

Corpus-based measures discriminate inflection and derivation cross-linguistically

Anonymous

ABSTRACT

In morphology, a distinction is commonly drawn between inflection and derivation. However, a precise definition of this distinction which captures the way the terms are used 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 how much and how variably a morphological construction changes aspects of the lexical entry, specifically, the lexeme's form and the lexeme's semantic and syntactic properties (as operationalised by distributional word embeddings). Based on a sample of 26 languages, we find that we can reconstruct 90% 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 signatures 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 distribution. We find 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, there are still many

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

constructions near the model's decision boundary between the two categories, indicating a gradient, rather than categorical, distinction.

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. 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, which organises constructions into inflections and derivations based on typical traditions for a given language. The inflection–derivation distinction in Unimorph is therefore determined by what Haspelmath terms *traditional comparative concepts* (Haspelmath forthcoming), relating to the ways in which traditional Western dictionaries and grammar books are structured. Unimorph has been largely successful in annotating

inflectional paradigms from a wide range of languages in a cross-linguistically consistent way – indicating a high degree of consistency in what morphosyntactic features are considered inflectional.

Despite this relative consistency at the level of annotation, there is considerable disagreement amongst linguists about the fundamental properties that might underlie or explain these traditional categorizations – such as the degree of syntactic or semantic change, or the creation of new words. As an example, Plank (1994) covers no fewer than 28 tests for inflectional and derivational status. Upon applying them to just 6 English morphological constructions, Plank (1994) finds significant contradictions between the 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.* forthcoming; 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 forthcoming).

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, yet it remains unclear to what extent they successfully capture the traditional concepts of inflection and derivation on a large scale – perhaps measures based on these subjective tests can 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 languages studied and the measures used. In particular, Bonami and Paperno (2018) and Copot *et al.* (forthcoming) explored semantic and frequency-based measures of *variability* in French, aiming to test the claim that derivation tends to introduce more *idiosyncratic* (variable) changes than inflection. Meanwhile, Rosa and Žabokrtský (2019) looked at the *magnitude* of orthographic and se-

mantic change between morphologically related forms in Czech, following the claim that derivation tends to introduce *larger* changes than inflection. All of these studies found significant 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 considered, 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, including measures of *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). We then ask whether, for a given construction, knowing just these measures is sufficient to predict its inflectional versus derivational status in Unimorph. In other words, to what extent can purely quantitative information about wordforms and corpus distribution recapitulate the linguistic intuitions, subjective tests, and comparative concepts encapsulated in the Unimorph annotations? If, across a variety of languages, belonging to different grammatical traditions, language families, and morphological typologies, the Unimorph annotations can be predicted with high accuracy based on our four measures, this would provide evidence that traditional concepts of inflection and derivation *do* closely correspond to intuitions about the different *types* of changes inflection and derivation induce.

To explore this question, we train two different types of machine learning models (a logistic regression classifier and a multilayer perceptron). For each construction in our training set, the models are trained to predict whether the construction is inflectional or derivational, given just four input features: our measures of the magnitude and variability of the changes in wordform and distributional representations. Since we are interested in the cross-linguistic consistency of these predictors, the models are not given access to the input language or any of its typological features. In experiments on 26 languages (including five from non-Indo-European families) and 2,772 constructions, we find that both models are able to predict with high accuracy whether a held-out construction is listed as inflection or derivation in Unimorph (86% and 90%, 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.

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 for a characterisation of inflection and derivation based on a notion of changes to a “lexical entry.” Under this view, a wordform is represented by four components: 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, on the other hand, argues that both inflection and derivation may change *any of these 4 components*, but canonical inflections correspond to small changes of 1 or 2 of the components while canonical derivations correspond to larger changes of more of the components.

This description of the distinction naturally unifies many diagnostics, capturing and generalising notions like derivations causing larger changes in the semantics or changing part of speech, while also suggesting ones that are less frequently claimed, such as derivational relations typically involving larger changes to the form of a word.¹ 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

¹This is suggested, though not explicitly, by criteria like Plank (1994)’s “derivational morphemes resemble free morphs.”

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 significant amounts of information about a word's syntactic category, as operationalised by its part of speech (Pimentel *et al.* 2020; Lin *et al.* 2015). Because of the distributional nature of meaning, it is in fact difficult to induce a space from pure language data where distance corresponds to *syntactic* similarity entirely independently from *semantic* similarity. While there is prior work on inducing such representational spaces (e.g., He *et al.* 2018; Ravfogel *et al.* 2020), due to our complex and highly multilingual setting, we instead choose to *collapse* the distinction of syntactic and semantic change made by Spencer, focusing on what is captured by embeddings designed primarily for capturing semantics but which also capture syntactic information. In particular, we use FastText embeddings, described in more detail in Section 3.2.

To capture the notion of a change in lexical index, we look at the variability of various aspects of the distribution of the word. Words with different lexical indices are thought to have processes like semantic drift apply separately from each other (Spencer 2013; Copot *et al.* forthcoming; Bonami and Paperno 2018); derivational relations (which have historically been argued to either always or usually change the lexical index) have been claimed to have more variability in the changes made to the form (Plank 1994).

