A Grounded Typology of Word Classes

Anonymous ACL submission

Abstract

In this work, we propose a grounded approach to meaning in language typology. Using images captioned across languages, we can treat the images as an empirical language agnostic representation of meaning, allowing the quantification of language function and semantics. Using principles from information theory, we define "groundedness", an empirical measure of contextual semantic contentfulness which can be computed using multilingual (vision-and-)language models. As an initial application, we apply this measure to the typology of word classes. We find 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. 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 which occur across the world's languages (Croft, 2002, pp. 1–2). In order 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 language, 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).

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).

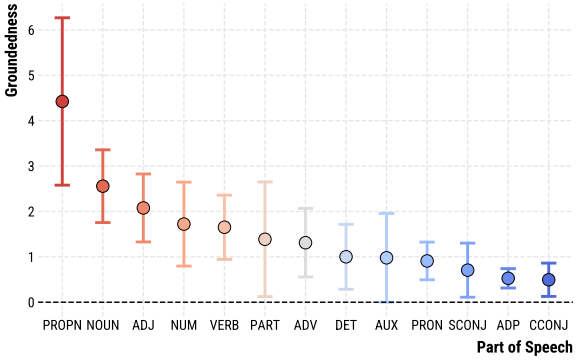


Figure 1: Mean and standard deviation of per-language mutual information estimates between word class and image. Across 30 languages, we see clear and consistent tendencies about which parts of speech are more "grounded", corresponding to a distinction between lexical and functional classes.

In this work, 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 cross-linguistically at sentences produced as captions of the same image, we can use the image as an objective, language-agnostic representation of the shared semantics underlying these utterances, analogous to the objective acoustic signal in the study of vowel spaces.

In this work, we specifically focus on semantic contentfulness–how semantically informative a given word token is. We introduce a way to empiri-

cally quantify contentfulness, *groundedness*, which relies only on self-supervised vision-and-language models. Groundedness quantifies 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 traditional language model is an estimate of the pointwise mutual information: the greater this difference (LM surprisal > captioning surprisal), 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 (better known within the field of natural language processing as "parts of speech"). Literature from language evolution, cognitive linguistics, pyscho- and neurolinguistics convergently 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; Strik Lievers et al., 2021; Chiarello et al., 1999). Nevertheless, there has been no cross-linguistic quantitative 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 norms like imageability and concreteness in English, it provides an intuitive ranking between nouns, verbs, and adjectives (noun > adjectives > verbs) across languages but contradicts the view of adpositions as a "semi-lexical" class. However, our results suggest grammatical word classes still carry semantic content. These results validate intuitions about word class contentfulness and 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 measures for further study.[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 analogous/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). In this work, we investigate word classes as operationalised 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 generalised 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

[1] https://osf.io/bdhna/?view_only=cf5322aae1d04d1287821d1d9ab0c372

of particular members of a given word class can be quite variable. For example, a noun like "factor" 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 interrogated related concepts.

While theoretical linguistics has focused on a distinction between content words and function words, psycholinguistics has focused on semantic dimensions like imageability, concreteness, and strength of perceptual experience. These have also been found to be highly relevant to 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). Given that we operationalise meaning as an image, notions such as imageability seem even more clearly related to our groundedness measure. However, as discussed in Section 5.4, these concepts are different from our measure in that informativity is not a major factor in their definition.

Another related concept studied in computational psycholinguistics is surprisal. Similar to 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 (Staub, Forthcoming). However, surprisal entangles formal and functional information in language. As such, valid cross-linguistic comparisons based on surprisal can be challenging, since form is language specific (Park et al., 2021; Mielke et al., 2019). We here aim to focus on information due to language *function*, separated from form.

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 image.

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

While the true probabilities in Equation 1 are not available, using an image as a meaning representation makes both quantites straightforwardly estimable with existing self-supervised neural techniques: $p_{\boldsymbol{\phi}}(w_t \mid m, \mathbf{w}_{<t})$ corresponds to the probability of the token under an image captioning model, while $p_{\boldsymbol{\theta}}(w_t \mid \mathbf{w}_{<t})$ corresponds to its probability under a vanilla language model.[2] This also allows us to understand groundedness as a *difference in surprisal*, with the value corresponding to how much more expected the token is in the grounded model than in the textual model. As such, in principle the PMI should rarely take on negative values–because the captioning model has strictly more information than the language model. However, some tokens, such as those that are highly grammatical or structural, should be close to 0 (independence).

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:

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

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:

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

[2] These are trained by minimising their cross-entropy with respect to the empirical data distribution, so they provide an upper bound to the entropy of the true distribution.

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 $\mathcal{D}$:

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

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 out mutual information estimates are conditional on *context*. As such, words which are very grounded in one context could have a very low groundedness 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, weakening this effect.

4 Experimental setup

Captioning model $p_{\boldsymbol{\phi}}(w_t \mid \mathbf{w}_{<t}, m)$ 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 this model is a general-purpose multimodal vision and language model, capable of handling a wide range of tasks, it is designed to be fine-tuned on and used for a single task. As such, we use the released `paligemma-3b-ft-coco35-224` checkpoint for multilingual captioning, which has been fine-tuned on COCO-35L.

Language model $p_{\boldsymbol{\theta}}(w_t \mid \mathbf{w}_{<t})$ For our language model, our aim is to use a model as similar to our captioning model $p_{\boldsymbol{\phi}}(w_t \mid \mathbf{w}_{<t}, m)$ as possible. This is critical to getting good (P)MI estimates, which relies on estimating a difference in surprisal between the two models. For instance, 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 have seen 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 therefore maximally comparable. Table 1 summarizes the data matching between the two models.

Evaluation Datasets We also require multilingual image captioning datasets for evaluation which are not observed during training. For this, we use three separate datasets, each with their own strengths and weaknesses. First, we use **Crossmodal-3600**. This dataset includes captions for 3,600 images across a range of cultures, manually and independently captioned by 1-2 speakers of 36 typologically diverse languages. However, it is relatively small per language compared to other datasets, as many images have only one or two captions in a given language. Further, the independence of the captions means that there is greater diversity in what aspects of an image are being described across languages.

