Chapter 2

Corpus-based Measures Discriminate Inflection and Derivation Cross-Linguistically

2.1 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 (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 distinction is gradient rather than categorical (Bybee, 1985; Spencer, 2013; Copot et al., 2022; Dressler, 1989; Štekauer, 2015; Corbett, 2010; Bauer, 2004) or to take the even stronger position that the distinction carries no theoretical weight at all (Haspelmath, 2024).

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

One major issue in evaluating these theoretical claims is the lack of large-scale, cross-linguistic evidence based on quantitative measures (rather than subjective tests). Work in theoretical linguistics has established that the intuitions underlying subjective tests can be problematic in certain cases (Haspelmath, 2024; Plank, 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 2.7, 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.

2.2 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 2.3.

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. 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;

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

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 2.3.2.

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, 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 2.1. We define a construction here as a unique combination of a morpheme

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

(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 2.8.1.

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 2.4 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.

2.3 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.

2.3.1 Orthography-based measures

In this study, we use orthography as a proxy for phonological form, as discussed in Section 2.2. 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:

$$M_{\text{Form}} = \frac{1}{N} \sum_{i=1}^{N} \textsc{EditDistance}(b_i, c_i). \tag{2.1}$$

To measure the *variability* of the change in form V_{Form} (a measure of the construction's degree of allomorphy), we start by constructing an *edit template* for each word pair, which describes the changes made to the base in a way that abstracts away from specific string positions. For example, the pair (*tanzen*, *getanzt*) yields the edit template ge_XXt, meaning "start by writing ge, copy from the base form, delete the last two characters, and append t." Similarly, the edit template for the pair (Sohn, Söhne) produces the edit template _Xö_e. This example highlights two important design decisions for these edit templates. First, we abstract out any variation in length of the spans which are shared with the input. This is based on the assumption that these reflect variation in the base form itself rather than morphological allomorphy. In our dataset, which does not contain any languages with templatic morphology, this assumption works well; however, future studies wishing to extend to such languages should revisit this assumption. Secondly, because we operate over orthographic form rather than the true form phonetics/featural information, edits which are considered “the same” in linguistic theory may sometimes be considered different and vice-versa. Here, a linguist might describe this plural allomorph as adding +FRONT to the vowel’s features, which would cover the templates _Xö_e, _Xä_e, and _Xü_e. However, addressing this issue is outside the scope of this study.

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

$$V_{\text{Form}} = \sum_{i=1}^{M} \sum_{j=1}^{M} F_{T_i} \cdot F_{T_j} \cdot \textsc{EditDistance}(T_i, T_j). \tag{2.2}$$

For example, suppose we have a morpheme with two edit templates: _as, used 80% of the time, and _os, used 20% of the time. Then this measure would be $0.8 \cdot 0.2 \cdot \textsc{EditDistance}(\_as, \_os) + 0.2 \cdot 0.8 \cdot \textsc{EditDistance}(\_os, \_as) = 0.32$. 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.

2.3.2 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

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

[4] https://commoncrawl.org/

automatically tagged by language to train language-specific embedding models (Grave et al., 2018). These FastText vectors are known to correlate well with human semantic similarity scores (Vulić et al., 2020; Bojanowski et al., 2017), and are more commonly used as models of semantics than syntax.[5] However, there is evidence from the literature in unsupervised part-of-speech tagging (He et al., 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 2.8.1.

In principle, this issue of interpretability could be avoided by using alternative

[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 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/.

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 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:

$$M_{\text{Embed}} = \frac{1}{N} \sum_{i=1}^{N} \left\| E(c_i) - E(b_i) \right\|. \tag{2.3}$$

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 $\mathbf{D} = [\mathbf{d}_1 \dots \mathbf{d}_N]$. We then make the simplifying assumption that the covariance between the dimensions of $\mathbf{D}$ is zero, which allows us to estimate the variance of $\mathbf{D}$ (and thereby V_{Embed}) as the sum of the variances of the individual dimensions k:

$$V_{\text{Embed}} = \sum_{k=1}^{K} \text{Var}(\mathbf{D}_{k,*}), \tag{2.4}$$

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 2.5 and 2.6, 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.

2.4 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 deriva-

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

> tions 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 2.5.1.[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 2.8.1, 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 in other typological variables such as syllable structure, phoneme inventory, and syntactic variables,

[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.

which could affect our measures of formal or distributional change.

2.4.1 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 2.3.2, 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.

2.5 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 2.1. 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 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 2.8.1.

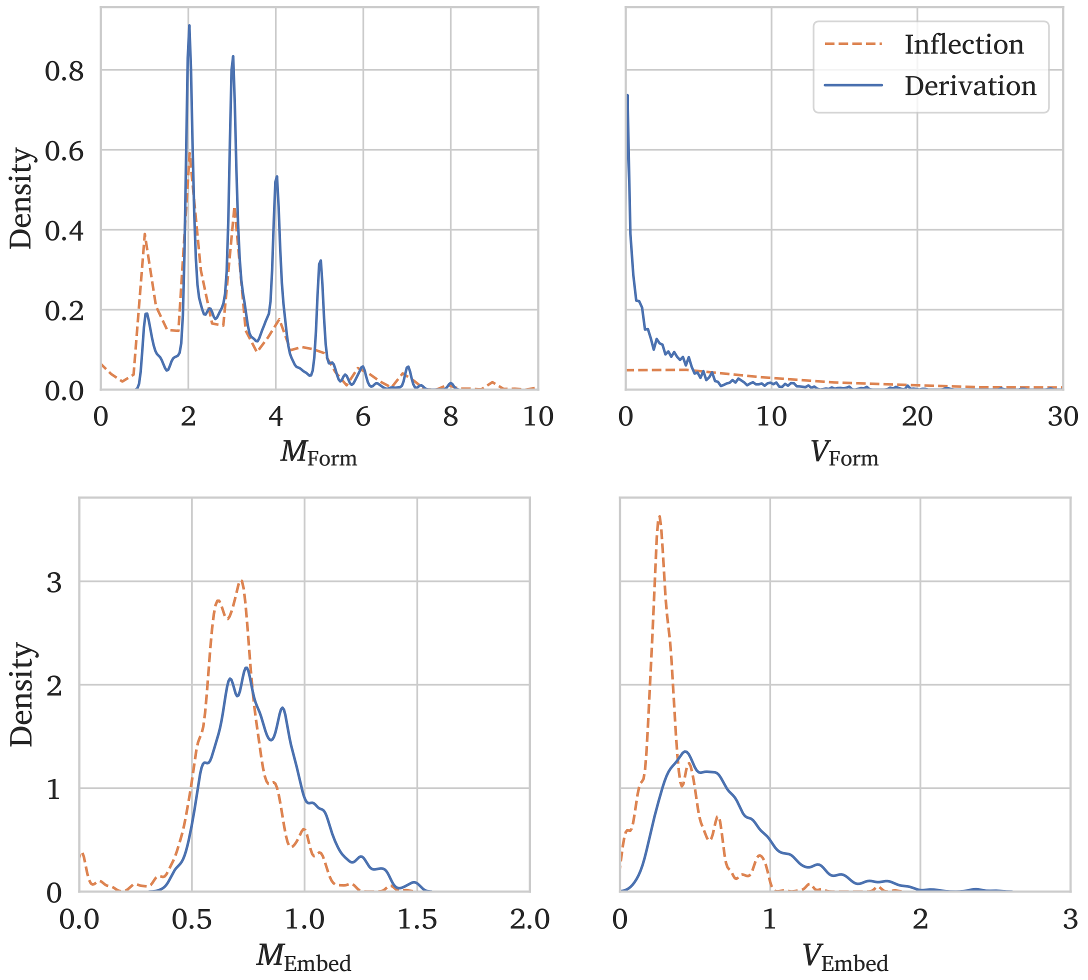

Figure 2.1: The empirical distributions of our four measures (quantifying the magnitude M and variability V of changes in Form and in Embedding space) for inflections and derivations in UniMorph

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.

