Corpus-based measures discriminate inflection and derivation cross-linguistically

Coleman Haley, Edoardo M. Ponti, and Sharon Goldwater
University of Edinburgh

ABSTRACT

In morphology, a distinction is commonly drawn between inflection and derivation. However, a precise definition of this distinction which reflects the way it manifests across languages remains elusive within linguistic theory, typically being based on subjective tests. In this study, we present 4 quantitative measures which use the statistics of a raw text corpus in a language to estimate to what extent a given morphological construction changes the form and distribution of lexemes. In particular, we measure both the average and the variance of this change across lexemes. Crucially, distributional information captures syntactic and semantic properties and can be operationalised by word embeddings.

Based on a sample of 26 languages, we find that we can reconstruct $89 \pm 1\%$ of the classification of constructions into inflection and derivation in UniMorph using our 4 measures, providing large-scale cross-linguistic evidence that the concepts of inflection and derivation are associated with measurable signatures in terms of form and distribution that behave consistently across a variety of languages.

We also use our measures to identify in a quantitative way whether categories of inflection which have been considered non-canonical in the linguistic literature, such as inherent inflection or transpositions, appear so in terms of properties of their form and

Keywords: inflection, derivation, morphology, distributional semantics, typology

distribution. We find that while combining multiple measures reduces the amount of overlap between inflectional and derivational constructions, there are still many constructions near the model's decision boundary between the two categories. This indicates a gradient, rather than categorical, distinction.

1 INTRODUCTION

In the field of morphology, a distinction is commonly drawn between inflection and derivation. This distinction is intended to capture the notion that sometimes morphological processes form a "new" word (derivation), whereas other morphological processes merely create a "form" thereof (inflection) (Booij 2007). While the theoretical underpinnings and nature of this distinction are a subject of significant and ongoing debate, it is nevertheless employed throughout theoretical linguistics (Perlmutter 1988; Anderson 1982), computational and corpus linguistics (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

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

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 (1994) covers no fewer than 28 tests for inflectional and derivational status. Upon applying them to just six English morphological constructions, Plank (1994) finds considerable contradictions between the results based on different criteria. Such difficulties in producing a cross-linguistically consistent definition have led many researchers to conclude that the inflection–derivation distinction is gradient rather than categorical (Bybee 1985; Spencer 2013; Copot *et al.* 2022; Dressler 1989; Štekauer 2015; Corbett 2010; Bauer 2004) or to take the even stronger position that the distinction carries no theoretical weight at all (Haspelmath 2024).

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

Several previous studies have shared our goal of 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) ex-

plored 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 1994).

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 fami-

lies, 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 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 MOTIVATION FOR OUR MEASURES

In order to explore our question of interest, we need to operationalise some of the linguistic properties that have been argued to differentiate inflection from derivation. This section briefly reviews some of those properties and explains, at a high level, how they relate to corpus-based measures. We defer the detailed definitions of these measures to Section 3.

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., *kkahki*, 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; Ravfogel *et al.* 2020), due to our complex and highly multilingual setting, we instead choose to *collapse* the distinc-

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

tion of syntactic and semantic change made by Spencer, focusing on what is captured by embeddings designed primarily for capturing semantics but which also capture syntactic information. In particular, we use FastText embeddings, described in more detail in Section 3.2.

In addition to considering the size of the changes made to these aspects of words by a construction, we also consider the *variability* of these changes. Words with different lexical indices are thought to have processes like semantic drift apply separately from each other (Spencer 2013; Copot *et al.* 2022; Bonami and Paperno 2018), which Copot *et al.* (2022) carefully links to variability in semantics. We also consider variability in the changes made to the form. This aspect has been under-explored in prior computational work. Following Plank's (1994) claim that formal 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 1. We define a construction here as a unique combination of a morpheme (given in a canonical form like *–ion* for derivation or as morpho-syntactic features for inflection), initial part-of-speech, constructed part-of-speech, and language. That is, we do not group morphemes across languages, nor do we group derivations with identical canonical forms which apply to or produce different parts of speech. This decision is motivated by examples like agentive *–er* vs. comparative *–er* in English, which differ

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

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 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 4 for details). While individual word pairs within a construction may have quite variable distributional properties, the *general tendencies* of that construction may paint a picture that is more clearly in line with notions of inflection and derivation.

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

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

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

3 METHOD

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

3.1 *Orthography-based measures*

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

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

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

$$V_{\text{Form}} = \sum_{i=1}^{M} \sum_{j=1}^{M} F_{T_i} \cdot F_{T_j} \cdot \textsc{EditDistance}(T_i, T_j). \tag{2}$$

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 $0.8 \cdot 0.2 \cdot \textsc{EditDistance}(\texttt{\_as}, \texttt{\_os}) + 0.2 \cdot 0.8 \cdot \textsc{EditDistance}(\texttt{\_os}, \texttt{\_as}) = 0.32$. 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.

Distributional-embedding-based measures 3.2

To approximate the semantic and syntactic properties of the words in our study, we use type-based (non-contextual) distributional word embeddings. Specifically, we use the FastText vectors for each language released by Bojanowski *et al.* (2017);[3] these were trained on Common Crawl[4] and Wikipedia data, which was automatically tagged by language to train language-specific embedding models (Grave *et al.*

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

[4] https://commoncrawl.org/

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 8.1.

[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/`.

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

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

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

$$M_{\text{Embed}} = \frac{1}{N}\sum_{i=1}^{N}\left\|E(c_i) - E(b_i)\right\|. \tag{3}$$

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:

$$V_{\text{Embed}} = \sum_{k=1}^{K} \text{Var}(\mathbf{D}_{k,*}), \tag{4}$$

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 5 and 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.

