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**

2 Background **11**

2.1 Typology and Comparative Concepts 11

2.2 Approaches to Comparative Concepts in Computational Typology 13

2.2.1 Comparative concepts in typological databases 13

2.2.2 Phonological typology 14

2.2.3 Semantic category systems 17

2.2.4 Multidimensional scaling 20

2.2.5 Deep learning models of comparative concepts 23

2.3 The Distinctness of Lexicon and Syntax 25

2.3.1 Splitting the dimensions of the distinction 26

2.3.2 Relevant phenomena at different formal levels 29

2.3.3 Theories that reify the distinction 30

2.3.4 Doing without the distinction 30

2.3.5 Evidence from psychology and neuroscience 30

2.3.6 Computational operationalization 30

2.4 The Gradient Nature of Lexicality 30

2.5 Finding Meaning in Computational Models 31

3 Corpus-based Measures Discriminate Inflection and Derivation Cross-Linguistically 33

3.1 Introduction 33

3.2 Motivation for our measures 38

3.3 Method 42

3.3.1 Orthography-based measures 43

3.3.2 Distributional-embedding-based measures 44

3.4 Data 48

3.4.1 Data selection and summary 50

3.5 Distribution of the individual measures 51

3.5.1 Effects of Frequency 53

3.6 Predicting inflection and derivation 54

3.7 Classification of Linguistic Types of Inflection 57

3.7.1 Categories of inflectional meaning 59

3.7.2 Inherent vs. contextual inflection and transpositions . . 62

3.7.3 Summary 65

3.8 Discussion 65

3.8.1 The role of our individual measures 65

3.8.2 Language generality 69

3.8.3 The classification approach 71

3.8.4 Inflection and derivation: gradient or categorical? 72

3.8.5 Are inflection and derivation identifiable from the statistics of language? 75

3.8.6 Future work 76

3.9 Conclusion 79

3.10 The role of syntactic information 81

4 Groundedness and the Lexical–Functional Distinction 85

4.1 Introduction 85

4.2 Background 88

4.2.1 Contentfulness and word class 89

4.2.2 Measuring contentfulness 89

4.3 Method 91

4.4 Experimental setup 93

4.5 Results 96

4.5.1 Which word classes are grounded? 96

4.5.2 Which word classes are more grounded? 98

4.5.3 How consistent is word class groundedness across languages? 100

4.5.4 Semantic dimension of the measure 102

4.6 Discussion and Conclusion 104

5 Splitting and lumping: Visual groundedness as an organizing factor among lexical classes 109

5.1 Introduction 109

5.2 Continua among lexical word classes 111

5.3 Japanese adjectives 116

5.3.1 Method 118

5.3.2 Results 120

5.3.3 Discussion 123

5.4 The Tensedness Correlation 125

5.4.1 The typological finding 126

5.4.2 Theoretical explanation of the finding 128

5.4.3 Methodological background 129

5.4.4 Results 134

5.4.5 Discussion 137

5.5 Conclusion 140

Bibliography **143**

List of Figures

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 52

3.2 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 58

3.3 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 61

3.4 Probability and Odds ratio with 95% confidence intervals of being classified as derivation for inherent inflections and transpositions 63

3.5 Probability and Odds ratio with 95% confidence intervals of being classified as derivation for inherent vs. contextual noun inflections 64

3.6 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 74

3.7 The mean cosine similarity between FastText embeddings of words of the same and different parts of speech in UniMorph. . 82

4.1 Mean and standard deviation of per-language mutual information estimates between word class and image. Across 30 languages, we see clear and consistent tendencies about which parts of speech are more “grounded”, corresponding to a distinction between lexical and functional classes. 86

4.2 Heatmap of mutual information estimates across parts of speech in thirty languages. Cells show the statistical significance of a word class's groundedness (MI > 0). Unattested classes are white. Some functional classes display non-significant levels of groundedness in several languages, while lexical classes dominantly show highly significant grounding. 97

4.3 Word token level distributions of the groundedness measure (PMI) across all languages and datasets, grouped by part of speech (word class). We also report the estimated marginal mean and ranking of each word class. Colors are based on the ranking of classes, rather than their average PMIs. Overall, the distribution and estimated ranking of word classes strongly suggest our groundedness measure quantitatively captures the distinction between lexical and functional classes. 99

4.4 Correlation between human concreteness ratings and type-level groundedness (PMI; left, $\rho = 0.368$) or uncertainty coefficent (right, $\rho = 0.609$): i.e., the average ratio between LM surprisal and captioning model surprisal. 101

5.1 Groundedness scores for *na*-adjective *makka* (completely red; right) and *i*-adjective *akai* (red; left) in the STAIR-full-dev dataset. 124

5.2 Goundedness 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. 135

5.3 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. 138

5.4 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 disproporionally "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. 139

List of Tables

3.1 Sample of an inflectional construction (upper table, German nominative plural) and derivational construction (lower table, English verbal nominalisation with *–ion*) in our data 41

4.1 We match the data points on which the language model and image captioning model were trained. The three datasets are the Gemma pre-training mixture (PT), PaliGemma multimodal data for continued training (CT), and COCO image–caption pairs for fine-tuning (FT). Symbols indicate whether models are trained on text data (**A**) or on multimodal data (🖼**A**). 93

5.1 Croft (2001)'s analysis of the conceptual categories of the major parts of speech and their semantic properties (transposed). . . . 113

5.2 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**) 120

5.3 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**) 122

Chapter 1

Introduction

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.

One major challenge of typology is the large-scale nature of the problem. While important typological insights have been gained from the study of individual languages with interesting properties, robust typological generalizations generally require data from a wide and diverse set of languages. However, manually collecting and comparing data across many languages is inherently challenging, requiring much time investment and breadth and depth of expertise.

In recent years, the field has increasingly turned to computational methods to address these challenges (Futrell et al., 2015; Cotterell et al., 2019; Levshina, 2020; Östling, 2015; Gerdes et al., 2021). These techniques can serve to automate and systemize manual comparisons done by typologists, enabling new scales of analysis. Large typological datasets such as Universal Dependencies (de Marneffe et al., 2021) and UniMorph (Batsuren et al., 2022) which use a hybrid of manual and automatic annotations have facilitated large-scale comparison, and natural language processing (NLP) tools have been used to extract and analyze linguistic features from corpora in a more scalable way.

To make cross-linguistic comparisons and identify cross-linguistic variation, both traditional and computational typological work 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:

(1.1)	*Paul*	*loves*	*Peter.*
	Subj	Verb	Obj

(1.2)	*pooru-wa*	*piitaa-wo*	*aishiteiru*
	Paul-TOPIC	Peter-ACC	love-PRES.CONT
	Subj	Obj	Verb

English and Japanese 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."

These cross-linguistic concepts used for typological comparison (like "subject," "object," and "verb" in the previous example) have been the subject of substantial discussion in typology, with typologist debating both the nature and theoretical status of these concepts. Much early work and work in the generative linguistics traditions have sought *universal grammatical concepts*, which not only describe the variation across languages, but also have descriptive validity for the grammmars of individual languages (Greenberg, 1966a; Chomsky, 1957, p. 50;

Newmeyer, 2007; Wiltschko, 2014). A growing body of work has questioned the existence of such universal grammatical concepts, instead attempting to define cross-linguistically valid *comparative concepts*—which need not necessarily map onto the structure of individual language's grammar (Haspelmath, 2007; Croft, 2001, pp. 32–34).

Croft (2016), following Haspelmath (2010), identifies two major types of these cross-linguistic comparative concepts. The first, purely functional concepts, are similar to those argued for by Haspelmath (Haspelmath, 2010, 2003, 2012) Relying only on general semantics and discourse structure, they achieve cross-linguistic validity under fairly weak assumptions about the degree of cross-linguistic semantic relativity. As a prototypical example, take Haspelmath's suggestion to study the cross-linguistic syntax of words which denote "properties" instead of "adjectives." In contrast to this first type, Croft identifies "hybrid" comparative concepts as the other major type; these concepts have both a formal and a functional component (e.g., "adjectives"). These categories often have an intuitive appeal stemming from their similarities to language-specific grammatical categories and superficial broad cross-linguistic similarity. However, their hybrid nature makes cross-linguistic consistency incredibly challenging. Form generally varies from language to language: while in English, nouns are words which follow determiners, other languages may not have separate determiner words.

It is such hybrid comparative concepts with which this thesis concerns itself. While mainstream typologists have long been debating and investigating these concepts directly, computational work has dominantly employed hybrid comparative concepts while treating them as primitive or well-defined, using datasets such as UniMorph and Universal Dependencies and implicitly following whatever practices diverse dataset creators used to produce categorisations. In this thesis, I seek to directly interrogate hybrid cross-linguistic comparative concepts as instantiated in these typological datasets, and directly measure the cross-linguistic consistency of their application through computational techniques. Contrasting with the dominant paradigm in typology which has qualitatively identified prototypical semantics and formal properties of these concepts, I seek quantitative measures which do not rely on human judgements.

Besides scalability, another major advantage of the computational approach is the tools it provides for formalising functional and semantic aspects of language. In recent decades, a range of deep learning models of language (including word embedding methods, (vision-and-)language models, and so-called masked language models) have been introduced in the field of natural language processing, which have enabled formal theories to work with more intangible, continuous aspects of language like semantics (Copot et al., 2022). These models learn a representational space which provides rich representations of semantics and the world. Further, they do so without requiring direct instruction on this structure, but rather learn it implicitly from learning to predict words in a (linguistic and/or visual) context. This capacity makes these tools ideal for cross-linguistically operationalizing semantic dimensions of comparative concepts.

In this thesis, then, I provide the first explicit computational study of the cross-linguistic application of several comparative concepts **(Contribution 1)**. While some theoretical linguists have argued that these concepts are underspecified and inconsistently deployed, **(Claim 1)** I find that the application of deep learning models to provide concrete measures of the semantic dimensions of these concepts shows a high degree of cross-linguistic consistency in the way linguists use these concepts.

Another major contribution of this thesis is the development of a multimodal approach to typological investigation **(Contribution 2)**. Recent advances in deep learning applications have produced sophisticated cross-lingual joint models of image and text. A central challenge for computational typological investigation of semantics is the difficulty of having semantics which is *aligned* across languages. Prior studies attempt to use multilingual models with a shared representation space or to align representation spaces across multiple models. However, these approaches are fairly ad-hoc. An image *grounds* language in a language-agnostic model of the world state in which the language was produced. While this representation is necessarily imperfect, it also enables new directions.

I also demonstrate that the notion of *semantic contentfulness* underlies a wide range of hybrid cross-linguistic concepts **(Claim 2)**. In this thesis, I apply deep learning models to quantify semantic contentfulness and investigate how it underlies and organizes these cross linguistic concepts. Starting with simple, decontextual, word-vector based methods in Chapter 3, I then introduce a new, contextual measure, groundedness, which I investigate in Chapters 4 and 5. Moving beyond more straightforward applications of SSLMs within typology, it uses multimodal language models to provide a model of semantics which goes beyond the distributions of words alone. In Chapter 5, I demonstrate the potential of my approach to provide new cross-linguistic comparative concepts and insights, showing that groundedness can help explain edge cases of word class organization in a way that existing comparative concepts within typology cannot.

In **Chapter 3**, I focus on the concept of the **inflection-derivation distinction**. While a cross-linguistically consistent definition of the terms inflection and derivation has been elusive in theoretical linguistics, I demonstrate that by making linguist intuitions about the distinction concrete, much of the way the distinction has been made across languages can be explained, quantifying the cross-lingual consistency of these concepts. In so doing, I both show how linguistic intuitions about the distinction hold up in a corpus-based setting, and provide an indication of why these concepts have been so useful and attractive, despite issues with their cross-lingual descriptive validity.

I introduce a set of four quantitative measures of morphological constructions, including measures of both the magnitude and the variability of the changes introduced by each construction. Crucially, these measures can be computed directly from a linguistic corpus, allowing us to consistently operationalise them across many languages and morphological constructions. In experiments on 26 languages[1] (including five from non-Indo-European families) and 2,772 constructions, I find that both models are able to predict with high accuracy whether a held-out construction is listed as inflection or derivation in UniMorph. **This work has been published in 2024 in the Journal of Language Modelling and should require minimal adaptation for the thesis.**

In **Chapter 4**, I introduce *groundedness*, a new semantic-contentfulness measure based on multimodal models. Focusing on the domain of image captions, I am able to treat an image as a proxy for a caption's meaning. Using a language model and an image captioning model, I am able to estimate the pointwise mutual information between a token and the image as a surprisal difference under the two models. In this chapter, I focus on the **lexical-functional distinction** in parts of speech, showing that there is a broad degree of cross-linguistic consistency in the relationship between semantic content and parts of speech closely corresponding to the lexical-functional distinction.

Using image captioning data in 30 languages from 10 language families, I find this groundedness measure largely rediscovers the distinction between lexical and functional word classes across 30 languages. Further, though it correlates only weakly with norms like imageability and concreteness in English, it provides a ranking suggested by cognitive linguists between nouns, verbs, and adjectives (noun > adjectives > verbs) across languages but contradicts

[1] cat, ces, dan, deu, eng, ell, fin, fra, gle, hun, hye, ita, kaz, lat, lav, mon, nob, nld, pol, por, ron, rus, spa, swe, tur, ukr

the view of adpositions as a "semi-lexical" class. However, our results suggest grammatical word classes still carry semantic content. These results validate intuitions about word class contentfulness and suggest the utility of this measure as a general tool for studying contentfulness in linguistics, and of taking a grounded approach to typological problems. We release the model used to estimate our measure and a dataset of groundedness measures for further study. **This work has recently been presented as a main conference paper at NAACL 2025 (see attached). I intend to spend ≈ 2 weeks updating it and harmonizing it with the rest of the thesis.**

In **Chapter 5**, I show the relationship between visual groundedness and cross-linguistic variation in **parts of speech**. While parts of speech have often been considered a cross-linguistic universal, a range of different organisational principals appear to emerge cross-linguistically. I show that deviations from prototypical part-of-speech organizations (as established in Chapter 4) are associated with groundedness that corresponds with the deviation, indicating an iconic link between groundedness and form.

To establish this, I first investigated 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 hypothesise 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 investigated two manually captioned dataset and one machine translated dataset, finding significantly higher groundedness for *na*-adjectives in both manually captioned datasets.