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

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

| 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 | re-decoration | -ion | V | N | ENG |
| elide | elision | -ion | V | N | ENG |
| ... | ... | ... | ... | ... | ... |

to or produce different parts of speech. This is motivated by examples like agentive *-er* vs. comparative *-er* in English, which differ only in the parts of speech which they apply to and produce.

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:

- $\|\Delta_{\text{form}}\|$ and $\text{var}(\Delta_{\text{form}})$: the average magnitude of the change in form induced by a construction, and the variability of that change.
- $\|\Delta_{\text{embed}}\|$ and $\text{var}(\Delta_{\text{embed}})$: the average magnitude of the change in semantic/syntactic embedding space induced by a construction, and the variability of that change.

These measures, while inspired by Spencer's factorisation of the inflection–derivation distinction, are also predicted to be relevant by criteria from other accounts. For example, in Table 2, we relate Plank's survey of criteria to what predictions they make about the relative values of our metrics for inflectional and derivational constructions.

Table 2: Plank's tests for derivational status. The "In corpus" column indicates whether this measure could be determined from the information in the linguistic signal alone, while the columns "Formal" and "Distributional" indicate whether a test implies measureable properties about the structure of word forms or the linguistic distribution of that word, respectively. An upward arrow in a column indicates that the test implies the corresponding measure should be greater for derivations than inflections, while a downward arrow indicates the opposite relation.

| # | Description | In corpus | Formal | $\ \Delta_{\text{form}}\ $ | $\text{var}(\Delta_{\text{form}})$ | Distributional | $\ \Delta_{\text{embed}}\ $ | $\text{var}(\Delta_{\text{embed}})$ |
|----|-------------------------------------|-----------|--------|----------------------------|------------------------------------|----------------|-----------------------------|-------------------------------------|
| 1 | different concept from base | ✓ | -- | -- | -- | ✓ | ↑ | ↑ |
| 2 | different lexeme | ✓ | -- | -- | -- | ✓ | -- | ↑ |
| 3 | concrete meaning | ✓ | -- | -- | -- | ✓ | ↑ | ↓ |
| 4 | non-obligatory | -- | -- | -- | -- | -- | -- | -- |
| 5 | non-paradigmatic | -- | -- | -- | -- | -- | -- | -- |
| 6 | non-relational | -- | -- | -- | -- | -- | -- | -- |
| 7 | no agreement on form | ✓ | -- | -- | -- | ✓ | -- | -- |
| 8 | no agreement with form | ✓ | -- | -- | -- | ✓ | -- | -- |
| 9 | not assigned by syntax | -- | -- | -- | -- | -- | -- | -- |
| 10 | replaceable by same POS | ✓ | -- | -- | -- | ✓ | -- | -- |
| 11 | located close to base in form | ✓ | ✓ | -- | -- | -- | -- | -- |
| 12 | base allomorphy | ✓ | ✓ | ↑ | -- | -- | -- | -- |
| 13 | idiosyncratic base allomorphy | ✓ | ✓ | -- | ↑ | -- | -- | -- |
| 14 | loosely-regulated morph allomorphy | ✓ | ✓ | -- | ↑ | -- | -- | -- |
| 15 | morphemes can have same meaning | ✓ | -- | -- | -- | ✓ | -- | -- |
| 16 | blocked by homonyms | ✓ | ✓ | -- | -- | -- | -- | -- |
| 17 | not fusional | -- | -- | -- | -- | -- | -- | -- |
| 18 | diverse semantic contribution | ✓ | -- | -- | -- | ✓ | -- | ↑ |
| 19 | opaque semantic relationship | ✓ | -- | -- | -- | ✓ | -- | ↑ |
| 20 | limited applicability | ✓ | ✓ | -- | -- | -- | -- | -- |
| 21 | unattested base | ✓ | ✓ | -- | -- | -- | -- | -- |
| 22 | changes POS | ✓ | -- | -- | -- | ✓ | ↑ | -- |
| 23 | applies to multiple POS | ✓ | -- | -- | -- | -- | -- | ↑ |
| 24 | creates multiple POS | ✓ | -- | -- | -- | ✓ | -- | ↑ |
| 25 | can be repeatedly applied | ✓ | ✓ | -- | -- | -- | -- | -- |
| 26 | dissimilar form from related morphs | ✓ | ✓ | -- | -- | ✓ | -- | -- |
| 27 | resembles free morph | ✓ | ✓ | ↑ | ↓ | -- | -- | -- |
| 28 | cross-linguistically uncommon | -- | -- | -- | -- | -- | -- | -- |

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

3

METHOD

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 $\|\Delta_{\text{form}}\|$ using the Levenshtein distance (Levenshtein 1966): we simply compute the average distance between each pair of words in the construction (assuming all edits count equally).

To measure the *variability* of the change in form $\text{var}(\Delta_{\text{form}})$, 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 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 average pairwise edit distance as our measure of variability.

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);² these were trained on Common Crawl³ and Wikipedia data which was automatically tagged by language to train language-specific embedding models (Grave *et al.* 2018). These FastText vectors are known to correlate well with human semantic similarity scores (Vulić *et al.* 2020; Bojanowski *et al.* 2017). While these word vectors are more commonly used as models of semantics than syntax, 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 encode syntactic information. We investigate the relationship between our embedding measures and syntax further in Section 8.2.

