounts for only an additional 1.6% of the total variation ($F_{330,10^7} = 602.5$, $p < 0.001$), despite having many degrees of freedom. So cross-linguistically consistent tendencies comprise the bulk of the explainable variance in the overall PMI distribution across these three datasets—5 times as much as language and dataset, and 7.5 times as much as language differences in POS groundedness.[6]

5.4 Semantic dimension of the measure

In this section we explore the semantic properties of the groundedness measure introduced here, comparing it to semantic norms related to contentfulness that are widely used in psycholinguistics. One potential advantage of our method is the ease with which it allows the rating of individual word tokens in context; however, existing ratings tend to be for words in isolation (word types). We focus our analysis here on English and on word types which occur at least 30 times in the COCO(-35L)[7] validation set, averaging across occurrences to obtain an estimate of the average type-level groundedness.

We compare to three different psycholinguistic norms: imageability, concreteness, and strength of visual experience. Such norms are measured by providing a definition and examples of low- and high-value words to raters, who then rate words on a Likert Scale. For imagability, we use the Glasgow Psycholinguistic Norms (Scott et al., 2019). For concreteness, we use the Brysbaert et al. (2014) norms. For strength of visual experience, we use

[6] The token-level interaction models and their ANOVA statistics are computationally intensive (512GB RAM; 6hrs).

[7] While COCO-35L is mostly machine translated data, the English data is fully human generated.

the Lancaster Sensorimotor Norms (Lynott et al., 2020). Results for concreteness are shown in Figure 4 (left). We observe fairly weak (though significant, $p < 0.001$) correlations with groundedness using Spearman's ρ (Imageability: $\rho = 0.288$, Concreteness: $\rho = 0.368$, Visual strength: $\rho = 0.212$).

We find these weak correlations are partly due to to the *informativity* aspect of our measures, which seems not to play as large of a role in human ratings (e.g. woman is just as concrete as skateboard, but less informative and also less grounded by our measure). To account for differences in baseline (LM) word informativity, we can normalize the PMI scores by the LM surprisal, yielding the uncertainty coefficient (Theil, 1970): the proportion of the LM surprisal explained by the PMI. Regressing this value against the psycholinguistic norms, stronger correlations emerge (Imagability: $\rho = 0.548$, Concreteness: $\rho = 0.609$ as shown in Figure 4 (right), Visual strength: $\rho = 0.320$). This suggests that the differences between groundedness and surprisal are associated with concreteness. However, this measure collapses differences between word classes in overall informativity/surprisal.

In some cases, outliers are due to contextual effects. For example, in our data the word "polar" (high groundedness, moderate concreteness) occurs exclusively as the first word in the multiword expression "polar bear" which is highly concrete, imageable, and visual; while ratings based on the word type are for the more abstract geographical concept. Other words with divergent scores between human-based and model-based methods tend to be those which frequently occur in contexts where they are highly expected (e.g. "shore" which tends to occur in limited syntactic contexts and after the appearance of words like "boat," "lake," or "surfers"), or words which are often used nonspecifically in the image captioning context (e.g. "photo" exhibits very low PMIs, because captions frequently begin with "A photo of . . . ").

6 Discussion and Conclusion

We have proposed a grounded approach to typology, using images as a proxy for sentence meaning. Using information theory and neural models, we define *groundedness*, a measure of a token's association with the meaning expressed in a sentence Our results demonstrate that word classes display consistent patterns in terms of their groundedness across a typologically diverse sample of languages. We find these patterns can be described as a continuous cline which generalizes the traditionally dichotomous distinction between lexical and functional word classes into a gradient one. However, our results suggest grammatical word classes still carry semantic content. We find that nouns > adjectives > verbs, in line with a view of these classes as a continuum; yet, our results contradict claims that adpositions are more lexical than other functional classes. Our measure is related to surprisal, but diverges from it, particularly for concrete words.

While this work has focused on word classes, groundedness enables the exploration of other aspects of how languages express function through form. Future work could investigate in detail under what conditions "functional" items have higher groundedness. For example, do more spatial adpositions and determiners have higher groundedness than less spatial ones? Humans tend to have difficulty scoring highly abstract and grammaticalized words, and getting contextual scores is difficult with existing psycholinguistic approaches: groundedness opens new ways to address these questions.

Our approach is also suitable for studying nonprototypical word class organizations, such as languages which do not clearly distinguish between adjectives and verbs (Korean; Maling and Kim, 1998), or languages that split individual word classes into distinct sub-classes (Japanese adjectives; Backhouse, 1984). Future work should look at both formal and semantic sub-classes of parts of speech—such as gerunds, participles, and different semantic classes of verbs (as in VerbNet; Kipper Schuler et al., 2009)—investigating their groundedness and how it aligns with or varies from existing metrics. In particular, we conjecture that boundary classes (e.g. gerunds) may display intermediate groundedness (between nouns and verbs) compared to prototypical members of those classes. Groundedness makes it possible to test this conjecture with reference to the contexts in which words appear, which is needed for distinguishing syncretic forms.

Our approach can also cover any classes which can be defined over linguistic units, such as morphemes, phrases, or semantic classes. For instance, future work could explore the claim that inflections are more "grammatical" than derivations (Booij, 2007; Haley et al., 2024). Similarly, our measure could be used to study the lexicalization or grammaticalization of constructions (as a decrease in groundedness over time). To support such work, we release our groundedness scores online.[8]

Going beyond the details of the approach here, our work generally suggests a role for multimodal models in computational typology similar to the one played by language models in the past decade (e.g. Pimentel et al., 2023; Cotterell et al., 2018; Ackerman and Malouf, 2013). While language coverage remains more limited than text models, the latest multimodal models and datasets cover enough typologically and culturally diverse languages to make them worth studying—and we anticipate coverage will only improve. Further, the ability of multimodal models to provide an empirically grounded (if imperfect) representation of meaning makes them uniquely valuable for quantitatively addressing questions about the relation between form and function in language. Our work provides the first study of this kind, and we hope that by demonstrating the utility of this approach and releasing our groundedness scores we will inspire other researchers to follow suit.

Limitations

Our approach has a number of important limitations. These limitations should inform the interpretation of results here, as well as any future studies considering using these techniques.

First, our operationalisation of meaning as an image is necessarily a simplification and has numerous implications for our results. Notably, the choice of images rather than videos (motivated by model quality and availability) as the representation of meaning has major implications for verbs, which tend to have meanings which are more temporally extended. This choice also has substantial implications about the variety of language which can be analyzed–many types of language use, such as metaphoric extension, are likely to be much less frequent in image captions than in other domains of language use: such phenomena are perhaps best studied using a different technique. This problem is compounded by the fact that existing multilingual corpora for these datasets remain fairly small–thus the analysis of long-tail phenomena in language using these methods is likely not yet possible.

Compared to existing methods in typology, this method trades human effort for computational resources. While we make both our models and data available, significantly lessening the burden on future studies, the models here contain between two and three billion parameters, and the image models have very long sequence lengths due to the image tokens. Inference on new data is therefore fairly expensive with current technologies.

Further, there remain significant limitations on the languages which can be studied with these approaches. Currently available models cover just 16 languages outside of the Indo-European language family, and entire areal typological regions like the Americas are not covered. We hope that the quality and coverage of these models can continue to improve, and that findings based on current models can be revisited and replicated with newer models.