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 [3], yielding 25,000 captions per language. Further, the captions are translations of each other, ensuring more comparable semantic content across languages. However, the captions are machine translated, which presents potential quality issues.

Finally, we consider **Multi30K**. This dataset comprises 30,000 images captioned 5 times each in English, with a single caption per image addition-

[3] Crossmodal-3600 and COCO-35L cover the same languages with the exception of Quechua, which is omitted from COCO-35L due to the lack of a translation model.

Model	Gemma PT (Gemma, 2024)	PaliGemma CT (Beyer et al., 2024)	COCO-35L FT (Thapliyal et al., 2022)
Image captioning model	**A**	🖼 **A**	🖼 **A**
Language model	**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**).

ally manually translated into each of 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 convergent evidence, we obtain a picture that is robust to the weaknesses of the individual datasets.

Part-of-Speech Annotation Note that none of these datasets are annotated with word class information. We adopt the Univeral 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.

Estimating word-level probabilities Because the tokenizer of the present model does not cross orthographic word boundaries, we are able to sum the PMI of their constituent subword tokens to obtain a word-level rather than token-level PMI estimate. Ordinarily, languages without spaces in orthography such as Japanese pose a complication for this type of approach, but the pretraining data and evaluation datasets (Crossmodal-3600 and COCO-35L) are word-tokenized, so orthographic word boundaries are generally reliable for wordhood status in these corpora. Finally, we use the correction proposed by Oh and Schuler (2024); Pimentel and Meister (2024) to correctly estimate word-level surprisals for leading-whitespace models.

5 Results

One of the major generalizations about parts of speech or word classes within linguistics is that they are, broadly, divided into contentful, semantically rich categories; and semantically poor, grammatically-driven categories. Semantically-rich words, we hypothesize, are likely to demonstrate a stronger linkage with the image, while more grammatical words will have a weaker link. Overall, we found that contentful parts of speech had higher mutual information than functional/grammatical parts of speech. Figure 1 summarizes the mutual information (MI) estimates for word classes across 30 languages and our three datasets (Multi30K, COCO-35L-Dev, and Crossmodal-3600). Figure 2 shows the overall token-level distribution of our groundedness measure across all three datasets (with results for all individual languages and datasets in Appendices C, D, and E). Both figures seem to show a soft yet clear tendency for traditionally lexical parts of speech to have a higher mutual information with the image they describe. In the following sections, we aim to quantify these trends, and explore the semantics of our measure.

5.1 Which word classes are grounded?

We first investigate whether there are any parts of speech which are *not* significantly grounded–that is, their estimated mutual information with the image is not greater than 0.

To do so, we use a 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 randomly permute their signs, averaging them 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 probability that the true MI is greater than 0.[4]

We find that most word classes have a MI significantly greater than 0 in all languages.[5] Further,

[4]We set our significance threshold at $\alpha = 0.01$ and use the Benjamini and Yekutieli (2001) corrections.

[5]Detailed results in Appendix B.

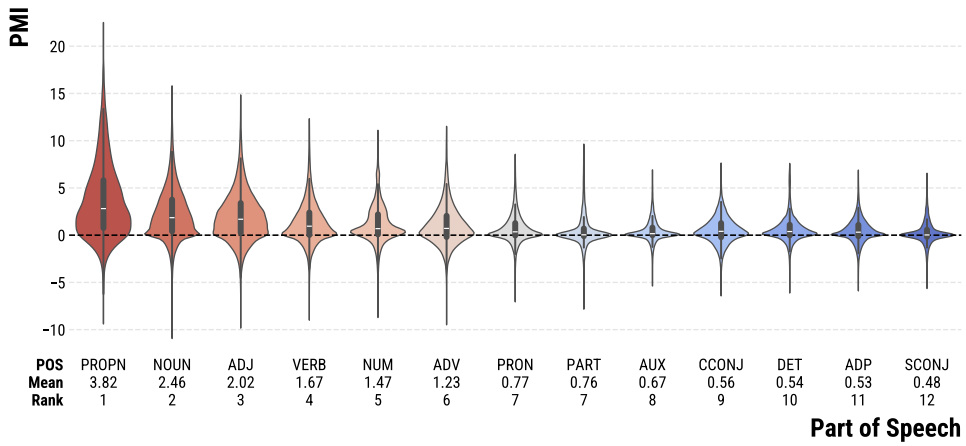


Figure 2: Word token level distributions of the groundedness measure (PMI) across all languages and datasets, grouped by part of speech (word class). We also report the estimated marginal mean and ranking of each word class. Colors are based on the ranking of classes, rather than their average PMIs. Overall, the distribution and estimated ranking of word classes strongly suggest our groundedness measure quatitatively captures the distinction between lexical and functional classes.

all word classes have a mutual information estimate significantly greater than 0 in most langauges. Word classes for which we cannot reject the null hypothesis in some languages are largely highly grammaticalized/functional: particles (5 langs.), subordinating conjunctions (4 langs.), coordinating conjunctions (4 langs.), auxiliary verbs (4 langs.). A few traditionally lexical classes also fail to reject the null hypothesis in some languages: adverbs (`he` and `te`), numerals (`en`), and proper nouns (`ar`). Overall, these 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).

5.2 What 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[6], and perform a post-hoc pairwise comparison test of the means.[7] The results of this analysis are displayed in Figure 2. 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). Pronouns occupy an intermediate position, having a an EMM which is not significantly different than particles. The ranking corroborates ideas from cognitive linguistics which place nouns, adjectives, and verbs in a continuous grammaticalization cline, with nouns > adjectives > verbs[8]. 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).

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 (analyzing the values summarized in Figure 1), and token-level PMI (summarized in Figure 2). In both cases, we use ANOVA to estimate the amount of the variance in groundedness explained by word class.