Even though FastText is capable of producing vectors for words not seen at training time, we find that the inclusion of 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.

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

²<https://fasttext.cc/docs/en/crawl-vectors.html>

³<https://commoncrawl.org/>

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. In particular, one strategy for getting type-level semantics from these models is to simply present the wordform with no surrounding context. While this has been shown to work well for monolingual models (Bommasani *et al.* 2020; Vulić *et al.* 2020), for several of our languages, no monolingual contextual model is available. This strategy has been shown to perform poorly when used on multilingual models, and while we are aware of other techniques such as the averaging technique of Bommasani *et al.* (2020), these remain unvalidated for languages other than English and multilingual models.

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 $\|\Delta_{\text{embed}}\|$ as the average Euclidean distance across all pairs in the construction. (Preliminary experiments indicated that Euclidean distance correlates more strongly with the inflection–derivation distinction than cosine distance.)

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 d_i between the embeddings, i.e., $d_i = E(b_i) - E(c_i)$. For a construction with n pairs, this yields a matrix of differences $D = d_1 \dots d_n$. We then make the simplifying assumption that the covariance between the dimensions of D is zero, which allows us to estimate the variance of D as the sum of the variances of the individual dimensions. While this assumption 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, sug-

gesting that this simplifying assumption does not prevent the measure from capturing relevant aspects of variability in the embedding space.

DATA

4

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

The data and annotations in Unimorph 4.0 are semi-automatically extracted from Wiktionary,⁴ 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).

⁴<https://en.wiktionary.org/>

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

Due to the semi-automatic annotation in Unimorph 4.0, and the community-led construction of the source data in Wiktionary, there could be some errors or systematic issues with the data – for example, if the annotation fails to collapse 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 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

Table 3: 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 |

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 MorphyNet, and 1,185 (42.7%) are considered derivational. Table 3 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.

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.

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 Welch's t -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 (Figure 1), we see relatively small effects of inflection-hood and derivation-hood: Cohen's d for $\|\Delta_{\text{form}}\|$ is 0.15, while for $\text{var}(\Delta_{\text{form}})$ it is 0.32. Despite the small difference in $\|\Delta_{\text{form}}\|$ between inflection and derivation, the difference does go in the expected direction, with $\|\Delta_{\text{form}}\|$ higher on average for derivation and inflection. However, on average, $\text{var}(\Delta_{\text{form}})$ is *lower*

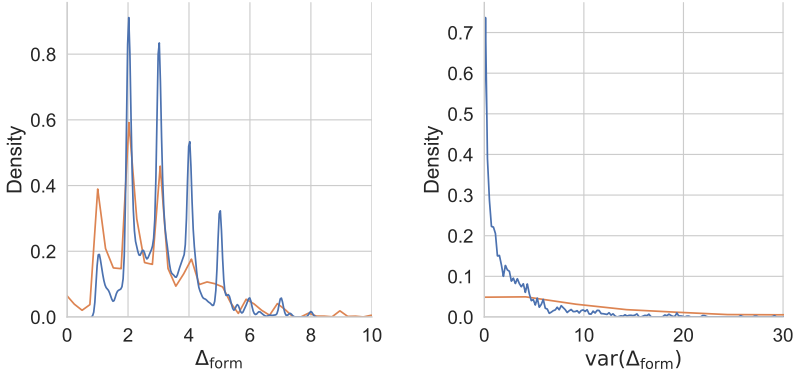


Figure 1:
The empirical distributions of the magnitude and variability in the change in form for inflections (■) and derivations (■) in Unimorph

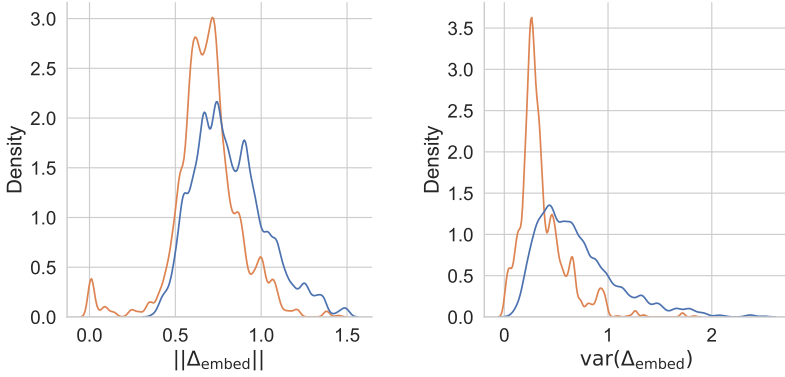


Figure 2:
The empirical distributions of the magnitude and variability in the change in embedding space for inflections (■) and derivations (■) in Unimorph

for derivation than for inflection – the opposite of what is suggested by Plank (1994) and Spencer (2013). This is discussed in Section 8.1.

In comparison to the form measures, the embedding-based semantics/syntax measures (Figure 2) are more strongly correlated with the inflection–derivation distinction. For $\|\Delta_{\text{embed}}\|$, we observe a Cohen's d of 0.67, indicating a moderately large effect of inflection- or derivation-hood on this measure; while for $\text{var}(\Delta_{\text{embed}})$ 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 frequency rather than to the linguistic properties of interest.⁵ 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

⁵The 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.

consideration reveals that none of them have a significant linear association with frequency, despite the large number of word forms. While there is a significant 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.