While I see this as a core result of this chapter, I am continuing to work on additional experiments. Most essentially, I plan to spend the next 4 weeks conducting a large-scale, cross-lingual experiment about non-prototypical word-class use. Building on the word-level groundedness score dataset from Chapter 4, I am annotating the data with information about whether a word is a PROPERTY, ENTITY, CONCEPT, ACTION, EVENT, or OTHER–the protype functions of the major word classes (nouns, adjectives, and verbs). I am using BabelNet 4.0 to first annotate unambiguous instances of each label. I will combine BabelNet with XL-WSD to get more annotations for most of our languages and an evaluation set for those languages, including some ambiguous instances. I then need a method for classification. I will first try in-context learning with an off-the-shelf LLM. If performance is poor, I will train an XLM-R or Aya-101 based classifier on the XL-WSD train and BabelNet annotated data. After annotating the dataset, I will compare parts of speech with semantic class to see which is more strongly associated with groundedness. Further, I will investigate whether captions are more likely to use a non-prototypical part of speech (e.g. using a noun to describe an event) in more highly grounded contexts. Finally, while Japanese is a language where a major part of speech is "split", other languages in our dataset have major parts of speech "joined"–notably Korean (verbs and adjectives) and Tagalog (nouns and verbs). I plan to investigate the patterns of groundedness in these "joined" parts of speech more closely–possibly with some finer grained semantics, and compare to other languages–e.g. are Korean property words typically less grounded?

If time permits, I also hope to carry out some additional experiments on adpositions. In Chapter 4, I found that adpositions have very low average groundedness, despite having previously been described as "semi-lexical." I aim to break apart this heterogenous class. I will first focus on identifying spatial adposition use (following the same techniques used in the previous semantic class experiment with BabelNet) as distinct from more "abstract" adposition use, and investigate whether spatial adpositions are more grounded than others, and how they compare to other functional word classes cross-linguistically. If I find that spatial adpositions are not in fact more grounded, this may be due to the noted poor spatial reasoning capacities of VLMs. To test this, I could leverage the image segmentation data of COCO, using a classifier to probe for a set of basic spatial relations based on the captions and seeing whether correct identification of the spatial relation by the classifier is correlated with increased groundedness of the associated spatial adposition. I anticipate that the basic experiments here would take 2 weeks, with the spatial relation probing experiments taking 2-4 additional weeks.

It may be that I begin to run into limits of the quality of current groundedness scores, as the estimation can be somewhat noisy. I have two proposals for how to get improved score estimations. The first is to leverage advancements in multilingual multimodal models. Gemma 3 is capable of handling both mono- and multimodal inputs, and so using it to produce scores should be straightforward (compared to the fine-tuning routine required for PaliGemma). Additionally, I am considering running some experiments with LaBSE sentence embeddings as the meaning representation, rather than images. While they may be less language-agnostic than images in practice, they would enable this work to scale to domains beyond image captioning text and may help answer some of the research questions in the chapter. Using LaBSE as a meaning representation would require fine-tuning a simple model on top of a base LLM which takes the LaBSE embedding as a soft prompt. I plan to use Aya-101 for such an experiment. I am budgeting 2-4 weeks for producing an improved model as needed.

Altogether, then, I estimate I need 2-3 months to complete the experimental work for Chapter 5. I estimate 2-4 additional weeks to write these results into a full content chapter. After this, I estimate 1 month for writing the background chapter (for which I have been reading and taken notes, but have not started writing), 2 weeks for writing discussion and conclusion, and 1 month for other changes to the introduction and other content chapters. I therefore anticipate submitting in October of this year.

Chapter 2

Background

2.1 Typology and Comparative Concepts

One of the primary aims of typological linguistics is to identify cross-linguistic generalizations about the ways that languages (do not) vary. This aim of typology was first clearly articulated by **?**, who... While other aims of typology include Y and Z, this thesis is primarily concerned with Greenbergian typology. This type of typology shares a similar aim to the generative linguistics approach, in that both aim to identify and explain universal properties of human language. Yet they vary substantially in their approach to this aim. Generative linguistics usually couches its claims in terms of a formal system of grammar which generates the set of possible utterances in a language, and encodes universals or variation limits in the parameter and structure of this grammar. Similarly, this style of linguistic analysis typically proceeds through the analysis of individual languages, with each language viewed as an instantiation and test case for the grammatical framework. Changes that are made to the grammatical framework in one language are, in principle posited to be universally available in other languages, but perhaps inert in that language. Central to the generalizations in such frameworks is the notion of universal categories. A generative grammar gains its generative capacity by positing rules that apply based on the membership of a linguistic unit to a particular category, such that members of that category share behaviour. However, categorical behaviour can vary substantially across languages. For example, some languages clearly distinguish adjectives from both nouns and verbs

Greenbergian typological linguistics, on the other hand, typically proceeds from observations across a wide and varied sample of languages, and attempts to characterize the patterns of variation and invariances that are observed in this sample of languages.

- **Formal**
  - Briefly discuss places where formal concepts are useful
- **Hybrid Concepts**
  - Point out how most "formal" concepts have some kind of semantic component
  - Even if they don't within a specific language…
  - Debates on if a universal concept/category exists
- **Semantic Concepts**
  - Haspelmath
  - Semantic maps
- **Key Challenge for Functional Comparative Concepts**
  - Gradience of meaning/prototypicality of categories
  - What is the relationship between a functional comparative concept and language-specific categories?
    - Croft's construction and strategy solution
    - My solution (empirical measurement)

2.2 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.

2.2.1 Comparative concepts in typological 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 **?**, 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 (**?**).

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 **??** includes a detailed discussion of some of the 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 **?**, 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.

2.2.2 Phonological typology

The earliest and most succesful application of empirical measures of underlying linguistic dimensions to typology has been in phonology. In phonology, it has long been understood that language-specific categories make a poor basis for cross-linguistic comparison. In the words of **?**, "phonemes are not fruitful universals": because phonemes are defined by language-internal contrasts, mapping phonemes across languages with different sets of contrasts is ill-defined. **?** introduced a splitting approach to phonetic universals, developing a distinctive feature theory where phonemes are decomposed into a set of binary phonological features which may have variable phonetic relaization. **?** influentially adopted this approach, proposing the additional notion of *natural classes*, the idea that phonemes which share one or more features (a "natural class") tend to pattern together in phonological processes. However, the idea that these features or natural classes are cross-linguistically universal has been challenged: the same sound can pattern with a different natural class depending on the language (**?**). However, the biggest challenge to a discrete, featural view of phonology and phonological typology comes from vowels.

Vowels are typically described in terms of three articulatory dimensions: height (high, mid, low), backness (front, central, back), and roundedness (rounded, unrounded). However, the phonetic space of vowels is continuous, and the mapping from this continuous space to discrete categories is highly variable across languages. The study of vowel system typology is as old as the distinctive theory feature of phonology, with **?** describing a set of implicational universals about vowel systems within his proto-featural theory. Yet within this featural approach, 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 comparative concepts used in vowel typology at this time were either discrete approximate phonemes, or featural decompositions of these phonemes, both of which lost critical aspects of the underlying continuous space.

In the nineteenth century, Helmholz (**?**) and Hermann (**?**) developed the *formant theory* of vowels: from the study of the resonances of tubes and vibrating filters, they ascertained that vowel qualities correspond perceptually to spectral properties of the acoustic signal. This is to say that vowels are distinguished by peaks of frequency in their sound spectrum, known as *formants*, especially the lowest formants, F1, F2, and F3. The development of enhanced spectrographic techniques and formant estimation techniques through the mid-twentieth century aided the development of better theories of vowel perception (**?**). This enabled the gradual replacement of discrete phones and feature values as the comparative concepts for vowels cross linguistically, with real-valued dimensions consisting of the first three formants. This yielded new theories of vowel system typology. First, there was the *quantal theory* of vowels (**??**), which identfies regions in the vowel space where formants are stable, yielding regions of relative perceptual stability that Stevens hypothesized would be preferred in vowel systems. Contemporaneously **?** developed the *dispersion theory* of vowel systems, which posits that vowel systems are organized to maximize the distance between vowels in formant space. With the formant-based continuous space in which to operationalize their theory, **?** were able to simulate optimal vowel systems and compare them to attested systems, finding that their optimal systems very closely matched attested systems.

Critically, the precise nature of the predictions given by **?**'s computational modelling approach demonstrated some critical ways attested systems diverge from the predictions of the dispersion theory (notably with incorrect predictions around central vowels). This led to the development of the *dispersion-focalization theory* of **??**, which combined the dispersion theory with aspects of the quantal theory. In so doing, they were not only able to better predict attested vowel systems, but also to provide precise statements about the relative effects of dispersion and focalization in attested systems. Finally **?** were able to provide a more complete model of vowel system typology by producing a model of the *full distribution* of attested vowel systems, rather than just the optimal systems. This provided further insights into the relative importance of different pressures on vowel systems.

At first blush, it may seem that these developments in the area of phonological typology and are of little relevance to the present work–vowels are ultimately formal and material, while I aim to study linguistic function, which is more abstract. While we can record the acoustic instantiation of a vowel to a close approximation, there is no equivalent physical instantiation of linguistic function. However, I argue that the issues are a matter of degree, not of kind, and that, therefore, the history of theories of vowel typology are instructive for what future theories of meaning in typology will look like. In particular, I believe the deep learning approach is enabling for developing an empirical underlying space for meaning, analogous to the development of the formant theory of vowels. Having an empirical underlying space makes all typological generalizations precise and testable, subject to the linking hypothesis between the empirical space and the cognitive/perceptual space. Similarly to how progress on the link between formant space and perceptual space occurred outside of and parallel to the development of vowel system typology, advances in the link between deep learning models and human semantic processing are substantial and ongoing (**??**). In the next section, I will show convergent evidence from the typology of semantic category systems that developing an empirical underlying space can enable substantial progress in understanding typological patterns and new types of typological generalizations.

2.2.3 Semantic category systems

The most analogous area of study to the vowel typology described above with respect to linguistiyic function is the study of so-called semantic category systems. In this literature, the partitioning of a complex underlying semantic space into discrete categories is studied cross-linguistically. Similarly to the case of vowel systems, there is an attempt to explain the cross-linguistic patterns in terms of simple, universal pressures on language as a system: economy/simplicity, iconicity, ease of learning, communicative fidelity, etc. With access to an underlying semantic space, these pressures can be tested through similar techniques as with vowel systems. Us can identify the optimal trade-off between several of these pressures, and compare attested systems to this optimum. Studies in this area also frequently rely on the simulation of hypothetical sub-optimal systems to compare to attested systems, in order to test the hypothesis that attested systems are more optimal than chance. Different pressures can be compared to see which ones are most predictive of attested systems.

This type of study heavily relies on the establishment of a detailed underlying semantic map. I will first focus on how the development of this underlying conceptual map for colour systems has enabled substantial progress in our understanding of universals in colour systems, because of the strong universals that were previously known in this domain, and how they were advanced by the development of an underlying space.

In the seminal study of **?**, the authors established vowel-system like generalizations about colour systems. In the same way that systems with a two-way vowel contrast are almost always based on a height contrast (e.g. /i/ vs. /a/), color systems with two basic colour terms seem to universally be based on a light-dark contrast. As three vowel systems are almost always /i/, /a/, and /u/, three-term color systems are almost always white, black, and red. Berlin and Kay couched these generalizations in terms of an implicational hierarchy of color terms, quite similar to the implicational hierarchies common in the typological literature more broadly. However, the authors could offer little in the way of explanation for these implicational hierarchies with the methodology they employed.

The foundational work on color system typology (**??**) uses discrete color chips to facilitate cross-linguistic elicitation. However, analogous to the progress in vowel system universals, the development of a detailed map of the underlying perceptual space revealed deeper universals about color system typology. Studying the relationship between the color space in terms of frequency and perceptual distance, **?** conjectured that this implicational hierarchy of color terms was based by the perceptual non-uniformity of the color space, with the universals relating to "peaks" in the perceptual color space—essentially, a color-system version of **?**'s quantal theory of vowels. However, to test such a hypothesis against the complex attested systems required a precise working definition of the underlying perceptual space and a computational formalization of the pressures on color systems. The first component was supplied by the development of the CIEL*a*b* color space (**?**), which transforms a color space based on physical properties of light to better align with facts we know about human color perception. Using CIEL*a*b*, **?** were able to model optimal systems with respect to the similarity within categories and dissimilarity between categories, finding that their predicted systems closely matched attested systems, and that attested systems were much more optimal than unattested systems. This initial effort has expanded into a rich literature, studying the trade off of different types of pressures such as communicative need, ease of learning, and modelling the change in systems over time. 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 color systems.[1]

Beyond color, other domains have been studied where an underlying map defining a distance metric and possible systems can be defined. These include kinship, number, quantifiers, modals and personal pronouns—all domains with an underlying discrete structure, making the construction of a map more straightforward. Two domains which are not underlyingly discrete have also been studied: spatial terms (**?**) and morphological marking **?**. However, these studies rely on a simplification of the underlying space. For example, **?** simplify the complex space of tense and aspectual marking semantics into 7 discrete

[1] See Appendix **??** for my own contribution to this literature during my Ph.D.

categories based only on the distance from the present. Further, using distance metrics based on a discrete map of the space enforces an assumption that all adjacencies in the discrete space are equally close–but this may not be the case. **?**'s study constructs an underlying 3-dimensional semantic map for spatial relations based on multidimensional scaling of term usage in a variety of scenarios and languages as collected by **?**. This direction is potentially promising for future research into other difficult to operationalize domains, but comes with the caveats of multi-dimensional scaling research I will discuss in the next section.

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.

2.2.4 Multidimensional scaling

The multidimensional scaling (MDS) approach to semantic maps proposed by **?** 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 **?**,[2] and to provide a stronger mathematical basis for the semantic map theory. These methods take in some input matrix of dissimilarities between linguistic items, and use matrix decomposition methods to provide a low-dimensional approximation. 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. **?** 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 **?** used to recreate **?**'s analysis of indefinites.Other examples of this type of map include... This type of map takes a set of N linguistic forms F, and a set of K underlying functions M. The binary input matrix $\mathbf{I} \in \{\mathrm{Y}, \mathrm{N}\}^{K \times N}$ is constructed such that

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

From this, an $N \times N$ similarity matrix $\mathbf{D}$ among the functions is constructed as:

$$\mathbf{D}_{i,j} = \big(\sum_{p=1}^{K} \mathbf{1}[I_{p,i} = I_{p,j} = \mathrm{Y}]\big)/K$$

to which the MDS technique is applied. In this type of analysis, the function is entirely manually posited by the typologist and abstracted away from the

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

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-occurences is modelled. Nevertheless, in terms of comparative concepts, the functions here retain the traditional approach to function in typology, with its known shortcomings.

? 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 **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:

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

The lessened reliance on manually posited functions allows for a 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 **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 study 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.[3]

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-occurence 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.