Finally, we rely on automatic part of speech tagging based on Universal Dependencies for the analyses here (see Appendix A for further information and Appendix B for per-language performance). Overall, the accuracy of the Stanza tagger is high for the Universal Dependencies corpora of the languages studied here (96% on average); however, it is not uniformly accurate across languages. Vietnamese has the lowest average accuracy, with 81.5% on their test set; however, our data is different in domain from many of the universal dependencies corpora, so the accuracy might be somewhat lower or higher. Universal Dependencies part of speech tags are not entirely without controversy as well—for instance, some linguists would argue that Korean does not have an adjective class, but UD uses one. It is possible that choices or inconsitencies in the assignment of POS tags according to UD could impact some MI estimates. In summary, noise due to POS tagging may have some influence on the results here, but is unlikely to affect our main conclusions.

Acknowledgments

The authors would like to thank Adam Lopez, Nina Gregorio, Burin Naowarat, Yen Meng, Amr Keleg, Oli Liu, Anna Kapron-King, and Kate McCurdy for their feedback on this work. This work was performed using resources provided by Edinburgh International Data Facility (EIDF), Edinburgh Compute and Data Facility (ECDF), and the Data-Driven Innovation Programme at the University of Edinburgh. This work was in part supported by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics and School of Philosophy, Psychology & Language Sciences.

[8] https://osf.io/bdhna/

References

Farrell Ackerman and Robert Malouf. 2013. Morphological organization: The low conditional entropy conjecture. *Language*, 89(3):429–464.

Malihe Alikhani and Matthew Stone. 2019. "Caption" as a coherence relation: Evidence and implications. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, pages 58–67, Minneapolis, Minnesota. Association for Computational Linguistics.

A.E. Backhouse. 1984. Have all the adjectives gone? *Lingua*, 62(3):169–186.

Barend Beekhuizen, Julia Watson, and Suzanne Stevenson. 2017. Semantic typology and parallel corpora: Something about indefinite pronouns. In *39th Annual Conference of the Cognitive Science Society (CogSci)*, pages 112–117.

Yoav Benjamini and Daniel Yekutieli. 2001. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188.

Uri Berger and Edoardo M. Ponti. 2024. Cross-lingual and cross-cultural variation in image descriptions. *Preprint*, arXiv:2409.16646.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. 2024. PaliGemma: A versatile 3B VLM for transfer. *Preprint*, arXiv:2407.07726.

Helen Bird, David Howard, and Sue Franklin. 2003. Verbs and nouns: The importance of being imageable. *Journal of Neurolinguistics*, 16(2):113–149.

Walter Bisang. 2017. Grammaticalization. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.

Geert Booij. 2007. Inflection. In Geert Booij, editor, *The Grammar of Words: An Introduction to Linguistic Morphology*, pages 99–124. Oxford University Press.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911.

Xi Chen, Xiao Wang, Soravit Changpinyo, A. J. Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V. Thapliyal, James Bradbury, and Weicheng Kuo. 2023. PaLI: A jointly-scaled multilingual language-image model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Christine Chiarello, Connie Shears, and Kevin Lund. 1999. Imageability and distributional typicality measures of nouns and verbs in contemporary English. *Behavior Research Methods, Instruments, & Computers*, 31(4):603–637.

Norbert Corver and Henk Van Riemsdijk. 2001. Semi-lexical categories. In Norbert Corver and Henk Van Riemsdijk, editors, *Semi-Lexical Categories*, pages 1–20. de Gruyter.

Ryan Cotterell and Jason Eisner. 2017. Probabilistic typology: Deep generative models of vowel inventories. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1182–1192, Vancouver, Canada. Association for Computational Linguistics.

Ryan Cotterell, Sabrina J. Mielke, Jason Eisner, and Brian Roark. 2018. Are all languages equally hard to language-model? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 536–541, New Orleans, Louisiana. Association for Computational Linguistics.

William Croft. 2002. *Typology and Universals*, 2nd edition. Cambridge University Press.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Catherine Dubé, Laura Monetta, María Macarena Martínez-Cuitiño, and Maximiliano A. Wilson. 2014. Independent effects of imageability and grammatical class in synonym judgement in aphasia. *Psicothema*, 26(4):449–456.

Francis Ferraro, Nasrin Mostafazadeh, Ting-Hao Huang, Lucy Vanderwende, Jacob Devlin, Michel Galley, and Margaret Mitchell. 2015. A survey of current datasets for vision and language research. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 207–213, Lisbon, Portugal. Association for Computational Linguistics.

Simeon Floyd. 2011. Re-discovering the Quechua adjective. *Linguistic Typology*, 15(1):25–63.

Spandana Gella, Rico Sennrich, Frank Keller, and Mirella Lapata. 2017. Image pivoting for learning multilingual multimodal representations. In *Proceedings of the 2017 Conference on Empirical Methods* *in Natural Language Processing*, pages 2839–2845, Copenhagen, Denmark. Association for Computational Linguistics.

Team Gemma. 2024. Gemma: Open models based on Gemini research and technology. *Preprint*, arXiv:2403.08295.

Talmy Givon. 1984. *Syntax: A Functional-Typological Introduction Vol I*. Amsterdam: Benjamins.

Joseph Harold Greenberg, editor. 1966. *Universals of Language*, 2nd edition. Number 37 in The M.I.T. Press Paperback Series. M.I.T Pr, Cambridge, Mass.

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*.

Coleman Haley, Edoardo M. Ponti, and Sharon Goldwater. 2024. Corpus-based measures discriminate inflection and derivation cross-linguistically. *Journal of Language Modelling*, 12(2):477–529.

Martin Haspelmath. 2007. Pre-established categories don't exist: Consequences for language description and typology. *Linguistic Typology*, 11(1):119–132.

Martin Haspelmath. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language*, 86(3):663–687.

Martin Haspelmath. 2012. How to compare major word-classes across the world's languages. *UCLA Working Papers in Linguistics*, 17:109–130.

Henrison Hsieh. 2019. Distinguishing nouns and verbs: A Tagalog case study. *Natural Language & Linguistic Theory*, 37(2):523–569.

Daniel Kaufman. 2009. Austronesian Nominalism and its consequences: A Tagalog case study. *Theoretical Linguistics*, 35(1):1–49.

Karin Kipper Schuler, Anna Korhonen, and Susan Brown. 2009. VerbNet overview, extensions, mappings and applications. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Tutorial Abstracts*, pages 13–14, Boulder, Colorado. Association for Computational Linguistics.

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

Johan Liljencrants, Björn Lindblom, and Bjorn Lindblom. 1972. Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language*, 48(4):839–862.

Kimberly R Lin, Lisa Wisman Weil, Audrey Thurm, Catherine Lord, and Rhiannon J Luyster. 2022. Word imageability is associated with expressive vocabulary in children with autism spectrum disorder. *Autism & Developmental Language Impairments*, 7.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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(3):1271–1291.

Joan Maling and So-Won Kim. 1998. Case assignment in the *sipta*-construction. In Ross King, editor, *Description and Explanation in Korean Linguistics*. East Asia Program, Cornell University, Ithaca, NY.

