

A computational approach to typological comparative concepts for lexicality

Coleman Haley



Doctor of Philosophy
Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh
2025

Abstract

One major dimension of linguistic organization is the notion that there are more lexical linguistic units, which express meanings, and more functional linguistic units, which are determined by syntax and/or discourse and serve to organize and clarify the relationships between lexical elements. This dichotomy has been described at levels of linguistic structure and motivates at least two classical distinctions in linguistics. At the level of words, it motivates the so-called lexical-functional distinction, while within morphology, a related distinction is drawn between derivation (which forms new lexical items) and inflection (which produces forms of lexical items). These dichotomies have many noted boundary cases, which have led to many linguists rejecting them, or treating them as gradient. In this thesis, I refer to this gradient of semantic weight at different levels of formal structure as *lexicality*.

There is substantial neurological and psychological evidence for the importance of *lexicality* to human language processing. Further, *lexicality* dichotomies also emerge in cross-linguistic trends in grammatical organization, such as asymmetries between inflection and derivation, or between the properties of functional and lexical word classes. Yet the *lexicality* of a particular linguistic unit varies contextually and diachronically. I develop quantitative methods to test the consistency of these concepts across typologically diverse languages. First, I show inflection vs. derivation can be predicted with high accuracy from formal and distributional properties.

In linguistic practices that proceed from analysis of language-particular data to a language-general analysis, issues of *lexicality* have played a role of central importance. However, in the functional—typological tradition, which proceeds from cross-linguistic analysis to the language particular, the relationship of this dimension to linguistic organization has had little theoretical impact. A major factor is that typological research must be conducted with cross-linguistically

applicable comparative concepts. In this thesis, I leverage deep learning models to produce empirically grounded measures for lexicality, which I argue can serve as interesting and useful comparative concepts for typological study.

In the first part of the thesis, I focus on inflection and derivation, operationalizing a four-dimensional framework for formal and distributional properties of the distinction. I show that formal and distributional variability are strong correlates of this traditional distinction across a sample of 26 languages, and that the four measures can predict inflection vs. derivation with 90% accuracy

In the second part of the thesis, I introduce a novel groundedness measure, which aims to provide a cross-linguistic empirical ground for language function to quantify contextual semantic contentfulness. To do so, I leverage image–caption datasets and vision–language models. This measure captures the lexical–functional distinction in word classes across 30 languages but diverges substantially from related measures like concreteness.

Interestingly, groundedness displays asymmetries not just between lexical and functional items, but also among the major lexical classes of nouns, verbs, and adjectives. I argue that this suggests a connection between ideas of lexical word class continua in cognitive linguistics and the lexical–functional distinction. I apply groundedness to deviations from prototypical lexical class organization. I show that groundedness predicts the split between Japanese *na-* and *i*-adjectives, which has previously been thought to have little synchronic relevance. On the other hand, an investigation of the Tensedness Hypothesis shows the challenges with certain types of cross-linguistic comparisons of groundedness with current methods.

Lay Summary

Lay summary here

Acknowledgements

Acknowledgements here

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(Coleman Haley)

Contents

List of Figures	xiii
List of Tables	xvii
1 Introduction	1
1.1 The role of lexicality in linguistic organization	1
1.2 Approach	3
1.3 Structure of the Thesis	5
2 (Computational) Comparative Concepts and Lexicality	13
2.1 Typology and Comparative Concepts	14
2.1.1 Defining Hybrid Comparative Concepts	17
2.2 Finding Meaning in Computational Models	23
2.2.1 Type-level Distributional Embeddings	24
2.2.2 Contextual Embeddings and Language Models	26
2.2.3 Vision-and-language models	29
2.2.4 Rich representations from large-scale pretraining	33
2.3 Approaches to Comparative Concepts in Computational Typology	35
2.3.1 Comparative concepts in multilingual databases	35
2.3.2 Phonological typology	36
2.3.3 Semantic category systems	38
2.3.4 Multidimensional scaling	39

2.3.5	Deep learning models of comparative concepts	42
2.4	Formal and functional dimensions of lexicality	44
2.4.1	The formal dimension	45
2.4.2	The functional-semantic dimension	48
2.4.3	Summary	52
2.5	Chapter Summary	53
I	Inflection and Derivation	55
3	Corpus-based Measures for Inflection and Derivation	57
3.1	Introduction	57
3.2	Motivation for our measures	61
3.3	Method	66
3.3.1	Orthography-based measures	66
3.3.2	Distributional-embedding-based measures	67
3.4	Data	71
3.4.1	Data selection and summary	73
3.5	Distribution of the individual measures	75
3.5.1	Effects of Frequency	77
3.6	The role of syntactic information	78
3.7	Conclusion	81
4	Predicting Inflection and Derivation Cross-linguistically	83
4.1	Predicting inflection and derivation	85
4.2	Classification of Linguistic Types of Inflection	88
4.2.1	Categories of inflectional meaning	90
4.2.2	Inherent vs. contextual inflection and transpositions . .	93
4.2.3	Summary	96
4.3	Discussion	96

4.3.1	The role of our individual measures	96
4.3.2	Language generality	100
4.3.3	The classification approach	102
4.3.4	Inflection and derivation: gradient or categorical?	103
4.3.5	Are inflection and derivation identifiable from the statistics of language?	106
4.3.6	Classification and syntactic change	107
4.3.7	Future work	109
4.4	Conclusion	111
II	Word Classes	117
5	Groundedness and the Lexical–Functional Distinction	119
5.1	Background	122
5.1.1	Contentfulness and word class	123
5.1.2	Measuring contentfulness	123
5.2	Method	125
5.3	Experimental setup	127
5.4	Results	131
5.4.1	Which word classes are grounded?	131
5.4.2	Which word classes are more grounded?	132
5.4.3	How consistent is word class groundedness across languages?	134
5.4.4	Semantic dimension of the measure	138
5.5	Conclusion	139
6	Splitting and lumping: Visual groundedness as an organizing factor among lexical classes	141
6.1	Introduction	141

6.2	Continua among lexical word classes	144
6.3	Japanese adjectives	148
6.3.1	Method	150
6.3.2	Results	152
6.3.3	Discussion	154
6.4	The Tensedness Correlation	157
6.4.1	The typological finding	158
6.4.2	Theoretical explanation of the finding	160
6.4.3	Methodological background	161
6.4.4	Results	168
6.4.5	Discussion	169
6.5	Conclusion	173
7	Conclusion	175
7.1	Contributions	176
7.2	Limitations and future directions	177
7.3	Finis	178
A	Chapter 4: Results using Word2Vec embeddings	179
B	Part II: Model performance by language	183
C	Groundedness correlation plots for other psycholinguistic norms	187
D	Groundedness distributions by language and dataset	189
D.1	Crossmodal-3600	189
D.2	Multi30K	197
D.3	COCO-35L Development Set	198
Bibliography		207

List of Figures

2.1	An example semantic map for the dative domain, adapted from Haspelmath (2003). Nodes represent different functions which “dative-like” elements can express. The boundaries for English <i>to</i> and French <i>à</i> are shown in pink and blue, respectively. Both terms cover contiguous regions of the map, satisfying the Semantic Map Connectivity Hypothesis.	19
2.2	A typical vision-and-language model architecture, in the process of captioning an image. A vision transformer produces a representation of the input image, which is linearly projected into embedding vectors for an autoregressive transformer language model. The model generates text one subword token at a time based on the image and the preceding tokens. Here, it has generated the token <i>ying</i> as the next token after <i>_a _cute _cat _is _pla</i> . The token <i>ying</i> will be added to the input at the next time step to continue generating a caption.	31
3.1	The empirical distributions of our four measures (quantifying the magnitude M and variability V of changes in Form and in Embedding space) for inflections and derivations in UniMorph	76
3.2	The mean cosine similarity between FastText embeddings of words of the same and different parts of speech in UniMorph. .	80

4.1	Cross-validation accuracy and standard error in reconstructing UniMorph’s inflection–derivation distinction by various supervised classifiers. Linguistically-motivated hypotheses referred to in the text are denoted with letters	89
4.2	Probability and Odds ratio with 95% confidence intervals of being classified as derivation for various kinds of inflectional meaning. Inflections to the right of the dotted line were disproportionately likely to be classified as derivation by our model	92
4.3	Probability and Odds ratio with 95% confidence intervals of being classified as derivation for inherent inflections and transpositions	94
4.4	Probability and Odds ratio with 95% confidence intervals of being classified as derivation for inherent vs. contextual noun inflections	95
4.5	Our two most predictive measures for inflection and derivation. Saturation represents overlapping constructions. With respect to these two variables, the inflection–derivation distinction appears gradient rather than categorical	105
5.1	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.	133

5.2 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 quantitatively captures the distinction between lexical and functional classes.	134
5.3 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 graded distinction between lexical and functional classes.	136
5.4 Correlation between human concreteness ratings and type-level groundedness (PMI; left, $\rho = 0.368$) or uncertainty coefficient (right, $\rho = 0.609$): i.e., the average ratio between LM surprisal and captioning model surprisal.	137
6.1 Groundedness scores for <i>na</i> -adjective <i>makka</i> (completely red; right) and <i>i</i> -adjective <i>akai</i> (red; left) in the STAIR-full-dev dataset.	156
6.2 Groundedness of the verbal categories across the 30 languages in this study. Error bars represent standard error in the mean groundedness across the datasets considered (COCO-35L, XM3600, and Multi30K where available). Contra theoretical predictions, verb-y languages do not exhibit higher mean groundedness of verbs, but are somewhat below average. However, this effect is confounded by model quality issues, as suggested by the lower groundedness of verbs in non-Latin script languages.	167

6.3 Z-scored groundedness of the verbal categories. Error bars represent standard error in the Z-scores across the datasets considered (COCO-35L, XM3600, and Multi30K where available). The results suggest verbs are not <i>relatively</i> more grounded than other words in verby languages. However, we observe a clear effect of script, with languages written in Latin script exhibiting relatively more grounded verbs.	170
6.4 Z-scored groundedness of the verbal categories, with adjectives included for verby languages. Error bars represent standard error in the Z-scores across the datasets considered (COCO-35L, XM3600, and Multi30K where available). Despite the higher groundedness of adjectives than verbs in general, and concerns that legitimate members of the verbal category could be disproportionately “lost” to the adjective tag in verby languages, we still observe lower groundedness for the verby languages. This suggests an disproportionate effect of captioning and language model quality on verbs.	171
A.1 Accuracy in reconstructing UniMorph’s inflection–derivation distinction by MLP classifiers using Word2Vec- vs. FastText-based distributional features. Hypotheses referred to in the main text are denoted with letters.	181
C.1 Correlation between English psycholinguistic norms and type-level groundedness (left) or uncertainty coefficient (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.	188

List of Tables

3.1	Sample of an inflectional construction (upper table, German nominative plural) and derivational construction (lower table, English verbal nominalization with <i>-ion</i>) in our data	64
3.2	Descriptive statistics of our filtered dataset by language.	74
5.1	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, PaliGemma multimodal data for continued training , and COCO-35L image–caption pairs for fine-tuning. Symbols indicate whether models are trained on text data (A) or on multimodal data (mA).	128
6.1	Croft (1991)’s analysis of the conceptual categories of the major parts of speech and their semantic properties.	145
6.2	Differences in groundedness between adjective classes across datasets. “MT?” indicates whether the captions were machine-translated from English. The effect size is the increase in groundedness (in bits) associated with <i>na</i> -adjective-hood, estimated using a linear mixed effects model with fixed effects of word class and position and a random effect for word type. Overall, <i>na</i> -adjectives tend to be more grounded than <i>i</i> -adjectives. <u>(Significant results)</u>	153

6.3 The effect of adjective class on LM surprisal and captioning surprisal. We find that <i>na</i> -adjectives tend to be more surprising in the language model than <i>i</i> -adjectives, but this effect is reduced by conditioning on the images, resulting in higher overall groundedness. (Significant results)	155
B.1 Per-language performance metrics for the models used. A) CIDEr scores on Crossmodal-3600 (XM3600) and COCO-35L for the <code>paliGemma-3b-ft-coco35-224</code> 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.	184

Chapter 1

Introduction

1.1 The role of lexicality in linguistic organization

The distinction between LEXICAL and FUNCTIONAL linguistic units has played a role in the theory and analysis of language for millennia. This is to say, a distinction can be drawn between two poles for the role that linguistic units play in communication. On one end, we have the LEXICAL: linguistic signs which carry specific meanings, often referring out to objects or events in the world—nouns like *cat* or *tree*. On the other end, we have the FUNCTIONAL: signs which do not so much carry specific meanings, but rather serve to organize and clarify the relationships between lexical elements—like the tense marker *-ing* or the word *to*.

For as long as people have been studying language, this distinction has been connected to the idea that functional elements bear little semantic content. In the Greek tradition, Aristotle distinguished *phōnē* *sēmantik* (sign-bearing sounds) from *phōn* *ásēmos* (non-sign-bearing sounds), such as the class of *árthron* which includes prepositions and preverbs (?). This distinction was not limited to the proto-linguistics of Indo-European languages: in the 12th century the *Wén zé* (文澤) of Chen Kui (陈葵) catalogued *zhùchí* (lit. “helping words”)—corresponding

to what we would today call function words. In the *Yìzhù* (1311) Lu Yiwei defines this class as words that do not have a “precise concrete meaning” (?), and future authors would adopt the term *yìcí 空* (lit. “empty words”) to refer to this class (?). Further, psycho- and neurolinguistics provide ample evidence for differential processing of lexical and functional elements (Segalowitz and Lane, 2000; Diaz and McCarthy, 2009; Boye and Bastiaanse, 2018; Chanturidze et al., 2019; Neville et al., 1992; Pulvermüller et al., 1995; Friederici et al., 2000).

Despite this long history, the direct study of semantic contentfulness on linguistic structure remains largely pre-theoretical in linguistics, especially in large-scale cross-linguistic study. A major cause of this theoretical lacuna is the difficulty of specifying *semantic contentfulness* in a principled, cross-linguistically applicable way. This has led many linguists to avoid this notion entirely, focusing instead on how notions like frequency shape grammatical expression (?). In this thesis, I seek to address this gap by developing measures of the *semantic* and formal properties of the distinction between lexical and functional elements, and applying these measures to traditional conceptualizations of this distinction.

While typical discussions of the distinction between lexical and functional units tend to focus on a contrast between lexical words or roots and functional words or affixes, a closely related distinction is drawn *within* the domain of morphology between DERIVATION and INFLECTION. Morphology described as *inflectional* typically have all the prototypical properties of functional units: inflection is typically closed-class, bound, and obligatory, and align closely with “possible grammatical concepts.” In contrast, *derivational* morphology may express rich and perhaps unconstrained meaning, interacting idiosyncratically with the meaning of roots, and is typically optional. However, there are a few key differences with the lexical–functional distinction at the level of words. First, there are strong cross-linguistic tendencies for derivations to occur closer to the root than inflections do (Greenberg’s Universal 28) (Greenberg, 1966; Lieber

and Štekauer, 2014). Second, simple conversions or transpositions of word class (e.g. converting an adjective to a noun: *happy*→*happiness*) tend to pattern more similarly to derivations than inflection (e.g. scoping inside inflections), despite their apparent lack of semantic weight and highly obligatory and productive nature. This has led to substantial debate over whether such morphological constructions are better considered inflection (Haspelmath, 1996) or derivation (ten Hacken, 1994). Further, where exactly the boundary between “transpositions of word class” and more semantically rich derivations lies is also unclear. For example, the *-er* nominalizer in English forms agentive nouns from verbs (*teach*→*teacher*). Again, we have encountered a distinction between two poles that seem to be related to semantic weight, but with unclear boundaries. While derivational morphology tends to be more semantically rich than inflectional morphology, it is typically less semantically rich than lexical roots; however, in highly agglutinative languages like Inuktitut, derivational morphemes can carry semantic content comparable to roots in other languages.

This thesis concerns itself with these divisions between more lexical and more functional linguistic units, at multiple levels of linguistic structure. I treat both the lexical–functional distinction at the level of words and the inflection–derivation distinction at the level of morphology, and connect them to prototype phenomena among the major lexical classes of nouns, verbs, and adjectives. *In this thesis, I refer to this gradient of semantic contentfulness at different levels of linguistic structure as the spectrum of LEXICALITY.*

1.2 Approach

Multiple Levels of Linguistic Structure This thesis spans a *wide range* of levels of linguistic structure. While previous work has largely treated lexicality distinctions at different formal and semantic levels as separate, unrelated problems (e.g.

the inflection–derivation distinction at the morphological level; the distinction between lexical and functional word classes at the word level; or the distinctions between the major lexical classes of nouns, verbs, and adjectives), I investigate lexicality across multiple levels of linguistic structure, showing new parallels and connections between them. That being said, I limit my focus to *sub-phrasal* linguistic units (morphemes and words), leaving phrases, semantic frames, and more schematic constructions to future work.

Cross-linguistic investigation This thesis focuses on the *cross-linguistic* consistency of lexicality distinctions.¹ A large body of work in linguistic typology has argued that language-specific categories do not map onto some clean set of universal grammatical categories (Haspelmath, 2007; Croft, 2001; Dixon, 1977). Instead, typologists have argued for the importance of cross-linguistically valid *comparative concepts*—which need not necessarily map onto the structure of individual language’s grammar (Haspelmath, 2010; Croft, 2016). Studies that focus on the distinction between inflection and derivation or between lexical and functional word classes which consider only a single language risk conflating language-particular properties and categories with cross-linguistic generalizations. While finding a consistent distinction between inflection and derivation, or between lexical and functional word classes in an individual language is interesting, it does *not* a-priori tell us whether such a distinction has cross-linguistic descriptive value. Thus, this thesis aims to cover a large and diverse sample of languages wherever possible.

Computational and quantitative methods To study the cross-linguistic consistency of lexicality distinctions, I take inspiration from the successes of empirical grounding in certain areas of typology (like vowel, colour, and kinship

¹While in Chapter 6, I do conduct a language-particular analysis of Japanese word classes, the motivation for this analysis is to investigate whether cross-linguistically validated measures of lexicality can explain unusual language-particular patterns.

systems; Schwartz et al., 1997; Zaslavsky et al., 2019; Kemp and Regier, 2012) and from recent advances in deep learning models of language. These models have been shown to learn rich representations of semantics and the world, without requiring direct instruction on this structure, but rather learning it implicitly from learning to predict word sequences associated with a (linguistic and/or visual) context, and have been successfully employed to study a range of linguistic phenomena (Petersen and Potts, 2023; Hamilton et al., 2016; Matthews et al., 2025; Boguraev et al., 2025; Scivetti and Schneider, 2025). This capacity makes these tools ideal for operationalizing semantic dimensions of lexicality distinctions. The computational approach also aids in our goal of cross-linguistic investigation—while psychological and neurological evidence for lexicality distinctions exists, scaling it to typological study is challenging. Computational methods require only corpus data, which enables the simultaneous study of many languages, though it biases the study towards languages with sufficient digital resources. Further, the quantitative focus of this thesis enables the study and quantification of consistency, in contrast with many previous studies that focus primarily on problematic cases.

1.3 Structure of the Thesis

I explore three key questions in this thesis:

Q1: How can we operationalize semantic contentfulness in a cross-linguistically applicable way? (Chapters 3, 5)

Q2: How *cross-linguistically consistent* are lexicality-related divisions like the division between the lexical and functional word classes, or the inflection–derivation distinction? (Chapters 4, 5)

Q3: What is the relationship between *form* and *semantics* across the lexicality

spectrum? (Chapters 4, 6)

Chapter 2: (Computational) Comparative Concepts and Lexicality In this chapter, I expand on the theoretical framework for the thesis. I introduce in more detail the problems of cross-linguistic category comparison, and the method of comparative concepts for resolving these issues in typological research. I review the ways in which semantic function has been handled in typological comparative concepts, highlighting a role for deep learning models of language in the creation of semantic comparative concepts. Next, I trace the history of how comparative concepts have been (at times, implicitly) employed, handled, and studied in computational typology. Through this background, I highlight how the creation of empirically grounded comparative concepts through the development of technologies and measures outside of typology (like perceptual theories of vowels and colours) has enabled major advances in typological research in the domains where it has been possible. This serves as further motivation for the computational approach to comparative concepts taken in this thesis. Finally, I provide a broad-scale overview of the lexicality spectrum, identifying different manifestations of a correlation between semantic contentfulness and formal linguistic structure, providing the connective tissue for the studies that follow. More detailed background on specific lexicality-related distinctions is provided in the relevant chapters.

Part I: Inflection and Derivation

Part I of this thesis consists of Chapters 3–4, which focus on the inflection–derivation distinction drawn in morphology. These chapters are based primarily on the following journal article:

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

Chapter 3: Corpus-based Measures for Inflection and Derivation In this chapter, we introduce a computational framework for studying the inflection–derivation distinction. Inspired by Spencer (2013)’s description of the distinction, we introduce a set of four quantitative measures of morphological constructions, including measures of both the magnitude and the variability of the changes to *form* and *distribution* introduced by each construction. These measures operationalize the intuition that derivations are more semantically contentful and semantically variable, as well as proposed formal differences between the categories. Crucially, these measures can be computed directly from a linguistic corpus, allowing us to consistently operationalize them across many languages and morphological constructions.

In contrast to prior computational studies that focus on a single language, we investigate 26 languages using the UniMorph 4.0 corpus (Batsuren et al., 2022). We demonstrate that the measures are not explained by simple frequency effects, and that the distributional measures capture a limited amount of syntactic information in addition to semantic information. Using these measures, we find differences between inflection and derivation for all four measures, but substantial overlap for each individual measure, and that the distributional measures are more strongly associated with the distinction than the formal measures. Interestingly, we find that inflectional constructions are *more* formally variable than derivational constructions, contrary to some received wisdom in the literature on the topic but in line with the idea of the difference between inflection and derivation being similar to other lexicality distinctions.

Chapter 4: Predicting Inflection and Derivation In this chapter, we investigate whether a *combination* of the measures from Chapter 3 can discriminate inflection and derivation cross-linguistically, in contrast to prior studies which focused on single measures. We train classifier models to predict whether a construction

is labelled as inflection or derivation in UniMorph. This novel classification approach allows us to quantify how much of the inflection–derivation distinction a combination of our measures explains. We find that language-agnostic classifiers based on our measures are able to predict inflectional–derivational status with high accuracy (90%)—indicating a high degree of cross-linguistic *consistency* in the application of the distinction. We then analyse various sub-categories of inflection, finding inflectional transpositions like participles are *not* more likely to be misclassified as derivational, in line with Haspelmath’s 1996 argument that these are best considered derivational. Overall, our results are in line with a *consistent, yet gradient* view of the inflection–derivation distinction. Our results suggest that distributional and formal *variability* are the most important dimensions for the distinction, but the *magnitude* of distributional and formal change also play a role. These results indicate that a *combination* of formal and distributional/semantic properties can explain much of the way linguists have applied the inflection–derivation distinction in UniMorph, in line with the view of lexicality as a combination of formal and functional properties.

Part II: Word Classes

Part II of this thesis consists of Chapters 5–6, which focus on lexicality among (functional *and* lexical) word classes.

Chapter 5: Groundedness and the Lexical-Functional Distinction This chapter introduces a quantitative, language-agnostic measure of semantic contentfulness called *groundedness*. This measure attempts to overcome the *structuralist* nature of traditional information-theoretic measures of contentfulness, which treat linguistic form as the only source of information about meaning, and thus conflate formal and semantic information. To quantify groundedness, I leverage image–caption corpora, treating the image as a language-neutral proxy for

meaning. Using a language model and an image captioning model, I define visual groundedness as the pointwise mutual information between a token and the image, which is estimated by taking the difference in surprisal of the token under the two models. This measure captures how much information about the image is provided by the token in context.

I apply this measure to investigate the visual groundedness of Universal Dependencies parts-of-speech across 30 languages from 10 language families. I find the relative groundedness of parts of speech is consistent across languages, reflecting the classical lexical–functional distinction in the form of a continuum. This continuum interestingly includes a consistent and linguistically interesting ranking between the traditionally *lexical* classes (nouns > adjectives > verbs), a fact which I return to in Chapter 6. Our results suggest that traditionally functional word classes still carry semantic content, in line with prior psycholinguistic findings. These results suggest the utility of this measure as a general tool for studying contentfulness in linguistics, and of taking a visually grounded approach to typological questions. This chapter is based on a conference paper at which appeared at NAACL 2025:

Haley, C., Goldwater, S., and Ponti, E. M. (2025). A Grounded Typology of Word Classes. In *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)*, pp. 10380–10399, Albuquerque, New Mexico. Association for Computational Linguistics.

Chapter 6: Splitting and Lumping In this chapter, I investigate the relationship between visual groundedness and cross-linguistic variation in **lexical parts of speech**. While there has been substantial work in linguistic typology investigating cross-linguistic variation in the expression of the major lexical categories of nouns, verbs, and adjectives, this work has previously been largely disconnected from work on semantic contentfulness and the lexical–functional distinction; however, my results in Chapter 5 suggest cross-linguistic consistency in the

relative visual groundedness of these classes. Building on this finding, I connect existing continuum and prototype theories of lexical categories and meanings with groundedness. I conjecture that the role of semantic contentfulness in lexical categories can help explain cross-linguistic variation in lexical category organization. In particular, I focus on languages which have been argued to “split” or “lump” major lexical categories.

To establish this, I first investigate Japanese. In Japanese, words denoting “properties” have the unusual property of constituting two formally very distinct word classes, rather than a single “adjective” class. Building on the insight that one of these classes is more formally “nominal” (*na*-adjectives) and one more “verbal” (*i*-adjectives), I hypothesize that we should see analogous trends in function: one class serving more prototypically nominal functions and one more prototypically verbal. In terms of visual groundedness, this corresponds to higher values for the nominal class. I investigate two manually captioned datasets and one machine translated dataset, finding significantly higher groundedness for *na*-adjectives in the manually captioned datasets, in line with the theoretical predictions. This stands in contrast to previous studies, which have indicated little synchronic functional difference between the two classes.

To investigate lumping phenomena, I turned to the Tensedness Correlation, which correlates the formal similarities of adjectives to verbs in languages with a lack of obligatory tense marking on verbs. In languages with obligatory tense marking, the expression of adjectives is more similar to nouns. Drawing on prior explanations of the Correlation, I argue that they suggest that languages that lack tense marking (“verby” languages) should have more grounded verbs. Testing this hypothesis involves direct comparison of groundedness estimates across languages, which can be challenging to interpret. After careful consideration of potential confounds, I find no significant relationship between verb groundedness and tensedness, highlighting some limitations of the kinds of

questions that can be answered with current methods for estimating visual groundedness.

Chapter 7: Conclusion In this chapter, I summarize the contributions of this thesis, discuss limitations, and outline directions for future work.

Chapter 2

(Computational) Comparative Concepts and Lexicality

In this chapter, I argue for a new perspective on comparative concepts in linguistic typology, which grounds operationalizations of complex hybrid concepts in empirical measures of underlying linguistic dimensions. I present the major goals of typology and the challenges of defining comparative concepts for typology, and review existing approaches to these challenges and how they compare to my proposed approach. I then review the development of deep learning models of language, describing how they provide new avenues for defining empirical measures of semantic dimensions of language, and the richness of the semantic and conceptual information they acquire. I then provide a review of the study and application of comparative concepts in *computational* typology, highlighting how building rich empirical models of underlying semantic and perceptual spaces have been key to successful computational typological research, and the parallels between these approaches and my proposed approach. I also highlight the shortcomings of current discrete approaches to semantics in computational typology. Together, this motivates the empirical grounding approach to comparative concepts and linguistic categories that I take in this

thesis.

Finally, I provide a high level overview of the lexicality spectrum, defining formal and functional dimensions of lexicality, and describing their interrelationships. This sets the stage for the remainder of the thesis, which focuses on defining empirical measures of these dimensions, leveraging deep learning models of language, and investigating how these dimensions relate to existing lexicality-related distinctions in multilingual databases.

2.1 Typology and Comparative Concepts

Linguistic typology is the study of variation across the world’s languages. Typologists perform cross-linguistic comparisons with the aim of making generalizations about this variation. Such generalizations may consist of identifying and classifying languages into a small set of types (typological classification) or identifying cross-linguistically consistent patterns in variation. By studying this variation, typologists aim to identify the limits on and universals of human languages, and, often, to identify simple, language-neutral explanations of these limits.

To make cross-linguistic comparisons and identify cross-linguistic variation, typological research has explicitly or implicitly had to identify a frame of *alignment* between languages—typically taking the form of shared concepts identified across languages. Take, for example, the study of basic word order typology:

- (2.1) *Paul* *kisses* *Peter.*
 SUBJ VERB OBJECT

- (2.2) *pooru-wa piitaa-wo kisu-shiteiru*
 Paul-TOPIC Peter-ACC kiss-DO-PRES.CONT
 SUBJ OBJECT VERB

English and Japanese, then, vary in their basic word orders. In English, the verb is proceeded by the subject and followed by the object (“SVO order”),

while in Japanese, the default ordering is subject-object-verb (“SOV”). This comparison, however, relies on the consistent cross-linguistic identification of the categories of SUBJECT, VERB, and OBJECT. Such concepts over which cross-linguistic comparisons can be made have been termed COMPARATIVE CONCEPTS (Haspelmath, 2010; Croft, 2016).

Many of these comparative concepts present serious methodological challenges. It is well known that many categories in linguistics are semantically *motivated*, but not semantically *defined*. Take, for example, the category of subject. While subject’s across languages are typically the agents of an action, in the specifics of individual languages, there is additional complexity. While English has “SVO” order, in passive constructions, it is the patient and not the agent which appears in this initial position:

- (2.3) *Paul is kissed by Peter.*
 SUBJ VERB OBJECT

Defining the subject in terms of the obvious distributional commonality between *Paul* in (2.1) and (2.3) would make the claim “English is an SVO language” circular. While getting around this particular problem is relatively straightforward (through the deployment of additional constructional tests), this type of issue is pervasive in typological analysis, and careless or inconsistent application of categories cross-linguistically can lead to generalizations or even debates which are vacuous.

For example, frequent debates have occurred over whether a particular language has the category ADJECTIVE: a syntactic category covering property words, distinct from nouns and verbs. The typical structure of such debates involves identifying the behaviour of words which denote properties in a particular language in various constructions, and comparing their behaviour to members of other classes. For example, in Korean, both adjectives and verbs (but not nouns) inflect for tense, leading some to argue that Korean lacks a class of ad-

jjectives, and to claim that in Korean, adjectives are a type of stative verb. On the other hand, some have argued that because adjectives in Korean are somewhat restricted in terms of the tense–aspect–mood constructions they can appear with, they are better analysed as a distinct class. However, cross-linguistically, adjectives rarely inflect for tense or aspect, so this distinction is being made on a very language-particular basis.

Croft (2001) calls this type of syntactic argumentation **METHODOLOGICAL OPPORTUNISM**: the application of arbitrary language-particular criteria to identify distinctions between supposedly universal categories. This approach cannot lead to consistent generalizations across languages. If we consider a generalization like “adjectives do not inflect for tense”, then Korean is a counterexample if adjectives are not a type of verb. If they are a type of verb, then Korean is a counterexample to the generalization “adjectives require some kind of copula-like element in prediction”. To understand what the actual generalizations in typology are and whether a particular language is or is not a counterexample, we need to base these comparisons on cross-linguistically consistent criteria.

What the best comparative concepts are for a given problem is an empirical question, based on their predictive power in terms of generalizations about language variation. All the comparative concepts I have discussed so far are **HYBRID CONCEPTS** (Croft, 2016): they combine aspects of formal distribution with semantics. As an alternative, we might consider **FUNCTIONAL¹** compar-

¹By simultaneously addressing both issues of comparative concepts and the lexical-functional distinction in this thesis, I am trapped into a very confusing overload of the term *functional*; it has two, almost diametrically opposed meanings in the literature. As discussed in Chapter 1 the context of the lexical–functional distinction, *functional* refers to a pole of a continuum of linguistic behaviour, where “functional” items/elements/units are those which serve primarily to organize and clarify relationships between other elements. These elements are often described as “grammatical” or “lacking meaning”. In the context of comparative concepts and typological theory, however, linguistic *function* refers to the communicative content of linguistic expressions, as contrasted with its *form*: the specific linguistic realization which conveys a function. In this sense, the function of an adjective is basically its intension: the property it denotes, while the function of a verbal tense marker is the temporal and aspectual information it conveys about the event describe by the verb. Thus, a “functional comparative concept” in this sense is one which is defined in terms of the communicative content of the linguistic expression, rather than its formal distribution. Here, the terms “function” and “functional” are

ative concepts: concepts that invoke only the communicative function of the linguistic expressions studied, regardless of their formal distribution. Returning to the adjective example, this would be making comparisons of all expressions of property meanings across languages, as suggested by Haspelmath (2012). However, as noted by Croft (2016), such broad functional comparative concepts may fail to capture important cross-linguistic distinctions that are relevant for typological generalizations. Different types of properties may have different distributions within a single language: the expression of colour properties may pattern differently from emotions, for example. Thus, increasingly fine-grained functional comparative concepts are often necessary to capture cross-linguistic variation. Even with purported “functional” comparative concepts, something of the formal tends to creep in. Ultimately, after all, the typological generalizations are about the behaviour of linguistic expressions, which are formal entities. There are therefore two major dimensions of challenge in identifying useful, valid comparative concepts for typology: selecting the right functions, and formalizing categories of linguistic expressions which align with these functions. In the next section, I review different meta-approaches to these challenges, and describe the general way I tackle these issues in the thesis, which diverges from prior work in important ways.

2.1.1 Defining Hybrid Comparative Concepts

Constructions and strategies ? provides a useful terminological and conceptual framework for describing how typological research can and has often implicitly combined form and function into useful cross-linguistic generaliza-

preferred to “semantics” and “semantic” because the later terms are often taken to refer only to truth-conditional content, while function is inclusive information structure.

Of course, these senses are related, but in a quasi-antonymic manner: more functional elements have more abstract, relational, and language-internal functions. In this thesis, when I use the word “functional” to describe words, morphemes, items, elements, or units, I mean it in the lexical-functional sense. Other uses of the words “function” or “functional” generally refer to the communicative content sense (function as opposed to form), unless otherwise specified.

tions, which I will use to frame the discussion in this section. We will follow Croft (2022) to define a **FUNCTIONAL CONSTRUCTION** (p. 17):

any pairing of form and function in a language (or any language) used to express a particular combination of semantic content and information packaging²

This type of construction is defined only by *what* it expresses. Croft contrasts this with the narrower **STRATEGY** Croft (2022, p. 19):

a construction in a language (or any language), used to express a particular combination of semantic structure and information packaging function (the *what*), that is further distinguished by certain characteristics of grammatical form that can be defined in a cross-linguistically consistent fashion (the *how*).

The separation of strategies from functional constructions is a useful conceptual tool for approaching conflicting cross-linguistic formal data about the expression of a particular function. For example, one argument that Korean Adjectives are a type of verb is that Korean Nouns require a copula to be predicated, while Adjectives do not. Rather than saying “Korean lacks adjectives”, we can say that Korean uses different strategies for predication nouns and adjectives, with nouns using a copula strategy, and verbs using a zero-copula strategy. This method of description foregrounds the actual distributional data that needs to be explained. Indeed, empirically there are many interesting questions about the cross-linguistic distribution and co-occurrence of different strategies for particular linguistic functions.

Semantic maps An extremely influential method for relating form to function in linguistic typology is the use of **SEMANTIC MAPS** (Haspelmath, 2003; Croft, 2002, pp. 133–139).³ As I discussed, a major problem with replacing traditional

²Here, “information packaging” refers to the discourse organization of semantic content within an utterance. It is not of central importance to the present discussion.

³These sources summarize the emerging literature around semantic maps and standardize terminology; the method was developed over gradually over a few decades by a number of linguists, as described in the referenced passages.

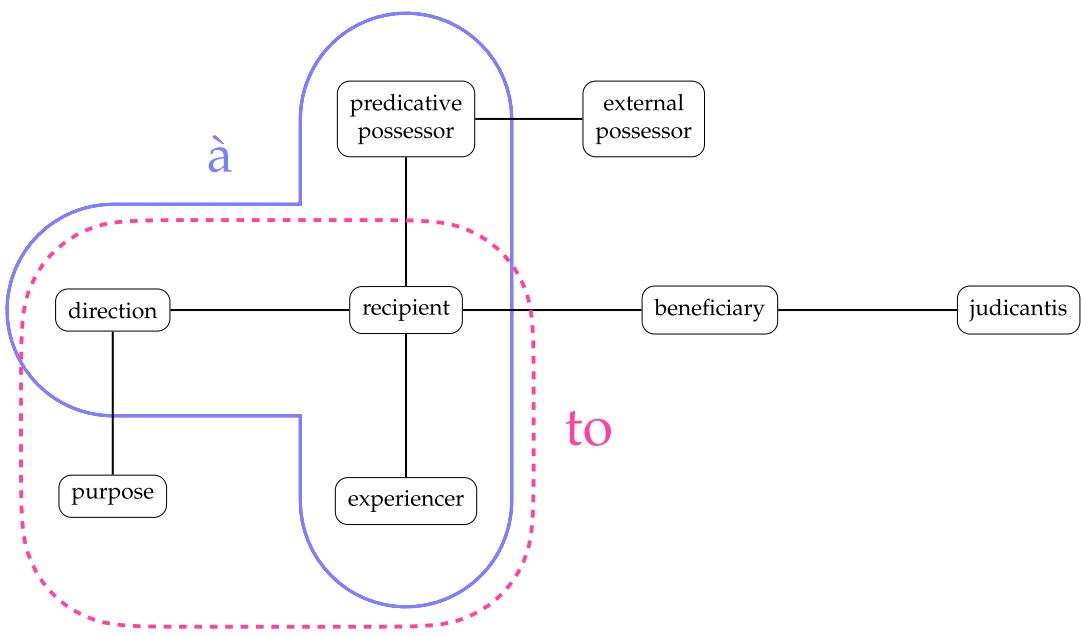


Figure 2.1: An example semantic map for the dative domain, adapted from Haspelmath (2003). Nodes represent different functions which “dative-like” elements can express. The boundaries for English *to* and French *à* are shown in pink and blue, respectively. Both terms cover contiguous regions of the map, satisfying the Semantic Map Connectivity Hypothesis.

comparative concepts like “adjective” with broad functional comparative concepts like “properties” is that languages may have different formal behaviour for different properties. As such, we may need to define more fine-grained functional comparative concepts. But what if we are still interested in a broader function like “properties”?

Semantic maps offer a solution by decomposing broad functional domains into a network of finer-grained functions, which can then be related to one another based on their observed co-expression⁴ across languages. Figure 2.1 shows an example of a semantic map. To construct such a map, the functional

⁴That is, using the same form or construction in a language. English co-expresses singular and plural second person as *you*.

categories are first selected on the basis of whether at least one pair of languages differ in their expression of a function. English uses *to* both for direction (“I’m going to the store”) and purpose (“I’m leaving early to be on time”), while French uses *à* for direction but *pour* for purpose. Thus, “purpose” must be a distinct functional category. The functional categories are then organized into a graph structure based on co-expression, with the aim to make language-particular strategies correspond to connected subgraphs of the semantic map. This desideratum has been termed the SEMANTIC MAP CONNECTIVITY HYPOTHESIS (Croft, 2001, p. 96), and has also been claimed as a universal property of human language. Designing maps to satisfy this property has the following corollary for the resulting map: if function A and function B are co-expressed in a language, and there is no path between A and B on the map that does not pass through function C, then C must also be co-expressed with A and B in that language. Croft (2001) calls this resulting graph the *conceptual space*, and this is the resulting language universal that is claimed by a particular semantic map analysis. On top of the conceptual space, we draw SEMANTIC MAPS for particular languages, which show the subgraphs of the conceptual space that are co-expressed in that language. The conceptual spaces, then, represent cross-linguistic *constraints* on the application of strategies to express particular functions cross-linguistically, thereby providing a comparative bridge between form and function. Semantic maps can be very useful for expressing the generalizations about problematic, broad hybrid comparative concepts like “dative” or “adjective.”

The semantic map method and the semantic map connectivity hypothesis underlying it have been extremely fruitful in identifying cross-linguistic co-expression patterns and universals. While conceptual spaces are not based on semantic similarity *per se*, rather facts of co-expression, in many domains where the semantic map method has been applied, the resulting conceptual spaces align closely with semantic similarity. In Section 2.3.4, I will describe

how computational methods have built on the theory and practice of semantic maps, as well as some of the limitations of existing approaches.

Retro-definitions Haspelmath (2021) proposes a radical approach to hybrid comparative concepts, which he terms RETRO-DEFINITION. In this approach, common but potentially problematic comparative concepts (“traditional comparative concepts”) are maintained as terms, but re-defined in a way that is cross-linguistically straightforward to operationalize and closely matches the traditional term. In this way, it represents a radical acceptance that comparative concepts need not relate to any “true” categories of language. As an example, Haspelmath proposes retro-defining adjectives as “property roots”, regardless of their syntactic behaviour in a given language. Even more radically, he proposes retro-defining “inflection” as morphemes that express a fixed set of meanings cross-linguistically, and “derivation” as any kind of word-formation process that expresses any other type of meaning (Haspelmath, 2024). This has the effect of allowing the precise usage of these terms in cross-linguistic comparison, but whether these definitions provide the most *useful generalizations* about language is an open question.

Prototype theory and fuzzy categories Another conceptual approach to dealing with the challenges of defining comparative concepts is to embrace the idea that categories are inherently fuzzy and gradient, and organized around a central *prototype*. This viewpoint was popularized in cognitive science by a series of seminal works by Eleanor Rosch, which demonstrated clear effects that people both agree which members of categories like “vegetable” or “furniture” are most prototypical, and that prototypicality influences processing (?). This finding was influential on early work in the development of cognitive linguistics, with theorists like Ronald Langacker and George Lakoff arguing that linguistic categories also have a prototype structure (Langacker, 1987; ?),

and has been proposed as a solution to conflicting cross-linguistic data about categories: unusual distributional behaviour is associated with less prototypical members of a category. However, the approach has come under fire for failing to account for apparent category boundaries or being unfalsifiable when applied to distributional data (Newmeyer, 1999; Haspelmath, 2024).

In many instances, prototype models are like a crude version of a semantic map model, because the semantic map connectivity hypothesis has similar implications: the longer the path between functions in conceptual space, the less likely they are to be co-expressed. However, rather than providing a fine-grained map of functions, prototype models focus on identifying central features, and suppose that less prototypical members should be less likely to be co-expressed.

My approach: grounding comparative concepts The main approach in this thesis does not fall neatly into any of the above widely-discussed approaches to hybrid comparative concepts, but takes a heterogeneous set of inspirations from them. The approach of the thesis is to define empirical measures which capture continuous dimensions of formal and functional variations, and relate them to *existing* category operationalizations. While this approach will be described in more detail later in the thesis, especially in ??, I will here briefly review how it relates to these existing approaches. Similar to the semantic map approach, I aim to take a complex, hybrid concept and decompose it into finer-grained dimensions. Because my measures are continuous, it shares with many versions of the prototype approach the idea that category membership is gradient. My approach takes from retro-definitions the idea that comparative concepts come from a linguistic tradition which needs to be questioned. However, rather than re-defining existing terms, I take an existing operationalization as a starting point, and investigate how well my empirical measures align with these operationalizations. In the thesis, I focus on standard, theory-neutral operation-

alizations from databases like UniMorph and Universal Dependencies. Because the relationship between my empirical measures and the operationalizations is the object of study, the claim does not rest on the a-priori validity of the databases—we aim to study the relationship of these comparative concepts *as they are used*. Future work can and should see if different operationalizations align better with the empirical measures I define. This empirical grounding approach also takes inspiration from literature in computational typology, which I will review in Section 2.3, but to my knowledge this literature has not been directly invoked in the context of comparative concepts.

Debates around comparative concepts often focus on problematizing existing attempts (Croft, 2016, p. 379). With this approach, I aim to reverse the discussion, by quantifying the consistency of existing operationalizations with respect to my empirical measures. In the context of this thesis’s focus on *lexicality*, this means relating distinctions among word classes or between inflection and derivation to highly *abstract* empirical measures; in contrast to the semantic map approach or many of Haspelmath’s retro-defintions, which focus on the *functions themselves*. In this way I aim to provide a new perspective on whether such distinctions really relate to the abstract properties which are often invoked to explain them, but have been difficult to measure directly. To do so, I rely heavily on emergent computational techniques for learning linguistic representations, which I will now describe in the next section.

2.2 Finding Meaning in Computational Models

Over the past two decades, deep learning models have revolutionized the field of natural language processing. These models come out of a history of brain-inspired cognitive modelling called *connectionism* (Rumelhart et al., 1986), which posited that important aspects of human cognition were best understood as

the emergent behaviour of large parallel and distributed networks of simple processing units. In natural language processing, deep learning models have demonstrated the ability to perform linguistic tasks at a level that would have once been thought to require human-level linguistic competence, like translating sentences or summarizing documents (Brown et al., 2020). While these models do not in-and-of-themselves provide a theory of human language processing, they provide a useful test bed for gradient and usage based theories of language (Futrell and Mahowald, 2025). In this section, I discuss several types of models, and the evidence that they acquire rich semantic and conceptual representations, and the importance of large-scale pretraining for this acquisition.

2.2.1 Type-level Distributional Embeddings

Much of the research on deep learning models of language is heavily indebted to the **DISTRIBUTIONAL HYPOTHESIS**, which posits that a word's meaning is determined by the contexts in which it occurs (Harris, 1954; Joos, 1950; Firth, 1957). This hypothesis led to the development of distributional semantic models, which represent the meaning of linguistic units (typically, words) as a function of their co-occurrence patterns with other words in large corpora. A distributional semantic model typically represents each word as a high-dimensional **VECTOR** (a point in high-dimensional Euclidean space), often referred to as the word's **EMBEDDING**. The application of neural networks to learn these embeddings led to a revolution in the field of distributional semantics in the early 2010s. The SkipGram or Word2Vec approaches (Mikolov et al., 2013a) learn vector representations of words by training a neural network to predict the context words surrounding a target word. The embeddings trained on this task are used as the word representations. With training on large corpora (consisting of millions or billions of words), these embeddings were quickly recognized to capture *semantic* similarity through their geometric properties: words which

humans judge as similar tend to have embeddings which are close in vector space (Mikolov et al., 2013b). Further, these embeddings were found to capture various types of semantic relationships through vector arithmetic: for example, the vector difference between the vector for *queen* is approximately equal to the vector for *king* minus the vector for *man* and plus the vector for *woman* (Mikolov et al., 2013b). These properties were improved upon by subsequent models like GloVe (Pennington et al., 2014) and FastText (Bojanowski et al., 2017), the latter of which incorporated distributional subword information to provide a back-off which is especially useful for representing rare words.

The rich semantic properties of distributional embeddings have been shown to have linguistic import. For example, Ettinger and Linzen (2016) showed that semantic priming effects can be predicted by embedding vector similarity, indicating that distributional information captured by these embeddings is predictive of human semantic processing. Distributional embeddings have also been shown to encode gradedness of properties of concepts—they encode the relative prototypical sizes of objects, as well as many other properties like speed, temperature, and gender (Grand et al., 2022).

A core hypothesis of the semantic map approach is that there is such a thing as universal conceptual similarity which is reflected through cross-linguistic co-expression patterns, and this literature has produced a rich body of findings supporting universal conceptual structure across a wide range of concepts (Youn et al., 2016; Haspelmath, 2003; Rogers, 2016; Regier et al., 2013). Distributional embeddings provide support for this view; the structure of embedding spaces learned from these co-occurrence patterns are similar enough cross-linguistically that embeddings trained on sufficient data can be largely aligned across languages using only simple linear translations (rotations, reflections, and scaling), even in an unsupervised manner (Vulić et al., 2020b; Hartmann et al., 2019). While of course, being word level representations, differences in lexicalization

and polysemy patterns across languages limit the degree of alignment that can be achieved, these results nevertheless provide additional evidence for universal conceptual structure underlying language.

The nature of SkipGram and related approaches implies that the embeddings capture *type-level* semantic information: each word has a single embedding, shared across contexts (this is often referred to as a STATIC embedding). This makes them especially well-suited for comparing to human experiments that don't provide utterance contexts, like the ones described above, or for modelling type-level data. However, this type-level nature prevents them from providing natural ways to capture *token-level* semantic variation. In Chapters 3–4, I use distributional embeddings because my data, like most morphological data, is type-level. Further, because at the word level, morphology can be thought of as a *relationship* between two types, I am able to study morphemes in combinations with many different roots, providing a diversity of contexts for each morpheme type, but not each word type.