MI estimates For the language-level MI estimates in Figure 1, we consider the separate effects language, dataset, and POS have on groundedness. Because the meanings (images) are matched across languages, this allows us to estimate and control for variation due to some languages having consistently larger or smaller MI estimates (due to language-specific variation in our multilingual neu-

[6] Averaged over values of language and dataset.

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

[8] Our operationalization of grammaticalization here likely somewhat under-estimates the groundedness of verbs, as they tend to be temporally extended.

ral estimators), and how much variation due to certain datasets generally varying in the groundedness of the language they contain. 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 ($\eta^2 = 0.005$, $F_{3,816} = 5.71$, $p < 0.01$). Language identity has a larger effect, explaining 8.2% of the variance ($\eta^2 = 0.082$, $F_{29,789} = 6.42$, $p < 0.001$). However, word class dominates, explaining most of the total variance (57.3%, $\eta^2 = 0.573$, $F_{12,806} = 775$, $p < 0.001$), and 62.8% of the remaining variance (partial $\eta^2 = 0.628$) 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 mutual information 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 2). Within a POS, contentfulness is expected to vary significantly, so 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 ($\eta^2 = 0.024$, $F_{64,10^7} = 4727$, $p < 0.001$). Word class accounts for 12.0% of the total variation ($\eta^2 = 0.120$, $F_{12,10^7} = 123583$, $p < 0.001$). 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 ($\eta^2 = 0.016\%$, $F_{330,10^7} = 602.5$, $p < 0.001$), despite having many degrees of freedom. So cross-linguistic 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.[9]

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 widely used in psycholinguistics. While one potential advantage of our method here is the ease with which it allows the rating of individual word tokens in context, existing ratings tend to be at the level of types. We focus our analysis here on English and on word types which occur at least 30 times in the COCO(-35L) validation set,[10] averaging across occurences to obtain an estimate of the average type-level groundedness.

We explore comparisons with 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 many words on a Likert Scale. For imagability, we use the Glasgow Psycholinguistic Norms (Scott et al., 2019), for concreteness we use the norms from Brysbaert et al. (2014) and for strength of visual experience, we use the values from the Lancaster Sensorimotor Norms (Lynott et al., 2020). Overall, we observe fairly weak (though significant, $p < 0.001$) correlations with these norms using Spearman's ρ (Imageability: $\rho = 0.288$, Concreteness: $\rho = 0.368$, Visual strength: $\rho = 0.212$).

We find this is in part related 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 according to our measure). To account for differences in baseline (LM) word informativity, we can normalize the PMI scores by dividing by the LM surprisal, yielding the uncertainty coefficient (Theil, 1970), which measures 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: $\rho = 0.609$, Visual strength: $\rho = 0.320$). This suggests that the differences between groundedness and surprisal are assoicated with concreteness. However, this measure necessarily 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 presumably based on the more abstract geographical concept. Other words with di-

[9] The token-level interaction models and their ANOVA statistics are computationally intensive, involving the repeated fitting of hundreds of parameters to millions of data points. We use 512GB of RAM and approximately 6 hours to compute these values.

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

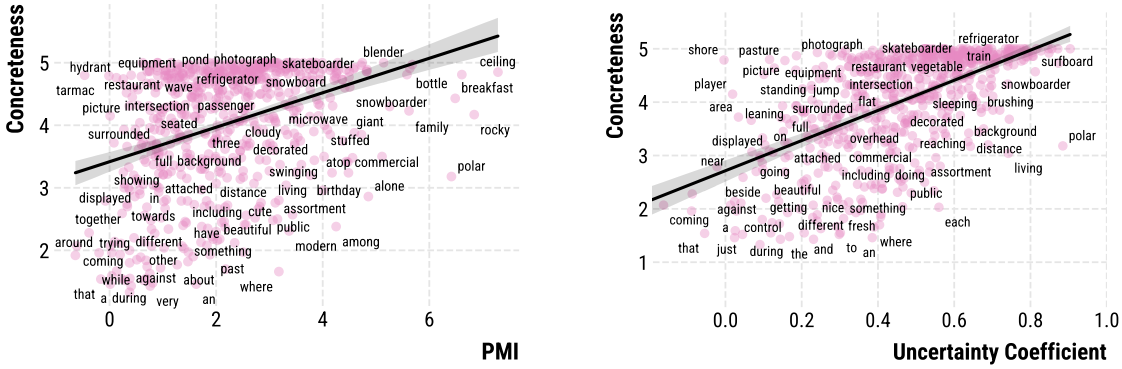


Figure 3: Correlation between human concreteness ratings and type-level groundedness (PMI; left, $\rho = 0.368$) or uncertainty coefficent (right, $\rho = 0.609$): i.e., the average ratio between LM surprisal and captioning model surprisal.

vergent 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 used more abstractly in the image captioning context (e.g. "photo" exhibits very low PMIs, because captions frequently begin with "A photo of...").

6 Discussion and Conclusion

In this study, we propose a grounded approach to typology, using images as a proxy for sentence meaning. Using information theory and neural models, we define a groundedness measure of a token's association with its meaning. Together, our results demonstrate that parts of speech vary systematically in terms of their groundedness across a typologically-diverse sample of languages. We find this variation can be described as a continuous cline generalizing the traditional distinction between lexical and functional word classes. However, our results suggest grammatical word classes still carry semantic content. We find that nouns > adjectives > verbs, in line with a continuum view of these classes, but 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 a single case study, future work can and should use the measures introduced here to investigate various typological questions about the way languages organise information. For instance, we anecdotally noticed that languages with grammatical gender on determiners tend to have higher than average groundedness for that class. Future work should explore this, investigating in detail what kinds of "functional" classes have higher groundedness and under what conditions. Additionally, our approach could be used to study non-prototypical word class organizations, such as languages which don't clearly distinguish between adjectives and verbs (Korean), or languages that split single word classes into distinct sub-classes (Japanese adjectives). Our approach can also cover any linguistic classes which can be defined over tokens, such as morphemes or semantic classes. For instance, future work could explore the claim that inflections are more "grammatical" than derivations (Booij, 2007; Haley et al., 2023). To support such work, we make the groundedness scores from the present study available online.[11]

Going beyond the details of the approach here, our work generally suggests a role for multimodal models like image captioning models in computational typology similar to the one played by language models in the past decade. While the range of languages covered and data availability is less than for pure text models, the latest multimodal models and datasets cover enough languages from a typologically and culturally diverse set of languages to make them worth studying–and we anticipate this will only improve from here. 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.