2.2.5 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 for typology. Recently, **?** used multilingual BERT (**?**) and Aya (**?**) to study animacy

[3] A wide range of domains and phenomena have been studied with this approach.

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.

? 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* (**?**), 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. **?** 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 nacent 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 touched in any of this literature, likely due to the difficulty of making precise comparative concepts for it beyond frequency.

2.3 The Distinctness of Lexicon and Syntax

> ...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 (⊠⊠)* of Chen Kui (⊠⊠) catalogued *zhùchí* ⊠⊠ (lit. "helping words")–corresponding to what we would today call function words. In the Y⊠zhù (1311) Lu Yiwei defines this class as words that do not have a "precise concrete meaning" (**?**), and future authors would adopt the term *y⊠cí* ⊠⊠ (lit. "empty words") to refer to this class (**?**). 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 this section, I review the differences in cross-linguistic behaviours of lexical and functional elements,

2.3.1 Splitting the dimensions of the distinction

In the introduction to this section, I deliberately used the vague terms "items," "elements," and "units" as the locus for the lexical–functional distinction. This is because this gradient is relevant at different levels of formal linguistic structure, and often goes by different names at different levels. Indeed, a critical aspect of the linguistic interest in this distinction is the *correlation* between the level of structure and the semantic distinction. For example, more grammatical meanings are more likely to be expressed as bound morphemes, while more lexical meanings tend to be expressed as whole words.[4] First I will discuss the

[4] Of course, "bound" and "word" are also problematic terms, which I will unpack further shortly.

gradient levels of formal structure, and then the gradient levels of semantic content. Finally, I will discuss the interrelation between the two.

2.3.1.1 The formal dimension

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.

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 **??**, 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 themeselves 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).[5] 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 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 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 [⊠z] depending on the phonological context). The more of these processes apply, the more phonologically bound the morpheme is. **?** argue that

[5] In 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.

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 instancesl. *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.

2.3.1.2 The semantic dimension

The semantic dimension

Relationality A major part of this distributionalsem

2.3.2 Relevant phenomena at different formal levels

2.3.2.0.1

2.3.2.0.2 The functional dimension

2.3.3 Theories that reify the distinction

2.3.4 Doing without the distinction

2.3.5 Evidence from psychology and neuroscience

2.3.6 Computational operationalization

The notion that syntax exists (colorless green ideas)

Structure of theories

- Syntax
- Morphology

Basic psychological and neurological evidence

- Syntax
- Morphology

Cross-linguistic comparison, labelling and describing data

- UniMorph, universal derivations

2.4 The Gradient Nature of Lexicality

- Information packing (syntactic conversion) vs. semantic content
    - Inflection and derivation
- Polysynthesis
- Adpositions, light nouns, light verbs
- Grammaticalization

2.5 Finding Meaning in Computational Models

- SkipGram family of approaches?
- Evidence of semantics
- Measurement (length, differences)
- Type level = no prototypicality :/
- Entanglement of form and function
- Briefly, conventional contextual representations
- Token-level, but still entangle form and function
- Challenge of finding semantics
- Evidence of semantics
- Multimodal models
  - How they work, grafting a vision model onto a language model
  - How the separate modality provides an avenue for language-agnostic function

Chapter 3

Corpus-based Measures Discriminate Inflection and Derivation Cross-Linguistically

??

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, 2007a). 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 languages (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 organises 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 categorisations – such as the degree of syntactic or semantic change, or the creation of new words. As an example, Plank (1994a) covers no fewer than 28 tests for inflectional and derivational status. Upon applying them to just six English morphological constructions, Plank (1994a) 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

[1] https://www.wiktionary.org/

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).

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, 1994a). 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 operationalising 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 paper, 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 operationalise 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, 1994a).

We then ask whether, for a given construction, knowing just these measures is sufficient to predict its inflectional versus derivational status in UniMorph. In other words, 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.

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 3.7, 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 characterisation 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.

3.2 Motivation for our measures

In order to explore our question of interest, we need to operationalise 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 characterised 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-emphasises 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 generalises 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 characterise 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., 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.

[2] This is suggested, though not explicitly, by criteria like Plank (1994a)'s "derivational morphemes resemble free morphs."

However, these vectors also capture substantial amounts of information about a word's syntactic category, as operationalised 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 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., 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 (1994a) claim that formal variablity 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.

Another thread of research inspiring this particular factorisation 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), as well as in unsupervised morphological segmentation, which typically covers both inflection and derivation (Schone and Jurafsky,

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 nominalisation with *–ion*) in our data

2000; Soricut and Och, 2015; Narasimhan et al., 2015; Bergmanis and Goldwater, 2017).

Because debates about inflectional and derivational status typically focus on *constructions* such as "the nominal plural in German" or "the addition of the *–ion* nominalisation 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 morpho-syntactic 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 3.8.1.

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:

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

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 $(\textit{tanzen}, \textit{getanzt})$ yields the edit template $\mathrm{ge\_XXt}$, 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 $\mathrm{\_X\ddot{o}\_e}$. 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 $+\mathrm{FRONT}$ to the vowel's features, which would cover the templates $\mathrm{\_X\ddot{o}\_e}$, $\mathrm{\_X\ddot{a}\_e}$, and $\mathrm{\_X\ddot{u}\_e}$. 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

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

For example, suppose we have a morpheme with two edit templates: $\_\text{as}$, used 80% of the time, and $\_\text{os}$, used 20% of the time. Then this measure would be $0.8 \cdot 0.2 \cdot \textsc{EditDistance}(\_\text{as}, \_\text{os}) + 0.2 \cdot 0.8 \cdot \textsc{EditDistance}(\_\text{os}, \_\text{as}) = 0.32$. 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 Crawl[4] and Wikipedia data, which was 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., 2020; Bojanowski et al., 2017), and are more commonly used as models of semantics than syntax.[5] However,

[3] https://fasttext.cc/docs/en/crawl-vectors.html

[4] https://commoncrawl.org/

[5] Recent studies have shown that embeddings from newer large language models such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) correlate even better than FastText

there is evidence from the literature in unsupervised part-of-speech tagging (He et al., 2018; Lin et al., 2015) and probing (Pimentel et al., 2020; Babazhanova et al., 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 3.8.1.

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., 2013) 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

embeddings with human judgements of semantic similarity (Bommasani et al., 2020; Vulić et al., 2020). 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. (2020) 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.

[6] Indeed, our own supplementary results suggests that these vectors encode substantial syntactic information, and that the addition of gold-standard syntactic category information provides little benefit over our proposed model. For further information, please see Section 2 of the supplementary material at https://osf.io/uztgy/.

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:

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

While cosine distance is more frequently used than Euclidean distance for semantic 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 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 $\mathbf{D} = [\mathbf{d}_1 \ldots \mathbf{d}_N]$. 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:

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

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 3.6, 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 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 deriv-

[7] https://en.wiktionary.org/

> ations 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 3.8.1, while noting here that our priority is to 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

[8] We 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.

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 morpho-syntactic/semantic features) in the present work.

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 **??** 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.

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

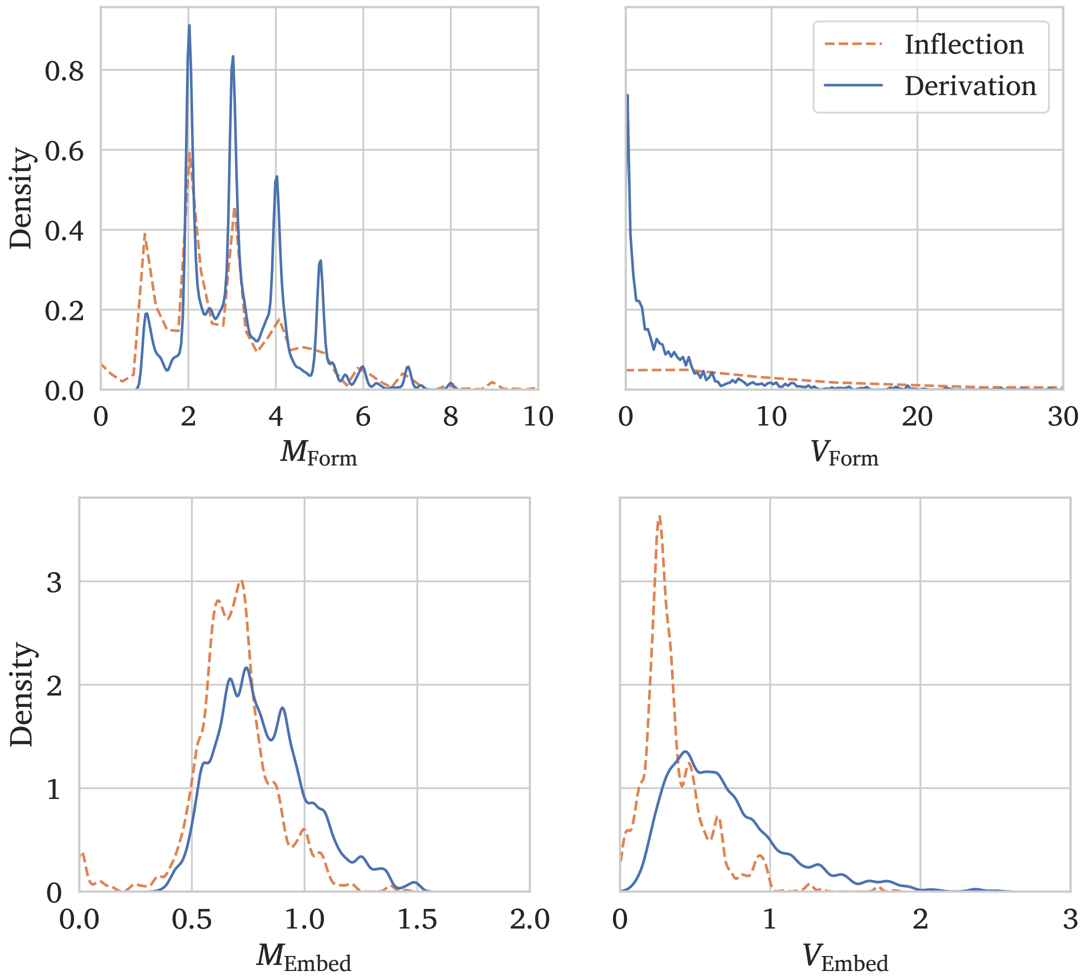

Figure 3.1: The empirical distributions of our four measures (quantifying the magnitude M and variability V of changes in Form and in Embedding space) for inflections and derivations in UniMorph

derivation than inflection. However, on average, V_{Form} is *lower* for derivation than for inflection – the opposite of what is suggested by Plank (1994a). This is discussed in Section 3.8.1.

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 of 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.

[9] The 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.

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 utilising a per-construction level of analysis: the effects we find here cannot be explained by frequency of constructed forms.

3.6 Predicting inflection and derivation

In this section, we investigate how well the characterisation 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 3.7 and Section 3.8.2. 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 characterise inflection- and derivationhood. First, we consider how much any **single** feature recovers the distinction from UniMorph. Secondly, we consider several combinations of two features: (A). **just variability** $(V_{\textbf{Form}}, V_{\textbf{Embed}})$: 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** $(M_{\textbf{Form}}, V_{\textbf{Form}})$ or (D) **syntax/semantics** $(M_{\textbf{Embed}}, V_{\textbf{Embed}})$ 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 characterisation of the distinction. If there are instead large regions with intermediate values, this corresponds to a gradient characterisation of the distinction.[10] 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 3.2, the logistic classifier results best support the **just variability** hypothesis (A),

[10] This issue of whether the distinction is gradient or categorical with respect to our measures is discussed further in Section 3.8.4.

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 optimiser (**?**) 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 categorisation of inflection/derivation, all four features contain unique information relevant to recreating this distinction (Hypothesis E).

3.7 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 operationalisation of the distinction (UniMorph 4.0) may also be of interest in the ways in which it *differs* from that operationalisation. Accordingly, in this section, we look at the trends in how our model classifies constructions which are labelled as inflection in UniMorph.

	Features				Accuracy (▨ = Logistic, ⊡ = MLP)
	Majority class (Inflection)				▨ 0.57
	M_{Form}	–	–	–	▨ 0.58 ± 0.01 ⊡ 0.58 ± 0.01
	–	M_{Embed}	–	–	▨ 0.66 ± 0.01 ⊡ 0.66 ± 0.01
	–	–	V_{Form}	–	▨ 0.68 ± 0.01 ⊡ 0.68 ± 0.02
	–	–	–	V_{Embed}	▨ 0.73 ± 0.01 ⊡ 0.74 ± 0.01
(A)	–	–	V_{Form}	V_{Embed}	▨ 0.83 ± 0.01 ⊡ 0.83 ± 0.01
(B)	M_{Form}	M_{Embed}	–	–	▨ 0.67 ± 0.01 ⊡ 0.67 ± 0.01
(C)	M_{Form}	–	V_{Form}	–	▨ 0.69 ± 0.01 ⊡ 0.73 ± 0.01
(D)	–	M_{Embed}	–	V_{Embed}	▨ 0.75 ± 0.01 ⊡ 0.78 ± 0.01
	M_{Form}	–	–	V_{Embed}	▨ 0.73 ± 0.01 ⊡ 0.75 ± 0.01
	–	M_{Embed}	V_{Form}	–	▨ 0.73 ± 0.01 ⊡ 0.73 ± 0.01
	M_{Form}	M_{Embed}	V_{Form}	–	▨ 0.73 ± 0.01 ⊡ 0.77 ± 0.01
	M_{Form}	M_{Embed}	–	V_{Embed}	▨ 0.76 ± 0.01 ⊡ 0.81 ± 0.01
	M_{Form}	–	V_{Form}	V_{Embed}	▨ 0.83 ± 0.01 ⊡ 0.84 ± 0.01
	–	M_{Embed}	V_{Form}	V_{Embed}	▨ 0.83 ± 0.01 ⊡ 0.85 ± 0.01
(E)	M_{Form}	M_{Embed}	V_{Form}	V_{Embed}	▨ 0.83 ± 0.01 ⊡ 0.89 ± 0.01

Figure 3.2: 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

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.