Alireza Mohammadshahi, Rémi Lebret, and Karl Aberer. 2019. Aligning multilingual word embeddings for cross-modal retrieval task. In *Proceedings of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*, pages 11–17, Hong Kong, China. Association for Computational Linguistics.

Byung-Doh Oh and William Schuler. 2024. Leading whitespaces of language models' subword vocabulary poses a confound for calculating word probabilities. *Preprint*, arXiv:2406.10851.

Hyunji Hayley Park, Katherine J. Zhang, Coleman Haley, Kenneth Steimel, Han Liu, and Lane Schwartz. 2021. Morphology matters: A multilingual language modeling analysis. *Transactions of the Association for Computational Linguistics*, 9:261–276.

Andrew K. Pawley. 2006. Where have all the verbs gone? Remarks on the organisation of languages with small, closed verb classes. In *11th Biennial Rice University Linguistics Symposium*.

Tiago Pimentel and Clara Meister. 2024. How to compute the probability of a word. *Preprint*, arXiv:2406.14561.

Tiago Pimentel, Clara Meister, Ethan Wilcox, Kyle Mahowald, and Ryan Cotterell. 2023. Revisiting the optimality of word lengths. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2240–2255, Singapore. Association for Computational Linguistics.

Frans Plank. 1994. Inflection and derivation. In *The Encyclopedia of Language and Linguistics*, pages 1671–1679. Elsevier Science and Technology, Amsterdam.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Janarthanan Rajendran, Mitesh M. Khapra, Sarath Chandar, and Balaraman Ravindran. 2016. Bridge correlational neural networks for multilingual multimodal representation learning. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 171–181, San Diego, California. Association for Computational Linguistics.

Alexander Rauhut. 2023. *Quantitative Aspects of the Word Class Continuum in English*. Ph.D. thesis, Freie Universität Berlin.

Norvin Richards. 2009. Nouns, verbs, and hidden structure in Tagalog. *Theoretical Linguistics*, 35(1):139–152.

Eva Schultze-Berndt. 2000. *Simple and Complex Verbs in Jaminjung: A Study of Event Categorisation in an Australian Language*. Ph.D. thesis, Radboud University, Nijmegen.

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(3):1258–1270.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Adrian Staub. Forthcoming. Predictability in language comprehension: Prospects and problems for surprisal. *Annual Review of Linguistics*.

Ashish V. Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A Massively Multilingual Multimodal Evaluation Dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 715–729, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Henri Theil. 1970. On the estimation of relationships involving qualitative variables. *American Journal of Sociology*, 76(1):103–154.

David John Weber. 1983. *A Grammar of Huallaga (Huanuco) Quechua.* Ph.D. thesis, University of California, Los Angeles, United States – California.

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.

Tianxing Wu, Chaoyu Gao, Lin Li, and Yuxiang Wang. 2022. Leveraging multi-modal information for cross-lingual entity matching across knowledge graphs. *Applied Sciences*, 12(19).

Andre Ye, Sebastin Santy, Jena D. Hwang, Amy X. Zhang, and Ranjay Krishna. 2024. Computer vision datasets and models exhibit cultural and linguistic diversity in perception. *Preprint*, arXiv:2310.14356.

X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. 2023. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11941–11952, Los Alamitos, CA, USA. IEEE Computer Society.

A Implementation details

A.1 Part of Speech annotations

Note that none of the datasets used here come annotated with word class information. We adopt the Universal Dependencies tagset, using Stanza (Qi et al., 2020, v.1.8.2) to tag words with their Universal Dependencies parts of speech. We remove single orthographic words that Stanza assigns multiple parts of speech, like English "don't" or German "zum" from our analysis, since it is unclear to which part of speech they should be assigned. Stanza does not cover Thai, Maori, Tagalog, Swahili, or Bengali for part of speech tagging, so they are excluded.

A.2 Word-level PMI Estimates

Because the tokenizer of the present model does not cross orthographic word boundaries, we are able to sum the log probabilities of their constituent subword tokens to obtain word-level rather than token-level log probability estimates. Ordinarily, some languages do not indicate word boundaries in their orthography, such as Japanese; however, the pretraining data and evaluation datasets (Crossmodal-3600 and COCO-35L) are word-tokenized, so this information is readily available. Further, because our language model uses sub-word tokenization with trailing whitespaces, we adopt the correction proposed by Oh and Schuler (2024); Pimentel and Meister (2024). Specifically, let $\mathbf{s}_{w_t}$ be the decomposition of word w_t into a sequence of subwords, and $\mathbf{s}_{\mathbf{w}_{<t}}$ be the decomposition of context $\mathbf{w}_{<t}$ into a sequence of subwords. Given $\mathcal{S}_{\text{bow}}$, the subset of the tokenizer vocabulary that contains subwords that are beginning-of-word (e.g., with a trailing whitespace):

$$p(w_t \mid \mathbf{w}_{<t}) = p(\mathbf{s}_{w_t} \mid \mathbf{s}_{\mathbf{w}_{<t}}) \cdot \frac{\sum_{s \in \mathcal{S}_{\text{bow}}} p(s \mid \mathbf{s}_{\mathbf{w}_{<t}} \odot \mathbf{s}_{w_t})}{\sum_{s \in \mathcal{S}_{\text{bow}}} p(s \mid \mathbf{s}_{\mathbf{w}_{<t}})} \tag{5}$$

where $\odot$ stands for concatenation.

A.3 Training details

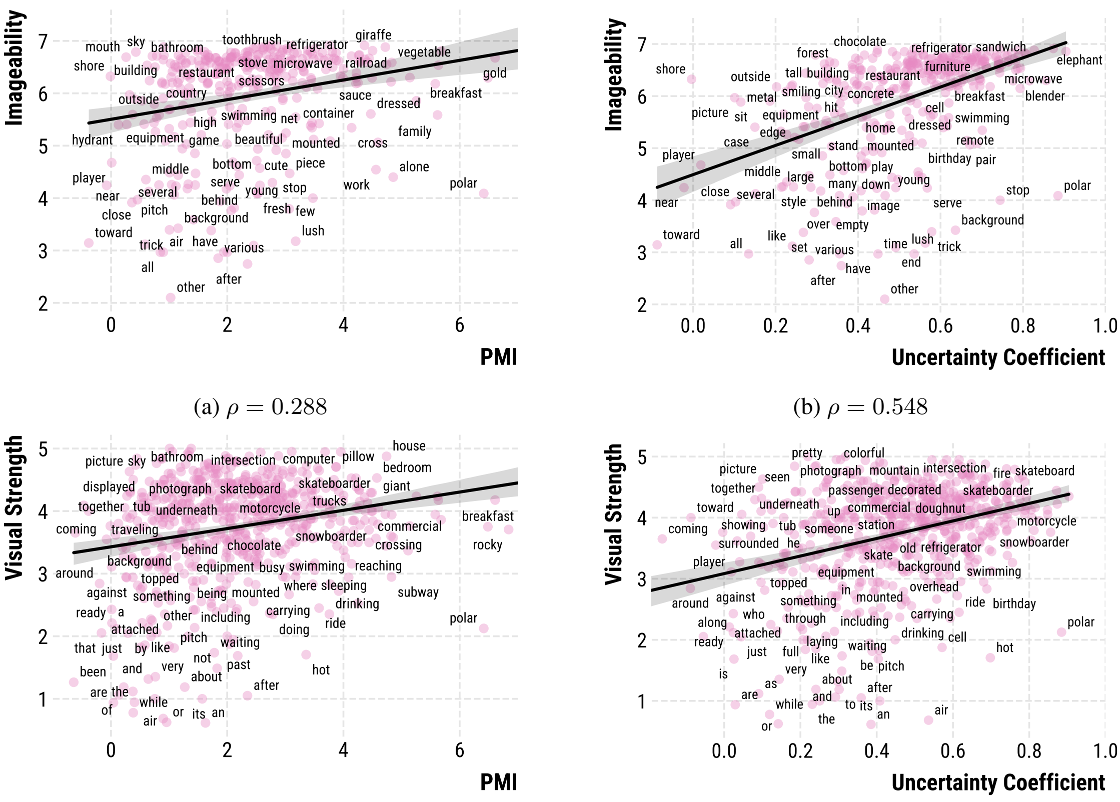
For training our language model, we did a grid search over learning rates and whether or not to use weight decay. We use a learning rate of 2×10^{-5} and weight decay of 1×10^{-6} with the Adam optimizer. To train the final model, we train on a single A100 with a batch size of 4 for 430,000 steps on COCO-35L ($\approx$ 50 hours of training, approximately 3 epochs). Our model achieves much lower perplexity on our evaluation datasets than Gemma-2B, suggesting successful domain adaptation.

