

## **Abstract**

## **Keywords**

morphology, inflection, derivation, distributional semantics

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Related studies . . . . .	1
1.2	Distributional semantics . . . . .	3
1.3	UniMorph . . . . .	3
1.4	The morphology of aspect in Polish . . . . .	5
<b>2</b>	<b>Methodology</b>	<b>9</b>
2.1	Building the datasets . . . . .	9
2.2	Obtaining the mean similarity . . . . .	12
2.2.1	Implementing a random baseline . . . . .	12
<b>3</b>	<b>Results</b>	<b>14</b>
3.1	Clean data . . . . .	14
	<b>References</b>	<b>15</b>

# 1. Introduction

Usually, a distinction between inflection and derivation is drawn. In introductory morphology books they are presented as different processes (see for instance Aronoff and Fudeman (2011), Booij (2012), and Haspelmath and Sims (2013) or academic literature like Booij (2006) and Stump (2005, 2017)). These authors and many others provide many different criteria in order to draw a boundary between both morphological processes such as: “inflection is relevant to the syntax while derivation is not”, “inflection does not change the part of speech of the base while derivation may change it” or “derivation expresses a new meaning different from the base while inflection does not”. These criteria are just a small sample of all of the them that are usually provided.

However, there have been authors that have challenged this sharp boundary view directly posing that we should not make a distinction between them (Haspelmath, 2024) or that there is at best a continuum or gradient in which (canonical/prototypical) inflection and derivation stand on opposite sides (Bybee, 1985; Štekauer, 2015). Even Haspelmath and Sims (2013) argue that if all criteria are given the same importance then a continuum is the best explanation, since we cannot draw a sharp boundary between both processes. In fact, the same authors that provide long lists of criteria also acknowledge that exceptions exist and that some inflectional processes appear derivational and some derivational processes appear inflectional. Booij (2006) also draws a distinction between two different inflectional processes, inherent (not required by syntax, but a semantic choice like the use of a plural form or infinitives and participles) and contextual (required by syntax, such as verb-subject agreement or a case choice in nouns). The author argues that inherent inflection is halfway between derivation and contextual inflection, since inherent inflection also may change the part of speech of a word.

## 1.1 Related studies

Despite extensive literature on inflection and derivation, there have not been many empirical quantitative studies trying to shed light on this theoretical question until relatively recently. Thanks to computational advancements in distributional semantics (see subsection 1.2) some

studies have tried to separate both processes using one of the commonly proposed criteria, the semantic regularity criterion. This criterion states that “Inflection is semantically more regular than derivation.” (Stump, 2005). Essentially, it says that derived lexemes stray further from the meaning of the base than inflected ones, while in inflection the core meaning stays the same. To date and to my understanding, the studies that have examined this debate in a quantitative way are the following:

Bonami and Paperno (2018) use their own French embeddings model in order to assess if inflection is semantically more regular than derivation (semantic regularity criterion). Using a French lexicon, they construct a triplets dataset consisting of a pivot, an inflectional comparandum and a derivational comparandum based on different frequency measures. In their experiment, they measure shifts in meaning, for that they measure vector offset variance using Euclidean distance between the inflectional comparandum and the derivational comparandum. The authors clarify that using cosine similarity gives similar results on their data. They conclude that a categorical boundary between inflection and derivation cannot be found, although inflectional relations are more stable on average.

Rosa and Žabokrtský (2019) use a FastText Czech model (although they purposefully ignore words that do not appear in the model) to automatically separate inflection and derivation using a different criterion, the lexical meaning change criterion. They aim to determine if two morphologically related words belong to the same inflectional paradigm or are linked by derivation. Using a Czech database of word formation relations they measure string similarity (Jaro-Winkler edit distance) and cosine similarity. They find that inflectional forms are more similar to each other than derivational forms and that some derivational processes behave like inflection and vice versa, supporting the idea of a continuous scale and no strict boundary.

Haley et al. (2024) conducted the most complete of all three studies in quantity of measures (4) and languages (26). Like the previous study, they also used FastText models for all of the languages. The authors computed four measures, two based on the orthographic form and two on the distributional characteristics, to predict whether a given construction is inflectional or derivational using UniMorph data and two types of machine learning models. Their results also indicate that there is no strict boundary between inflection and derivation, but

that they belong to a gradient.