PREDICTING INFLECTION AND DERIVATION

6

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

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 an accuracy of 57% on the test set, 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 70%. In other words, a model could capture up to, but no more than, 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.

While we tested all possible combinations of features for each of our classification models, we here focus primarily on combinations which correspond to clear hypotheses about the factors which are descriptive of 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: 1. **Form/embed variability**: Perhaps it is the case that only variability matters, as investigated in the embedding case by Bonami and Paperno (2018). Or perhaps 2. **Form/embed magnitude**: 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 3. **form (Form magnitude + variability)** or 4. **syntax/semantics (Embed magnitude + variability)** alone can recover as much information as all the metrics combined. Finally, of course, there is the hypothesis 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.⁶ If the non-linear model is required to recover the distinction,

⁶This issue of whether the distinction is gradient or categorical with respect to our measures is discussed further in Section 8.5.

then discontinuous areas in the feature space may fall in a certain category, which would not neatly correspond with linguistic intuitions.

First, we consider a logistic regression classifier. As shown in Table 2 and discussed in Spencer (2013), the expectation from linguistic theory is that greater values of any measure should be associated with that construction being derivational. While our analysis in Section 5 largely backs up this relation (with the relationship being inverted for form variability), it is not clear to what degree this relationship is strictly linear; however, this type of linear modelling decreases the chance of our model picking up on statistical noise.

Due to our highly-restricted selection of measures, we are able to create classifiers with all possible combinations of features, selecting the best models on our validation data. As shown in Table 4, we find that the conjunction of all 4 features performs best, achieving a final test-set accuracy of 86%; however, it is closely followed by the **just variability** hypothesis at 84%. Similarly, the best possible combination of the three features ($\|\Delta_{\text{embed}}\|$, $\text{var}(\Delta_{\text{embed}})$, and $\text{var}(\Delta_{\text{form}})$) achieves a test-set accuracy of 85%, suggesting that $\|\Delta_{\text{form}}\|$ provides little information which is not redundant with the other measures in a linear-modelling setting.

While our logistic classification model can capture 29 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 layers and layer size was chosen using validation set performance, while the number of steps was chosen based on loss convergence on the training set. We found similar patterns of performance for most combinations of predictors. However, we found substantial improvements in performance for combinations of features which included both form features; for example, form magnitude + variability improving from 70% to 75%. Perhaps as a result of this, we achieve a test-set accuracy of 90%, when using all 4 predictors – representing a 4-point improvement over the best linear model, as well as a 4-point improvement over the best combination of 3 measures using the MLP ($\|\Delta_{\text{embed}}\|$, $\text{var}(\Delta_{\text{embed}})$, $\text{var}(\Delta_{\text{form}})$). This therefore suggests that

Table 4:
Accuracy in
reconstructing
Unimorph's
inflection-
derivation
distinction by
various
supervised
classifiers

| Features | Logistic | MLP |
|-------------------------------------|-------------|-------------|
| Majority class (Inflection) | 0.57 | – |
| $\ \Delta_{\text{embed}}\ $ | 0.67 | 0.68 |
| $\ \Delta_{\text{form}}\ $ | 0.59 | 0.60 |
| $\text{var}(\Delta_{\text{embed}})$ | 0.76 | 0.76 |
| $\text{var}(\Delta_{\text{form}})$ | 0.71 | 0.71 |
| Form/embed magnitude* | 0.66 | 0.67 |
| Form/embed variability* | 0.84 | 0.84 |
| Form magnitude + variability* | 0.70 | 0.75 |
| Embed magnitude + variability* | 0.77 | 0.77 |
| All measures* | 0.86 | 0.90 |

while the variability features are the most descriptive of Unimorph's categorisation of inflection/derivation, all 4 features contain unique information relevant to recreating this distinction.

7

CLASSIFICATION OF LINGUISTIC CATEGORIES OF INFLECTION

Given the substantial amounts of controversy over what should be considered inflection and derivation, a model which largely aligns with a typical operationalisation of the distinction (Unimorph 4.0) may also be of interest in the ways in which it *differs*. Accordingly, in this section, we look at the trends in our model's classification of constructions which are labelled as inflection in Unimorph. We consider a number of inflectional categories which we believe to be of linguistic interest. In particular, we consider the type of meanings expressed by an inflection, whether an inflection changes part of speech, and whether that inflection is *contextual inflection* or *inherent inflection* (as described by Booij (1996)). For each of these categories, we consider whether constructions belonging to these categories are more likely than the average construction to be classified as derivational under our best model (the MLP with all 4 measures).

Our focus on the classification of inflections is due to the structure of our data. While derivations are not categorised in any cross-linguistically consistent ways (other than by their start and end parts

of speech in Unimorph) we can use the featural annotations on *inflections* in Unimorph as the basis of our categorisation scheme.