4 DATA

To perform our analysis, we require a multilingual resource that labels pairs of words with the inflectional or derivational construc-

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

Due to the semi-automatic annotation in UniMorph 4.0, and the community-led construction of the source data in Wiktionary, there

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

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 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 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, which could affect our measures of formal or distributional change.

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

[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.

UniMorph derivation data is not publicly available; Serbo-Croatian because the UniMorph data is in Latin script while the vast majority of Serbo-Croatian text used in the construction of the FastText vectors is written in Cyrillic; and Nynorsk because FastText does not distinguish between Nynorsk and Bokmål, and Bokmål is the large majority of written Norwegian.

As mentioned in Section 3.2, 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 2 contains descriptive statistics about the representation of languages, morphological typologies, and language families within our filtered dataset. Indo-European languages and, accordingly, languages with fusional typology are heavily represented in our data; however, we also have data from five languages which are not Indo-European, representing four major language families; and six languages with an agglutinative typology. We acknowledge that many language families with distinctive morphological typologies, such as the Niger-Congo languages, the Inuit-Yupik languages, and the Semitic languages, are not represented in the present study. Nevertheless, even results on a broad range of Indo-European languages plus a few others is a substantial advance in the typological coverage of existing work in the area.

Table 2: Descriptive statistics of our filtered dataset by language.

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

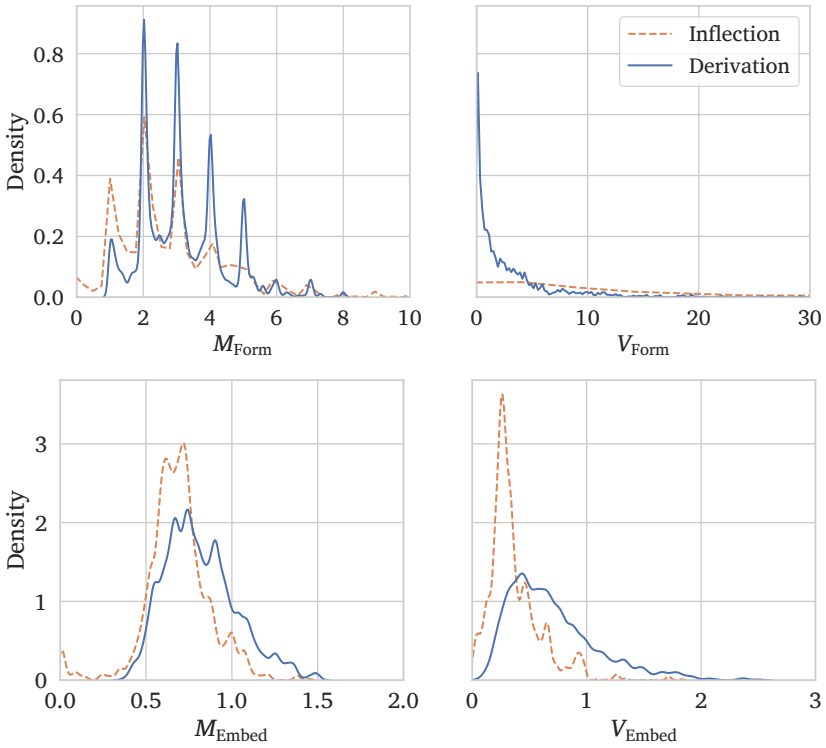


Figure 1: The empirical distributions of our four measures (quantifying the magnitude M and variability V of changes in Form and in Embedding space) for inflections and derivations in UniMorph

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 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 (1994). This is discussed in Section 8.1.

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

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

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 fre-

quency rather than to the linguistic properties of interest.[9] Accordingly, it is desirable to quantify these relationships with frequency.

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

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,

[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.

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 7 and Section 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** $\left(V_{\textbf{Form}}, V_{\textbf{Embed}}\right)$: 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** $\left(M_{\textbf{Form}}, M_{\textbf{Embed}}\right)$: 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** $\left(M_{\text{Form}}, V_{\text{Form}}\right)$ or (D) **syntax/semantics** $\left(M_{\text{Embed}}, V_{\text{Embed}}\right)$ 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.