3.7.1 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 semantically marked across languages(Greenberg, 1966b; 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,[11] 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 3.3 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 model (using all 4 features) in these results, which are presented in Figure 3.3

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

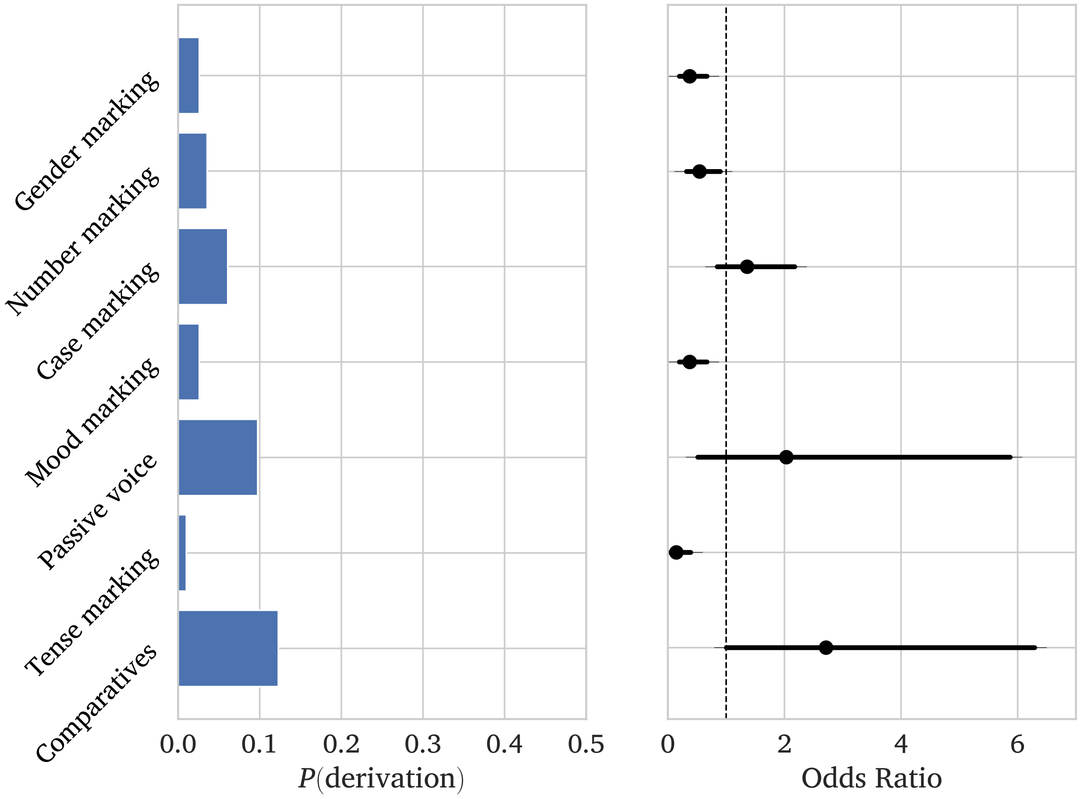

Figure 3.3: Probability and Odds ratio with 95% confidence intervals of being classified as derivation for various kinds of inflectional meaning. Inflections to the right of the dotted line were disproportionately likely to be classified as derivation by our model

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.[12] 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.

3.7.2 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: inherent inflection and transpositions, for which results are displayed in Fig-

[12] For example, they are listed as derivations in English, but as inflections in German.

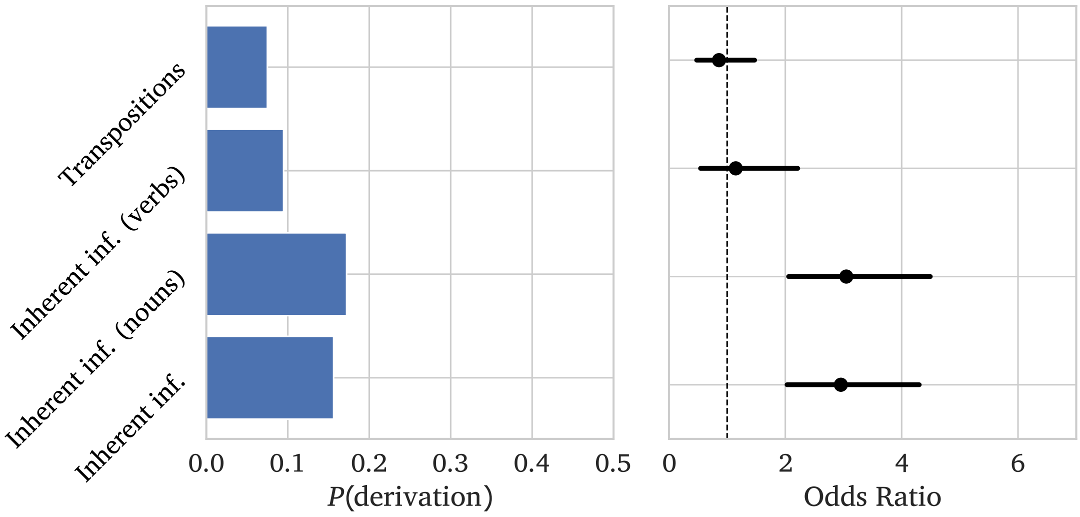

Figure 3.4: Probability and Odds ratio with 95% confidence intervals of being classified as derivation for inherent inflections and transpositions

ure 3.4.

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 operationalise this in a simple, cross-linguistically consistent way, we associate number, gender, and case[13] 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 3.4, we find that inherent inflectional constructions are not more likely to be classified

[13] Booij (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.

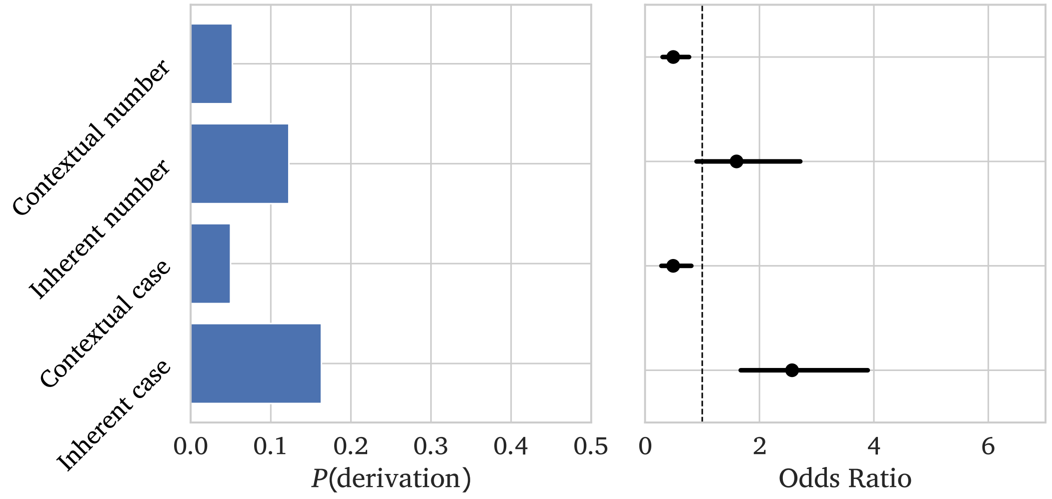

Figure 3.5: Probability and Odds ratio with 95% confidence intervals of being classified as derivation for inherent vs. contextual noun inflections

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 3.5, 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 3.4. Transpositions have often been argued to be non-canonical inflection or even derivation because transpositions change the part of speech (Spencer, 2013; Plank, 1994a; 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 or 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.

3.7.3 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.

3.8 Discussion

3.8.1 The role of our individual measures

As shown in Section 3.6, all four of our measures can be used to achieve better discrimination between traditional concepts of inflection and derivation; however, 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 categorisations 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 (1994a) and Dressler (1989). 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 $(M_{\text{Form}}, M_{\text{Embed}}, V_{\text{Embed}})$ 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 (Pearon'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 useful.[14] 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.

3.8.2 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

[14] For more details about these experiments, see the supplementary material at https://osf.io/uztgy/.

morphological traits, as well as Armenian, Latvian, Irish, and Greek, covering 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.

3.8.3 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 measures. 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.[15] 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.

3.8.4 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,

[15] As described in Section 3.8.4 and shown in Figure 3.6, it is this situation in which we find ourselves.

and processing implicitly or explicitly employ the "split morphology hypothesis," 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; Štekauer, 2015; 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 3.8.3, 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 principle 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

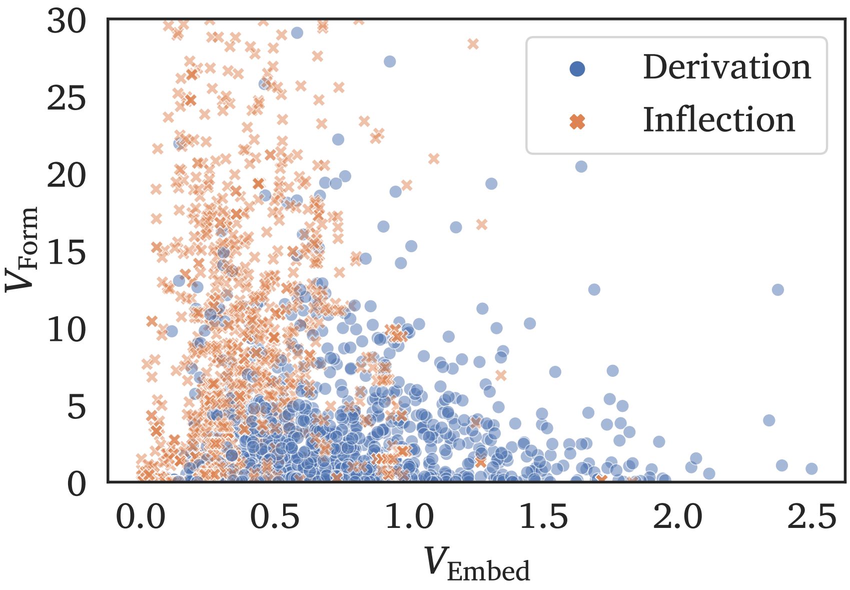

Figure 3.6: Our two most predictive measures for inflection and derivation. Saturation represents overlapping constructions. With respect to these two variables, the inflection–derivation distinction appears gradient rather than categorical

predictive measures, as shown in Figure 3.6. Recall that a logistic classifier using only these 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 operationalise 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.

3.8.5 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 morpho-syntactic 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 categorise 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 SIGMORPHON 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 lemmatisation (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.

3.8.6 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 3.8.2), 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 normalisation 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 (Kyjánek 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 harmonisation 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. (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 3.8.1), or other as-of-yet undiscovered systematic biases.[16]

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. (2018) and Ravfogel et al. (2020) – though there are significant challenges to operationalising 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 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

[16] See 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.

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 decontextualisation 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.

3.9 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 hypothesised 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 categorisations 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 3.6, 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, 1994a; 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 *generalise* to unseen constructions.

With these measures, we are also able to identify in a quantitative way *how canonical* different categories of inflections are (Section 3.7) 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. 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 3.8.4). 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.

3.10 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 human semantic similarity scores has been extensively studied (Vulić et al., 2020; Bojanowski et al., 2017), they are not usually utilised to capture syntactic similarity. Indeed, some studies have attempted to produce more syntactically-aligned embeddings from vectors like FastText (He et al., 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 figures/pos_dist.png

Figure 3.7: The mean cosine similarity between FastText embeddings of words of the same and different parts of speech in UniMorph.

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 3.7. 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).

To more directly study the syntactic information captured by our embedding-based measures, we fit a logistic regression classifier which uses the two embedding measures $(M_{\text{Embed}}, V_{\text{Embed}})$ 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 choose to use a logistic regression classifier because our findings in Section 6 indicate that an MLP may not be necessary for these features, and it is less prone to spurious overfitting than an MLP. 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.

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 derivations. 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.

Chapter 4

Groundedness and the Lexical–Functional Distinction

4.1 Introduction

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

While empirically grounding language form (surface structure like vowels) is typically straightforward, language is not just a formal system, but also a functional one. Many questions within typology relate to the relationship between form and *meaning*, especially in domains like morphology and syntax. Typically, typologists manually identify semantic/functional roles such as "subject", and

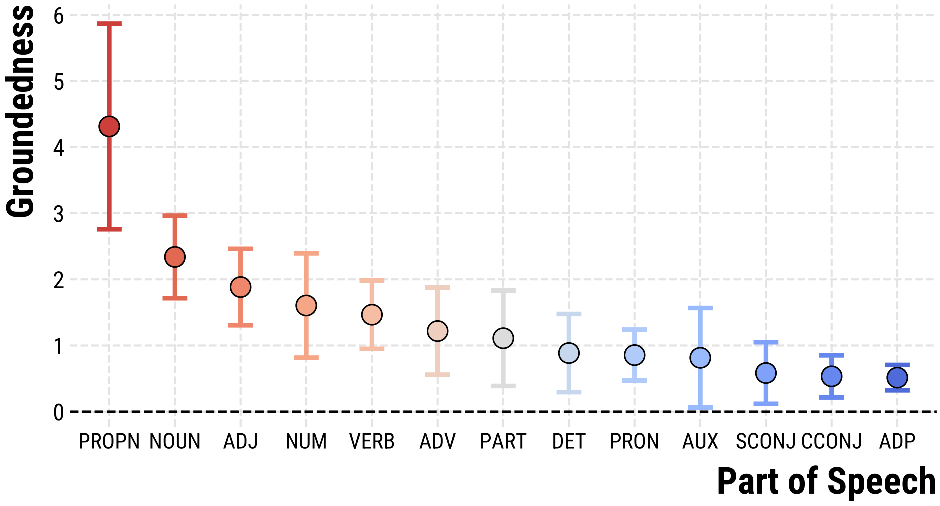

Figure 4.1: Mean and standard deviation of per-language mutual information estimates between word class and image. Across 30 languages, we see clear and consistent tendencies about which parts of speech are more "grounded", corresponding to a distinction between lexical and functional classes.

"causative" and study their expression across languages (Haspelmath, 2010; Greenberg, 1966a). Unlike with many definitions based on form, definitions based on meaning are left up to subjective discretion, leading to debates which reduce to the definition of particular terms cross-linguistically (Haspelmath, 2007, 2012; Plank, 1994b).

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

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

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

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

4.2 Background

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

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

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

4.2.1 Contentfulness and word class

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

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

4.2.2 Measuring contentfulness

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

While theoretical linguistics has focused on a distinction between content and function words, psycholinguistics has focused on semantic dimensions like imageability, concreteness, and strength of perceptual experience. Measures of these dimensions have relied on subjective, decontextualized human judgements, but nevertheless predict processing differences between word classes, such as asymmetries in the processing of nouns and verbs in certain aphasias (Bird et al., 2003; Dubé et al., 2014; Lin et al., 2022). Because we operationalize meaning as images, notions such as imageability seem especially related to our groundedness measure. However, as discussed in Section 4.5.4, these concepts differ from our measure in that informativity is not a major factor in their definition. For example, while both "zebra" and "woman" are highly concrete nouns, the former has higher groundedness on average, because although both are often strongly associated with an image, "zebra" is more informative/surprising, especially if the image is unavailable—thus, the image adds more information in that case.