## 1.2 Distributional semantics

In order to explore the semantic regularity criterion we need to be able to capture the meaning of the words and nowadays, in order to capture it, we make use of distributional semantics. They are based on the distributional hypothesis which essentially states that similar words appear in similar contexts (Boleda, 2020). Distributional semantics represent the meaning of words as vectors, that is points in a multidimensional space, and similar words will also have similar vector thus occupying a similar space. In the semantic space created by the distributional model we can see the relation between specific words looking at the geometric relations of their vectors, using either Euclidean distance (the length of the straight line between them) or, more commonly, cosine similarity (the cosine of the angle between the vectors) (Boleda, 2020; Chandrasekaran & Mago, 2021). The vectors in this context are also often called word embeddings. For the current study two different static word embedding models were used, namely Word2vec (Mikolov et al., 2013) and FastText (Bojanowski et al., 2017). The main difference between both is that Word2vec operates on word level and FastText on a subword level. This means FastText can compute word representations for words that were not in the training data (Bojanowski et al., 2017). For details on the specific models used in this study see section 2.

## 1.3 UniMorph

UniMorph (Universal Morphology) is a resource that provides annotated morphological inflection tables for hundreds of languages (Batsuren et al., 2022). It offers a standardized schema for encoding morphological features such as tense, mood, number, person, case, and gender. The project compiles and normalizes inflectional paradigms extracted primarily from Wiktionary, producing tabular datasets that list all known inflected forms of a lemma alongside their corresponding morphological features (Table 1). With the release of UniMorph 4.0 they included derivational information and an annotation schema to represent derivational processes (Table 2).

universalizar	universalizarás	V;IND;FUT;2;SG;INFM
desguanguañar	desguanguañamos	V;IND;PRS;1;PL
desencorvar	desencorvo	V;IND;PRS;1;SG
nutrir	nutrían	V;IND;PST;IPFV;3;PL
innovar	innovaré	V;IND;FUT;1;SG
zakazać	zakazał	V;PST;3;SG;MASC
ścignąć	ścigną	V;FUT;3;PL
zadzwońić	zadzwoniliście	V;PST;2;PL;MASC;HUM
umówić się	umówiło	V;PST;3;SG;NEUT
zaniechać	zaniechałyście	V;PST;2;PL

**Table 1.** Sample from the Spanish (top) and Polish (bottom) UniMorph inflectional dataset.

lluvia	lluvioso	N:ADJ	-oso
transformar	transformación	V:N	-ción
comparable	comparablemente	ADJ:ADV	-mente
pirita	calcopirita	N:N	calco-
desintoxicar	desintoxicante	V:ADJ	-ante
namierzyć	namierzenie	V:N	-enie
żywczanin	żywczańka	N:N	-ka
wykrwawić	wykrwawiać	V:V	-ać
hejnał	hejnałowy	N:ADJ	-owy
zadrzeć	zadarcie	V:N	-cie

**Table 2.** Sample from the Spanish (top) and Polish (bottom) UniMorph derivational dataset.

## 1.4 The morphology of aspect in Polish

Take for instance the verb *pisać* (write.IPFV). In order to fully express what you would using one verb in English, you also need to know its aspectual pair *napisać* (write.PFV). The prefix *na-* is what is called an empty prefix, that is a prefix that changes the aspect of the verb and does not alter its meaning. However, the existence of empty prefixes and what aspectual pairs really are is a contentious topic that will be addressed later. In Polish, the perfective aspect refers to completed actions that result in changes in the state of affairs, while the imperfective refers to other types of actions: on-going, habitual, or complete but not causing a change in the state of affairs (Swan, 2002). The meaning behind the imperfective-perfective opposition also depends on the meaning of the verb itself, take for instance *stukać* (knock.IPFV) – *stuknąć* (knock.PFV). In the perfective aspect, this verb implies just one knock, but in the imperfective it implies there were many knocks. Compare it to *pisać* (write.IPFV) – *napisać* (write.PFV) where *napisać* does not mean ‘write just once’ but ‘write until the end, finish writing’. With other types of verbs such as *rozumieć* (understand.IPFV) – *zrozumieć* (understand.PFV), the perfective form does not imply an ending to the understanding, but rather a beginning (Nagórko, 2010). To form the perfective of an imperfective verb a prefix may be employed, like the previously mentioned *pisać* (write.IPFV) – *napisać* (write.PFV). It can also be done with the semelfactive suffix *-nąć* (*krzyczeć* (shout.IPFV) – *krzyknąć* (shout.PFV)). On the contrary, to form the imperfective from a perfective verb a suffix or morphophonological modification is employed (*kupić* (buy.PFV) – *kupować* (buy.IPFV); *ocenić* (mark.PFV) – *oceniać* (mark.IPFV)) (Bloch-Trojnar, 2015). Prefixation can also generate more verbs with a different lexical meaning, all in the perfective aspect (generally), see Table 3. A thing to note about perfective verbs is that they are defective, since they cannot be used to refer to the present only to the past and the future, while imperfective verbs have all three tenses. The imperfective future tense is formed using *być* (be.IPFV) as an auxiliary in the future tense, while in perfective verbs it is formed synthetically (Willim, 2006).