2.2.2 Contextual Embeddings and Language Models

Researchers sought to overcome the type-level limitations of type-level embeddings like Word2Vec by developing CONTEXTUAL EMBEDDINGS, which provide a unique embedding for each token in context. Through developments in deep learning, larger and more complex models have been created to learn these embeddings, utilizing the transformer architecture (Vaswani et al., 2017). The most prominent approach to learning contextual embeddings is BERT, which uses a masked language modelling objective, learning to predict missing words in context (Devlin et al., 2019). This objective is used for PRETRAINING: fitting the model on an extremely large corpus for a task. From pretraining BERT model learns an embedding for each token in context, which provides a solid basis for FINE-TUNING on a range of downstream tasks; that is to say, training further

on a smaller task-specific dataset and objective to create a better task-specific model. Examples of linguistically-oriented tasks where BERT embeddings achieve new levels of performance include part-of-speech tagging and word sense disambiguation (Tenney et al., 2019b; Wiedemann et al., 2019; Chronis and Erk, 2020).

However, the transition to token-level embeddings produced new challenges for interpreting representations. Today’s models are *deep*, meaning they have many *layers*, each of which provides a different representation of a token. It is not always clear which layer is the *right* one for a given task—representations become more contextualized at deeper layers (Ethayarajh, 2019), but features like parts of speech are better represented in early layers (Tenney et al., 2019a). Today’s models are also *subword-level*—they represent rare words as sequences of multiple tokens, each representing a subword selected by a statistical learning algorithm (usually byte-pair encoding or a similar method; Sennrich et al., 2016), so the units which have vector representations are not always aligned with linguistic structure. Finally, distance between embeddings is dominated by *rogue dimensions*, a handful of dimensions of large magnitude and obscure semantic similarity (Timkey and van Schijndel, 2021). After better understanding these issues, researchers were able to extract rich semantic information from contextual embeddings by pooling across contexts and subwords (Bommasani et al., 2020; Eyal et al., 2022), standardizing dimensions (Timkey and van Schijndel, 2021), and selecting layers based on task. All this is to say, as modelling approaches have become more complex, it requires more care to identify and extract the semantic information encoded in these models. After such discoveries, contextual embeddings have been shown to capture interesting semantic information, about, e.g., different senses of *break* in English (Petersen and Potts, 2023) and constructions like *a beautiful five days* or *day by day* (Scivetti and Schneider, 2025; Rozner et al., 2025)

For the next class of models I discuss, identifying a semantic space is still an area of extremely active research. These are the so-called AUTOREGRESSIVE LANGUAGE MODELS (or simply language models), which are trained to predict the next token in a sequence, given all prior tokens. These models have received enormous attention recently due to their ability to generate fluent and coherent text and solve complex tasks “in-context”, meaning by sequentially generating to complete a sequence which includes instructions for the task (Brown et al., 2020). In line with these capabilities, these models are typically very large, with billions to hundreds of billions of parameters, and are trained on massive corpora of billions or trillions of tokens. Their impressive capabilities have led to a surge of interest in understanding what kinds of linguistic and world knowledge they acquire during training (Futrell and Mahowald, 2025).

Despite impressive performance, we are a ways off from a clear understanding of how these models represent meaning. Their autoregressive nature means the vectors “representing” a particular token are only conditioned on the *preceding* context—which means sense information may be distributed onto later disambiguating tokens (e.g. in *break the law* vs *break the news*). Nevertheless, these models have been shown to be impressive predictors of incremental processing in humans: both of psycholinguistic performance through reading times (Staub, Forthcoming; Wilcox et al., 2023), and of neural activity during language processing (Schrimpf et al., 2021; AlKhamissi et al., 2025). Further, the nascent literature on interpreting the internal representations of these models has shown evidence for rich conceptual representations. Using feature and circuit extraction techniques like sparse autoencoders, researchers have identified highly abstract features, like an eye feature that responds to mentions of eyes across languages, and in code and ASCII drawings of faces and eyes (Tarng et al., 2025). Other work has identified circuits which copy abstract concepts even when words or tokens differ (Feucht et al., 2025) and shown linear gradability

effects with vectors representing properties like BEAUTY (Kozlowski et al., 2025). Many of these models are *multilingual*, being able to generate text in a range of languages, and there is some evidence that they use *shared* representations across languages, for relations among lexical concepts (Lindsey et al., 2025), and for traditionally “grammatical” or “morphological” concepts like tense, case, and number (Brinkmann et al., 2025).

Together, these results provide evidence that deep learning models of language acquire rich semantic and conceptual representations, at both the type and token levels. These representations can be extracted and studied to provide insights into human language processing, as well as the nature of linguistic meaning. These representations are also naturally gradient, making them well-suited for studying gradient theories of language and meaning, something that has been challenging with the traditional approaches in linguistics (Petersen and Potts, 2023; Futrell and Mahowald, 2025).

2.2.3 Vision-and-language models

While language models and embedding models have demonstrated impressive capabilities in acquiring semantic representations from text alone, their basis in a strong form of the distributional hypothesis makes fully segregating form and function difficult. Because the aim of typology is to identify cross-linguistic generalizations about how different languages express the same communicative functions, typologists usually aim to identify functions which are as form-agnostic as possible (functional constructions) for at least some types of cross-linguistic comparison. In the second part of this thesis, I aim to bring this spirit into computational approaches to typology, by utilizing recent advances in VISION-AND-LANGUAGE MODELS (VLMs). These models can process a combination of visual and textual input, allowing them to model scenarios where language use is grounded in and modulated by visual perception—such

as visual storytelling, visual question answering, and image captioning (Lin et al., 2014; Antol et al., 2015; Huang et al., 2016). In this way, VLMs provide a way to combine the computational power of language models with a language-and form-agnostic source of meaning/function: the visual modality.

While VLMs have been the subject of less interpretability⁵ research and linguistic analysis than pure language models (due to the rapidly changing nature of the field), the evidence so far suggests a similarly rich picture to the other models of language discussed in this section. VLMs have been shown to exhibit better understanding of hypernyms and categorical structure than pure language models (Qin et al., 2025), to possess units that respond to the same concept presented in different visual forms (e.g. images, text, and drawings; Goh et al., 2021), and to better predict hippocampal activity for “concepts” than unimodal models (Choksi et al., 2022). VLMs have also been shown to share representations *both* across languages *and* across the visual and textual modalities (Wu et al., 2025), and to acquire representations of lexical similarity that align with human judgements (Yun et al., 2021). These results suggest that VLMs have rich internal representations which combine information from both modalities, and that these representations can be studied to provide insights into human language processing and meaning. I will now provide a high-level overview of how these models are constructed to provide context for their use in Part II.

Making a vision-and-language model

Today’s state-of-the-art VLMs are typically built out of two key components: a **VISION ENCODER**, which processes the visual input, and a **LANGUAGE MODEL**, which produces text output conditioned on the visual input. Figure 2.2 shows a schematic of this architecture, in the process of generating a caption for an

⁵Interpretability is the field of research which aims to understand the representations and processing of machine learning models.

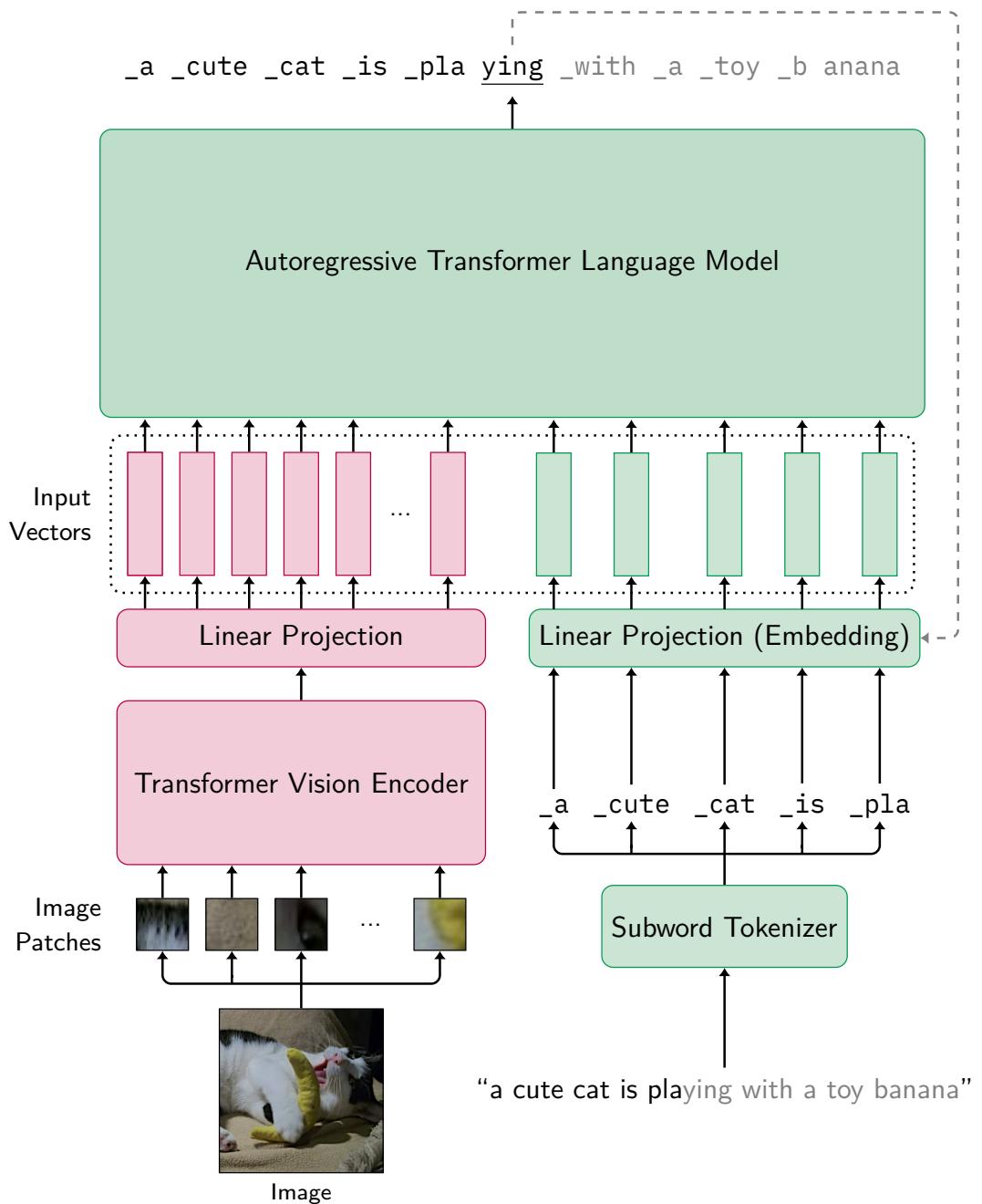


Figure 2.2: A typical vision-and-language model architecture, in the process of captioning an image. A vision transformer produces a representation of the input image, which is linearly projected into embedding vectors for an autoregressive transformer language model. The model generates text one subword token at a time based on the image and the preceding tokens. Here, it has generated the token `ying` as the next token after `_a _cute _cat _is _pla`. The token `ying` will be added to the input at the next time step to continue generating a caption.

input image. In this thesis, I use the PaliGemma VLM (Beyer et al., 2024), which follows this general approach.

The vision encoder in a VLM typically uses a vision transformer to provide a high-dimensional representation of an input image. A vision transformer is a variant of the transformer architecture (Vaswani et al., 2017) which can process images by dividing them into patches which are embedded using a learned linear transformation and then processed similarly to tokens in a text transformer (Dosovitskiy et al., 2020). This model is usually trained together with a transformer text encoder on an extremely large dataset of image–caption pairs. In training, the model maximizes the similarity between embeddings of images and their corresponding captions, while minimizing the similarity of embeddings of non-corresponding image–caption pairs (Radford et al., 2021; Zhai et al., 2023). After training, the vision encoder can be used independently to produce high-dimensional vector representations of images without the need for captions.⁶

The language model component is typically initialized from a pretrained autoregressive transformer language model of the kind discussed in Section 2.2.2. The vision encoder is connected to the language model in the following way: the output of the vision encoder (a high-dimensional vector) is linearly projected into a series of vectors which are used as “prefix” tokens to condition the language model’s text generation (Tsimpoukelli et al., 2021; Karpathy and Fei-Fei, 2017).

With continued training on image–text pairs for a variety of tasks (image captioning, optical character recognition, visual question answering, etc.) the output of the vision encoder is adapted to condition the language model’s text generation. In this way, the language model can generate text which is grounded in the visual input. This approach also means that most VLMs *include* a language

⁶Vision encoders used to be trained without text encoders, using objectives like image classification (Deng et al., 2009), but the use of contrastive learning with text has been shown to produce substantially better representations for a diverse range of target tasks (Radford et al., 2021).

model which can be used independently of the vision encoder, a fact which I exploit in Chapter 5.

2.2.4 Rich representations from large-scale pretraining

In this section, I have reviewed several classes of deep learning models of language, highlighting key similarities and difference among them as pertains to their representation of semantic information. Distributional embeddings like Word2Vec and FastText leverage collocational patterns to learn a high-dimensional vector space where geometric relationships between vectors correspond to semantic similarity. Contextual embeddings like BERT extend this approach to provide token-level representations which can capture contextual nuances and constructional meanings, but require more careful interpretation to extract conceptual semantic information. Autoregressive language models like the GPT family learn to predict the next token in a sequence, acquiring representations that solve complex tasks and are especially useful for modelling human language processing, but their internal representations are still not well understood. Finally, VLMs combine visual and textual information to ground language in perceptual content, preserving the strengths of language models while providing a form-agnostic source of meaning.

The major theme of all this research, and indeed of most of the research in natural language processing in the twenty-first century, is the value of *large-scale pretraining* on massive datasets (Saphra et al., 2024). In between each of the approaches discussed here, there were many intermediate models and data-efficient approaches proposed, but these models achieved their success and long-standing impact through pretraining on large corpora. The advent of the fine-tuning and few-shot learning paradigms further cemented the hegemony of pretrained models, as increasingly even small-data tasks are best solved by leveraging large pretrained models (Brown et al., 2020; Devlin et al., 2019).

Not only has large-scale pretraining been the key to improving performance on natural language processing tasks, but it has also paralleled findings in typology about the universal conceptual structure driving differences and similarities across languages. In the same way as coexpression patterns reveal a conceptual space respected across languages, sufficiently large-scale pretraining on datasets induces isomorphic word vector spaces across languages (Vulić et al., 2020b), and produces convergent representations across different model architectures, languages, and modalities (Jha et al., 2025; Huh et al., 2024; Brinkmann et al., 2025). Indeed, a growing body of work suggests that representational alignment between deep learning models and *the human brain* is driven primarily by the discovery of a shared, universal conceptual structure (Kauf et al., 2024; Hosseini et al., 2024; ?; Antonello and Huth, 2024). And on top of this, the models that result from large-scale pretraining have been shown to be sensitive to extremely fine-grained semantic and distributional distinctions (Petersen and Potts, 2023; Rozner et al., 2025; Goldberg, 2024), making them exceptionally rich sources of linguistic representations.

Therefore, I argue that the representations learned by large-scale pretrained models provide a promising avenue for studying function in typology. While the large-scale pretraining paradigm presents challenges and limitations for typology (as most of the world's languages lack sufficient data for current methods), the linguistic capabilities of these models will be essential for bringing a stronger empirical basis to the notion of semantic contentfulness in this thesis, and presents exciting avenues for future work in typology more broadly. In the following section, I will review how comparative concepts have been used in computational typology to date, and how the representations learned by large-scale pretrained models can provide new avenues for defining, studying and operationalizing functional comparative concepts.

2.3 Approaches to Comparative Concepts in Computational Typology

This section reviews the roles that comparative concepts have played in computational approaches to linguistic typology. In this section, I aim to highlight how technological advances can enable new types of comparative concepts, and how the choice of comparative concepts can shape the types of questions that can be asked and generalizations that can be formed. In particular, I will argue that computational modelling thrives when fine-grained distance metrics are defined, be this in a high-dimensional discrete space or especially a continuous space.

2.3.1 Comparative concepts in multilingual databases

The predominant approach to comparative concepts in computational typology has simply been to follow the lead of whatever annotation scheme is used in the typological or cross-linguistic database used in the study. For example, influential works like Futrell et al. (2015), which demonstrated minimization of dependency length as a universal pressure on word order, used the data of Universal Dependencies (?), as-is. Similarly, cross-linguistic studies of morphological complexity and inflectional paradigms have relied on a combination of feature values from databases like the World Atlas of Language Structures (WALS) (?), or the encodings of grammatical features in UniMorph (Batsuren et al., 2022).

Databases like UniMorph and Universal Dependencies attempt to use a single cross-linguistic annotation scheme for their comparative concepts, but these schemes usually much more closely resemble the hybrid categories of language-particular analyses. For example, UniMorph’s feature set includes values like TENSE=PAST, with no information about what constructions that

form is used in. Chapter 3 includes a detailed discussion of some limitations of UniMorph annotations, which are largely based on language-specific grammatical traditions rather than typological best practices. Similarly, Universal Dependencies part-of-speech tags are cross-linguistically “universal”, but they are deployed with the “methodological opportunism” described by Croft (2002), where categories are defined in a language-particular way. Very recently, Universal Dependencies has begun a process of aligning their representational scheme with more fine-grained and cross-linguistically valid comparative concepts (??), but this process is still ongoing.

That the annotation of these databases is *not* consistently based on cross-linguistically valid universal comparative concepts actually provides an interesting opportunity, however, and one which this thesis exploits. A major aim of this thesis is to define computational comparative concepts of dimensions hypothesized to underlie grammatical category distinctions cross-linguistically, and then investigate how these dimensions relate to the categories used in these databases. In this way, I can investigate the extent to which these databases distinctions, while flawed, nevertheless align with the underlying dimensions of meaning that motivate these distinctions.

2.3.2 Phonological typology

Phonology offers the earliest and clearest example of how empirical, continuous measures can advance typology. It has long been recognized that language-specific categories like phonemes are ill-suited for cross-linguistic comparison (“phonemes are not fruitful universals”; ?) because they are defined by language-internal contrasts. To address this, ? introduced a *feature-based* account of phonetic universals, later refined by ? through the notion of *natural classes*.

Vowels have long been central to typological inquiry. ? proposed early implicational universals about vowel systems within his proto-featural frame-

work. Yet within this featural approaches, universals of vowel systems were understood as complex, varying substantially on the number of vocalic contrasts within a language. The true generalizations, which turn out to be extremely simple when properly understood in a continuous space, required substantial developments in the acoustic theory of vowels. The decisive shift came with the formant theory (??), which linked vowel quality to acoustic resonances (F1–F3). The development of better technologies for measuring and recording formants through the twentieth century ultimately provided a real-valued acoustic representation (?) that allowed vowels to be compared across languages in a shared empirical space, enabling the development of theories that made quantitative, testable predictions. The quantal theory (?Stevens, 1989) proposed that languages prefer perceptually stable regions of this space, while the dispersion theory (Liljencrants et al., 1972) modelled vowel inventories as systems maximizing perceptual distance. By simulating optimal vowel systems and comparing them to attested inventories, these models offered a precise computational account of typological tendencies. Subsequent dispersion–focalization models (??) and probabilistic analyses of entire vowel-system distributions (?) further refined these predictions, revealing the relative influence of competing pressures such as distinctiveness and perceptual stability.

While vowel typology concerns formal acoustic dimensions, its trajectory exemplifies a broader lesson: defining an empirical, continuous underlying space can transform typological theory. Just as formant space enabled precise and falsifiable generalizations about vowel systems, deep learning models may provide an analogous empirical grounding for meaning. Linking model-derived semantic spaces to human cognition could allow typology of linguistic function to achieve the same level of quantitative precision. In the next section, I turn to evidence from semantic category systems supporting this view.

2.3.3 Semantic category systems

The domain most closely paralleling vowel typology in its treatment of function is the study of semantic category systems—cross-linguistic analyses of how languages partition a shared underlying semantic space. As with vowels, researchers model these systems as optimizing trade-offs among universal pressures such as simplicity, communicative efficiency, and learnability. Once an underlying space is defined, these pressures can be formalized, simulated, and quantitatively tested against attested systems.

The case of colour terms provides the clearest illustration. In their seminal study, ? identified robust implicational hierarchies—two-term systems distinguishing light from dark, three-term systems adding red. Berlin and Kay couched these generalizations in terms of an implicational hierarchy of colour terms; however, their methodology could offer little insight into *why* these hierarchies exist. The breakthrough came with the development of a precise underlying perceptual space for colour. Studying the relationship between the colour space in terms of frequency and perceptual distance, Later work showed that a continuous perceptual space was key. Building on the CIEL^a*b^b* colour space (?), which aligns physical and perceptual properties of colour, Regier et al. (2007) modelled how systems maximize within-category similarity and between-category distinctiveness. These simulations closely predicted attested colour inventories and revealed that real systems are significantly more optimal than chance. This initial effort has expanded into a rich literature, modelling trade-offs among different pressures such as communicative need, perceptual structure, and learnability. While this remains a more rapidly evolving area of research than vowel system typology, studies continue to refine hypotheses and distinguish between pressures with increasingly fine-grained predictions about colour systems.

Comparable approaches have since been applied to other semantic do-

mains—kinship, number, quantifiers, modals, and pronouns—where discrete conceptual structure allows for tractable mapping. More challenging are domains with continuous meaning spaces, such as spatial terms (?) and tense–aspect marking (?), which have required simplifying the space into coarse discrete categories or low-dimensional projections (e.g. multidimensional scaling of usage data; ?). Therefore, work in these frameworks has primarily focused on domains where an underlying conceptual space can be more straightforwardly defined.

Overall, both the study of vowel systems and semantic category systems show that by creating an operationalization of an underlying space, we can test and discover parsimonious and predictive theories of the fundamental data of linguistic typology. In each case, substantial typological progress was able to be made without access to the “true” space—we are still learning about the perceptual dimensions of vowels (), and CIEL^a^b has known shortcomings ()—but the development of some empirical model of the underlying space that could be coded across languages was critical for this progress. However, the limited scope of functions studied in these types of computational frameworks points to the challenge of defining an underlying space for more complex semantics. In this thesis, I argue that recent advances in deep learning models of language are likely an early step in this direction, analogous to the early stages of development of the formant theory of vowels.

2.3.4 Multidimensional scaling

The multidimensional scaling (MDS) approach to semantic maps proposed by Croft and Poole (2008) represents the most well-developed technique for modelling a continuous functional space for typological comparison. This approach was developed to address the challenge of translating large typological

datasets into a traditional semantic map following the methodology of ?,⁷ and to provide a stronger mathematical basis for the semantic map theory. These methods take as input a high-dimensional discrete matrix of linguistic data, and produce a low-dimensional continuous representation of the data, using either optimal unfolding, or in some studies, matrix decomposition techniques. The resulting map provides an approximation where Euclidean distance approximates the frequency with which two functions are expressed with the same form. In this way, an underlying semantic/conceptual space is being inferred from typological data about form—representing a move away from the discrete identification of functional comparative concepts to a richer emergent empirical representation of the complexities of meaning, towards the desiderata of this thesis. Studies vary primarily in the way they construct the input dissimilarity matrix, and thereby in how much they allow for an emergent representation of function. van der Klis and Tellings (2022) provide a recent overview of the different methods used in the literature, which I briefly summarize here to illustrate how differing approaches rely on different comparative concepts. They provide a three-way typology of approaches to constructing the input for an MDS analysis.

First, there is the classical input representation, which Croft and Poole (2008) used to recreate ?'s analysis of indefinites.⁸ This type of map takes a set of N linguistic forms \mathcal{F} , and a set of K underlying functions \mathcal{M} . The binary input matrix $\mathbf{I} \in \{\text{Y}, \text{N}\}^{K \times N}$ is constructed such that

$$\mathbf{I}_{i,j} = \begin{cases} \text{Y} & \text{if form } f_i \text{ from language } l \text{ conveys (or is used to express) function } m_j, \\ \text{N} & \text{otherwise.} \end{cases}$$

to which the optimal unfolding technique is applied. In this type of analysis, the functions are entirely manually posited by the typologist and abstracted

⁷While an algorithm for producing graph-based maps from large-scale data was later introduced by ?, the MDS techniques retain a number of advantages.

⁸Other examples of this type of map include...

away from the constructions on which the functional claim is based. As a result, the MDS analysis cannot capture fine-grained variations around the prototypes of a given abstract function, but can capture differences in the closeness of two functions in a more fine-grained way than the classical approach to semantic maps—frequency of form-function co-occurrences is modelled. Nevertheless, in terms of comparative concepts, the functions here retain the traditional approach to function in typology, with its known shortcomings.

Croft and Poole (2008) also introduce a second method for producing an MDS map, which allows more fine-grained study of the prototype structure of functions. This second map relies on ?'s tense-aspect data. Here, rather than manually determining in a binary manner whether a particular linguistic form can or cannot encode a particular function, a range of specific constructions are included in the analysis. In ?, informants across languages translated sentences in a specific temporal and observational context (e.g., you saw someone writing a letter yesterday). These data were assigned to tense-aspect prototypes, so the input matrix \mathbf{I} now has K sentential contexts $c_j \in \mathcal{C}$, belonging to a smaller number of abstract function prototypes, and takes the following form:

$$\mathbf{I}_{i,j} = \begin{cases} Y & \text{if form } f_i \text{ was used for sentential context } c_j \text{ in language } l, \\ N & \text{otherwise.} \end{cases}$$

The lessened reliance on manually posited functions allows for the prototype structure to emerge from the data. However, the contexts are still manually selected by the typologist, and the semantic information still only comes from co-occurrence with forms in the sample itself—so the study necessarily cannot represent the full complexity of the studied forms across the vast space of possible meanings, nor can it fully capture frequency effects. This type of study has also been applied to other aspectual constructions (?), and to verb-specific semantic roles (?).

Finally, the third major type of MDS analysis relies on a fully bottom-up

approach to function, using parallel corpora rather than reference grammars or elicited survey data. In this type of study, relevant parallel clauses for a particular phenomenon are identified in a parallel corpus, and the input matrix \mathbf{I} is constructed such that each row contains the construction used in that clause in each language studied, producing a K -tuple where K is the number of languages. To compute distance in this type of study, the Hamming distance between the tuple for two clauses is computed, yielding a similarity matrix based on the number of languages that use the same construction in both clauses. This type of analysis removes the manual positing of functions and contexts entirely, proceeding bottom-up, and is thus the most in the spirit of our present inquiry. However, the approach still only captures similarity based on translations in corpora. The fact that contemporary deep learning models capture extremely fine-grained semantic distinctions is not leveraged in this approach, and so the semantic space captured is only as good as the evidence directly given by the co-occurrence of translations, rather than language-internal evidence about meaning.⁹

Overall, the MDS approach allows us to both study the rich gradient structure underlying linguistic function, and decrease the dependence on manually posited functions. However, the state of the art still relies entirely on parallel co-occurrence data. In the next section, I will discuss the small literature that leverages recent advances in deep learning to provide a rich representation of function in typology.

2.3.5 Deep learning models of comparative concepts

Despite the rich body of evidence that deep learning models of language capture fine-grained semantic distinctions, there has been relatively little work leveraging these models to provide empirically grounded comparative concepts

⁹A wide range of domains and phenomena have been studied with this approach:

for typology. Recently, Gregorio et al. (2025) used multilingual BERT (Devlin et al., 2019) and Aya (?) to study animacy cross-linguistically. Specifically, they identify which syntactic roles and clausal positions are most associated with animacy of the referent. However, the role of the models used here is not truly gradient, nor is the function emergent—the models are used to produce a 3-way classification (human, animate, and inanimate) based on an annotated corpus, and the analysis is conducted over these discrete categories. While the rich representations of the models are critical for creating an accurate classifier, the comparative concepts are still discrete and manually posited.

Papadimitriou et al. (2021) study grammatical subjecthood with a less discrete approach. Specifically, they train a multi-layer perceptron classifier on multilingual BERT representations to distinguish between the embeddings of transitive subjects and objects, then examine the classifier’s categorization of intransitive subjects, finding that intransitive subjects are categorized as more subject-like than object-like, and that classifiers transfer across languages, including languages with different morphosyntactic alignment (e.g. ergative-absolutive vs. nominative-accusative languages). However, they found that animate non-subjects and passive subjects were more likely to be classified as subjects and objects respectively, indicating a semantic dimension to this cross-linguistically robust representation of subjecthood.

Another technique for obtaining a gradient representation of a semantic dimension is to use *semantic projection* (Grand et al., 2022), which has been shown to capture human judgements about object features. This technique uses exemplars at extremes of a semantic dimension (e.g. “huge” and “tiny”), using a deep learning model to embed them in a Euclidean space. All possible embedding pairs across the two sets of exemplars are subtracted from each other, and these difference vectors are averaged to produce a single vector representing the semantic dimension. This vector can then be used to project other words

onto this dimension by taking the dot product of their embedding with the dimension vector. Li (2025) uses this technique to study models' representations of animacy. The authors claim that their results show that models represent animals as more animate than humans, in line with psychological findings in humans (). They suggest that this indicates inductive biases in humans that shape grammatical animacy by focusing on certain constructions. However, their operationalization of animacy is questionable, as the exemplars they use to define high animacy are exclusively non-human animals. Nevertheless, the techniques here show how deep learning models can be used to provide a gradient representation of a semantic dimension which can be used to study cross-linguistic patterns in form-function mappings.

Altogether, the results in this nascent literature are promising, but there are still many dimensions of deep learning representations that have not been explored. In this thesis, I will focus on how deep learning models help provide new and better models of lexicality, which has so far not been directly addressed in any of this literature.

2.4 Formal and functional dimensions of lexicality

...we may be quite sure of the analysis of the words in a sentence, and yet not succeed in acquiring that inner “feel” of its structure that enables to tell infallibly what is “material content” and what is “relation”

— Edward Sapir (?)

Some units of language are more meaningful than others. This basic insight is almost as old as the study of language itself. In the Greek tradition, Aristotle distinguished *phōnē*⊗ *sēmantik*⊗ (sign-bearing sounds) from *phōn*⊗ *ásēmos* (non-sign-bearing sounds), such as the class of *árthron* which includes prepositions and preverbs (?). This distinction was not limited to the proto-linguistics of Indo-European languages: in the 12th century the *Wén zé* (文澤) of Chen Kui (陈葵)

catalogued *zhùchí* 帮助 (lit. “helping words”)—corresponding to what we would today call function words. Across the world’s languages, we see asymmetries between elements that express content and those that express grammatical function. It is little wonder then that the distinction between contentful and functional elements continues to have relevance across linguistic theories and domains. Yet boundary cases abound and the nature of the distinction has made it challenging to formalize. In Chapter 1, I sketched the idea of a LEXICALITY SPECTRUM as a general way of conceptualizing the correlation between formal expression and (degree of) semantic content. In this section, I describe the formal and functional dimensions of this spectrum in more detail.

2.4.1 The formal dimension

As I argued in Chapter 1, the lexicality spectrum is operant at multiple levels of formal linguistic structure, with different names being used for similar distinctions at different levels. But the distinctions between “words”, “clitics”, “morphemes”, and “affixes” are all theoretically tenuous at best (Zwicky, 1994; Bruening, 2018). As a theory of these terms themselves is beyond the scope of this thesis, I will here focus on the general formal trends that underlie these distinctions—the formal dimension of the lexicality spectrum.

The basic idea underlying all these terms is this: some linguistic units have “bigger” forms than others. Of course, this is implied by the compositional nature of linguistic structure: a phrase may be composed of several words, a word may be composed of several morphemes, and each morpheme can contain a variable number of phonemes. Yet even comparing morphemes to morphemes, some are formally bigger than others. Here are some ways in which this can manifest:

Boundness One aspect of formal size is the notion of **BOUNDNESS**. While Haspelmath has argued for a sharp cross-linguistic definition of boundness ?? as “unable to occur in isolation”, I share ?’s scepticism of the utility of this as a cross-linguistic criteria and share his feeling that this is better understood as a gradient notion. There are many languages where no morpheme can occur in isolation—surely our comparative concepts should apply to them! Further, the notion of “isolation” is itself problematic, as language always occurs in a discursive context. Here, I sketch boundness as a continuum of cluster properties. In Chapter 3, I operationalize some relevant aspects of this continuum, but here I will simply focus on what the formal trends *are*.

At one end of the spectrum of boundness are free morphemes. In many languages, these morphemes can form whole utterances by themselves in the right context (Consider “Cat.” as a response to “What is your favourite animal?”). In some languages, even the freeest morphemes may not be able to stand alone, requiring some obligatory bound marking (e.g. case or tense marking), but the free morpheme behaves in some way like the “root” of the word. This often takes the form of the free morpheme occurring at the periphery of the word (usually the beginning).¹⁰ Further, they typically occur immediately coincident to that host morpheme. If morphemes occur between a bound morpheme and its root, those morphemes are typically intermediate in terms of these formal properties—that is, the most bound morphemes occur furthest from the root morpheme.

Boundness is a similar notion to **VALENCY**, the notion that certain linguistic units require a certain number of arguments. For example, nouns typically have a valency of zero. Most verbs, on the other hand, have a valency of one or more, requiring a subject and one or more objects. Syntactic valency is different

¹⁰In cases where the free morpheme cannot stand alone, a critical aspect of the argument for the *freeness* of this morpheme is typically semantic. However, I am here focusing exclusively on the *formal* properties of boundness.

from semantically valency. For example the verb *to rain* arguably requires no semantic arguments in English, but it still requires a subject syntactically (*It rains*). Putting things in mathematical terms, we can view such words as functions, which require certain arguments to form a complete expression. Similarly, bound morphemes are often formalized as predicates which take in a root morpheme to produce a combined expression. However, prototypical free but valent morphemes (e.g. verbs) are distinguished from bound morphemes by their degree of syntagmatic integration with their arguments. Either the arguments are themselves free morphemes, able to move around depending on the construction or take their own bound morphemes, or else they are expressed through bound morphemes on the verb itself.

Allomorphy More bound forms are also phonetically more variable. They may be subject to special phonological processes that do not apply to other morphemes in the language (such as the English plural -s being realized variably as [s], [z], or [əz] depending on the phonological context). The more of these processes apply, the more phonologically bound the morpheme is. ? argue that lexically- or morphologically-conditioned variants of morphemes (*allomorphs*) are a formal sign of greater bondedness.

Length Perhaps the most obvious dimension of formal size can be seen in the number of phonemes in a morpheme—which can vary dramatically. The length parameter, at the short end, is intimately tied up with the other dimensions of formal size. Allomorphy may reduce the number of shared phonemes between morpheme variants. Morphemes can also get a length shorter than one phoneme in some instances. *Portmanteau* morphemes share multiple (unrelated) features in a single marker, meaning that the phonological material dedicated to any one of them can be the equivalent of less than a single segment. Tightly bound morphemes can become shorter than a segment by becoming suprasegmental

or process morphemes, e.g. by changing tone or root morpheme vowel quality (*Ablaut*). We can therefore think of all dimensions of formal size outlined here as related to the concept of length.

2.4.2 The functional-semantic dimension

Information and Frequency At the semantic core of the lexicality spectrum is the notion of **CONTENTFULNESS**.¹¹ The basic intuition is that some linguistic items contribute more to the overall meaning of an utterance than others. This is a notoriously difficult notion to pin down, as it relates to the deep question of what *meaning* and *content* are in the first place. **INFORMATION THEORY** both provides a mathematical formalization of this notion and a demonstration of why separating content from form is difficult. Shannon (1948) introduced the notion of *entropy* as a measure of information in bits. The entropy of a random variable X with possible values $\{x_1, x_2, \dots, x_n\}$ and probability mass function $P(X)$ is defined as:

$$H(X) = - \sum_{i=1}^n P(x_i) \log P(x_i). \quad (2.4)$$

The entropy is related to the notion of “information” in an important respect: it provides a lower bound¹² on the number of bits needed to encode the value of X . That is, it tells you the most efficient encoding/compression of X must be at least $H(X)$ bits long. Shannon’s entropy and information-theory more broadly have been widely and successfully applied to linguistic theory (). With respect to “words”/“morphemes”, the information-theoretic perspective implies that the information content of a morpheme is its negative log-probability.¹³ Under this perspective, more frequent morphemes carry less information. Combining this with the idea from Section 2.4.1 that the formal dimension is all related to length

¹¹The terms “semantic weight” and “semantic force” are also common

¹²Technically, an average lower bound over many samples.

¹³This follows from the assumption that morphemes are generated independently, which is not true, however, later work has extended the information-theoretic insight to account for dependencies, with a similar overall conclusion (Piantadosi et al., 2011).

or formal size, we can see the lexicality spectrum as a generalization of ZIPF'S LAW OF ABBREVIATION, one of the most famous and foundational discoveries in all of computational linguistics (?), which states that more frequent words are much shorter than less frequent ones. The information-theoretic corollary of this is that more informative morphemes are longer than less informative ones. Frequency certainly plays a central role in the structure of language and the lexicality spectrum. It has been argued to be a driving force in grammaticalization (Bybee, 1985), and the so-called ICONICITY OF COMPLEXITY (itself closely related to the notion of a lexicality spectrum) has been argued to be an effect of frequency pressures on efficient communication (?). Nevertheless, debates around the nature of the lexical–functional distinction, inflection–derivation distinction, and related phenomena continue unabated, indicating that frequency and thus standard information-theoretic notions of content are insufficient as a full account of the relationship between meaning and function.

Abstractness and Polysemy Another common way of characterizing the functional dimension of the lexicality spectrum is in terms of ABSTRACTNESS. Functional elements are often described as being more abstract than lexical elements; on the other hand, functional elements are also often more POLYSEMOUS than lexical elements (Haspelmath, 2003). The difference between abstractness and polysemy is not always clear. In Figure 2.1, it is fairly clear that *to* has multiple senses, such as the purpose sense in *I went to the store to buy milk* and the recipient sense in *I gave the book to Mary*. Haspelmath (2003) points out that data is often ambiguous between a *monosemic* position where there is a single vague, abstract meaning that interacts with contexts to serve different functions, and a *polysemic* or even *homonymic* interpretation where there are multiple more specific meanings which share a surface form for motivated or ideo-syntactic reasons. Under such an interpretation, the meaning of functional elements is

not vague at all, and perhaps not very abstract. Nevertheless, words like *in* which have spawned whole literatures in linguistics and psychology attempting to characterize their use, and which seem to cover a continuous space of meaning rather than discrete scenarios, are more straightforwardly characterized as abstract than polysemous.¹⁴

Another problem for abstractness of meaning is that lexical elements can also be highly abstract (e.g. *idea, angry*) and some traditionally functional elements can be thought of as being fairly concrete (e.g. plural markers). This problem is sometimes waved away by invoking prototypicality, but such an account leaves some serious questions unanswered (?Croft, 2002, p. 225). ? notes that *concrete* nouns which are frequent do get shorter forms (e.g. *dog, car*), indicating that length is more a function of predictability than abstractness. Nevertheless, while these forms are shorter, they do share many formal properties with prototypical concrete lexical items, such as their lack of boundness. So there is still something *different* about these forms compared to functional items, even if abstractness is not the right way to capture it.

Relationality Perhaps one way out of this conundrum is to focus on semantic RELATIONALITY rather than abstractness *per se*. In the simplest terms, a relational meaning is one that inherently implies the existence of at least one other entity (?., p. 67). For example, the concept ROUND is relational, as roundness can only be defined with respect to some entity. On the other hand, CIRCLE is non-relational, despite referring to the same properties. In my view, a key property of relationality of meaning is that when composed with entities, the resulting meaning is less abstract than the relational meaning alone. For example *round* is more abstract than *a round rock*. This can help explain some of the more “concrete” grammatical functions: while plural *forms* are not very abstract (e.g.

¹⁴I refer here to the spatial sense(s) of *in*.

cats is likely to be very concrete), the plurality is itself highly abstract.

There are several barriers to associating relationality with the functional dimension of the lexicality spectrum. A first objection is that because verbs and adjectives are both traditionally considered lexical and relational, relationality cannot be the defining property of functional items. In Part II of this thesis, I will argue that relationality is a key dimension for shaping the lexicality spectrum, and that this *includes* adjectives and verbs as in some sense closer to functional elements than nouns. I believe the definition of relationality I have given here helps explain this: prototypical *lexical* relational concepts such as GIVING, ROUND, or RED are, I believe, more concrete and more full of intension than functional relational concepts such as PLURALITY or ANIMACY. A key correlate of this objection is confusing relationality with semantic valency. The number of entities required to specify a meaning (or the number of syntactic arguments) is *not* the sense of relationality I am referring to here. Valency manifests iconically in syntax, but relationality as I define it is more closely related to notions like boundness. For example, RED is less relational than PLURAL, despite the fact that both are monovalent, because RED is more conceptualizable on its own.

A second problem for this view is that meaning itself is relational. This view goes under the CONCEPTUAL ROLE THEORY of meaning (Block, 1998) in philosophy of mind, and also undergirds structuralist and distributionalist views of meaning in linguistics (de Saussure, 1916; ?). Piantadosi and Hill (2022) provide a useful example, pointing out that despite a clear prototypical concrete referent, the concept of POSTAGE STAMP is fundamentally relational, and we can easily imagine *virtual* postage stamps so long as the abstract referent fulfills the role of tracking payment for delivery. Plausibly, this tension could be resolved by countering that the conceptualization of POSTAGE STAMP is not as relational as the inferential or functional extensions it affords—that is, we imagine a concrete, bounded, non-relational object, which has been selected

by the complex networks of meaning in our minds to fulfill a relational role. Nevertheless, I think this is a serious challenge to the entity–relation distinction which should be further investigated, but stands outside the scope of this thesis.

Lastly, a key objection to both relationality and abstractness is that they are hard to specify and often subjectively defined. On this point I agree. For example, the sense that **RED** is somehow less relational than **ROUND**, or **PLURAL** is less relational than **GIVING**, despite the higher valency of the latter, is intuitive, but difficult to give precise criteria for. This is why a key goal of this thesis is to provide empirical and computational tools for investigating these notions rigorously. While the above discussion is theoretical and subjective, and one can dispute the abstractness I assign to various concepts, the empirical facts in the rest of the thesis remain. The argument here should be seen as a motivation and interpretive lens for the empirical work that follows.

2.4.3 Summary

In this section, I have attempted to sketch the separate formal and functional dimensions of the lexicality spectrum. On the formal side, I have argued that boundness, allomorphy, and length are all correlated dimensions of formal size. On the functional side, I have argued for the importance of relationality alongside information content, and discussed how the former relates to boundness iconically, and the latter relates to formal size via economy principles. Together, this provides a high-level overview of the lexicality spectrum, from nouns at one extreme to inflectional affixes at the other. In the rest of this thesis, I will explore specific aspects of this spectrum and distinctions drawn along it in more detail. Therefore properties specific to inflection and derivation, for example, will be discussed in the relevant chapters.

2.5 Chapter Summary

This chapter has provided a theoretical background for the thesis. In Section 2.1, I reviewed the problem of comparative concepts in linguistic typology, discussing approaches like semantic maps and retro-definitions, comparing and contrasting them with my approach of *grounding* problematic distinctions in empirical measures. My approach allows for understanding how consistent existing or proposed operationalizations of comparative concepts are in terms of empirical dimensions. In ??, I provided an overview of a decade of advancements in deep learning models of language, focusing on how these models capture fine-grained and potentially universal semantic distinctions, and how large-scale pretraining is useful for the acquisition of rich linguistic structure. I argued that these models provide a promising avenue for separating meaning from frequency in typology. In ??, I connected the computational literature in typology to the notion of comparative concepts, which have largely been implicit in this literature. From phonological typology and semantic category systems, I argued that computational models have been able to advance typological theory and understanding as technologies develop that allow the empirical grounding of the underlying space, further motivating my approach.

Finally, in Section 2.4, I provided a high-level theoretical overview of the lexicality spectrum, separating the formal and functional dimensions, and arguing for the role of informativity and relationality in shaping the formal dimension. In the rest of this thesis, I will use deep learning models to operationalize and investigate specific lexicality-related phenomena, showing consistent patterns across distinctions that have been difficult to formalize in the past.

Part I

Inflection and Derivation

Chapter 3

Corpus-based Measures for Inflection and Derivation

3.1 Introduction

In the field of morphology, a distinction is commonly drawn between inflection and derivation. This distinction is intended to capture the notion that sometimes morphological processes form a “new” word (derivation), whereas other morphological processes merely create a “form” thereof (inflection) (Booij, 2007). While the theoretical underpinnings and nature of this distinction are a subject of significant and ongoing debate, it is nevertheless employed throughout theoretical linguistics (Perlmutter, 1988; Anderson, 1982), computational and corpus linguistics (ten Hacken, 1994; McCarthy et al., 2020; Wiemerslage et al., 2021), and even psycholinguistics (Laudanna et al., 1992; MacKay, 1978; Cutler, 1981).

To a large degree, dictionaries and grammars roughly agree on which morphological relationships are inflectional and which are derivational within a given language. There is even a degree of cross-linguistic consistency in the constructions which are typically/traditionally considered inflections—e.g. tense marking on verbs is considered to be inflectional across a wide range of lan-

guages (Haspelmath, 2024; Bybee, 1985, pp. 21–22). This cross-linguistic consistency is highlighted by the development of resources such as UniMorph (Batsuren et al., 2022), a multilingual resource which annotates inflectional constructions across over a hundred languages using a unified feature scheme and, more recently, also includes derivational constructions from 30 languages. UniMorph data is extracted from the Wiktionary open online dictionary,¹ which organizes constructions into inflections and derivations based on typical descriptive grammars for a given language, rather than any particular linguistic theory. The inflection–derivation distinction in UniMorph is therefore determined by what Haspelmath terms *traditional comparative concepts* (Haspelmath, 2024), which are informed by the traditional structure of Western dictionaries and grammar books. The success of this initiative indicates a high degree of cross-linguistic overlap in what morphosyntactic features are considered inflectional.

Despite this relative consistency at the level of annotation, there is considerable disagreement among linguists about the fundamental properties that might underlie or explain these traditional categorizations—such as the degree of syntactic or semantic change, or the creation of new words. As an example, Plank (1994) covers no fewer than 28 tests for inflectional and derivational status. Upon applying them to just six English morphological constructions, Plank (1994) finds considerable contradictions between the results based on different criteria. Such difficulties in producing a cross-linguistically consistent definition have led many researchers to conclude that the inflection–derivation distinction is gradient rather than categorical (Bybee, 1985; Spencer, 2013; Copot et al., 2022; Dressler, 1989; Štekauer, 2015; Corbett, 2010; Bauer, 2004) or to take the even stronger position that the distinction carries no theoretical weight at all (Haspelmath, 2024).

¹<https://www.wiktionary.org/>

One major issue in evaluating these theoretical claims is the lack of large-scale, cross-linguistic evidence based on quantitative measures (rather than subjective tests). Work in theoretical linguistics has established that the intuitions underlying subjective tests can be problematic in certain cases (Haspelmath, 2024; Plank, 1994). Even so, it is possible that measures based on these subjective tests could indeed be used to classify the vast majority of morphological relationships across languages in a way that is consistent with traditional distinctions. If so, a large-scale empirical study could also provide evidence regarding the gradient versus categorical nature of the inflection–derivation distinction.

Several previous studies have shared our goal of operationalizing linguistic intuitions about the inflection–derivation distinction and applying them on a large scale, but these studies have been limited in terms of both the sample size and diversity of the languages studied and the comprehensiveness and generality of the measures used. In particular, Bonami and Paperno (2018) and Copot et al. (2022) explored semantic and frequency-based measures of *variability* in French, aiming to test the claim that derivation tends to introduce more *idiosyncratic* (variable) changes than inflection. Meanwhile, Rosa and Žabokrtský (2019) looked at the *magnitude* of orthographic and semantic change between morphologically related forms in Czech, following the claim that derivation tends to introduce *larger* changes than inflection. All of these studies found differences *on average* between (traditionally defined) inflectional and derivational constructions but also considerable overlap. That is, results so far are consistent with the view that although quantitative measures do align to some extent with the two traditional categories, the distinction between inflection and derivation is at best gradient. Moreover, these studies provide little evidence that quantitative measures would be sufficient to determine the inflectional versus derivational status of a new construction with any accuracy. However, it is possible that the picture could change when a wider variety of languages is

included, especially if we also consider a larger number of measures at once.