B Model performance by language

See Table 2 for per-language captioning performance, POS tagging accuracy, and perplexity of the base Gemma-2B model, the PaliGemma captioning model, and our fine-tuned LM.

C Correlation plots for other psycholinguistic norms

Figure 5 shows the relationship between our measure and concreteness, as well as the uncertainty coefficient, which normalizes our measure by the language model surprisal. While concreteness is most strongly associated with our measure/its normalized variant, for completeness we show the relationships between our measure and the other psycholinguistic norms (imageability and strength of visual experience) we investigate here.



(a) $\rho = 0.288$

(b) $\rho = 0.548$

(c) $\rho = 0.212$

(d) $\rho = 0.320$

Figure 5: Correlation between English psycholinguistic norms and type-level groundedness (left) or uncertainty coefficent (right): i.e., the average ratio between LM surprisal and captioning model surprisal. Type-level measures were computed by averaging scores across the COCO-dev dataset for types which occur at least 30 times.

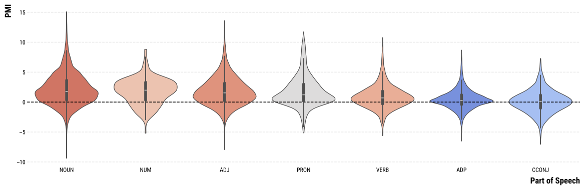
Language	ISO 639-1	CIDEr		Perplexity (COCO-35L)			Perplexity (XM3600)			Perplexity (Multi30K)			Tagging Acc.
		COCO-35L	XM3600	Gemma	PG	FT-LM	Gemma	PG	FT-LM	Gemma	PG	FT-LM	
Arabic	ar	93.73	33.20	4.86	1.48	3.09	5.12	2.87	4.63	4.51	1.94	3.46	95.18
Bengali	bn	91.23	24.07	2.85	0.88	1.61	2.65	1.56	2.16	–	–	–	–
Czech	cs	85.57	30.12	5.07	1.40	3.04	4.94	2.45	4.38	4.61	2.24	4.04	98.31
Danish	da	117.94	47.57	5.79	1.46	3.02	5.74	2.96	5.06	–	–	–	98.30
German	de	93.78	33.13	5.23	1.59	3.47	5.50	3.14	5.55	4.73	2.16	4.22	96.96
Greek	el	119.99	21.90	3.54	2.13	3.55	3.32	0.90	1.75	–	–	–	97.12
English	en	138.15	68.30	4.74	1.73	3.62	4.88	3.51	5.72	4.13	3.02	4.79	97.56
Spanish	es	138.51	48.69	4.85	1.55	3.36	5.40	3.23	5.51	–	–	–	98.01
Persian	fa	122.99	45.62	4.86	1.45	2.88	4.96	2.84	4.47	–	–	–	97.43
Finnish	fi	35.76	10.86	5.31	1.39	2.91	4.95	2.70	4.49	–	–	–	97.20
French	fr	137.79	53.35	4.96	1.44	3.15	5.13	3.12	5.08	4.36	2.73	4.50	97.55
Hebrew	he	97.94	36.59	4.36	1.34	2.71	3.84	2.30	3.74	–	–	–	90.84
Hindi	hi	104.52	26.98	3.75	1.19	2.28	3.86	2.68	3.54	–	–	–	97.95
Croatian	hr	89.42	25.95	5.24	1.37	2.88	4.68	2.49	4.33	–	–	–	98.21
Hungarian	hu	78.90	21.96	4.94	1.46	3.05	4.88	2.84	4.88	–	–	–	95.80
Indonesian	id	146.38	37.46	6.01	1.63	3.51	4.98	3.16	5.18	–	–	–	95.03
Italian	it	131.15	37.98	5.21	1.50	3.34	5.44	3.36	5.43	–	–	–	96.98
Japanese	ja	125.07	35.90	5.95	1.34	2.81	6.07	2.60	4.60	–	–	–	95.74
Korean	ko	112.40	42.82	4.89	1.29	2.61	4.80	2.37	3.95	–	–	–	95.86
Norwegian	no	118.02	39.67	6.13	1.50	3.07	5.70	2.90	4.75	–	–	–	98.38
Dutch	nl	114.76	47.19	4.96	1.54	3.24	5.34	3.15	5.55	–	–	–	96.71
Polish	pl	86.99	29.50	5.10	1.41	3.06	4.70	2.45	4.66	–	–	–	98.80
Portuguese	pt	136.40	42.76	5.52	1.53	3.30	5.56	3.38	5.49	–	–	–	97.74
Romanian	ro	118.57	22.36	5.15	1.30	2.73	4.62	2.63	4.18	–	–	–	97.98
Russian	ru	98.45	28.23	4.67	1.39	3.21	4.21	2.50	5.12	–	–	–	97.34
Swedish	sv	120.08	45.93	5.77	1.51	3.11	6.03	2.99	5.37	–	–	–	97.81
Swahili	sw	111.15	29.45	5.59	1.28	2.57	5.17	2.96	4.10	–	–	–	–
Maori	mi	156.26	40.81	5.59	1.07	2.14	5.78	3.12	3.96	–	–	–	–
Telugu	te	76.35	25.80	2.93	0.79	1.48	2.98	1.60	2.32	–	–	–	93.97
Thai	th	146.17	67.49	4.80	1.08	2.00	4.60	1.70	2.90	–	–	–	–
Turkish	tr	86.26	27.58	6.05	1.62	3.42	5.61	3.00	5.00	–	–	–	95.26
Ukrainian	uk	92.90	22.47	4.26	1.23	2.67	4.01	2.48	4.38	–	–	–	97.52
Vietnamese	vi	159.82	51.57	4.83	1.48	3.02	4.66	3.02	4.86	–	–	–	81.48
Chinese	zh	103.19	26.41	6.01	1.55	3.21	5.86	3.06	4.97	–	–	–	88.82

Table 2: Per-language performance metrics for the models used. A) CIDEr scores on Crossmodal-3600 (XM3600) and COCO-35L for the `paligemma-3b-ft-coco35-224` model. B) Perplexity scores for the base Gemma-2B model (Gemma), PaliGemma (PG) and our finetuned PaliGemma-based LM. As expected, PaliGemma has the lowest perplexity, and our fine-tuned model particularly improves perplexity on COCO-35L and for languages with different orthographies. C. Average POS tagging accuracy for the Stanza models on the Universal Dependencies treebank test sets for each language.