[11] https://osf.io/bdhna/?view_only=cf5322aae1d04d1287821d1d9ab0c372

7 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 which are not part 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.

References

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.

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.

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.

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.

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

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

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.

Coleman Haley, Edoardo M. Ponti, and Sharon Goldwater. 2023. Corpus-based measures discriminate inflection and derivation cross-linguistically. *Society for Computation in Linguistics*, 6(1):403–407.

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.

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.

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.

Sabrina J. Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2019. What kind of language is hard to language-model? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4975–4989, Florence, Italy. 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.

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.

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.

Adrian Staub. Forthcoming. Predictability in Language Comprehension: Prospects and Problems for Surprisal. *Annual Review of Linguistics*.

Francesca Strik Lievers, Marianna Bolognesi, and Bodo Winter. 2021. The linguistic dimensions of concrete and abstract concepts: Lexical category, morphological structure, countability, and etymology. *Cognitive Linguistics*, 32(4):641–670.

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.

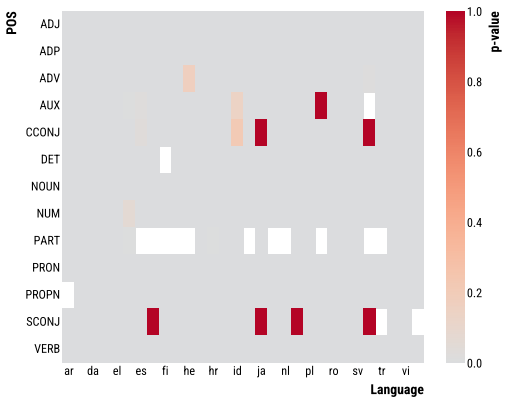
X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. 2023. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11941–11952, Los Alamitos, CA, USA. IEEE Computer Society.

A 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.