In this chapter, we take inspiration from both linguistic theory and the studies above to develop a set of four quantitative measures of morphological constructions, which capture *both* the magnitude and the variability of the changes introduced by each construction. Crucially, our measures can be computed directly from a linguistic corpus, allowing us to consistently operationalize them across many languages and morphological constructions. That is, given a particular morphological construction (such as “the nominative plural in German”) and examples of word pairs that illustrate that construction (e.g. “*Frau, Frauen*”, “*Kind, Kinder*”), we compute four corpus-based measures—two based on orthographic form and two based on distributional characteristics—which quantify the idea that derivations produce *larger* and *more variable* changes to words compared to inflections (Spencer, 2013; Plank, 1994).

We show that while inflection and derivation are significantly different on *average* for all four measures, there is considerable overlap between the two categories, but that distributional characteristics show larger differences between the categories than formal characteristics. The utility of the distributional measures is shown to be unrelated to frequency differences between inflected and derived forms. We show that the distributional embeddings capture some limited syntactic category information in addition to their noted ability to capture semantic similarity. In line with prior studies, we find substantial overlap between inflectional and derivational constructions on all measures—suggesting that indeed, cross-linguistically, the inflection–derivation distinction is not well explained by any single one of these dimensions of variations. This chapter sets the groundwork for Chapter 4, where we explore the ability of combinations of these measures to predict inflectional vs. derivational status when combined.

3.2 Motivation for our measures

In order to explore our question of interest, we need to operationalize some of the linguistic properties that have been argued to differentiate inflection from derivation. This section briefly reviews some of those properties and explains, at a high level, how they relate to corpus-based measures. We defer the detailed definitions of these measures to Section 3.3.

We take inspiration from the framing of Spencer (2013), who argues that morphological processes are characterized by changes to one or more of the four components of a wordform: 1. its *form* (the string of phonemes which make up its pronunciation), 2. its *semantics* 3. its *syntax* (e.g. part of speech and argument structure), and 4. its “*lexical index*”, a number corresponding to the abstract “word” to which the wordform belongs. Within this framework, a traditional view of the inflection–derivation distinction would be that inflections are those morphological relations between entries that differ in a number of aspects but have the *same* lexical index; whereas derivation corresponds to regular transformations that produce words with a *different* lexical index. Spencer argues instead for a taxonomy of morphological processes that focuses not just on lexical index, but on changes to any of these four components. Within this taxonomy, canonical inflections tend to produce small changes to one or a few components, whereas canonical derivations make large changes to more components. Indeed, in Spencer’s view, some cases classically considered derivational, such as transpositions, do not change the lexical index. Furthermore, words may be related by an inflectional process, yet (through semantic drift) have distinct lexical indices (e.g. *khaki*, a colour, and *khakis*, a type of pants). While this may seem counter-intuitive under traditional views of inflection and derivation, it is important to note that the concept of lexical index goes beyond the inflection-derivation distinction, but rather aims also to capture

empirical effects observed within psycholinguistics, such as priming effects in lexical decision tasks. While it has been argued that these effects align with the inflection-derivation distinction (Laudanna et al., 1992; Kirkici and Clahsen, 2013), this represents an independent basis for notions of words being the “same” or “different”.

While Spencer de-emphasizes the classical distinction between inflection and derivation, we treat his taxonomy of morphological processes as a continuous extension of the inflection and derivation distinction. Doing so naturally unifies many existing diagnostics. It both captures and generalizes correlations like derivations causing larger changes in the semantics or changing part of speech, and also suggests less frequently discussed correlations, such as derivational relations typically involving larger changes to the form of a word.² The notion of lexical index, while not directly observable, captures the notion of being the “same” or “different” word.

Importantly, it is (at least theoretically) possible to characterize a great deal of information about each of these aspects from text corpora alone. For languages with alphabetic writing systems, such as those we consider here, form is largely encoded in the orthography. Syntactic part of speech can be determined with high accuracy by the context in which words appear (He et al., Tenth century–2018). Finally, the distributional semantic hypothesis (Harris, 1954) holds that semantically similar words appear in similar types of contexts; this hypothesis is supported by the empirically impressive correlation of similarities in word embedding models like FastText (Bojanowski et al., 2017) with human semantic similarity judgements. However, these vectors also capture substantial amounts of information about a word’s syntactic category, as operationalized by its part of speech (Pimentel et al., 2020; Lin et al., 2015). Because of the distributional nature of meaning, it is in fact difficult to induce a space from pure language data where

²This is suggested, though not explicitly, by criteria like Plank (1994)’s “derivational morphemes resemble free morphs.”

distance corresponds to *syntactic* similarity entirely independently from *semantic* similarity. While there is prior work on inducing such representational spaces (e.g. He et al., Tenth century–2018; Ravfogel et al., 2020), due to our complex and highly multilingual setting, we instead choose to *collapse* the distinction of syntactic and semantic change made by Spencer, focusing on what is captured by embeddings designed primarily for capturing semantics but which also capture syntactic information. In particular, we use FastText embeddings, described in more detail in Section 3.3.2.

In addition to considering the size of the changes made to these aspects of words by a construction, we also consider the *variability* of these changes. Words with different lexical indices are thought to have processes like semantic drift apply separately from each other (Spencer, 2013; Copot et al., 2022; Bonami and Paperno, 2018), which Copot et al. (2022) carefully links to variability in semantics. We also consider variability in the changes made to the form. This aspect has been under-explored in prior computational work. Following Plank’s (1994) claim that formal variability is greater for derivations than inflections, we would expect that allomorphy is greater for derivations than inflections, perhaps relating to the idiosyncrasies in the application of derivational allomorphs, as well as the semantic inconsistencies of derivation. **On the other hand, the discussion of the formal dimensions of the lexicality spectrum and boundedness in Section 2.4.1 suggests that allomorphy is associated with boundness, integration and shorter “length”, thereby being associated with inflection rather than derivation. This apparent contradiction highlights the complexity of the inflection-derivation distinction, and motivates our empirical approach.**

Another thread of research inspiring this particular factorization comes from the field of natural language processing. There, the interplay between formal and distributional aspects within morphology has been widely investigated, both in derivational morphology (Cotterell and Schütze, 2018; Deutsch et al.,

2018; Hofmann et al., 2020), and in unsupervised morphological segmentation, which typically covers both inflection and derivation (Schone and Jurafsky, 2000; Soricut and Och, 2015; Narasimhan et al., 2015; Bergmanis and Goldwater, 2017).

Base	Constructed	Morph.	Start POS	End POS	Lang.
Frau	Frauen	NOM;PL	N	N	DEU
Auge	Augen	NOM;PL	N	N	DEU
Lehrerin	Lehrerinnen	NOM;PL	N	N	DEU
Kind	Kinder	NOM;PL	N	N	DEU
...

Base	Constructed	Morph.	Start POS	End POS	Lang.
protrude	protrusion	-ion	V	N	ENG
defenestrate	defenestration	-ion	V	N	ENG
redecorate	redecoration	-ion	V	N	ENG
elide	elision	-ion	V	N	ENG
...

Table 3.1: Sample of an inflectional construction (upper table, German nominative plural) and derivational construction (lower table, English verbal nominalization with *-ion*) in our data

Because debates about inflectional and derivational status typically focus on *constructions* such as “the nominal plural in German” or “the addition of the *-ion* nominalization morpheme to verbs in English,” this is the level at which we perform our analysis. Examples of constructions from our dataset are shown in Table 3.1. We define a construction here as a unique combination of a morpheme (given in a canonical form like *-ion* for derivation or as morphosyntactic

features for inflection), initial part-of-speech, constructed part-of-speech, and language. That is, we do not group morphemes across languages, nor do we group derivations with identical canonical forms which apply to or produce different parts of speech. This decision is motivated by examples like agentive *-er* vs. comparative *-er* in English, which differ only in the parts of speech which they apply to and produce. While there is some asymmetry in the way this grouping is handled between inflection and derivation, we do not believe this substantially affects our results. For further discussion, see Section 4.3.1.

Choosing to analyse constructions, rather than individual pairs of words, also has the advantage that any unusual behaviour of individual pairs will tend to get smoothed out as we are looking at a large number of pairs for each construction (see Section 3.4 for details). While individual word pairs within a construction may have quite variable distributional properties, the *general tendencies* of that construction may paint a picture that is more clearly in line with notions of inflection and derivation.

Given that we are working at the level of constructions, the four quantities we wish to measure for each construction are:

- M_{Form} and V_{Form} : the average magnitude of the change in form induced by a construction, and the variability of that change.
- M_{Embed} and V_{Embed} : the average magnitude of the change in semantic/syntactic embedding space induced by a construction, and the variability of that change.

The following section describes how these measures are computed for each construction.

3.3 Method

In this section, we define M_{Form} , V_{Form} , M_{Embed} , and V_{Embed} for constructions with N pairs of words (b_i, c_i) , where b_i is the base word, and c_i the constructed word which results from applying the morphological construction.

3.3.1 Orthography-based measures

In this study, we use orthography as a proxy for phonological form, as discussed in Section 3.2. For each construction, we measure the *magnitude* of the change in form M_{Form} using the Levenshtein edit distance (Levenshtein, 1966): we simply compute the average distance between each pair of words in the construction (assuming all edits count equally). For a construction with N word pairs (b_i, c_i) , this metric is given as follows:

$$M_{\text{Form}} = \frac{1}{N} \sum_{i=1}^N \text{EDITDISTANCE}(b_i, c_i). \quad (3.1)$$

To measure the *variability* of the change in form V_{Form} (a measure of the construction’s degree of allomorphy), we start by constructing an *edit template* for each word pair, which describes the changes made to the base in a way that abstracts away from specific string positions. For example, the pair (*tanzen*, *getanzt*) yields the edit template *ge_XXt*, meaning “start by writing *ge*, copy from the base form, delete the last two characters, and append *t*.“ Similarly, the edit template for the pair (*Sohn*, *Söhne*) produces the edit template *_Xö_e*. This example highlights two important design decisions for these edit templates. First, we abstract out any variation in length of the spans which are shared with the input. This is based on the assumption that these reflect variation in the base form itself rather than morphological allomorphy. In our dataset, which does not contain any languages with templatic morphology, this assumption works well; however, future studies wishing to extend to such languages should revisit this assumption. Secondly, because we operate over orthographic form rather

than the true form phonetics/featural information, edits which are considered “the same” in linguistic theory may sometimes be considered different and vice-versa. Here, a linguist might describe this plural allomorph as adding +FRONT to the vowel’s features, which would cover the templates `_Xö_e`, `_Xä_e`, and `_Xü_e`. However, addressing this issue is outside the scope of this study.

Having so defined a description of the change in form with a sensible equality metric (i.e., not reliant on the length of the base), it remains to measure how much this change *varies* within a given construction. We take the edit template for each word-pair in a construction and compute its edit distance with each of the other edit templates in the construction, reporting the frequency-weighted pairwise edit distance as our measure of variability. That is, if an edit template T_i appears at a rate F_{T_i} , and there are M edit templates for a construction, this metric is computed as

$$V_{\text{Form}} = \sum_{i=1}^M \sum_{j=1}^M F_{T_i} \cdot F_{T_j} \cdot \text{EDITDISTANCE}(T_i, T_j). \quad (3.2)$$

For example, suppose we have a morpheme with two edit templates: `_as`, used 80% of the time, and `_os`, used 20% of the time. Then this measure would be $0.8 \cdot 0.2 \cdot \text{EDITDISTANCE}(\text{_as}, \text{_os}) + 0.2 \cdot 0.8 \cdot \text{EDITDISTANCE}(\text{_os}, \text{_as}) = 0.32$. This measure goes beyond simply counting allomorphic variants by weighting them both in terms of how different they are from each other, and by how widely they are applied in the lexicon.

3.3.2 Distributional-embedding-based measures

To approximate the semantic and syntactic properties of the words in our study, we use type-based (non-contextual) distributional word embeddings. Specifically, we use the FastText vectors for each language released by Bojanowski et al. (2017);³ these were trained on Common Crawl⁴ and Wikipedia data, which was

³<https://fasttext.cc/docs/en/crawl-vectors.html>

⁴<https://commoncrawl.org/>

automatically tagged by language to train language-specific embedding models (Grave et al., 2018). These FastText vectors are known to correlate well with human semantic similarity scores (Vulić et al., 2020a; Bojanowski et al., 2017), and are more commonly used as models of semantics than syntax.⁵ However, there is evidence from the literature in unsupervised part-of-speech tagging (He et al., Tenth century–2018; Lin et al., 2015) and probing (Pimentel et al., 2020; Babazhanova et al., Tenth century–2021) that they also encode syntactic information.⁶

One complicating aspect of our use of FastText vectors is that they include distributional information not only at the word, but the sub-word level. The nature of this information is itself purely distributional, relating not to the characters within those subwords, but rather the context in which the subwords appear. Nevertheless, it means that the distance between words in this distributional embedding space can be influenced by how similar they are in terms of form, when they share subwords. The primary goal of our study is identifying whether there are signals present in a raw text corpus which can reliably distinguish between inflection and derivation. As such, while the inclusion of FastText embeddings is *motivated* by their ability to represent semantic and syntactic similarity, that they include some formal information is not an issue to this primary question. It does somewhat complicate the question of assigning relative importance to formal vs distributional features, an issue we return to in Section 4.3.1.

⁵Recent studies have shown that embeddings from newer transformer language models such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) correlate even better than FastText embeddings with human judgements of semantic similarity (Bommasani et al., 2020; Vulić et al., 2020a). However, these context-dependent token-level embeddings would require further processing to produce the type-level similarities needed for our study, and we know of no strategy to do so that is validated to work with the type of resources available for our data. For example, the methods explored by Bommasani et al. (2020); Vulić et al. (2020a) are either shown to work well only for monolingual context models (which are not available for all of our languages), or only for English and multilingual models.

⁶Indeed, our own results suggests that these vectors encode some syntactic information, and that the addition of gold-standard syntactic category information provides little benefit over our proposed model. For further information, please see Sections 3.6 and 4.3.6.

In principle, this issue of interpretability could be avoided by using alternative embeddings that do not include sub-word distributional information, such as Word2Vec (Mikolov et al., 2013b) or GloVe (Pennington et al., 2014). However, FastText has several benefits over these alternatives that we feel outweigh this issue. First, FastText models produce more accurate semantic representations of rare words (Bojanowski et al., 2017), which is important since many morphological variants are rare. In addition, publicly available pre-trained FastText embeddings are available for a much wider range of languages than Word2Vec or GloVe embeddings. Using these pre-trained embeddings makes our study easier to replicate and less computationally intensive, since pre-trained Word2Vec and GloVe vectors are not available for all the languages we include. It also makes our work easier to extend to other languages when relevant morphological resources become available.

Even though FastText is capable of producing vectors for words not seen at training time, we find that including these words biases low-frequency constructions to have artificially large average distances in semantic space, so we exclude all word pairs where the constructed form does not explicitly appear in the vocabulary of the FastText model. This serves as an implicit cut-off for very low-frequency forms, without requiring explicit frequency information for all of our languages.

Given the FastText embeddings, we measure changes in syntax/semantics for a construction as distances in the embedding space between the word pairs in that construction. Specifically, for each (base form, constructed form) pair (b_i, c_i) , we find the Euclidean distance between their embeddings $(E(b_i), E(c_i))$ and we compute M_{Embed} as the average Euclidean distance across all N pairs in the construction:

$$M_{\text{Embed}} = \frac{1}{N} \sum_{i=1}^N \|E(c_i) - E(b_i)\|. \quad (3.3)$$

While cosine distance is more frequently used than Euclidean distance for se-

mantic similarity, this is typically because the vector norm is perceived as less relevant for semantic similarity, in part because it encodes some frequency information, at least for Word2Vec (Schakel and Wilson, 2015). However, frequency information may be useful in our case, since (as noted by Copot et al. 2022) the frequency of a word is correlated with the frequency of other morphological variants of that word, and more so when these variants have similar semantics. Perhaps as a result, we find this metric works as well as or better than cosine distance empirically.

To measure the variability of syntactic/semantic changes within a construction, for each word pair (b_i, c_i) in the construction, we first compute the difference vector \mathbf{d}_i between the embeddings, i.e., $\mathbf{d}_i = E(b_i) - E(c_i)$. For a construction with N pairs and K dimensional embeddings, this yields a $K \times N$ matrix of differences $\mathbf{D} = [\mathbf{d}_1 \dots \mathbf{d}_N]$. We then make the simplifying assumption that the covariance between the dimensions of \mathbf{D} is zero, which allows us to estimate the variance of \mathbf{D} (and thereby V_{Embed}) as the sum of the variances of the individual dimensions k :

$$V_{\text{Embed}} = \sum_{k=1}^K \text{Var}(\mathbf{D}_{k,*}), \quad (3.4)$$

where $\mathbf{D}_{k,*}$ is the k -th row of \mathbf{D} .

While assuming zero covariances is not necessarily realistic (we do observe covariances which are non-zero), accurately estimating the full covariance matrix and/or its determinant requires at least as many data points as the number of dimensions in the matrix (Hu et al., 2017). As the number of dimensions in the FastText embeddings is 300, fulfilling such a criterion would severely limit which constructions and even languages we would be able to study here. Further, as described in Sections 3.5 and 4.1, we observe a strong empirical correlation between our measure of semantic/syntactic variability and inflectional/derivational status in UniMorph, and find this feature highly useful in

creating classifiers of inflection and derivation, suggesting that this simplifying assumption does not prevent the measure from capturing relevant aspects of variability in the embedding space.

3.4 Data

To perform our analysis, we require a multilingual resource that labels pairs of words with the inflectional or derivational construction that relates them. While there are many resources that provide such construction-level information for inflectional morphology (e.g. Hathout et al., 2014; Ljubešić et al., 2016; Beniamine et al., 2020; Oliver et al., 2022), most high-quality derivational morphology resources (e.g. Kyjánek et al., 2020) only indicate which pairs of words are related, but not what construction relates them. An exception is the recently released UniMorph 4.0 resource, which we use in our study because it includes annotation of inflectional constructions for 182 languages as well as annotation of derivational constructions for 30 of those languages.

The data and annotations in UniMorph 4.0 are semi-automatically extracted from Wiktionary,⁷ a collection of online community-built dictionaries available for multiple languages. Inflectional and derivational information are extracted as follows:

- To identify and label inflectional constructions covering most cases, tables with the HTML `class` property `inflection-table` are extracted; some additional manual parsing is used to extract relations which are not tabular in some languages (e.g. English noun plurals). These tables are categorised based on their structure, and one table from each category is hand-annotated with the UniMorph feature set for inflectional features.

Inflectionally related pairs, and the construction to which they belong, are

⁷<https://en.wiktionary.org/>

then obtained from the base word associated with the entry, the particular contents of a cell, and the inflectional feature set with which that cell was annotated (McCarthy et al., 2020).

- To identify and label derivational constructions, the set of candidate derivations to consider for each base form A is found by looking at the *Derived terms* section of A’s Wiktionary entry. The page for each derived term typically contains an etymology of the form A + -B, where -B is a derivational morpheme. In such cases, this information is added to UniMorph, together with the parts of speech of the base form and the derived term (Batsuren et al., 2022, 2021).

Due to the semi-automatic annotation in UniMorph 4.0, and the community-led construction of the source data in Wiktionary, there could be some errors or even systematic issues with the data. In particular, low-frequency forms in the inflectional data are better represented than low-frequency forms in the derivational data, because inflectional forms are constructed using paradigm tables which include all inflections of a given wordform, whereas derivational forms are added on an individual basis. However, since we necessarily exclude low-frequency forms due to the nature of our measures, this concern is somewhat mitigated. We also check for possible frequency confounds in Section 3.5.1.⁸

Another potential systematic issue is that the annotation may fail to collapse derivational allomorphs into a single construction. We comment further on this possible issue in Section 4.3.1, while noting here that our priority is to

⁸We note that data sparsity is a problem for derivational resources in general, not just UniMorph 4.0. For example, in Batsuren et al. (2021)’s evaluation of MorphyNet, the resource on which the derivational data in UniMorph 4.0 builds, the authors find the resource tends to have low recall and high precision when evaluated against derivational networks like Démonette (Hathout and Namer, 2016), despite having comparable numbers of morphological relations. However, manual evaluation revealed that these false positives in an overwhelming majority of cases represent real morphological relationships, indicating sparsity affects both MorphyNet/UniMorph and other derivational resources. Our own manual and against-derivational-network analysis of the extended UniMorph 4.0 data showed similar trends.

include as many languages and constructions as possible so that our sample will represent a wider range of linguistic typologies—UniMorph 4.0 contains languages with a range of morphological typologies, uncommon inflectional features, and different ratios of inflections and derivations; as well as variation in other typological variables such as syllable structure, phoneme inventory, and syntactic variables, which could affect our measures of formal or distributional change.

3.4.1 Data selection and summary

Of the 30 languages for which UniMorph 4.0 provides both inflectional and derivational constructions, some are not suitable for our current purposes. We exclude Galician because at time of writing its UniMorph derivation data is not publicly available; Serbo-Croatian because the UniMorph data is in Latin script while the vast majority of Serbo-Croatian text used in the construction of the FastText vectors is written in Cyrillic; and Nynorsk because FastText does not distinguish between Nynorsk and Bokmål, and Bokmål is the large majority of written Norwegian.

As mentioned in Section 3.3.2, we exclude all word pairs where the constructed form does not explicitly appear in the vocabulary of the FastText model, due to low-quality estimates of semantic similarity for these vectors. We also exclude constructions which have fewer than 50 forms remaining after pre-processing, to ensure robust estimates of the quantities of interest. Finally, we exclude constructions where <1% of the transformed word forms are different from the base word forms, because UniMorph data is non-contextual and we would need context to distinguish the base and transformed forms. On the other hand, we ignore the problem of across-construction syncretism (where the transformed forms are identical but express different morphosyntactic/semantic features) in the present work.

Language family	Language	Morph. typology	# inf.	# der.	Tot. wordpairs
Indo-European (IE)	Armenian	Agglutinative	67	7	41,053
IE: Romance	Catalan	Fusional	52	31	52,329
	French	Fusional	45	104	110,643
	Italian	Fusional	50	79	127,251
	Latin	Fusional	65	23	52,175
	Portuguese	Fusional	69	35	122,622
	Romanian	Fusional	43	28	41,442
	Spanish	Fusional	121	88	337,923
IE: Germanic	Danish	Fusional	23	12	18,343
	German	Fusional	53	68	298,068
	Dutch	Fusional	21	19	36,077
	English	Fusional	7	225	119,543
	Bokmål	Fusional	14	12	50,847
	Swedish	Fusional	40	28	76,226
IE: Slavic	Czech	Fusional	96	76	103,325
	Polish	Fusional	92	104	164,837
	Russian	Fusional	94	46	292,479
	Ukrainian	Fusional	25	13	17,680
IE: Baltic	Latvian	Fusional	66	23	64,571
IE: Celtic	Irish	Fusional	21	10	21,894
IE: Hellenic	Greek	Fusional	84	3	105,358
Uralic	Finnish	Agglutinative	116	65	328,869
	Hungarian	Agglutinative	143	65	272,760
Mongolic	Mongolian	Agglutinative	16	4	15,840
Turkic	Turkish	Agglutinative	164	9	75,873
	Kazakh	Agglutinative	0	8	643
Total			1587	1185	2,948,671

Table 3.2: Descriptive statistics of our filtered dataset by language.

After performing the filtering steps above, we exclude Scottish Gaelic from our analysis, due to a lack of constructions that meet the inclusion criteria. This leaves us with 2,772 constructions from 26 languages: 1,587 (57.3%) of these are considered inflectional by UniMorph, and 1,185 (42.7%) are considered derivational. Table 3.2 contains descriptive statistics about the representation of languages, morphological typologies, and language families within our filtered dataset. Indo-European languages and, accordingly, languages with fusional typology are heavily represented in our data; however, we also have data from five languages which are not Indo-European, representing four major language families; and six languages with an agglutinative typology. We acknowledge that many language families with distinctive morphological typologies, such as the Niger-Congo languages, the Inuit-Yupik languages, and the Semitic languages, are not represented in the present study. Nevertheless, even results on a broad range of Indo-European languages plus a few others is a substantial advance in the typological coverage of existing work in the area.

3.5 Distribution of the individual measures

In this section, we compare the distributions of our individual measures of constructions labelled as inflections to those of constructions labelled as derivations in UniMorph.

The distributions of the four measures for inflectional and derivational constructions in our data are shown in Figure 3.1. For all measures considered, thanks to the large amount of data in the study there is a significant difference between the mean values for inflectional and derivational constructions ($p < 0.001$ under the Mann-Whitney U test). However, we are more concerned with the direction and magnitude of those differences, which vary across the four measures.

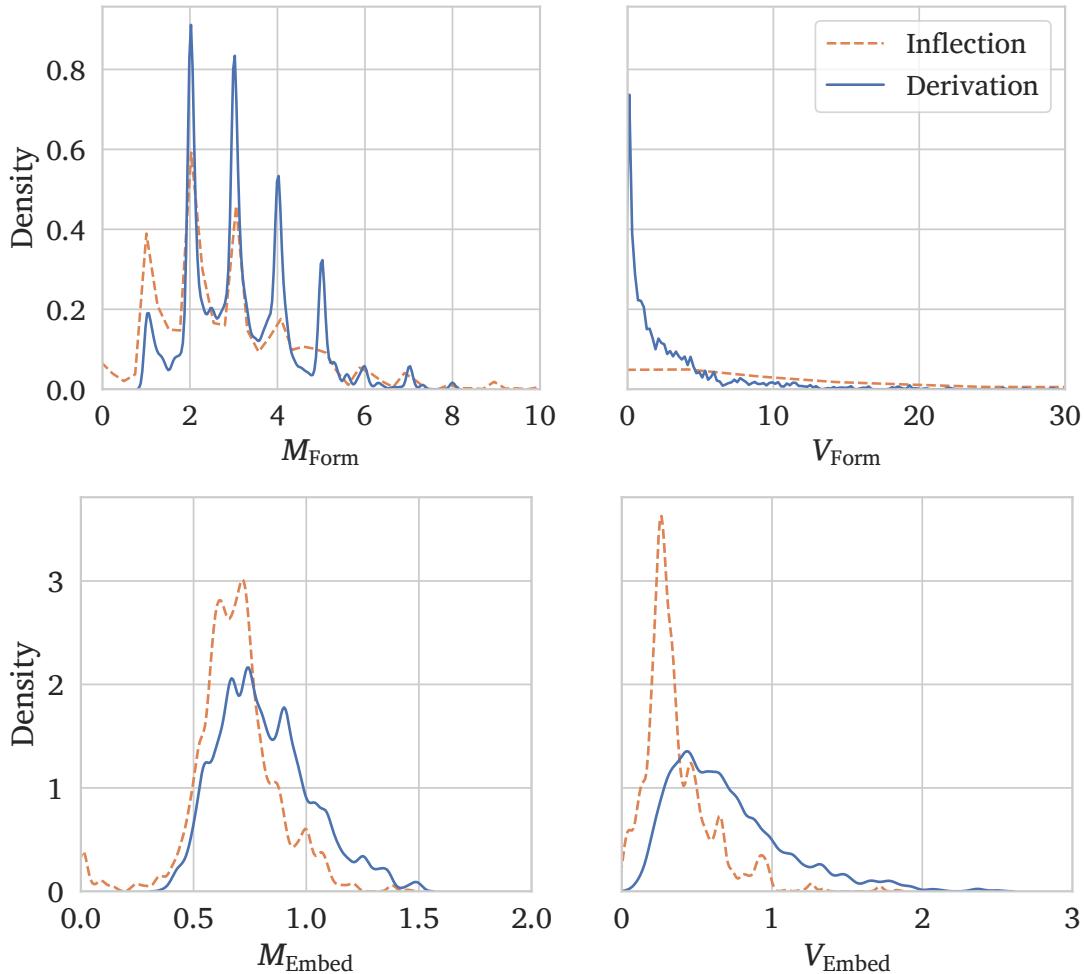


Figure 3.1: The empirical distributions of our four measures (quantifying the magnitude M and variability V of changes in Form and in Embedding space) for inflections and derivations in UniMorph

First, looking at the form measures, we see relatively small effects of inflection-hood and derivation-hood: Cohen’s d for M_{Form} is 0.15, while for V_{Form} it is 0.32. Despite the small difference in M_{Form} between inflection and derivation, the difference does go in the expected direction, with M_{Form} higher on average for derivation than inflection. However, on average, V_{Form} is *lower* for derivation than for inflection—the opposite of what is suggested by Plank (1994) **and much received wisdom in the literature, but in line with my prediction in Section 2.4.1.**

This is discussed further in Section 4.3.1.

In comparison to the form measures, the embedding-based semantics/syntax measures are more strongly correlated with the inflection–derivation distinction. For M_{Embed} , we observe a Cohen’s d of 0.67, indicating a moderately large effect of inflection- or derivation-hood on this measure; while for V_{Embed} we observe a Cohen’s d of 1.09, indicating a large effect. In both cases, we observe larger values on average for derivations than inflections, which indicates that relative to inflections, derivations tend to change a word’s linguistic distribution by a larger amount, and that the direction of this change is more variable. Both of these results are consistent with standard linguistic claims about inflection and derivation.

Prior work on French and Czech has suggested that any single one of these measures will show substantial overlapping regions for inflection and derivation (Bonami and Paperno, 2018; Rosa and Žabokrtský, 2019). Our results confirm this on a larger number of constructions and languages for all the measures we consider.

3.5.1 Effects of Frequency

A potential confounder for our measures on word embeddings is frequency, since the relative frequencies of two words tend to affect their distance in distributional embedding spaces, potentially dominating or complicating meaning-related similarities (Wartena, 2013). In fact, Bonami and Paperno (2018) suggested that differences in frequency may obfuscate measures of semantic distance based on current distributional embedding methods (with low-frequency constructed forms producing larger distances to a given base form than high-frequency constructed forms). If our measures are correlated with frequency, and frequency is also correlated with inflection- or derivation-hood, then any correlation we find between our measures and the inflection–derivation distinction could simply

be due to this discrepancy in frequency rather than to the linguistic properties of interest.⁹ Accordingly, it is desirable to quantify these relationships with frequency.

Unfortunately, for some languages considered here, word frequency information is not readily available. As a result, we restrict ourselves to the 19 languages in our data which are available through the `wordfreq` Python package. We estimate the frequency of unattested word forms as 0. We find the mean frequency of constructed inflectional word forms is less than that of derivational word forms cross-linguistically, with Cohen's $d = 0.71$, indicating a moderately large effect. However, computing Pearson's r statistic for the relationship between constructed form frequency and the four measures under consideration reveals that none of them have a significant linear association with frequency, despite the large number of word forms. While there is a sizeable relationship between some of these measures at the level of an individual distance measure (e.g. the distance between $E(\text{dog})$ and $E(\text{dogs})$), these correlations do not surface when averaged over constructions as we do in this study (e.g. the average distance between a noun and its plural form in English). As such, while our results do not contradict the concerns of Bonami and Paperno (2018), we find we are able to sidestep them in our present study by utilizing a per-construction level of analysis: the effects we find here cannot be explained by frequency of constructed forms.

3.6 The role of syntactic information

Our study uses FastText embeddings as a proxy for both semantic and syntactic similarity. While the ability of such embedding vectors to capture hu-

⁹The reverse could also be a problem: that is, if our measures are correlated with frequency, but inflection and derivation are *not* correlated with frequency, then frequency would introduce an irrelevant confound into our measures and weaken their statistical power.

man semantic similarity scores has been extensively studied (Vulić et al., 2020a; Bojanowski et al., 2017), they are not usually utilized to capture syntactic similarity. Indeed, some studies have attempted to produce more syntactically-aligned embeddings from vectors like FastText (He et al., Tenth century–2018), though replicating these techniques in a highly multilingual setting with low-resource languages is challenging. In this section, we analyse how much syntactic information FastText vectors are able to capture in our dataset, and how much more of UniMorph’s inflection–derivation distinction we might be able to capture with a better representation of syntactic similarity.

To investigate the extent to which distances between FastText vectors encode syntactic information, we consider the mean cosine similarity between embeddings of words in UniMorph that have different parts of speech (using the UniMorph part of speech annotations as shown in Table 1). We take a random sample of up to 5000 words of each part of speech for each language in our data. We then compute mean pairwise cosine similarity within and across these groups per language, and then weighted by number of words of the part of speech per language and averaged across languages. These results are presented in Figure 3.2. As can be seen in the figure, words with the same part of speech exhibit greater mean pairwise cosine similarity than pairs of words with different parts of speech, across all pairs of parts of speech. However, different parts of speech seem to be segregated to different degrees in vector space. On one extreme, we have adverbs where the mean cosine similarity observed between adverbs within a language was 64% greater than with any other part of speech. However, nouns are on average only 6.6% closer to each other than to the average word of their most similar part of speech (adjectives).

However, there are average semantic differences between parts of speech—nouns typically denote objects, while verbs denote events, which could explain some of the above results. To more directly study the syntactic change informa-



Figure 3.2: The mean cosine similarity between FastText embeddings of words of the same and different parts of speech in UniMorph.

tion captured by our embedding-based measures, we fit a logistic regression classifier which uses the two embedding measures (M_{Embed} , V_{Embed}) to classify whether a derivation changes part of speech—essentially using the difference between the base and derived forms in embedding space and the variability of its direction to determine whether the part of speech has been changed or not.

We use 70% of the derivations as a training set, 10% as validation, and 20% as test. We find the classifier is able to predict whether a given construction changes the part of speech with 61% accuracy. Simply predicting the majority class (POS does not change) achieves a test-set accuracy of 53%, so this represents a 9-point improvement. Accordingly, we conclude that our embedding measures capture some information relevant to syntactic change. **However, the relatively modest size of this improvement does not suggest that syntactic information is the primary driver of the correlation we observe between our embedding-based**

measures and the inflection–derivation distinction in UniMorph.

3.7 Conclusion

In this chapter, we have unified and extended prior work on formal and distributional properties of inflection and derivation, defining four measures which quantify the magnitude and variability of changes in form and distribution induced by morphological constructions, inspired by Spencer’s (2013) factorization of the dimensions of morphological change. Across 26 languages and 2,772 constructions, we have found significant differences between inflectional and derivational constructions for all four measures, with derivational constructions tending to induce larger and more variable changes in distribution especially.

The pattern of results for the formal measures is more surprising: derivations are only very slightly longer than inflections on average, but they tend to be *less* variable in their formal changes than inflections. This latter result runs counter to received wisdom in the literature, but is in line with the notion of inflection and derivation as a lexicality distinction in terms of the formal properties discussed in Section 2.4.1. On the other hand, the relatively small difference in length between the categories (M_{Form}) is not what one would expect under the lexicality framing I developed in Chapters 1–2, which suggests that the inflection–derivation distinction is part of general correlation between formal size and semantic contentfulness. This suggests an enhanced role for formal variability in conceptualizing lexicality.

We also demonstrated that our embedding-based measures are not explained by frequency confounds, despite their correlation with inflection- and derivation-hood. This suggests that our measures are capturing relevant linguistic properties of inflection and derivation, beyond the noted assymmetries in frequency across lexicality distinctions. Finally, we quantified the extent to which our

embedding-based measures capture syntactic information, finding that while they can predict part-of-speech changes above chance, the effect is relatively small. This suggests that while syntactic information may play some role in the correlation we observe between our embedding-based measures and the inflection–derivation distinction, it is unlikely to be the primary driver of this correlation.

In line with prior work on single languages, we find substantial overlap between inflectional and derivational constructions for all four measures we consider. However, prior studies have typically focused on individual measures—it may be that the inflection–derivation distinction is better explained by an *interaction* of multiple formal and distributional factors. In the next chapter, I will investigate this hypothesis by building a composite model which *combines* these measures as predictors of inflection and derivation across languages.

Chapter 4

Predicting Inflection and Derivation Cross-linguistically

Despite the centrality of inflection and derivation to morphological description and many linguistic theories, there is substantial disagreement among linguists about the boundaries between these categories and what principles, if any, underly the distinction. A common position is that the distinction is fundamentally **GRADIENT** (Bybee, 1985; Spencer, 2013; Dressler, 1989; Štekauer, 2015; Bauer, 2004). This view has been supported by the existing literature which proposes computational measures of the properties of inflection and derivation (Bonami and Paperno, 2018; Copot et al., 2022; Rosa and Žabokrtský, 2019; Bonami and Strnadová, 2019), finding that while certain measures correlate the inflection–derivation distinction, there is substantial overlap between the categories. However, these studies have focused on one or two measures at a time, which is at odds with another common position in the linguistic literature: that inflection and derivation are **MULTI-DIMENSIONAL** concepts (Spencer, 2013; Plank, 1994; Dressler, 1989; Štekauer, 2015), combining formal and functional aspects. Could inflection and derivation be better characterized by considering a more complete set of formal and distributional factors? In Chapter 3, we

proposed four measures, inspired by (Spencer, 2013), which align with many proposed properties of inflection and derivation, and found that these measures do indeed distinguish inflection and derivation to some extent, but with substantial overlap between the categories. In this chapter, we investigate whether, for a given construction, knowing just these measures is sufficient to predict its inflectional versus derivational status in UniMorph 4.0 (Batsuren et al., 2022).

To what extent can purely quantitative information about wordforms and corpus distribution recapitulate the linguistic intuitions, subjective tests, and comparative concepts encapsulated in the UniMorph annotations? If, across a variety of languages, belonging to different grammatical traditions, language families, and morphological typologies, the UniMorph annotations can be predicted with high accuracy based on our four measures, this would provide evidence that traditional concepts of inflection and derivation *do* closely correspond to intuitions about the different *types* of changes inflection and derivation induce, contra claims that the distinction carries no theoretical weight (Haspelmath, 2024).

To explore this question, we train two different types of machine learning models (a logistic regression classifier and a multilayer perceptron). For each construction in our training set, the models are trained to predict whether the construction is inflectional or derivational, given just four input features: our measures of the magnitude and variability of the changes in wordform and distributional representations. Since we are interested in the cross-linguistic consistency of these predictors, the models are not given access to the input language or any of its typological features. In experiments on 26 languages (including five from non-Indo-European families) and 2,772 constructions, we find that both models are able to predict with high accuracy whether a held-out construction is listed as inflection or derivation in UniMorph (83% and 89%, respectively, for the two models, compared to a majority-class baseline of 57%).

We additionally find that our distributional measures alone are more predictive than our formal ones, and our variability measures alone are more predictive than our magnitude ones; nevertheless, combining all four features yields the best results. Additionally, in Section 4.2, we investigate which *inflectional categories* are particularly likely or unlikely to be classified as inflection by our model, notably finding that inherent inflection is particularly likely to be classified as derivation by our model, in line with Booij (1996)'s characterization of inherent inflection as non-canonical.

Together, these results provide large-scale cross-linguistic evidence that despite the apparent difficulty in designing subjective tests to definitively identify inflectional versus derivational relations, the comparative concepts of inflection and derivation are nevertheless associated with distinct and measurable formal and distributional signatures that behave relatively consistently across a variety of languages. Further analysis of our results does not, however, support the view of these concepts as clearly discrete categories. Although combining multiple measures reduces the amount of overlap in feature space between inflectional and derivational constructions, we still find a gradient pattern, with many constructions near the model's decision boundary between the two categories.

4.1 Predicting inflection and derivation

In this section, we investigate how well the characterization of inflection and derivation given by the UniMorph dataset can be captured by our measures. To do so, we use these measures as input features to simple classification models, which are trained to predict whether a given construction is listed as inflection or derivation in UniMorph, based only on those features. We created a train-validation-test split, randomly selecting 10% of the constructions to reserve for validation and 20% of the constructions for test. We used the validation

set for model selection and hyper-parameter tuning, and the test set was used exclusively for evaluation of the model accuracy. We use the best model trained on this split for the analyses in Section 4.2 and Section 4.3.2. Within the current section, we evaluate our classification methods using stratified 5-fold cross-validation, to ensure the robustness of our findings to dataset splits.

To understand the scenario in which these classifiers are operating, it is helpful to consider some simple baselines. First, we note that simply predicting the majority class across languages, inflection, achieves a cross-validation accuracy of 57%, as there are simply more inflectional constructions than derivational ones in the UniMorph data. However, languages have a highly variable ratio of inflection to derivation constructions in UniMorph; classifying all the morphemes in a given *language* with the majority class for the language instead achieves an accuracy of $69 \pm 1\%$. In other words, a model could capture up to, but no more than, $\approx 70\%$ of the variation in the UniMorph data purely by capturing which language a construction is in—without achieving any ability to distinguish between inflections and derivations within a language. Note, however, that our models must predict whether a construction is inflectional or derivational without access to the language that construction comes from, so even reaching an accuracy of 70% would indicate that the input features encode cross-linguistically informative distinctions.

We tested all possible combinations of features for each of our classification models, but we focus our discussion mainly on combinations corresponding to clear hypotheses about the factors that characterize inflection- and derivation-hood. First, we consider how much any **single** feature recovers the distinction from UniMorph. Secondly, we consider several combinations of two features: (A). **just variability** ($V_{\text{Form}}, V_{\text{Embed}}$): Perhaps it is the case that only variability matters, as investigated in the embedding case by Bonami and Paperno (2018). Or perhaps (B) **just magnitude** ($M_{\text{Form}}, M_{\text{Embed}}$): only the magnitude of the

changes in the components of the lexical entry matters, and variability is in practice a weak correlate or essentially redundant with magnitude. Further, it could be the case that the two measures of either (C) **form** ($M_{\text{Form}}, V_{\text{Form}}$) or (D) **syntax/semantics** ($M_{\text{Embed}}, V_{\text{Embed}}$) alone can recover as much information as all the metrics combined. Finally, of course, there is the hypothesis (E) that **all four features** are important—each contributing some amount of unique information for recovering the distinction from UniMorph.

We explored these features with two types of models: a simple logistic regression classifier, which captures only linear relationships, and a multi-layer perceptron (MLP), which can capture non-linear relationships between features. The logistic regression classifier encodes the assumption that inflection and derivation can be separated by a hyperplane in feature space. If the feature values cluster, without intermediate regions, this corresponds to a categorical characterization of the distinction. If there are instead large regions with intermediate values, this corresponds to a gradient characterization of the distinction.¹ If the non-linear model is required to recover the distinction, then discontinuous areas in the feature space may fall in a certain category, which would not neatly correspond with linguistic intuitions.

First, we consider the logistic regression classifier. As described in Section 3.2, the expectation from linguistic theory is that greater values of any measure should be associated with that construction being derivational. Our analysis in Section 3.5 largely backs up this relation (with the relationship being inverted for form variability), though it is not clear to what degree this relationship is strictly linear.

Due to our highly-restricted selection of measures, we are able to create classifiers with all possible combinations of features. As shown in Figure 4.1, the logistic classifier results best support the **just variability** hypothesis (A),

¹This issue of whether the distinction is gradient or categorical with respect to our measures is discussed further in Section 4.3.4.

with no notable performance gains achieved by adding other features in a linear-modelling setting.

While our best logistic classification model can capture 26 points of variation more than predicting the majority class, it may be missing non-linear interactions between independent variables, or between an individual independent variable and the dependent variable. To account for such non-linear relationships, we fit a multi-layer perceptron (MLP) with a hidden layer size of 100, using the Adam optimizer (Kingma and Ba, 2015) and training for 3000 steps. The number of layers and layer size was chosen using validation set performance, while the number of steps was chosen based on loss convergence on the training set. We find similar patterns of performance for most combinations of predictors. However, we see substantial improvements in performance for combinations of features which include both magnitude and variability features; for example, $(M_{\text{Form}}, V_{\text{Form}})$ improving from $69 \pm 1\%$ to $73 \pm 1\%$. Perhaps as a result of this, we achieve a test-set accuracy of $89 \pm 1\%$, when using all four predictors—representing a 6-point improvement over the best linear model, as well as a 4-point improvement over the best combination of three measures using the MLP $(M_{\text{Embed}}, V_{\text{Embed}}, V_{\text{Form}})$. This therefore suggests that while the variability features are the most descriptive of UniMorph’s categorization of inflection/derivation, all four features contain unique information relevant to recreating this distinction (Hypothesis E).

4.2 Classification of Linguistic Types of Inflection

Given the controversy over what should be considered inflection and derivation, a model that largely aligns with a typical operationalization of the distinction (UniMorph 4.0) may also be of interest in the ways in which it *differs* from that operationalization. Accordingly, in this section, we look at the trends in how

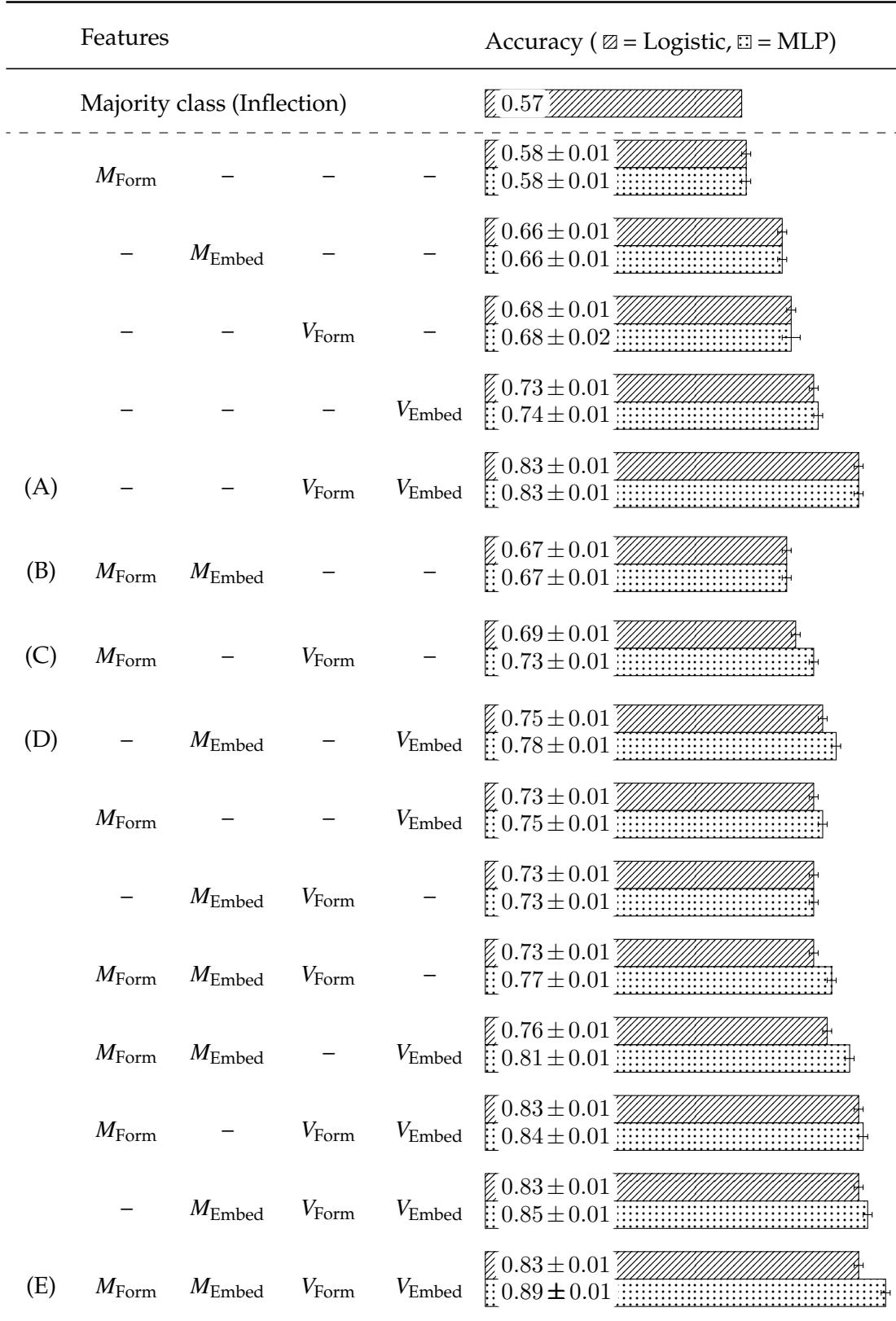


Figure 4.1: Cross-validation accuracy and standard error in reconstructing UniMorph’s inflection–derivation distinction by various supervised classifiers. Linguistically-motivated hypotheses referred to in the text are denoted with letters

our model classifies constructions which are labelled as inflection in UniMorph. We consider several distinctions which we believe to be of linguistic interest, specifically: what kind of meaning is expressed by an inflection; whether it is *transpositional* (changes the part of speech); and whether it is *contextual* or *inherent* (as described by Booij 1996). We ask whether these distinctions affect how likely an inflectional construction is to be classified correctly under our best model (the MLP with all four measures). We focus only on inflectional constructions because UniMorph has cross-linguistically consistent featural annotations on inflections that we can use for the analysis; no such cross-linguistically consistent annotation exists for derivation.