As shown by the prior example, our measure is also closely related to another concept widely studied in computational psycholinguistics: *surprisal.* Like our groundedness measure, surprisal has an intuitive link to contentfulness from an information theoretic perspective, and has been extensively studied in relation to processing difficulty (Hale, 2001; Levy, 2008; Smith and Levy, 2013; **?**; Staub, ming). 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).

4.3 Method

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

$$\mathrm{PMI}(w_t; m \mid \mathbf{w}_{<t}) = \log \frac{p(w_t \mid m, \mathbf{w}_{<t})}{p(w_t \mid \mathbf{w}_{<t})} \tag{4.1}$$

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 4.1 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; **?**). Further, a major strength of images as a meaning representation is that estimating both quantities in Equation 4.1 becomes straightforward with neural models: $p_{\boldsymbol{\phi}}(w_t|m, \mathbf{w}_{<t})$ corresponds to the probability of the token under an image captioning model, while $p_{\boldsymbol{\theta}}(w_t|\mathbf{w}_{<t})$ corresponds to its probability under a language model.

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

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

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

$$\begin{aligned} &p(\mathcal{C}_i, m \mid \mathbf{w}_{<t}) = \\ &\qquad \sum_{w_t \in \mathcal{W}} \left[p(\mathcal{C}_i \mid w_t, \mathbf{w}_{<t}) p(w_t, m \mid \mathbf{w}_{<t}) \right]. \end{aligned} \tag{4.2}$$

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

$$I[\mathcal{C}_i; m | \mathbf{w}_{<t}] = \mathop{\mathbb{E}}_{p(\mathcal{C}_i, m, \mathbf{w}_{<t})} \left[\log \frac{p(w_t | \mathbf{w}_{<t}, m)}{p(w_t | \mathbf{w}_{<t}))} \right]. \tag{4.3}$$

Given our factorization of the joint, we can perform a Monte Carlo estimation of the expectation by simply averaging groundedness over all the tokens tagged

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

Table 4.1: We match the data points on which the language model and image captioning model were trained. The three datasets are the Gemma pre-training mixture (PT), PaliGemma multimodal data for continued training (CT), and COCO image–caption pairs for fine-tuning (FT). Symbols indicate whether models are trained on text data (**A**) or on multimodal data (🖼**A**).

with $\mathcal{C}_i$ in the data $\mathcal{D}$:

$$\hat{I}[\mathcal{C}_i; m \mid \mathbf{w}_{<t}] = \sum_{(m,\mathbf{w}_{<t})\in\mathcal{D}} \frac{\mathbb{1}_{\mathcal{C}_{w_t}=\mathcal{C}_i} \log \frac{p_{\phi}(w_t|\mathbf{w}_{<t},m)}{p_{\theta}(w_t|\mathbf{w}_{<t})}}{\sum_{w_t\in\mathcal{D}} \mathbb{1}_{\mathcal{C}_{w_t}=\mathcal{C}_i}} \tag{4.4}$$

where $\mathbb{1}_{\mathcal{C}_{w_t}=\mathcal{C}_i}$ is 1 when a token's class is $\mathcal{C}_i$ 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.

4.4 Experimental setup

4.4.0.0.1 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 (**?**), and the CC3M-35L dataset consisting of 3 million image-caption pairs in each of 35 languages (Thapliyal et al., 2022).

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

4.4.0.0.2 Language model $p_{\boldsymbol{\theta}}(w_t|\mathbf{w}_{<t})$ Our aim is to use a language model as similar to our captioning model $p_{\boldsymbol{\phi}}(w_t|\mathbf{w}_{<t}, m)$ 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 *cap-*

tions 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 4.1 summarizes the data matching between the two models. Further implementational and POS tagging details are in Appendix **??**.

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

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

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

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 groundedness measure across languages and word classes. We begin by examining which word classes exhibit significant groundedness (Section 4.5.1), followed by an analysis of cross-linguistic trends and their consistency (4.5.2 and 4.5.3). Finally, we relate our findings to contentfulness-related psycholinguistic norms (4.5.4).

4.5 Results

4.5.1 Which word classes are grounded?

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

To compute significance levels, we use a one-sample permutation test. Taking the set of PMIs for a part of speech (POS) in a language, we sample up to 500 PMIs at a time from all datasets and randomly permute their signs (assign + or - with equal probability to each PMI value), then average these values to produce a new estimate of mutual information (MI). We repeat this process to produce 10^5 permuted estimates. By measuring how often our estimate based on the observed data is greater than the permuted estimate, we obtain the p-value,[3] i.e., the probability that our observations would have occurred under the null hypothesis of MI = 0.

Results are shown in Figure 4.2. Overall, the results suggest most or all

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

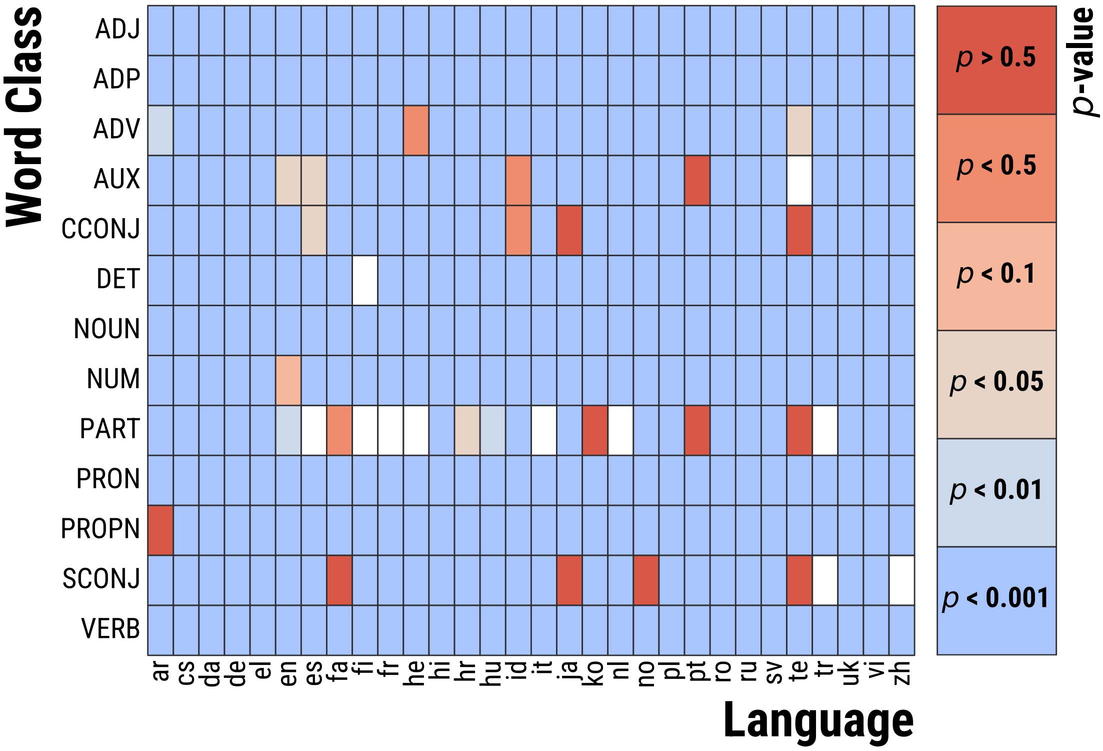

Figure 4.2: Heatmap of mutual information estimates across parts of speech in thirty languages. Cells show the statistical significance of a word class's groundedness (MI > 0). Unattested classes are white. Some functional classes display non-significant levels of groundedness in several languages, while lexical classes dominantly show highly significant grounding.

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 t raditionally 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.

4.5.2 Which word classes are more grounded?

We hypothesize that the cross-linguistically consistent trends in word class groundedness correspond to a cline which is a continuous analogue of the lexical–functional word class distinction. To isolate the contribution of word class identity to mutual information cross-linguistically, we compute estimated marginal means (EMMs) for each word class's groundedness,[4] and perform a post-hoc pairwise comparison test of the means.[5] The results of this analysis are displayed in Figure 4.3. All pairwise comparisons except between pronouns and particles are statistically significant, leading to a near total ranking of word classes. We find that lexical word classes (Proper nouns, nouns, adjectives, verbs, numbers, and adverbs) have higher groundedness than functional word classes (particles, auxiliaries, conjunctions, determiners, and adpositions), with pronouns ranking together with particles at the upper end of the functional categories. The ranking corroborates ideas from cognitive linguistics which place nouns, adjectives, and verbs along a lexical–functional continuum, with nouns > adjectives > verbs (**?**). 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

[4] Averaged over values of language and dataset.

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

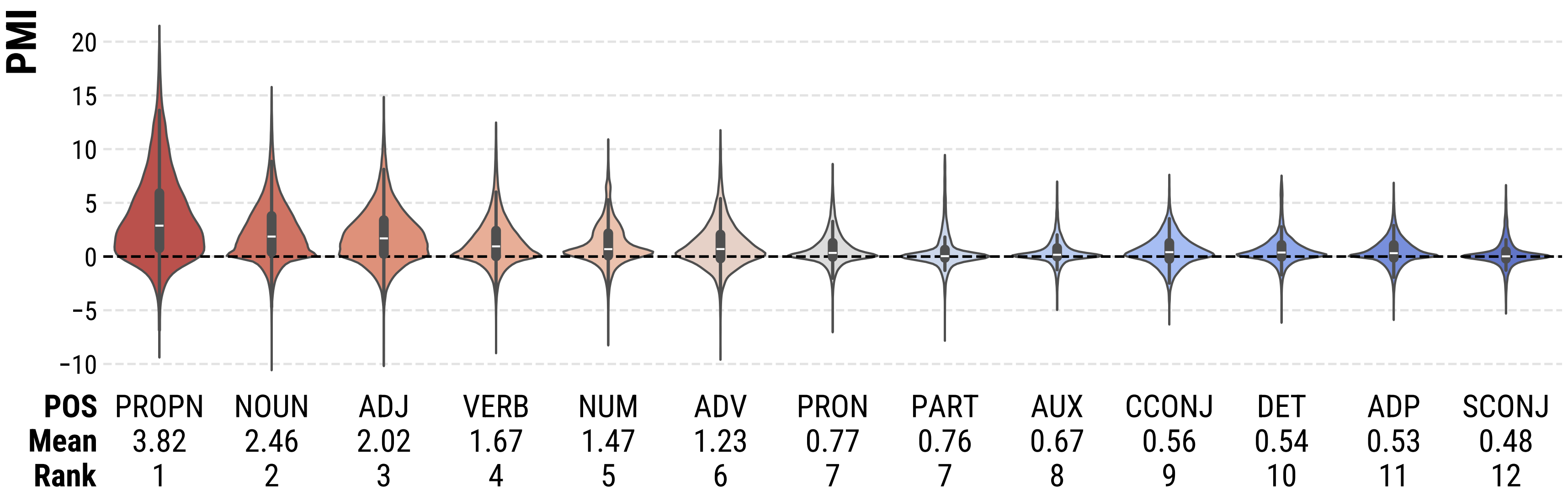

Figure 4.3: Word token level distributions of the groundedness measure (PMI) across all languages and datasets, grouped by part of speech (word class). We also report the estimated marginal mean and ranking of each word class. Colors are based on the ranking of classes, rather than their average PMIs. Overall, the distribution and estimated ranking of word classes strongly suggest our groundedness measure quantitatively captures the distinction between lexical and functional classes.

information for other functional classes, suggesting they could be more meaning-bearing than traditionally viewed.

4.5.3 How consistent is word class groundedness across languages?

We quantify the strength of the association between groundedness and word class on two levels: language-level MI estimates (Figure 4.1), and token-level PMI (Figure 4.3). The first level quantifies how consistent languages are in the groundedness of word classes, while the second level quantifies how much word class drives the groundedness of individual tokens. In both cases, we use ANOVA to estimate the amount of the variance in groundedness explained by word class.

4.5.3.0.1 MI estimates For the language-level MI estimates in Figure 4.1, we consider the separate effects of language, dataset, and POS on groundedness. Because the meanings (images) are matched across languages, this allows us to estimate and control for some languages having consistently larger or smaller MI estimates (due to language-specific variation in our neural estimators). We find significant effects of all 3 factors, but they differ dramatically in how much variation they explain. The effect of dataset is extremely small, explaining 0.5% of the observed variance ($F_{3,816} = 5.71$, $p < 0.01$). Language identity has a larger effect, explaining 8.2% of the variance ($F_{29,789} = 6.42$, $p < 0.001$). However, word class dominates, explaining most of the total variance (57.3%, $F_{12,806} = 775$, $p < 0.001$), and 62.8% of the remaining variance after controlling for variance due to dataset and language. Altogether, these factors explain 65.6% of the variance, leaving the remaining variance to cross-linguistic differences in the MI of specific parts of speech.

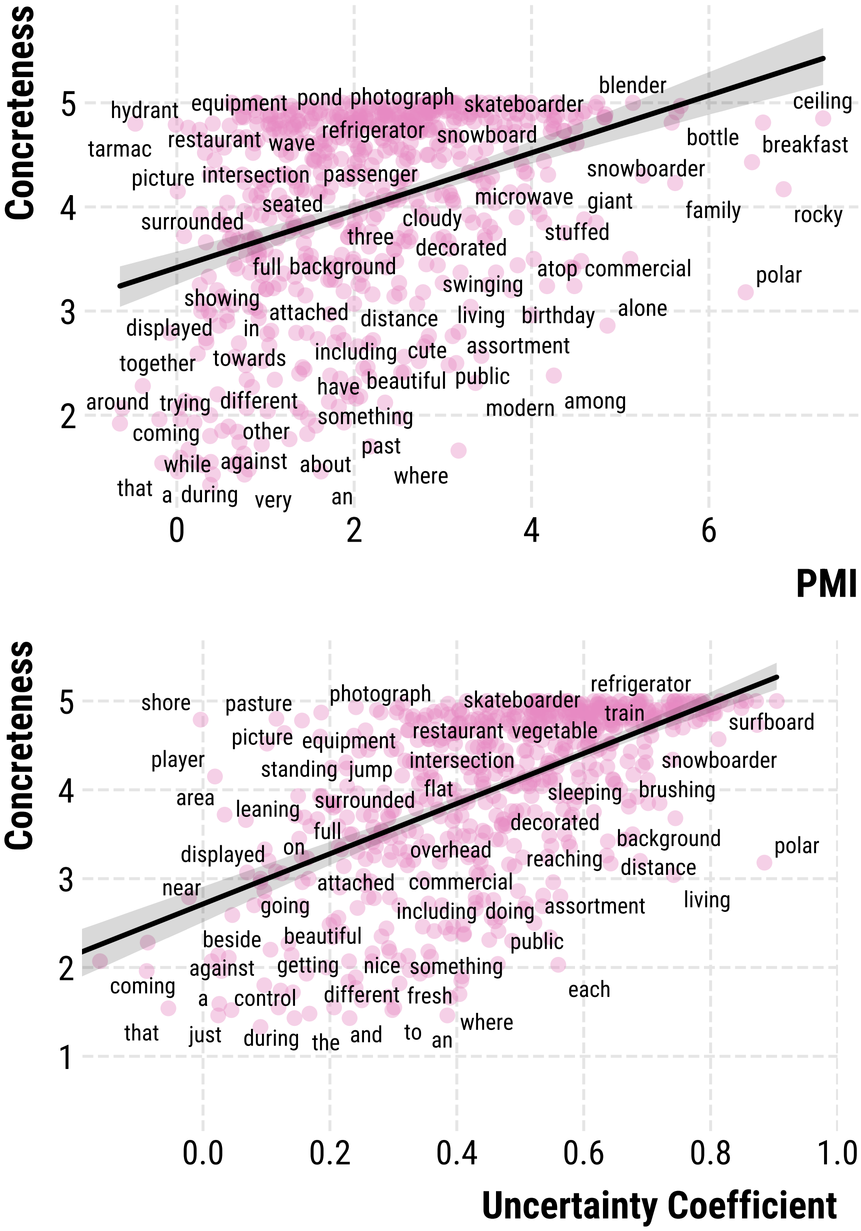

Figure 4.4: Correlation between human concreteness ratings and type-level groundedness (PMI; left, $\rho = 0.368$) or uncertainty coefficent (right, $\rho = 0.609$): i.e., the average ratio between LM surprisal and captioning model surprisal.