2.5.1 Effects of Frequency

A potential confounder for our measures on word embeddings is frequency, since the relative frequencies of two words tend to affect their distance in distributional embedding spaces, potentially dominating or complicating meaning-related similarities (Wartena, 2013). In fact, Bonami and Paperno (2018) suggested that differences in frequency may obfuscate measures of semantic distance based on current distributional embedding methods (with low-frequency constructed forms producing larger distances to a given base form than high-frequency constructed forms). If our measures are correlated with frequency, and frequency is also correlated with inflection- or derivation-hood, then any correlation we find between our measures and the inflection–derivation distinction could simply be due to this discrepancy in frequency rather than to the linguistic properties of interest.[9] Accordingly, it is desirable to quantify these relationships with frequency.

Unfortunately, for some languages considered here, word frequency information is not readily available. As a result, we restrict ourselves to the 19 languages in our data which are available through the wordfreq Python package. We estimate the frequency of unattested word forms as 0. We find the mean frequency of

[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.

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.

2.6 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 2.7 and Section 2.8.2. 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 derivation-hood. First, we consider how much any **single** feature recovers the distinction from UniMorph. Secondly, we consider several combinations of two features: (A). **just variability** $\left(V_{\mathbf{Form}}, V_{\mathbf{Embed}}\right)$: 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** $\left(M_{\text{Form}}, M_{\text{Embed}}\right)$: 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** $\left(M_{\mathbf{Form}}, V_{\mathbf{Form}}\right)$ or (D) **syntax/semantics** $\left(M_{\mathbf{Embed}}, V_{\mathbf{Embed}}\right)$ 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 2.2, the expectation from linguistic theory is that greater values of any measure should be associated with that construction being derivational. Our analysis in Section 2.5 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 2.2, the logistic classifier results best support the **just variability** hypothesis (A), 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

[10] This issue of whether the distinction is gradient or categorical with respect to our measures is discussed further in Section 2.8.4.

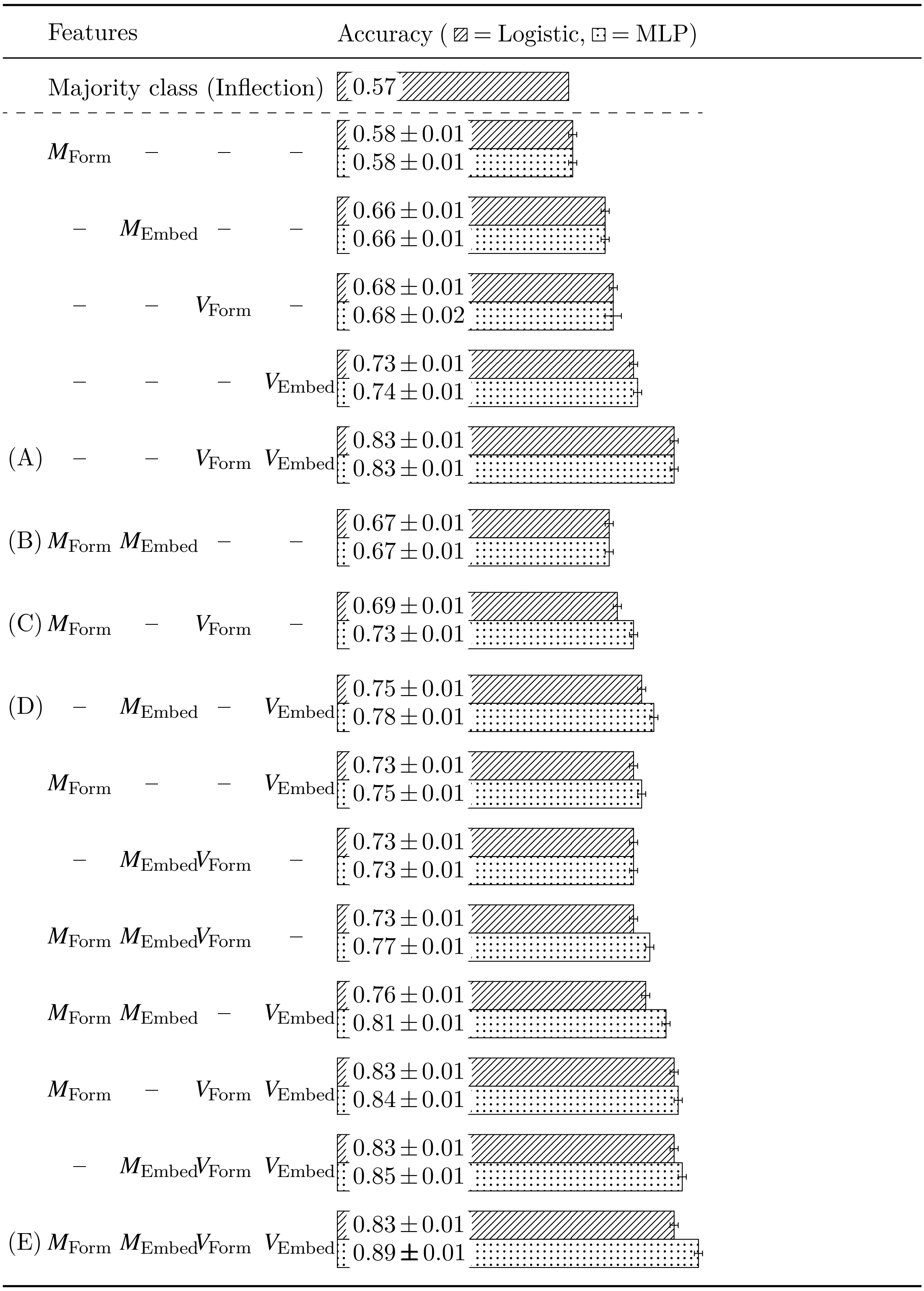

Figure 2.2: Cross-validation accuracy and standard error in reconstructing UniMorph's inflection–derivation distinction by various supervised classifiers. Linguistically-motivated hypotheses referred to in the text are denoted with letters

fit a multi-layer perceptron (MLP) with a hidden layer size of 100, using the Adam optimiser (Kingma and Ba, 2015) and training for 3000 steps. The number of layers and layer size was chosen using validation set performance, while the number of steps was chosen based on loss convergence on the training set. We find similar patterns of performance for most combinations of predictors. However, we see substantial improvements in performance for combinations of features which include both magnitude and variability features; for example, $(M_{\text{Form}}, V_{\text{Form}})$ improving from $69 \pm 1\%$ to $73 \pm 1\%$. Perhaps as a result of this, we achieve a test-set accuracy of $89 \pm 1\%$, when using all four predictors – representing a 6-point improvement over the best linear model, as well as a 4-point improvement over the best combination of three measures using the MLP $(M_{\text{Embed}}, V_{\text{Embed}}, V_{\text{Form}})$. This therefore suggests that while the variability features are the most descriptive of UniMorph's categorisation of inflection/derivation, all four features contain unique information relevant to recreating this distinction (Hypothesis E).