Many of the derived verbs relate to the meaning of *pisać* (write.IPFV) in that the actions are also done in writing, *dopisać* (add.PFV) is ‘adding while writing’, *przepisać* (copy.PFV) is ‘copying while writing’, while some others do not, in *opisać* (describe.PFV) the meaning is

Imperfective	Perfectives
<i>pisać</i> ‘write’	<i>dopisać</i> ‘add’ <i>popisać</i> ‘scribble’ <i>wpisać</i> ‘write in’ <i>podpisać</i> ‘sign’ <i>zapisać</i> ‘fill’ <i>prepisać</i> ‘copy’ <i>opisać</i> ‘describe’ <i>odpisać</i> ‘write back’

**Table 3.** Examples of derived verbs of the verb *pisać* (.IPFV).

not limited to a writing context. After seeing these examples, one might come to the conclusion right away that the prefixes contain predictable meaning. This is partly true since these prefixes are also prepositions, so the meaning can be predictable at times, although it is often complicated. Take for instance *za-*, which has 8 meanings assigned to it (Śmiech, 1986, as cited in Kita, 2017): *grać* (play.IPFV) – *zagrać* (play.PFV); *bić* (hit.IPFV) – *zabić* (kill.PFV); *brać* (take.IPFV) – *zabrać* (earn.PFV); *dać* (give.IPFV) – *zadać* (ask.PFV); *trzymać* (hold.IPFV) – *zatrzymać* (stop.PFV) (Perlin, 2005). Another thing that is not predictable is which prefix is the one that functions as an empty prefix, so one must always know it by heart.

Regarding now suffixation and morphophonological modification, these are often associated with imperfectivizing perfective verbs. This generalization does not hold when it comes to the previously mentioned semelfactive suffix *-ną*, which is used in the creation of perfective verbs, like it was pointed out in *krzyczeć* (shout.IPFV) – *krzyknąć* (shout.PFV) or *stukać* (knock.IPFV) – *stuknąć* (knock.PFV). Derived verbs, that is verbs with a prefix that have a different lexical meaning, shown in Table 3, can be further imperfectivized. These imperfective forms of derived perfective verbs are called secondary imperfectives and they can be formed either with the suffix *-yw-/-iw-*, alternating the thematic vowel or suffix, or with a stem alternation, like we can see in the following examples provided by Willim (2006):

Imperfective	Perfective	Secondary impf.
<i>kupić</i> ‘buy’	<i>przekupić</i> ‘bribe’	<i>przekupywać</i>
<i>bić</i> ‘hit’	<i>przebić</i> ‘pierce’	<i>przebijać</i>
<i>służyć</i> ‘serve’	<i>zasłużyć</i> ‘deserve’	<i>zasługiwać</i>
<i>brać</i> ‘take’	<i>obrać</i> ‘peel’	<i>obierać</i>

**Table 4.** Examples of secondary imperfectives.



Under this simple overview, it seems prefixed forms are formed by derivational processes and suffixed forms by inflectional processes. Some authors (Grzegorzczkowska et al. (1999), Nagórko (2010), Perlin (2005), and Włodarczyk and Włodarczyk (2006) among others) support this dual view of aspect, with some caveats. One issue this analysis has is aspectual pairs formed by prefixation (*pisać* (write.IPFV) – *napisać* (write.PFV)), since the only difference between both is the aspect, while there is no difference in lexical meaning. Due to this, some of them, like Nagórko (2010), say that the formation of aspectual pairs is an inflectional process, even those formed by prefixation, although others still place that specific type of aspectual pairs under derivation, like Grzegorzczkowska et al. (1999), who argue that the choice of an empty prefix is lexically motivated since you cannot predict which one a given verb will take, thus they list both aspect forms under different lexemes. They also note that some verbs have more than one aspectual pair with an empty prefix, like *brudzić* (soil.IPFV) – *zabrudzić* (soil.PFV) – *pobrudzić* (soil.PFV). Perlin (2005) adds some more arguments to this, such as the change in meaning when adding a prefix is irregular and that the addition of it does not always change the aspect of the verb:  *pływać* (swim/sail.IPFV) – *wypływać* (flow out/surface.IPFV), *chodzić* (go/walk.IPFV) – *dochodzić* (arrive.IPFV), *wieszać* (hang.IPFV) – *rozwieszać* (hang clothes.IPFV). This author also points out that sometimes some prefixes change the aspect while some others do not, on the same verb: *biegać* (run.IPFV) – *wybiegać* (run out.IPFV) but *pobiegać* (run for a while.PFV). In a similar way, some other authors, like Włodarczyk and Włodarczyk (2006) among others, reject the existence of empty prefixes since all verbal prefixes have some semantic weight. Given that the perfective may be expressed by many verbs derived from a single one (see all the derived perfectives in Table 3), there is no true aspectual pair formed by empty prefixes since this type of pairs have a slightly different meaning, they argue, although Grzegorzczkowska et al. (1999) do not go that far. Włodarczyk and Włodarczyk (2006) also consider only true aspectual pairs suffixed imperfective forms or morphophonologically altered forms, like *przepisywać* (copy.IPFV) from *przepisać* (copy.PFV) or *zamawiać* (order.IPFV) from *zamówić* (order.PFV). Contrary to this, Grzegorzczkowska et al. (1999) claim that imperfective forms of derived prefixed forms are not true aspectual pairs, since the meaning changes to iterative. Under their analysis, unprefixed pairs like *kupić* (buy.PFV) – *kupować* (buy.IPFV)