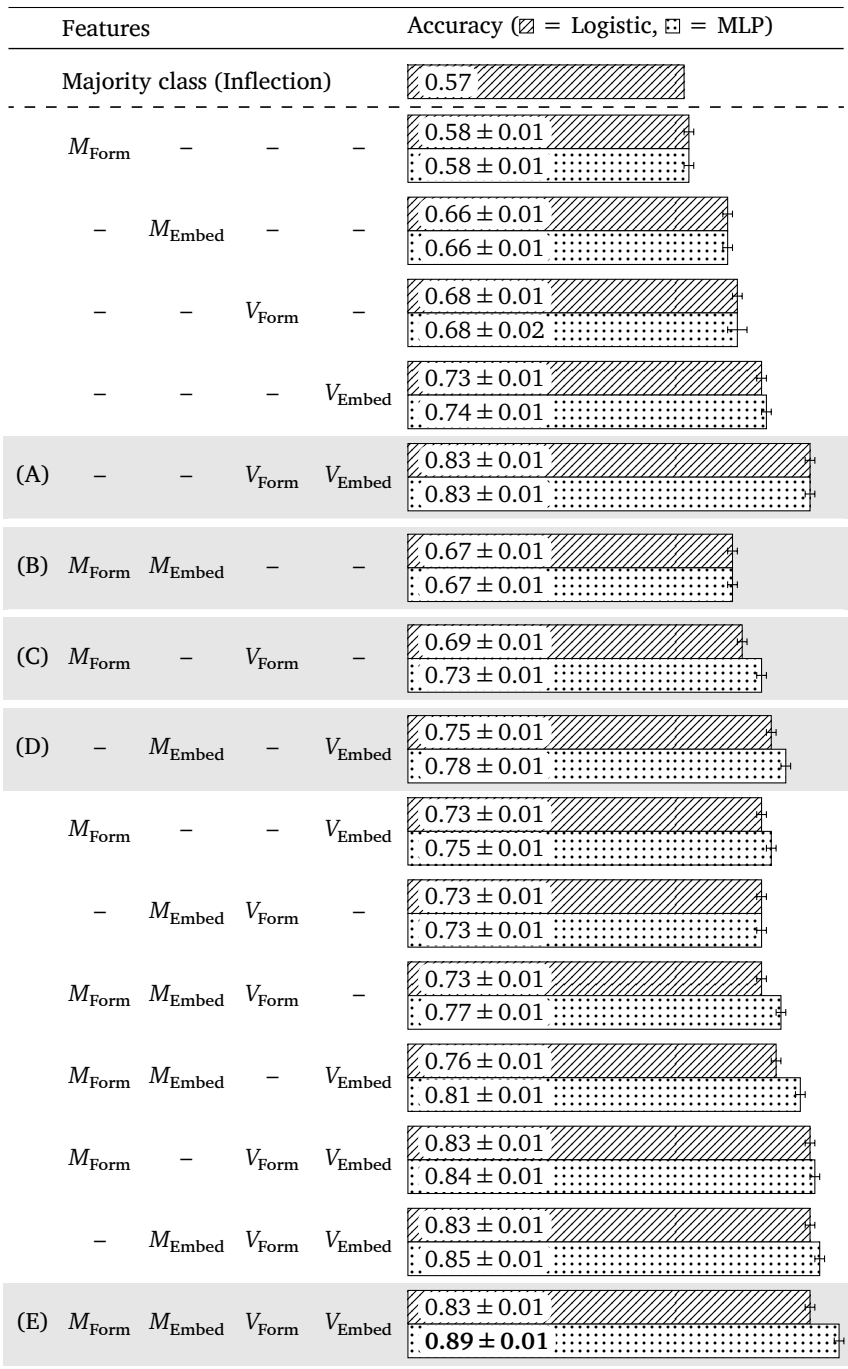
First, we consider the logistic regression classifier. As described in Section 2, the expectation from linguistic theory is that greater values of any measure should be associated with that construction being derivational. Our analysis in Section 5 largely backs up this relation (with the relationship 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, 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 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 lay-

[10] This issue of whether the distinction is gradient or categorical with respect to our measures is discussed further in Section 8.4.

Figure 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



ers 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, $\left(M_{\text{Form}}, V_{\text{Form}}\right)$ 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 $\left(M_{\text{Embed}}, V_{\text{Embed}}, V_{\text{Form}}\right)$. 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).

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.

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 1966; Silverstein 1986; Croft 2002; Ackema and Neeleman 2019), and semantic markedness often corresponds to morphological marking. For example, past tense is generally considered more semantically marked than present, and in many languages the past tense requires an affix while the present tense does not. However, the UniMorph annotations include both the semantically marked and unmarked inflections (e.g., V;PAST;PL and V;PST;PL for Ukrainian verbs). Therefore, for the purposes of this analysis, we consider active voice, singular number, nominative case,[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

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

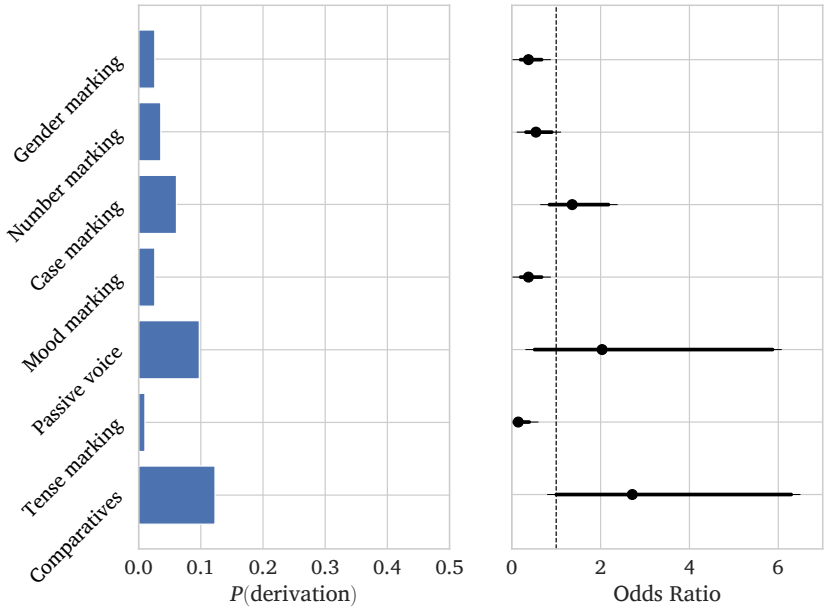


Figure 3: Probability and Odds ratio with 95% confidence intervals of being classified as derivation for various kinds of inflectional meaning. Inflections to the right of the dotted line were disproportionately likely to be classified as derivation by our model

not in the featural descriptions. For the category of gender, we simply consider nouns not to be marked, as their gender is typically not a morphological process but a lexical property.