2.7 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. 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.

2.7.1 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, 2002a; 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 2.3 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 2.3 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

[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.

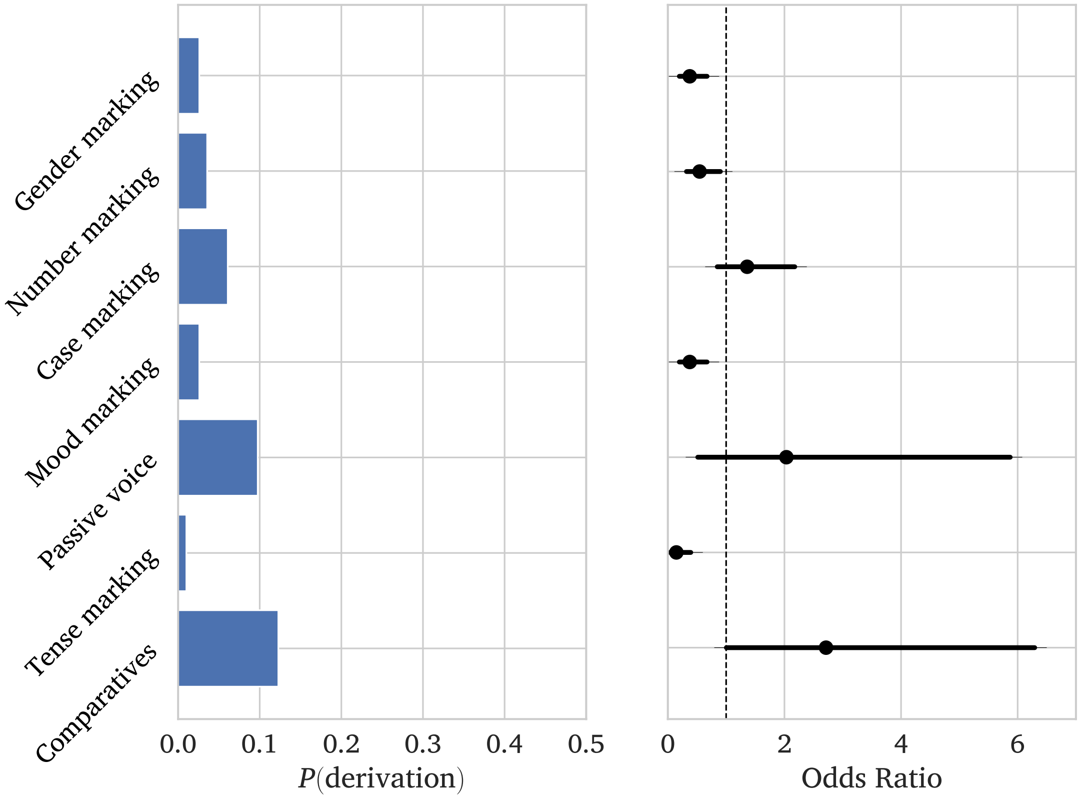

Figure 2.3: Probability and Odds ratio with 95% confidence intervals of being classified as derivation for various kinds of inflectional meaning. Inflections to the right of the dotted line were disproportionately likely to be classified as derivation by our model

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.

2.7.2 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 Figure 2.4.

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

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

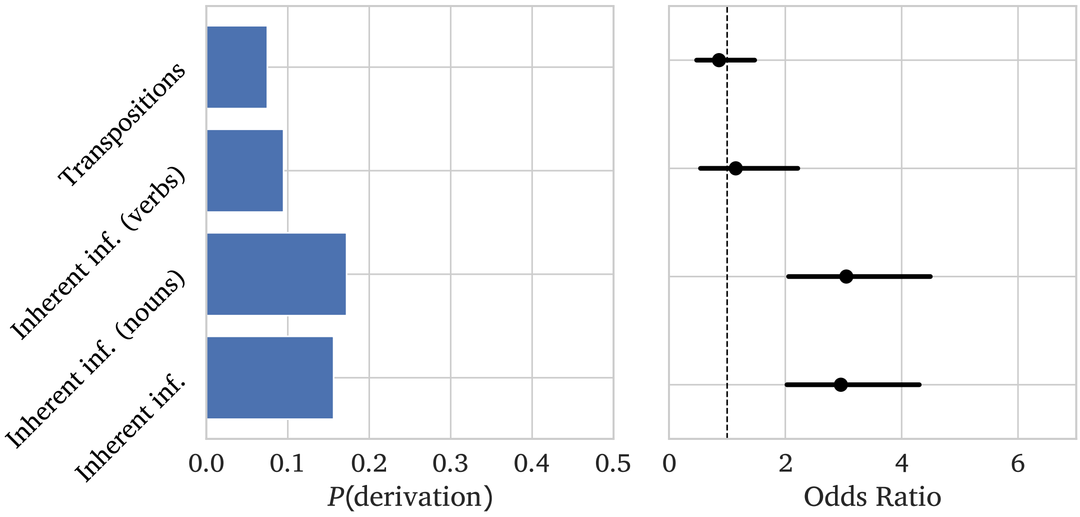

Figure 2.4: Probability and Odds ratio with 95% confidence intervals of being classified as derivation for inherent inflections and transpositions

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 2.4, we find that inherent inflectional constructions are not more likely to be classified as derivation than inflection; however, they *are* significantly more likely to be classified as derivation compared to other types of inflections, as quantified by the odds ratio ($p = 6 \times 10^{-9}$). Interestingly, though, we find this to be almost entirely due to nominal inherent inflection ($p = 2 \times 10^{-8}$), rather than verbal inherent inflection ($p = 0.7$). We see this exemplified in Figure 2.5, 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

[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.