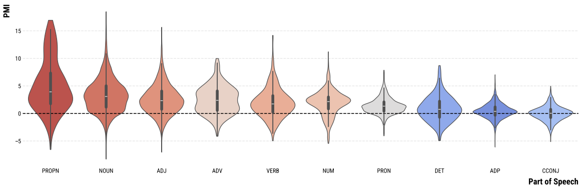
D Groundedness distribution for Crossmodal-3600

Results are ordered by descending mutual information estimate within the dataset (average groundedness/PMI). Hue indicates the average cross-linguistic ranking of a part of speech.

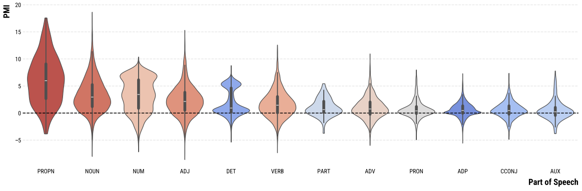
Arabic (ar)



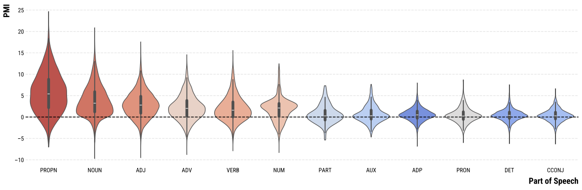
Czech (cs)



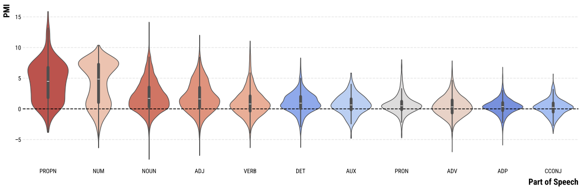
Danish (da)



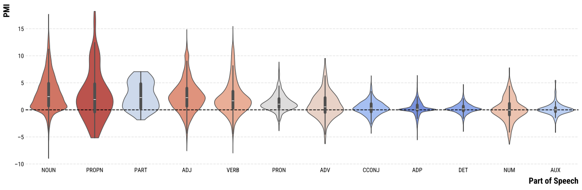
German (de)



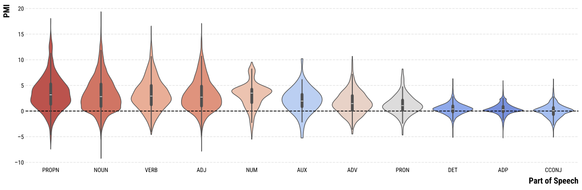
Modern Greek (el)



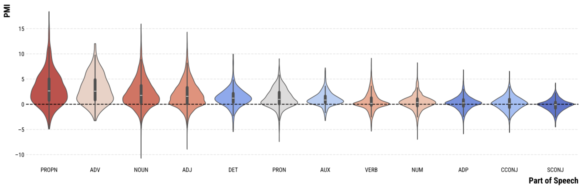
English (en)



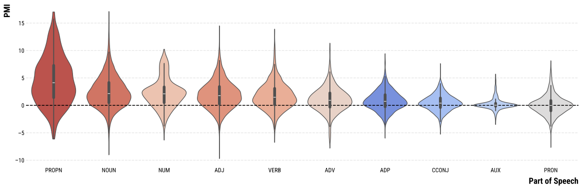
Spanish (es)



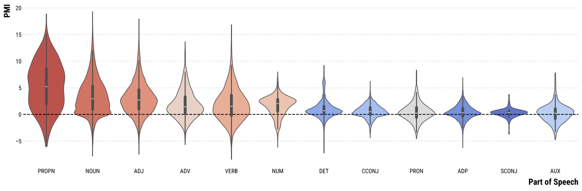
Persian (fa)



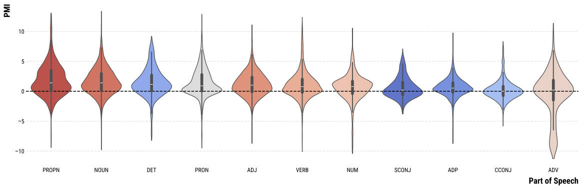
Finnish (fi)



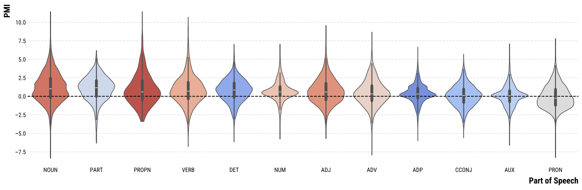
French (fr)



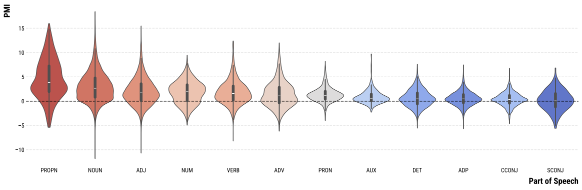
Hebrew (he)



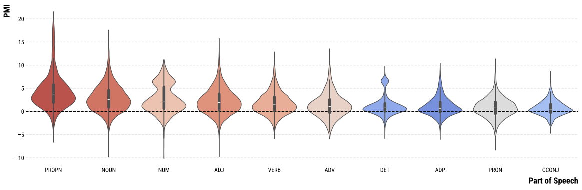
Hindi (hi)



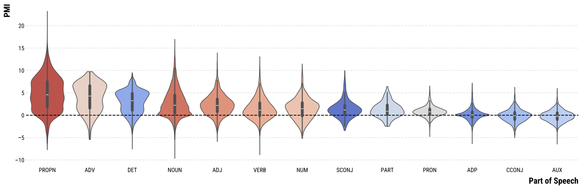
Croatian (hr)