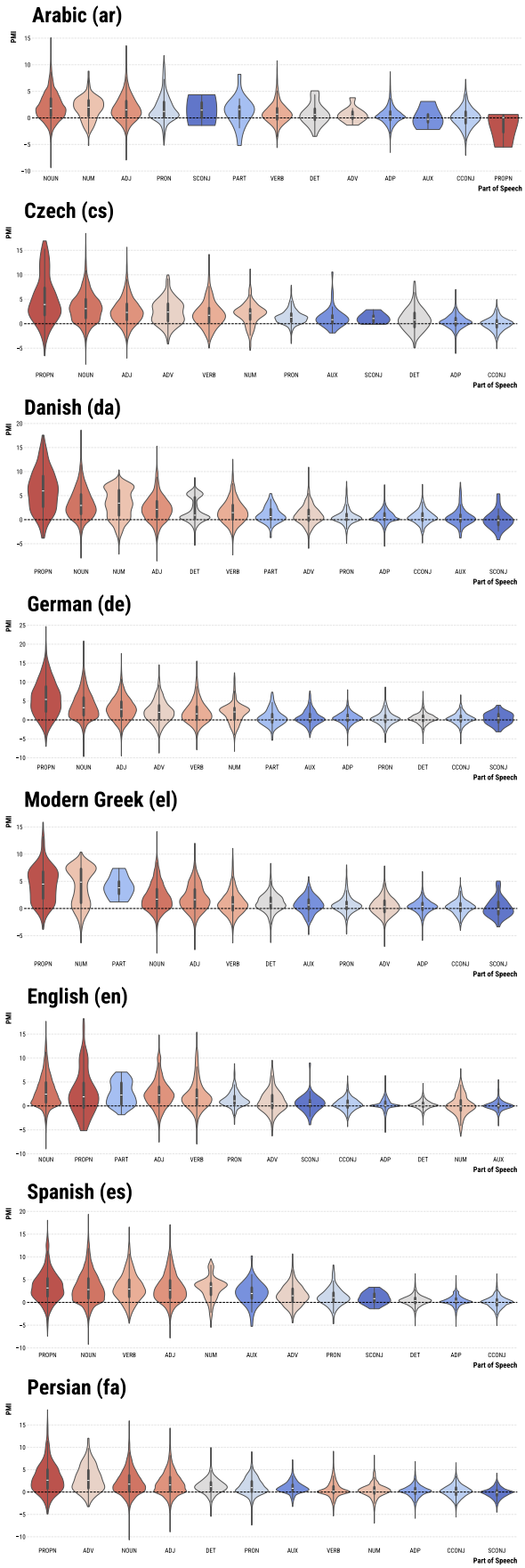
B Detailed results of permutation tests

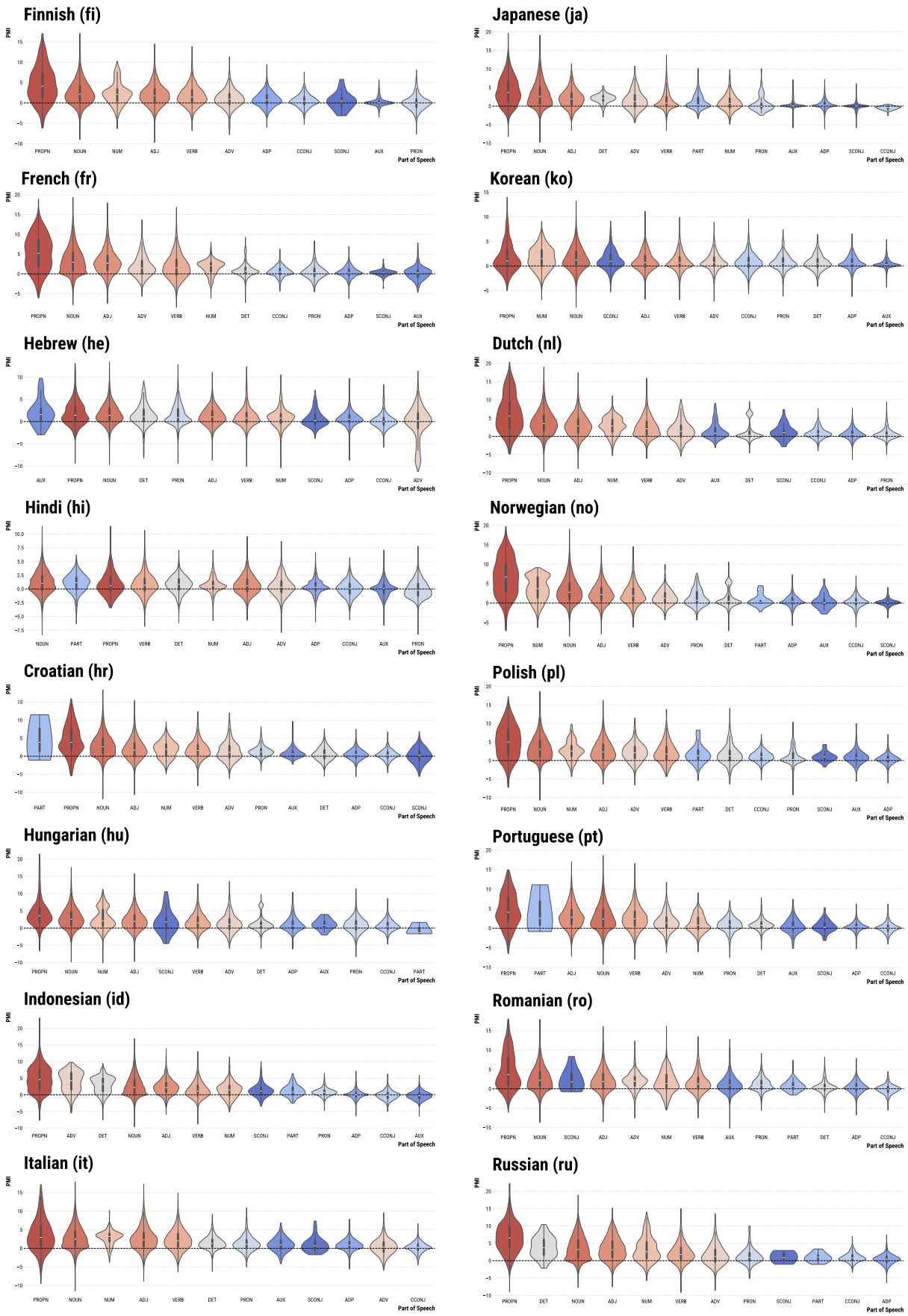
The permutation tests are described in detail in Section 5.1. White squares indicate combinations of part of speech and language which are not attested. Grey combinations are significant, while red/orange combinations are not significantly different from 0 based on our permutation test. The parts of speech which are not significantly different from 0 are predominately functional.

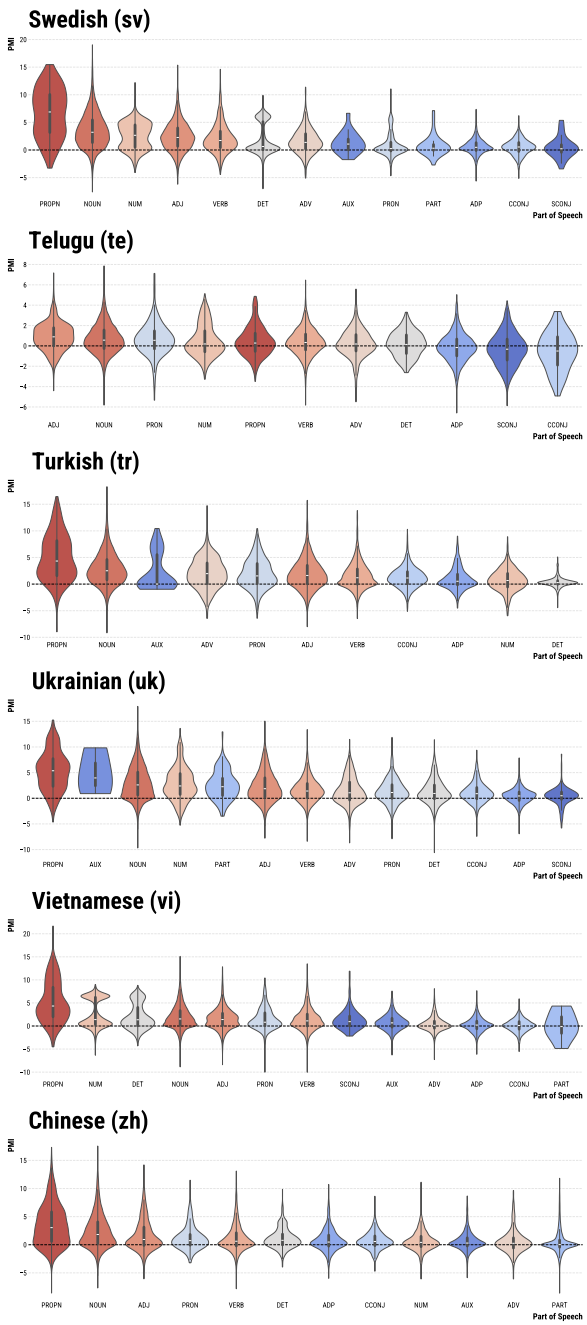


C Groundedness distribution for Crossmodal-3600

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.