4.2.1 Categories of inflectional meaning

We first consider several categories of inflectional meanings: features for mood (e.g. indicative, subjunctive); tense (present, past...); number (singular, dual, plural...); voice (active, passive); comparison (comparative, absolute/relative superlative, equative); gender, and case. These categories of meaning are often used to structure accounts of inflection, such as UniMorph’s description of its feature set (Sylak-Glassman, 2016) as well as theoretical accounts like Anderson (1985) and even Haspelmath (2024)’s retro-definition of inflection. It is, however, worth noting that not all sources agree on all of these categories as being inflectional. For example, Haspelmath rejects voice as inflectional, and comparison is often omitted from discussions of major cross-linguistic inflectional categories (as is the case in both Anderson, 1985 and even Haspelmath, 2024), and is considered *inherent inflection* (which is less canonical) by Booij (1996). One might reasonably expect constructions which are semantically marked for these controversial categories to be *more likely to be classified as derivation* by our model.

Note that linguists generally agree on which categories of meaning are se-

mantically marked across languages (Greenberg, 1966; Silverstein, 1986; Croft, 2002; Ackema and Neeleman, 2019), and semantic markedness often corresponds to morphological marking. For example, past tense is generally considered more semantically marked than present, and in many languages the past tense requires an affix while the present tense does not. However, the UniMorph annotations include both the semantically marked and unmarked inflections (e.g. V;PAST;PL and V;PST;PL for Ukrainian verbs). Therefore, for the purposes of this analysis, we consider active voice, singular number, nominative case,² and present tense unmarked values, even when present in the featural description of a construction. For example, in Ukrainian verb annotations, V;PAST;PL would be considered marked for tense and number, while V;PST;SG would be considered unmarked for both; both verbs would be unmarked for voice and mood since these are not in the featural descriptions. For the category of gender, we simply consider nouns not to be marked, as their gender is typically not a morphological process but a lexical property.

Figure 4.2 displays the probability that a construction marking for one of these inflection types will be classified as derivation by our best-performing model. As can be seen in the figure, our model does not classify any of these major kinds of inflection as *more derivational than inflectional*; each is substantially more likely to be classified as inflection than derivation. This finding is perhaps unsurprising given our model’s cross-linguistic test set classification accuracy of 90%—it classifies 92% of inflections correctly in general. Accordingly, classifying just 15-20% of constructions belonging to a particular inflectional category as derivations has the potential to be significant.

In order to answer the question “Are constructions which mark for this inflection type significantly more likely to be classified as derivational than others?”, we compute the odds ratio. We focus on the best performing MLP

²While some languages have been argued to mark for nominative case with accusative being unmarked (König, 2006) no such language is present in our study.

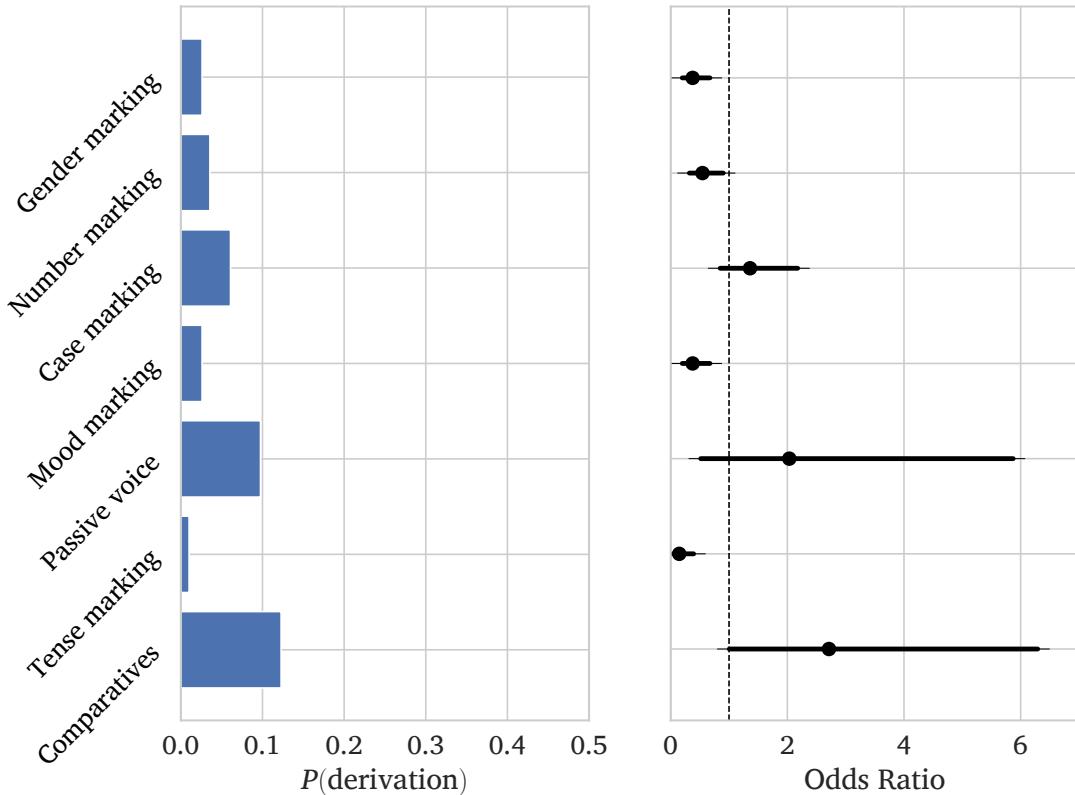


Figure 4.2: Probability and Odds ratio with 95% confidence intervals of being classified as derivation for various kinds of inflectional meaning. Inflections to the right of the dotted line were disproportionately likely to be classified as derivation by our model

model (using all 4 features) in these results, which are presented in Figure 4.2 with 95% confidence intervals. Constructions with an odds ratio significantly greater than 1, while not more likely to be classified as derivation than inflection, can nevertheless be thought of as particularly *non-canonical* types of inflection under our model, while those with odds ratios significantly below 1 are *canonical* with respect to our model.

We apply the Boschloo exact test (Boschloo, 1970) to the results and correct for multiple comparisons with the Bonferroni correction, which yields a significance level of $0.05/7 = 0.007$. We find the odds ratios for gender ($p = 1 \times 10^{-7}$), tense ($p = 3 \times 10^{-7}$), and mood ($p = 1 \times 10^{-7}$) significant. This identifies gender, mood, and tense as particularly canonical inflectional distinctions under our model—all of which are well in line with the claims of Haspelmath and others.

While we do not identify any inflectional meaning categories which are significantly more likely to be classified as derivations than the average inflections, the categories of passive voice ($p = 0.03$) and comparatives ($p = 0.08$) each have 95% confidence intervals which are almost exclusively larger than 1. Each of these categories has been discussed as less canonical kinds of inflection, with comparatives even occasionally being listed as derivations within UniMorph.³ As these are the two least common categories in our sample (consisting of just 57 comparative constructions and 41 passives), it may be that these effects would be significant with a larger sample; alternatively, their relatively high likelihood of being classified as derivation could be an artefact of their rarity in our sample.

4.2.2 Inherent vs. contextual inflection and transpositions

While we do not find any categories of inflectional *meaning* as non-canonical under our model, we also consider two other major categories of inflection that have been discussed in the linguistic literature as potentially non-canonical:

³For example, they are listed as derivations in English, but as inflections in German.

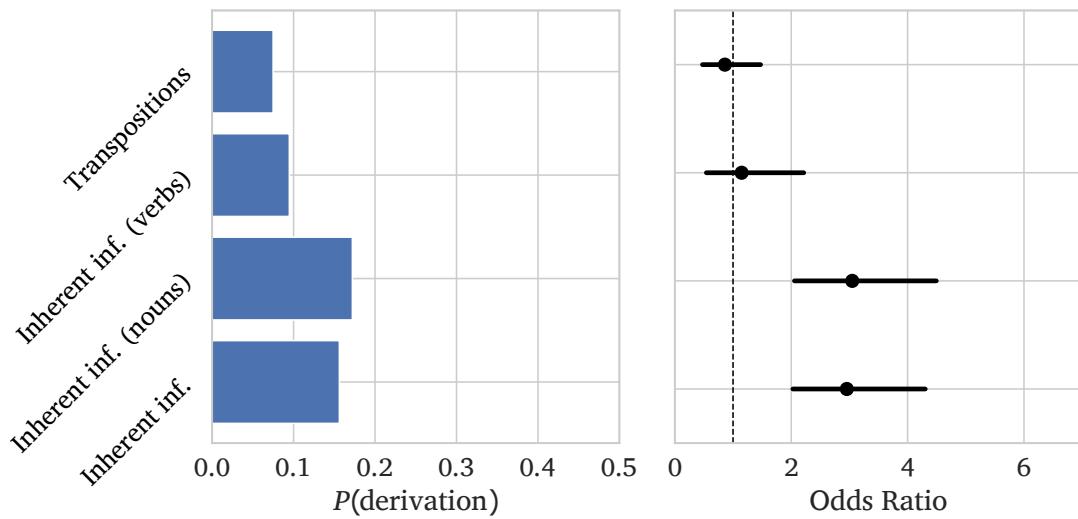


Figure 4.3: Probability and Odds ratio with 95% confidence intervals of being classified as derivation for inherent inflections and transpositions

inherent inflection and transpositions, for which results are displayed in Figure 4.3.

First, we consider Booij (1996)'s notion of inherent and contextual inflection. Booij describes contextual inflection as canonical: it is determined by the syntactic context in which a word appears and indicates agreement (e.g. plural marking on a verb, which is controlled by its subject). In contrast, inherent inflection is non-canonical: it contributes to the meaning of the word itself (e.g. the plural noun). To operationalize this in a simple, cross-linguistically consistent way, we associate number, gender, and case⁴ with nouns—meaning that when those features appear on other parts of speech, we consider them contextual inflections. Analogously, we associate mood, tense, and voice with verbs. We then may consider whether an inflection is *inherent* or not, where we define inherency as not marking *any* contextual features. As shown in Figure 4.3, we

⁴Booij (1996) makes the distinction between structural and semantic case, with the former being contextual inflection and the latter inherent. However, due to the complexity in drawing a line between these categories, we treat all case marking on nouns as inherent.

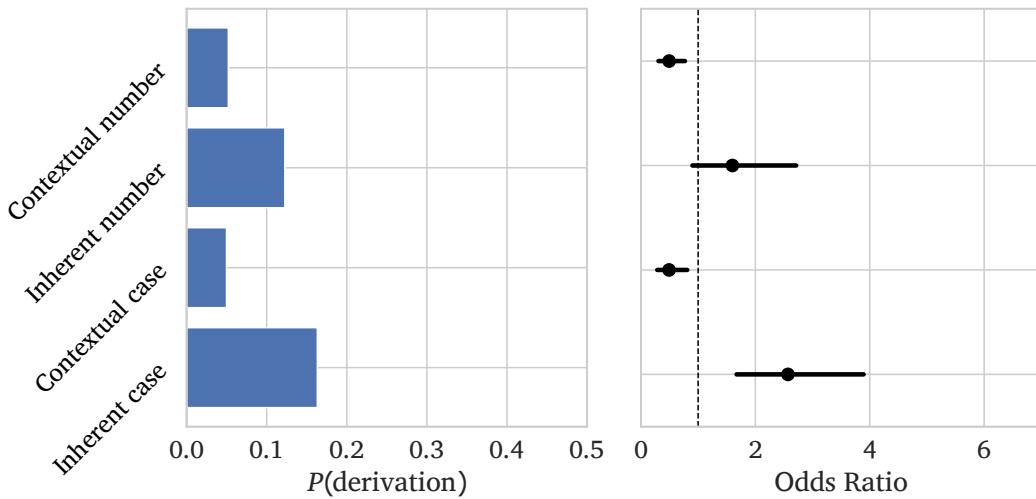


Figure 4.4: Probability and Odds ratio with 95% confidence intervals of being classified as derivation for inherent vs. contextual noun inflections

find that inherent inflectional constructions are not more likely to be classified as derivation than inflection; however, they *are* significantly more likely to be classified as derivation compared to other types of inflections, as quantified by the odds ratio ($p = 6 \times 10^{-9}$). Interestingly, though, we find this to be almost entirely due to nominal inherent inflection ($p = 2 \times 10^{-8}$), rather than verbal inherent inflection ($p = 0.7$). We see this exemplified in Figure 4.4, which shows that inherent case is significantly associated with being classified as derivation ($p = 1 \times 10^{-5}$), while contextual case ($p = 0.003$) and contextual number ($p = 0.0008$) are significantly associated with being classified as inflection.

Finally, we consider inflectional transpositions, denoted in UniMorph as participles (deverbal adjectives), converbs (deverbal adverbs), and masdars (deverbal nouns), shown in Figure 4.3. Transpositions have often been argued to be non-canonical inflection or even derivation because transpositions change the part of speech (Spencer, 2013; Plank, 1994; Haspelmath, 2024). We here find under our model that transpositions appear neither significantly more or

less likely to be classified as derivations than inflections by our model—neither particularly canonical nor non-canonical. This may be due to the non-contextual nature of our embedding model: many inflectional transpositions are syncretic with a non-transpositional form, and our model must assign these the same location in embedding space. Thus, our null result here should not be taken as strong evidence against considering transpositions as non-canonical.

4.2.3 Summary

In this section, we have investigated different kinds of inflectional constructions discussed in the linguistics literature to see whether any of these are particularly *canonical* or *non-canonical* under our model. That is, we looked at whether our model is more (or less) likely to correctly classify these constructions as inflectional, relative to the average inflectional construction.

We identify mood, tense, and gender as *canonical inflections* under our model, but we do not find any categories of inflectional meaning which are significantly *non-canonical* in our sample. We find that inherent inflections are significantly more likely to be classified as derivations, in line with Booij (1996)'s view of them as non-canonical inflection. Interestingly, we find this is driven by inherent nominal inflections rather than inherent verbal inflections. Finally, we investigate transpositions (typically thought of as non-canonical inflection), finding no evidence that they are either canonical or non-canonical under our model.

4.3 Discussion

4.3.1 The role of our individual measures

As shown in Section 4.1, all four of our measures can be used to achieve better discrimination between traditional concepts of inflection and derivation; how-

ever, not every feature plays an equally large role. In this section, we discuss the roles played by each of our features and their connection to linguistic theory.

Among our four measures, our results point to variability of the change in distributional embedding V_{Embed} being the most relevant to traditional categorizations of inflection and derivation. This is in line with the findings of Bonami and Paperno (2018) and Copot et al. (2022) in French, who focus on similar measures as a proxy for semantic drift, as part of a theory where traditional concepts of inflection and derivation reflect higher or lower *paradigmatic predictability*. Indeed, it is possible that this measure could be (roughly) equivalent to Copot et al. (2022)'s predictability of frequency, as it is motivated from a similar theoretical basis. On the other hand, our measure is much simpler to define and compute: attempting to produce a measure of *predictability* immediately raises complex issues around on *what basis* such predictions should be made, complicating the interpretation of results.

In addition, we find a clear and complementary influence of the variability of the change in form, V_{Form} : adding this feature to our model produces a large increase in performance, even when V_{Embed} is already included. This measure (described in Section 3.3.1) can be thought of as a weighted measure of allomorphy, capturing not just the number of distinct patterns, but also their similarity. Our results point to a much higher degree of formal variability/allomorphy for inflections than derivations across a wide range of languages, contrary to the predictions of Plank (1994) and Dressler (1989), but in line with the argument in Section 2.4.1. Although work on French has suggested little difference in the *predictability* of form for derivational and inflectional constructions (Bonami and Strnadová, 2019), we clearly find within our sample of languages evidence that the *actual degree of variation* is very different.

Superficially, this finding could appear to be caused by the fact that derivational allomorphs are sometimes not collapsed in UniMorph data (e.g. *-heit*

and *-keit* being listed as different morphemes in German). However, when we looked into this issue, we found that most derivations had 0–1 such uncollapsed allomorphs. Combining two allomorphs in this way would add at most half the edit distance between the morphs to our measure. In most cases, the edit distance between these allomorphs is 1–2, adding just 0.5–1.0 to the value of V_{Form} . This is much less than the difference between the means of the two categories in this feature, suggesting that failure to collapse allomorphs is not the primary source of this finding. Returning to the example of *-heit* and *-keit* within German, we find *-heit* has V_{Form} of 1.53 and *-keit* has V_{Form} of 1.25. The two morphemes occur 27% and 73% of the time respectively. When combined, they have a V_{Form} of 2.43—still well within the derivational range.

Similarly, one might object that not only such straightforwardly-conditioned allomorphs must be accounted for, but also more idiosyncratic variants that express the same meanings. For example, in French, such formally distinct forms as *-age*, *-ance*, and *-ure* could be argued to be allomorphs of a single action-noun forming morpheme. Copot et al. (2022) handle this by grouping morphemes with similar semantics, by computing average difference vectors in embedding space between base and constructed form for each morpheme, and agglomeratively clustering morphemes with difference vectors with cosine similarity over 0.7. We find such clustering of our data does not sufficiently align with semantic categories of morphemes across our full range of languages to reformat our analysis around it. However, even when clustering derivations with this threshold of similarity, we still find a much lower degree of formal variability for derivations than inflections. On average across languages, 38% of derivational constructions cluster with nothing else at all, without increasing variability. The average cluster contains just 1.8 morphemes, with inflectional morphemes, which are not clustered in this way, exhibiting still 208% more allomorphs on average than derivational clusters.

Future studies should explore the relevance of the variability of form further, to see if it is robust to different languages, and focus directly on the validity of this measure. However, we note that our best performing model without this feature, the MLP with the features (M_{Form} , M_{Embed} , V_{Embed}) achieves a classification accuracy of $81 \pm 1\%$, which is still 23 points above predicting the majority class.

Finally, our results show smaller influence of the magnitude measures M_{Form} and M_{Embed} . This finding seems to contrast with Spencer’s general claim that derivations are associated with larger changes to the properties of a lexeme, but it is not entirely contradictory. In particular, M_{Embed} still displays a fairly strong correlation with inflection and derivation on its own, and likely does not contribute as much to our models due to its substantial correlation (Pearson’s $r: 0.86$) with the more strongly predictive V_{Embed} . In the case of M_{Form} , we find little evidence here that derivations have a tendency to produce larger changes to the form; however, this may be in part related to our need to remove constructions which are orthographically syncretic between the base form and constructed form (which are dominantly considered inflectional in our sample of languages). The length of the change in form does seem to play a small role as a part of a composite set of factors based on its use in our best-performing MLP model.

As noted in Section 3.3.2, our use of FastText somewhat complicates the interpretation of the role of the distributional measures, in the sense that embeddings based on sub-words may capture some formal similarity between words as well as semantic and syntactic similarity. However, we note that if the embeddings do capture formal similarity, at least some of this information must be complementary to that captured by our form-based measures, since including both types of features yields a better classifier than either alone. We also performed some supplementary experiments with Word2Vec embeddings to check that distributional features without sub-word information are also use-

ful.⁵ While overall performance of the classifier was lower (likely due to overall worse quality of the embeddings, for the reasons described in Section 3.3.2), we still found a non-trivial contribution from the distributional features. So, while we can say that both formal and distributional properties are associated with the inflection–derivation distinction, further work is needed to clearly distinguish semantic, syntactic, and formal properties.

4.3.2 Language generality

An important aspect of our model is its language-generality. A major limitation of existing computational studies of the inflection–derivation distinction (Copot et al., 2022; Rosa and Žabokrtský, 2019; Bonami and Paperno, 2018) is their focus on single European languages. In particular, Haspelmath (2024) argues that many properties of inflection and derivation are not proven to apply in a consistent way across languages (especially non-European and non-Indo-European languages). Our model achieves high accuracy across languages, while using no language-specific features. As such, it suggests that across the languages in our sample, inflection and derivation show cross-linguistically similar distributional properties.

Given the large number of European languages in our sample, this result clearly suggests that, at least in the Indo-European family, inflection and derivation are associated with distinct signatures in terms of both their distribution and their form (at least, as expressed in orthography). While evidence for such claims has been provided in specific languages by Copot et al. (2022), Bonami and Paperno (2018), and Rosa and Žabokrtský (2019), many large sub-families within the Indo-European language family had previously been untouched by this literature. Our study includes several Germanic languages with distinctive morphological traits, as well as Armenian, Latvian, Irish, and Greek, covering

⁵For more details about these experiments, see Appendix A.

many smaller European branches of the Indo-European family. We also expand the evidence for consistency in the application of the terms “inflection” and “derivation” within the Romance and Slavic language families. This broad coverage overall provides quantitative evidence for the cross-linguistically consistent application of the inflection–derivation distinction within the languages of Europe—not only in terms of the morpho-syntactic traits of these constructions, as framed by Haspelmath (2024), but also in terms of corpus-based measures which are a proxy for the linguistic intuitions and subjective tests Haspelmath argues should be abandoned.

In addition to this robust evidence that these properties can discriminate inflection and derivation within Indo-European languages, we also show evidence of a degree of applicability to a wider range of languages. On this subset of languages, our best MLP classifier averages 82% accuracy on the test set, lower than for the Indo-European languages (average 91% accuracy). While this is still well above the majority class baseline (74% accuracy on this subset), it suggests that the application of the inflection–derivation distinction to non-Indo-European languages may indeed be less consistent, as suggested by Haspelmath. Of particular note are the results for Turkish. Turkish is a highly agglutinative language with, according to traditional descriptions, an exceptionally rich inflectional system—reflected by an extremely large number of inflectional constructions and relatively small number of derivations in our dataset. Our classifier over-uses the label derivation for this language—classifying all derivations correctly, but also classifying many inflections as derivations. This suggests a mis-alignment between the orthographic and distributional tendencies observed in European languages, and the way linguists typically operationalise inflection and derivation in this language. On a theoretical level, then, our results are therefore compatible with either a view where we should think of some of these so-called inflections in Turkish as more derivational, or a

view where these corpus-based measures are less accurate indicators of what “should” be considered inflection for Turkish.

Due to the relatively small number of non-Indo-European languages and constructions from these languages we are able to consider in the present work, we are unable to draw definitive general conclusions about cross-linguistic consistency in our measures with languages outside Europe. Our results here seem to point to an intermediate view where these corpus-quantifiable correlates of inflection and derivation are *less reliable* descriptors of the way the distinction is made outside of Indo-European languages but still explain *substantial amounts* of the distinction.

4.3.3 The classification approach

Another key differentiating aspect of our work from previous computational studies is our focus on classification of constructions. This method allows us to quantify *how much* of the inflection–derivation distinction, as operationalised across a wide range of languages, can be explained by our simple set of corpus-based correlates. Our focus on a wide range of languages necessitates the use of a quantitative method such as classification, and contrasts with the single-language studies of Bonami and Paperno (2018) or Copot et al. (2022), who focus more on discussing individual constructions.

Further, our goal of looking at whether *multiple features* produces a more clear-cut and less gradient view of inflection compared to the single correlates examined by Bonami and Paperno (2018) or Copot et al. (2022) prevents us from simply doing a statistical test of correlation between a feature and inflection/derivation. While we avoid this by training a classification model, Rosa and Žabokrtský (2019) solve this problem by using clustering. We believe doing so conflates two questions about the measures under consideration. First is the question of how *consistent* linguists’ categorisations are in terms of the meas-

ures. Secondly, there is the question of how *natural* the traditional categories of inflection and derivation appear with respect to these measures. This first question is a lower bar than the latter: it may be possible to use these measures to determine inflectional or derivational status, regardless of whether they form natural clusters in the feature space.

Nevertheless, a finding of *consistency* without *naturalness* is still interesting, given that decisions about what to consider inflection and derivation were made without access to these measures. For example, consistency with respect to these measures could make them a successful “retro-definition” in the terms of Haspelmath (2024). The clustering approach may also fail to identify a distinction where inflection and derivation are predominately located in only slightly overlapping regions of the feature space but do not necessarily form natural clusters.⁶ It is this question of consistency which we primarily consider in this paper, leading us to eschew the unsupervised clustering approach for supervised classification.

Another advantage of our focus on classification is that it naturally lends itself to testing the *generalisability* of our claims: by holding out a random subset of our constructions for testing data and computing accuracy on that set, we confirm that our results do not over-fit to the constructions in the training set.

4.3.4 Inflection and derivation: gradient or categorical?

Whether the inflection–derivation distinction is principally a gradient or categorical phenomenon is a longstanding debate within linguistic theory with potentially wide-ranging implications about the nature of linguistic representations. Many theories of morphological grammatical organisation, production, and processing implicitly or explicitly employ the “split morphology hypo-

⁶As described in Section 4.3.4 and shown in Figure 4.5, it is this situation in which we find ourselves.

thesis,” which holds that inflection and derivation are separated in the grammar (Perlmutter, 1988; Anderson, 1982). Those who propose such separate structures rely on both the distinction between inflection and derivation being discrete and the specifics of that distinction—i.e., what morphological constructions in what languages are considered either inflectional or derivational.

On the other hand, a growing body of linguistic theory rejects a hard distinction (e.g. Bybee, 1985; Spencer, 2013; Dressler, 1989; ?; Corbett, 2010; Bauer, 2004). In its place, they often treat inflection and derivation as a gradient, perhaps emergent out of deeper phenomena. This view has been borne out in the computational work of Bonami and Paperno (2018) and Copot et al. (2022) who find clear continuous gradience with respect to their metrics and the categories of inflection and derivation.

While, as discussed in 4.3.3, we focus primarily on the *consistency* of traditional categories of inflection and derivation, in this section we briefly investigate whether, under our measures, the distinction between inflection and derivation appears more *gradient* or more *categorical*. If the former is the case, we expect a relatively even distribution of constructions in feature space, which (perhaps gradually) transition from being traditionally classified as inflection to being traditionally classified as derivation. In the categorical case, however, we expect *clusters* within feature space with relatively few constructions lying in intermediate ambiguous regions.

We focus on four measures in this study, so we are unable to directly visualise in the feature space. While we applied principal component analysis to produce a two-dimensional representation of our full feature space, the principal components did not pattern into inflectional and derivational regions. This is certainly evidence against *naturalness* of the traditional distinction with respect to our measures. However, we may also look at our two most strongly predictive measures, as shown in Figure 4.5. Recall that a logistic classifier using only these

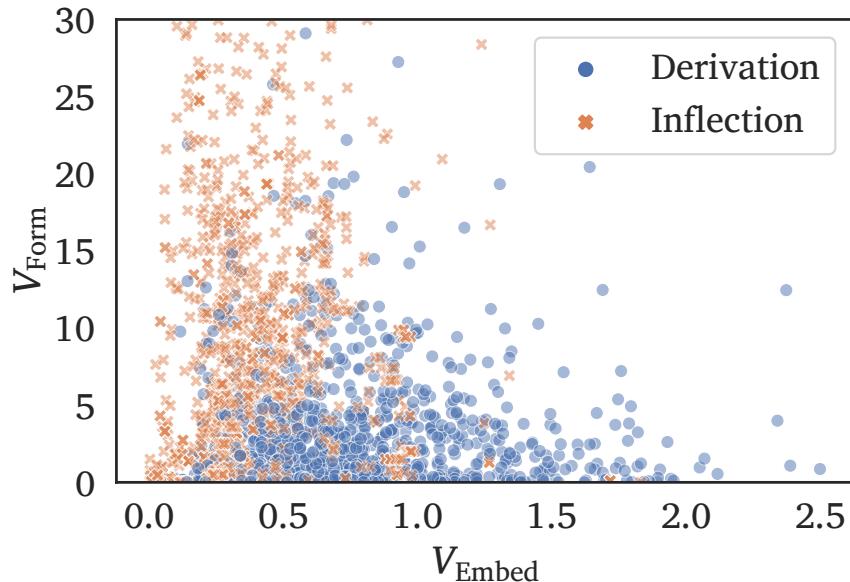


Figure 4.5: Our two most predictive measures for inflection and derivation. Saturation represents overlapping constructions. With respect to these two variables, the inflection–derivation distinction appears gradient rather than categorical

features was able to correctly classify $83 \pm 1\%$ of constructions. Our results with our measures are here consistent with the existing findings of a gradient, rather than categorical, distinction between inflection and derivation with respect to traditional linguistic tests/measures which operationalize them—we observe a spread of constructions in the two-dimensional feature space with a smooth transition between regions containing almost exclusively inflections and regions containing almost exclusively derivations.

4.3.5 Are inflection and derivation identifiable from the statistics of language?

In this work, we have focused on identifying cross-linguistically applicable corpus-based measures, which have a consistent relationship with the traditional concepts of inflection and derivation. While we have primarily motivated the use of these corpus-based measures in terms of quantifying how consistently these categories are applied across languages or making concrete subjective linguistic tests, the fact that they are built purely from the statistics of natural language corpora allows us to consider another important question: is the inflection–derivation distinction something which is present in the statistics of language itself?

If the retro-definition given by Haspelmath (2024) is the right one, for instance, the answer to this question would superficially appear to be *no*. Haspelmath casts the distinction in terms of morphosyntactic feature values, which themselves refer in many cases to the *meaning* expressed by a morphological exponent. If the specific meaning expressed by a morphological relation is necessary to distinguish which relations are inflectional in nature and which are derivational, then the typical inflection–derivation distinction requires *grounding* the meanings of sentences to solve—for example, no amount of raw text input in a language can tell you whether the relationship between two words is “agentive” or “plural.”

The answer to this question has implications within psycholinguistics as well as computational linguistics. Psycholinguistics provides some empirical evidence that inflection and derivation are processed differently (Laudanna et al., 1992; Kirkici and Clahsen, 2013), which seems to imply learners have some implicit ability to categorize constructions into inflection and derivation. How might a learner learn what processing to apply to a given morphological

construction in this case? A substantial body of literature indicates that humans can and do perform purely statistical learning within language acquisition (Swingley, 2005; Saffran et al., 1996; Thiessen et al., 2013; Thompson and Newport, 2007; Thiessen and Saffran, 2003). Without using or even having access to the references of sentences in some cases, learners uncover important aspects of the structure of language. Our results therefore suggest the possibility that statistical learning may play a role in learning to process canonical inflection differently from canonical derivation.

This is also relevant for the validity of several constructs within natural language processing. For example, the paradigm clustering task from SIG-MORPHON 2021 (Wiemerslage et al., 2021), which requires identifying inflectional paradigms from raw text, can only be solved if inflections and derivations can be distinguished from the statistics of such a corpus. Otherwise, derivational relations would be outputted by even the best possible system. Similarly, the task of unsupervised lemmatization (Kasthuri et al., 2017; Rosa and Zabokrtský, 2019) also relies on the distinction between inflection and derivation being evident within a text corpus. Our results point to these types of construct being largely valid for Indo-European languages given the high degree of discriminability between the categories, but our slightly lower results for non-Indo-European languages suggests the need for further investigation into the validity of such constructs for typologically-distant languages to those considered here.

4.3.6 Classification and syntactic change

To place an upper bound on how many of the model’s errors can be explained by syntactic information, we consider how many errors can be explained by a syntactic change oracle variable. Using the annotations for part of speech in UniMorph, we produce a binary variable for whether a given construction changes the part of speech, using the start and end parts of speech for deriva-

tions. For inflections, we assume the part of speech does not change unless it is annotated by UniMorph as one of a participle, masdar, or converb. We add this oracle variable to the input to the classifier. We achieve a test-set accuracy of 84% with the logistic classifier and 92% with the MLP when combined with our four distributional measures. This represents a performance decrease of 2 points and increase of 2 points, respectively, suggesting little-to-no improvement to be found by a feature so closely aligned to linguistic notions of a change in part of speech.

However, this oracle measure captures only a very restricted notion of syntactic change: change in coarse-grained part of speech. For instance, while we treat inflectional transpositions, such as participles, as changing the part of speech in the creation of our oracle variable, this is a contentious point due to some syntactic similarities they share with verbs, which might be reflected in such a measure. On the other hand, some derivations which do not change part of speech may nevertheless change something about the syntactic context (e.g., verbal argument-structure alterations or passive constructions), and may thereby yield greater values in such a measure. A more fine-grained syntax measure which captures this might map more neatly onto the categories of inflection and derivation. Finally, since UniMorph part-of-speech annotations are only at the construction-level, there is no variability in this syntactic information; a distributional account of syntactic information could represent individual pair variation within a construction (due to semantic drift, for example), which might be informative for reconstructing the distinction.

Despite these caveats, these results suggest that syntactic transposition has little added predictive power over and above our corpus-based measures. This is in line with a view of the inflection-derivation distinction where syntactic change is not definitionally related to the distinction, but epiphenomenally correlated with it.

4.3.7 Future work

We believe our study presents a number of interesting avenues for expansion. One such possibility is the extension of the present work to a larger and more diverse sample of languages. In this work, we have taken advantage of the recently produced UniMorph 4.0 dataset to validate claims based on individual languages that corpus-based measures can capture traditional notions of inflection and derivation, and quantify how many intermediate constructions exist under such measures, but our results mostly bear on languages of Europe belonging to the Indo-European language family. While this still represents a substantial advancement in knowledge, and we do find some evidence that our results are applicable to non-Indo-European languages (as described in Section 4.3.2), the evidence presented here cannot yet fully refute Haspelmath (2024)'s claim that inflection and derivation are much less applicable to languages outside Europe. Relatively few (590) of the constructions in our data belong to non-Indo-European languages, with even fewer (201) coming from languages spoken outside Europe, and no representation of languages from outside Eurasia. As argued by Dryer (1989), typological claims must be made not just with normalization with respect to language families or small geographical areas, but even large geographical areas—which is not possible with available data. In order to properly understand to what degree the concepts of inflection and derivation map onto language generally, there is a critical need for the expansion of resources like UniMorph 4.0 and Universal Derivations (Kyzanek et al., 2020) to cover a larger and more representative set of languages. While UniMorph increasingly covers the inflectional morphology of a wide range of languages throughout the world, having added 65 languages from 9 non-European language families in the 4.0 release alone, no unified derivational resource covers a large number of non-European languages. The harmonization and integration of resources like derivational networks such as Hebrewnette

(Laks and Namer, 2022) and finite-state morphological transducers which cover derivation such as Arppe et al. (2014–2019), Larasati et al. (2011), Strunk (2020), or Vilca et al. (2012) into multilingual resources is essential to answering truly general typological questions with these resources in the future.

Another limitation of this study that future work could address is indeed our use of the UniMorph 4.0 dataset. While UniMorph 4.0 provides the largest-scale multilingual dataset of inflection and derivation presently available, it is limited by factors related to its semi-automated construction, which may affect the way allomorphy is represented (as discussed in Section 4.3.1), or other as-of-yet undiscovered systematic biases.⁷

Additionally, we have limited ourselves to a small set of measures here. Future work could seek to improve these measures, or look at other or additional measures. Many previously suggested properties of these categories, such as affix ordering, have directly observable effects on the statistics of text. Future works could test corpus-based measures of distance from the stem or limitedness of applicability, for example. Particularly interesting, we believe, would be the investigation of a syntactic distance and variability component, drawing on works such as He et al. (Tenth century–2018) and Ravfogel et al. (2020)—though there are significant challenges to operationalizing these embeddings in a multilingual, low-resource domain.

There is also room for refinement of our measures and classification techniques. For example, extension to many other languages would likely require a re-assessment of our use of orthography as a proxy for linguistic form. The assumption that orthography is a reasonable proxy for form is not accurate in many languages—however, at present UniMorph does not include phonological

⁷See Malouf et al. (2020) for a discussion of potential pitfalls of the UniMorph dataset for typological research. UniMorph represents not exactly a consensus of highly-trained linguists, but rather largely of the amateur lexicographers that make up the Wiktionary community. Accordingly, as more large-scale multilingual datasets are available, future work should investigate the degree to which these findings are robust to the method of data collection as well as the source of the data.

transcriptions, and automated grapheme-to-phoneme conversion across a broad range of languages is the subject of very active research (Ashby et al., 2021). These difficulties would need to be overcome in order to use phonological transcriptions. Future work should also investigate to what degree our variability of embedding measure is equivalent to or complementary to Copot et al. (2022)’s predictability of frequency measure, as both are motivated from semantic drift due to a change in lexical index. Similarly, future work could clarify the contribution of distributional semantics by using a model such as Word2Vec or GloVe, or newer models of distributional semantics, such as XLM-R (Conneau et al., 2020)—though in the latter case they would have to overcome the difficulties of multilingual decontextualization as described in Section 3.3.2. Further, as we use only two simple classification techniques (logistic regression and an MLP), it is possible that further hyperparameter tuning or use of other techniques, such as random forests or gradient boosting, could improve on classification accuracy.

4.4 Conclusion

In this work, we have presented the first multilingual computational study of the inflection–derivation distinction. In Section 3.3 we define a small set of measures capturing the hypothesized tendency of derivation to produce bigger and more variable changes to the base form in terms of form, syntax, and semantics. We then systematically study the relationship between these measures and traditional categorizations of morphological constructions into inflection and derivation, which we derive from the UniMorph 4.0 dataset. In Section 3.5, we show that these measures each correlate, in some cases strongly, with whether a construction is listed as inflectional or derivational in UniMorph 4.0. We show evidence that these correlations are not due to systematic differences in the

frequency of inflectional and derivational constructions. In Section 4.1, we show that both logistic regression and multi-layer perceptron classifiers which use these measures as inputs can be trained to reconstruct most of the UniMorph inflection–derivation distinction, with logistic classifier achieving a classification accuracy of $83 \pm 1\%$ and the MLP achieving a classification accuracy of $89 \pm 1\%$, improving by 26 and 32 points over predicting the majority class, respectively. We identify the variability of the change in distributional embedding space V_{Embed} and the variability of the change of form V_{Form} as particularly strong correlates of the distinction, together able to classify $83 \pm 1\%$ of constructions as they are classified in UniMorph.

Overall, these results show that much of the categories of inflection and derivation as used in UniMorph can be accounted for by corpus-based measures which make concrete the subjective tests suggested by linguists. In so doing, we have also validated in a larger, multilingual context the core findings of Bonami and Paperno (2018) and Rosa and Žabokrtský (2019), finding that these properties hold across 26 languages (21 Indo-European and 5 others), with a model that uses no language-specific features. These well-defined, empirical measures avoid the often-discussed subjectivity and vagueness of existing criteria (Haspelmath, 2024; Plank, 1994; Bybee, 1985), and enable us to produce the first large-scale quantification of how consistently the categories of inflection and derivation are applied, and validate that these measures can *generalize* to unseen constructions.

With these measures, we are also able to identify in a quantitative way *how canonical* different categories of inflections are (Section 4.2) in terms of properties of their form and distribution. We determine, that, as suggested by Booij (1996), inherent inflection is a *non-canonical inflectional category* under our model: inflectional constructions which are purely inherent are significantly more likely to be classified as derivations than other inflections under our

model. We find in our sample this seems to be particularly due to *nominal* inherent inflections, like case and number. Furthermore, we find no traditional categories of inflectional meaning significantly non-canonical, providing some validation accounts of inflection which are structured around these categories like Haspelmath (2024) or Sylak-Glassman (2016), though we find weak evidence that voice and comparatives could be such categories.

Finally, we note that while there is a high degree of consistency in the use of the terms inflection and derivation in terms of our measures and combining multiple measures reduces the amount of overlap between inflectional and derivational constructions, we still find many constructions near the model’s decision boundary between the two categories, indicating a gradient, rather than categorical, distinction (Section 4.3.4). This gradient region is relatively small, as suggested by our high accuracies, but does not suggest inflection and derivation as categories *naturally emerging* from our measures.

The view from lexicality

I now return to the broader questions and arguments in this thesis. The results in this study, finding a high degree of consistency in the application of the inflection–derivation distinction across a wide range of languages, suggest that the deployment of this distinction is perhaps somewhat less fraught than some have suggested (Haspelmath, 2024; Štekauer, 2015), supporting cross-linguistic consistency in the application of this lexicality-related distinction. Further, the finding that a combination of corpus-based measures can reconstruct the inflection–derivation distinction with high accuracy, in addition to providing tools for analyzing boundary cases, supports the argument in Chapter 2 that defining a multidimensional empirical space for complex comparative concepts like inflection and derivation is a fruitful approach to understanding such concepts—similar to the development of empirical spaces for vowel typology or

color systems.

Do the results here support the view of inflection and derivation as a distinction of lexicality? The picture here remains somewhat mixed. We do see a moderate correlation of our measure of distributional change M_{Embed} with the inflection–derivation distinction, suggesting that derivations do tend to produce larger changes in distributional behavior than inflections, as would be expected if derivations are more semantically contentful than inflections. However, this measure is highly correlated with and slightly less predictive than our measure of the variability of the change in distribution V_{Embed} . This measure, inspired by ideas from the morphological literature about semantic drift and changes in lexical index, raises interesting questions about other levels on the lexicality continuum. If we think of inflectional and derivational constructions as two-item *compounds* between a lexical root and the morphological material, then the importance of V_{Embed} could suggest that the lexicality of units at higher levels of formal structure (such as roots and words) could be reflected in the variability of distributional behavior of these compounds. For example, the prediction is that meanings / distributions of noun–noun or adjective–noun compounds vary more than determiner–noun “compounds”. (e.g. *red line* vs. *the line*). This would be an interesting avenue for future work.

The contribution of the formal measures is also interesting from the perspective of lexicality. M_{Form} contributes very little to our best models, suggesting that the size of the change in form is not strongly related to the inflection–derivation distinction, and against the notion of lexicality as a correlation between formal size/length and semantic content. However, V_{Form} is a useful predictor in our models, suggesting that inflections are more formally variable than derivations. Under the view I sketched in Section 2.4.1, this could be interpreted as inflections having less of a shared formal core than derivations, and being “smaller” in that sense. Overall, the results support a view of inflection and derivation

as a complex combination of form and distribution, with some connections to semantic content.

Part II

Word Classes

Chapter 5

Groundedness and the Lexical–Functional Distinction

In the past century of linguistics, a key question has been whether language is characterized primarily as a formal system of signs abstracted from their signification, or as a functional system for communication. The former view has its roots in the STRUCTURALIST tradition of Saussure (de Saussure, 1916), and has been prominently represented in the GENERATIVE tradition founded by Chomsky (Chomsky, 1957). Beginning in the 1970s and 1980s, there was a reaction against this view, with the rise of FUNCTIONAL and COGNITIVE linguistics (Langacker, 1987; Givón, 1979; Haiman, 1980, i.a.), which emphasized the role of the communicative function of language, and the relationship between linguistic form and meaning. The latter view has gained substantial traction in recent years, with cross-linguistic work demonstrating the importance of functional pressures in shaping cross-linguistic patterns (Croft, 2022; Givón, 1979; Stassen, 1997), as well as language change and variation (Kirby, 1999; Croft, 2000; Zaslavsky et al., 2018).

Paralleling this, one of the most important mathematical advances in the twentieth century was the development of Shannon’s INFORMATION THEORY

(Shannon, 1948), which provided a framework for quantifying information. This framework has been widely applied to the study of language, especially by functionalists (Futrell and Hahn, 2022). Yet Shannon’s theory of information is fundamentally **STRUCTURALIST** in nature; it is concerned with the statistical properties of symbols in a communications system, ignoring the signified. The Shannon’s Information of a linguistic unit combines uncertainty due to both *what* is being expressed (function) and *how* it is being expressed (form). This entanglement is particularly problematic for cross-linguistic comparison, since form is language-specific.

Can we disentangle semantic contentfulness from linguistic form? This is made challenging by the empirical success of both information theory and the **DISTRIBUTIONAL HYPOTHESIS** (Harris, 1954), which together suggest that meaning is inextricably linked to form. In Part I of this thesis, we explored the inflection–derivation distinction through the lens of distributional semantics. Because inflection and derivation are morphological processes which modify word form, using distributional representations from FastText allowed us to study this distinction through the geometry of word vectors. However, this approach entangles form and function; as shown in Chapter 3, word vector similarities also include substantial syntactic information.

In this chapter, we propose a simple mathematical framework for separating form and function in the information-theoretic study of language. By introducing a language-neutral representation of meaning, we can quantify the information due to function alone, which we term **GROUNDNESS**. In this chapter, we specifically focus on **VISUAL GROUNDEDNESS**: 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.

Visual 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 ($\text{LM} > \text{captioning}$), the more *grounded* the word is in that context.

As a case study, we apply this measure to the study of the **LEXICAL–FUNCTIONAL DISTINCTION**. Literature from cognitive, psycho- 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

¹<https://osf.io/bdhna/>

5.1 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 operationalization of word class. In future work, we aim to investigate the relationship between these measures and non-prototypical parts of speech.

5.1.1 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*” 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.

5.1.2 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 judgments, 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 5.4.4, 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 (being the pointwise version of the Shannon Information), 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).

5.2 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 $\mathbf{w}_{<t}$. The groundedness of a token is given by its *pointwise mutual information* (PMI) with the meaning:

$$\text{PMI}(w_t; m | \mathbf{w}_{<t}) = \log \frac{p(w_t | m, \mathbf{w}_{<t})}{p(w_t | \mathbf{w}_{<t})} \quad (5.1)$$

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 5.1 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 5.1 becomes straightforward with neural models: $p_\phi(w_t | m, \mathbf{w}_{<t})$ corresponds to the probability of the token under an image captioning model, while $p_\theta(w_t | \mathbf{w}_{<t})$ 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 (?) 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 modelling 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:

$$\text{PMI}(w_t; m \mid \mathbf{w}_{\leq t}) = \log \frac{p(w_t \mid m, \mathbf{w}_{\leq t})}{p(w_t \mid \mathbf{w}_{\leq t})} \quad (5.2)$$

$$= \log p(w_t \mid m, \mathbf{w}_{\leq t}) - \log p(w_t \mid \mathbf{w}_{\leq t}) \quad (5.3)$$

$$= \text{Surprisal}(w_t \mid \mathbf{w}_{\leq t}) - \text{Surprisal}(w_t \mid m, \mathbf{w}_{\leq t}). \quad (5.4)$$

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 visual groundedness of *word classes*. Drawing inspiration from functionalist typology, we treat a word class 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:

$$p(C_i, m \mid \mathbf{w}_{\leq t}) = \sum_{w_t \in \mathcal{W}} [p(C_i \mid w_t, \mathbf{w}_{\leq t}) p(w_t, m \mid \mathbf{w}_{\leq t})]. \quad (5.5)$$

We can then formulate the mutual information between a word class and meaning as the expected value of the PMI between each token labelled with that class,

and the token’s associated image:

$$I[C_i; m | \mathbf{w}_{<t}] = \mathbb{E}_{p(C_i, m, \mathbf{w}_{<t})} \left[\log \frac{p(w_t | \mathbf{w}_{<t}, m)}{p(w_t | \mathbf{w}_{<t}))} \right]. \quad (5.6)$$

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 C_i in the data \mathcal{D} :

$$\hat{I}[C_i; m | \mathbf{w}_{<t}] = \sum_{(m, \mathbf{w}_{<t}) \in \mathcal{D}} \frac{\mathbb{1}_{C_{w_t} = C_i} \log \frac{p_\phi(w_t | \mathbf{w}_{<t}, m)}{p_\theta(w_t | \mathbf{w}_{<t})}}{\sum_{w_t \in \mathcal{D}} \mathbb{1}_{C_{w_t} = C_i}} \quad (5.7)$$

where $\mathbb{1}_{C_{w_t} = C_i}$ 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.

5.3 Experimental setup

Captioning model $p_\phi(w_t | \mathbf{w}_{<t}, m)$ 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

	Gemma	PaliGemma	COCO-35L
	Pretraining	Continued training	Fine-tuning
Captioning Model	A	☒ A	☒ A
Language Model	A	☒ A	A

Table 5.1: 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, PaliGemma multimodal data for continued training , and COCO-35L image–caption pairs for fine-tuning. Symbols indicate whether models are trained on text data (**A**) or on multimodal data (**☒ A**).