Categories of inflectional meaning

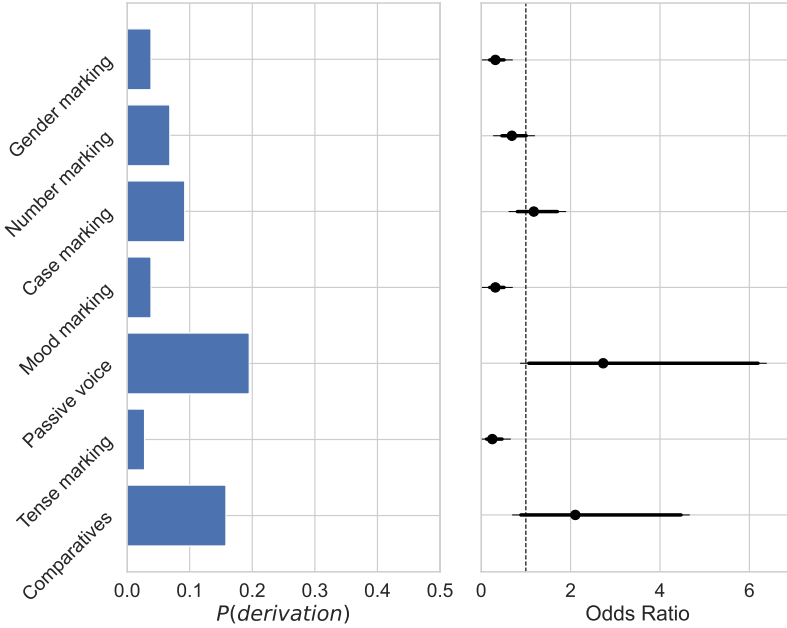
7.1

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 (forthcoming)'s retro-definition of inflection. Not all sources agree on all of these categories – Haspelmath rejects voice as inflectional, and comparison is often undiscussed as a major cross-linguistic inflectional category (as is the case in both Anderson 1985 and even Haspelmath forthcoming), and is considered *inherent inflection* (which is less canonical) by Booij (1996). One might reasonably expect constructions which mark for these controversial categories to be *more likely to be classified as derivation* by our model.

Consider the construction labelled ``V;PST;PL" in Ukrainian. Since no features belonging to the categories of, for example, voice and mood are in the featural description of the construction, we consider it *unmarked* for these categories. However, some categories such as number are marked obligatorily in some or all languages in Unimorph; it may not be desirable to, for example, consider the singular form of a noun as ``marked" for number (at least in a semantic sense). For such obligatory categories of inflectional meaning, asking if *all* values of the feature are classified more or less frequently as derivation is simply asking about that whole *part of speech*. To avoid this, we select sensible values of these features to consider unmarked even when in the featural description of a construction. Specifically, for the purposes of this analysis, we consider active voice, singular number, nominative case,⁷ and present tense unmarked values, even when

⁷ 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
categories of
inflectional
meaning.
Inflections to the
right of the
dotted line were
disproportion-
ately likely to be
classified as
derivation by our
model



present in the featural description of a construction. Returning to our ``V;PST;PL" example, this is then marked for tense, while a construction like ``V;PRES;PL" would be unmarked for tense in our analysis. For the category of gender, we simply consider nouns unmarked.

Figure 3 displays the probability that a construction marking for one of these categories 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 categories of inflection as *more derivational than inflectional*; each category is substantially more likely to be classified as inflection than derivation. This finding is perhaps unsurprising given our model's cross-linguistic 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 category 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 below 1 are *canonical* with respect to our model.

We apply the Boschloo exact test (Boschloo 1970) to the results and correct for multiple comparisons with the Bonferroni correction, which yields a significance level of $0.05/7 = 0.007$. We find the odds ratios for gender ($p = 1 \times 10^{-7}$), tense ($p = 3 \times 10^{-7}$), and mood ($p = 1 \times 10^{-7}$) significant. This identifies gender, mood, and tense as particularly canonical inflectional categories 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 categories of inflection, with comparatives even occasionally being listed as derivations within Unimorph⁸. 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.

Inherent vs. contextual inflection and transpositions

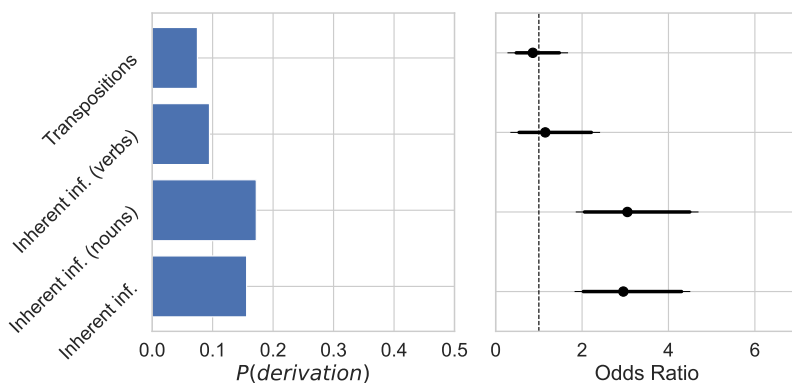
7.2

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 inflection which is

⁸For 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



determined by the syntactic context in which a word appears, while inherent inflection contributes to the meaning of a word itself. Booij identifies contextual inflection as canonical, with inherent inflection being non-canonical. In order to capture this notion, we here make a distinction between constructions such as ``N;PL" and the aforementioned ``V;PST;PL". The plural noun is an instance of what Booij (1996) terms *inherent* inflection: it is not required by syntax. On the other hand, the plural verbal form is controlled by its subject – it is selected by syntax for the purposes of agreement. We here consider forms such as ``V;PST;PL" *unmarked* for number, but as marked as participants in *agreement*. To operationalise this in a simple, cross-linguistically consistent way, we associate number, gender, and case⁹ 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

⁹Booij (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.