4.5.3.0.2 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 4.3). Within a POS, groundedness is expected to vary substantially: for example, some (concrete, visually distinct) nouns have much higher PMI with the image than others, and tokens of the same word type also have different groundedness (e.g. "lot" referring to a location vs. "lot" as a quantity expression) Therefore, we expect word class to explain much less variance than in the overall MI estimates. Language, dataset, and their interaction account for 2.4% of the total variation in PMIs across the three datasets ($F_{64,10^7} = 4727$, $p < 0.001$). Word class accounts for 12.0% of the total variation ($F_{12,10^7} = 123583$, $p < 0.001$). Additionally, the interaction between word class and language (cross-linguistic variation in the means of word classes) accounts for only an additional 1.6% of the total variation ($F_{330,10^7} = 602.5$, $p < 0.001$), despite having many degrees of freedom. So cross-linguistically consistent tendencies comprise the bulk of the explainable variance in the overall PMI distribution across these three datasets—5 times as much as language and dataset, and 7.5 times as much as language differences in POS groundedness.[6]

4.5.4 Semantic dimension of the measure

In this section we explore the semantic properties of the groundedness measure introduced here, comparing it to semantic norms related to contentfulness that are widely used in psycholinguistics. One potential advantage of our method is the ease with which it allows the rating of individual word tokens in context; however, existing ratings tend to be for words in isolation (word types). We focus our analysis here on English and on word types which occur at least 30

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

times in the COCO(-35L)[7] validation set, averaging across occurrences to obtain an estimate of the average type-level groundedness.

We compare to three different psycholinguistic norms: imageability, concreteness, and strength of visual experience. Such norms are measured by providing a definition and examples of low- and high-value words to raters, who then rate words on a Likert Scale. For imagability, we use the Glasgow Psycholinguistic Norms (Scott et al., 2019). For concreteness, we use the Brysbaert et al. (2014) norms. For strength of visual experience, we use the Lancaster Sensorimotor Norms (Lynott et al., 2020). Results for concreteness are shown in Figure 4.4 (left). We observe fairly weak (though significant, $p < 0.001$) correlations with groundedness using Spearman's ρ (Imageability: $\rho = 0.288$, Concreteness: $\rho = 0.368$, Visual strength: $\rho = 0.212$).

We find these weak correlations are partly due to to the *informativity* aspect of our measures, which seems not to play as large of a role in human ratings (e.g. woman is just as concrete as skateboard, but less informative and also less grounded by our measure). To account for differences in baseline (LM) word informativity, we can normalize the PMI scores by the LM surprisal, yielding the uncertainty coefficient (Theil, 1970): the proportion of the LM surprisal explained by the PMI. Regressing this value against the psycholinguistic norms, stronger correlations emerge (Imagability: $\rho = 0.548$, Concreteness: $\rho = 0.609$ as shown in Figure 4.4 (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

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

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 …").

4.6 Discussion and Conclusion

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

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

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

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

Going beyond the details of the approach here, our work generally suggests a role for multimodal models in computational typology similar to the one played by language models in the past decade (e.g. **?**Cotterell et al., 2018; **?**). While language coverage remains more limited than text models, the latest

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

multimodal models and datasets cover enough typologically and culturally diverse languages to make them worth studying—and we anticipate coverage will only improve. Further, the ability of multimodal models to provide an empirically grounded (if imperfect) representation of meaning makes them uniquely valuable for quantitatively addressing questions about the relation between form and function in language. Our work provides the first study of this kind, and we hope that by demonstrating the utility of this approach and releasing our groundedness scores we will inspire other researchers to follow suit.

Limitations

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

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

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

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

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

Chapter 5

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

The detail of the pattern is
movement.

T.S. Eliot, *Four Quartets*

5.1 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 nouns and verbs. But for each of these distinctions, there are languages where it is not clearly relevant

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

In Chapter 4, 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 4, 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 items 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 4. Can a groundedness continuum help explain how and why some languages split a major 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 present evidence from Japanese, where adjectives are split into two formally distinct classes, 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 an investigation inspired by Wetzer (1996) 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 (1996) identified languages with a more stative conception of verbs as those that do not obligatorally 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 (1996) 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.

5.2 Continua among lexical word classes

One of the major findings of Chapter 4, 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 an 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 assymmetry 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 critized 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 (1991a) 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 5.1: Croft (2001)'s analysis of the conceptual categories of the major parts of speech and their semantic properties (transposed).

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 handfull of "adjectives" with morphosyntax distinguished from nouns and verbs. For example, Bemba has less than twenty adjectives accoring 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 (1991b, p. 281) and Uehara (1995, pp. 214–215), persistince 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 (**?**). However, 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

[1] This 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.

a coherent visual object. Indeed, in **?**, Givón ammended 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.

5.3 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:

(5.1)	*yama-ga*	*takai*	/	*takakatta.*
	mountain-NOM	high	/	high.PAST

"The mountain is/was tall." **(*i*-adjective)**

(5.2)	*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 their syntactic position as in (1). Both *i*-adjectives and verbs can modify nouns simply by appearing pre-nominally. However, their inflectional paradigm 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 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.

[3] Japanese publisher Sanseido's Word of the Year in 2015 (**?**).

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 betwen *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 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]

[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.

5.3.1 Method

I use the models and methods introduced in Chapter 4 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 $W = w_1, w_2, w_3 ... w_t ...$, we formalise groundedness as:

$$\text{Groundedness}(w_t) = \log p(w_t \mid I, \mathbf{w}_{<t}) - \log p(w_t \mid \mathbf{w}_{<t}), \tag{5.3}$$

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. 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 4, 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 , 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 4, 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 ideosyncratic effects on groundedness (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

[5] https://pypi.org/project/SudachiPy/

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 (**?**).

5.3.2 Results

			Types		Tokens			
Dataset	MT?	# Captions	*i-*	*na-*	*i-*	*na-*	bits Effect(*na-*)	*p*-value
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 5.2: 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**)

Results are shown in Table 5.2. 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 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 patterns. 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 effectgs 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 discusssed above.

		LM surprisal		Captioning surprisal	
Dataset	MT?	bits Effect(*na*-)	*p*-value	bits Effect(*na*-)	*p*-value
COCO-35L-dev	Yes	1.34±0.71	0.063	**1.15±0.46**	**0.014**
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	**1.26±0.58**	**0.030**	0.35±0.40	0.38

Table 5.3: 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**)

5.3.3 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 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 lemmas 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

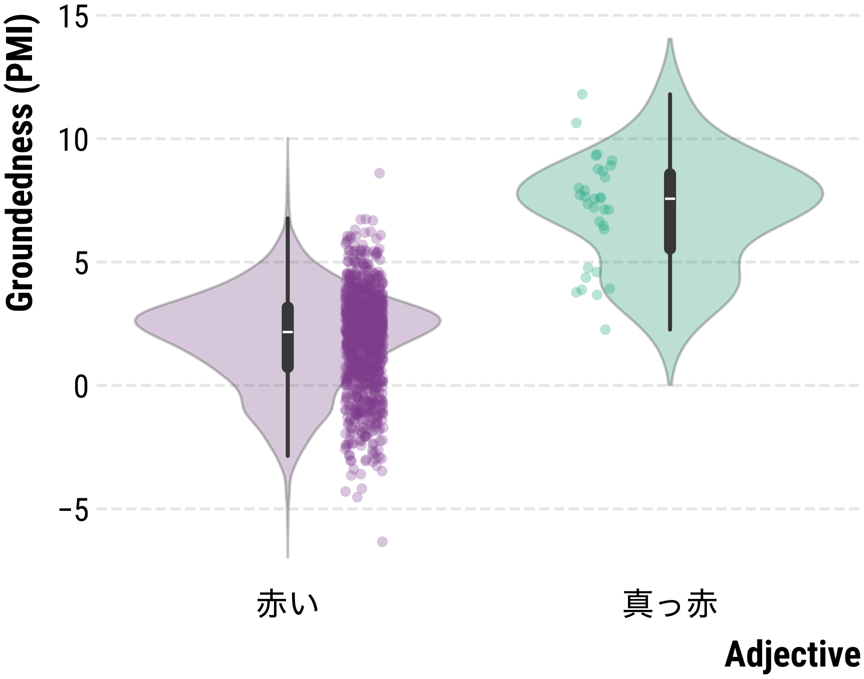

Figure 5.1: Groundedness scores for *na*-adjective *makka* (completely red; right) and *i*-adjective *akai* (red; left) in the STAIR-full-dev dataset.

white) or *koudai* (*na*-adjective, vast) and *hiroi* (i-*adjective, wide)*.

While it is clear that the *na*-adjective/*i*-adjective distinction is not syncronically 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 ideosyncratic splits in word classes across languages to see if similar groundedness effects are observed.

5.4 The Tensedness Correlation

In the previous section, I showed evidence from Japanese that groundedness could provide a novel explanation for seemingly ideosyncratic *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.

5.4.1 The typological finding

Wetzer (1996) 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:

(5.4) *Der* *Mann* *ist* *alt*
The-Masc man is old
"The man is old." **(Property predication)**

(5.5) *Der* *Mann* *ist* *Arzt.*
The-Masc man is doctor
"The man is a doctor." **(Nominal predication)**

(5.6) *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):

(5.7) *Zhāngsān* *shì* *yi-ge* *hùshiì.*
Zhangsan cop one-clf nurse
"Zhangsan is a nurse." **(Nominal predication)**

(5.8) *tā* *pàng*
3sg fat
"She is fat." **(Property predication)**

(5.9) *tā* *yóuyo⌧ng*
3sg swim
"She swims." **(Verbal predication)**

Wetzer (1996), 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.

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 (1996) argue extensively that exceptions to this generalization can largely be understood as artifacts of recent diacronic changes in languages

[6] Given 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 prety 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.

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 ().

5.4.2 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 (1996) 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 (1996) 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 instantion 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.

5.4.3 Methodological background

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

5.4.3.1 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 very 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.

5.4.3.2 Measuring a groundedness shift

Comparing surprisals across languages is fraught with complexity. Previous studies have occaisionally 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.

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

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

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 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

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

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:

$$\text{Quality Ratio} = \frac{\text{NLL}_{\text{Captioning}}}{\text{NLL}_{\text{Language Model}}} \tag{5.10}$$

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]

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

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 4.

Boundary between verbs and adjectives Finally, it is worth noting that the UPOS[10] 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 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

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

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.

5.4.4 Results

Figure 5.2 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

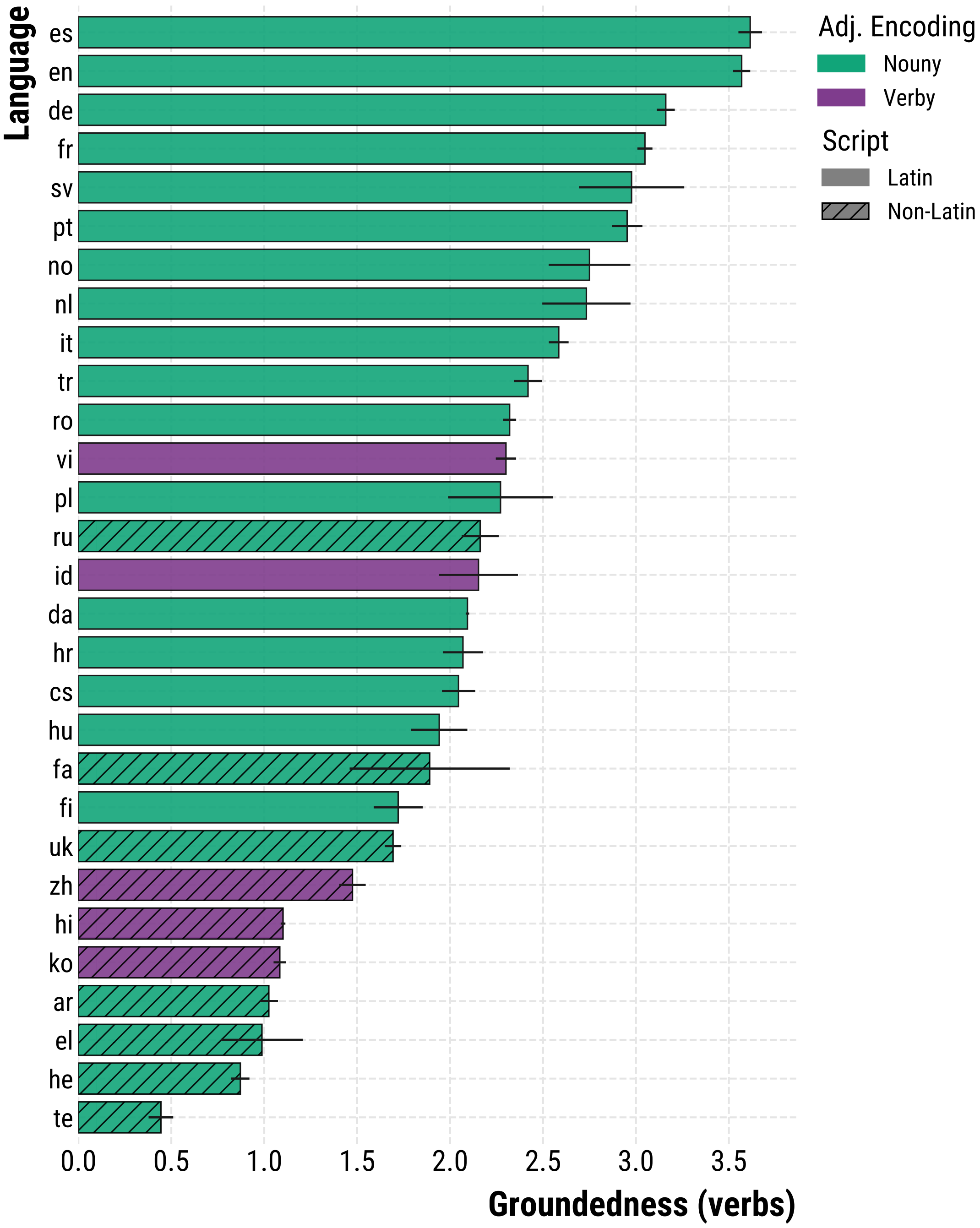

Figure 5.2: Goundedness of the verbal categories across the 30 languages in this study. Error bars represent standard error in the mean groundedness across the datasets considered (COCO-35L, XM3600, and Multi30K where available). Contra theoretical predictions, verby languages do not exhibit higher mean 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.