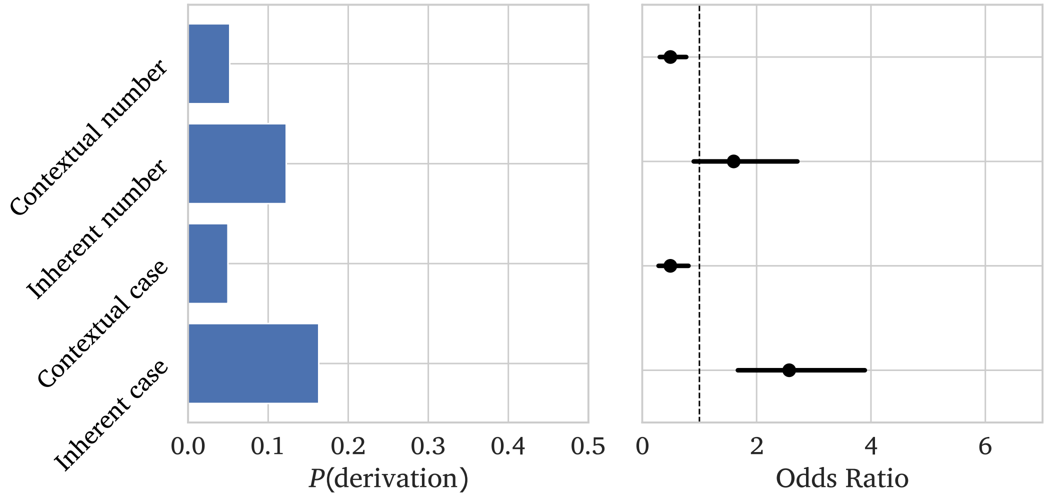

Figure 2.5: Probability and Odds ratio with 95% confidence intervals of being classified as derivation for inherent vs. contextual noun inflections

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 2.4. 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.

2.7.3 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.

2.8 Discussion

2.8.1 The role of our individual measures

As shown in Section 2.6, 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 2.3.1) 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 $\left(M_{\text{Form}}, M_{\text{Embed}}, V_{\text{Embed}}\right)$ 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 2.3.2, 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 2.3.2), 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.

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

2.8.2 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 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.

2.8.3 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.

2.8.4 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, 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.

[15] As described in Section 2.8.4 and shown in Figure 2.6, it is this situation in which we find ourselves.

While, as discussed in 2.8.3, 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 predictive measures, as shown in Figure 2.6. 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.

2.8.5 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

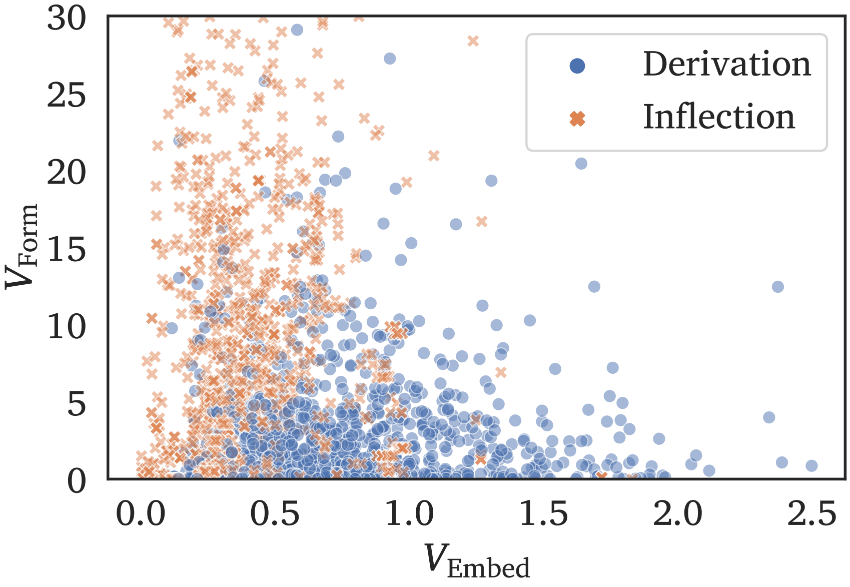

Figure 2.6: Our two most predictive measures for inflection and derivation. Saturation represents overlapping constructions. With respect to these two variables, the inflection–derivation distinction appears gradient rather than categorical

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.

2.8.6 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 2.8.2), 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 2.8.1), 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

[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.

difficulties would need to be overcome in order to use phonological transcriptions. Future work should also investigate to what degree our variability of embedding measure is equivalent to or complementary to Copot et al. (2022)'s predictability of frequency measure, as both are motivated from semantic drift due to a change in lexical index. Similarly, future work could clarify the contribution of distributional semantics by using a model such as Word2Vec or GloVe, or newer models of distributional semantics, such as XLM-R (Conneau et al., 2020) – though in the latter case they would have to overcome the difficulties of multilingual decontextualisation as described in Section 2.3.2. 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.

2.9 Conclusion

In this work, we have presented the first multilingual computational study of the inflection–derivation distinction. In Section 2.3 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 2.5, 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 2.6, 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 2.7) 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 2.8.4). 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.

Chapter 3

A Grounded Typology of Word Classes

In this work, we propose a grounded approach to meaning in language typology. Using images captioned across languages, we can treat the images as an empirical language-agnostic representation of meaning, allowing the quantification of language function and semantics. Using principles from information theory, we define "groundedness", an empirical measure of contextual semantic contentfulness which can be computed using multilingual (vision-and-)language models. As an initial application, we apply this measure to the typology of word classes. We find our measure captures the contentfulness asymmetry between functional (grammatical) and lexical (content) classes across languages, but contradicts the view that functional classes do not convey content. We release a dataset of groundedness scores for 30 languages. Our results suggest that the grounded typology approach can provide quantitative evidence about semantic function in language.

3.1 Introduction

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

While empirically grounding language form (surface structure like vowels) is typically straightforward, language is not just a formal system, but also a functional one. Many questions within typology relate to the relationship between form and *meaning*, especially in domains like morphology and syntax. Typically, typologists manually identify semantic/functional roles such as "subject", and "causative" and study their expression across languages (Haspelmath, 2010; Greenberg, 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).

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

In this work, we specifically focus on semantic contentfulness–how semantically informative a given word token is. We introduce a way to empirically quantify contentfulness, *groundedness*, which relies only on self-supervised vision-and-pos_all.pdf

Figure 3.1: 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.

language models. Groundedness quantifies how much less surprising a word is when we know the perceptual stimuli (i.e. the image) it describes. This *surprisal difference* between the surprisal of the word token in an image captioning model versus its surprisal in a traditional language model is an estimate of the pointwise mutual information: the greater this difference (LM surprisal > captioning surprisal), the more *grounded* the word is in that context.

As a case study, we apply this measure to the study of the typology of word classes (better known within the field of natural language processing as "parts of speech"). Literature from language evolution, cognitive linguistics, pyscho- and neurolinguistics convergently point to contentfulness being an organizing factor in word class processing and even formation and structure: low-content (functional) word classes have many different properties from high-content (lexical) classes(Dubé et al., 2014; Bird et al., 2003; Strik Lievers et al., 2021; Chiarello et al., 1999). Nevertheless, there has been no cross-linguistic quantitative study of the relationship between contentfulness and word class.

Using our groundedness measure to quantify semantic contentfulness, we can estimate the mutual information of a word class with a caption's meaning (image). We find our measure largely rediscovers the distinction between lexical and functional word classes across 30 languages. Further, though it correlates only weakly with norms like imageability and concreteness in English, it provides an intuitive ranking between nouns, verbs, and adjectives (noun > adjectives > verbs) across languages but contradicts the view of adpositions as a "semi-lexical" class. However, our results suggest grammatical word classes still carry semantic content. These results validate intuitions about word class contentfulness and suggest the utility of this measure as a general tool for studying contentfulness in linguistics, and of taking a grounded approach to typological problems. We release the model used to estimate our measure and a dataset of groundedness measures for further study.[1]

3.2 Background

3.2.1 Typology

Within linguistics, *typology* is the subfield focused on the study of patterns which occur across the world's languages–that is, the facts examined within typology are cross-linguistic patterns (Croft, 2002b, pp. 1–2). So while, for example, phonology may study the patterns of sounds within a language, phonological typology is concerned with cross-linguistic trends, generalizations, universals, and restrictions on sound patterns across languages. In order to identify such patterns, linguists must carefully specify systems of interest within languages, and then align them with one another.