are true aspectual pairs formed by inflection and prefixed pairs like *przepisać* (copy.PFV) – *przepisywać* (copy.IPFV) are formed by derivation and are not true aspectual pairs. On the other hand, Perlin (2005) argues for an inflectional analysis of aspectual pairs, although he calls it the imperfective-perfective opposition and supports the existence of empty prefixes. The author lists six morphological processes involved in aspectual pairs: (1) suppletism, *brać* (take.IPFV) – *wziąć* (take.PFV), *widzieć* (see.IPFV) – *zobaczyć* (see.PFV); (2) prefixation, *myć* (wash.IPFV) – *umyć* (wash.PFV), *czytać* (read.IPFV) – *przeczytać* (read.PFV); (3) suffixation, *skakać* (jump.IPFV) – *skoczyć* (jump.PFV), *klaskać* (clap.IPFV) – *klasnąć* (clap.PFV); (4) prefixation and suppletism, *kłaść* (put.IPFV) – *położyć* (put.PFV); (5) infixation (or thematic alternation), *nazywać* (be called.IPFV) – *nazwać* (be called.PFV); and (6) infixation and suffixation, *wyżerać* (eat away.IPFV) – *wyżreć* (eat away.PFV). Authors also cite native speaker intuition that verbs come in pairs as an argument in favour of this inflectional analysis (Młynarczyk, 2004; Perlin, 2005).

A summary to this general, short overview is that there is no consensus on what aspectual pairs in Polish really are and what is the morphological process behind them. Even authors that agree on the process behind aspectual pairs, might disagree on what they are. Some authors maintain that aspectual pairs belong to the same lexeme, so there is no change of meaning, while others argue that there is a change of meaning, among other arguments. The two problematic ways of forming aspectual pairs are (1) adding an empty prefix (*pisać* (write.IPFV) – *napisać* (write.PFV), *czytać* (read.IPFV) – *przeczytać* (read.PFV)) and (2) suffixation or alternation of the thematic vowel/suffix (*podpisać* (sign.PFV) – *podpisywać* (sign.IPFV), *obrać* (peel.PFV) – *obierać* (peel.IPFV)). The very existence of empty prefixes is another topic of debate, those that do not support that analysis only view secondary imperfectives and their perfective counterparts as true aspectual pairs (2), while others that do support the existence of empty prefixes disagree completely with that view and argue that true aspectual pairs are those that follow the process in (2) but excluding secondary imperfectives (so only pairs like *kupić* (buy.PFV) – *kupować* (buy.IPFV)), and those that agree on the existence of empty prefixes might disagree on what is the morphological process that makes use of them to form aspectual pairs.

## 2. Methodology

This study explored the distinction between inflection and derivation in Polish and Spanish using two static word embedding models Word2vec and FastText. The Spanish Word2vec embeddings model was trained on the Spanish Billion Words<sup>1</sup> (SBW) corpus (Cardellino, 2016). For Polish, the IPIPAN Word2vec embeddings model *nkjp+wiki-forms-all-300-skipg-ns*<sup>2</sup> was used. This one was specifically chosen because it resembles the SBW model in vector dimension and training method. It was trained with the `gensim` library using the skip-gram algorithm with negative sampling on the National Corpus of Polish (NKJP<sup>3</sup>) and Wikipedia, includes all parts of speech and word forms, and produces 300-dimensional vectors. Regarding the FastText embeddings, the ones used are the newest version available in the FastText site<sup>4</sup> (Grave et al., 2018). The data that was used was taken from UniMorph. It was filtered and fixed and used without further changes (see subsection 2.1 for a detailed explanation).

To determine whether empty prefixes and suffixation in Polish behave more like inflection or derivation, we ran a test case. For this, we designed a dataset, consisting of three types of pairs: an imperfective verb and the perfective pair with an empty prefix, an imperfective verb and a derived (perfective) verb, and a derived prefixed perfective verb and the corresponding imperfective (so called secondary imperfective), as a sort of baseline since these verbs are a point of agreement among all authors.