($\beta = -0.22 \pm 0.11, p = 0.046$). 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 stuggles 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 5.3 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.

5.4.5 Discussion

Across all three experiments, we find no evidence that verby languages have a more grounded verbal prototype than nouny languages, contrary to the theorretical 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 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

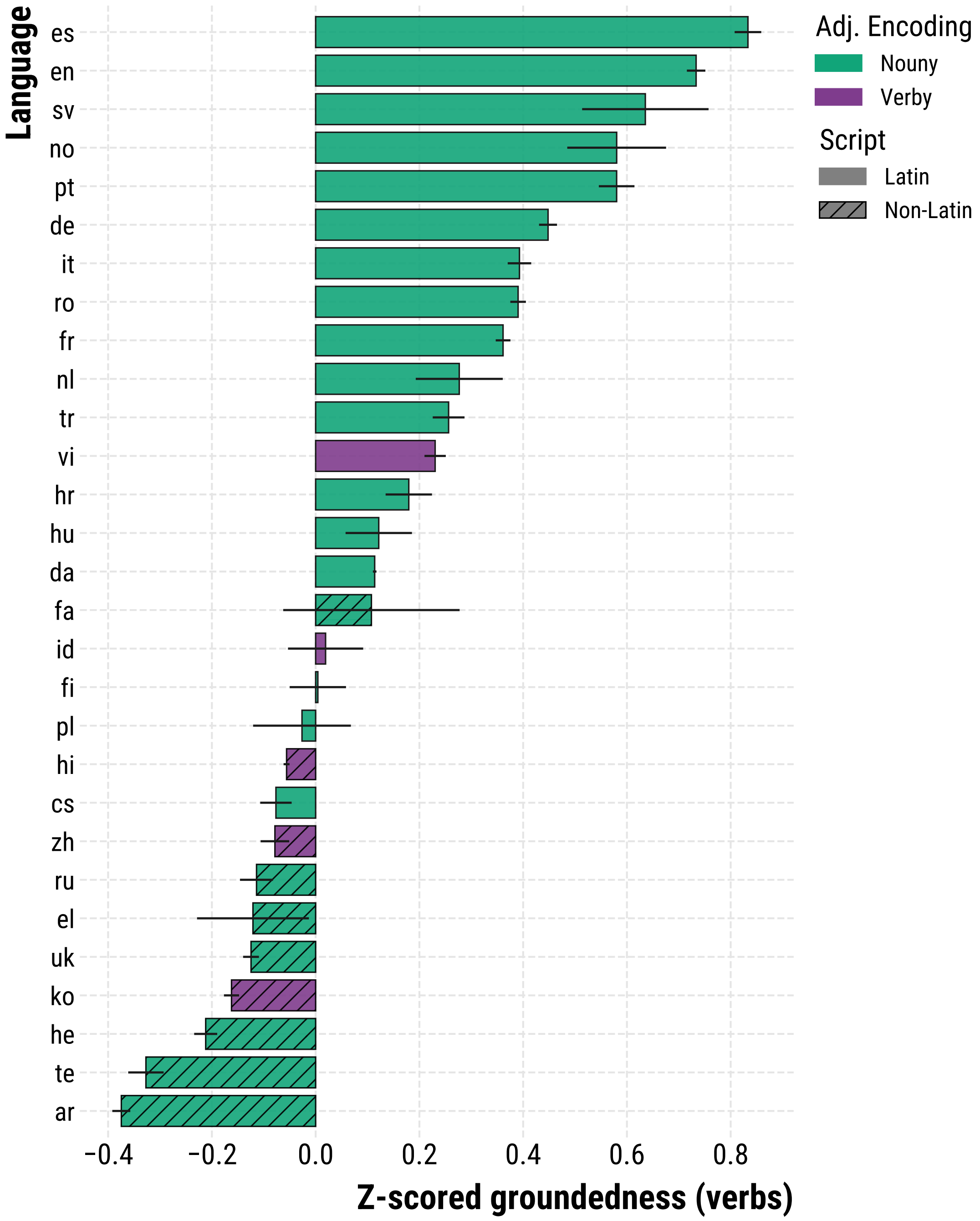

Figure 5.3: Z-scored groundedness of the verbal categories. Error bars represent standard error in the Z-scores across the datasets considered (COCO-35L, XM3600, and Multi30K where available). The results suggest verbs are not *relatively* more grounded than other words in verby languages. However, we observe a clear effect of script, with languages written in Latin script exhibiting relatively more grounded verbs.

,

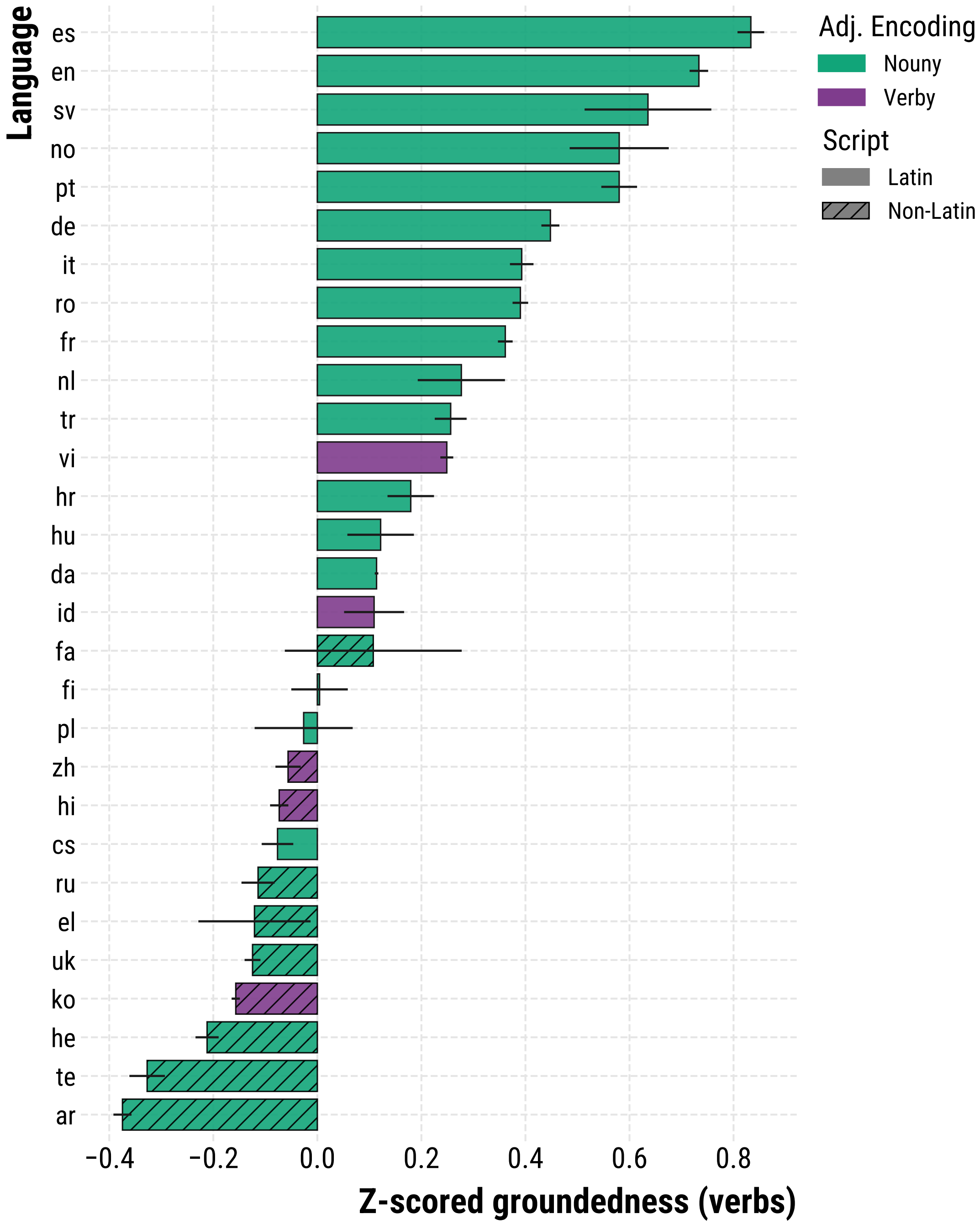

Figure 5.4: Z-scored groundedness of the verbal categories, with adjectives included for verby languages. Error bars represent standard error in the Z-scores across the datasets considered (COCO-35L, XM3600, and Multi30K where available). Despite the higher groundedness of adjectives than verbs in general, and concerns that legitimate members of the verbal category could be disproporionally “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.

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 5.3, we compared groundedness within a single language, while in Chapter 4, 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 multilingual vision–language pretraining may help alleviate some of the model quality issues I have observed here.

5.5 Conclusion

In this chapter, I have argued that groundedness and meaning content are operant in grammatical organization not only across the lexical–functional

[11] In Chapter **??** I lay out a number of such directions.

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 4, 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.

Bibliography

Ackema, P. and Neeleman, A. (2019). *Default person versus default number in agreement*, pages 21–54. Open Generative Syntax. Language Science Press.

Alikhani, M. and Stone, M. (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.

Anderson, S. R. (1982). Where's morphology? *Linguistic Inquiry*, 13:571–612.

Anderson, S. R. (1985). Inflectional morphology. In *Language Typology and Syntactic Description*, volume 3, pages 150–201. Cambridge University Press, 1 edition.

Arppe, A., Harrigan, A., Schmirler, K., Antonsen, L., Trosterud, T., Nørstebø Moshagen, S., Silfverberg, M., Wolvengrey, A., Snoek, C., Lachler, J., Santos, E. A., Okimāsis, J., and Thunder, D. (2014–2019). Finite-state transducer-based computational model of Plains Cree morphology.

Ashby, L. F., Bartley, T. M., Clematide, S., Del Signore, L., Gibson, C., Gorman, K., Lee-Sikka, Y., Makarov, P., Malanoski, A., Miller, S., Ortiz, O., Raff, R., Sengupta, A., Seo, B., Spektor, Y., and Yan, W. (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.

Babazhanova, M., Tezekbayev, M., and Assylbekov, Z. (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.

Batsuren, K., Bella, G., and Giunchiglia, F. (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.

Batsuren, K., Goldman, O., Khalifa, S., Habash, N., Kieraś, W., Bella, G., Leonard, B., Nicolai, G., Gorman, K., Ate, Y. G., Ryskina, M., Mielke, S., Budianskaya, E., El-Khaissi, C., Pimentel, T., Gasser, M., Lane, W. A., Raj, M., Coler, M., Samame, J. R. M., Camaiteri, D. S., Rojas, E. Z., López Francis, D., Oncevay, A., López Bautista, J., Villegas, G. C. S., Hennigen, L. T., Ek, A., Guriel, D., Dirix, P., Bernardy, J.-P., Scherbakov, A., Bayyr-ool, A., Anastasopoulos, A., Zariquiey, R., Sheifer, K., Ganieva, S., Cruz, H., Karahóg☒a, R., Markantonatou, S., Pavlidis, G., Plugaryov, M., Klyachko, E., Salehi, A., Angulo, C., Baxi, J., Krizhanovsky, A., Krizhanovskaya, N., Salesky, E., Vania, C., Ivanova, S., White, J., Maudslay, R. H., Valvoda, J., Zmigrod, R., Czarnowska, P., Nikkarinen, I., Salchak, A., Bhatt, B., Straughn, C., Liu, Z., Washington, J. N., Pinter, Y., Ataman, D., Wolinski, M., Suhardijanto, T., Yablonskaya, A., Stoehr, N., Dolatian, H., Nuriah, Z., Ratan, S., Tyers, F. M., Ponti, E. M., Aiton, G., Arora, A., Hatcher, R. J., Kumar, R., Young, J., Rodionova, D., Yemelina, A., Andrushko, T., Marchenko, I., Mashkovtseva, P., Serova, A., Prud'hommeaux, E., Nepomniashchaya, M., Giunchiglia, F., Chodroff, E., Hulden, M., Silfverberg, M., McCarthy, A. D., Yarowsky, D., Cotterell, R., Tsarfaty, R., and Vylomova, E. (2022). UniMorph 4.0: Universal Morphology. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855, Marseille, France. European Language Resources Association.

Bauer, L. (2004). The function of word-formation and the inflection-derivation distinction. *Words and their Places. A Festschrift for J. Lachlan Mackenzie. Amsterdam: Vrije Universiteit*, pages 283–292.

Beniamine, S., Maiden, M., and Round, E. (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.

Benjamini, Y. and Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics*, 29(4):1165–1188.

Berger, U. and Ponti, E. M. (2024). Cross-lingual and cross-cultural variation in image descriptions.

Bergmanis, T. and Goldwater, S. (2017). From segmentation to analyses: a probabilistic model for unsupervised morphology induction. In *Proceedings of EACL*, Valencia, Spain.