Figure 3 displays the probability that a construction marking for one of these inflection types will be classified as derivation by our best-performing model. As can be seen in the figure, our model does not classify any of these major kinds of inflection as *more derivational than inflectional*; each is substantially more likely to be classified as inflection than derivation. This finding is perhaps unsurprising given our model's cross-linguistic test set classification accuracy of 90% – it classifies 92% of inflections correctly in general. Accordingly, classifying just 15-20% of constructions belonging to a particular inflectional category as derivations has the potential to be significant.

In order to answer the question “Are constructions which mark for this inflection type significantly more likely to be classified as derivational than others?”, we compute the odds ratio. We focus on the best performing MLP model (using all 4 features) in these results, which are presented in Figure 3 with 95% confidence intervals. Constructions with an odds ratio significantly greater than 1, while not more

likely to be classified as derivation than inflection, can nevertheless be thought of as particularly *non-canonical* types of inflection under our model, while those with odds ratios significantly below 1 are *canonical* with respect to our model.

We apply the Boschloo exact test (Boschloo 1970) to the results and correct for multiple comparisons with the Bonferroni correction, which yields a signficance 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.

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 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

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

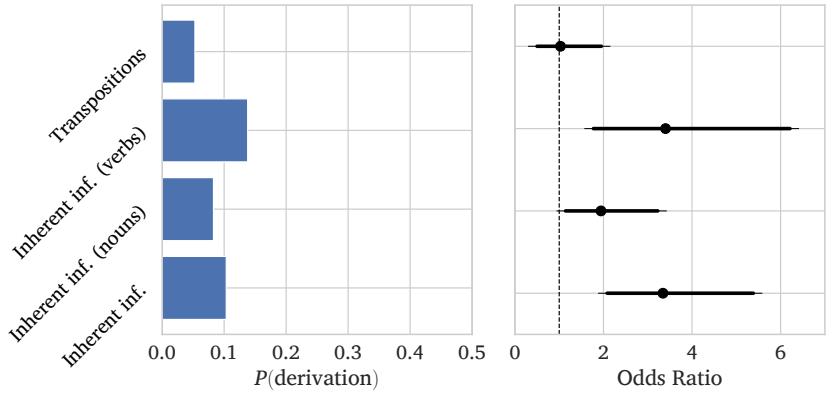


Figure 4: Probability and Odds ratio with 95% confidence intervals of being classified as derivation for inherent inflections and transpositions

by its subject). In contrast, inherent inflection is non-canonical: it contributes to the meaning of the word itself (e.g. the plural noun). To operationalise this in a simple, cross-linguistically consistent way, we associate number, gender, and case[13] with nouns – meaning that when those features appear on other parts of speech, we consider them contextual inflections. Analogously, we associate mood, tense, and voice with verbs. We then may consider whether an inflection is *inherent* or not, where we define inherency as not marking *any* contextual features. As shown in Figure 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 5, which shows that inherent case is significantly associated with being classified as derivation ($p = 1 \times 10^{-5}$), while contextual case ($p = 0.003$) and contextual number ($p = 0.0008$) are significantly associated with being classified as inflection.

Finally, we consider inflectional transpositions, denoted in UniMorph as participles (deverbal adjectives), converbs (deverbal ad-

[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.

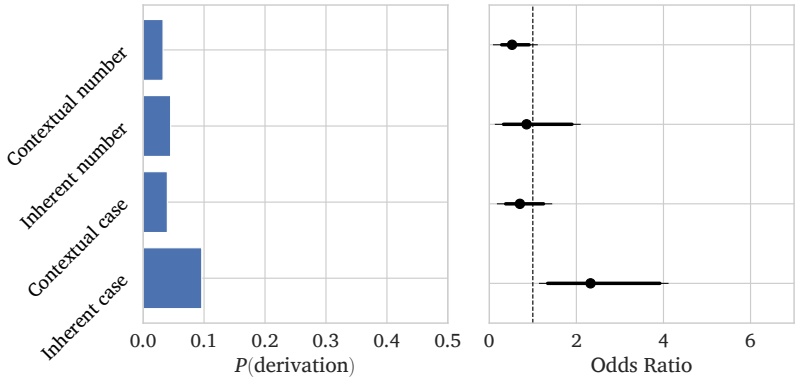


Figure 5: Probability and Odds ratio with 95% confidence intervals of being classified as derivation for inherent vs. contextual noun inflections

verbs), and masdars (deverbal nouns), shown in Figure 4. Transpositions have often been argued to be non-canonical inflection or even derivation because transpositions change the part of speech (Spencer 2013; Plank 1994; Haspelmath 2024). We here find under our model that transpositions appear neither significantly more or less likely to be classified as derivations than inflections by our model – neither particularly canonical 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.

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.

8 DISCUSSION

8.1 *The role of our individual measures*

As shown in Section 6, all four of our measures can be used to achieve better discrimination between traditional concepts of inflection and derivation; however, not every feature plays an equally large role. In this section, we discuss the roles played by each of our features and their connection to linguistic theory.

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

In addition, we find a clear and complementary influence of the variability of the change in form, V_{Form}: adding this feature to our model produces a large increase in performance, even when V_{Embed} is already included. This measure (described in Section 3.1) can be thought of as a weighted measure of allomorphy, capturing not just

the number of distinct patterns, but also their similarity. Our results point to a much higher degree of formal variability/allomorphy for inflections than derivations across a wide range of languages, contrary to the predictions of Plank (1994) and Dressler (1989). 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 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 informa-

tion are also useful.[14] While overall performance of the classifier was lower (likely due to overall worse quality of the embeddings, for the reasons described in Section 3.2), 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.