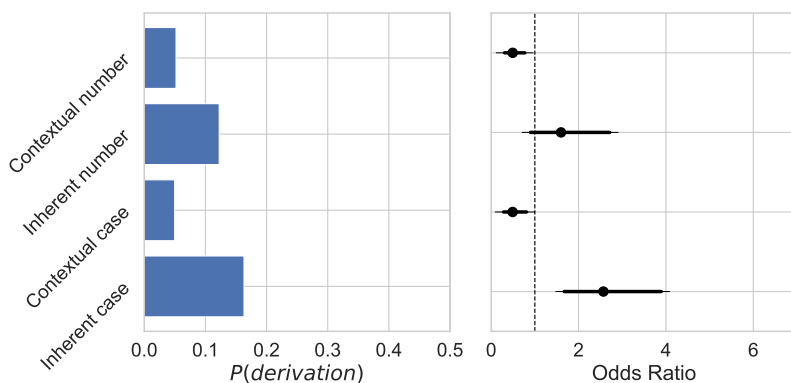


Figure 5:
Probability and
Odds ratio with
95% confidence
intervals of being
classified as
derivation for
inherent vs.
contextual noun
inflections

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 the category of inflectional transpositions, denoted in Unimorph as participles (deverbal adjectives), converbs (deverbal adverbs), and masdars (deverbal nouns), shown in Figure 4. This category has often been argued to be non-canonical inflection or even derivation because transpositions change the part of speech (Spencer 2013; Plank 1994; Haspelmath forthcoming). 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.

Summary

7.3

In this section, we have investigated which inflectional categories from theoretical linguistics are particularly *canonical* or *non-canonical*

under our model. While we find all such categories of inflection are more likely to be classified as inflection than derivation due to the high accuracy of our classifier, we instead investigate which categories are classified as *inflection* more often than the average inflectional construction, identifying these as *canonical categories* of inflection with respect to our model, and which are classified as *derivation* more frequently than average, identifying these as *non-canonical categories* of inflection. The *non-canonical categories* can be thought of as having less typical values of our measures, resulting in a lower chance to be classified as inflection.

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 categories rather than inherent verbal categories. 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 clearly point to variability of the change in distributional embedding $\text{var}(\Delta_{\text{embed}})$ 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.* (forthcoming) 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.* (forthcoming)'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.

Similarly, we find a clear and complementary influence of the variability of the change in form $\text{var}(\Delta_{\text{form}})$ – adding this feature to our model produces a large increase in performance, even when $\text{var}(\Delta_{\text{embed}})$ is already included. This measure can be thought of as describing the complexity of the structural relationship between base forms and constructed forms. Our results point to this relationship being much more complex for inflections than derivations across a wide range of languages, which is contrary to the predictions of Plank (1994), Dressler (1989), and others. While 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 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 $\text{var}(\Delta_{\text{form}})$. 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 $\text{var}(\Delta_{\text{form}})$ of 1.53 and *-keit* has $\text{var}(\Delta_{\text{form}})$ of 1.25. The two morphemes occur 27% and 73% of the time respectively. When combined, they have a $\text{var}(\Delta_{\text{form}})$ of 2.43 – still well within the derivational range.

Future studies should explore this aspect of our results 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 $(\|\Delta_{\text{form}}\|, \|\Delta_{\text{embed}}\|, \text{var}(\Delta_{\text{embed}}))$ achieves a classification accuracy of 82%, which is still 25 points above predicting the majority class.

On the other hand, our results show smaller influence of the magnitude measures $\|\Delta_{\text{form}}\|$, and $\|\Delta_{\text{embed}}\|$. This contrasts with Spencer's general notion that derivations are associated with larger changes to the properties of a lexeme. However, this is not entirely contradictory; $\|\Delta_{\text{embed}}\|$ still displays a fairly strong correlation with inflection and derivation on its own, and likely does not contribute as much to our models due to its substantial correlation (Pearson's r : 0.86) with the more strongly predictive $\text{var}(\Delta_{\text{embed}})$. In the case of $\|\Delta_{\text{form}}\|$, we find little evidence that derivations here 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.

8.2

The role of syntactic information

In this study, we use FastText embeddings as a proxy for both semantic and syntactic similarity. While the ability of such embedding vectors to capture human semantic similarity scores has been extensively studied (Vulić *et al.* 2020; Bojanowski *et al.* 2017), they are not usually utilised to capture syntactic similarity. Indeed, some studies have attempted to produce more syntactically-aligned embeddings from vectors like FastText (He *et al.* 2018), though replicating these techniques in a highly multilingual setting with low-resource languages is challenging. In this section, we analyse how much syntactic information FastText vectors are able to capture in our dataset, and how much more of Unimorph's inflection–derivation distinction we might be able to capture with a better representation of syntactic similarity.