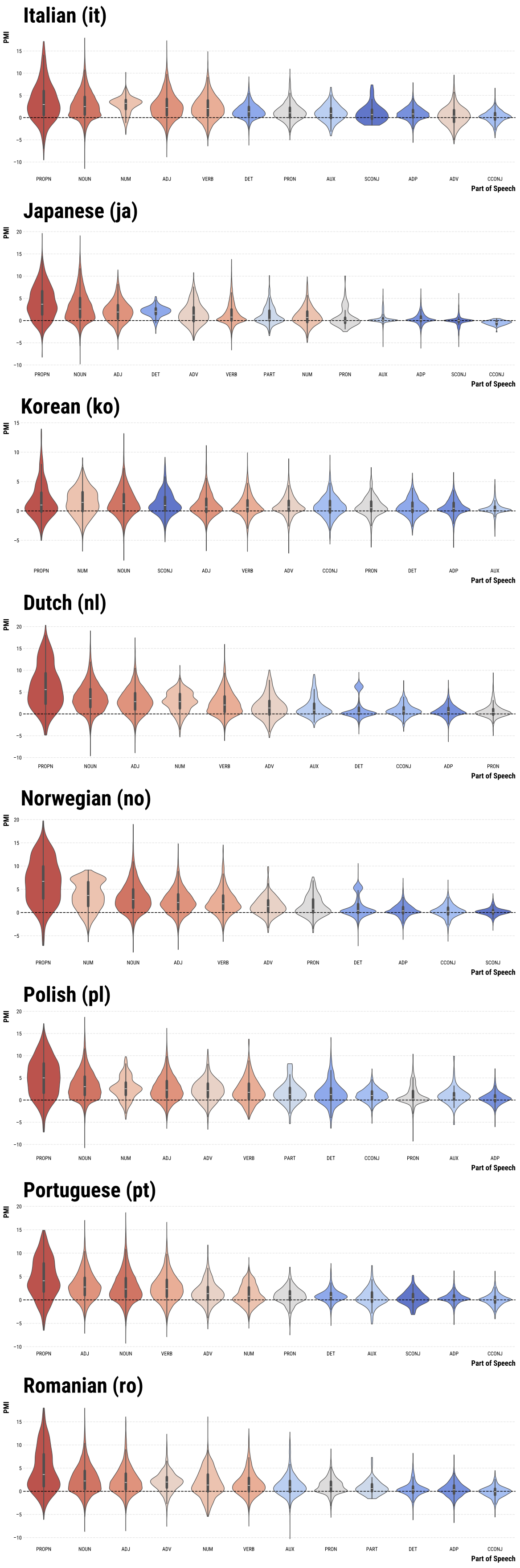


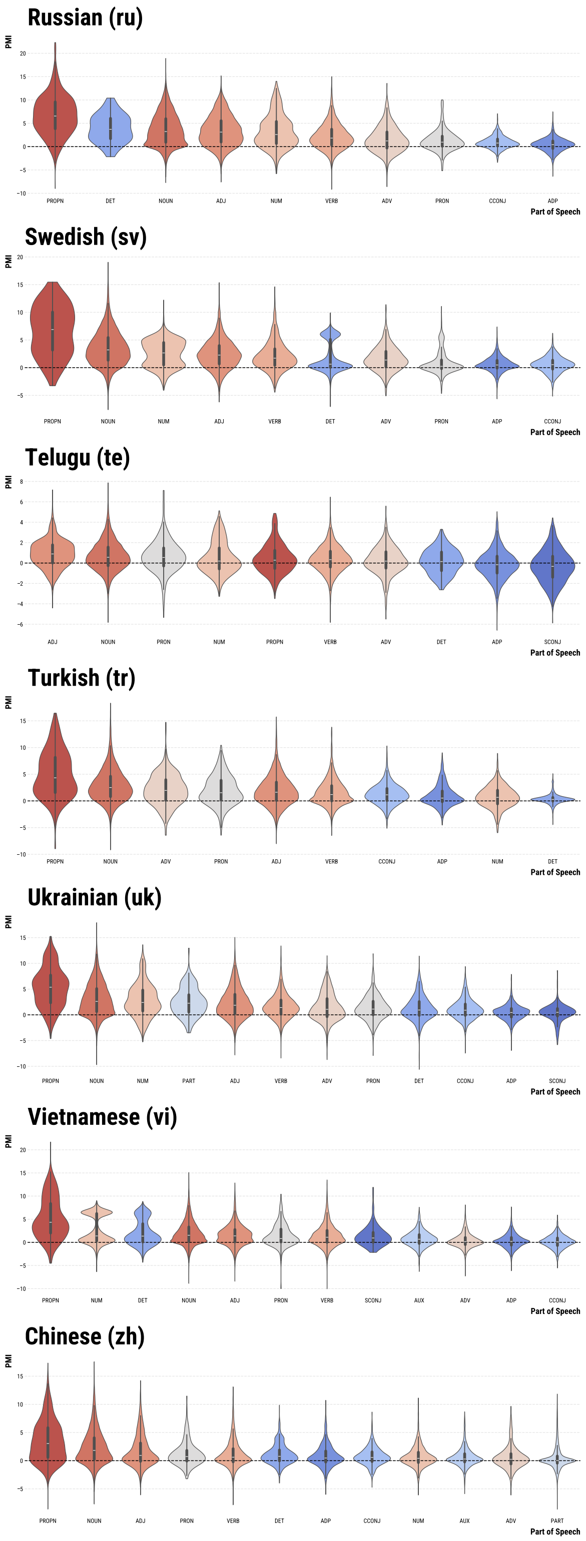
Hungarian (hu)



Indonesian (id)

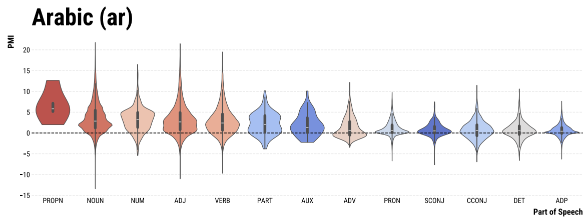


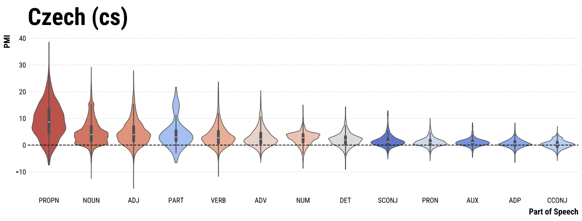


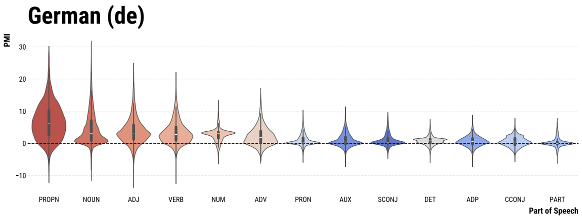


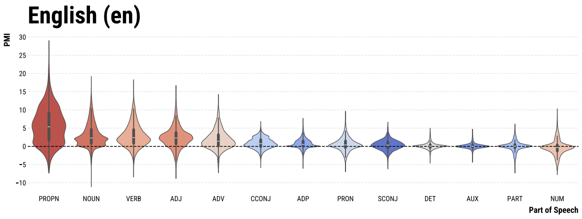
E Groundedness distribution for Multi30K

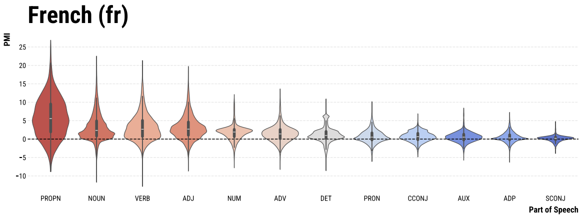
Results are ordered by descending mutual information estimate within the dataset (average groundedness/PMI). Hue indicates the average cross-linguistic ranking of a part of speech.