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 $p_{\theta}(w_t | \mathbf{w}_{<t})$ 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 pre-

trained 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 5.1 summarizes the data matching between the two models.

Training details When 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 (≈ 50 hours of training, approximately 3 epochs). Our model achieves lower or similar perplexity on our evaluation datasets than Gemma-2B, suggesting successful domain adaptation (see Appendix B for a perplexity comparison).

Part-of-speech tagging Note that none of the datasets used here come annotated with word class information. We adopt the Universal Dependencies tagset, using Stanza (? , 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 from the present study.

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 leading whitespaces, we adopt the correction proposed by Oh and Schuler (2024) and Pimentel and Meister (2024). Specifically, let \mathbf{s}_{w_t} be the decomposition of word w_t into a sequence of subwords, and $\mathbf{s}_{\mathbf{w}_{}}$ be the decomposition of context $\mathbf{w}_{}$ into a sequence of subwords. Given S_{bow} , the subset of the tokenizer vocabulary that contains subwords that are beginning-of-word (e.g., with a leading whitespace):

$$p(w_t | \mathbf{w}_{}) = p(\mathbf{s}_{w_t} | \mathbf{s}_{\mathbf{w}_{}}) \cdot \frac{\sum_{s \in S_{bow}} p(s | \mathbf{s}_{\mathbf{w}_{} \odot \mathbf{s}_{w_t}})}{\sum_{s \in S_{bow}} p(s | \mathbf{s}_{\mathbf{w}_{}})} \quad (5.8)$$

where \odot stands for concatenation.

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 lan-

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

guages (?) 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.

The following sections quantitatively investigate the trends in our visual groundedness measure across languages and word classes. We begin by examining which word classes exhibit significant groundedness (Section 5.4.1), followed by an analysis of cross-linguistic trends and their consistency (5.4.2 and 5.4.3). Finally, we relate our findings to contentfulness-related psycholinguistic norms (5.4.4).

5.4 Results

5.4.1 Which word classes are grounded?

We first investigate the evidence for groundedness in each word class—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 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 $\text{MI} = 0$.

Results are shown in Figure 5.1. 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.

5.4.2 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 5.2. 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

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

⁴Averaged over values of language and dataset.

⁵Using Šidák corrections; significance threshold = 0.01.

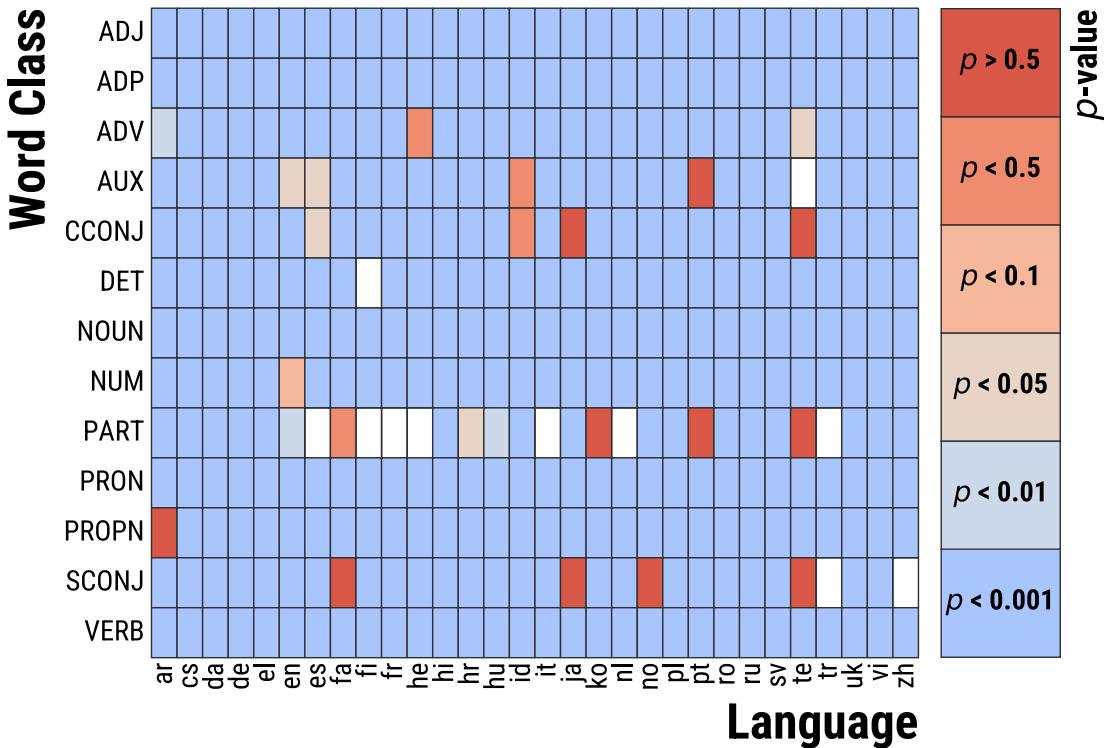


Figure 5.1: Heatmap of mutual information estimates across parts of speech in thirty languages. Cells show the statistical significance of a word class's groundedness ($MI > 0$). Unattested classes are white. Some functional classes display non-significant levels of groundedness in several languages, while lexical classes dominantly show highly significant grounding.

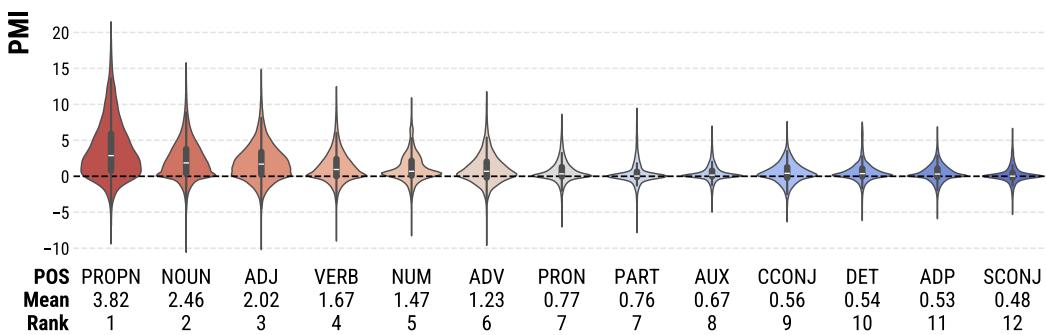


Figure 5.2: Word token level distributions of the groundedness measure (PMI) across all languages and datasets, grouped by part of speech (word class). We also report the estimated marginal mean and ranking of each word class. Colors are based on the ranking of classes, rather than their average PMIs. Overall, the distribution and estimated ranking of word classes strongly suggest our groundedness measure quantitatively captures the distinction between lexical and functional classes.

categories. The ranking corroborates ideas from cognitive linguistics which place nouns, adjectives, and verbs along a lexical–functional continuum, with nouns > adjectives > verbs (?). 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.

5.4.3 How consistent is word class groundedness across languages?

We quantify the strength of the association between visual groundedness and word class on two levels: language-level MI estimates (Figure 5.3), and token-

level PMI (Figure 5.2). 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.

MI estimates For the language-level MI estimates in Figure 5.3, 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 5.2). 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

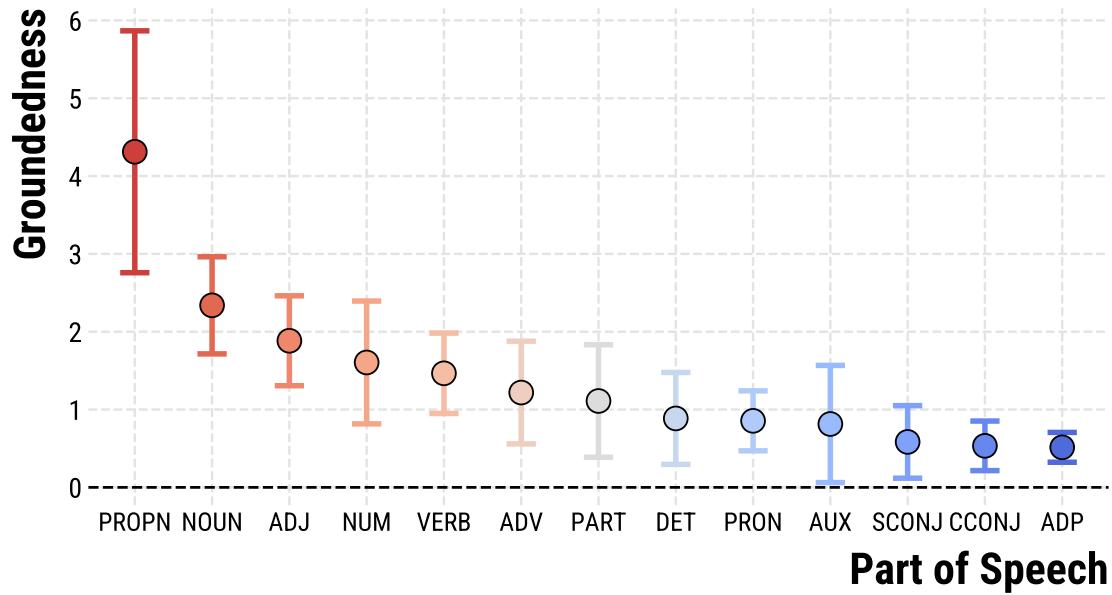


Figure 5.3: Mean and standard deviation of per-language mutual information estimates between word class and image. Across 30 languages, we see clear and consistent tendencies about which parts of speech are more “grounded”, corresponding to a graded distinction between lexical and functional classes.

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

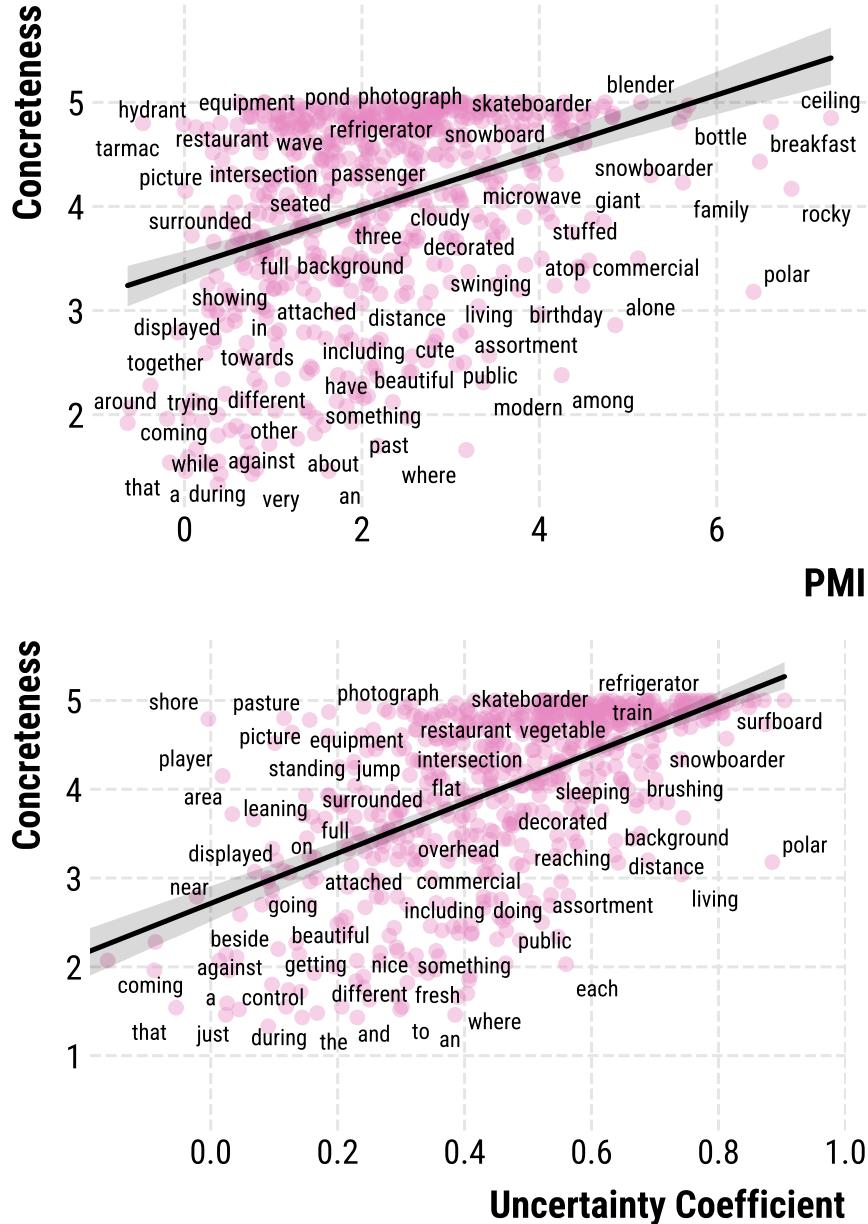


Figure 5.4: Correlation between human concreteness ratings and type-level groundedness (PMI; left, $\rho = 0.368$) or uncertainty coefficient (right, $\rho = 0.609$): i.e., the average ratio between LM surprisal and captioning model surprisal.

5.4.4 Semantic dimension of the measure

In this section we explore the semantic properties of the visual 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 imageability, 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 the Lancaster Sensorimotor Norms (Lynott et al., 2020). Results for concreteness are shown in Figure 5.4 (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 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

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

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

explained by the PMI:

$$U(w_t, m | w_t) = \frac{\text{PMI}(w_t; m | \mathbf{w}_{})}{-\log p(w_t | \mathbf{w}_{})} = 1 - \frac{-\log p(w_t | m, \mathbf{w}_{})}{\log p(w_t | \mathbf{w}_{})} \quad (5.9)$$

Regressing this value against the psycholinguistic norms, stronger correlations emerge (Imageability: $\rho = 0.548$, Concreteness: $\rho = 0.609$ as shown in Figure 5.4 (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 non-specifically in the image captioning context (e.g. “photo” exhibits very low PMIs, because captions frequently begin with “A photo of ...”).

5.5 Conclusion

In this chapter, we introduced GROUNDEDNESS, a simple measure of semantic contentfulness which goes beyond the structuralist and distributionalist nature of traditional information theory and the approach in Part I of this thesis. Utilizing images as a language-agnostic representation of function, we use neural models to measure visual groundedness at both a token and word class level. Our results demonstrate that word classes display *cross-linguistically 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.

In Chapter 6, I extend this work to further study how groundedness relates to cross-linguistic *variation* in word class systems, particularly among lexical word classes. I argue that variation among lexical classes is driven by similar factors to the lexical-functional distinction.

Chapter 6

Splitting and lumping: Visual groundedness as an organizing factor among lexical classes

The detail of the pattern is
movement.

T.S. Eliot, *Four Quartets*

6.1 Introduction

What is the theoretical status of the relationship between meaning and word class? Within any word class in a given language, exceptions to their semantic properties abound. Nevertheless, there is a great degree of cross-linguistic consistency in the relationship between the meaning of lexical items and their syntactic behaviour—the vast majority of languages clearly handle object words differently from action words. Property words also tend to have special morpho-syntactic expression across languages, differing from both objects and actions. But for each of these distinctions, there are languages where it is not clearly

formally and distributionally relevant (Bisang, 2010). How can a theory explain both these strong universal tendencies and well-established deviations from them?

In Chapter 5, we investigated the lexical–functional distinction: the distinction between word classes that are semantically rich and referential (lexical), and those that serve grammatical and syntactic functions. As discussed previously, this distinction has played an important role in theoretical, traditional, and experimental linguistics, but a clear definition is elusive. In chapter 5, I proposed a computational measure, visual groundedness, which could help to clarify this distinction. Visual groundedness shows a clear relationship to the distinction between lexical and functional word classes across 30 languages, demonstrating substantial cross-linguistic consistency—the same classes have similar groundedness across languages.

However, the distinction between lexical and functional classes identified by groundedness is not categorical, but gradient. Traditionally “functional” items sometimes exhibit high groundedness, and “lexical” items range substantially in how grounded they are. In the rest of this thesis, I investigate whether groundedness has the potential to explain not just the cross-linguistic consistency in which classes are lexical and which are functional, but also deviations and gradations within word class organization. In this chapter, I focus on the traditionally “lexical” side of classes in the lexical–functional distinction. The three “major” word classes—nouns, adjectives, and verbs—have often been argued to form a continuum organized around semantic prototypes (). I found a similar continuum between nouns, adjectives, and verbs in Chapter 5. Because these continuum and prototype theories have been argued to explain *deviations* from typical lexical class organization, a question naturally arises: “Can a groundedness continuum help explain how and why some languages split a major lexical class, or collapse two classes together?” In this chapter, I focus on the adjective

class, which has an especially variable cross-linguistic expression and status. I conduct two studies, one focused on word class “splitting”, and one focused on word class “lumping.” With respect to splitting, the first study presents evidence from Japanese, where adjectives are split into two formally distinct classes, *na*-adjectives and *i*-adjectives, which are formally similar to nouns and verbs respectively. While prior work has failed to find a semantic distinction between these classes, I show that their differences in groundedness are iconic of their formal similarities to nouns and verbs, respectively.

To study when and how languages collapse two major word classes together, I present a second study inspired by Wetzer (2013) and Stassen (1997)’s *Tensedness Correlation*, which proposes that more verb-like encoding vs. more noun-like encoding of adjectives in a language is representative of a difference in how *statively* they conceive of verbs—with languages that have a more stative conceptions of verbs using a verb-like encoding for adjectives. Wetzer (2013) identified languages with a more stative conception of verbs as those that do not obligatorily mark tense on verbs, and showed this is strongly associated with “noun-y” vs “verb-y” encoding of adjectives. The proxy of tense expression was necessary because Wetzer (2013) did not have access to the conceptual prototype of verbs; however, I investigate the hypothesis that groundedness, which is higher for more stative concepts like adjectives and nouns, could display a similar pattern, with “verb-y” languages having higher verbal groundedness than “noun-y” languages. However, using present models and corpora, I am unable to find such convergent evidence for the Tensedness Correlation. This study highlights potential difficulties in comparing groundedness values between languages.

6.2 Continua among lexical word classes

One of the major findings of Chapter 5 was that nouns exhibit significantly higher groundedness than adjectives, and both are significantly more grounded than verbs cross-linguistically—despite all being traditionally lexical classes. While many linguistic theories have treated these categories as entirely separate, there is a substantial literature in cognitive linguistics and typology which explores the idea that these categories constitute some kind of continuum within and across languages, especially that adjectives represent an intermediate category between nouns and verbs.

An early influential work in this direction is Ross (1972), who suggested a continuum with adjectives between nouns and verbs, based on syntactic behaviour. In particular, his argument hinges on further intermediate categories, such as different participle uses and “adjectives used as nouns” (e.g. *fun*). He shows an asymmetry and continuum across the application of several phenomena, like preposition deletion and postponing. Subsequent works have built on this idea with different types of evidence. Ross (1972)’s approach to treating the major parts of speech as a continuum through a “category squish” was criticized on a number of fronts (Newmeyer, 1999). Firstly, the ordering within/across categories was motivated formally, but lacked any functional justification. Secondly, the squish being formalized as positions on a real number line between 0 and 1 was criticized as arbitrary—there was no clear external criteria for assigning a particular word/noun-phrase/element its real-valued position in the squish. Structurally, the groundedness approach expanded upon in this part of the thesis is very Rossian in its approach, addressing these two criticisms by adding a functional formalization for assigning real-valued positions (groundedness) to linguistic elements, but ultimately maintaining the unidimensional flavor of Ross’s approach.

Subsequent work built on Ross's ideas by adding functional justifications to both category prototypicality effects and fuzzy boundaries among the lexical classes, and by creating multifactorate accounts. For example, Givón, while considering multiple factors, gives a central role to the notion of *temporal stability* in Givón (1979), citing a cline between nouns, adjectives, and verbs in terms of their prototypical temporal stability, with verbs being the least prototypically stable. Thompson (1988) proposes a view on which adjectives are intermediate between nouns and verbs in terms of discourse function: they are both prototypically *referent introducing* (like nouns) and *predicative* (like verbs). Croft (1991) takes a more multifactorate approach, defining four dimensions across which objects, properties, and actions (the semantic prototypes of nouns, adjectives, and verbs respectively) vary. Noteably, most of Croft's properties have a monotonic continuum between nouns, adjectives, and verbs—the exception being gradability.

	Objects	Properties	Actions
Prototypical Class	Noun	Adjective	Verb
Relationality	nonrelational	relational	relational
Stativity	state	state	process
Transitoriness	permanent	permanent	transitory
Gradability	nongradable	gradable	nongradable
Valency	0	1	≥ 1

Table 6.1: Croft (1991)'s analysis of the conceptual categories of the major parts of speech and their semantic properties.

While these accounts differ in the specific way they break down the parts of speech into a continuum, they are unified in the idea that adjectives represent a position which is in some important way(s) intermediate to nouns and verbs. This idea is supported not just by monolingual evidence, like Ross (1972)'s

English data, but also by a plethora of typological data. Dixon (1977) presents a seminal survey, investigating 17 languages, and proposing 7 categories of properties which vary with how likely they are to pattern with nouns or verbs in the sample. In some languages, there are only a handful of “adjectives” with morphosyntax distinguished from nouns and verbs. For example, Bemba has less than twenty adjectives according to Dixon. Dixon identifies a cline of semantic categories of properties which are more or less likely to pattern with nouns or verbs. For example, MATERIAL properties (e.g. *wood(en)*, *metal*) tend to pattern with nouns,¹ while HUMAN PROPENSITIES (e.g. *kind*, *angry*) tend to pattern with verbs. Other semantic categories fall between these extremes.

This typological evidence suggests a universal conceptual space between nouns, verbs, and adjectives—a semantic map.² However, while evidence for the fine-grained tendencies of semantic categories to pattern with nouns or verbs is compelling, the evidence for more abstract *motivations* for these tendencies is less clear. For example, Stassen (1997) links his own Dixon-like hierarchy of property meanings to Givón (1979)’s temporal stability idea, but the direct typological evidence is for these semantic categories, not for temporal stability itself. As noted by Croft (1991, p. 281) and Uehara (1995, pp. 214–215), persistence and transitoriness is often more complex than such hierarchies suggest, with certain property predicates being persistent for certain entities but not others (e.g. *hard* for a rock vs. *hard* for bread). Further, time-stability is necessarily gradient, and depends on the scale of reference. As such, the more fine-grained generalizations about semantic categories stand on firmer ground than the more abstract generalizations that motivate them.

The notion of temporal stability has often been treated as the key dimension distinguishing nouns, adjectives, and verbs (Verkerk and Lestrade, 2008). How-

¹This was actually identified in refinements to Dixon’s work by ?.

²? investigated this tendency using a multi-dimensional scaling analysis on eleven languages, finding a rich two-dimensional prototype structure which largely aligned with previously proposed semantic dimensions.

ever, this account faces serious challenges. Words such as *lightning*, *explosion*, *puff*, *snowflake*, *bubble*, and *glimmer* describe highly ephemeral phenomena, yet they function naturally as nouns. Their success as nouns suggests that temporal stability alone cannot explain word-class distinctions.

What these “ephemeral nouns” have in common is that, despite their brevity in time, they are spatially contained and identifiable. An explosion, for instance, may last only a moment, but it occupies a bounded region in space and forms a coherent visual object. Indeed, in ?, Givón amended his account to implicate SPATIAL COMPACTNESS—not just temporal persistence—in the nominal prototype, with spatial diffuseness tending to characterize verbs.

Yet temporal and spatial properties alone do not capture the full conceptual space of word classes. As discussed in Chapter 2, RELATIONALITY of meaning as distinct from (but iconically related to) formal valency provides another crucial dimension. Taken together, temporal stability, spatial compactness/diffuseness, and relationality jointly shape how concepts are realized in lexical categories.

These dimensions are not independent. For instance, a concept with high relationality (e.g. *give*) tends to involve multiple participants distributed across space and time, thus exhibiting greater spatial and temporal diffuseness. Conversely, temporally compact experiences that are perceptually salient (e.g. *explosion*, *blink*) often form spatially bounded wholes, encouraging nominalization. Temporally compact experiences which are interesting enough to give a name to often involve motion, which spreads out the reference spatially. Temporal instability also means that, at any given moment, not all the information to fully pin down an event’s category can always be perceived—what appears to be a kick could be someone standing still *as if* kicking, for example. The interrelation of these dimensions has an intuitive connection to visual grounding in images, as they influence how readily a concept can be visually identified.

Importantly, groundedness is not limited to lexical categories. It extends

into the functional domain, organizing the continuum from content words to grammatical morphemes. On this view, the familiar cline from nouns to verbs to adjectives reflects just one region of a broader GROUNDEDNESS and LEXICALITY CLINE that also encompasses function words and (in principal) affixes. This offers a unified framework for connecting lexical class organization with the broader architecture of morphology and syntax.

6.3 Japanese adjectives

The two word classes in Japanese typically described as adjectives are *i*-adjectives and *na*-adjectives. These classes are clearly distinguished from each other in Japanese in terms of their syntax and morphology:

- (6.1) *yama-ga takai / takakatta.*
mountain-NOM high / high.PAST
“The mountain is/was tall.” (*i*-adjective)
- (6.2) *Taroo-ga sizuka da / sizuka datta*
Taro-NOM quiet COP / quiet COP.PAST
“Taro is/was quiet.” (*na*-adjective)

i-adjectives have an analogous inflectional paradigm to verbs (inflecting for aspect and polarity) and can take the syntactic position of verbs as in (1). Both *i*-adjectives and verbs can modify nouns simply by appearing pre-nominally. However, the inflectional paradigm of *i*-adjectives exhibits some differences from verbs, and to be used in reference requires a different construction from verbs.

As shown in (2), *na*-adjectives must be combined with the copula in predication like nouns. But nouns and *na*-adjectives require an attributive marker, *-no* for nouns and *-na* for *na*-adjectives, to modify nouns. Formally, then, *na*-adjectives and *i*-adjectives are more distinct from each other than either is from nouns or verbs respectively.

According to Uehara (1995) and traditional accounts, *i*-adjectives and verbs are closed class, in contrast to *na*-adjectives and nouns, which are open class. However, a recent survey (?) found that new *i*-adjectives have been entering the language at an increasing rate in the past century and a half, including loanwords like *abui* (“abnormal”) and *emoi* (“emotional”)³, suggesting that the class is less closed than traditionally thought.

These categories are not necessarily strictly dichotomous, but rather have fuzzy boundaries. ? showed in his sample that as many of 70% of *na*-adjectives exhibit nominal behaviour in some contexts, such as being used with the nominal attributive marker *-no* rather than *-na*. Further, the boundary between *i*-adjectives and *na*-adjectives itself is not rigid; some stems can be used as either class, like *tisa* (“small”), *ooki* (“big”), or *atataka* (“warm”). ? performed a corpus study on social media which suggested that there might be more fluidity between the two classes in practice, particularly for infrequent or long adjectives. Nevertheless, ambiguity between *i*-adjectives and *na*-adjectives is quite limited; most adjectives belong clearly to one class or the other.

Despite their clear formal differences, prior work has struggled to find a clear semantic distinction between *i*-adjectives and *na*-adjectives. Various semantic distinctions have been proposed. Oshima et al. (2019) found that *na*-properties and *i*-properties both tend to be gradable, but properties that take *-no* in modification (like nouns) tend not to be gradable. Morita (2010) relates it to semantic hierarchies of adjectives, but finds mixed results (e.g. colors are split between the two classes). ? conducted a survey proposing a persistent-transitory distinction between the two classes, but failed to find a significant correlation in corpus data. Overall, semantic accounts of the distinction have proven inconclusive. While ? provides a compelling diachronic account of the origin of the two classes, suggesting that almost all *na*-adjectives arose from

³Japanese publisher Sanseido’s Word of the Year in 2015 (?).

nouns through the recruitment of locational modification constructions, the prevailing view is that there is no synchronically relevant non-formal distinction between the two classes.⁴

6.3.1 Method

I use the models and methods introduced in Chapter 5 to compute visual groundedness scores. Groundedness is formally defined as the pointwise mutual information between a word/linguistic unit in the context of an utterance, and the meaning of that utterance. I focus on *visual groundedness*—representing meaning with an image. As a reminder, for an image I and word w_t in an utterance $W = w_1, w_2, w_3 \dots w_t \dots$, we formalise groundedness as:

$$\text{Groundedness}(w_t) = \log p(w_t | I, \mathbf{w}_{<t}) - \log p(w_t | \mathbf{w}_{<t}), \quad (6.3)$$

Which allows us to compute groundedness as a *difference in surprisal* between an image captioning model and a (domain-matched) language model.

We focus on three datasets: the Japanese subsets of COCO-35L and Crossmodal-3600 (Thapliyal et al., 2022), and STAIR (Yoshikawa et al., 2017). Each of these datasets consists of images paired with one or more captions. COCO-35L is machine-translated from English using Google’s translation service (c.a. 2022), but STAIR and Crossmodal-3600 are human-captioned by native Japanese speakers. Importantly, STAIR is a Japanese re-captioning effort for COCO, so the same images are captioned manually in STAIR that were captioned automatically in COCO-35L. I consider two splits of STAIR: STAIR-dev, which is a set of captions for exactly the same images as COCO-35L-dev, and STAIR-dev-full, a larger split of STAIR that includes additional images from the COCO dataset.

⁴? claims the distinction is purely phonological in the native (non-loaned) lexical stratum, analogous to the distinction between adjectives which inflect for degree in English (“hard”) and those that do not (“difficult”). Uehara (1995) finds Backhouse’s generalization holds for a large portion of adjectives, but suggests that it is due to diachronic factors around the phonological structure of nouns and verbs, rather than representing a synchronic phonological distinction.

This allows me to consider the effect of caption quality and human choice on groundedness estimates for *i*-adjectives and *na*-adjectives. For COCO-35L and Crossmodal-3600 I use the groundedness scores computed in Chapter 5, while for STAIR I compute the scores using the same methods and models to ensure comparability between the datasets.

All datasets are first tagged by the Stanza part of speech tagger to coarsely identify adjectives. However, because this tagger doesn't support the Japanese-specific classes of *i*-adjectives and *na*-adjectives, I use the Sudachi part of speech tagger (Takaoka et al., 2018), as implemented in the `sudachipy`⁵ Python package, to tag identified adjectives with these fine-grained labels. I use this two-stage approach because, while, to my knowledge, Sudachi is the best performing tagger for Japanese that supports *i*-adjectives and *na*-adjectives, it is a simpler, rule-based model, and its overall POS tagging accuracy is much lower than Stanza's (73.7% vs. 95.8%—though note the datasets and tagsets are not directly comparable). Manual inspection revealed that all *i*-adjective and *na*-adjective lemmas identified by Sudachi were correctly classified—as expected given the large differences between the classes in terms of form and formal distribution.

As noted in Chapter 5, single groundedness estimates can be noisy, so we filter for only adjective types which occur at least 5 times in our corpus. This is especially important as *na*-adjectives are less frequent than *i*-adjectives in our corpora.

Statistical model As our datasets are unbalanced and we have multiple observations per word type, we use a linear mixed effects model to estimate the effect of word class on groundedness. We include fixed effects for word class (*i*-adjective vs *na*-adjective).

I have found that position often has ideoyncratic effects on groundedness

⁵<https://pypi.org/project/SudachiPy/>

(e.g. first tokens having a unique groundedness distribution), so I include it as a categorical fixed effect. This control is conservative; positions may not be uniformly distributed across word classes due to their distinct distributional properties, so in the presence of a true effect of word class, this positional control may reduce the estimated effect size.

Finally, I include a random intercept for word type to account for repeated measures. This very strong control allows each word to have its own baseline groundedness, with the only restriction being that all these intercepts are drawn from the same distribution. This accounts for the fact that we have repeated measures for each word type, and that our dataset might be biased towards certain types of words. A significant effect in this regime suggests that even if we had a different sample of word types, we would see the same effect. We fit this model using the `nlme` package in R (?).

6.3.2 Results

Results are shown in Table 6.2. We observe a consistent trend of higher groundedness across all datasets for *na*-adjectives as opposed to *i*-adjectives, though this trend is not significant in all datasets. However, the estimated effect size is remarkably consistent across STAIR and XM3600, hovering around 1 bit. The exception is COCO-35L, where the effect is very small and not significant ($p = 0.68$, $\beta = 0.16 \pm 0.29$). COCO-35L was produced by machine translation from English. Thus, their selection of when to use *i*-adjectives and *na*-adjectives to describe images is likely to be heavily influenced by the English captions, which were not made with awareness of such a distinction. In contrast, the other datasets were captioned manually by native Japanese speakers. We get some indication of this difference by looking at the number of *i*-adjective and *na*-adjective tokens in each dataset. In COCO-35L, *na*-adjectives make up 29% of adjective tokens, while in the three other samples, *na*-adjectives make up

Dataset	MT?	# Captions	Types		Tokens		bits Effect(<i>na-</i>)	<i>p</i> -value
			<i>i</i> -	<i>na</i> -	<i>i</i> -	<i>na</i> -		
COCO-35L-dev	Yes	5316	55	56	4060	1655	0.16	0.68
XM3600	No	2810	42	26	3058	399	0.90	0.029
STAIR-dev	No	6139	60	33	6292	632	1.07	0.12
STAIR-full-dev	No	51805	142	142	52828	6424	0.94	0.015

Table 6.2: Differences in groundedness between adjective classes across datasets. “MT?” indicates whether the captions were machine-translated from English. The effect size is the increase in groundedness (in bits) associated with *na*-adjective-hood, estimated using a linear mixed effects model with fixed effects of word class and position and a random effect for word type. Overall, *na*-adjectives tend to be more grounded than *i*-adjectives. (**Significant results**)

9–12%—so *na*-adjectives are over-represented in the captions translated from English. COCO-35L-dev and STAIR-dev caption the same images, so we can directly compare their results. While in neither case do we find a significant effect on this set of images, the estimated size of the effect is much larger in STAIR-dev ($\beta = 1.07 \pm 0.68$, $p = 0.12$) than in COCO-35L-dev ($\beta = 0.16 \pm 0.29$, $p = 0.68$). Finally, STAIR-full-dev, which includes additional images captioned by native speakers, shows a significant effect ($p = 0.015$, $\beta = 0.94 \pm 0.39$), again with an effect size similar to XM3600 ($p = 0.029$, $\beta = 0.90 \pm 0.41$) and STAIR-dev.

Decomposing groundedness Two terms are used to compute our visual groundedness measure: surprisal under a language model and surprisal under an image captioning model. While we have found a consistent effect of adjective class on groundedness, this could correspond to several different underlying pat-

terns. It could be that *na*-adjectives are more surprising in the linguistic signal, but become equally surprising to *i*-adjectives when the image is provided (that is, adjective class predicts language model surprisal, but not captioning surprisal). Alternatively, *na*-adjectives and *i*-adjectives could become more predictable than *i*-adjectives when the image is provided (class predicting captioning surprisal), which might drive the groundedness effect. We carried out the same mixed effects analysis as before, but with captioning surprisal and LM surprisal as the dependent variables. These results are shown in Table ??.

Generally, we do not find significant effects of adjective class on either LM surprisal or captioning surprisal alone. Our estimates suggest that *na*-adjectives tend to be more surprising in the language model than *i*-adjectives, in line with their overall lower frequency, but this effect is only at $p < 0.05$ in STAIR-full-dev, our largest dataset. However, this surprisal difference is not reflected in the captioning surprisal in 3 out of 4 datasets. This suggests that the greater groundedness of *na*-adjectives is driven by their greater surprisal in the language model, which is then largely mitigated by the image information in the captioning model. On COCO-35L-dev, where we saw the least evidence for a groundedness difference, we see that *na*-adjectives are significantly more surprising under the captioning model as well, suggesting that the image information does not mitigate their greater surprisal in the language model. This may be related to the unnatural use of *na*-adjectives in COCO-35L, as discussed above.

6.3.3 Discussion

Overall, our results suggest that *na*-adjectives express more visually grounded meanings than *i*-adjectives in Japanese. This is in line with the formal similarities of *na*-adjectives to nouns and *i*-adjectives to verbs, as nouns tend to be more grounded than verbs cross-linguistically. This finding contrasts with prior work which failed to find evidence for a semantic distinction between these

Dataset	MT?	LM surprisal		Captioning surprisal		
		bits	Effect(<i>na-</i>)	<i>p</i> -value	bits	Effect(<i>na-</i>)
COCO-35L-dev	Yes	1.34 ± 0.71		0.063	<u>1.15±0.46</u>	<u>0.014</u>
XM3600	No	1.13 ± 0.78		0.154	0.278 ± 0.61	0.65
STAIR-dev	No	2.01 ± 1.15		0.085	0.96 ± 0.78	0.22
STAIR-full-dev	No	<u>1.26±0.58</u>	<u>0.030</u>		0.35 ± 0.40	0.38

Table 6.3: The effect of adjective class on LM surprisal and captioning surprisal. We find that *na*-adjectives tend to be more surprising in the language model than *i*-adjectives, but this effect is reduced by conditioning on the images, resulting in higher overall groundedness. (**Significant results**)

classes (Morita, 2010; Oshima et al., 2019; Uehara, 1995), suggesting that visual groundedness may be a useful tool for uncovering semantic distinctions that are not easily captured by traditional semantic features.

The results suggest that both categories of adjectives have similar levels of predictability when the image is provided, but *na*-adjectives are more a-priori surprising. This suggests that *na*-adjectives are used to express properties which are less frequent and more specific, but still highly salient in the visual context.

While STAIR-dev and COCO-35L-dev caption the same images, XM3600 and STAIR-full-dev cover very different image distributions and were captioned by different people, so the similarity between the findings in these datasets is encouraging. The differences between COCO-35L-dev and STAIR-dev suggest that naturalistic use of *na*-adjectives results in a stronger groundedness effect, as COCO-35L was machine-translated from English captions which do not make the *i*-adjective/*na*-adjective distinction.

In a few instances, there are closely-related *i*-adjective and *na*-adjective lemm-

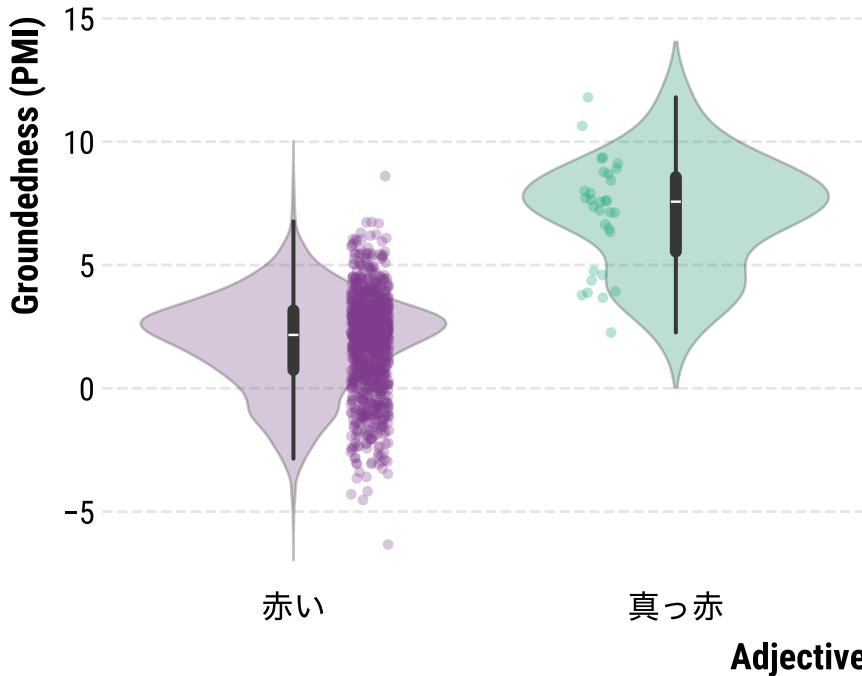


Figure 6.1: Groundedness scores for *na*-adjective *makka* (completely red; right) and *i*-adjective *akai* (red; left) in the STAIR-full-dev dataset.

mas which appear in the datasets. For example, with color terms, both *akai* (*i*-adjective, red) and *makka* (*na*-adjective, completely red) appear. Figure X shows the groundedness scores for these two words in STAIR-full-dev. We can see that *makka* has consistently higher groundedness scores than *akai*, suggesting that even for very similar meanings, *na*-adjectives exhibit higher visual groundedness than *i*-adjectives. We observe a similar pattern with other closely-related pairs, such as *shiroi* (*i*-adjective, white) and *masshiro* (*na*-adjective, completely white) or *koudai* (*na*-adjective, vast) and *hiroi* (*i*-adjective, wide).

While it is clear that the *na*-adjective/*i*-adjective distinction is not synchronically purely semantic, our results suggest that visual groundedness plays a role in how these classes are organized and used by speakers. This finding supports the broader hypothesis that groundedness plays a role in how word classes are

organized cross-linguistically. Future work should explore other idiosyncratic splits in word classes across languages to see if similar groundedness effects are observed.

6.4 The Tensedness Correlation

In the previous section, I showed evidence from Japanese that groundedness could provide a novel explanation for seemingly idiosyncratic *splits* within the major word classes. Could groundedness also account for similarities or lumping behaviour among the major word classes?

Forming a quantitative hypothesis for testing whether groundedness is operative in how word classes split was relatively straightforward. If there are multiple classes for a single class, it follows that an influence of groundedness on word class structure implies a difference in groundedness between those classes. Due to the formal and distributional differences which are constitutive of a split in word classes, these classes are further easily identified with existing classifiers. Finally, the general prototype theory of cognitive science, as has been applied in cognitive linguistics, gave a clear hypothesis for the directionality of the groundedness effects: greater formal similarity to nouns should imply higher groundedness, while similarity to verbs has the reverse implication.

However, identifying a hypothesis for the behaviour of lumped classes is less clear. If a language lumps together verbs and adjectives, or adjectives and nouns, how could this be predicted by groundedness? Simply measuring the groundedness of the combined verb–adjective class, for example, seems potentially tautological: we already know that adjectives, and by extension, property meanings, tend to have higher groundedness than verbs cross-linguistically; therefore, an observation of “higher” groundedness with respect to some comparative base (to be specified) is simply to be expected, and such an observation

of the mean does not seem on its face informative about why *this language* has organized its part of speech in this way.

Despite this fundamental difference between lumping and splitting, in this section, I expound on a theory from typology and cognitive linguistics which purports to explain a type of “lumping” behaviour among the major classes and demonstrate that it can be translated into a number of more specific hypotheses about groundedness, which invoke different interpretations of this hypothesis.

6.4.1 The typological finding

Wetzer (2013) first noted that in predication, languages rarely employ a unique strategy for adjectives/property words; rather, languages generally fall into two camps: those which encode property predication identically to/similarly to nouns, and those which encode property predication like (intransitive verbs). Wetzer calls such languages “nouny” and “verby” with regard to adjectivals respectively. As a canonical example of a nouny language, we can consider a language like German:

- (6.4) *Der Mann ist alt*
 The-MASC man is old
 “The man is old.” (**Property predication**)

- (6.5) *Der Mann ist Arzt.*
 The-MASC man is doctor
 “The man is a doctor.” (**Nominal predication**)

- (6.6) *Der Mann läuft.*
 The-MASC man walk-PRES.3SG
 “The man is walking.” (**Verbal predication**)

As we can see, the same strategy (in German, copular marking) is used for nominal predication, but not On the other hand, Mandarin Chinese is a canonical verby language (? , p. 148, 143):

- (6.7) *Zhāngsān shì yi-ge hùshī.*
 Zhangsan COP one-CLF nurse
 “Zhangsan is a nurse.” (**Nominal predication**)

- (6.8) *tā pàng*
 3SG fat
 “She is fat.” (**Property predication**)

- (6.9) *tā yóuyoòng*
 3SG swim
 “She swims.” (**Verbal predication**)

Wetzer (2013), and subsequently and more comprehensively Stassen (1997), identify that most languages (85% in Stassen (1997)'s sample of 410 languages) exhibit only a single strategy for all adjectives—either nouny or verby.⁶ The remaining languages exhibit some kind of *mixed* strategy (Japanese representing an extreme of this type of language).

Wetzer and Stassen show extensive cross-linguistic evidence for what they call the *Tense* or *Tensedness Correlation* (going forward, I will refer to it as Stassen does, as the “Tensedness Correlation”). They define the typological parameter of “Tensedness” for a language. A language is **tensed** if it has obligatory morphologically bound marking which distinguishes (at least) between past and non-past time reference. If such marking does not exist, it is expressed as something other than a bound form, or it is not obligatory, the language is non-tensed. The Tensedness Correlation claims:

1. A language is nouny if and only if it is tensed.
2. A language is verby if and only if it is non-tensed.

⁶Given that these strategies do not cover all constructions involving adjectives, it is not “lumping” in the sense of those that seek to identify if a language “has” or “lacks” adjectives. However, such questions fall pretty to what ? calls “methodological opportunism”: the fact is that it is not clear in which constructions adjectives need to differ from nouns and verbs to count as a distinct class. Rather than studying whether a language “has” or “doesn’t have” adjectives, I am studying whether the similarity of adjectives in key constructions to other classes reflects something about their groundedness.

Stassen (1997) shows overwhelming cross-linguistic evidence for this claim. While exceptions exist, in the vast majority of cases we see a bidirectional relationship of tensedness and nouny coding of adjectives in predication. Stassen (1997) and Wetzer (2013) argue extensively that exceptions to this generalization can largely be understood as artifacts of recent diachronic changes in languages or cases on the margins of being a tense system. For example, the *i*-adjective category in Japanese uses verby encoding, but Japanese in the present day seems to have a tense system, though its tense properties are more recent and the status of tense as opposed to aspect in the language is a matter of some debate ().

6.4.2 Theoretical explanation of the finding

While the typological finding is widely considered to be robust, on its own it lacks motivation—why should it be that these factors are correlated?

Situated in the cognitive linguistics literature around the prototype and continuum structure of parts of speech I summarized in Section ??, both Wetzer (2013) and Stassen (1997) focus on the dimension of *time-stability*. Drawing on Givón (1979), they argue that adjectives/properties represent an intermediate level of time-stability between nouns and verbs. Stassen (1997) gives a particularly detailed argument that in many languages with some degree of mixed encoding for properties, more time-stable properties more likely to be encoded nounily. Wetzer (2013) argues that given the intermediate time-stability of properties, the Tensedness Correlation reflects the prototypes of verbs in different languages. Specifically, Wetzer claims that languages that have a more stative, temporally extended, and stable verbal prototype reflect this through their lack of tense marking, while languages that conceptualize verbs as more time-bound and less stative reflect this through their obligatory morphological tense.

Stassen does not directly invoke a conceptual verbal prototype, but makes a similar argument. Bybee (1985) argued that morphological boundness reflects

the *semantic relevance* of a morpheme to the stem. Events (the prototype of verbs), as the least time-stable predicate type, “attract” bound tense morphemes, in Stassen’s view. This is, he argues, a more specific instantiation of Haiman (1980)’s *Structural Iconicity*: the tendency of linguistic structure to reflect the conceptual structure of human experience. Obligatory, bound tense marking is iconically motivated by a conceptual closeness/entanglement between an event and its location in time. He goes on to argue that, for prototypical properties (e.g. forms, dimensions, colors)—bound tense marking is at best non-iconic, and possibly ‘anti-iconic’: given their time-stability, marking them with tense is inappropriate. That is, rather than a shift in verbal prototype per se, he sees the emergence of bound tense marking as a boundary which initiates a process of kicking property meanings out of the verbal category and towards a new, more noun-like expression.

6.4.3 Methodological background

With this theoretical groundwork laid, I will now argue for an analogous testable hypothesis about visual groundedness.

6.4.3.1 Shifted prototypes

While Wetzer argues explicitly for a shift in the time stability of a verbal prototype towards nouns, such an explicit argument is lacking from Stassen’s exposition. Stassen removes the causal role of the prototype shift in the Tensedness Correlation, replacing it with the interacting, conflicting forces of iconicity for the temporal specificity of events and the temporal extendedness of prototypical properties.