### 2.1 Building the datasets

In order to conduct the analysis two separate datasets were built, one for inflection and another one for derivation, filtering the necessary data from UniMorph and cleaning it afterwards. For the inflection analysis, a pivot-inflection-category dataframe was constructed. The data was filtered to include only the following verb tenses in Spanish: Present Indicative (UniMorph tag: V;IND;PRS), Past Imperfect (UniMorph tag: V;IND;PST;IPFV) and Future Indicative (UniMorph tag: V;IND;FUT). A second filter was applied on the resulting data to remove *vos* and

---

<sup>1</sup><https://crscardellino.net/SBWCE/>

<sup>2</sup><https://dsmodels.nlp.ipipan.waw.pl/>

<sup>3</sup><https://nkjp.pl/>

<sup>4</sup><https://fasttext.cc/>

*usted* forms. *Voseo* forms do not have a high occurrence and *usted* forms are essentially duplicated forms since they are the same as 3rd person singular forms. In Polish the filtering included the same verb tenses: Present (UniMorph tag: V;PRS), Past (UniMorph tag:V;PST) and Future (UniMorph tag:V;FUT).

For the derivation analysis, the label U (unknown) in UniMorph was either fixed or removed in the final dataframe. In Spanish there are 20 rows that contain a derivation that results in U (i.e. X:U, where the pivot position is to the left of the colon and the derivation position is on the right) and 107 in Polish. On the other hand, there are even more derivations in which the pivot is tagged with U (U:X), 36 in Spanish and 253 in Polish. In order to clean the Spanish derivations dataset a new category according to the affix was assigned to the pivot position. All the pivots that end in *-ero*, *-ez*, *-ismo*, *-í* and *-illa* were changed from U (U:X) to N (N:X), since all these are noun affixes. The category was also changed to V:X in those that contain the affixes *-ar* and *-ear*. As a result 6 rows were left over and eliminated from the final dataframe because they simply contained mistakes. The rows which were tagged U in the derivation position (X:U) were dropped because they contained numerals and words that are not nouns, adjectives, adverbs or verbs. The latter were simply dropped, and since numerals could be tagged as either N or ADJ, they can also be dropped. In total, only 36 rows were discarded, which will not have major influence on the results.

Regarding the fixes in the Polish data, some rows incorrectly labelled U in the derivation position (X:U) were fixed by looking at the affixes. The rows containing affixes *-any*, *-ony*, *-ty*, *-y*, or *-qcy*, *-ęty* were changed to X:ADJ. These endings are taken by participles. Some formatting issues were also fixed, like rows in the pivot position which contained both the pivot and the derived form (i.e. *mylićpomylić pomylić*). Some rows that contained verbs in both columns but were not correctly labeled in the derivation form (V:U) were fixed as well. Three specific rows incorrectly labeled U:U were changed to ADJ:ADJ since they contained adjectives. The left over rows can finally be removed as well as they do not contain any nouns, adjectives, adverbs or verbs. Tackling the pivot position, there were many rows (163) that contained infinitives in both columns but were incorrectly labelled U:V. Since infinitives in Polish end in *-ć* or *-c* this was a quick fix. After this simple fix, there are few rows (21) with issues. Just like

in previous instances, these rows contain words that are not nouns, adjectives, adverbs or verbs and they were dropped. In the end, all four dataframes (inflection and derivation in Spanish and Polish) do not contain any row labeled with U anymore.

For the construction of the Polish aspectual dataset we wrote a Python script. The main resource was once again the Grammatical Dictionary of Polish (SGJP). We extracted the verbs from the most common 8000 lexemes, which consisted of 1819 verbs after removing biaspectual verbs, verbs that only exist in imperfective (*imperfectiva tantum*) and verbs that only exist in the perfective (*perfectiva tantum*). In order to group them by pairs, we filtered the imperfectives and scraped their perfective counterpart from the same site, obtaining 992 rows of imperfective-perfective pairs. We tagged them automatically using a Python script based on their prefix and their ending. The tags are either “empty” (for empty prefixed pairs) and “suffixed” for secondary imperfectives (AND ALSO VERBS THAT HAVE A ROOT CHANGE OR SOMETHING LIKE THAT, CHECK TI). Afterwards, the whole dataset was manually checked, in order to fix possible inaccuracies that the script might have introduced. After that, we automatically extracted all the derived verbs of the imperfectives, that appear in Wikisłownik<sup>5</sup> (Polish Wiktionary) under *wyrazy pokrewne* ‘related expressions’, tagging them “derived”. Once again, the data was manually checked, mainly because the Python script cannot account for all the formatting inconsistencies of the website. An example of the final dataset can be seen in Table 5.