Take, for example, the typology of vowel systems. While dozens of acoustically distinct vowels exist, languages vary dramatically in both how many and which vowels they use, with some languages distinguishing as few as two vowels (Colarusso, 1988) and others distinguishing as many as 46 vowels. Yet across the world's languages, vowel systems are not at all uniformly distributed. Most languages have systems with 5-7 vowels, and only an extremely small subset of possible 2, 3, and 4 vowel systems is attested (Gordon, 2016, p. 44). Attested systems strongly tend e.g. to maximally separate the acoustic properties of their constituent vowels (Liljencrants et al., 1972; Schwartz et al., 1997).

Such generalisations about what sorts of vowel systems are attested across the languages of the world require defining cross-linguistically consistent categories. Vowels exist in a continuous acoustic and perceptual space, without clear boundaries between them. To define consistent vowel categories and align systems across language, linguists rely largely on acoustic properties of the speech signal–reducing

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

the problem to a physically grounded, empirical one (Liljencrants et al., 1972; Cotterell and Eisner, 2017).

Empirical grounding with respect to the typology of language *form* (its surface structure) has been successful and is often relatively straightforward to operationalise (e.g. in phonology). However, human 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. The traditional approach has been to manually identify semantic/functional roles such as ”subject”,”passive”, and ”causative”, and study how they are expressed across a range of languages (Haspelmath, 2010; Greenberg, 1966a).

Unlike with definitions based on form, these definitions based on meaning are difficult to empirically ground, and boundary cases are left up to subjective discretion, leading to debates which are strongly influenced by simply the definition of particular terms cross-linguistically (Haspelmath, 2007, 2012; Plank, 1994b). Further, this approach is not suitable for the *discovery* of functional roles, as it requires strict prior definition by a linguist. By introducing grounded typology, we aim to empirically ground functional concepts in typology. Analogously to the objective, measurable acoustic signal in the study of vowel spaces, we treat images as an (albeit imperfect) objective form for language semantics/function, allowing the quantitative approaches applied to formal typology to be extended to functional typology.

3.2.2 Word class

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

The elusiveness of a cross-linguistic definition for word classes leads to many debates about particular languages ”having” or ”not having” a distinction between (e.g.) nouns and verbs on the basis of a mix of formal and semantic criteria (cf. Kaufman, 2009; Hsieh, 2019; Richards, 2009; Weber, 1983; Floyd, 2011). On the other hand, some languages separate a cross-linguistically common word class into multiple clearly distinguished formal/distributional categories. A canonical example here is Japanese adjectives, which are partitioned into two major categories, *na*-adjectives and *i*-adjectives, which differ in their morphology, relationship with the copula, and syntax–with *i*-adjectives behaving more like verbs and *na*-adjectives behaving more like nouns. It is not clear that *na*- and *i*-adjectives together form a natural class in Japanese, yet these sub-categories have no general cross-linguistic parallels (Backhouse, 2004). In this work, we investigate word classes as operationalised in a framework where there is a fixed set of *universally applicable* word classes, as set out in the Universal Dependencies project (de Marneffe et al., 2021) and implemented in the form of the Stanza part-of-speech tagger (Qi et al., 2020). 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.

3.2.3 Contentfulness and word class

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

Complications about these generalised categories and tendencies abound, however. For example, in some languages like Jaminjung, prototypically lexical categories like verbs are closed class (Schultze-Berndt, 2000; Pawley, 2006). Further, both the abstraction and semantic contentfulness 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.

When considering diachronic change, the picture becomes even more complicated, as grammaticalisation acts to change the meaning and syntactic behaviour of contentful words to become more abstract. Grammaticalisation tends to be unidirectional, such that e.g. words for expressing spatial relationships develop from words for body parts (Bisang, 2017). Under such a view, the distinctions between word classes blur and shift over time, with the syntactic distribution in part being a function of a word’s semantic contentfulness.

3.2.4 Measuring contentfulness

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

While theoretical linguistics has focused on a distinction between content words and function words, psycholinguistics has focused on semantic dimensions like imageability, concreteness, and strength of perceptual experience. These have also been found to be highly relevant to processing differences between word classes, such as asymmetries in the processing of nouns and verbs in certain aphasias (Bird et al., 2003; Dubé et al., 2014; Lin et al., 2022). Given that we operationalise meaning as an image, notions such as imageability seem even more clearly related to our groundedness measure. However, as discussed in Section 3.5.4, these concepts are different from our measure in that informativity is not a major factor in their definition.

Recent work at the intersection of natural language processing and computer vision has also shared a goal of quantifying contentfulness. Existing works focus on estimating concreteness and/or imageability norms in a data-driven way (Hessel et al., 2018; Ljubešić et al., 2018; Wu and Smith, 2023; Martínez et al., 2024; Köper and Schulte im Walde, 2016). Unlike the approach here, existing approaches cannot estimate *contextual* scores for individual words, allowing an analysis only at the level of word types, while we are able to analyze at a word-token level in this work. Further, many previous approaches either lack the data or models to be extended to the multilingual context, or rely on supervised training data, inheriting the weaknesses of existing norms.

Another related concept studied in computational psycholinguistics is surprisal. Similar to our groundedness measure, surprisal has an intuitive link to contentfulness from an information theoretic perspective, and has been extensively studied in relation to processing difficulty (Staub, ming). However, surprisal entangles formal and functional information in language. As such, valid cross-linguistic comparisons based on surprisal can be challenging, since form is language specific (Park et al., 2021; Mielke et al., 2019). We here aim to focus on information due to language *function*, separated from form.

3.3 Method

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

$$\text{PMI}(w_t; m \mid \mathbf{w}_{<t}) = \log \frac{p(w_t \mid m, \mathbf{w}_{<t})}{p(w_t \mid \mathbf{w}_{<t})} \tag{3.1}$$

While the true probabilities in Equation 4.3 are not available, using an image as a meaning representation makes both quantities straightforwardly estimable with existing self-supervised neural models: $p_{\boldsymbol{\phi}}(w_t \mid m, \mathbf{w}_{<t})$ corresponds to the probability of the token under an image captioning model, while $p_{\boldsymbol{\theta}}(w_t \mid \mathbf{w}_{<t})$ corresponds to its probability under a vanilla language model.[2] This also allows us to understand groundedness as a *difference in surprisal*, with the value corresponding to how much more expected the token is under the grounded model than under the textual model. As such, in principle the PMI should rarely take on negative values–because the captioning model has strictly more information than the language model. However, some tokens, such as those that are highly grammatical or structural, should be close to 0 (independence).