I do not wish to present a picture in which Wetzer argues for a prototype shift and Stassen argues against it. If the verbal prototype represents some kinds of summary of the types or tokens in the verbal class, Stassen’s argument, I argue,

also suggests some shift in the prototype of verbs. First, though verby encoding is not the same thing as adjectives being morphosyntactically undistinguished from verbs, in many verby languages this is (roughly) the case. In such a case, the ejection of adjectives from the verbal class proper should shift the verb prototype. Further, verbs can vary substantially in their temporal stability: prototypical verbs are punctual, like *jump*, *kick*, or *hit*; however, verbs can be durative to varying degrees, like *rain*, *dwell*, *believe*, and *sit* in English. Durative meanings should also be more likely to abandon a verbal encoding if tensedness is required.⁷ Overall, such individual attritions could, over time, accumulate into a shifted underlying verbal prototype in time-stability.

The typological evidence and argumentation I have presented has focused on the temporal stability dimension of the noun–adjective–verb cline, as this fits cleanly with the notion of tensedness. However, given the issues with temporal stability of the spectrum discussed in Section ??, and the positive findings in Japanese with groundedness in contrast to previous negative findings with temporal stability, I propose investigating the Tensedness Correlation from the perspective of groundedness. I argue that the same logic applies: if a language has a shifted verbal prototype towards more grounded meanings, this should be reflected in both the absence of bound tense marking and the encoding of properties as verbal.

6.4.3.2 Measuring a groundedness shift

Comparing surprisals across languages is fraught with complexity. Previous studies have occasionally assumed surprisal First, it must be noted that because we are unable to train independent language and captioning models on multi-parallel data, the raw surprisals (and thus groundedness scores) of our model may not be comparable.

⁷Factors such as relationality (e.g. the transitivity of *dwell* and *believe*, which is non-prototypical for properties) can block this transition.

If the models were trained on a parallel corpus, then the only difference between the models should be the differences in the way a language encodes the same content. However, when the corpus is not parallel, the models have different exposure to words, constructions, and concepts, and so may learn different distributions. If we assume there is an underlying “true” distribution for a given language that these models are approximating, then as the quantity of data grows, this effect should diminish. However, we must definitely take this possibility seriously here, as the languages in our sample vary widely in their resource level, and the quality of captions the model is trained and evaluated on may also vary (both because of differences in the quality of machine translation, and differences between captioners in the XM3600 dataset).

Identifying which languages might have worse models is also tricky. The full language composition of the multilingual pretraining corpus for PaliGemma (WebLI) is not public, so I cannot directly measure the amount of data per language. Further, while we have CIDEr scores for the captioning model on each language’s test set, these are not directly comparable across languages, as CIDEr is based on n -grams, so they are sensitive to the amount of information borne by an orthographic word in a language (which varies considerably)⁸. There is also the issue of *language relatedness*: modelling of less-resourced languages may be improved by transfer from related higher-resource languages in the pretraining corpus, which complicates the relationship between resource level and model quality. Additionally, orthography may play a role: languages with non-Latin scripts may be disadvantaged by suboptimal tokenization in the multilingual model.

Control variables To address these concerns, I introduce a few (imperfect) controls. First, I include a binary variable in my statistical model indicating

⁸Accordingly, I observe the lowest CIDEr scores in the dataset for Finnish.

whether the language is **written in a Latin script or not**. This should enable us to understand how much of the cross-linguistic variation in groundedness could be related to orthographic differences. We would also like to control for model quality more generally. While the word is not a comparable unit across the languages in the study, the data is parallel at the level of sentences, so sentence-level metrics should be comparable. I use the ratio of sentence-level negative log-likelihood (NLL) under the captioning model to that under the language model as a proxy for model quality—the smaller this ratio, the larger the effect of the image on surprisal:

$$\text{Quality Ratio} = \frac{\text{NLL}_{\text{Captioning}}}{\text{NLL}_{\text{Language Model}}} \quad (6.10)$$

This measure was chosen because it is independent of the absolute surprisal values, and incorporates both language model and captioning model performance. Intuitively, one might think that higher NLL under either the captioning or language model would indicate a worse model, but this is not necessarily the case here. I observe some of the highest NLL values in the dataset for very high-resource languages which also achieve high CIDEr scores (e.g. English, Spanish). This suggests that for some of the lower-resource languages, the model is overconfident in its predictions, leading to *artificially low* NLL values (e.g., we observe the lowest sentence-level NLL for Telugu, which was designated as one of the five lowest-resource languages in the corpus by the authors of XM3600). The ratio metric we use here is independent of the absolute NLL values, and captures how much of the surprisal is being explained by the image. A lower ratio indicates that the image is having a larger effect on surprisal, which should indicate better captioning and language models. I use the ratio as computed over sentences in COCO-35L-dev as a control variable, as the captions in this set are direct translations across languages.

Finally, another factor that could influence verbal surprisal specifically is **word order**. In languages where the object proceeds the verb, this could make

verbs more predictable from the linguistic context when the object is prototypically associated with the verb. Therefore, I include a binary variable indicating whether the language has Object–Verb or Verb–Object word order.⁹

Relative shift While the previously mentioned controls should help us assess the true effect of nouny/verby encoding on verbal groundedness, it looks only for evidence of an *absolute increase* in verbal groundedness in verby languages. However, the hypothesis could manifest more weakly as a difference in the groundedness of verbs in a language *relative to other parts of speech*. To test this idea, I perform Z-score normalization on the groundedness scores in each language, measuring how many standard deviations away a word token is from the mean overall groundedness of tokens in a language. Then, an estimate of the groundedness dimension of the verbal prototype was computed as the token-wise average of groundedness within the class, as in Chapter 5.

Boundary between verbs and adjectives Finally, it is worth noting that the UPOS¹⁰ verb category may not perfectly capture the verbal prototype in all languages. In particular, in verby languages, adjectives are often very similar to verbs formally and distributionally, and it is possible that some legitimate members of the verbal prototype are being tagged as adjectives instead of verbs. This could artificially lower the estimated groundedness of the verbal prototype in verby languages.

For example, in Korean, the vast majority of adjectives behave as a type of verb in general (in attribution as well as predication, e.g.), yet in Universal Dependencies (and consequently Stanza), these are always annotated as adjective. Some of these “adjectives” could be more stative verbs. This could be further compounded by tagger behavior—the less formally distinct verbs and adjectives

⁹I code `fa`, `te`, `ko`, `hi`, and `tr` as Object–Verb languages, based on ?.

¹⁰Universal Parts of Speech; the categories utilized in Universal Derivations which are deployed by the Stanza tagger we use in these analyses.

are, the easier it becomes to mis-tag adjectives as verbs. Some (potentially poorly defined, and unknown) amount of members of the verb class could be getting “lost” to the adjective class in the annotation scheme used here, and it is possible that an analysis that carefully identified them would demonstrate that verby languages *in fact* have a more grounded prototype.

In this final experiment, I combine the UPOS categories of verb and adjective *for the verby languages only*. Adjectives are in general more grounded than verbs, so adding them to the verbal prototype should make verbs more grounded than languages which do not include them. This result should provide an upper bound on the true groundedness estimate of the verbal prototype in verby languages in this dataset. The true effect of verby encoding should lie somewhere between the results of this analysis and the previous one. Relatedly, a finding that verby languages still have lower groundedness even with adjectives included would suggest that the effect of model quality is disproportionately affecting verbs, rather than resulting from an artifact of misclassification of parts of speech.

Nouny and verby languages We follow Stassen (1997)’s analysis of which languages are nouny and verby. Out of the sample, Hindi (`hi`), Indonesian (`id`), Chinese (`zh`), Korean (`ko`), and Vietnamese (`vi`) are classified as verby languages. Japanese, having a mixed strategy, is excluded from this analysis. The remaining 24 languages in the sample are classified as nouny languages. Noteably, Stassen identified Korean as a slightly problematic case, as it meets his criteria for being tensed, but has clear verby encoding of adjectives. He suggests this is due to recent diacronic changes around tense in Korean.

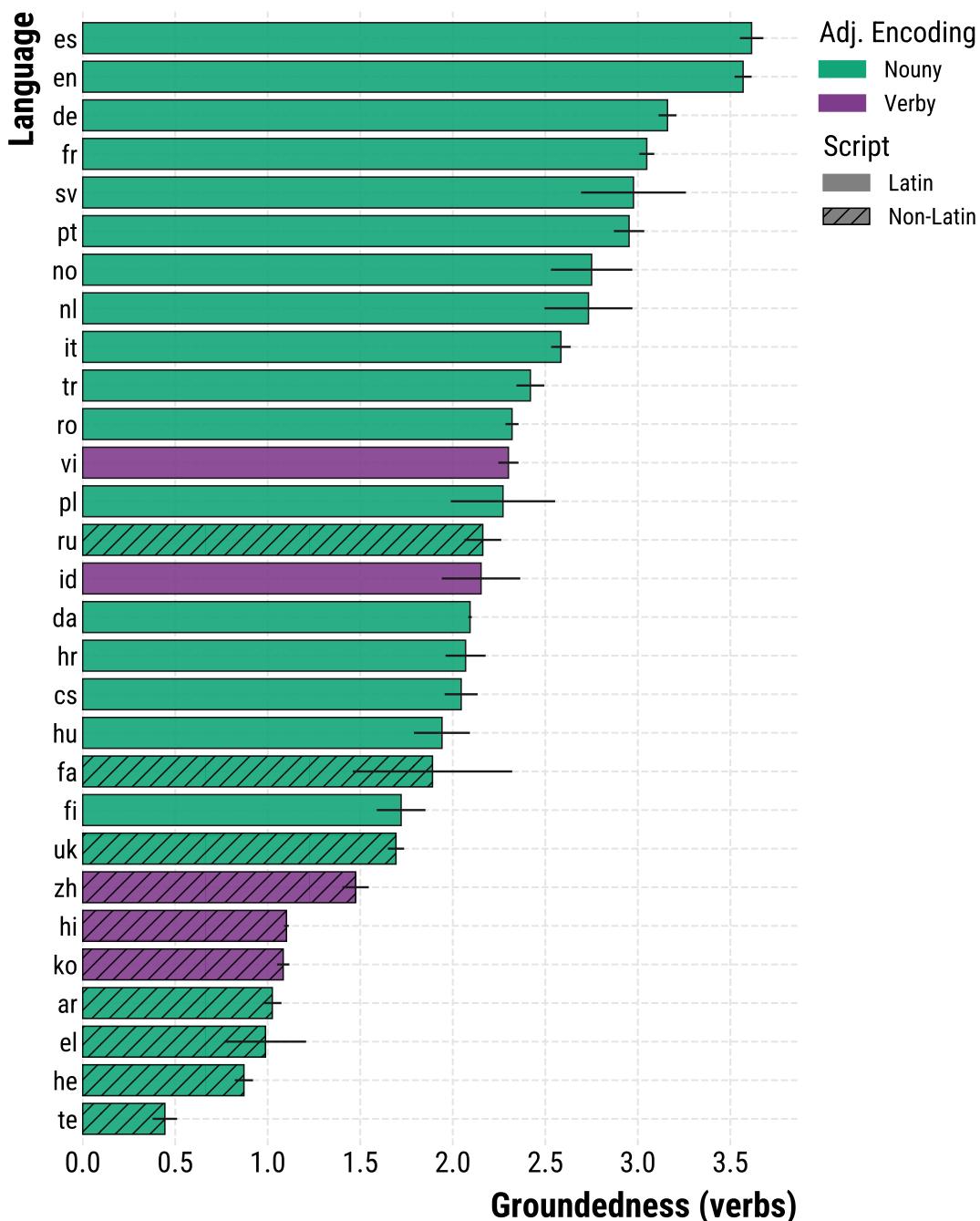


Figure 6.2: Groundedness of the verbal categories across the 30 languages in this study. Error bars represent standard error in the mean groundedness across the datasets considered (COCO-35L, XM3600, and Multi30K where available). Contra theoretical predictions, Verby languages do not exhibit higher mean groundedness of verbs, but are somewhat below average. However, this effect is confounded by model quality issues, as suggested by the lower groundedness of verbs in non-Latin script languages.

6.4.4 Results

Figure 6.2 shows the result of the absolute groundedness analysis for verbs across the 29 languages in the sample. I fit a mixed effect model with a random effect for dataset, and fixed effects for the quality ratio, word order (OV vs. VO), and script (Latin vs. non-Latin). This model supports a small effect of verby languages exhibiting *lower* verbal groundedness than nouny languages ($\beta = -0.22 \pm 0.11, p = 0.046$). However, we see a clear effect of quality ratio ($\beta = -0.27 \pm 0.05, p < 0.001$), and effects of script ($\beta = 0.34 \pm 0.11, p = 0.002$; Latin script) and word order ($\beta = -0.26 \pm 0.11, p = 0.027$; OV order). However, AIC does not support the inclusion of the verby/nouny variable (AIC: 61.17 with vs. 60.85 without; relative likelihood: 0.85). This is *counter* to the theoretical prediction that verby languages should have a *more* grounded verbal prototype than nouny languages. However, we find much stronger support for the negative effect of a language being written in a non-Latin script (AIC: -18.5), and find little remaining predictive effect after this simple heuristic for languages where the model struggles more (AIC: -20.5 with vs. -18.5 without). These results do not suggest that there is a difference in the average groundedness of the verbal prototype between nouny and verby languages.

Z-scored groundedness While the previous experiment did not show a clear difference between nouny and verby languages in terms of absolute groundedness after controlling for model quality, we might find an effect on the *relative* groundedness of verbs in these languages. Figure 6.3 shows the results of this analysis. A fixed effects model with the same model formula was applied. The model no longer supports an effect of the quality ratio ($\beta = -0.03 \pm 0.03, p = 0.24$), or word order ($\beta = -0.05 \pm 0.07, p = 0.23$; OV order). However, I observe a clear effect of script. ($\beta = 0.21 \pm 0.06, p < 0.001$; Latin script). This model does *not* support an effect of verby/nouny status ($\beta = -0.07 \pm 0.06, p = 0.24$). So I find

no evidence that verby languages have relatively more grounded verbs than nouny languages.

Including “Adjective” tags in the verbal prototype One possible cause for the lack of a groundedness shift in verby languages is the “loss” of more stative verbal meanings to the the ADJ UPOS tag. To provide an upper bound on verbal groundedness in these languages, I combined the ADJ and VERB UPOS tags for these languages only. Verby languages do show an increase in groundedness, with Indonesian and Hindi shifting up in the ranking—in line with the higher average groundedness of adjectives. Nevertheless, the relative groundedness of the verby languages is still less overall than the most grounded nouny languages, and I still observe an association with script—notably, now the two most grounded verby languages are exactly those two which use a Latin script. Replicating the same mixed effects analysis (with random effect for dataset) on this data, nevertheless I still only find a significant effect of script ($\beta = 0.22 \pm 0.06, p = 0.001$; Latin script), and no effect of verby/nouny status ($\beta = -0.04 \pm 0.06, p = 0.51$). Our results, then, do not suggest that the lack of a groundedness shift in verby languages is due to misclassification of parts of speech.

6.4.5 Discussion

Across all three experiments, we find no evidence that verby languages have a more grounded verbal prototype than nouny languages, contrary to the theoretical predictions based on the Tensedness Correlation. Instead we find evidence that where models struggle more (as approximated by non-Latin script and the ratio of captioning to language model NLL), verbs are *less* grounded both in absolute and relative terms. This effect persists even when adjectives are included in the verbal prototype for verby languages, suggesting that the lack

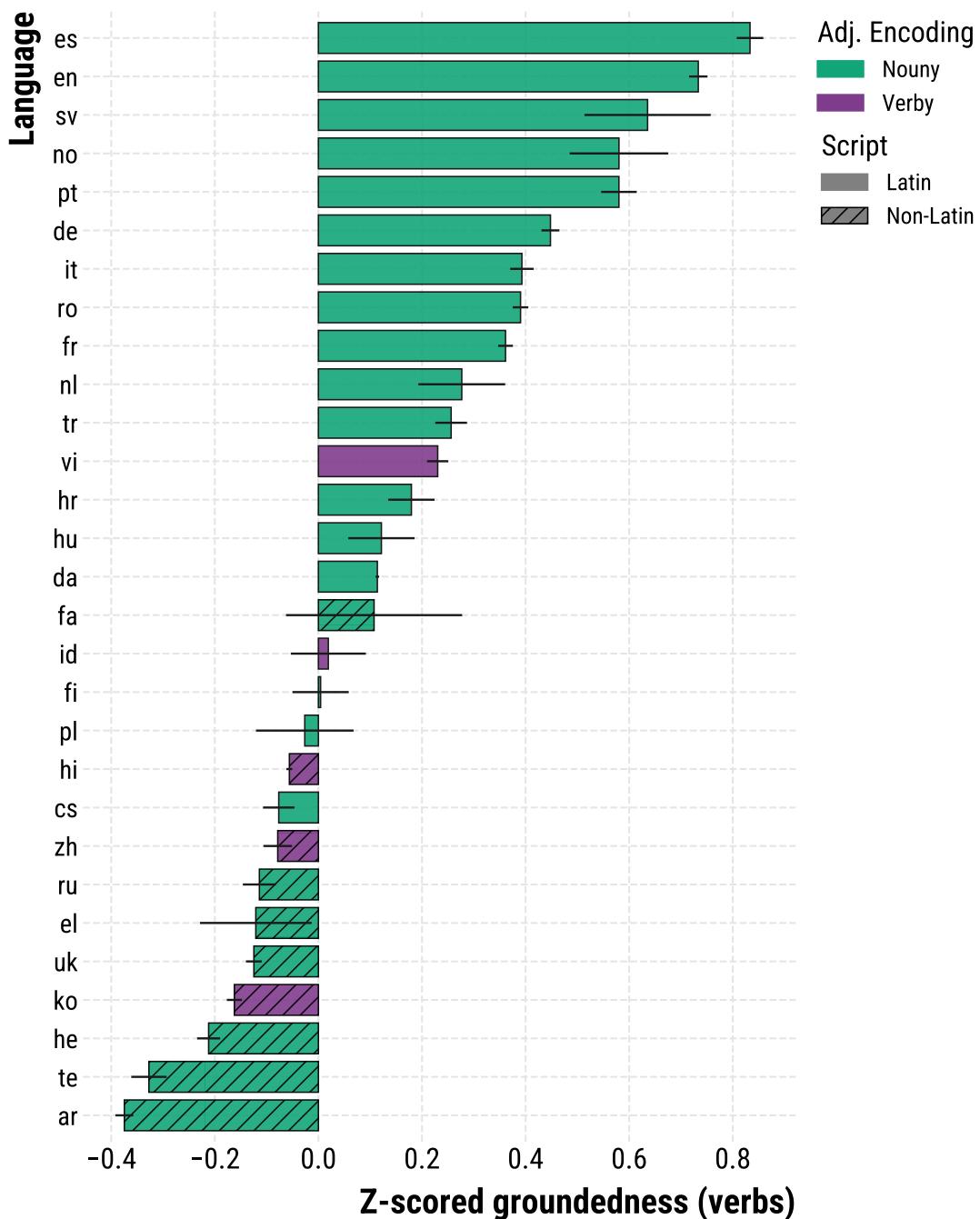


Figure 6.3: Z-scored groundedness of the verbal categories. Error bars represent standard error in the Z-scores across the datasets considered (COCO-35L, XM3600, and Multi30K where available). The results suggest verbs are not *relatively* more grounded than other words in verby languages. However, we observe a clear effect of script, with languages written in Latin script exhibiting relatively more grounded verbs.

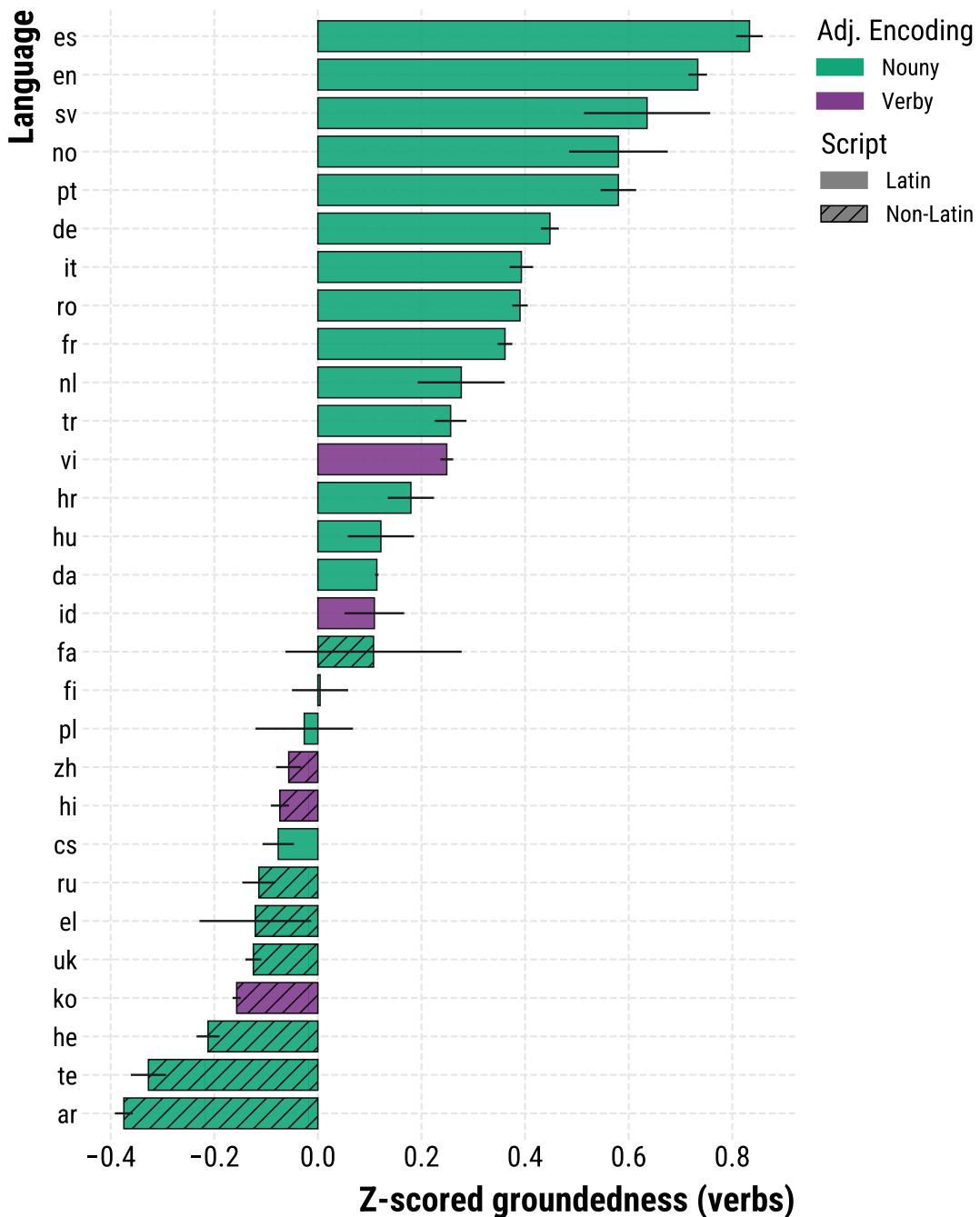


Figure 6.4: Z-scored groundedness of the verbal categories, with adjectives included for verby languages. Error bars represent standard error in the Z-scores across the datasets considered (COCO-35L, XM3600, and Multi30K where available). Despite the higher groundedness of adjectives than verbs in general, and concerns that legitimate members of the verbal category could be disproportionately “lost” to the adjective tag in verby languages, we still observe lower groundedness for the verby languages. This suggests an disproportionate effect of captioning and language model quality on verbs.

of a groundedness shift for verby languages is not due to incomparable part of speech tagging.

Rejecting the explanation that I have “lost” genuine verbs to the ADJ tag, I am left with evidence that verbs are “more difficult” to ground–poor models have a greater effect on verbs than at least some other parts of speech. This seems plausible—the exact factors identified as candidates for their lower groundedness than nouns (spatial diffuseness, temporal instability, relationality) suggests that learning verbs from images is harder than nouns, and so may take a bigger hit when models are poor. In the face of results which seem to pattern with orthography and training data availability, I cannot draw a strong conclusion about whether an estimate of visual groundedness computed with more comparable corpora would show correlations with tensedness and verby encoding of adjectives.

This highlights a key limitation of the present study: the lack of large-scale multi-parallel image captioning datasets. While such datasets represent a gold-standard for cross-linguistic comparison, they are very expensive to create, and difficult-to-impossible to extend to the “long tail” of the world’s languages. As such, I caution future studies to be mindful of the confounding effects of model quality when comparing groundedness estimates across languages. This is not to say that we cannot explore typological questions with existing models. The other groundedness analyses in this thesis have been carefully designed to avoid direct cross-linguistic comparison of groundedness scores. In Section 6.3, we compared groundedness within a single language, while in Chapter 5, we fit a statistical model of cross-linguistic trends *within* languages. While these approaches certainly constrain the kinds of questions we can ask, I believe there are still many exciting avenues for typological research on groundedness that can be pursued with existing data and model.¹¹ Further, improvements in

¹¹In Chapter 7 I lay out a number of such directions.

multilingual vision–language pretraining may help alleviate some of the model quality issues I have observed here.

6.5 Conclusion

In this chapter, I have argued that groundedness and meaning content are operant in grammatical organization not only across the lexical–functional divide, but also among the lexical classes of noun, adjective, and verb themselves. Drawing on the literature from cognitive linguistics on the continuum and prototype structure of lexical classes, I demonstrated that some aspects of this continuum are interestingly similar to the lexical–functional distinction. In particular, nouns, adjectives, and verbs vary in their prototypical *relationality*, with nouns being the least relational and verbs the most relational. Functional elements are also more relational than lexical elements. In so doing, I propose the study of a unified *lexicality spectrum*, which connects variation between lexical classes to functional classes.

Building on the work in Chapter 5, I have used groundedness as a computational measure of this relationality dimension. I argue that groundedness has the potential to combine dimensions like time-stability and spatial compactness into a single information-theoretic measure. I then aimed to show that groundedness can provide new evidence about variation in lexical class organization cross-linguistically.

Focusing first on Japanese, which has a well-studied “split” among its adjectives which has long been argued to be synchronically arbitrary, I showed that the two adjective classes differ in their groundedness when Japanese speakers chose how to describe images. The more formally noun-like *na*-adjectives are more grounded than the more verb-like *i*-adjectives, suggesting that the split still encodes a synchronic lexicality distinction.

Finally, I investigate “lumping” behaviour between the lexical classes through the lens of the Tensedness Correlation, which links obligatory tense marking to nouny encoding of adjectives. Building on cognitive-linguistic theories of this correlation, I proposed that the correlation reflects a shift in the verbal prototype away from nouns in tensed languages in terms of groundedness. However, I found no evidence for this hypothesis in a cross-linguistic comparison of visual groundedness of verbs in 29 languages. Instead, I found that verbs are less grounded in languages where the captioning and language models struggle more, suggesting that verbs are more difficult to ground than other parts of speech. This highlights the challenges of direct cross-linguistic comparison of groundedness estimates.

These results are nevertheless promising initial evidence for the role of groundedness and relationality across the whole lexicality spectrum, including up into the lexical classes themselves. Future work should continue to explore these connections, especially in more split-class languages, and with new measures and datasets.

Chapter 7

Conclusion

In this thesis, I have argued for an empirical, quantitative, and computational approach to defining and studying comparative concepts in linguistic typology. Drawing on the history of empirical and computational approaches in phonological and semantic category typology, I have proposed that developments from deep learning can take on an analogous role for other functional comparative concepts to the empirical measures that allowed these problems to be computationally modelled, defining an empirical conceptual space over which we can form generalizations.

I apply this idea to the notion of a spectrum of *lexicality*—a correlation between the formal size and independence of linguistic signs, and their semantic contentfulness and relationality. The notion of semantic contentfulness, while often appealed to in intuitions about putative binaries like the lexical–functional distinction or the inflection–derivation distinction, has been difficult to formalize, with traditional linguistic approaches not offering a principled basis for measuring it, and conventional computational approaches reducing to frequency. Partly as a result, these binaries have been argued by some to hold no theoretical or empirical weight (Haspelmath, 2024; Bruening, 2018). Deep learning offers a new way forward, by providing high-dimensional semantic representations

that can be used to operationalize semantic contentfulness in new ways.

I have provided computational studies at three different “levels” of lexicality. Part I provides a detailed study of the inflection–derivation distinction, Chapter 5 studies the lexical–functional distinction, and Chapter 6 studies distinctions between the major lexical classes of nouns, verbs, and adjectives. In each case, I have shown that deep learning-based measures reveal consistent cross-linguistic patterns across these problematic distinctions. This suggests that these distinctions have been relatively consistently applied cross-linguistically, even if they are gradient and fuzzy.

7.1 Contributions

To summarize, the main contributions of this thesis are:

1. Argument comparative concepts
2. A review of comparative concepts in computational typology
3. Argument lexicality
4. A four measure framework for
5. Classification experiments demonstrating
6. Groundedness
7. Evaluation of relationship between groundedness and lexical–functional
8. A theoretical argument relating the groundedness findings to lexical
9. An empirical study of a language that splits a major lexical class
10. Tensedness Correlation

7.2 Limitations and future directions

Language coverage

From a typological perspective, probably the most important limitation of the work here, and to a certain extent the methods proposed, is the language coverage.

Distinction operationalizations

Throughout this thesis, I have related my empirical measures to distinctions as they have been used in existing tools and datasets. For example, in Part I, I study the relationship between my four measures and the inflection–derivation distinction *as reflected in the UniMorph 4.0 dataset* (Batsuren et al., 2022), and in Chapter 5, I study word classes *as operationalized in Universal Dependencies* (de Marneffe et al., 2021). A natural objection to certain findings here, then, is that I have selected the *wrong* operationalization of a particular distinction, and that if I had grouped constructions differently I would have seen different results (e.g. a more clear-cut distinction between inflection and derivation). This is certainly a limitation of the present work, but it is not necessarily a weakness.

For example, the inflectional vs. derivational status of a construction in UniMorph is based on how Wiktionary editors for a language choose to organize morphological information. In this way, it directly reflects the most cynical interpretation of Haspelmath (2024) in describing inflection and derivation as *traditional comparative concepts*—that it is, effectively, a convention of dictionary organization. That we find a high degree of cross-linguistic consistency in terms of our measures is thus all the more surprising given the lack of linguistic attention paid to the classification of constructions. Further, the choice of using the categories unmodified from UniMorph or Universal Dependencies is motivated by the high likelihood that these operationalizations will be *used* as the

comparative concepts in future computational studies—this thesis thus serves as a form of external empirical *validation*. All this being said, an interesting direction for future work would certainly be to try different operationalizations of the distinctions with the approaches in this thesis, contrasting them with the findings here. Indeed, this direction is actively being pursued by researchers investigating how different ways of grouping derivational constructions influence the empirical differences between inflection and derivation in my framework (Kyzánek and Bonami, 2025).

Data comparability

The classification approach

Questions of groundedness

Images as meaning(?)

7.3 Finis

Appendix A

Chapter 4: Results using Word2Vec embeddings

As discussed in Sections 3.3.2 and 4.3.1, our study’s reliance on FastText vectors is potentially problematic for dissociating the contribution of form from syntax and semantics, because FastText includes distributional sub-word information that could cause distance metrics in embedding space to measure formal and not just word-level distributional similarity. While we still use their vectors for the benefits they provide in terms of quality, reliability, and representation of rare and morphologically complex forms, we here compare performance of FastText vectors to Word2Vec vectors, which do not have this sub-word level distributional information. As such, the Word2Vec results presented here can be seen as a lower bound on how much semantic/syntactic factors explain the performance of our FastText results.

Unlike FastText, a large multilingual pre-trained set of Word2Vec vectors does not exist. As a result, we must train our own Word2Vec representations, for which we use the 2021 dump of the OSCAR corpus (Abadji et al., 2022). Resource constraints prevent us from training on over 10 GB of data for any given language, significantly less than many of the languages in our dataset

were trained on for FastText. We use as much data as was available (up to 10 GB) per language and train for 5 epochs. Due to this limited amount of data, the performance of these vectors is not directly comparable. Further compounding this issue is one of vocabulary; the precise minimum count used in FastText models is unclear, but appears to be greater than the default 5, which we used for our models, judging by the ratio of vocabulary size to data size. As such, our Word2Vec vectors may contain more lower-quality vectors. Finally, due to differences in vocabulary, we must exclude some of our training and testing sets. We train and evaluate both Word2Vec and FastText MLP classifiers on only the portions of the train and test sets which are in-vocab for both models. This naturally leads to somewhat worse performance for the FastText-based MLP than the results in the paper.

From the set of results in Figure A.1, we do see non-trivial contributions of word-level distributional information, particularly for the V_{Embed} measure, even with this lower bound, indicating our FastText measure captures some semantic and syntactic information.

Features						Accuracy (\square = Logistic, \blacksquare = MLP)
Majority class (Inflection)						$\blacksquare 0.58 \pm 0.0$
M_{Form}	-	-	-	-		$\blacksquare 0.61 \pm 0.0$ $\square 0.61 \pm 0.0$
-	M_{Embed}	-	-	-		$\blacksquare 0.62 \pm 0.0$ $\square 0.70 \pm 0.0$
-	-	V_{Form}	-	-		$\blacksquare 0.71 \pm 0.0$ $\square 0.71 \pm 0.0$
-	-	-	V_{Embed}	-		$\blacksquare 0.70 \pm 0.0$ $\square 0.70 \pm 0.0$
(A)	-	-	V_{Form}	V_{Embed}		$\blacksquare 0.76 \pm 0.0$ $\square 0.80 \pm 0.0$
(B)	M_{Form}	M_{Embed}	-	-		$\blacksquare 0.60 \pm 0.0$ $\square 0.67 \pm 0.00$
(C)	M_{Form}	-	V_{Form}	-		$\blacksquare 0.75 \pm 0.0$ $\square 0.75 \pm 0.0$
(D)	-	M_{Embed}	-	V_{Embed}		$\blacksquare 0.71 \pm 0.0$ $\square 0.73 \pm 0.0$
(E)	M_{Form}	M_{Embed}	V_{Form}	V_{Embed}		$\blacksquare 0.79 \pm 0.0$ $\square 0.85 \pm 0.0$

Figure A.1: Accuracy in reconstructing UniMorph’s inflection–derivation distinction by MLP classifiers using Word2Vec- vs. FastText-based distributional features. Hypotheses referred to in the main text are denoted with letters.

Appendix B

Part II: Model performance by language

See Table B.1 for per-language captioning performance, part-of-speech (POS) tagging accuracy, and perplexity of the base Gemma-2B model, the PaliGemma captioning model, and our fine-tuned language model (LM).

Table B.1: 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.

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

Table B.1: Per-language performance metrics for the models used (continued).

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	
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

Table B.1: Per-language performance metrics for the models used (continued).

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	
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

Appendix C

Groundedness correlation plots for other psycholinguistic norms

Figure C.1 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.

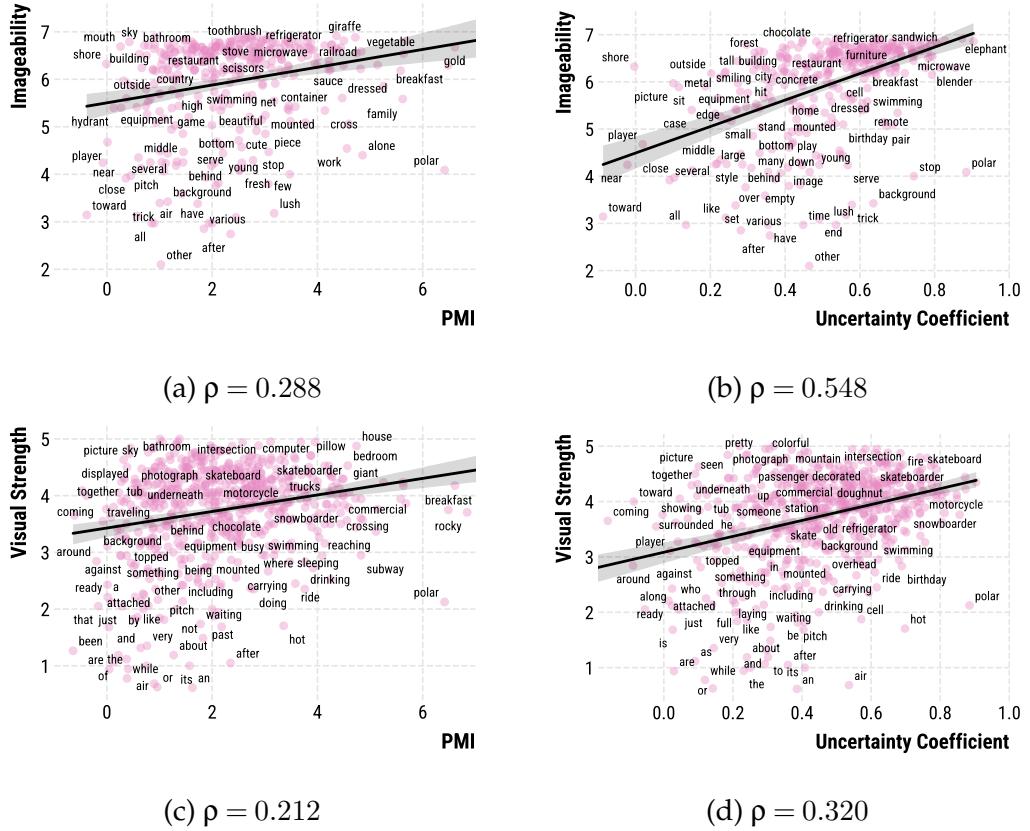


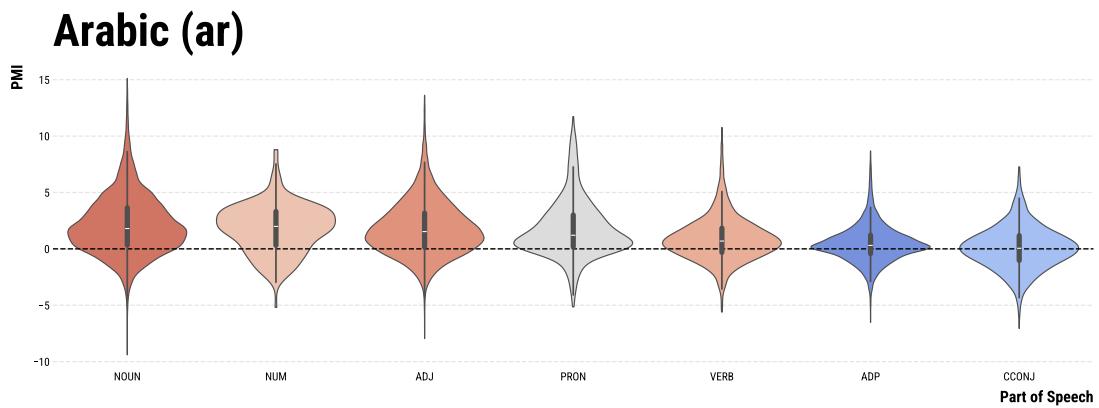
Figure C.1: Correlation between English psycholinguistic norms and type-level groundedness (left) or uncertainty coefficient (right): i.e., the average ratio between LM surprisal and captioning model surprisal. Type-level measures were computed by averaging scores across the COCO-dev dataset for types which occur at least 30 times.

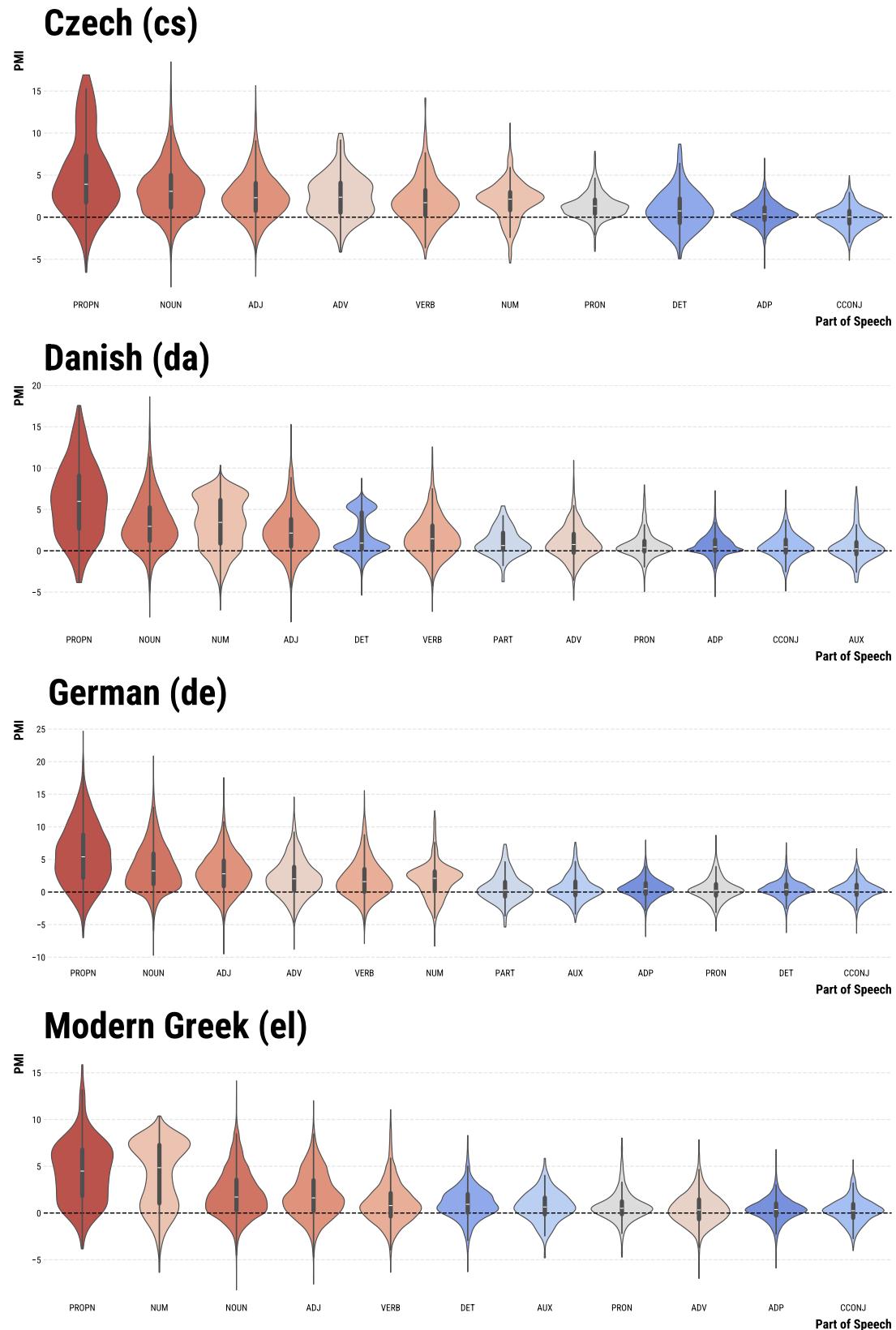
Appendix D

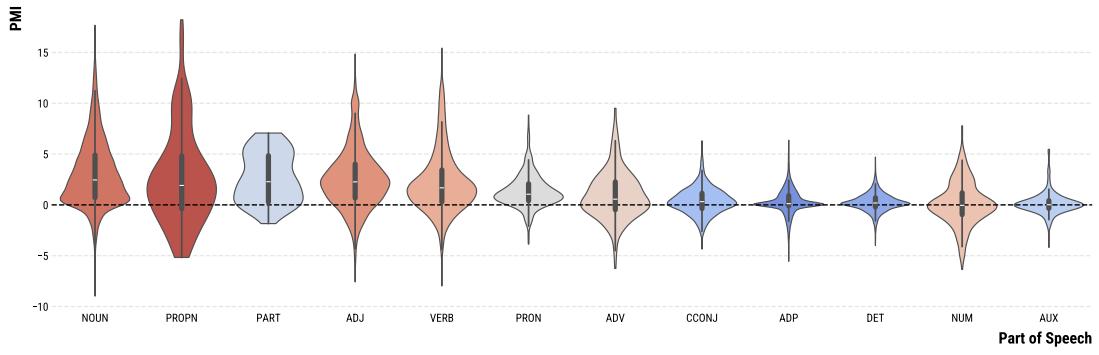
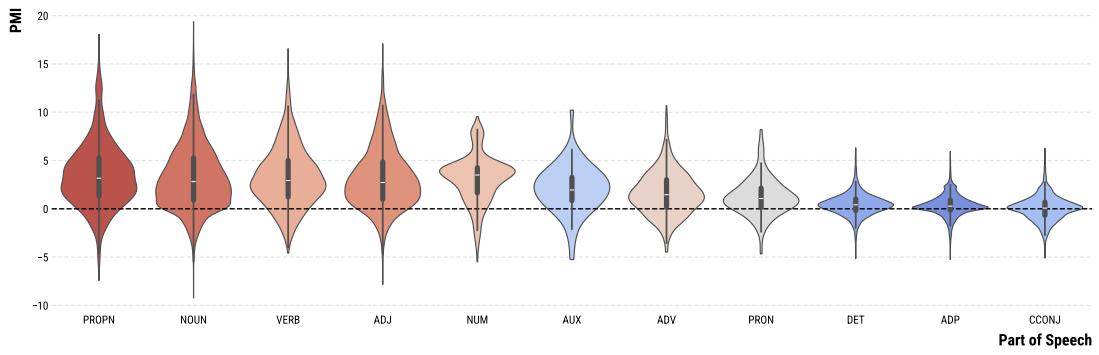
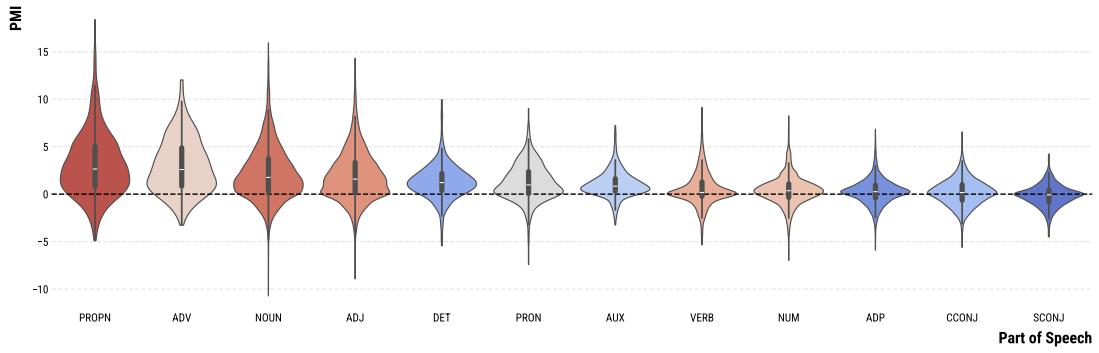
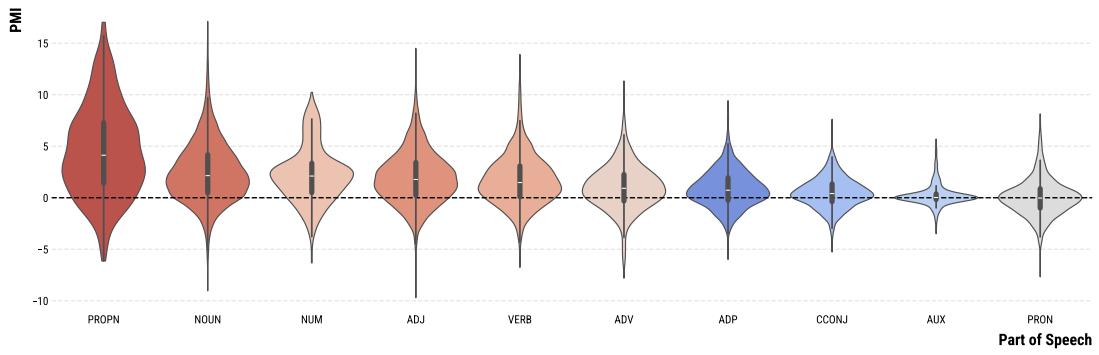
Groundedness distributions by language and dataset