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

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

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.

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.

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.

While, as discussed in 8.3, we focus primarily on the *consistency* of traditional categories of inflection and derivation, in this section

[15] As described in Section 8.4 and shown in Figure 6, it is this situation in which we find ourselves.

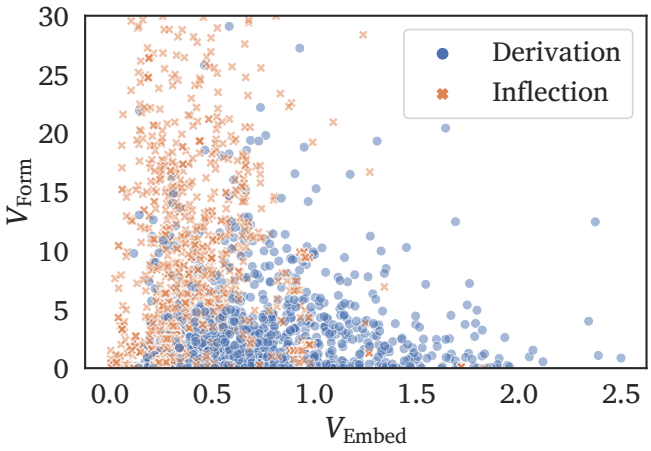


Figure 6: Our two most predictive measures for inflection and derivation. Saturation represents overlapping constructions. With respect to these two variables, the inflection–derivation distinction appears gradient rather than categorical

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 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.

8.5 *Are inflection and derivation identifiable from the statistics of language?*

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

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

The answer to this question has implications within psycholinguistics as well as computational linguistics. Psycholinguistics provides some empirical evidence that inflection and derivation are processed differently (Laudanna *et al.* 1992; Kirkici and Clahsen 2013), which seems to imply learners have some implicit ability to 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.

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 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.* (2014--2019), Larasati *et al.* (2011), Strunk (2020), or Vilca *et al.* (2012) into multilingual resources is essential to answering truly general typological questions with these resources in the future.

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

[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.

example. Particularly interesting, we believe, would be the investigation of a syntactic distance and variability component, drawing on works such as He *et al.* (2018) and Ravfogel *et al.* (2020) – though there are significant challenges to operationalising these embeddings in a multilingual, low-resource domain.

There is also room for refinement of our measures and classification techniques. For example, extension to many other languages would likely require a re-assessment of our use of orthography as a proxy for linguistic form. The assumption that orthography is a reasonable proxy for form is not accurate in many languages – however, at present UniMorph does not include phonological transcriptions, and automated grapheme-to-phoneme conversion across a broad range of languages is the subject of very active research (Ashby *et al.* 2021). These difficulties would need to be overcome in order to use phonological transcriptions. Future work should also investigate to what degree our variability of embedding measure is equivalent to or complementary to Copot *et al.* (2022)'s predictability of frequency measure, as both are motivated from semantic drift due to a change in lexical index. Similarly, future work could clarify the contribution of distributional semantics by using a model such as Word2Vec or GloVe, or newer models of distributional semantics, such as XLM-R (Conneau *et al.* 2020) – though in the latter case they would have to overcome the difficulties of multilingual decontextualisation as described in Section 3.2. 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.

9 CONCLUSION

In this work, we have presented the first multilingual computational study of the inflection–derivation distinction. In Section 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 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 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 1994; Bybee 1985), and enable us to produce the first large-scale quantification of how consistently the categories of inflection and derivation are applied, and validate that these measures can *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 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 cat-

egories 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 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.

ACKNOWLEDGEMENTS

The authors would like to thank Paul J.W. Schauenburg, Albert Haley, Itamar Kastner, Kate Mccurdy, and Francis Mollica for their comments on this work. This work was performed using resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service (`www.csd3.cam.ac.uk`), provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/T022159/1), and DiRAC funding from the Science and Technology Facilities Council (`www.dirac.ac.uk`). This work was in part supported by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics and School of Philosophy, Psychology & Language Sciences.

REFERENCES

Peter Ackema and Ad Neeleman (2019), *Default person versus default number in agreement*, pp. 21–54, Open Generative Syntax, Language Science Press, ISBN 9783961102013, doi:10.5281/zenodo.3458062.

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

Stephen R. Anderson (1985), Inflectional Morphology, in *Language Typology and Syntactic Description*, volume 3, pp. 150–201, Cambridge University Press, 1 edition.

Antti Arppe, Atticus Harrigan, Katherine Schmirler, Lene Antonsen, Trond Trosterud, Sjur Nørstebø Moshagen, Miikka Silfverberg, Arok Wolvengrey, Conor Snoek, Jordan Lachler, Eddie Antonio Santos, Jean Okimāsis, and Dorothy Thunder (2014--2019), Finite-State Transducer-Based Computational Model of Plains Cree Morphology, https://giellalt.uit.no/lang/crk/PlainsCreeDocumentation.html.

Lucas F.E. Ashby, Travis M. Bartley, Simon Clematide, Luca Del Signore, Cameron Gibson, Kyle Gorman, Yeonju Lee-Sikka, Peter Makarov, Aidan Malanoski, Sean Miller, Omar Ortiz, Reuben Raff, Arundhati Sengupta, Bora Seo, Yulia Spektor, and Winnie Yan (2021), Results of the Second SIGMORPHON Shared Task on Multilingual Grapheme-to-Phoneme Conversion, in *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pp. 115–125, Association for Computational Linguistics, Online, doi:10.18653/v1/2021.sigmorphon-1.13, https://aclanthology.org/2021.sigmorphon-1.13.