Verb	Aspect	Pair	Pair aspect	Pair type
<i>pisać</i>	ndk	<i>napisać</i>	dk	empty
<i>pisać</i>	ndk	<i>podpisać</i>	dk	derived
<i>pisać</i>	ndk	<i>odpisać</i>	dk	derived
<i>pisać</i>	ndk	<i>wpisać</i>	dk	derived
<i>podpisywać</i>	ndk	<i>podpisać</i>	dk	suffixed

**Table 5.** Example rows of the polish aspectual dataset.

---

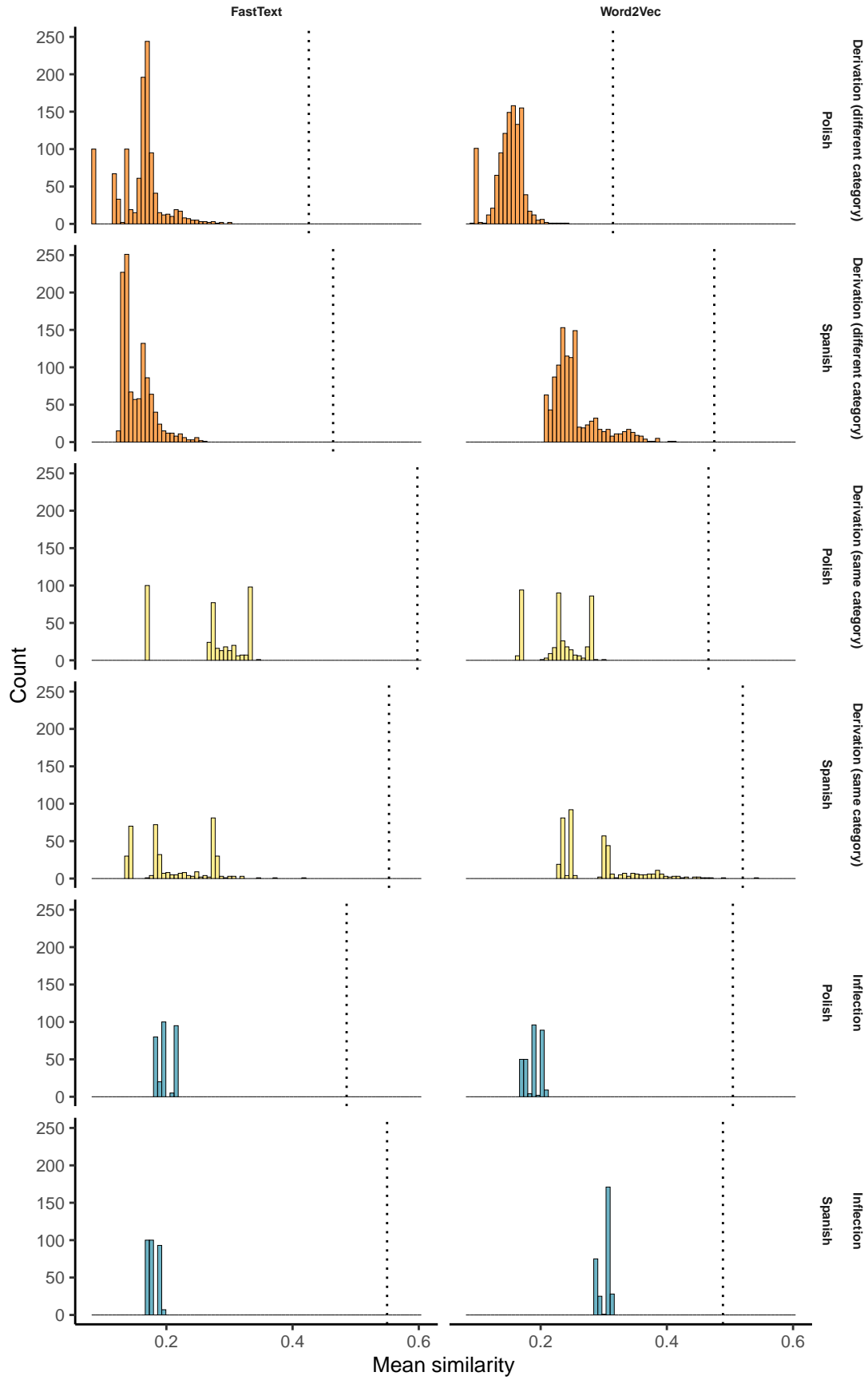
<sup>5</sup><https://pl.wiktionary.org/>

## **2.2 Obtaining the mean similarity**

In order to explore the distinction between inflection and derivation, a Python script was written and used on the four datasets (inflection and derivation in Spanish and Polish). In this script, the vector of the pivot and the vector of the second item (either the inflected or the derived form, depending on the dataset) were obtained row by row. Afterwards the cosine similarity was calculated using the `cosine` function of the `scipy` library and it was added to the dataframe. This data was exported afterwards for future processing in R. This automatic process was run using both word embedding models and both languages in inflection and derivation.