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

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

$$p(\mathcal{C}_i, m \mid \mathbf{w}_{<t}) = \sum_{w_t \in \mathcal{W}} \left[p(\mathcal{C}_i \mid w_t, \mathbf{w}_{<t}) p(w_t, m \mid \mathbf{w}_{<t}) \right]. \tag{3.2}$$

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

$$I[\mathcal{C}_i; m | \mathbf{w}_{<t}] = \mathbb{E}_{p(\mathcal{C}_i, m, \mathbf{w}_{<t})} \left[\log \frac{p(w_t | \mathbf{w}_{<t}, m)}{p(w_t | \mathbf{w}_{<t}))} \right]. \tag{3.3}$$

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

$$\hat{I}[\mathcal{C}_i; m \mid \mathbf{w}_{<t}] = \sum_{(m, \mathbf{w}_{<t}) \in \mathcal{D}} \frac{\mathbb{1}_{\mathcal{C}_{w_t} = \mathcal{C}_i} \log \frac{p_{\boldsymbol{\phi}}(w_t | \mathbf{w}_{<t}, m)}{p_{\boldsymbol{\theta}}(w_t | \mathbf{w}_{<t})}}{\sum_{w_t \in \mathcal{D}} \mathbb{1}_{\mathcal{C}_{w_t} = \mathcal{C}_i}} \tag{3.4}$$

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

3.4 Experimental setup

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

While this model is a general-purpose multimodal vision and language model, capable of handling a wide range of tasks, it is designed to be fine-tuned on and used for a single task. As such, we use the released paligemma-3b-ft-coco35-224 checkpoint for multilingual captioning, which has been fine-tuned on COCO-35L.

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

To do so, we initialize our language model with the weights from the pretrained PaliGemma model paligemma-3b-pt-224. However, out of the box, the decoder

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

Table 3.1: 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**).

behaves degenerately when no image is provided, so we need to adapt the model to not expect image information and to match the training data of the captioning model. To do so, we fine-tune the language model on the *captions only* from the COCO-35L dataset. In this way, we ensure the models have observed the same data during training and are adapted to the same domain, and are therefore maximally comparable. Table 3.1 summarizes the data matching between the two models.

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

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

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

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

3.4.0.0.5 Estimating word-level probabilities Because the tokenizer of the present model does not cross orthographic word boundaries, we are able to sum the PMI of their constituent subword tokens to obtain a word-level rather

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

than token-level PMI estimate. Ordinarily, some languages do not indicate word boundaries in their orthography, such as Japanese; however, the pretraining data and evaluation datasets (Crossmodal-3600 and COCO-35L) are word-tokenized, so this information is readily available. Finally, we use the correction proposed by Oh and Schuler (2024); Pimentel and Meister (2024) to correctly estimate word-level surprisals for leading-whitespace models, which we report in ??.

$$\begin{aligned} p(w_t \mid \mathbf{w}_{<t}) =& p(\mathbf{s}_{w_t} \mid \mathbf{s}_{\mathbf{w}_{<t}}) \cdot \\ & \cdot \frac{\sum_{s \in \mathcal{S}_{\text{bow}}} p(s \mid \mathbf{s}_{\mathbf{w}_{<t}} \odot \mathbf{s}_{w_t})}{\sum_{s \in \mathcal{S}_{\text{bow}}} p(s \mid \mathbf{s}_{\mathbf{w}_{<t}})} \end{aligned} \tag{3.5}$$

3.5 Results

One of the major generalizations about parts of speech or word classes within linguistics is that they are, broadly, divided into contentful, semantically rich categories; and semantically poor, grammatically-driven categories. Semantically-rich words, we hypothesize, are likely to demonstrate a stronger linkage with the image, while more grammatical words will have a weaker link. Overall, we found that contentful parts of speech had higher mutual information than functional/grammatical parts of speech. Figure 3.1 summarizes the mutual information (MI) estimates for word classes across 30 languages and our three datasets (Multi30K, COCO-35L-Dev, and Crossmodal-3600). Figure 3.2 shows the overall token-level distribution of our groundedness measure across all three datasets (with results for all individual languages and datasets in Appendices ??, ??, and ??). Both figures seem to show a soft yet clear tendency for traditionally lexical parts of speech to have higher mutual information with the image they describe. In the following sections, we aim to quantify these trends, and explore the semantics of our measure.

sum_all.pdf

Figure 3.2: Word token level distributions of the groundedness measure (PMI) across all languages and datasets, grouped by part of speech (word class). We also report the estimated marginal mean and ranking of each word class. Colors are based on the ranking of classes, rather than their average PMIs. Overall, the distribution and estimated ranking of word classes strongly suggest our groundedness measure quantitatively captures the distinction between lexical and functional classes.

3.5.1 Which word classes are grounded?

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

To do so, we use a permutation test. Taking the set of PMIs for a part of speech (POS) in a language, we sample up to 500 PMIs at a time from all datasets and randomly permute their signs, averaging them to produce a new estimate of mutual information (MI). We repeat this process to produce 10^5 permuted estimates. By measuring how often our estimate based on the observed data is greater than the permuted estimate, we obtain the probability that the true MI is greater than 0.[4]

We find that most word classes have a MI significantly greater than 0 in all languages.[5] Further, all word classes have a mutual information estimate significantly greater than 0 in most languages. Word classes for which we cannot reject the null hypothesis in some languages are largely highly grammaticalized/functional: particles (5 langs.), subordinating conjunctions (4 langs.), coordinating conjunctions (4 langs.), auxiliary verbs (4 langs.). A few traditionally lexical classes also fail to reject the null hypothesis in some languages: adverbs (he and te), numerals (en), and proper nouns (ar). Overall, these results suggest most or all word classes contribute some information about the image they describe–in line with theories in linguistics that emphasize the lexical aspects of categories which are traditionally considered functional Corver and Riemsdijk (2001); Bisang (2017).

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

[5] Detailed results in Appendix **??**.

3.5.2 What word classes are more grounded?

We hypothesize that the cross-linguistically consistent trends in word class groundedness correspond to a cline which is a continuous analogue of the lexical–functional word class distinction. To isolate the contribution of word class identity to mutual information cross-linguistically, we compute estimated marginal means (EMMs) for each word class's groundedness,[6] and perform a post-hoc pairwise comparison test of the means.[7] The results of this analysis are displayed in Figure 3.2. We find that lexical word classes (Proper nouns, nouns, adjectives, verbs, numbers, and adverbs) have higher groundedness than functional word classes (particles, auxiliaries, conjunctions, determiners, and adpositions). Pronouns occupy an intermediate position, having an EMM which is not significantly different from particles. The ranking corroborates ideas from cognitive linguistics which place nouns, adjectives, and verbs in a continuous grammaticalization cline, with nouns > adjectives > verbs.[8]On the other hand, it does not neatly align with ideas in linguistic theory about adpositions as a semi-lexical class Corver and Riemsdijk (2001).

3.5.3 How consistent is word class groundedness across languages?

We quantify the strength of the association between groundedness and word class on two levels: language-level MI estimates (analyzing the values summarized in Figure 3.1), and token-level PMI (summarized in Figure 3.2). In both cases, we use ANOVA to estimate the amount of the variance in groundedness explained by word class.