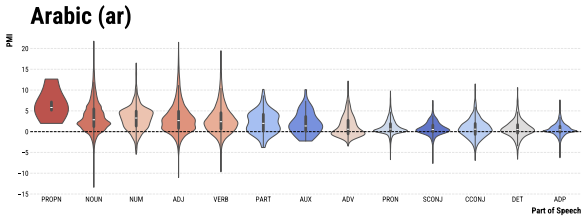


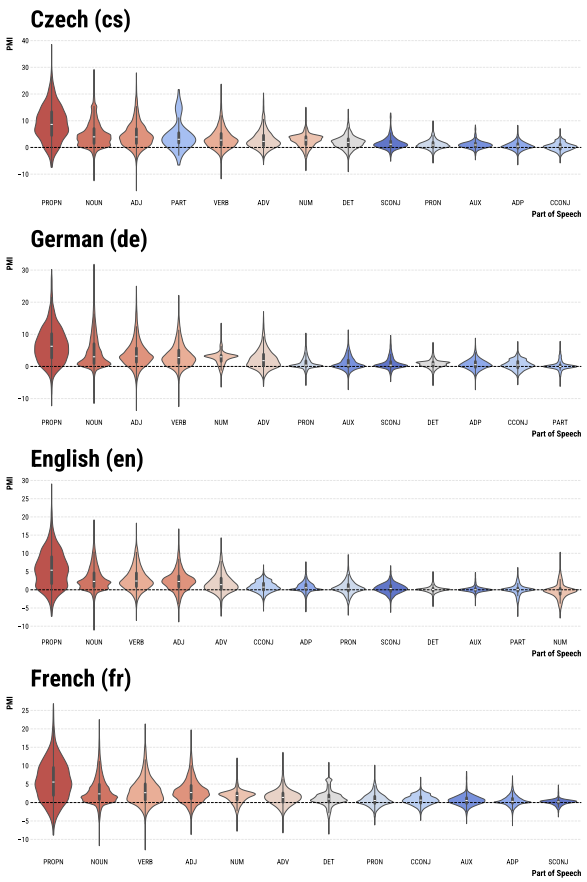

Finnish (fi)
PMI
PROPN NOUN NUM ADJ VERB ADV ADP CCONJ SCONJ AUX PRON
Part of Speech
Japanese (ja)
PMI
PROPN NOUN ADJ DET ADV VERB PART NUM PRON AUX ADP SCONJ CCONJ
Part of Speech
French (fr)
PMI
PROPN NOUN ADJ ADV VERB NUM DET CCONJ PRON ADP SCONJ AUX
Part of Speech
Korean (ko)
PMI
PROPN NUM NOUN SCONJ ADJ VERB ADV CCONJ PRON DET ADP AUX
Part of Speech
Hebrew (he)
PMI
AUX PROPN NOUN DET PRON ADJ VERB NUM SCONJ ADP CCONJ ADV
Part of Speech
Dutch (nl)
PMI
PROPN NOUN ADJ NUM VERB ADV AUX DET SCONJ CCONJ ADP PRON
Part of Speech
Hindi (hi)
PMI
NOUN PART PROPN VERB DET NUM ADJ ADV ADP CCONJ AUX PRON
Part of Speech
Norwegian (no)
PMI
PROPN NUM NOUN ADJ VERB ADV PRON DET PART ADP AUX CCONJ SCONJ
Part of Speech
Croatian (hr)
PMI
PART PROPN NOUN ADJ NUM VERB ADV PRON AUX DET ADP CCONJ SCONJ
Part of Speech
Polish (pl)
PMI
PROPN NOUN NUM ADJ ADV VERB PART DET CCONJ PRON SCONJ AUX ADP
Part of Speech
Hungarian (hu)
PMI
PROPN NOUN NUM ADJ SCONJ VERB ADV DET ADP AUX PRON CCONJ PART
Part of Speech
Portuguese (pt)
PMI
PROPN PART ADJ NOUN VERB ADV NUM PRON DET AUX SCONJ ADP CCONJ
Part of Speech
Indonesian (id)
PMI
PROPN ADV DET NOUN ADJ VERB NUM SCONJ PART PRON ADP CCONJ AUX
Part of Speech
Romanian (ro)
PMI
PROPN NOUN SCONJ ADJ ADV NUM VERB AUX PRON PART DET ADP CCONJ
Part of Speech
Italian (it)
PMI
PROPN NOUN NUM ADJ VERB DET PRON AUX SCONJ ADP ADV CCONJ
Part of Speech
Russian (ru)
PMI
PROPN DET NOUN ADJ NUM VERB ADV PRON SCONJ PART CCONJ ADP
Part of Speech



D Groundedness distribution for Multi30K

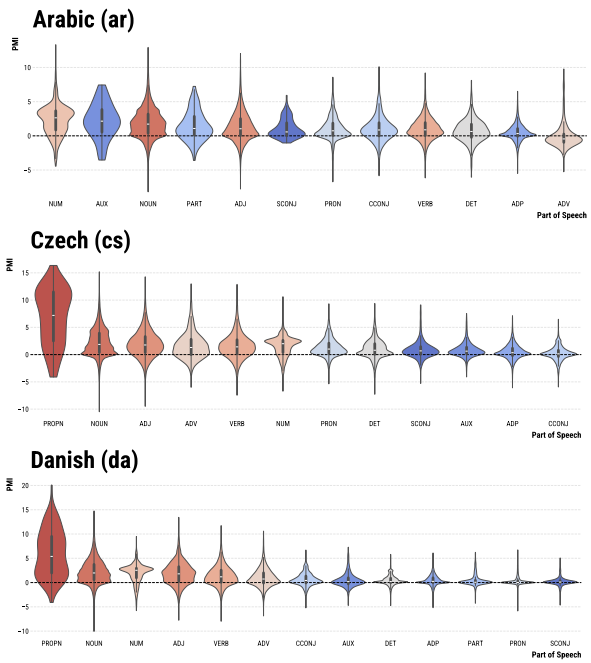
Results are ordered by descending mutual information estimate within the dataset (average groundedness/PMI). Hue indicates the average cross-linguistic ranking of a part of speech.