### **2.2.1 Implementing a random baseline**

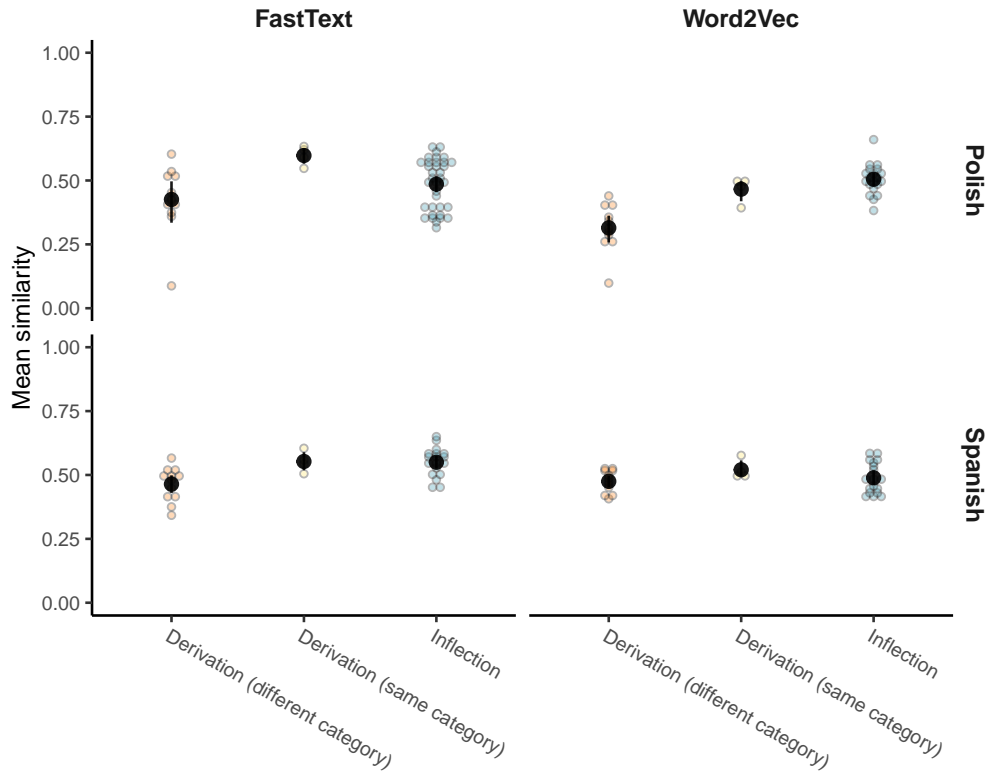
In order to evaluate the effectiveness of the methods and provide a point of comparison, a baseline was implemented. It filtered the data separately, inflections by tense (present, past or future) and derivations by category (N:ADJ, V:N...). After filtering, each individual part was shuffled 100 times while extracting the mean similarity of every row and the average of all rows. In other words, the mean similarity was extracted 100 times by type of morphological process (inflection or derivation), tense/category, model and language. The distribution is shown in Figure 1.



**Figure 1.** Distribution of the random baseline results. The dotted line represents the mean similarity value on the clean data.

### 3. Results

#### 3.1 Clean data



**Figure 2.** Mean similarity between pivot and form in inflection and derivation by model and language. Each dot represents a category. The black dot represents the mean similarity of all the categories and the standard deviation around it.

Inflection			Derivation		
Model	Language	Mean similarity	Model	Language	Mean similarity
FastText	Spanish	0.55	FastText	Spanish	0.51
	Polish	0.48		Polish	0.54
Word2Vec	Spanish	0.53	Word2Vec	Spanish	0.50
	Polish	0.51		Polish	0.40