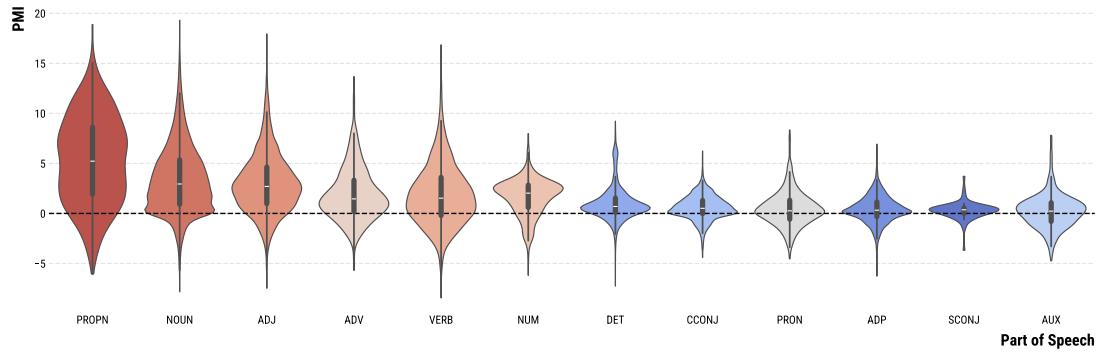
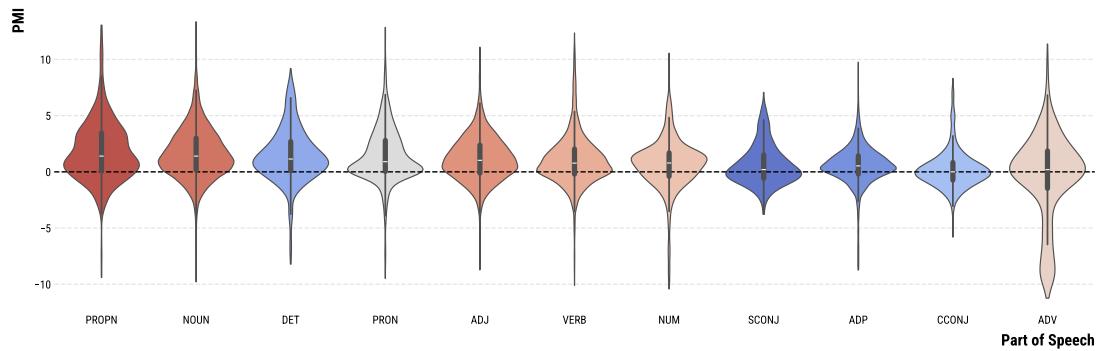
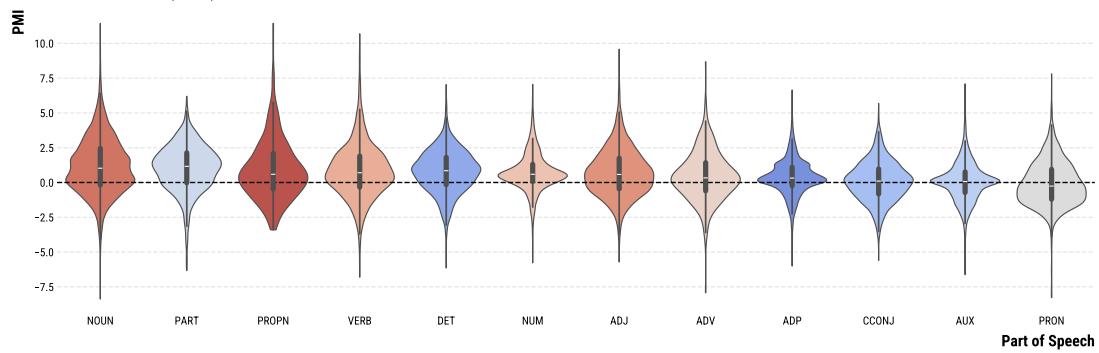
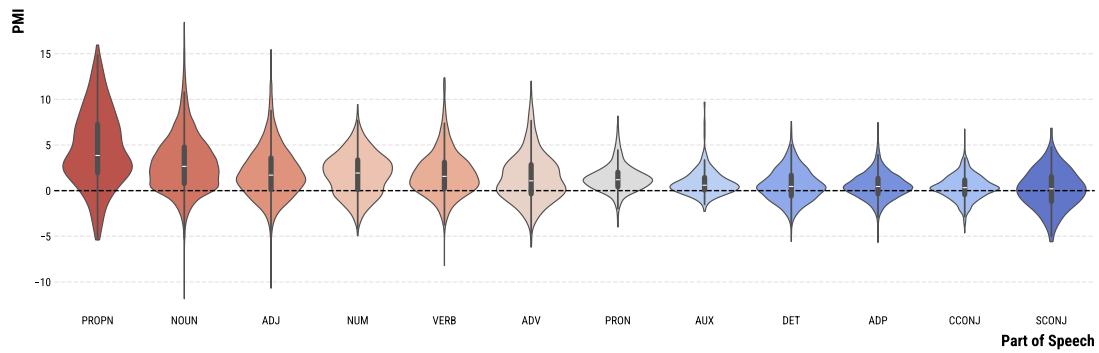
D.1 Crossmodal-3600

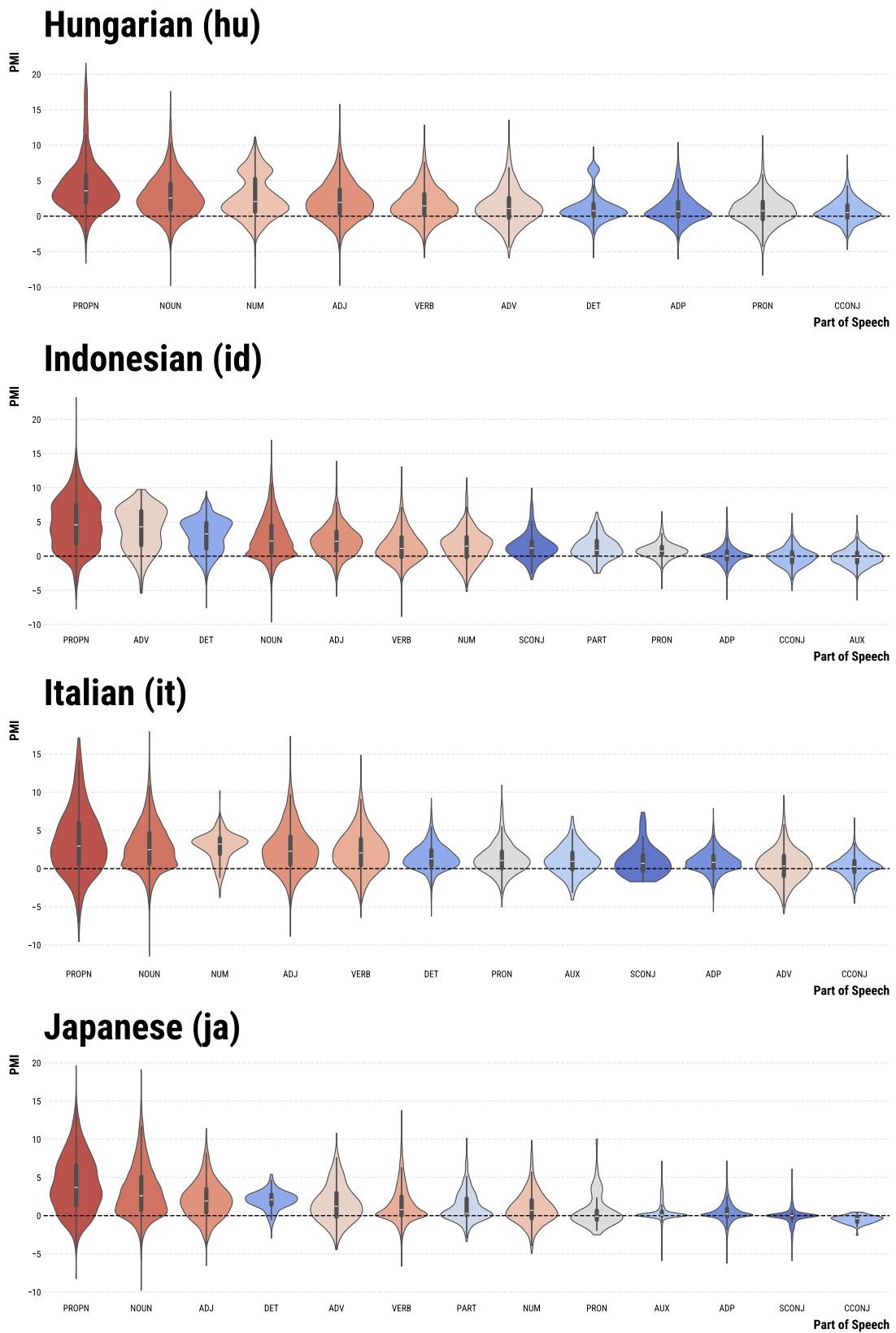
Results are ordered by descending mutual information estimate within the dataset (average groundedness/PMI). Hue indicates the average cross-linguistic ranking of a part of speech.

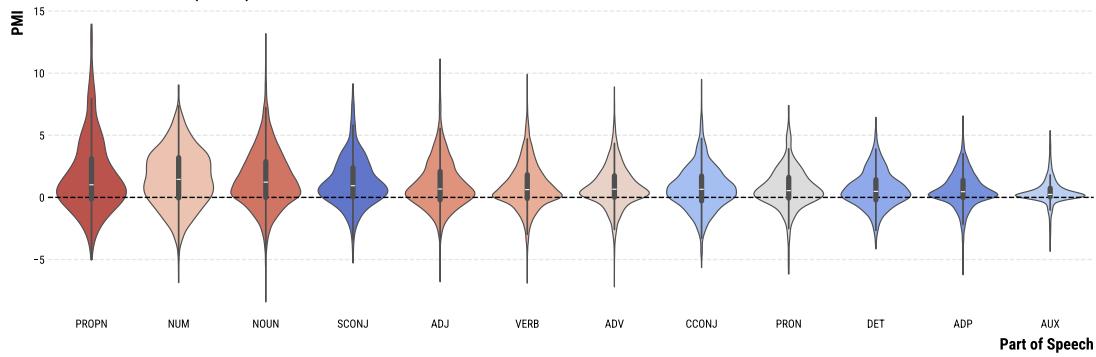
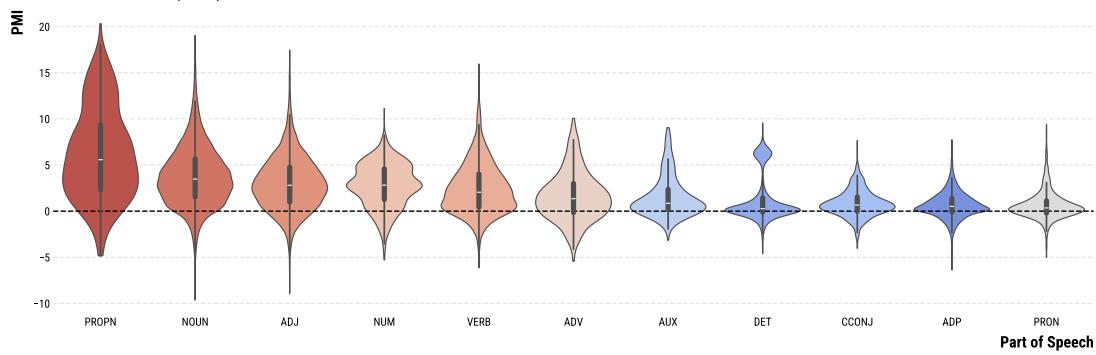
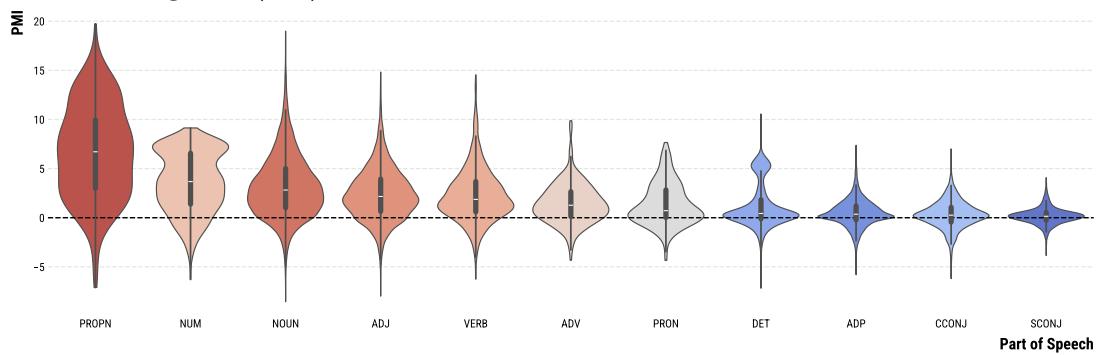
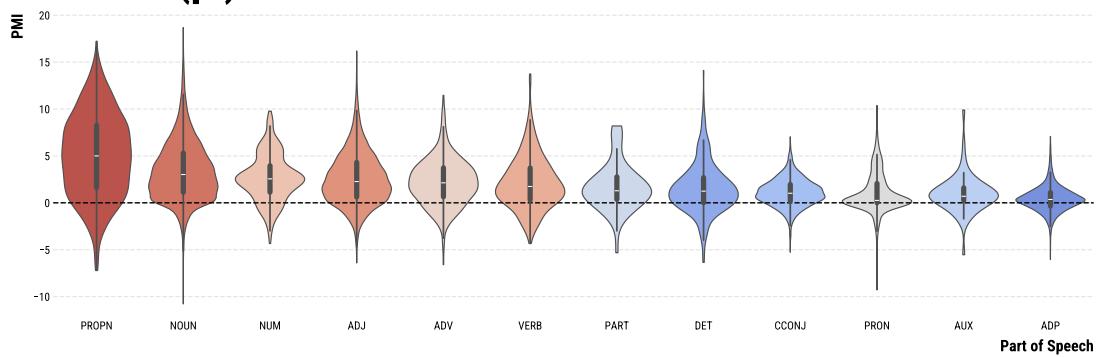


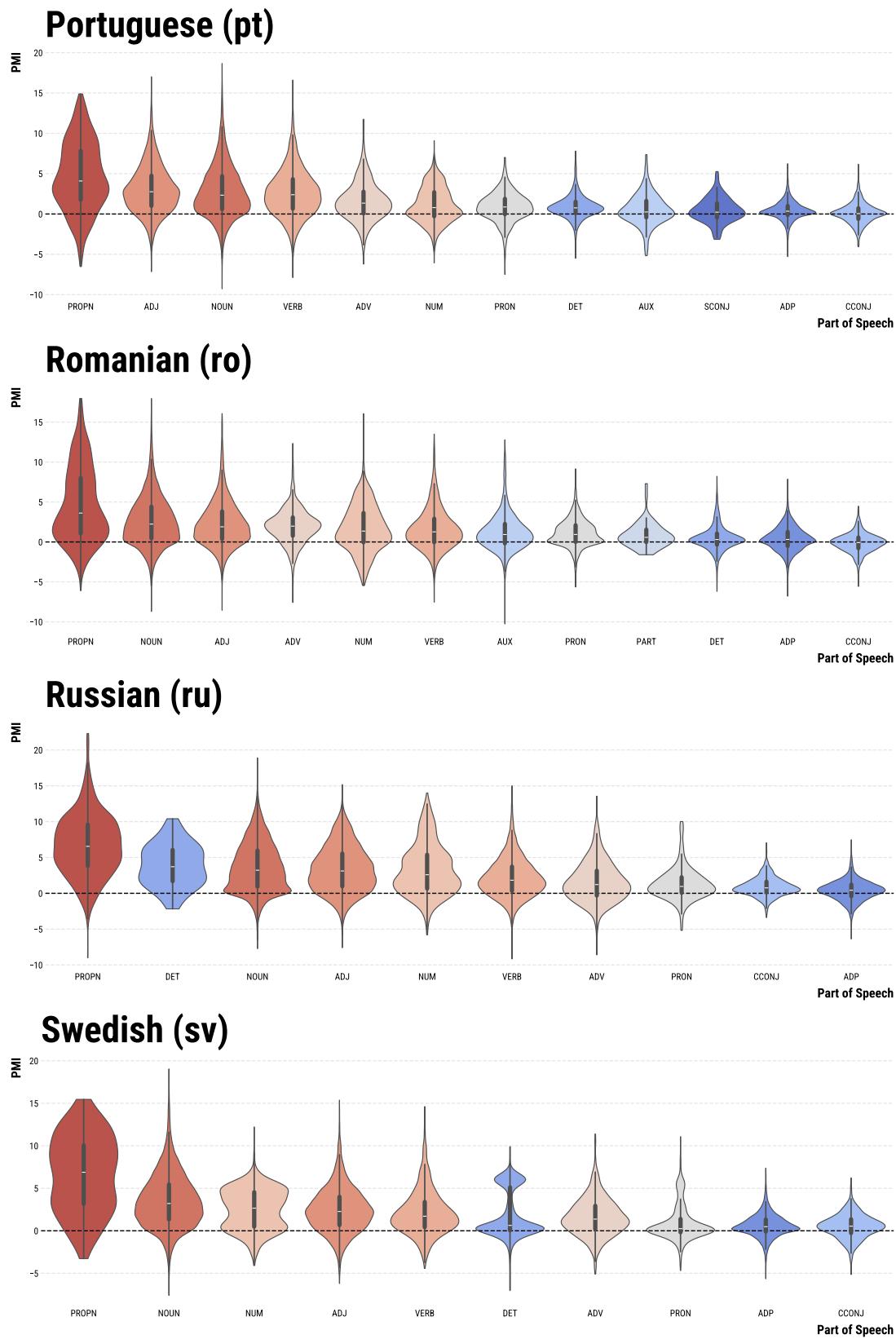


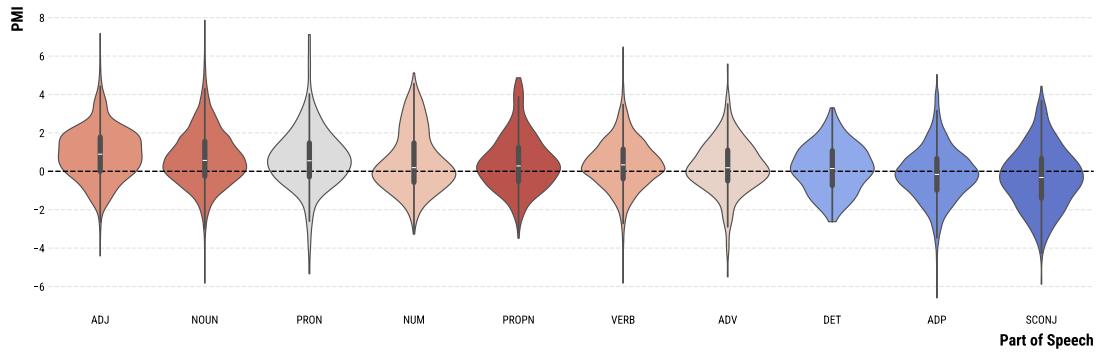
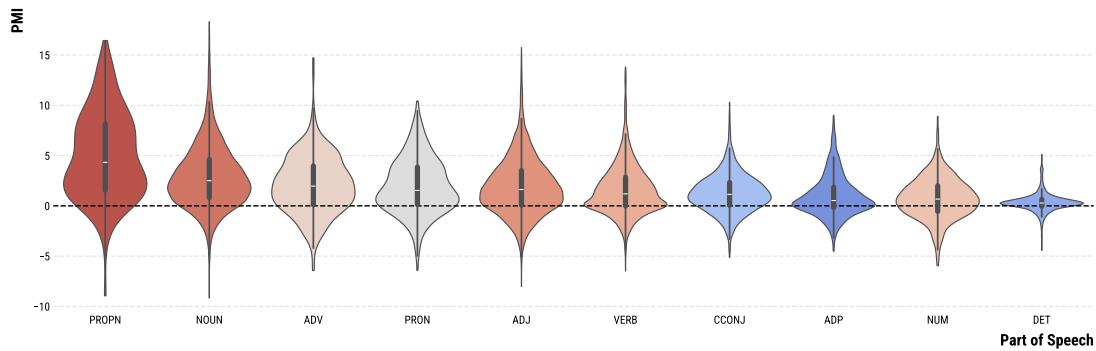
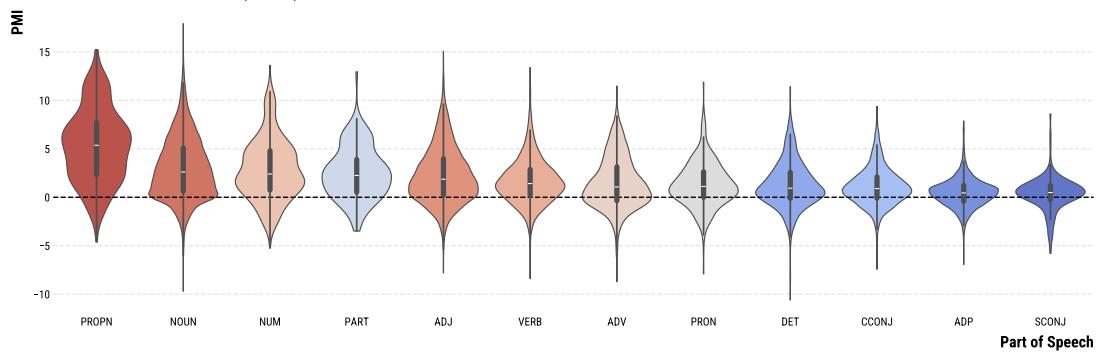
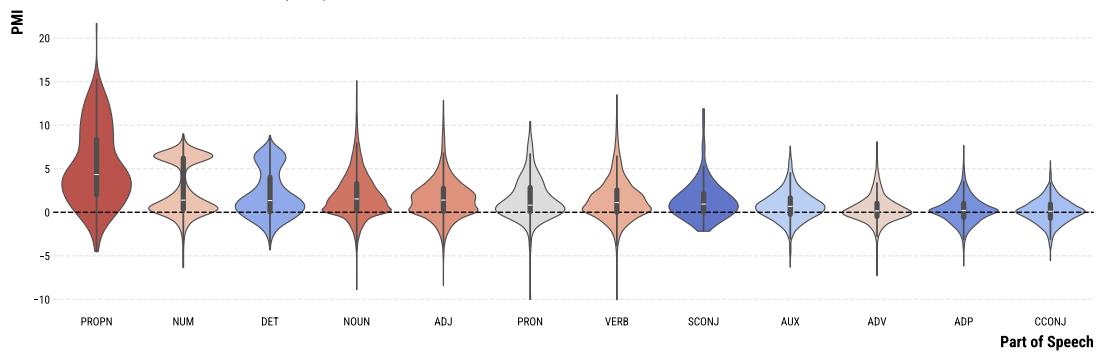
English (en)**Spanish (es)****Persian (fa)****Finnish (fi)**

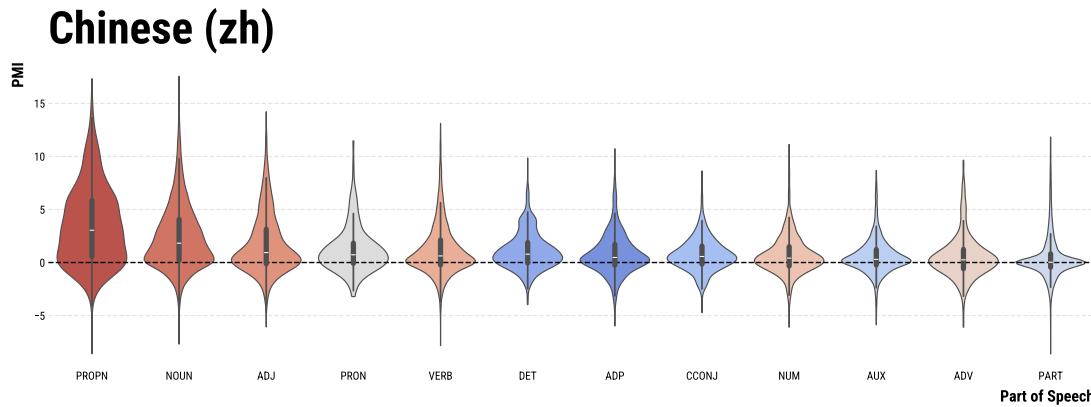
French (fr)**Hebrew (he)****Hindi (hi)****Croatian (hr)**



Korean (ko)**Dutch (nl)****Norwegian (no)****Polish (pl)**

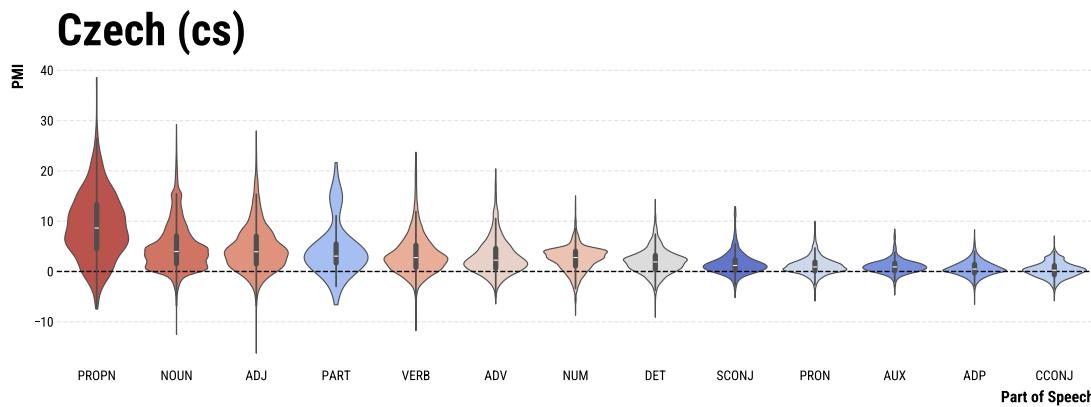
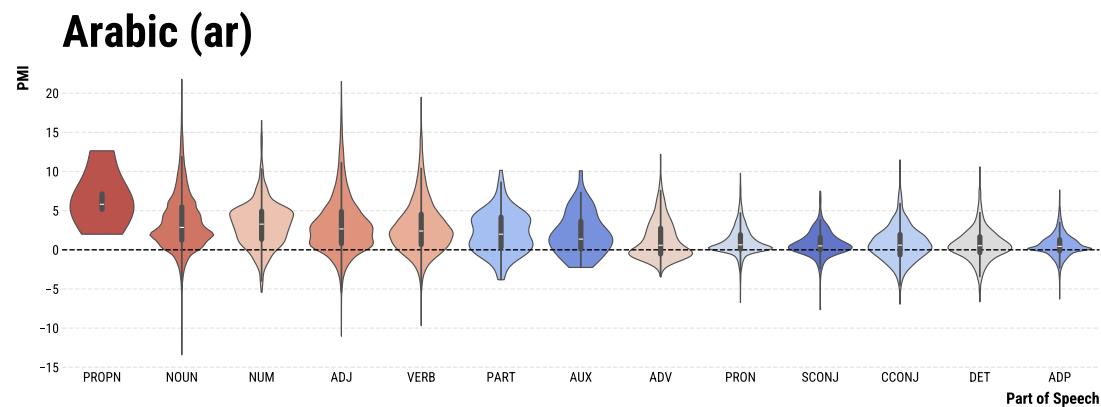


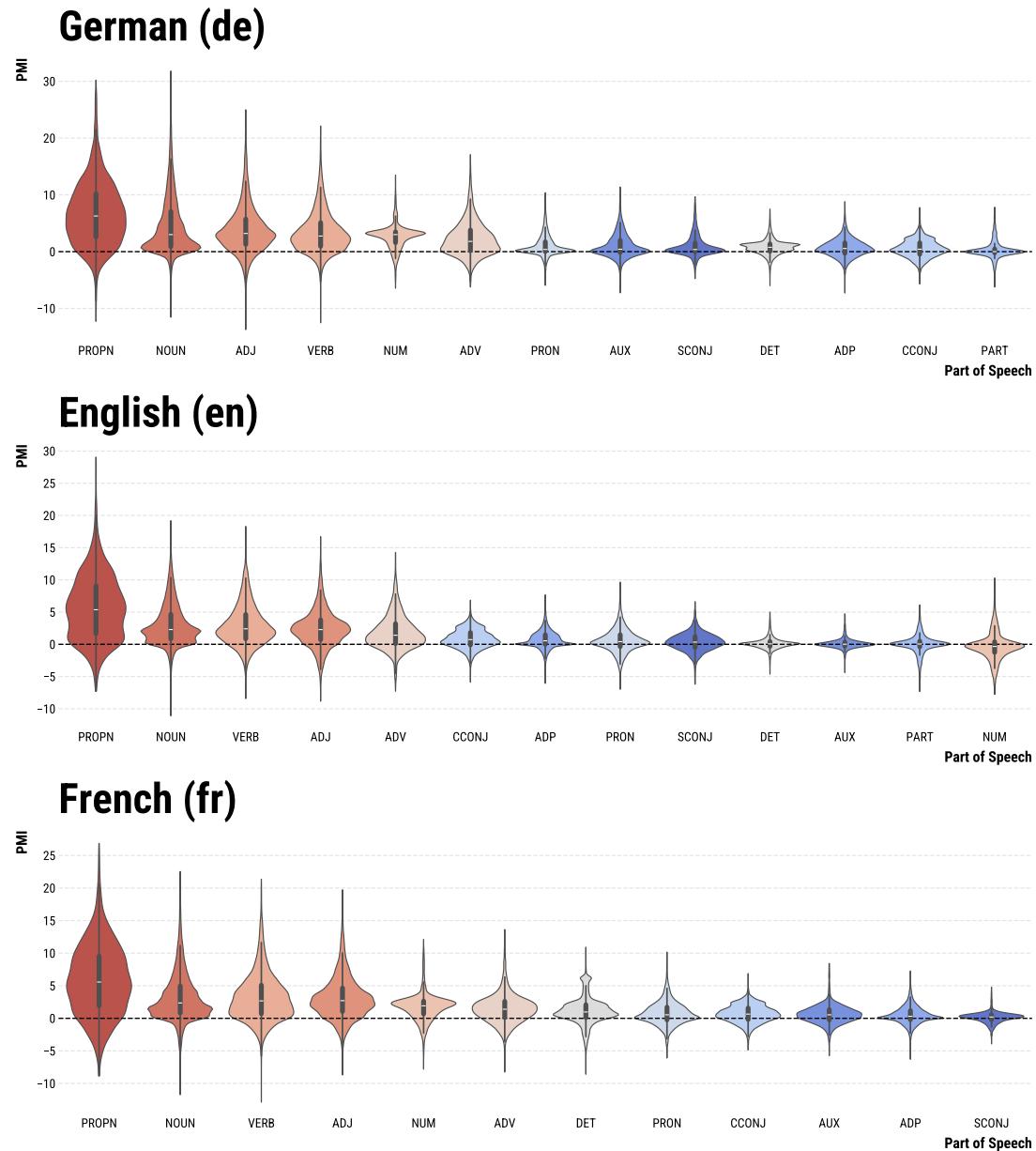
Telugu (te)**Turkish (tr)****Ukrainian (uk)****Vietnamese (vi)**



D.2 Multi30K

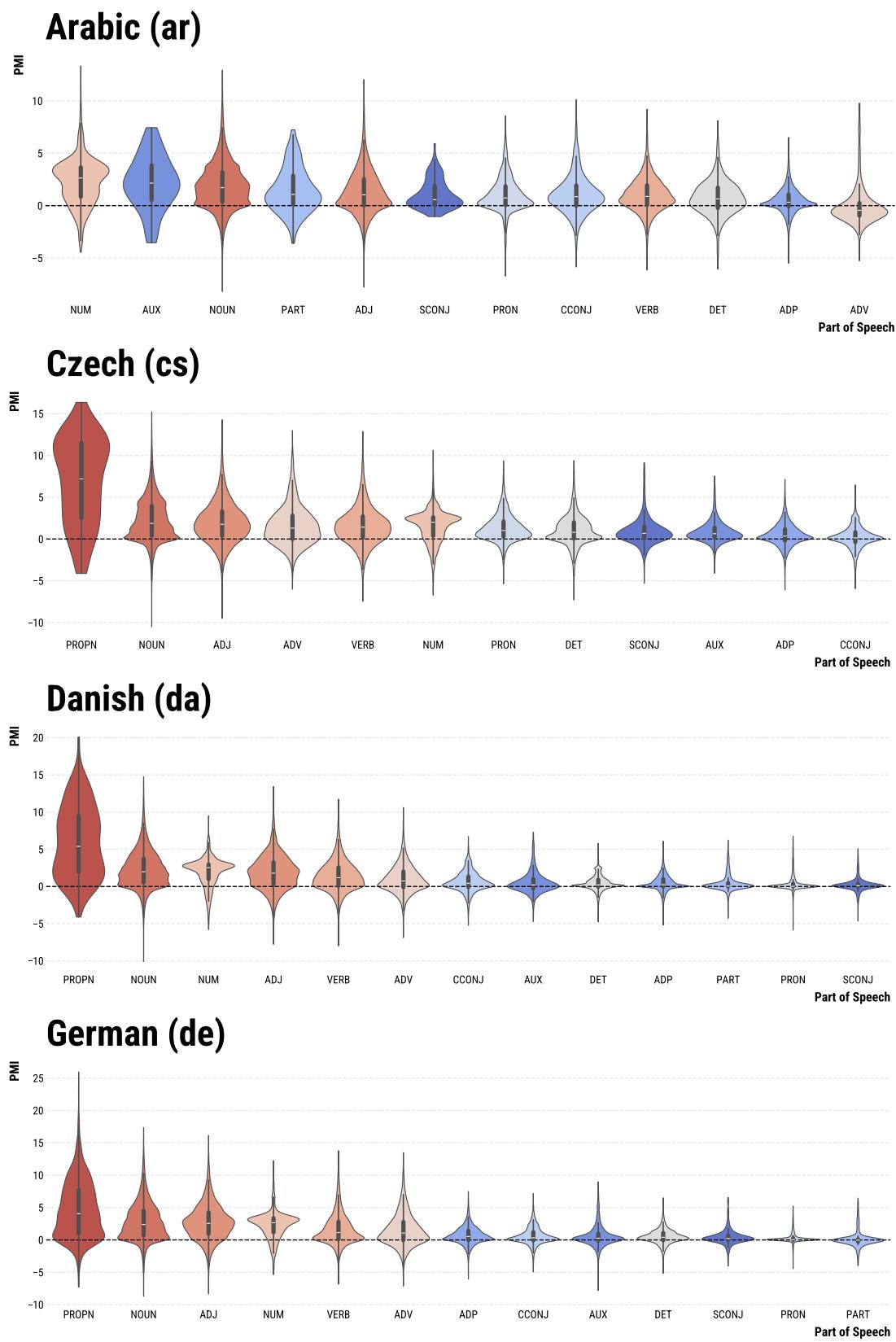
Results are ordered by descending mutual information estimate within the dataset (average groundedness/PMI). Hue indicates the average cross-linguistic ranking of a part of speech.



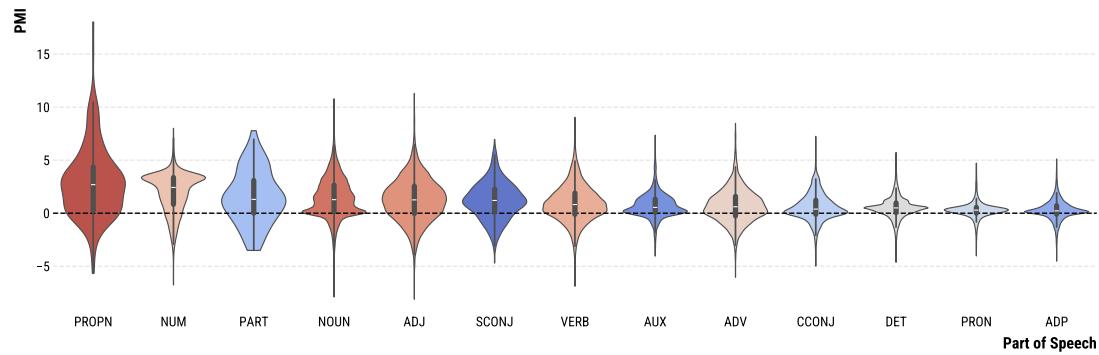


D.3 COCO-35L Development Set

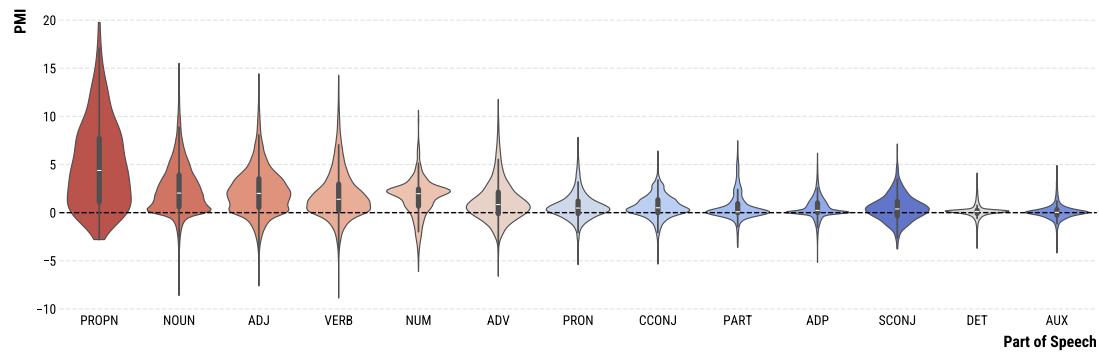
Results are ordered by descending mutual information estimate within the dataset (average groundedness/PMI). Hue indicates the average cross-linguistic ranking of a part of speech.



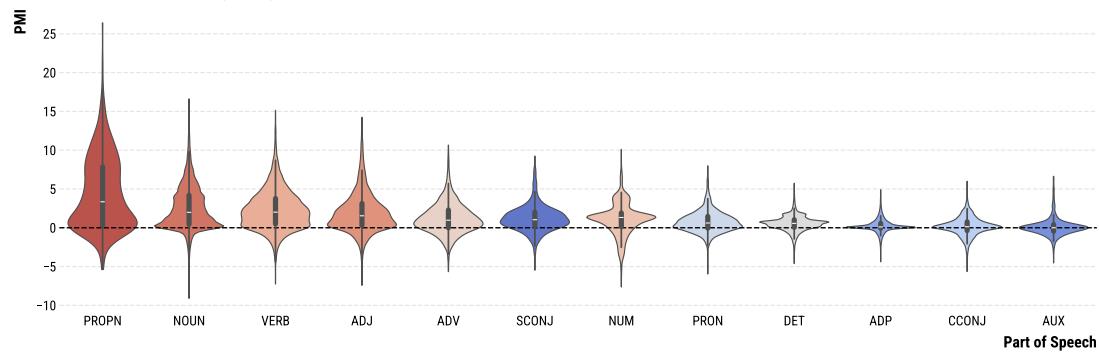
Modern Greek (el)



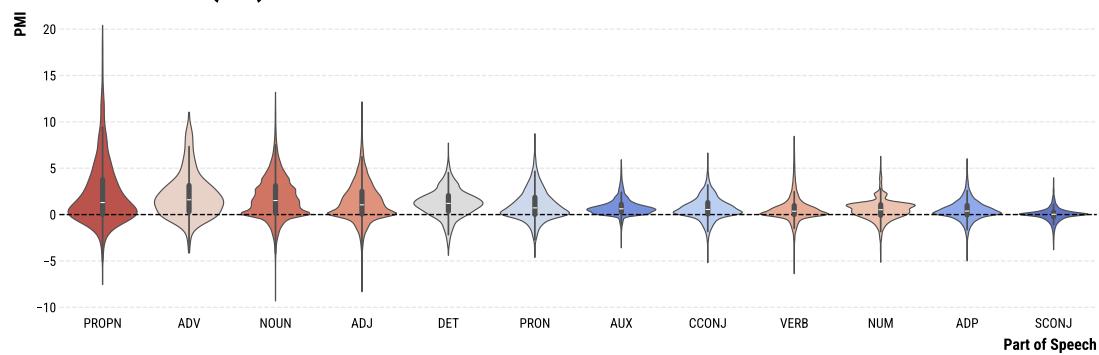
English (en)

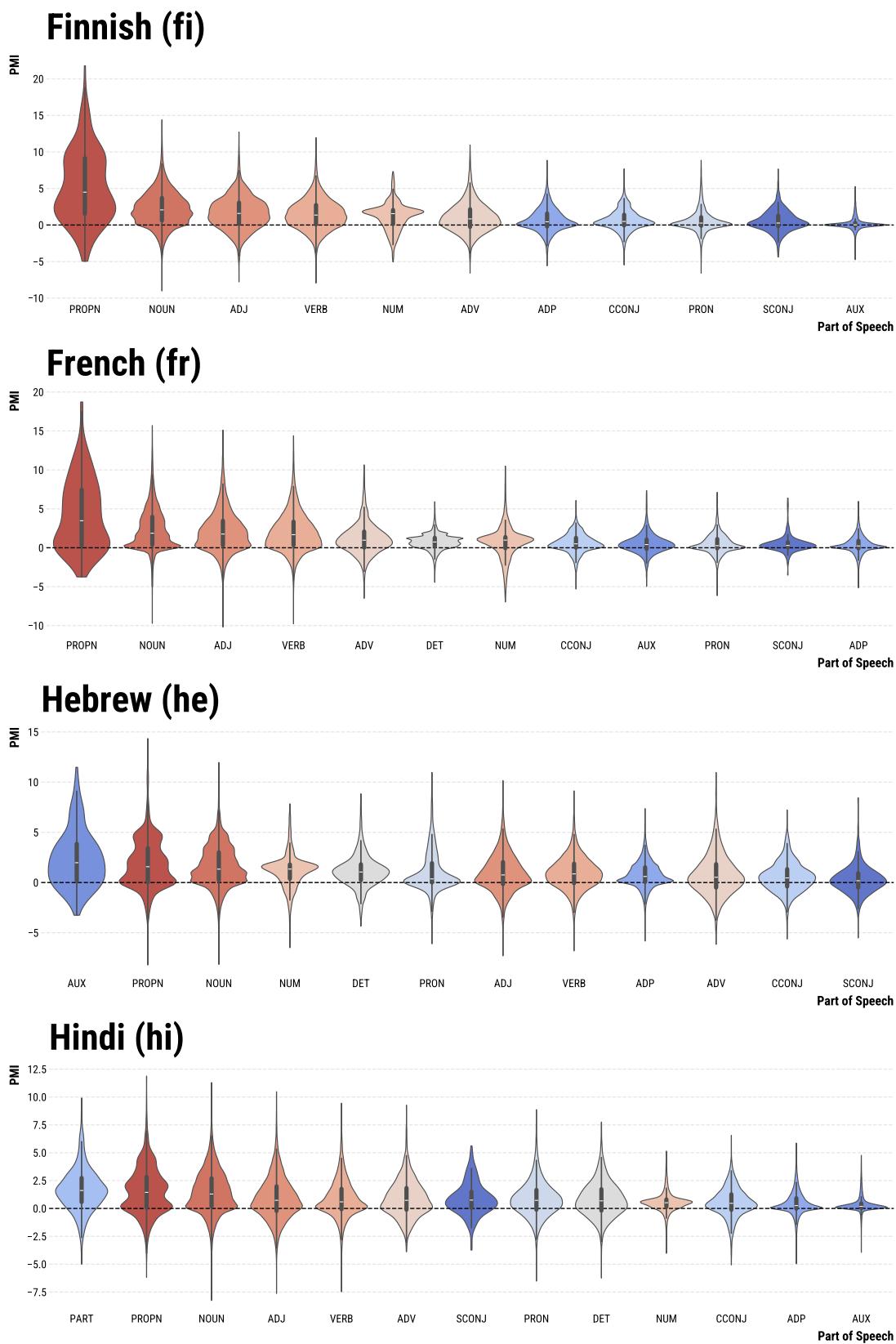


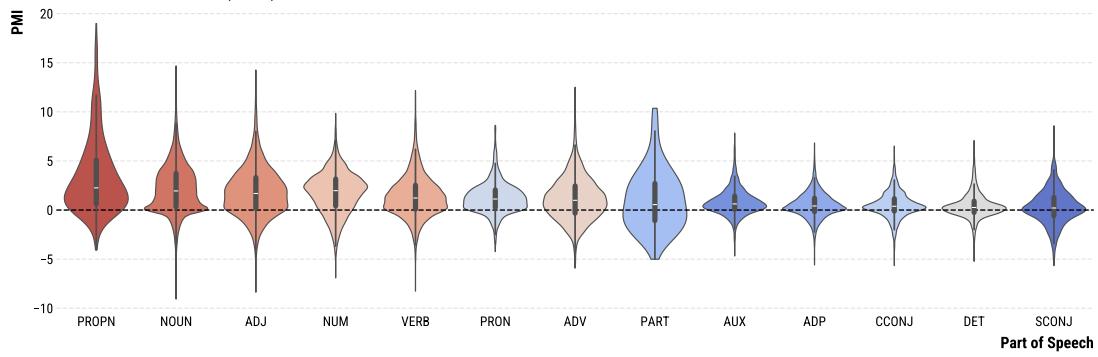
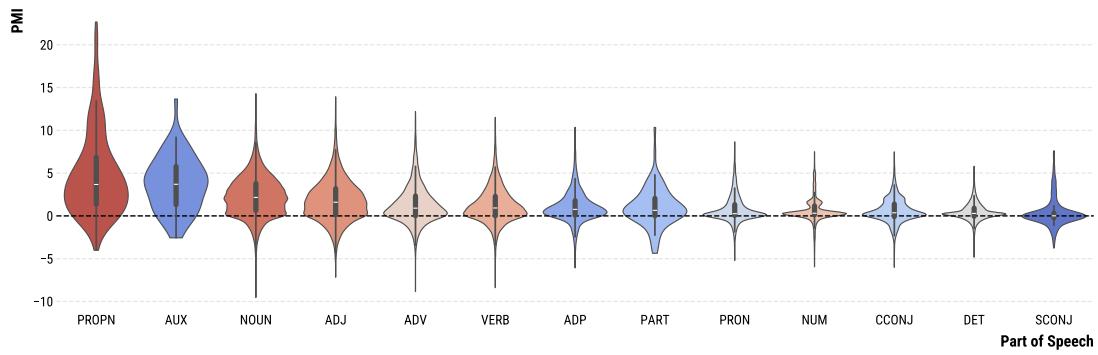
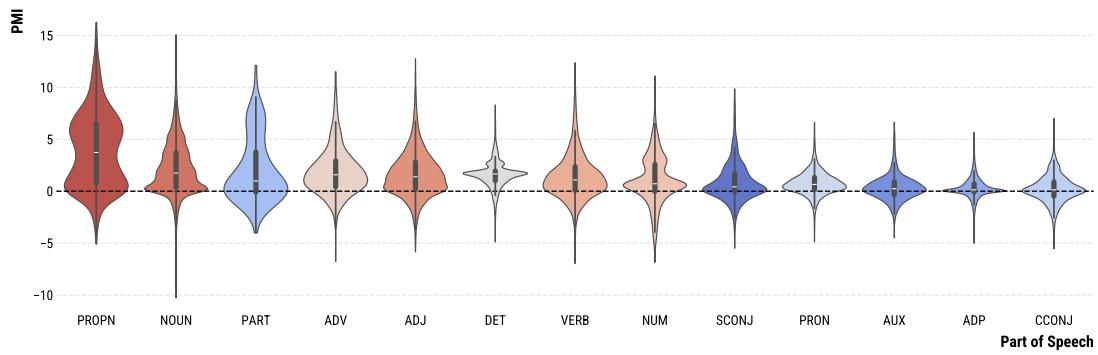
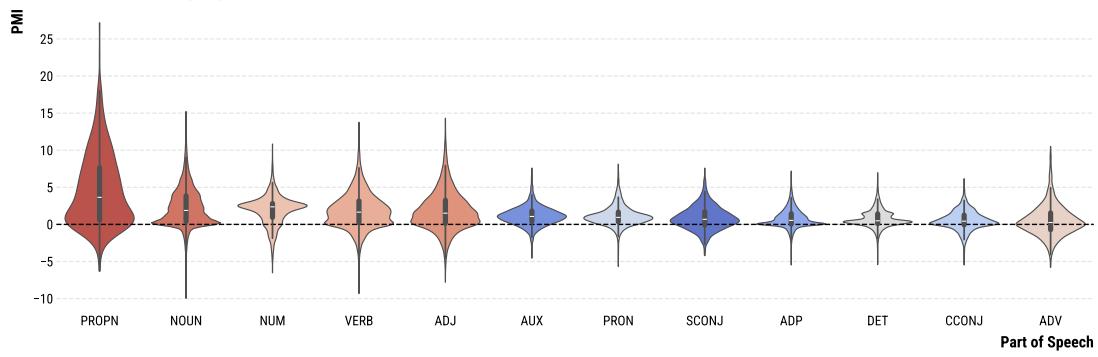
Spanish (es)