Beyer, L., Steiner, A., Pinto, A. S., Kolesnikov, A., Wang, X., Salz, D., Neumann, M., Alabdulmohsin, I., Tschannen, M., Bugliarello, E., Unterthiner, T., Keysers, D., Koppula, S., Liu, F., Grycner, A., Gritsenko, A., Houlsby, N., Kumar, M., Rong, K., Eisenschlos, J., Kabra, R., Bauer, M., Bošnjak, M., Chen, X., Minderer, M., Voigtlaender, P., Bica, I., Balazevic, I., Puigcerver, J., Papalampidi, P., Henaff, O., Xiong, X., Soricut, R., Harmsen, J., and Zhai, X. (2024). PaliGemma: A versatile 3b VLM for transfer.

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

Bisang, W. (2010). Word Classes. In *The Oxford Handbook of Linguistic Typology*. Oxford University Press.

Bisang, W. (2017). Grammaticalization. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bommasani, R., Davis, K., and Cardie, C. (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.

Bonami, O. and Paperno, D. (2018). Inflection vs. derivation in a distributional vector space. *Lingue e linguaggio*, 17(2):173–196.

Bonami, O. and Strnadová, J. (2019). Paradigm structure and predictability in derivational morphology. *Morphology*, 29(2):167–197.

Booij, G. (1996). Inherent versus contextual inflection and the split morphology hypothesis. In *Yearbook of Morphology 1995*, pages 1–16. Springer.

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

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

Boschloo, R. (1970). Raised conditional level of significance for the 2× 2-table when testing the equality of two probabilities. *Statistica Neerlandica*, 24(1):1–9.

Brysbaert, M., Warriner, A. B., and Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911.

Bybee, J. L. (1985). *Morphology: A Study of the Relation between Meaning and Form*. John Benjamins, Amsterdam.

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

Chomsky, N. (1957). *Syntactic Structures*. De Gruyter Mouton.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (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.

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

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

Corver, N. and Riemsdijk, H. V. (2001). Semi-lexical categories. In Corver, N. and Riemsdijk, H. V., editors, *Semi-Lexical Categories*, pages 1–20. de Gruyter.

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

Cotterell, R., Kirov, C., Hulden, M., and Eisner, J. (2019). On the Complexity and Typology of Inflectional Morphological Systems. *Transactions of the Association for Computational Linguistics*, 7:327–342.

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

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

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

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

Croft, W. (2001). *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford University Press.

Croft, W. (2002). *Typology and Universals*. Cambridge Textbooks in Linguistics. Cambridge University Press, 2nd edition.

Croft, W. (2016). Comparative concepts and language-specific categories: Theory and practice. *Linguistic Typology*, 20(2):377–393.

Cutler, A. (1981). Degrees of transparency in word formation. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, 26(1):73–77.

de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Deutsch, D., Hewitt, J., and Roth, D. (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.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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.

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

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

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

Dubé, C., Monetta, L., Martínez-Cuitiño, M. M., and Wilson, M. A. (2014). Independent effects of imageability and grammatical class in synonym judgement in aphasia. *Psicothema*, 26(4):449–456.

Ferraro, F., Mostafazadeh, N., Huang, T.-H., Vanderwende, L., Devlin, J., Galley, M., and Mitchell, M. (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.

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

Futrell, R., Mahowald, K., and Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

Gella, S., Sennrich, R., Keller, F., and Lapata, M. (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.

Gemma, T. (2024). Gemma: Open models based on Gemini research and technology.

Gerdes, K., Kahane, S., and Chen, X. (2021). Typometrics: From Implicational to Quantitative Universals in Word Order Typology. *Glossa: a journal of general linguistics*, 6(1).

Givón, T. (1979). *On Understanding Grammar*. Perspectives in Neurolinguistics and Psycholinguistics. Academic Press.

Givon, T. (1984). *Syntax: A Functional-Typological Introduction Vol I*. Amsterdam: Benjamins.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (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).

Greenberg, J. H., editor (1966a). *Universals of Language*. Number 37 in The M.I.T. Press Paperback Series. M.I.T Pr, Cambridge, Mass., 2nd edition.

Greenberg, J. H., editor (1966b). *Universals of language*. M.I.T. Press, 2 edition.

Haiman, J. (1980). The Iconicity of Grammar: Isomorphism and Motivation. *Language*, 56(3):515–540.

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

Harris, Z. (1954). Distributional structure. *Word*, 10(23):146–162.

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

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

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

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

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

Hathout, N. and Namer, F. (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).

Hathout, N., Sajous, F., and Calderone, B. (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).

He, J., Neubig, G., and Berg-Kirkpatrick, T. (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.

Hofmann, V., Schütze, H., and Pierrehumbert, J. (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.

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

Hu, Z., Dong, K., Dai, W., and Tong, T. (2017). A comparison of methods for estimating the determinant of high-dimensional covariance matrix. *The International Journal of Biostatistics*, 13(2):20170013.

Kasthuri, M., Kumar, S. B. R., and Khaddaj, S. (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.

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

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

Kirkici, B. and Clahsen, H. (2013). Inflection and derivation in native and non-native language processing: Masked priming experiments on turkish. *Bilingualism: Language and Cognition*, 16(4):776–791.

König, C. (2006). Marked nominative in africa. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, 30(4):655–732.

Kyjánek, L., Žabokrtský, Z., Ševčíková, M., and Vidra, J. (2020). Universal Derivations 1.0, a growing collection of harmonised word-formation resources. *The Prague Bulletin of Mathematical Linguistics*, 2(115):333–348.

Laks, L. and Namer, F. (2022). Hebrewnette–a new derivational resource for non-concatenative morphology: Principles, design and implementation. *The Prague Bulletin of Mathematical Linguistics*, 118:25–53.

Larasati, S. D., Kuboň, V., and Zeman, D. (2011). Indonesian morphology tool (MorphInd): Towards an indonesian corpus. In Mahlow, C. and Piotrowski, M., editors, *Systems and Frameworks for Computational Morphology*, pages 119–129. Springer Berlin Heidelberg, Berlin, Heidelberg.

Laudanna, A., Badecker, W., and Caramazza, A. (1992). Processing inflectional and derivational morphology. *Journal of Memory and Language*, 31(3):333–348.

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10:707.

Levshina, N. (2020). How tight is your language? A semantic typology based on Mutual Information. In Evang, K., Kallmeyer, L., Ehren, R., Petitjean, S., Seyffarth, E., and Seddah, D., editors, *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories*, pages 70–78, Düsseldorf, Germany. Association for Computational Linguistics.

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

Liljencrants, J., Lindblom, B., and Lindblom, B. (1972). Numerical Simulation of Vowel Quality Systems: The Role of Perceptual Contrast. *Language*, 48(4):839–862.

Lin, C.-C., Ammar, W., Dyer, C., and Levin, L. (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.

Lin, K. R., Wisman Weil, L., Thurm, A., Lord, C., and Luyster, R. J. (2022). Word imageability is associated with expressive vocabulary in children with autism spectrum disorder. *Autism & Developmental Language Impairments*, 7.

Liu, F., Bugliarello, E., Ponti, E. M., Reddy, S., Collier, N., and Elliott, D. (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.

Ljubešić, N., Klubička, F., Agić, Ž., and Jazbec, I.-P. (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).

Lynott, D., Connell, L., Brysbaert, M., Brand, J., and Carney, J. (2020). The Lancaster Sensorimotor Norms: Multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, 52(3):1271–1291.

MacKay, D. G. (1978). Derivational rules and the internal lexicon. *Journal of verbal learning and verbal behavior*, 17(1):61–71.

Malouf, R., Ackerman, F., and Semenuks, A. (2020). Lexical databases for computational analyses: A linguistic perspective. In Ettinger, A., Jarosz, G., and Pater, J., editors, *Proceedings of the Society for Computation in Linguistics 2020*, pages 446–456, New York, New York. Association for Computational Linguistics.

McCarthy, A. D., Kirov, C., Grella, M., Nidhi, A., Xia, P., Gorman, K., Vylomova, E., Mielke, S. J., Nicolai, G., Silfverberg, M., Arkhangelskiy, T., Krizhanovsky, N., Krizhanovsky, A., Klyachko, E., Sorokin, A., Mansfield, J., Ernštreits, V., Pinter, Y., Jacobs, C. L., Cotterell, R., Hulden, M., and Yarowsky, D. (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.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). 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, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Mohammadshahi, A., Lebret, R., and Aberer, K. (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.

Morita, C. (2010). The Internal Structures of Adjectives in Japanese. *Linguistic research: working papers in English linguistics*, 26:105–117.

Narasimhan, K., Barzilay, R., and Jaakkola, T. (2015). An unsupervised method for uncovering morphological chains. *Transactions of the Association for Computational Linguistics*, 3:157–167.

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

Newmeyer, F. J. (2007). Linguistic typology requires crosslinguistic formal categories. 11(1):133–157.

Oliver, B., Forbes, C., Yang, C., Samir, F., Coates, E., Nicolai, G., and Silfverberg, M. (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.

Oshima, D. Y., Akita, K., and Sano, S.-i. (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).

Östling, R. (2015). Word Order Typology through Multilingual Word Alignment. In Zong, C. and Strube, M., editors, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 205–211, Beijing, China. Association for Computational Linguistics.

Park, H. H., Zhang, K. J., Haley, C., Steimel, K., Liu, H., and Schwartz, L. (2021). Morphology matters: A multilingual language modeling analysis. *Transactions of the Association for Computational Linguistics*, 9:261–276.

Pawley, A. K. (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*.

Pennington, J., Socher, R., and Manning, C. (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.

Perlmutter, D. (1988). The split morphology hypothesis: Evidence from Yiddish. *Theoretical morphology*, pages 79–100.

Pimentel, T., Valvoda, J., Maudslay, R. H., Zmigrod, R., Williams, A., and Cotterell, R. (2020). Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. Association for Computational Linguistics.

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

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

Rajendran, J., Khapra, M. M., Chandar, S., and Ravindran, B. (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.

Ravfogel, S., Elazar, Y., Goldberger, J., and Goldberg, Y. (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.

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

Rosa, R. and Žabokrtský, Z. (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.

Rosa, R. and Zabokrtský, Z. (2019). Unsupervised lemmatization as embeddings-based word clustering. *CoRR*, abs/1908.08528.

Ross, J. R. (1972). The category squish: Endstation hauptwort. In Peranteau, P. M., Levi, J. N., and Phares, G. C., editors, *Proceedings of the Eighth Regional Meeting of the Chicago Linguistic Society*, pages 316–328, Chicago, Illinois. Chicago Linguistic Society, University of Chicago.

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.

Schakel, A. M. J. and Wilson, B. J. (2015). Measuring word significance using distributed representations of words. *Computing Research Repository*, arXiv:1508.02297.

Schone, P. and Jurafsky, D. (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*.

Schultze-Berndt, E. (2000). *Simple and Complex Verbs in Jaminjung: A Study of Event Categorisation in an Australian Language*. PhD thesis, Radboud University, Nijmegen.

Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., and Sereno, S. C. (2019). The Glasgow Norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods*, 51(3):1258–1270.

Silverstein, M. (1986). *7. Hierarchy of Features and Ergativity*, pages 163–232. De Gruyter Mouton, Berlin, Boston.

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

Soricut, R. and Och, F. (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.

Spencer, A. (2013). *Lexical Relatedness*. Oxford University Press, Oxford.

Stassen, L. (1997). *Intransitive Predication*. Oxford University Press.

Staub, A. (Forthcoming). Predictability in Language Comprehension: Prospects and Problems for Surprisal. *Annual Review of Linguistics*.

Štekauer, P. (2015). 14. the delimitation of derivation and inflection. In Müller, P. O., Ohnheiser, I., Olsen, S., and Rainer, F., editors, *Volume 1 Word-Formation*, pages 218–235. De Gruyter Mouton.

Strunk, L. A. (2020). *A Finite-State Morphological Analyzer for Central Alaskan Yup'Ik*. University of Washington.

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive psychology*, 50(1):86–132.

Sylak-Glassman, J. (2016). The composition and use of the universal morphological feature schema (unimorph schema).

Takaoka, K., Hisamoto, S., Kawahara, N., Sakamoto, M., Uchida, Y., and Matsumoto, Y. (2018). Sudachi: A Japanese Tokenizer for Business. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

ten Hacken, P. (1994). *Defining Morphology: A Principled Approach to Determining the Boundaries of Compounding, Derivation, and Inflection*. Altertumswissenschaftliche Texte Und Studien. G. Olms Verlag.

Thapliyal, A. V., Pont Tuset, J., Chen, X., and Soricut, R. (2022). Crossmodal-3600: A Massively Multilingual Multimodal Evaluation Dataset. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 715–729, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Theil, H. (1970). On the Estimation of Relationships Involving Qualitative Variables. *American Journal of Sociology*, 76(1):103–154.

Thiessen, E. D., Kronstein, A. T., and Hufnagle, D. G. (2013). The extraction and integration framework: a two-process account of statistical learning. *Psychological bulletin*, 139(4):792.

Thiessen, E. D. and Saffran, J. R. (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.

Thompson, S. (1988). A discourse approach to the cross-linguistic category 'Adjective'. In Corrigan, R., Eckman, F., and Noonan, M., editors, *Linguistic Categorization: Proceedings of an International Symposium in Milwaukee, Wisconsin, April 10–11, 1987*, pages 245–265. John Benjamins Publishing Company.

Thompson, S. P. and Newport, E. L. (2007). Statistical learning of syntax: The role of transitional probability. *Language learning and development*, 3(1):1–42.

Uehara, S. (1995). *Syntactic Categories in Japanese: A Typological and Cognitive Introduction*. PhD thesis, University of Michigan, United States – Michigan.

Vilca, H. D. C., Mariñó, F. C. C., and Calderon, E. F. M. (2012). Analizador morfólogico de la lengua Quechua basado en software libre Helsinkifinite-statetransducer (HFST).

Vulić, I., Baker, S., Ponti, E. M., Petti, U., Leviant, I., Wing, K., Majewska, O., Bar, E., Malone, M., Poibeau, T., Reichart, R., and Korhonen, A. (2020). Multi-SimLex: A large-scale evaluation of multilingual and crosslingual lexical semantic similarity. *Computational Linguistics*, 46(4):847–897.

Wartena, C. (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.

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

Wetzer, H. (1996). *The Typology of Adjectival Predication*. De Gruyter Mouton.

Wiemerslage, A., McCarthy, A. D., Erdmann, A., Nicolai, G., Agirrezabal, M., Silfverberg, M., Hulden, M., and Kann, K. (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.

Wiltschko, M. (2014). *The Universal Structure of Categories*. Cambridge Studies in Linguistics. Cambridge University Press, Cambridge.

Ye, A., Santy, S., Hwang, J. D., Zhang, A. X., and Krishna, R. (2024). Computer vision datasets and models exhibit cultural and linguistic diversity in perception.

Yoshikawa, Y., Shigeto, Y., and Takeuchi, A. (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.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. (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.