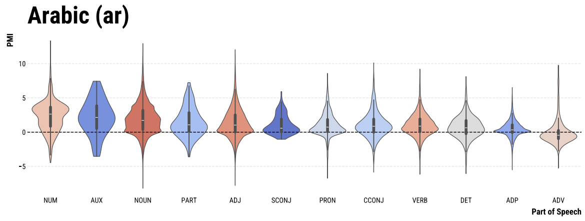


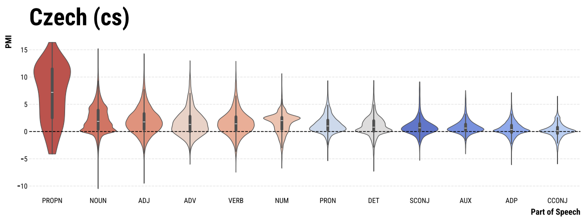


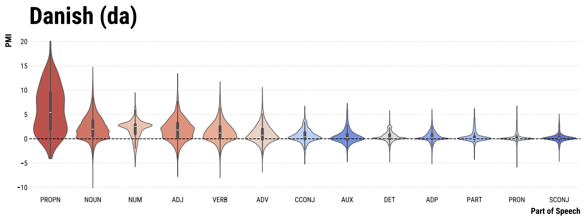


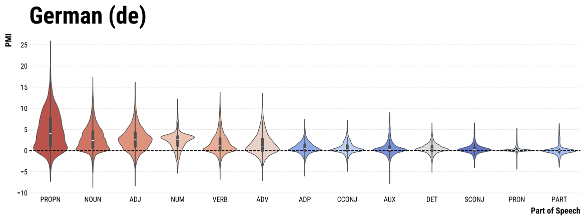
F Groundness distribution for COCO-35L Development Set

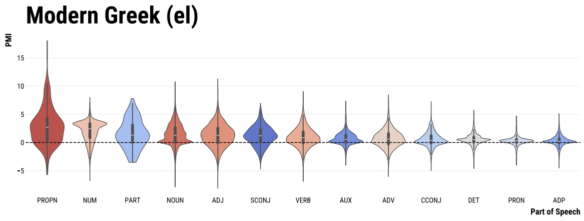
Results are ordered by descending mutual information estimate within the dataset (average groundedness/PMI). Hue indicates the average cross-linguistic ranking of a part of speech.