**Table 6.** Mean similarity between pivot and form in inflection and derivation by model and language.



## References

- Aronoff, M., & Fudeman, K. A. (2011). *What is morphology?* Wiley-Blackwell. <https://thuvienso.hoasen.edu.vn/handle/123456789/8789>
- Batsuren, K., Goldman, O., Khalifa, S., Habash, N., Kieraś, W., Bella, G., Leonard, B., Nicolai, G., Gorman, K., Ate, Y. G., Ryskina, M., Mielke, S. J., Budianskaya, E., El-Khaissi, C., Pimentel, T., Gasser, M., Lane, W., Raj, M., Coler, M., ... Vylomova, E. (2022). *UniMorph 4.0: Universal Morphology* (3). <https://doi.org/10.48550/ARXIV.2205.03608>
- Bloch-Trojnar, M. (2015). Grammatical aspect and the lexical representation of Polish verbs. *Poznan Studies in Contemporary Linguistics*, 51(4), 487–510. <https://doi.org/10.1515/psicl-2015-0015>
- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017, June 19). *Enriching Word Vectors with Subword Information*. arXiv: 1607.04606 [cs]. <https://doi.org/10.48550/arXiv.1607.04606>
- Boleda, G. (2020). Distributional Semantics and Linguistic Theory. *Annual Review of Linguistics*, 6(1), 213–234. <https://doi.org/10.1146/annurev-linguistics-011619-030303>
- Bonami, O., & Paperno, D. (2018). Inflection vs. derivation in a distributional vector space. *Lingue e linguaggio*, (2/2018). <https://doi.org/10.1418/91864>
- Booij, G. (2006). Inflection and derivation. In *Encyclopedia of Language & Linguistics* (pp. 654–661, Vol. 5). Elsevier.
- Booij, G. (2012). *The Grammar of Words: An Introduction to Linguistic Morphology*. Oxford University Press.
- Bybee, J. L. (1985). *Morphology: A study of the relation between meaning and form*. John Benjamins.
- Cardellino, C. (2016). *Spanish Billion Word Corpus and Embeddings*. <https://crscardellino.github.io/SBWCE/>.
- Chandrasekaran, D., & Mago, V. (2021). Evolution of Semantic Similarity—A Survey. *ACM Comput. Surv.*, 54(2), 41:1–41:37. <https://doi.org/10.1145/3440755>

- Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018, March 28). *Learning Word Vectors for 157 Languages*. arXiv: 1802.06893 [cs]. <https://doi.org/10.48550/arXiv.1802.06893>
- Grzegorczykowa, R., Laskowski, R., & Wróbel, H. (Eds.). (1999). *Morfologia* (Wyd. 3 popr.). Wydaw. Naukowe PWN.
- Haley, C., Ponti, E. M., & Goldwater, S. (2024). Corpus-based measures discriminate inflection and derivation cross-linguistically. *Journal of Language Modelling*, 12(2), 477–529. <https://doi.org/10.15398/jlm.v12i2.351>
- Haspelmath, M. (2024). Inflection and derivation as traditional comparative concepts. *Linguistics*, 62(1), 43–77. <https://doi.org/10.1515/ling-2022-0086>
- Haspelmath, M., & Sims, A. (2013). *Understanding Morphology* (2nd ed.). Routledge. <https://doi.org/10.4324/9780203776506>
- Kita, M. (2017). *Wybieram gramatykę! dla cudzoziemców zaawansowanych na poziomie C i dla studentów kierunków filologicznych* (Wydanie 3., poprawione). Wydawnictwo Uniwersytetu Śląskiego.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013, September 7). *Efficient Estimation of Word Representations in Vector Space*. arXiv: 1301.3781 [cs]. <https://doi.org/10.48550/arXiv.1301.3781>
- Młynarczyk, A. K. (2004). *Aspectual pairing in Polish*. LOT.
- Nagórko, A. (2010). *Podręczna gramatyka języka polskiego*. Wydawnictwo Naukowe PWN.
- Perlin, J. (2005). Ile jest aspektów w języku polskim oraz próba dowodu na fleksyjność opozycji dokonany / niedokonany. *Biuletyn Polskiego Towarzystwa Językoznawczego*, 61(2005), 49–58.
- Rosa, R., & Žabokrtský, Z. (2019). Attempting to separate inflection and derivation using vector space representations. In M. Ševčíková, Z. Žabokrtský, E. Litta, & M. Passarotti (Eds.), *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. <https://aclanthology.org/W19-8508/>

- Štekauer, P. (2015). The delimitation of derivation and inflection. In P. O. Müller, I. Ohnheiser, S. Olsen, & F. Rainer (Eds.), *Word-Formation* (pp. 218–235). DE GRUYTER. <https://doi.org/10.1515/9783110246254-016>
- Stump, G. T. (2005). Word-Formation and Inflectional Morphology. In P. Štekauer & R. Lieber (Eds.), *Handbook of Word-Formation* (pp. 49–71). Springer Netherlands. [https://doi.org/10.1007/1-4020-3596-9\\_3](https://doi.org/10.1007/1-4020-3596-9_3)
- Stump, G. T. (2017). Inflection. In *The Handbook of Morphology* (pp. 11–43). John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781405166348.ch1>
- Swan, O. E. (2002). *A Grammar of Contemporary Polish*. Slavica Publishers.
- Willim, E. (2006). *Event, individuation and countability: A study with special reference to English and Polish*. Wydawnictwo Uniwersytetu Jagiellońskiego.
- Włodarczyk, H., & Włodarczyk, A. (2006). Semantic Structures of Aspect (A Cognitive Approach). In *Od fonemu do tekstu : Prace dedykowane profesorowi Romanowi Laskowskiemu* (pp. 389–408). Wydawnictwo Lexis.