To investigate the extent to which distances between FastText vectors encode syntactic information, we consider the mean cosine

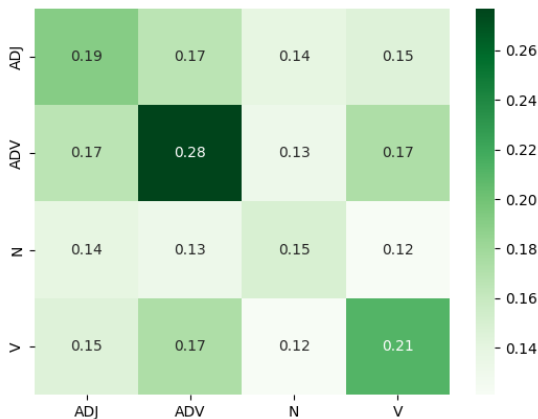


Figure 6:
The mean cosine similarity
between FastText
embeddings of words of the
same and different parts of
speech in Unimorph

similarity between embeddings of words in Unimorph that have different parts of speech (using the Unimorph part of speech annotations as shown in Table 1). We take a random sample of up to 5000 words of each part of speech for each language in our data. We then compute mean pairwise cosine similarity within and across these groups per language, and then weighted by number of words of the part of speech per language and averaged across languages. These results are presented in Figure 6. As can be seen in the figure, words with the same part of speech exhibit greater mean pairwise cosine similarity than pairs of words with different parts of speech, across all pairs of parts of speech. However, different parts of speech seem to be segregated to different degrees in vector space. On one extreme, we have adverbs where the mean cosine similarity observed between adverbs within a language was 64% greater than with any other part of speech. However, nouns are on average only 6.6% closer to each other than to the average word of their most similar part of speech (adjectives).

To more directly study the syntactic information captured by our embedding-based measures, we fit a logistic regression classifier which uses the two embedding measures ($\|\Delta_{\text{embed}}\|$, $\text{var}(\Delta_{\text{embed}})$) to classify whether a derivation changes part of speech – essentially using the difference between the base and derived forms in embedding space and the variability of its direction to determine whether the part of speech has been changed or not. We choose to use a logistic

regression classifier because our findings in Section 6 indicate that an MLP may not be necessary for these features, and it is less prone to spurious overfitting than an MLP. As before, we use 70% of the derivations as a training set, 10% as validation, and 20% as test. We find the classifier is able to predict whether a given construction changes the part of speech with 61% accuracy. Simply predicting the majority class (POS does not change) achieves a test-set accuracy of 53%, so this represents a 9-point improvement. Accordingly, we conclude that our embedding measures capture some information relevant to syntactic change.

To place an upper bound on how many of the model's errors can be explained by syntactic information, we consider how many errors can be explained by a syntactic change oracle variable. Using the annotations for part of speech in Unimorph, we produce a binary variable for whether a given construction changes the part of speech, using the start and end parts of speech for derivations. For inflections, we assume the part of speech does not change unless it is annotated by Unimorph as one of a participle, masdar, or converb. We add this oracle variable to the input to the classifier. We achieve a test-set accuracy of 84% with the logistic classifier and 92% with the MLP when combined with our four distributional measures. This represents a performance decrease of 2 points and increase of 2 points, respectively, suggesting little-to-no improvement to be found by a feature so closely aligned to linguistic notions of a change in part of speech.

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

syntactic information could represent individual pair variation within a construction (due to semantic drift, for example), which might be informative for reconstructing the distinction.

Language generality

8.3

A striking and important aspect of our model is its language-generality. A major limitation of existing computational studies of the inflection–derivation distinction (Copot *et al.* forthcoming; Rosa and Žabokrtský 2019; Bonami and Paperno 2018) is their focus on single European languages. In particular Haspelmath (forthcoming), argues that many properties of inflection and derivation are not proven to apply in a consistent way across languages (particularly non-European and non-Indo-European languages). Our model achieves high accuracy across languages with no language-specific features – in an entirely general way. 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 clearly suggests that in this family inflection and derivation are associated with distinct signatures in terms of their distribution, and their form (at least, as expressed in orthography). While evidence for such claims has been provided by Copot *et al.* (forthcoming), Bonami and Paperno (2018), and Rosa and Žabokrtský (2019) in specific languages, many large sub-families within the Indo-European language family had previously been completely untouched by this literature. In the case of the Germanic language family, we are able to look at several languages with distinctive morphological traits. We are also able to look at Armenian, Latvian, Irish, and Greek, covering many smaller European branches of the Indo-European family. We also expand the evidence for consistency in the application of the terms “inflection” and “derivation” within the Romance and Slavic language families. This 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 (forthcoming), 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 providing robust evidence for the ability of these properties to capture the application of the inflection–derivation distinction within Indo-European languages, we also provide initial evidence for the applicability of these measures to non-Indo-European languages. Our performance is lower on the whole than for the Indo-European languages, achieving only 82% accuracy compared to 91% accuracy for the Indo-European languages. While this is still well above simply predicting the majority class (which gets 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 low 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.

Another key differentiating aspect of our results from previous computational studies is our focus on classification of constructions. By doing so, we are able to quantify *how much* of the inflection–derivation distinctions, as operationalised across a wide range of languages, can be explained by our simple set of corpus-based correlates of inflection and derivation. This is to some degree *necessitated* by the large number of constructions and languages we consider, preventing us from discussing constructions purely at an individual level as in Bonami and Paperno (2018) or Copot *et al.* (forthcoming). We require quantification to make general statements, and classification provides a clear path to this.