[6] Averaged over values of language and dataset.

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

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

3.5.3.0.1 MI estimates For the language-level MI estimates in Figure 3.1, we consider the separate effects language, dataset, and POS have on groundedness. Because the meanings (images) are matched across languages, this allows us to estimate and control for variation due to some languages having consistently larger or smaller MI estimates (due to language-specific variation in our multilingual neural estimators), and how much variation due to certain datasets generally varying in the groundedness of the language they contain. We find significant effects of all 3 factors, but they differ dramatically in how much variation they explain. The effect of dataset is extremely small, explaining 0.5% of the observed variance ($\eta^2 = 0.005$, $F_{3,816} = 5.71$, $p < 0.01$). Language identity has a larger effect, explaining 8.2% of the variance ($\eta^2 = 0.082$, $F_{29,789} = 6.42$, $p < 0.001$). However, word class dominates, explaining most of the total variance (57.3%, $\eta^2 = 0.573$, $F_{12,806} = 775$, $p < 0.001$), and 62.8% of the remaining variance (partial $\eta^2 = 0.628$) after controlling for variance due to dataset and language. Altogether, these factors explain 65.6% of the variance, leaving the remaining variance to cross-linguistic differences in the mutual information of specific parts of speech.

3.5.3.0.2 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 3.2). Within a POS, contentfulness is expected to vary significantly, so we expect word class to explain much less variance than in the overall MI estimates. Language, dataset, and their interaction account for 2.4% of the total variation in PMIs across the three datasets ($\eta^2 = 0.024$, $F_{64,10^7} = 4727$, $p < 0.001$). Word class accounts for 12.0% of the total variation ($\eta^2 = 0.120$, $F_{12,10^7} = 123583$, $p < 0.001$). Additionally, the interaction between word class and language (cross-linguistic variation in the means of word classes) accounts for only an additional 1.6% of the total variation ($\eta^2 = 0.016\%$, $F_{330,10^7} = 602.5$, $p < 0.001$), despite having many degrees of freedom. So cross-linguistic consistent tendencies comprise the bulk of the explainable variance in the overall PMI distribution across these three datasets–5 times as much as language and dataset, and 7.5 times as much as language differences in POS groundedness.[9]

3.5.4 Semantic dimension of the measure

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

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

We find this is in part related to the *informativity* aspect of our measures, which seems not to play as large of a role in human ratings (e.g. woman is just as

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

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

concrete as skateboard, but less informative and also less grounded according to our measure). To account for differences in baseline (LM) word informativity, we can normalize the PMI scores by dividing by the LM surprisal, yielding the uncertainty coefficient (Theil, 1970), which measures the proportion of the LM surprisal explained by the PMI. Regressing this value against the psycholinguistic norms, stronger correlations emerge (Imagability: $\rho = 0.548$, Concreteness: $\rho = 0.609$, Visual strength: $\rho = 0.320$). This suggests that the differences between groundedness and surprisal are associated with concreteness. However, this measure necessarily collapses differences between word classes in overall informativity/surprisal. In some cases, outliers are due to contextual effects. For example, in our data the word "polar" (high groundedness, moderate concreteness) occurs exclusively as the first word in the multiword expression "polar bear" which is highly concrete, imageable, and visual; while ratings based on the word type are presumably based on the more abstract geographical concept. Other words with 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 used more abstractly in the image captioning context (e.g. "photo" exhibits very low PMIs, because captions frequently begin with "A photo of...").

3.6 Discussion and Conclusion

In this study, we propose a grounded approach to typology, using images as a proxy for sentence meaning. Using information theory and neural models, we define a groundedness measure of a token's association with its meaning. Together, our results demonstrate that parts of speech vary systematically in terms of their groundedness across a typologically diverse sample of languages. We find

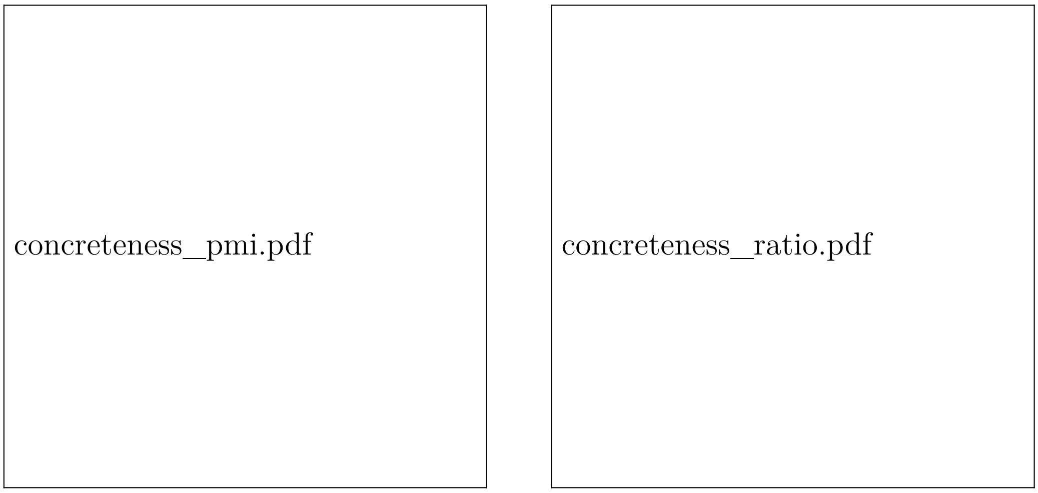

Figure 3.3: Correlation between human concreteness ratings and type-level groundedness (PMI; left, $\rho = 0.368$) or uncertainty coefficent (right, $\rho = 0.609$): i.e., the average ratio between LM surprisal and captioning model surprisal.

this variation can be described as a continuous cline generalizing the traditional 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 continuum view of these classes; 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 a single case study, future work can and should use the measures introduced here to investigate various typological questions about the way languages organise information. For instance, we anecdotally noticed that languages with grammatical gender on determiners tend to have higher than average groundedness for that class. Future work should explore this, investigating in detail what kinds of "functional" classes have higher groundedness and under what conditions. Additionally, our approach could be used to study non-prototypical word class organizations, such as languages which don't clearly distinguish between adjectives and verbs (Korean), or languages that split single word classes into distinct sub-classes (Japanese adjectives). Our approach can also cover any linguistic classes which can be defined over tokens, such as morphemes or semantic classes. For instance, future work could explore the claim that inflections are more “grammatical” than derivations (Booij, 2007b; Haley et al., 2023). To support such work, we make the groundedness scores from the present study available online.[11]

Going beyond the details of the approach here, our work generally suggests a role for multimodal models like image captioning models in computational typology similar to the one played by language models in the past decade. While the range of languages covered and data availability is less than for pure text models, the latest multimodal models and datasets cover enough languages from a typologically and culturally diverse set of languages to make them worth studying—and we anticipate this will only improve from here. Further, the ability of multimodal models to provide an empirically grounded (if imperfect) representation of meaning makes them uniquely valuable for quantitatively addressing questions about the relation between form and function in language. Our work provides the first study of this kind, and we hope that by demonstrating the utility of this approach and releasing our groundedness scores we will inspire other researchers to follow suit.