Madina Babazhanova, Maxat Tezekbayev, and Zhenisbek Assylbekov (2021), Geometric Probing of Word Vectors, in *ESANN 2021 Proceedings - 29th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pp. 587–592, i6doc.com publication, Virtual, Online, Belgium, doi:10.14428/esann/2021.ES2021-105.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia (2021), MorphyNet: a Large Multilingual Database of Derivational and Inflectional Morphology, in *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pp. 39–48, Association for Computational Linguistics, Online, doi:10.18653/v1/2021.sigmorphon-1.5, https://aclanthology.org/2021.sigmorphon-1.5.

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya

Ganieva, Hilaria Cruz, Ritván Karahóǧa, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud'hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova (2022), UniMorph 4.0: Universal Morphology, in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 840–855, European Language Resources Association, Marseille, France, https://aclanthology.org/2022.lrec-1.89.

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

Sacha Beniamine, Martin Maiden, and Erich Round (2020), Opening the Romance Verbal Inflection Dataset 2.0: A CLDF lexicon, in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 3027–3035, European Language Resources Association, Marseille, France, ISBN 979-10-95546-34-4, https://aclanthology.org/2020.lrec-1.370.

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

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov (2017), Enriching Word Vectors with Subword Information, *Transactions of the Association for Computational Linguistics*, 5:135–146, doi:10.1162/tacl_a_00051, https://aclanthology.org/Q17-1010.

Rishi Bommasani, Kelly Davis, and Claire Cardie (2020), Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4758–4781, Association for Computational Linguistics, Online, doi:10.18653/v1/2020.acl-main.431, https://aclanthology.org/2020.acl-main.431.

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

Olivier Bonami and Jana Strnadová (2019), Paradigm structure and predictability in derivational morphology, *Morphology*, 29(2):167–197, ISSN 1871-5656, doi:10.1007/s11525-018-9322-6, https://doi.org/10.1007/s11525-018-9322-6.

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

Geert Booij (2007), Inflection, in *The Grammar of Words: An Introduction to Linguistic Morphology*, Oxford University Press, ISBN 9780199226245, doi:10.1093/acprof:oso/9780199226245.003.0005, https://doi.org/10.1093/acprof:oso/9780199226245.003.0005.

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

Joan L Bybee (1985), *Morphology: A study of the relation between meaning and form*, John Benjamins, Amsterdam.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov (2020), Unsupervised Cross-lingual Representation Learning at Scale, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Association for Computational Linguistics, Online, doi:10.18653/v1/2020.acl-main.747, https://aclanthology.org/2020.acl-main.747.

Maria Copot, Timothee Mickus, and Olivier Bonami (2022), Idiosyncratic frequency as a measure of derivation vs. inflection, *Journal of Language Modelling*, 10(2):193–240, doi:10.15398/jlm.v10i2.301, https://jlm.ipipan.waw.pl/index.php/JLM/article/view/301.

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

Ryan Cotterell and Hinrich Schütze (2018), Joint Semantic Synthesis and Morphological Analysis of the Derived Word, *Transactions of the Association for Computational Linguistics*, 6:33–48, doi:10.1162/tacl_a_00003, https://aclanthology.org/Q18-1003.

William Croft (2002), *Typology and Universals*, Cambridge Textbooks in Linguistics, Cambridge University Press, 2 edition, doi:10.1017/CBO9780511840579.

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

Daniel Deutsch, John Hewitt, and Dan Roth (2018), A Distributional and Orthographic Aggregation Model for English Derivational Morphology, in

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1938–1947, Association for Computational Linguistics, Melbourne, Australia, doi:10.18653/v1/P18-1180, https://aclanthology.org/P18-1180.

Jacob DEVLIN, Ming-Wei CHANG, Kenton LEE, and Kristina TOUTANOVA (2019), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Association for Computational Linguistics, Minneapolis, Minnesota, doi:10.18653/v1/N19-1423, https://aclanthology.org/N19-1423.

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

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

Edouard GRAVE, Piotr BOJANOWSKI, Prakhar GUPTA, Armand JOULIN, and Tomas MIKOLOV (2018), Learning Word Vectors for 157 Languages, in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, European Language Resources Association (ELRA), Miyazaki, Japan, https://aclanthology.org/L18-1550.

Joseph H. GREENBERG, editor (1966), *Universals of language*, M.I.T. Press, 2 edition.

P. HACKEN (1994), *Defining Morphology: A Principled Approach to Determining the Boundaries of Compounding, Derivation, and Inflection*, Altertumswissenschaftliche Texte Und Studien, G. Olms Verlag, ISBN 9783487098913, https://books.google.co.uk/books?id=E8mWh_6mRAcC.

Zellig HARRIS (1954), Distributional structure, *Word*, 10(23):146–162.

Martin HASPELMATH (2024), Inflection and derivation as traditional comparative concepts, *Linguistics*, 62(1):43–77, doi:doi:10.1515/ling-2022-0086, https://doi.org/10.1515/ling-2022-0086.

Nabil HATHOUT and Fiammetta NAMER (2016), Giving Lexical Resources a Second Life: Démonette, a Multi-sourced Morpho-semantic Network for French, in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 1084–1091, European Language Resources Association (ELRA), Portorož, Slovenia, https://aclanthology.org/L16-1173.