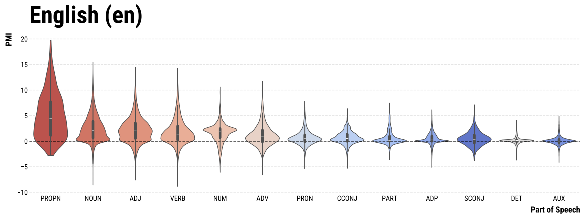


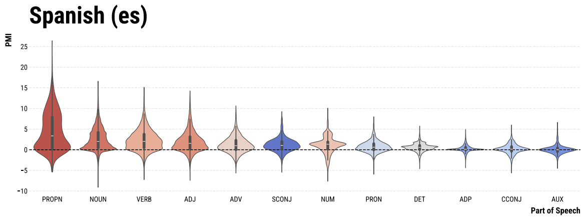


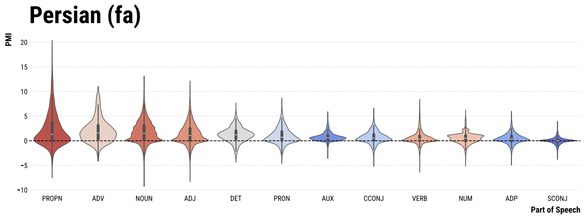


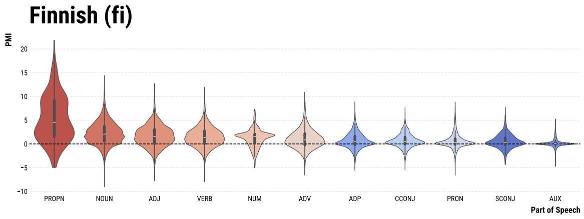


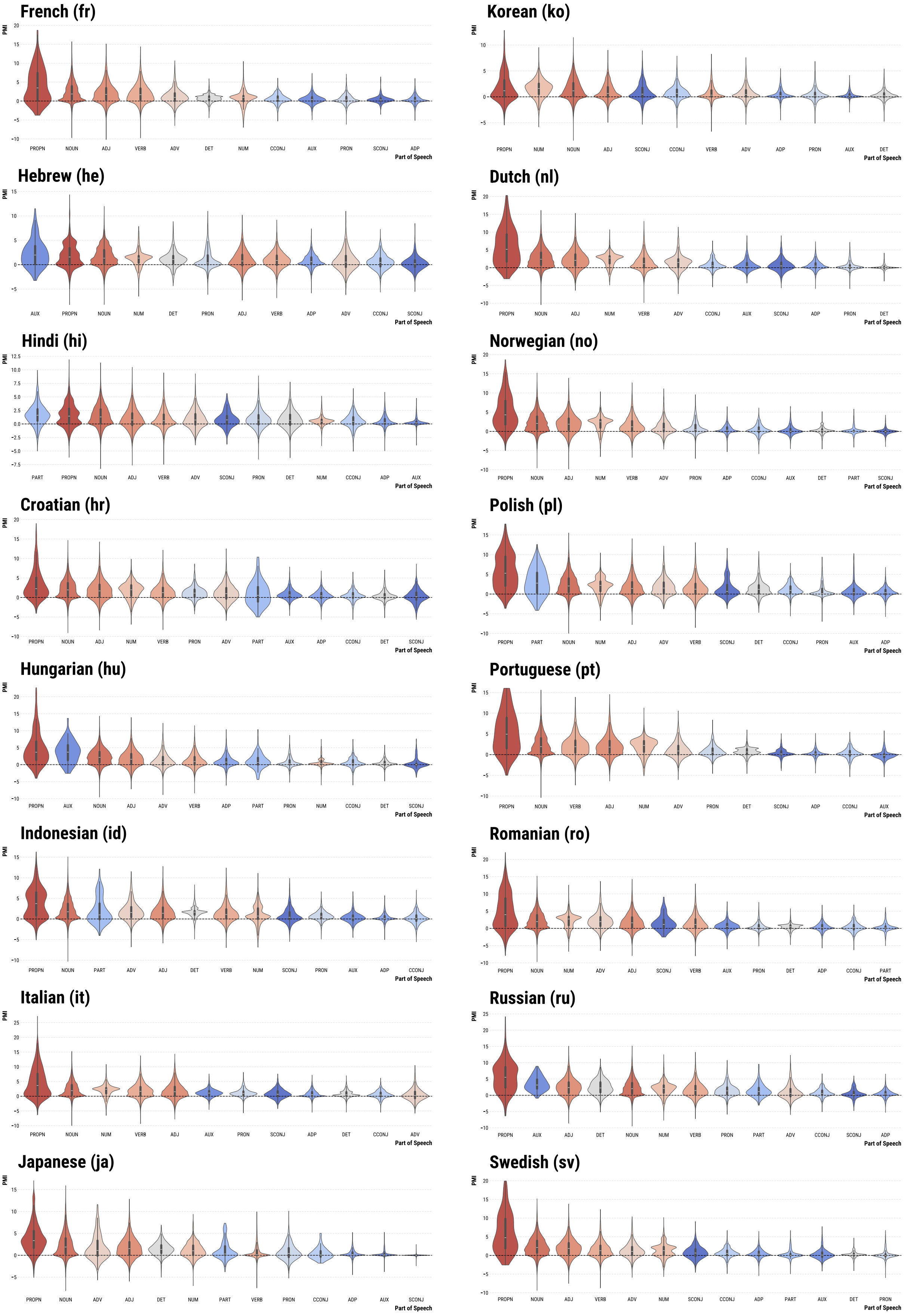


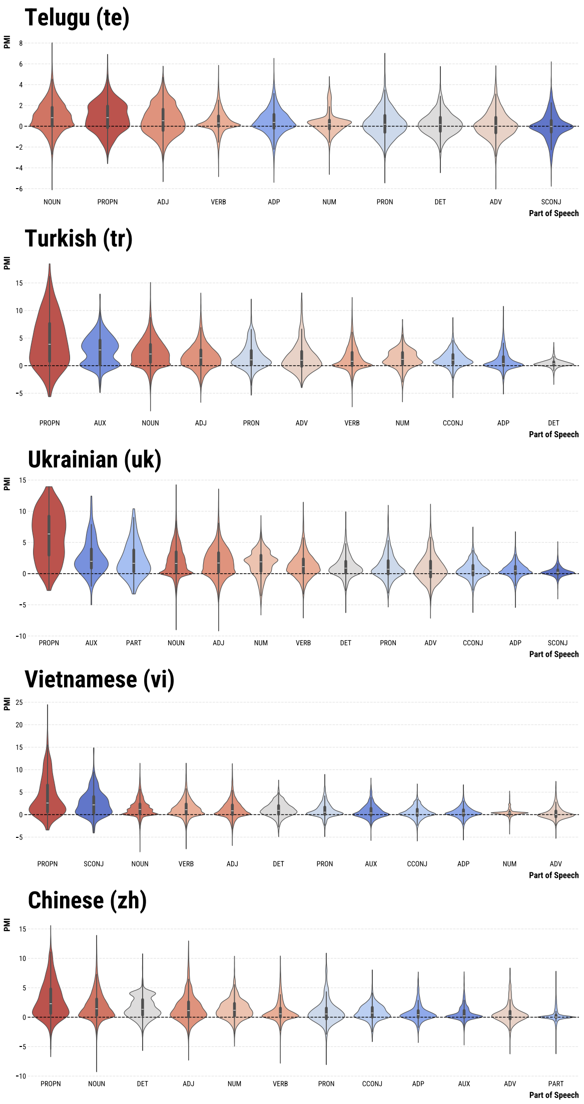









French (fr)
PMI
20
15
10
5
0
-5
-10
PROPN
NOUN
ADJ
VERB
ADV
DET
NUM
CCONJ
AUX
PRON
SCONJ
ADP
Part of Speech
Korean (ko)
PMI
10
5
0
-5
PROPN
NUM
NOUN
ADJ
SCONJ
CCONJ
VERB
ADV
ADP
PRON
AUX
DET
Part of Speech
Hebrew (he)
PMI
15
10
5
0
-5
AUX
PROPN
NOUN
NUM
DET
PRON
ADJ
VERB
ADP
ADV
CCONJ
SCONJ
Part of Speech
Dutch (nl)
PMI
20
15
10
5
0
-5
-10
PROPN
NOUN
ADJ
NUM
VERB
ADV
CCONJ
AUX
SCONJ
ADP
PRON
DET
Part of Speech
Hindi (hi)
PMI
12.5
10.0
7.5
5.0
2.5
0.0
-2.5
-5.0
-7.5
PART
PROPN
NOUN
ADJ
VERB
ADV
SCONJ
PRON
DET
NUM
CCONJ
ADP
AUX
Part of Speech
Norwegian (no)
PMI
20
15
10
5
0
-5
-10
PROPN
NOUN
ADJ
NUM
VERB
ADV
PRON
ADP
CCONJ
AUX
DET
PART
SCONJ
Part of Speech
Croatian (hr)
PMI
20
15
10
5
0
-5
-10
PROPN
NOUN
ADJ
NUM
VERB
PRON
ADV
PART
AUX
ADP
CCONJ
DET
SCONJ
Part of Speech
Polish (pl)
PMI
15
10
5
0
-5
-10
PROPN
PART
NOUN
NUM
ADJ
ADV
VERB
SCONJ
DET
CCONJ
PRON
AUX
ADP
Part of Speech
Hungarian (hu)
PMI
20
15
10
5
0
-5
-10
PROPN
AUX
NOUN
ADJ
ADV
VERB
ADP
PART
PRON
NUM
CCONJ
DET
SCONJ
Part of Speech
Portuguese (pt)
PMI
15
10
5
0
-5
-10
PROPN
NOUN
VERB
ADJ
NUM
ADV
PRON
DET
SCONJ
ADP
CCONJ
AUX
Part of Speech
Indonesian (id)
PMI
15
10
5
0
-5
-10
PROPN
NOUN
PART
ADV
ADJ
DET
VERB
NUM
SCONJ
PRON
AUX
ADP
CCONJ
Part of Speech
Romanian (ro)
PMI
20
15
10
5
0
-5
-10
PROPN
NOUN
NUM
ADV
ADJ
SCONJ
VERB
AUX
PRON
DET
ADP
CCONJ
PART
Part of Speech
Italian (it)
PMI
25
20
15
10
5
0
-5
-10
PROPN
NOUN
NUM
VERB
ADJ
AUX
PRON
SCONJ
ADP
DET
CCONJ
ADV
Part of Speech
Russian (ru)
PMI
25
20
15
10
5
0
-5
-10
PROPN
AUX
ADJ
DET
NOUN
NUM
VERB
PRON
PART
ADV
CCONJ
SCONJ
ADP
Part of Speech
Japanese (ja)
PMI
15
10
5
0
-5
PROPN
NOUN
ADV
ADJ
DET
NUM
PART
VERB
PRON
CCONJ
ADP
AUX
SCONJ
Part of Speech
Swedish (sv)
PMI
20
15
10
5
0
-5
-10
PROPN
NOUN
ADJ
VERB
ADV
NUM
SCONJ
CCONJ
ADP
PART
AUX
DET
PRON
Part of Speech


Telugu (te)
PMI
8
6
4
2
0
−2
−4
−6
NOUN PROPN ADJ VERB ADP NUM PRON DET ADV SCONJ
Part of Speech
Turkish (tr)
PMI
15
10
5
0
−5
PROPN AUX NOUN ADJ PRON ADV VERB NUM CCONJ ADP DET
Part of Speech
Ukrainian (uk)
PMI
15
10
5
0
−5
−10
PROPN AUX PART NOUN ADJ NUM VERB DET PRON ADV CCONJ ADP SCONJ
Part of Speech
Vietnamese (vi)
PMI
25
20
15
10
5
0
−5
PROPN SCONJ NOUN VERB ADJ DET PRON AUX CCONJ ADP NUM ADV
Part of Speech
Chinese (zh)
PMI
15
10
5
0
−5
−10
PROPN NOUN DET ADJ NUM VERB PRON CCONJ ADP AUX ADV PART
Part of Speech