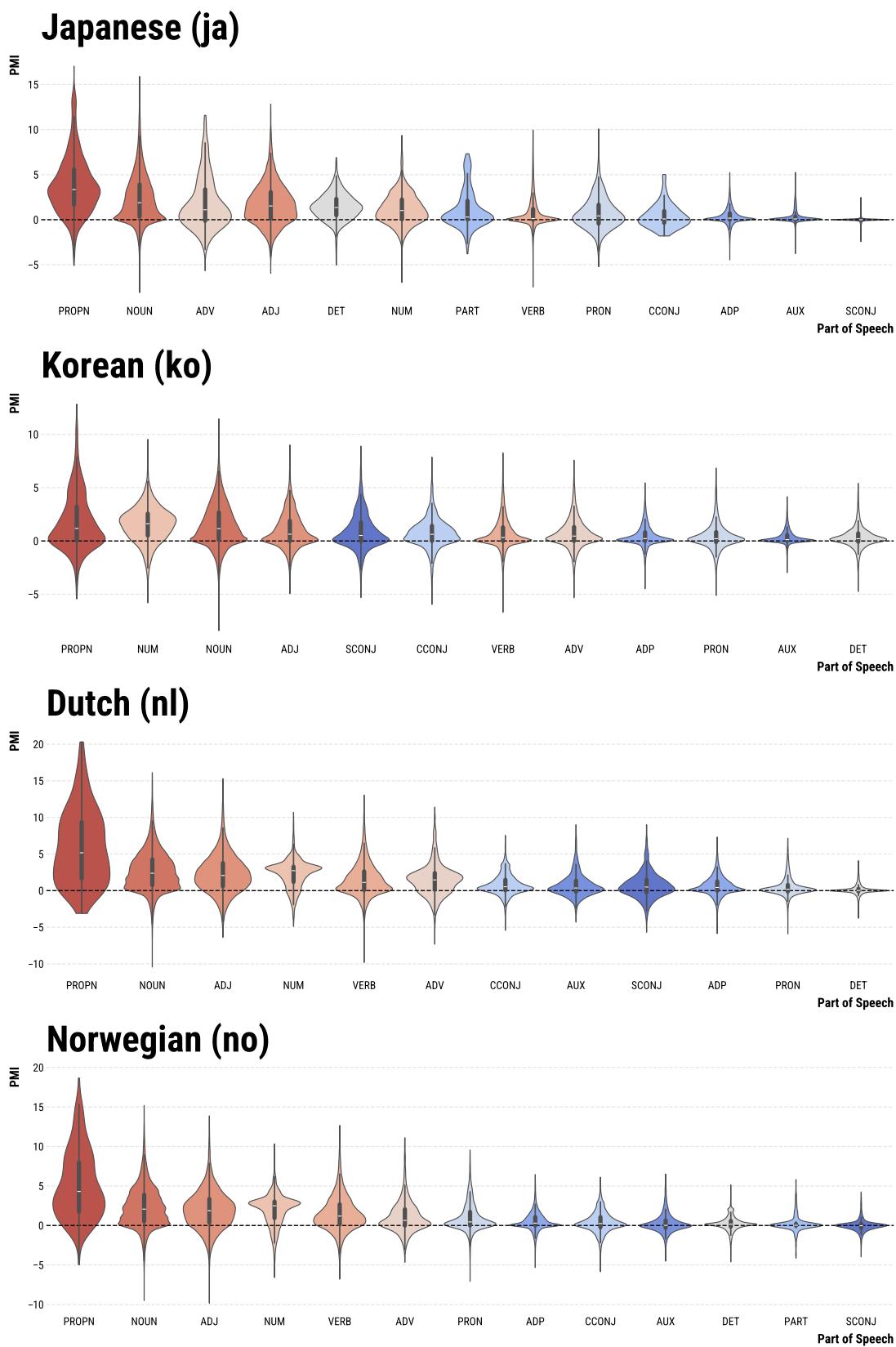


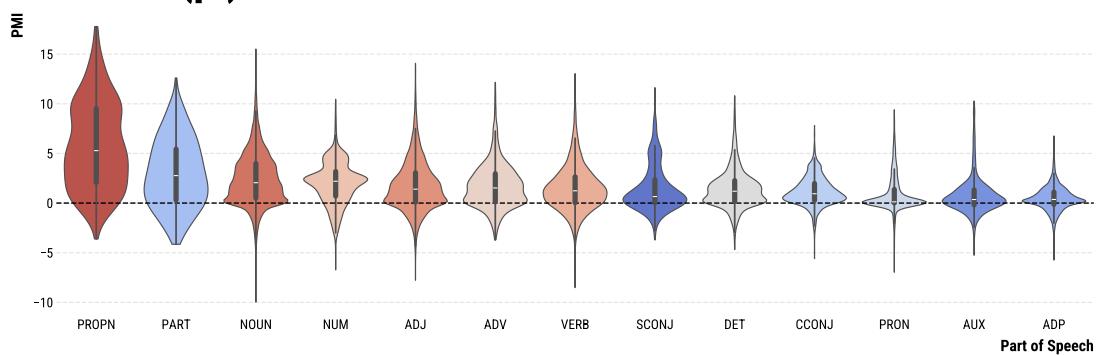
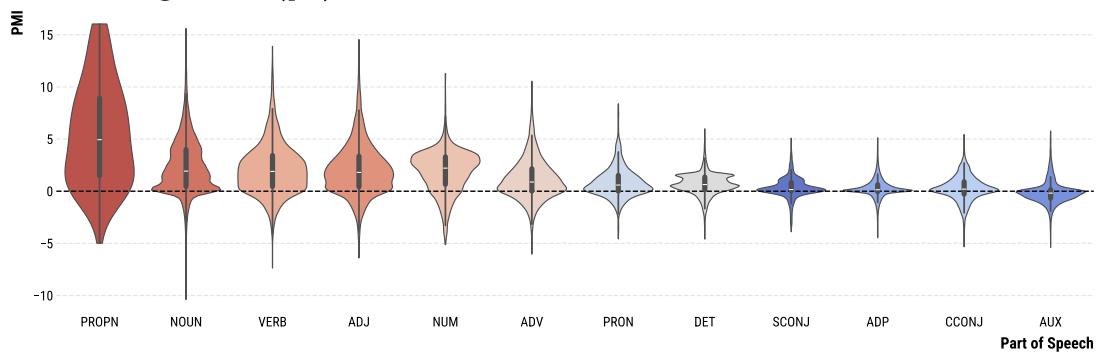
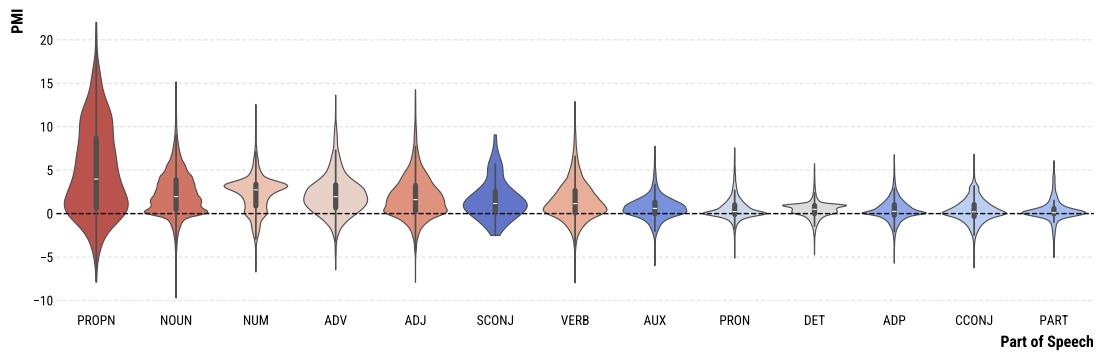
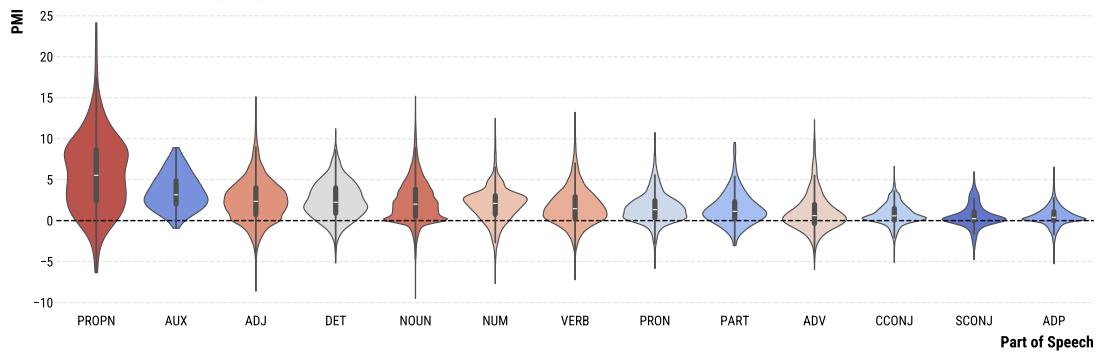
Persian (fa)

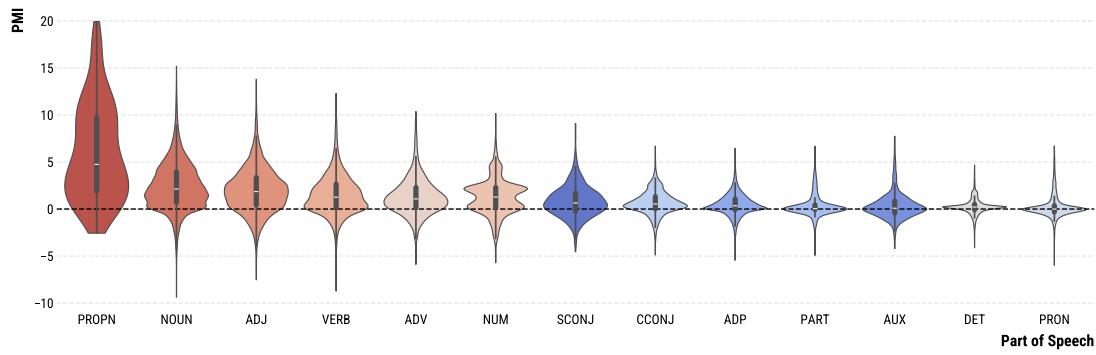
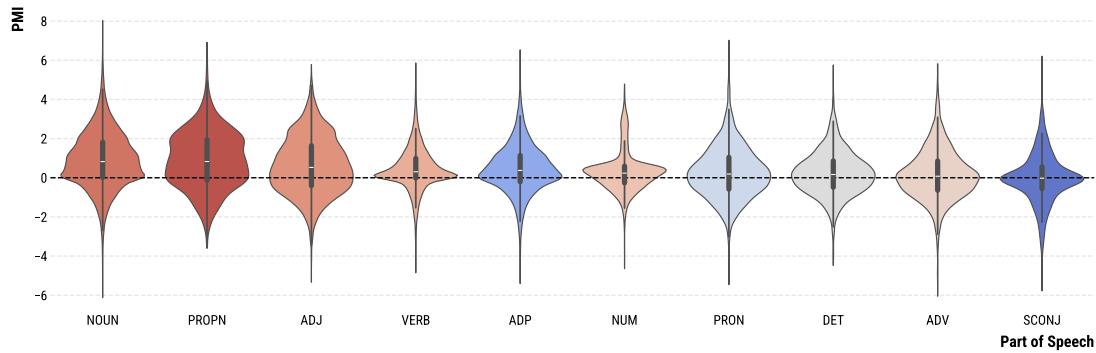
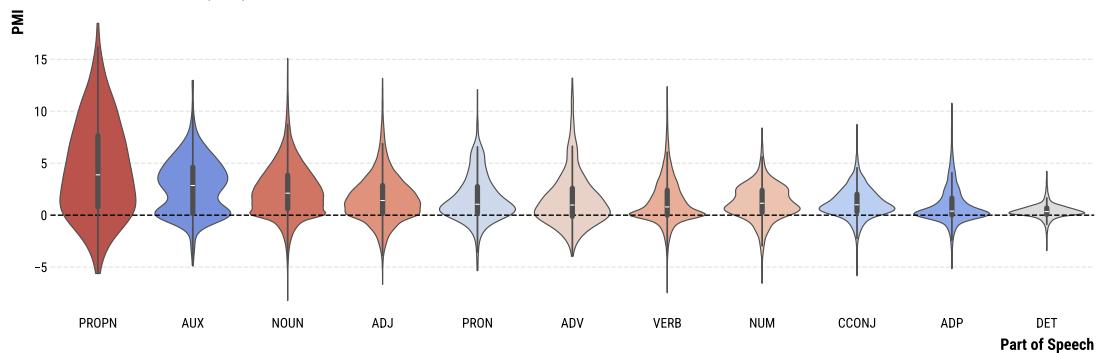
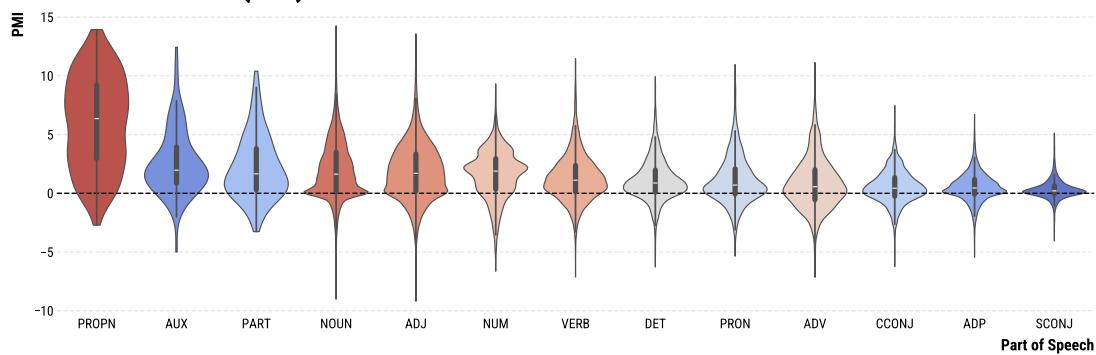


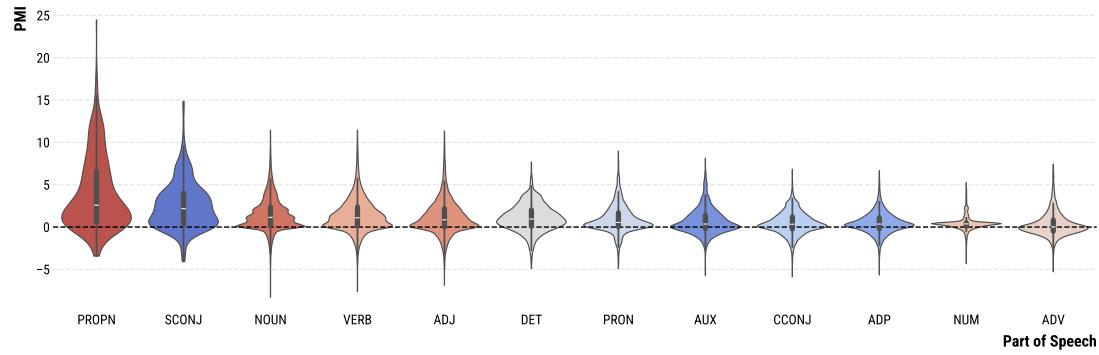
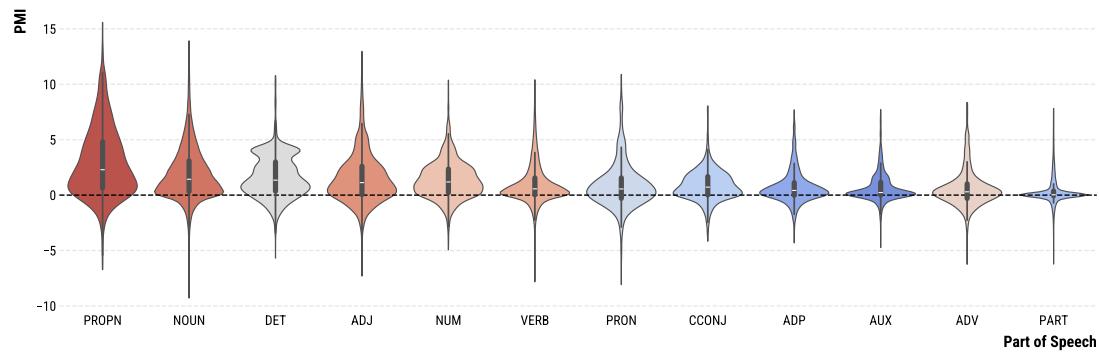


Croatian (hr)**Hungarian (hu)****Indonesian (id)****Italian (it)**



Polish (pl)**Portuguese (pt)****Romanian (ro)****Russian (ru)**

Swedish (sv)**Telugu (te)****Turkish (tr)****Ukrainian (uk)**

Vietnamese (vi)**Chinese (zh)**

Bibliography

- Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France. European Language Resources Association.
- Peter Ackema and Ad Neeleman. 2019. Default person versus default number in agreement. In *Agreement, Case and Locality in the Nominal and Verbal Domains, Open Generative Syntax*, pages 21–54. Language Science Press.
- 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.
- Badr AlKhamissi, Greta Tuckute, Yingtian Tang, Taha Osama A Binhuraib, Antoine Bosselut, and Martin Schrimpf. 2025. From Language to Cognition: How LLMs Outgrow the Human Language Network. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24332–24350, Suzhou, China. Association for Computational Linguistics.
- Stephen R. Anderson. 1982. Where’s morphology? *Linguistic Inquiry*, 13:571–612.
- Stephen R. Anderson. 1985. Inflectional morphology. In Timothy Shopen, editor, *Language Typology and Syntactic Description*, 1 edition, volume 3, pages 150–201. Cambridge University Press, Cambridge.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433.

Richard Antonello and Alexander Huth. 2024. Predictive Coding or Just Feature Discovery? An Alternative Account of Why Language Models Fit Brain Data. *Neurobiology of Language*, 5(1):64–79.

Antti Arppe, Atticus Harrigan, Katherine Schmirler, Lene Antonsen, Trond Trosterud, Sjur Nørstebø Moshagen, Miikka Silfverberg, Arok Wolvegrey, Conor Snoek, Jordan Lachler, Eddie Antonio Santos, Jean Okimāsis, and Dorothy Thunder. 2014–2019. Finite-state transducer-based computational model of Plains Cree morphology.

Lucas F.E. Ashby, Travis M. Bartley, Simon Clematide, Luca Del Signore, Cameron Gibson, Kyle Gorman, Yeonju Lee-Sikka, Peter Makarov, Aidan Malanoski, Sean Miller, Omar Ortiz, Reuben Raff, Arundhati Sengupta, Bora Seo, Yulia Spektor, and Winnie Yan. 2021. Results of the second SIGMORPHON shared task on multilingual grapheme-to-phoneme conversion. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 115–125, Online. Association for Computational Linguistics.

Madina Babazhanova, Maxat Tezekbayev, and Zhenisbek Assylbekov. Tenth century–2021. Geometric probing of word vectors. In *ESANN 2021 Proceedings - 29th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 587–592, Virtual, Online, Belgium. i6doc.com publication.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2021. MorphyNet:

- A large multilingual database of derivational and inflectional morphology. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 39–48, Online. Association for Computational Linguistics.
- Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóglua, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud'hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova. 2022. UniMorph 4.0: Universal Morphology. In *Proceedings of*

- the Thirteenth Language Resources and Evaluation Conference, pages 840–855, Marseille, France. European Language Resources Association.
- Laurie Bauer. 2004. The function of word-formation and the inflection-derivation distinction. In *Words in Their Places. A Festschrift for J. Lachlan Mackenzie*, pages 283–292. Vrije Universiteit, Amsterdam.
- Sacha Beniamine, Martin Maiden, and Erich Round. 2020. Opening the Romance verbal inflection dataset 2.0: A CLDF lexicon. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3027–3035, Marseille, France. European Language Resources Association.
- 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.
- Toms Bergmanis and Sharon Goldwater. 2017. From segmentation to analyses: A probabilistic model for unsupervised morphology induction. In *Proceedings of EACL*, Valencia, Spain.
- Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keyser, 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 Balazovic, 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. 2010. Word Classes. In Jae Jung Song, editor, *The Oxford Handbook of Linguistic Typology*, page 0. Oxford University Press.
- Walter Bisang. 2017. *Grammaticalization*. Oxford University Press.
- Ned Block. 1998. Semantics, conceptual role. In Edward Craig, editor, *Routledge Encyclopedia of Philosophy*, volume 8, pages 652–657. Routledge, London.
- Sasha Boguraev, Christopher Potts, and Kyle Mahowald. 2025. Causal Interventions Reveal Shared Structure Across English Filler–Gap Constructions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25032–25053, Suzhou, China. Association for Computational Linguistics.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781, Online. Association for Computational Linguistics.
- Olivier Bonami and Denis Paperno. 2018. Inflection vs. derivation in a distributional vector space. *Lingue e linguaggio*, 17(2):173–196.
- Olivier Bonami and Jana Strnadová. 2019. Paradigm structure and predictability in derivational morphology. *Morphology*, 29(2):167–197.
- Geert Booij. 1996. Inherent versus contextual inflection and the split morphology

- hypothesis. In Geert Booij and Jaap van Marle, editors, *Yearbook of Morphology* 1995, pages 1–16. Springer, Kluwer, Dordrecht.
- Geert Booij. 2007. Inflection. In Geert Booij, editor, *The Grammar of Words: An Introduction to Linguistic Morphology*, pages 99–124. Oxford University Press.
- RD Boschloo. 1970. Raised conditional level of significance for the 2× 2-table when testing the equality of two probabilities. *Statistica Neerlandica*, 24(1):1–9.
- Kasper Boye and Roelien Bastiaanse. 2018. Grammatical versus lexical words in theory and aphasia: Integrating linguistics and neurolinguistics. *Glossa: a journal of general linguistics*, 3(1).
- Jannik Brinkmann, Chris Wendler, Christian Bartelt, and Aaron Mueller. 2025. Large Language Models Share Representations of Latent Grammatical Concepts Across Typologically Diverse Languages. In *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 6131–6150, Albuquerque, New Mexico. Association for Computational Linguistics.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

- Benjamin Bruening. 2018. The lexicalist hypothesis: Both wrong and superfluous. *Language*, 94(1):1–42.
- 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.
- Joan L Bybee. 1985. *Morphology: A Study of the Relation between Meaning and Form*. John Benjamins, Amsterdam.
- Mari Chanturidze, Rebecca Carroll, and Esther Ruigendijk. 2019. Prepositions as a hybrid between lexical and functional category: Evidence from an ERP study on German sentence processing. *Journal of Neurolinguistics*, 52:100857.
- 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.
- Bhavin Choksi, Milad Mozafari, Rufin VanRullen, and Leila Reddy. 2022. Multimodal neural networks better explain multivoxel patterns in the hippocampus. *Neural Networks*, 154:538–542.
- Noam Chomsky. 1957. *Syntactic Structures*. De Gruyter Mouton.

Gabriella Chronis and Katrin Erk. 2020. When is a bishop not like a rook? when it's like a rabbi! multi-prototype BERT embeddings for estimating semantic relationships. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 227–244, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Maria Copot, Timothee Mickus, and Olivier Bonami. 2022. Idiosyncratic frequency as a measure of derivation vs. inflection. *Journal of Language Modelling*, 10(2):193–240.

Greville G Corbett. 2010. Canonical derivational morphology. *Word structure*, 3(2):141–155.

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 Hinrich Schütze. 2018. Joint semantic synthesis and morphological analysis of the derived word. *Transactions of the Association for Computational Linguistics*, 6:33–48.

W. Croft. 1991. *Syntactic Categories and Grammatical Relations: The Cognitive Organization of Information*. Emersion: Emergent Village Resources for Communities of Faith Series. University of Chicago Press.

- William Croft. 2000. *Explaining Language Change: An Evolutionary Approach*. Pearson Education.
- William Croft. 2001. *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford University Press.
- William Croft. 2002. *Typology and Universals*, 2 edition. Cambridge University Press.
- William Croft. 2016. Comparative concepts and language-specific categories: Theory and practice. *Linguistic Typology*, 20(2):377–393.
- William Croft. 2022. *Morphosyntax: Constructions of the World's Languages*. Cambridge University Press.
- William Croft and Keith T. Poole. 2008. Inferring universals from grammatical variation: Multidimensional scaling for typological analysis. *Theoretical Linguistics*, 34(1):1–37.
- Anne Cutler. 1981. Degrees of transparency in word formation. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, 26(1):73–77.
- Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.
- Ferdinand de Saussure. 1916. *Cours de Linguistique Générale*. Payot, Paris.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.
- Daniel Deutsch, John Hewitt, and Dan Roth. 2018. A distributional and orthographic aggregation model for English derivational morphology. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), pages 1938–1947, Melbourne, Australia. Association

for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019.

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

Michele T. Diaz and Gregory McCarthy. 2009. A comparison of brain activity evoked by single content and function words: An fMRI investigation of implicit word processing. *Brain Research*, 1282:38–49.

R. M. W. Dixon. 1977. Where have all the adjectives gone? *Studies in Language*, 1(1):19–80.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Wolfgang U Dressler. 1989. Prototypical differences between inflection and derivation. *STUF-Language Typology and Universals*, 42(1):3–10.

Matthew S Dryer. 1989. Large linguistic areas and language sampling. *Studies in Language. International Journal sponsored by the Foundation “Foundations of Language”*, 13(2):257–292.

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.

- Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.
- Allyson Ettinger and Tal Linzen. 2016. Evaluating vector space models using human semantic priming results. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 72–77, Berlin, Germany. Association for Computational Linguistics.
- Matan Eyal, Shoval Sadde, Hillel Taub-Tabib, and Yoav Goldberg. 2022. Large scale substitution-based word sense induction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4738–4752, Dublin, Ireland. Association for Computational Linguistics.
- 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.
- Sheridan Feucht, Eric Todd, Byron C. Wallace, and David Bau. 2025. The Dual-Route Model of Induction. In *Second Conference on Language Modeling*.
- J. R. Firth. 1957. A synopsis of linguistic theory, 1930–55. In *Studies in Linguistic Analysis*, pages 1–31. Blackwell, Oxford.
- Simeon Floyd. 2011. Re-discovering the Quechua adjective. *Linguistic Typology*, 15(1):25–63.

- A. D. Friederici, B. Opitz, and D. Y. von Cramon. 2000. Segregating semantic and syntactic aspects of processing in the human brain: An fMRI investigation of different word types. *Cerebral Cortex (New York, N.Y.: 1991)*, 10(7):698–705.
- Richard Futrell and Michael Hahn. 2022. Information Theory as a Bridge Between Language Function and Language Form. *Frontiers in Communication*, 7.
- Richard Futrell and Kyle Mahowald. 2025. How Linguistics Learned to Stop Worrying and Love the Language Models. *Behavioral and Brain Sciences*, pages 1–98.
- Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.
- 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.
- T. Givón. 1979. *On Understanding Grammar*. Perspectives in Neurolinguistics and Psycholinguistics. Academic Press.
- Gabriel Goh, Nick Cammarata †, Chelsea Voss †, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. 2021. Multimodal Neurons in Artificial Neural Networks. *Distill*, 6(3):e30.

- Adele E. Goldberg. 2024. Usage-based constructionist approaches and large language models. *Constructions and Frames*, 16(2):220–254.
- Gabriel Grand, Idan Asher Blank, Francisco Pereira, and Evelina Fedorenko. 2022. Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature Human Behaviour*, 6(7):975–987.
- Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- 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.
- Nina Gregorio, Matteo Gay, Sharon Goldwater, and Edoardo Ponti. 2025. The Cross-linguistic Role of Animacy in Grammar Structures. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7349–7363, Vienna, Austria. Association for Computational Linguistics.
- John Haiman. 1980. The Iconicity of Grammar: Isomorphism and Motivation. *Language*, 56(3):515–540.
- 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*.
- William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Zellig Harris. 1954. Distributional structure. *Word-journal of The International Linguistic Association*, 10(23):146–162.

Mareike Hartmann, Yova Kementchedjhieva, and Anders Søgaard. 2019. Comparing Unsupervised Word Translation Methods Step by Step. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Martin Haspelmath. 1996. Word-class-changing inflection and morphological theory. *Yearbook of morphology 1995*, pages 43–66.

Martin Haspelmath. 2003. The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In Michael Tomasello, editor, *The New Psychology of Language*, volume 2, pages 211–242. Lawrence Erlbaum Associates, Mahwah, NJ, USA.

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.

Martin Haspelmath. 2021. Towards standardization of morphosyntactic terminology for general linguistics. In Luca Alfieri, Giorgio Francesco Arcodia, and Paolo Ramat, editors, *Linguistic Categories, Language Description and Linguistic Typology*, pages 35–58. John Benjamins Publishing Company.

Martin Haspelmath. 2024. Inflection and derivation as traditional comparative concepts. *Linguistics*, 62(1):43–77.

Nabil Hathout and Fiammetta Namer. 2016. Giving lexical resources a second life: Démonette, a multi-sourced morpho-semantic network for French. In

- Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1084–1091, Portorož, Slovenia. European Language Resources Association (ELRA).
- Nabil Hathout, Franck Sajous, and Basilio Calderone. 2014. GLÀFF, a large versatile French lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1007–1012, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Junxian He, Graham Neubig, and Taylor Berg-Kirkpatrick. Tenth century–2018. Unsupervised learning of syntactic structure with invertible neural projections. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1292–1302, Brussels, Belgium. Association for Computational Linguistics.
- Valentin Hofmann, Hinrich Schütze, and Janet Pierrehumbert. 2020. A graph auto-encoder model of derivational morphology. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1127–1138, Online. Association for Computational Linguistics.
- Eghbal Hosseini, Colton Casto, Noga Zaslavsky, Colin Conwell, Mark Richardson, and Evelina Fedorenko. 2024. Universality of representation in biological and artificial neural networks.
- Henrison Hsieh. 2019. Distinguishing nouns and verbs: A Tagalog case study. *Natural Language & Linguistic Theory*, 37(2):523–569.
- Zongliang Hu, Kai Dong, Wenlin Dai, and Tiejun Tong. 2017. A comparison of methods for estimating the determinant of high-dimensional covariance matrix. *The International Journal of Biostatistics*, 13(2):20170013.
- Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet

- Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239, San Diego, California. Association for Computational Linguistics.
- Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. Position: The Platonic Representation Hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, pages 20617–20642. PMLR.
- Rishi Dev Jha, Collin Zhang, Vitaly Shmatikov, and John Xavier Morris. 2025. Harnessing the Universal Geometry of Embeddings. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Martin Joos. 1950. Description of Language Design. *The Journal of the Acoustical Society of America*, 22(6):701–707.
- Andrej Karpathy and Li Fei-Fei. 2017. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):664–676.
- M. Kasthuri, S. Britto Ramesh Kumar, and Souheil Khaddaj. 2017. PLIS: Proposed language independent stemmer for information retrieval systems using dynamic programming. In *2017 World Congress on Computing and Communication Technologies (WCCCT)*, pages 132–135.
- Carina Kauf, Greta Tuckute, Roger Levy, Jacob Andreas, and Evelina Fedorenko. 2024. Lexical-Semantic Content, Not Syntactic Structure, Is the Main Contributor to ANN-Brain Similarity of fMRI Responses in the Language Network. *Neurobiology of Language*, 5(1):7–42.
- Daniel Kaufman. 2009. Austronesian Nominalism and its consequences: A Tagalog case study. *Theoretical Linguistics*, 35(1):1–49.

- Charles Kemp and Terry Regier. 2012. Kinship Categories Across Languages Reflect General Communicative Principles. *Science*, 336(6084):1049–1054.
- Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Simon Kirby. 1999. *Function, Selection, and Innateness: The Emergence of Language Universals*. Oxford University Press.
- Bilal Kirkici and Harald Clahsen. 2013. Inflection and derivation in native and non-native language processing: Masked priming experiments on Turkish. *Bilingualism: Language and Cognition*, 16(4):776–791.
- Christa König. 2006. Marked nominative in africa. *Studies in Language. International Journal sponsored by the Foundation “Foundations of Language”*, 30(4):655–732.
- Austin C. Kozlowski, Callin Dai, and Andrei Boutyline. 2025. Semantic Structure in Large Language Model Embeddings. *Preprint*, arXiv:2508.10003.
- Lukáš Kyjánek and Olivier Bonami. 2025. Theoretical stances shape empirical generalizations on inflection vs. derivation: Quantitative evidence from Czech. In *Word-Formation Theories VII & Typology and Universals in Word-Formation V*, volume 25–26, Košice. Pavol Jozef Šafárik University.
- Lukáš Kyjánek, Zdeněk Žabokrtský, Magda Ševčíková, and Jonáš Vidra. 2020. Universal Derivations 1.0, a growing collection of harmonised word-formation resources. *The Prague Bulletin of Mathematical Linguistics*, 2(115):333–348.
- Lior Laks and Fiammetta Namer. 2022. Hebrewnette—a new derivational re-

- source for non-concatenative morphology: Principles, design and implementation. *The Prague Bulletin of Mathematical Linguistics*, 118:25–53.
- Ronald W. Langacker. 1987. *Foundations of Cognitive Grammar*. Stanford University Press, Stanford, Calif.
- Septina Dian Larasati, Vladislav Kuboň, and Daniel Zeman. 2011. Indonesian morphology tool (MorphInd): Towards an indonesian corpus. In Cерстин Mahlow and Michael Piotrowski, editors, *Systems and Frameworks for Computational Morphology*, pages 119–129. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Alessandro Laudanna, William Badecker, and Alfonso Caramazza. 1992. Processing inflectional and derivational morphology. *Journal of Memory and Language*, 31(3):333–348.
- Vladimir Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10:707.
- Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.
- Vivian G. Li. 2025. Embedding derived animacy rankings offer insights into the sources of grammatical animacy. In *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 1339–1351, Albuquerque, New Mexico. Association for Computational Linguistics.
- Rochelle Lieber and Pavol Štekauer. 2014. Universals in Derivation. In Rochelle Lieber and Pavol Štekauer, editors, *The Oxford Handbook of Derivational Morphology*, page 0. Oxford University Press.

- Johan Liljencrants, Björn Lindblom, and Bjorn Lindblom. 1972. Numerical Simulation of Vowel Quality Systems: The Role of Perceptual Contrast. *Language*, 48(4):839–862.
- Chu-Cheng Lin, Waleed Ammar, Chris Dyer, and Lori Levin. 2015. Unsupervised POS induction with word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1311–1316, Denver, Colorado. Association for Computational Linguistics.
- 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*, 7:23969415221085827.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.
- Authors Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall◊, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. 2025. On the Biology of a Large Language Model.
- 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.
- Nikola Ljubešić, Filip Klubička, Željko Agić, and Ivo-Pavao Jazbec. 2016. New inflectional lexicons and training corpora for improved morphosyntactic annotation of Croatian and Serbian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4264–4270, Portorož, Slovenia. European Language Resources Association (ELRA).
- 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.
- Donald G MacKay. 1978. Derivational rules and the internal lexicon. *Journal of verbal learning and verbal behavior*, 17(1):61–71.
- Robert Malouf, Farrell Ackerman, and Arturs Semenuks. 2020. Lexical databases for computational analyses: A linguistic perspective. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 446–456, New York, New York. Association for Computational Linguistics.
- Jacob A. Matthews, Laurent Dubreuil, Imane Terhmina, Yunci Sun, Matthew Wilkens, and Marten Van Schijndel. 2025. Disentangling language change: Sparse autoencoders quantify the semantic evolution of indigeneity in French. In *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 11208–11222, Albuquerque, New Mexico. Association for Computational Linguistics.
- Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka

- Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2020. UniMorph 3.0: Universal Morphology. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3922–3931, Marseille, France. European Language Resources Association.
- Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, Red Hook, NY, USA. Curran Associates Inc.
- 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.
- Chigusa Morita. 2010. The Internal Structures of Adjectives in Japanese. *Linguistic research: working papers in English linguistics*, 26:105–117.
- Karthik Narasimhan, Regina Barzilay, and Tommi Jaakkola. 2015. An unsupervised method for uncovering morphological chains. *Transactions of the Association for Computational Linguistics*, 3:157–167.
- Helen J. Neville, Debra L. Mills, and Donald S. Lawson. 1992. Fractionating

Language: Different Neural Subsystems with Different Sensitive Periods.

Cerebral Cortex, 2(3):244–258.

Frederick J. Newmeyer. 1999. The Discrete Nature Of Syntactic Categories: Against A Prototype-Based Account. chapter The Nature and Function of Syntactic Categories. Brill.

Byung-Doh Oh and William Schuler. 2024. Leading Whitespace of Language Models' Subword Vocabulary Pose a Confound for Calculating Word Probabilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3464–3472, Miami, Florida, USA. Association for Computational Linguistics.

Bruce Oliver, Clarissa Forbes, Changbing Yang, Farhan Samir, Edith Coates, Garrett Nicolai, and Miikka Silfverberg. 2022. An inflectional database for gitksan. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6597–6606, Marseille, France. European Language Resources Association.

David Y. Oshima, Kimi Akita, and Shin-ichiro Sano. 2019. Gradability, scale structure, and the division of labor between nouns and adjectives: The case of Japanese. *Glossa: a journal of general linguistics*, 4(1).

Isabel Papadimitriou, Ethan A. Chi, Richard Futrell, and Kyle Mahowald. 2021. Deep Subjecthood: Higher-Order Grammatical Features in Multilingual BERT. *Preprint*, arXiv:2101.11043.

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*.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- David Perlmutter. 1988. The split morphology hypothesis: Evidence from Yiddish. *Theoretical morphology*, pages 79–100.
- Erika Petersen and Christopher Potts. 2023. Lexical semantics with large language models: A case study of English “break”. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 490–511, Dubrovnik, Croatia. Association for Computational Linguistics.
- Steven T. Piantadosi and Felix Hill. 2022. Meaning without reference in large language models. *Preprint*, arXiv:2208.02957.
- Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.
- Tiago Pimentel and Clara Meister. 2024. How to Compute the Probability of a Word. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18358–18375, Miami, Florida, USA. Association for Computational Linguistics.
- Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Compu-*

- tational Linguistics*, pages 4609–4622, Online. 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.
- Friedemann Pulvermüller, Werner Lutzenberger, and Niels Birbaumer. 1995. Electrocortical distinction of vocabulary types. *Electroencephalography and Clinical Neurophysiology*, 94(5):357–370.
- Yulu Qin, Dheeraj Varghese, Adam Dahlgren Lindström, Lucia Donatelli, Kanishka Misra, and Najoung Kim. 2025. Vision-and-Language Training Helps Deploy Taxonomic Knowledge but Does Not Fundamentally Alter It. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- 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.
- Shauli Ravfogel, Yanai Elazar, Jacob Goldberger, and Yoav Goldberg. 2020. Unsupervised distillation of syntactic information from contextualized word

- representations. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 91–106, Online. Association for Computational Linguistics.
- Terry Regier, Paul Kay, and Naveen Khetarpal. 2007. Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences*, 104(4):1436–1441.
- Terry Regier, Naveen Khetarpal, and Asifa Majid. 2013. Inferring semantic maps. *Linguistic Typology*, 17(1):89–105.
- Norvin Richards. 2009. Nouns, verbs, and hidden structure in Tagalog. *Theoretical Linguistics*, 35(1):139–152.
- Phillip Rogers. 2016. Illustrating the prototype structures of parts of speech: A multidimensional scaling analysis.
- Rudolf Rosa and Zdeněk Žabokrtský. 2019. Attempting to separate inflection and derivation using vector space representations. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*, pages 61–70, Prague, Czechia. Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics.
- Rudolf Rosa and Zdenek Zabokrtský. 2019. Unsupervised lemmatization as embeddings-based word clustering. *CoRR*, abs/1908.08528.
- John R. Ross. 1972. The category squish: Endstation Hauptwort. In *Proceedings of the Eighth Regional Meeting of the Chicago Linguistic Society*, pages 316–328, Chicago, Illinois. Chicago Linguistic Society, University of Chicago.
- Joshua Rozner, Leonie Weissweiler, Kyle Mahowald, and Cory Shain. 2025. Constructions are Revealed in Word Distributions. In *Proceedings of the 2025 Con-*

- ference on Empirical Methods in Natural Language Processing, pages 2116–2138, Suzhou, China. Association for Computational Linguistics.
- David E. Rumelhart, James L. McClelland, and PDP Research Group. 1986. *Parallel Distributed Processing, Volume 1: Explorations in the Microstructure of Cognition: Foundations*. The MIT Press.
- Jenny R Saffran, Richard N Aslin, and Elissa L Newport. 1996. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.
- Naomi Saphra, Eve Fleisig, Kyunghyun Cho, and Adam Lopez. 2024. First Tragedy, then Parse: History Repeats Itself in the New Era of Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2310–2326, Mexico City, Mexico. Association for Computational Linguistics.
- Adriaan M. J. Schakel and Benjamin J. Wilson. 2015. Measuring word significance using distributed representations of words. *Computing Research Repository*, arXiv:1508.02297.
- Patrick Schone and Daniel Jurafsky. 2000. Knowledge-free induction of morphology using latent semantic analysis. In *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*.
- Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118.
- 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.
- Jean-Luc Schwartz, Louis-Jean Boë, Nathalie Vallée, and Christian Abry. 1997. The Dispersion-Focalization Theory of vowel systems. *Journal of Phonetics*, 25(3):255–286.
- Wesley Scivetti and Nathan Schneider. 2025. Construction Identification and Disambiguation Using BERT: A Case Study of NPN. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 365–376, Vienna, Austria. Association for Computational Linguistics.
- 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.
- Sidney J. Segalowitz and Korri C. Lane. 2000. Lexical access of function versus content words. *Brain and Language*, 75(3):376–389.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.
- C. E. Shannon. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423.
- Michael Silverstein. 1986. 7. Hierarchy of features and ergativity. In *Features and Projections*, pages 163–232. De Gruyter Mouton, Berlin, Boston.
- Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

- Radu Soricut and Franz Och. 2015. Unsupervised morphology induction using word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1627–1637, Denver, Colorado. Association for Computational Linguistics.
- Andrew Spencer. 2013. *Lexical Relatedness*. Oxford University Press, Oxford.
- Leon Stassen. 1997. *Intransitive Predication*. Oxford University Press.
- Adrian Staub. Forthcoming. Predictability in Language Comprehension: Prospects and Problems for Surprisal. *Annual Review of Linguistics*.
- Pavol Štekauer. 2015. The delimitation of derivation and inflection. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen, and Franz Rainer, editors, *Volume 1 Word-Formation*, pages 218–235. de Gruyter Mouton.
- Kenneth N. Stevens. 1989. On the quantal nature of speech. *Journal of Phonetics*, 17(1-2):3–45.
- Lonny Alaskuk Strunk. 2020. *A Finite-State Morphological Analyzer for Central Alaskan Yup'ik*. University of Washington.
- Daniel Swingley. 2005. Statistical clustering and the contents of the infant vocabulary. *Cognitive psychology*, 50(1):86–132.
- John Sylak-Glassman. 2016. The composition and use of the universal morphological feature schema (UniMorph schema).
- Kazuma Takaoka, Sorami Hisamoto, Noriko Kawahara, Miho Sakamoto, Yoshi-taka Uchida, and Yuji Matsumoto. 2018. Sudachi: A Japanese Tokenizer for Business. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

- Julius Tarng, Purvi Goel, and Isaac Kauvar. 2025. Visual features across modalities: SVG and ASCII art reveal cross-modal understanding. *Transformer Circuits Thread*.
- Pius ten Hacken. 1994. *Defining Morphology: A Principled Approach to Determining the Boundaries of Compounding, Derivation, and Inflection*. Altertumswissenschaftliche Texte Und Studien. G. Olms Verlag.
- Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.
- Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you learn from context? Probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.
- 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.
- Erik D Thiessen, Alexandra T Kronstein, and Daniel G Hufnagle. 2013. The extraction and integration framework: A two-process account of statistical learning. *Psychological bulletin*, 139(4):792.

- Erik D Thiessen and Jenny R Saffran. 2003. When cues collide: Use of stress and statistical cues to word boundaries by 7-to 9-month-old infants. *Developmental psychology*, 39(4):706.
- S. Thompson. 1988. A discourse approach to the cross-linguistic category 'Adjective'. In Roberta Corrigan, Fred Eckman, and Michael Noonan, editors, *Linguistic Categorization: Proceedings of an International Symposium in Milwaukee, Wisconsin, April 10–11, 1987*, pages 245–265. John Benjamins Publishing Company.
- Susan P Thompson and Elissa L Newport. 2007. Statistical learning of syntax: The role of transitional probability. *Language learning and development*, 3(1):1–42.
- William Timkey and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4527–4546, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, pages 200–212, Red Hook, NY, USA. Curran Associates Inc.
- Satoshi Uehara. 1995. *Syntactic Categories in Japanese: A Typological and Cognitive Introduction*. Ph.D. thesis, University of Michigan, United States – Michigan.
- Martijn van der Klis and Jos Tellings. 2022. Generating semantic maps through multidimensional scaling: Linguistic applications and theory. *Corpus Linguistics and Linguistic Theory*, 18(3):627–665.

- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Annemarie Verkerk and Sander Lestrade. 2008. The encoding of adjectives. *Linguistics in the Netherlands*, 25(1):157–168.
- Hugo David Calderon Vilca, Flor Cagniy Cárdenas Mariñó, and Edwin Fredy Mamani Calderon. 2012. Analizador morfológico de la lengua Quechua basado en software libre Helsinkifinite-statetransducer (HFST).
- Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart, and Anna Korhonen. 2020a. Multi-SimLex: A large-scale evaluation of multilingual and crosslingual lexical semantic similarity. *Computational Linguistics*, 46(4):847–897.
- Ivan Vulić, Sebastian Ruder, and Anders Søgaard. 2020b. Are All Good Word Vector Spaces Isomorphic? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3178–3192, Online. Association for Computational Linguistics.
- Christian Wartena. 2013. Distributional similarity of words with different frequencies. In *Proceedings of the 13th Edition of the Dutch-Belgian Information Retrieval Workshop (DIR 2013)*, pages 8–11. Hochschule Hannover.
- David John Weber. 1983. *A Grammar of Huallaga (Huanuco) Quechua*. Ph.D. thesis, University of California, Los Angeles, United States – California.
- Harrie Wetzer. 2013. *The Typology of Adjectival Predication*. De Gruyter Mouton.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019.

Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings. In *Proceedings of the 15th Conference on Natural Language Processing, KONVENS 2019, Erlangen, Germany, October 9-11, 2019*.

Adam Wiemerslage, Arya D. McCarthy, Alexander Erdmann, Garrett Nicolai, Manex Agirrezabal, Miikka Silfverberg, Mans Hulden, and Katharina Kann. 2021. Findings of the SIGMORPHON 2021 shared task on unsupervised morphological paradigm clustering. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 72–81, Online. Association for Computational Linguistics.

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):10107.

Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. 2025. The Semantic Hub Hypothesis: Language Models Share Semantic Representations Across Languages and Modalities. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

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.

Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. 2017. STAIR captions: Constructing a large-scale Japanese image caption dataset. In *Proceedings of*

- the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 417–421, Vancouver, Canada. Association for Computational Linguistics.
- Hyejin Youn, Logan Sutton, Eric Smith, Christopher Moore, Jon F. Wilkins, Ian Maddieson, William Croft, and Tanmoy Bhattacharya. 2016. On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences*, 113(7):1766–1771.
- Tian Yun, Chen Sun, and Ellie Pavlick. 2021. Does vision-and-language pretraining improve lexical grounding? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4357–4366, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Noga Zaslavsky, Charles Kemp, Terry Regier, and Naftali Tishby. 2018. Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, 115(31):7937–7942.
- Noga Zaslavsky, Charles Kemp, Naftali Tishby, and Terry Regier. 2019. Color Naming Reflects Both Perceptual Structure and Communicative Need. *Topics in Cognitive Science*, 11(1):207–219.
- 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.
- Arnold M. Zwicky. 1994. What is a clitic? In Joel A. Nevis, Brian D. Joseph, Dieter Wanner, and Arnold M. Zwicky, editors, *Clitics: A Comprehensive Bibliography 1892–1991*, pages xii–xx. John Benjamins, Amsterdam and Philadelphia.