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 looked at by Bonami and Paperno (2018) or Copot *et al.* (forthcoming) 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 (forthcoming). 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.¹⁰ 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.

8.5 *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.* (forthcoming) who find clear continuous gradience with respect to their metrics and the categories of inflection and derivation.

While, as discussed in 8.4, we focus primarily on the *consistency* of traditional categories of inflection and derivation, in this section we provide a brief investigation of whether, under our measures, the

¹⁰ As described in Section 8.5 and shown in Figure 7, it is this situation in which we find ourselves.

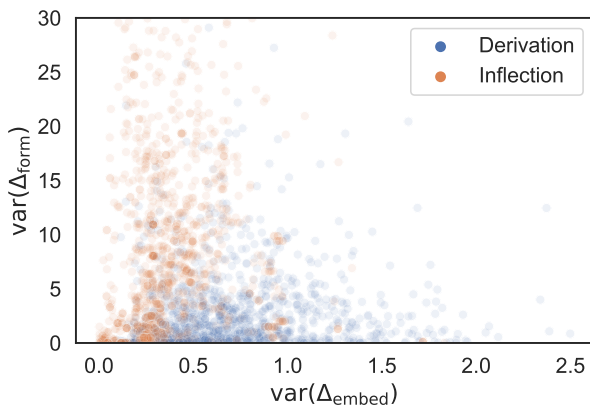


Figure 7:

Our two most predictive measures for inflection and derivation. With respect to these two variables, the inflection–derivation distinction appears gradient rather than categorical

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 7. Recall that a logistic classifier using only these features was able to correctly classify 84% 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.

8.6 *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 (forthcoming) is the right one, for instance, the answer to this question would superficially appear to be *no*. Haspelmath casts the distinction in terms of morpho-syntactic feature values, which themselves refer in many cases to the *meaning* expressed by a morphological exponent. If the specific meaning expressed by a morphological relation is necessary to distinguish which relations are inflectional in nature and which are derivational, then the typical inflection-derivation distinction requires *grounding* the meanings of sentences to solve – for example, no amount of raw language 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), 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 linguistic signal itself. 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 from the linguistic signal. 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.

Future work

8.7

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 very 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 8.3), the evidence presented here cannot yet fully refute Haspelmath (forthcoming)'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), truly general 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.

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. As shown in Table 2, we believe as many as 20 of Plank's subjective tests should have directly observable effects on the raw linguistic signal. Future works could test corpus-based measures of distance from the stem or limitedness of applicability, for example. Particularly interesting, we believe, would be the investigation of a syntactic distance and variability component, drawing on works such as He *et al.* (2018) and Ravfogel *et al.* (2020) – though there are significant challenges to operationalising these embeddings in a multilingual, low-resource domain.

There is also room for refinement of our measures. For example, extension to many other languages would likely require a reassessment 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.* (forthcoming)'s predictability of frequency measure, as both are motivated from semantic drift due to a change in lexical index. Similarly, future work could investigate the effects of using newer models of distributional semantics, such as XLM-R (Conneau *et al.* 2020), within our measurements – though they would have to overcome the difficulties of multilingual decontextualisation as described in Section 3.2.

CONCLUSION

9

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 86% and the MLP achieving a classification accuracy of 90%, improving by 29 and 33 points over predicting the majority class, respectively. We identify the variability of the change in distributional embedding space $\text{var}(\Delta_{\text{embed}})$ and the variability of the change of form $\text{var}(\Delta_{\text{form}})$ as particularly strong correlates of the distinction, together able to classify 84% of inflections in the same way in which they are classified in Unimorph.

Overall, these results show that much of the categories of inflection and derivation as used by linguistic resources 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 requires no language-specific features. These well-defined, empirical measures avoid the often-discussed subjectivity and vagueness of existing criteria (Haspelmath forthcoming; Plank 1994; Bybee 1985), while enabling us to produce the first large-scale quantification of how consistently traditional 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 that in our sample, this seems to be particularly due to *nominal* inherent inflections, like case and number. We find no traditional categories of inflectional meaning significantly non-canonical, providing some validation accounts of inflection which are structured around these categories like Haspelmath (forthcoming) 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 8.5). 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.

REFERENCES

- 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 KARAHÓGA, 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 MASHKOVTSOVA, 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>.

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 WENZKE, Francisco GUZMÁN, Edouard GRAVE, Myle OTT, Luke ZETTEMAYER, 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 (forthcoming), Idiosyncratic Frequency as a Measure of Derivation vs. Inflection, *Journal of Language Modelling*.

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

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

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>.

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 (forthcoming), Inflection and derivation as traditional comparative concepts, *Linguistics*.

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>.

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>.

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 LEVENSHTAIN (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.

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>.
- 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>.
- 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.
- 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 morfológico de la lengua Quechua basado en software libre Helsinki-finite-state-transducer (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.c1-4.5>.

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.

This work is licensed under the *Creative Commons Attribution 4.0 Public License*.

©  <http://creativecommons.org/licenses/by/4.0/>