Nabil HATHOUT, Franck SAJOUS, and Basilio CALDERONE (2014), GLÀFF, a Large Versatile French Lexicon, in *Proceedings of the Ninth International*

Conference on Language Resources and Evaluation (LREC'14), pp. 1007–1012, European Language Resources Association (ELRA), Reykjavik, Iceland, http://www.lrec-conf.org/proceedings/lrec2014/pdf/58_Paper.pdf.

Junxian He, Graham Neubig, and Taylor Berg-Kirkpatrick (2018), Unsupervised Learning of Syntactic Structure with Invertible Neural Projections, in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1292–1302, Association for Computational Linguistics, Brussels, Belgium, doi:10.18653/v1/D18-1160, https://aclanthology.org/D18-1160.

Valentin Hofmann, Hinrich Schütze, and Janet Pierrehumbert (2020), A Graph Auto-encoder Model of Derivational Morphology, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1127–1138, Association for Computational Linguistics, Online, doi:10.18653/v1/2020.acl-main.106, https://aclanthology.org/2020.acl-main.106.

Zongliang Hu, Kai Dong, Wenlin Dai, and Tiejun Tong (2017), A Comparison of Methods for Estimating the Determinant of High-Dimensional Covariance Matrix, *The International Journal of Biostatistics*, 13(2):20170013, doi:doi:10.1515/ijb-2017-0013, https://doi.org/10.1515/ijb-2017-0013.

M. Kasthuri, S. Britto Ramesh Kumar, and Souheil Khaddaj (2017), PLIS: Proposed Language Independent Stemmer for Information Retrieval Systems Using Dynamic Programming, in *2017 World Congress on Computing and Communication Technologies (WCCCT)*, pp. 132–135, doi:10.1109/WCCCT.2016.39.

Diederik P. Kingma and Jimmy Ba (2015), Adam: A Method for Stochastic Optimization, in Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, http://arxiv.org/abs/1412.6980.

Bilal Kirkici and Harald Clahsen (2013), Inflection and derivation in native and non-native language processing: Masked priming experiments on Turkish, *Bilingualism: Language and Cognition*, 16(4):776–791, doi:10.1017/S1366728912000648.

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

Lukáš Kyjánek, Zdeněk Žabokrtský, Magda Ševčíková, and Jonáš Vidra (2020), Universal Derivations 1.0, A Growing Collection of Harmonised Word-Formation Resources, *The Prague Bulletin of Mathematical Linguistics*, 2(115):333–348.

Lior Laks and Fiammetta Namer (2022), Hebrewnette--A New Derivational Resource for Non-concatenative Morphology: Principles, Design and Implementation, *The Prague Bulletin of Mathematical Linguistics*, 118:25–53.

Septina Dian Larasati, Vladislav Kuboň, and Daniel Zeman (2011), Indonesian Morphology Tool (MorphInd): Towards an Indonesian Corpus, in Cerstin Mahlow and Michael Piotrowski, editors, *Systems and Frameworks for Computational Morphology*, pp. 119–129, Springer Berlin Heidelberg, Berlin, Heidelberg, doi:10.1007/978-3-642-23138-4_8.

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

Vladimir Levenshtein (1966), Binary Codes Capable of Correcting Deletions, Insertions and Reversals, *Soviet Physics Doklady*, 10:707.

Chu-Cheng Lin, Waleed Ammar, Chris Dyer, and Lori Levin (2015), Unsupervised POS Induction with Word Embeddings, in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1311–1316, Association for Computational Linguistics, Denver, Colorado, doi:10.3115/v1/N15-1144, https://aclanthology.org/N15-1144.

Nikola Ljubešić, Filip Klubička, Željko Agić, and Ivo-Pavao Jazbec (2016), New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian, in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 4264–4270, European Language Resources Association (ELRA), Portorož, Slovenia, https://aclanthology.org/L16-1676.

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

Robert Malouf, Farrell Ackerman, and Arturs Semenuks (2020), Lexical databases for computational analyses: A linguistic perspective, in Allyson Ettinger, Gaja Jarosz, and Joe Pater, editors, *Proceedings of the Society for Computation in Linguistics 2020*, pp. 446–456, Association for Computational Linguistics, New York, New York, https://aclanthology.org/2020.scil-1.52.

Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky (2020), UniMorph 3.0: Universal Morphology, in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 3922–3931, European Language

Resources Association, Marseille, France, ISBN 979-10-95546-34-4, https://aclanthology.org/2020.lrec-1.483.

Tomas MIKOLOV, Ilya SUTSKEVER, Kai CHEN, Greg CORRADO, and Jeffrey DEAN (2013), Distributed Representations of Words and Phrases and Their Compositionality, in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, p. 3111–3119, Curran Associates Inc., Red Hook, NY, USA.

Karthik NARASIMHAN, Regina BARZILAY, and Tommi JAAKKOLA (2015), An unsupervised method for uncovering morphological chains, *Transactions of the Association for Computational Linguistics*, 3:157–167.

Bruce OLIVER, Clarissa FORBES, Changbing YANG, Farhan SAMIR, Edith COATES, Garrett NICOLAI, and Miikka SILFVERBERG (2022), An Inflectional Database for Gitksan, in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 6597–6606, European Language Resources Association, Marseille, France, https://aclanthology.org/2022.lrec-1.710.