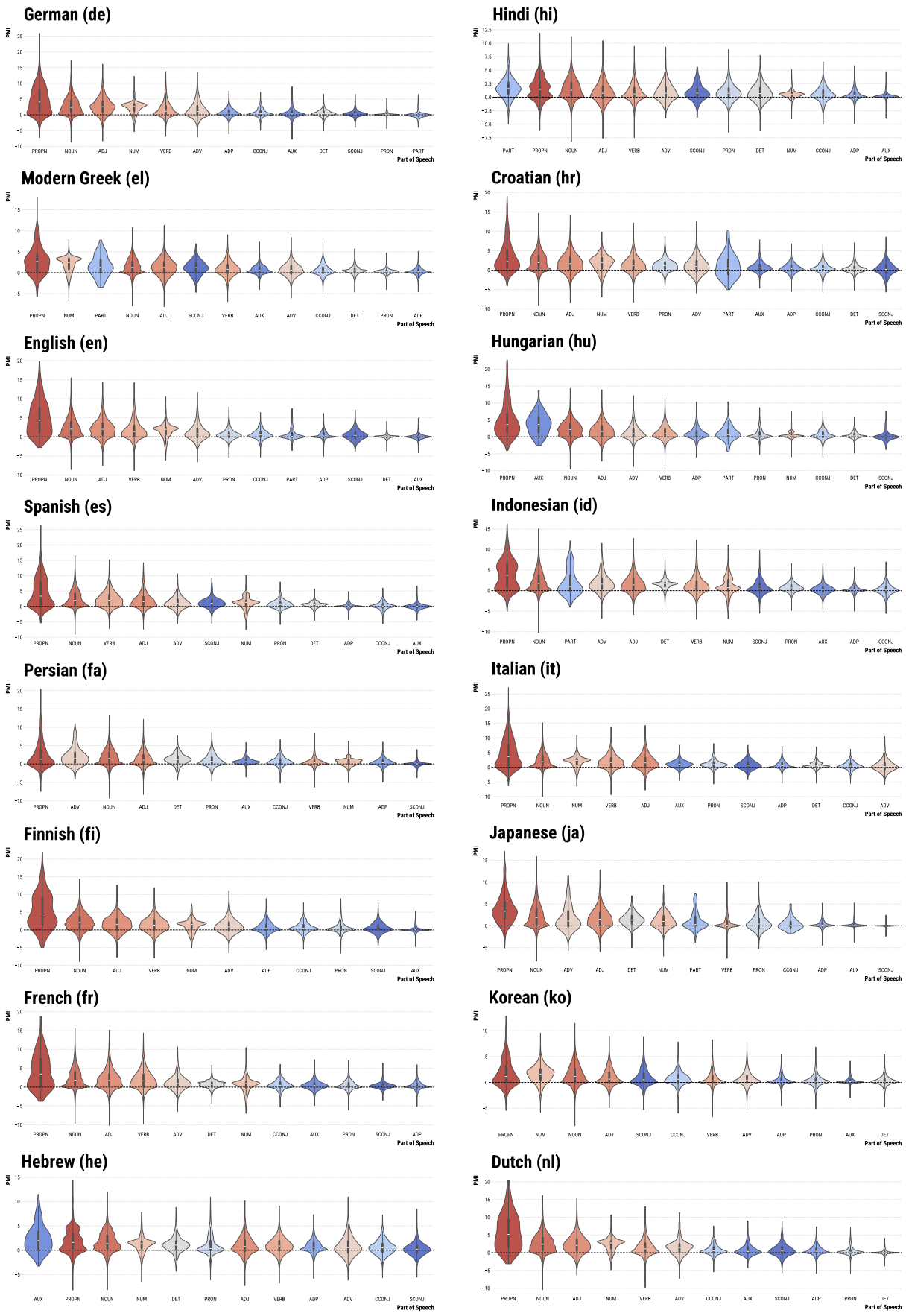


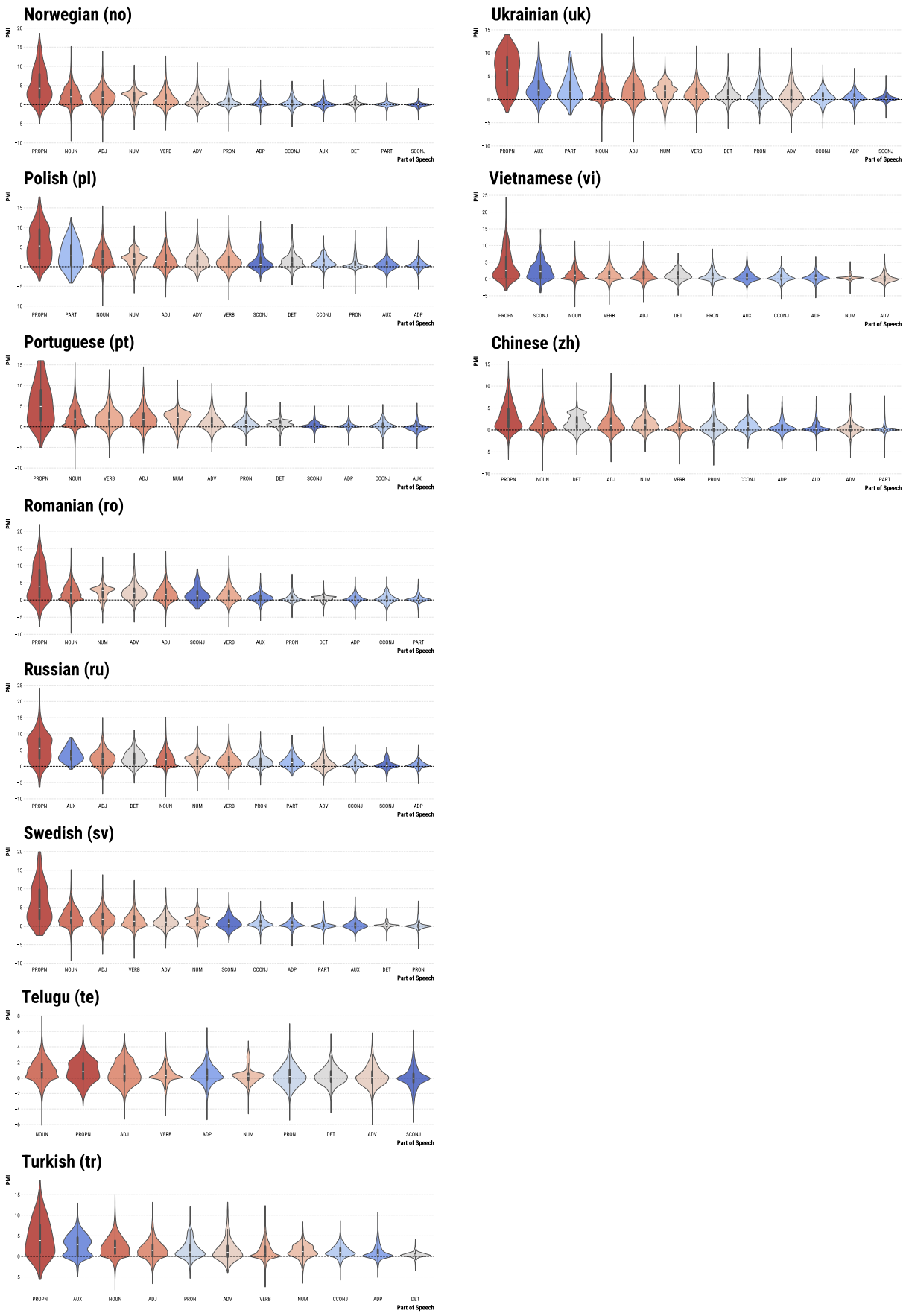


E Groundness distribution for COCO-35L Development Set

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.




German (de)
PMI
PROPN NOUN ADJ NUM VERB ADV ADP CCONJ AUX DET SCONJ PRON PART
Part of Speech
Modern Greek (el)
PMI
PROPN NUM PART NOUN ADJ SCONJ VERB AUX ADV CCONJ DET PRON ADP
Part of Speech
English (en)
PMI
PROPN NOUN ADJ VERB NUM ADV PRON CCONJ PART ADP SCONJ DET AUX
Part of Speech
Spanish (es)
PMI
PROPN NOUN VERB ADJ ADV SCONJ NUM PRON DET ADP CCONJ AUX
Part of Speech
Persian (fa)
PMI
PROPN ADV NOUN ADJ DET PRON AUX CCONJ VERB NUM ADP SCONJ
Part of Speech
Finnish (fi)
PMI
PROPN NOUN ADJ VERB NUM ADV ADP CCONJ PRON SCONJ AUX
Part of Speech
French (fr)
PMI
PROPN NOUN ADJ VERB ADV DET NUM CCONJ AUX PRON SCONJ ADP
Part of Speech
Hebrew (he)
PMI
AUX PROPN NOUN NUM DET PRON ADJ VERB ADP ADV CCONJ SCONJ
Part of Speech
Hindi (hi)
PMI
PART PROPN NOUN ADJ VERB ADV SCONJ PRON DET NUM CCONJ ADP AUX
Part of Speech
Croatian (hr)
PMI
PROPN NOUN ADJ NUM VERB PRON ADV PART AUX ADP CCONJ DET SCONJ
Part of Speech
Hungarian (hu)
PMI
PROPN AUX NOUN ADJ ADV VERB ADP PART PRON NUM CCONJ DET SCONJ
Part of Speech
Indonesian (id)
PMI
PROPN NOUN PART ADV ADJ DET VERB NUM SCONJ PRON AUX ADP CCONJ
Part of Speech
Italian (it)
PMI