3.7 Limitations

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

First, our operationalisation of meaning as an image is necessarily a simplification and has numerous implications for our results. Notably, the choice of

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

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

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

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

Chapter 4

Visual groundedness as an organizing principle for word class: Evidence from Japanese

4.1 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 morphosyntactic 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 ??, I 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 ??, 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 ??. 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 **?**'s *Tense Hypothesis*, 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. **?** 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 **?** 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 Tense Hypothesis. This study highlights potential difficulties in comparing groundedness values between languages.

4.2 Continua among lexical word classes

One of the major findings of Chapter **??**, 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 **?**, 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. **?**'s approach to treating the major parts of speech as a continuum through a "category squish" was criticized on a number of fronts (**?**). 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. Formally, 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 **?**, citing a cline between nouns, adjectives, and verbs in terms of their prototypical temporal stability, with verbs being the least prototypically stable. **?** 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). **?** 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.

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

Table 4.1: Croft (2001)'s analysis of the conceptual categories of the major parts of speech and their semantic properties.

While these accounts differ in the specific way they break down the parts of speech into a continuum, they are unified in the idea that adjectives represent a position which is in some important way(s) intermediate to nouns and verbs. This idea is supported not just by monolingual evidence, like **?**'s English data, but also by a plethora of typological data. **?** presents a seminal survey,

Theoretical foundation–semantic maps. Japanese adjectives.

Each of the previous accounts faces significant challenges when accounting for the cross-linguistic data, also the challenge to the prototype or gradient theory of categories generally.

Triangle or line or groundedness line.

4.3 Japanese adjectives

The major word classes of noun, verbs, and adjectives have often been argued to be cross-linguistically universal (). While clearly near-universals, there is also considerable variation in how these classes are organized within specific languages. Nevertheless, upholding many linguistic theories rely on the universality of these classes. Typologists often formulate implicational universals

The two[1] word classes in Japanese typically described as adjectives are *i-*

[1] Some linguists identify a third major class, which is identically syntactically distributed to

adjectives and *na*-adjectives. These classes are clearly distinguished from each other in Japanese in terms of their syntax and morphology:

(4.1) *yama-ga* *takai* / *takakatta.*
mountain-NOM high / high.PAST
"The mountain is/was tall." **(*i*-adjective)**

(4.2) *Taroo-ga* *sizuka* *da* / *sizuka* *datta*
Taro-NOM quiet COP / quiet COP.PAST
"Taro is/was quiet." **(*na*-adjective)**

While clearly distinct from nouns and verbs, *i*-adjectives have an analogous inflectional paradigm to verbs (inflecting for aspect and polarity) and can take their syntactic position as in (1), but as shown in (2), *na*-adjectives must be combined with the copula like nouns. Both *i*-adjectives and verbs can modify nouns simply by appearing pre-nominally, but nouns and *na*-adjectives require a (distinct) attributive marker to modify nouns.

This split is not attributable to phonology or semantics, nor is it a conjugation class. Some stems can belong to both classes. Attempts to describe it under existing semantic hierarchies (**??**) have proven largely unsuccessful.

4.3.1 Method

I use the models and methods introduced in Chapter **??** 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. In particular, for an image I and word w_t in an utterance $W = w_1, w_2, w_3 ... w_t ...$, we formalise groundedness as:

$$\begin{aligned}\text{Groundedness}(w_t) = \log p(w_t \mid I, \mathbf{w}_{<t}) \\ - \log p(w_t \mid \mathbf{w}_{<t})\end{aligned} \tag{4.3}$$

nouns, which we do not concern ourselves with here.

This allows us to compute groundedness as a *difference in surprisal* between an image captioning model and a (domain-matched) language model. In contrast to typical psycholinguistic norms like concreteness and imageability, groundedness is computed at the (word) *token* level. This implies the same word may be more or less grounded in different contexts.

The simplifying assumption of treating an image as meaning makes estimating (visual) groundedness with existing datasets and neural models tractable, and has interesting connections to relevant notions like imageability and concreteness.

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. 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 **??**, while for STAIR I compute the scores using the same methods and models to ensure comparability between the datasets: I use PaliGemma as an image-captioning model, and the fine-tuned PaliGemma language model with matched training data to the captioning model to achieve comparable surprisals between the two models for the PMI estimates, as argued in Chapter **??**.

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 (**?**), as implemented in the sudachipy[2] Python package, to tag identified adjectives

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

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 it's 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 **??**, single groundedness estimates can be noisy, so I 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 corpus.

4.3.2 Results

Our core results are presented in Table **??** Across our corpus of 7185 captions, we find 399 *na*-adjective tokens and 3058 *i*-adjective tokens. These tokens belong to 42 *i*-adjective types and 26 *na*-adjective types. On average, the *na*-adjectives display higher groundedness than *i*-adjectives (3.41 vs. 1.98). Our data has a nested structure, with many tokens of a single word type, and this word type influences groundedness independently of word class (*i*-adjective vs. *na*-adjective). To better estimate the effect of word class itself, we use a linear mixed effects model, with fixed effects of position and word class and a random effect for word type. Under this model, we find a significant effect of word class ($p = 0.029$). Specifically, we find that *na*-adjective-hood increases groundedness by 0.89 ± 0.40 bits.

Two terms are used to compute our visual groundedness measure: surprisal under a language model and surprisal under an image captioning model. Is the association between groundedness and the word class distinction above primarily due to one of these terms? Of particular concern is the first term: perhaps *na*-adjectives are just *a priori* more surprising in the linguistic signal (e.g. expressing lower-frequency concepts). If we find a strong correlation between word class and LM surprisal, it may be that the information provided by the image is dominated by these effects. Fitting the same fixed and random effects as before to instead predict LM surprisal, we do not find a significant effect ($p = 0.133, \beta = 1.17 \pm 0.77$). Similarly, we do not find a significant effect of word class on the captioning surprisal alone ($p = 0.591, \beta = 0.38 \pm 0.61$). So it is only through the interaction between these two factors (groundedness) that an association with word class emerges.

4.3.3 Conclusion

Together, our results suggest that *na*-adjectives are used to express more visually grounded meanings than *i*-adjectives in Japanese. In contrast to prior work which failed to find a semantic organizing principle for this distinction (**??**), our work suggests that the formal similarities *i*-adjectives and *na*-adjectives display to verbs and nouns respectively are not arbitrary, but reflect their semantic character.

While still exploratory, our results suggest an exciting role for groundedness in computational linguistics. Together with **?**, these results point to the utility of groundedness not just for explaining cross-linguistic *consistency* in word class organization, but also *variation*. Beyond this, groundedness can also be a useful tool for framing and answering questions about the relationship between form and meaning in a particular language, not just cross-linguistically. While groundedness is only somewhat correlated with norms like concreteness or imageability, concreteness allows the asking of related questions where such norms are not available–no relevant concreteness or imageability norms exist for Japanese adjectives. Future work should further validate these results on a larger array of words and datasets, and with new and improved models, and also explore such traditional, human-annotated norms.

4.4 The Tense Hypothesis

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?