Jeffrey PENNINGTON, Richard SOCHER, and Christopher MANNING (2014), GloVe: Global Vectors for Word Representation, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Association for Computational Linguistics, Doha, Qatar, doi:10.3115/v1/D14-1162, https://aclanthology.org/D14-1162.

David PERLMUTTER (1988), The split morphology hypothesis: Evidence from Yiddish, *Theoretical morphology*, pp. 79–100.

Tiago PIMENTEL, Josef VALVODA, Rowan Hall MAUDSLAY, Ran ZMIGROD, Adina WILLIAMS, and Ryan COTTERELL (2020), Information-Theoretic Probing for Linguistic Structure, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4609–4622, Association for Computational Linguistics, Online, doi:10.18653/v1/2020.acl-main.420, https://aclanthology.org/2020.acl-main.420.

Frans PLANK (1994), Inflection and Derivation, in *The Encyclopedia of Language and Linguistics*, pp. 1671–1679, Elsevier Science and Technology, Amsterdam.

Shauli RAVFOGEL, Yanai ELAZAR, Jacob GOLDBERGER, and Yoav GOLDBERG (2020), Unsupervised Distillation of Syntactic Information from Contextualized Word Representations, in *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 91–106, Association for Computational Linguistics, Online, doi:10.18653/v1/2020.blackboxnlp-1.9, https://aclanthology.org/2020.blackboxnlp-1.9.

Rudolf ROSA and Zdeněk ŽABOKRTSKÝ (2019), Attempting to separate inflection and derivation using vector space representations, in *Proceedings of the Second International Workshop on Resources and Tools for Derivational*

Morphology, pp. 61–70, Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Prague, Czechia, https://aclanthology.org/W19-8508.

Rudolf ROSA and Zdenek ZABOKRTSKÝ (2019), Unsupervised Lemmatization as Embeddings-Based Word Clustering, *CoRR*, abs/1908.08528, http://arxiv.org/abs/1908.08528.

Jenny R SAFFRAN, Richard N ASLIN, and Elissa L NEWPORT (1996), Statistical learning by 8-month-old infants, *Science*, 274(5294):1926–1928.

Adriaan M. J. SCHAKEL and Benjamin J. WILSON (2015), Measuring Word Significance using Distributed Representations of Words, *Computing Research Repository*, arXiv:1508.02297, http://arxiv.org/abs/1508.02297.

Patrick SCHONE and Daniel JURAFSKY (2000), Knowledge-Free Induction of Morphology Using Latent Semantic Analysis, in *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*, https://aclanthology.org/W00-0712.

Michael SILVERSTEIN (1986), *7. Hierarchy of Features and Ergativity*, pp. 163–232, De Gruyter Mouton, Berlin, Boston, ISBN 9783110871661, doi:doi:10.1515/9783110871661-008, https://doi.org/10.1515/9783110871661-008.

Radu SORICUT and Franz OCH (2015), Unsupervised Morphology Induction Using Word Embeddings, in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1627–1637, Association for Computational Linguistics, Denver, Colorado, doi:10.3115/v1/N15-1186, https://aclanthology.org/N15-1186.

Andrew SPENCER (2013), *Lexical Relatedness*, Oxford University Press, Oxford.

Pavol ŠTEKAUER (2015), 14. The delimitation of derivation and inflection, in Peter O. MÜLLER, Ingeborg OHNHEISER, Susan OLSEN, and Franz RAINER, editors, *Volume 1 Word-Formation*, pp. 218–235, De Gruyter Mouton.

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

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

John SYLAK-GLASSMAN (2016), The Composition and Use of the Universal Morphological Feature Schema (UniMorph Schema), https://unimorph.github.io/doc/unimorph-schema.pdf.

Erik D THIESSEN, Alexandra T KRONSTEIN, and Daniel G HUFNAGLE (2013), The extraction and integration framework: a two-process account of statistical learning., *Psychological bulletin*, 139(4):792.

Erik D Thiessen and Jenny R Saffran (2003), When cues collide: use of stress and statistical cues to word boundaries by 7-to 9-month-old infants., *Developmental psychology*, 39(4):706.

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

Hugo David Calderon Vilca, Flor Cagniy Cárdenas Mariñó, and Edwin Fredy Mamani Calderon (2012), Analizador morfólogico de la lengua Quechua basado en software libre Helsinkifinite-statetransducer (HFST).

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart, and Anna Korhonen (2020), Multi-SimLex: A Large-Scale Evaluation of Multilingual and Crosslingual Lexical Semantic Similarity, *Computational Linguistics*, 46(4):847–897, doi:10.1162/coli_a_00391, https://aclanthology.org/2020.cl-4.5.

Christian Wartena (2013), Distributional Similarity of Words with Different Frequencies, in *Proceedings of the 13th edition of the Dutch-Belgian information retrieval Workshop (DIR 2013)*, pp. 8–11, Hochschule Hannover.

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

Coleman Haley
ⓘ 0000-0003-3089-9558
coleman.haley@ed.ac.uk

Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh
Edinburgh, UK

Edoardo M. Ponti
ⓘ 0000-0002-6308-1050
eponti@ed.ac.uk

Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh
Edinburgh, UK

Sharon Goldwater
ⓘ 0000-0002-7298-0947
sgwater@inf.ed.ac.uk

Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh
Edinburgh, UK

Coleman Haley, Edoardo M. Ponti, and Sharon Goldwater (1970), *Corpus-based measures discriminate inflection and derivation cross-linguistically*, Journal of Language Modelling, $i^2(e^{i\pi})$:1–54
doi https://dx.doi.org/10.15398/jlm.vi^2i$e^{i\pi}$.351