PROPN NOUN NUM VERB ADJ AUX PRON SCONJ ADP DET CCONJ ADV
Part of Speech
Japanese (ja)
PMI
PROPN NOUN ADV ADJ DET NUM PART VERB PRON CCONJ ADP AUX SCONJ
Part of Speech
Korean (ko)
PMI
PROPN NUM NOUN ADJ SCONJ CCONJ VERB ADV ADP PRON AUX DET
Part of Speech
Dutch (nl)
PMI
PROPN NOUN ADJ NUM VERB ADV CCONJ AUX SCONJ ADP PRON DET
Part of Speech


Norwegian (no)
PMI
PROPN NOUN ADJ NUM VERB ADV PRON ADP CCONJ AUX DET PART SCONJ
Part of Speech
Polish (pl)
PMI
PROPN PART NOUN NUM ADJ ADV VERB SCONJ DET CCONJ PRON AUX ADP
Part of Speech
Portuguese (pt)
PMI
PROPN NOUN VERB ADJ NUM ADV PRON DET SCONJ ADP CCONJ AUX
Part of Speech
Romanian (ro)
PMI
PROPN NOUN NUM ADV ADJ SCONJ VERB AUX PRON DET ADP CCONJ PART
Part of Speech
Russian (ru)
PMI
PROPN AUX ADJ DET NOUN NUM VERB PRON PART ADV CCONJ SCONJ ADP
Part of Speech
Swedish (sv)
PMI
PROPN NOUN ADJ VERB ADV NUM SCONJ CCONJ ADP PART AUX DET PRON
Part of Speech
Telugu (te)
PMI
NOUN PROPN ADJ VERB ADP NUM PRON DET ADV SCONJ
Part of Speech
Turkish (tr)
PMI
PROPN AUX NOUN ADJ PRON ADV VERB NUM CCONJ ADP DET
Part of Speech
Ukrainian (uk)
PMI
PROPN AUX PART NOUN ADJ NUM VERB DET PRON ADV CCONJ ADP SCONJ
Part of Speech
Vietnamese (vi)
PMI
PROPN SCONJ NOUN VERB ADJ DET PRON AUX CCONJ ADP NUM ADV
Part of Speech
Chinese (zh)
PMI
PROPN NOUN DET ADJ NUM VERB PRON CCONJ ADP AUX ADV PART
Part of Speech