Abstract

Keywords

morphology, inflection, derivation, distributional semantics, aspect in Polish

Contents

1. Introduction

Inflectional and derivational morphology are commonly described as different processes. Inflection is involved in the formation of grammatical words, also called word-forms, of a lexeme. A lexeme is an abstract word that acts as the representative of the meaning that all its word-forms relate to, so BÓBR in the case of Table 1 is the lexeme, and all the words that appear inside the different cells are its grammatical words. Verbs are also often used as examples of inflection, which change depending on the tense, aspect and mood. In Table 2 we can see the Spanish verb lexeme COMER conjugated in three different tenses of the indicative mood. Number, case, gender, person, tense, aspect and mood are generally expressed through inflectional processes (Aronoff & Fudeman, 2011; Haspelmath & Sims, 2013).

	Singular	**Plural**
Nominative	bóbr	bobry
Genitive	bobra	bobrów
Dative	bobrowi	bobrom
Accusative	bobra	bobry
Instrumental	bobrem	bobrami
Locative	bobrze	bobrach
Vocative	bobrze	bobry

Table 1. Declension of the Polish noun *bóbr* 'beaver'

	Preterite	**Present**	**Future**
1SG	comí	como	comeré
2SG	comiste	comes	comerás
3SG	comió	come	comerá
1PL	comimos	comemos	comeremos
2PL	comisteis	coméis	comeréis
3PL	comieron	comen	comerán

Table 2. Conjugation of the Spanish verb *comer* 'eat'.

In the case of derivational morphology, it is involved in the creation of new lexemes. For instance, suffixes such as *-dor* or *-ero* in Spanish generate nouns, e.g. *abridor* 'opener' from *abrir* 'open', *casero* 'landlord' from *casa* 'house'; or the suffix *-ość* in Polis, which also generates nouns, e.g. *czystość* 'cleanliness' from *czysty* 'clean'. When creating a new lexeme,

the category (whether it is a noun, verb, adjective or adverb) may change, like in the previous examples of the deverbal noun *abridor* 'opener' and deadjectival noun *czystość* 'cleanliness' (Booij, 2012; Haspelmath & Sims, 2013).

In order to draw a boundary between both morphological processes authors propose many different criteria. For instance, it is said that inflection is relevant to the syntax but derivation is not. This means that what determines which word-form of a lexeme is used in a given position in a sentence is the relationship to the words around it (Aronoff & Fudeman, 2011; Booij, 2012). Still, inflection sometimes does not seem as relevant. Tense and aspect are rarely assigned by the syntax, so, following this criterion, they could be excluded from inflection (Haspelmath & Sims, 2013). Another criterion, a relevant one regarding this thesis, is that in inflection the part of speech or category of the base does not change. Derivational processes on the other hand typically change the category, although not always. There exist both category-preserving and category-changing derivation, since a change in lexical meaning does not mean a change of category (Booij, 2006; Haspelmath & Sims, 2013; Stump, 2017) like it was shown in the examples before. There are also some cases in which inflection is involved in a change of category, like in gerunds (infinitives in the case of Romance languages) or participles, which are part of the verbal paradigm but can also function as nouns or adjectives (Booij, 2012; Cappellaro & Meinschaefer, 2022; Haspelmath, 2024; Štekauer, 2015). Stump (2005) points out that, in many languages, participles never appear without the declension morphology of adjectives, e.g. Polish. The most relevant criterion in this thesis is the semantic regularity criterion. It states that derivation expresses a new meaning but inflection does not change the lexical meaning (Haspelmath & Sims, 2013) or, as put by Stump (2017), inflectional processes are semantically more regular than derivational processes. Lexemes formed by derivation are often hard to predict semantically, e.g. words ending in *-ize* in English such as *dollarize*, *hospitalize* or *vaporize* (Haspelmath, 2024; Stump, 2005, 2017). Sometimes, it is true that a related noun and adjective like *czystość* 'cleanliness' and *czysty* 'clean' are also clearly related in meaning, but this relation is not like that of a verb and a conjugated form, like *comer* 'eat' and *comiste* 'you ate' (Aronoff & Fudeman, 2011). In fact, Haspelmath and Sims (2013) argue examples like the first one fall in between inflection and derivation.

Like we have seen, the distinction between these two processes is not without debate. For this reason, 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 inflection and derivation stand on opposite sides (Bybee, 1985; Štekauer, 2015). Haspelmath and Sims (2013) argue that if all criteria are given the same importance then a continuum is the best explanation, since a sharp boundary between both processes cannot be drawn.

In this thesis, in order to shed light on the inflection-derivation debate we explore the semantic regularity criterion using distributional semantics, a computational method to represent the meaning of words, based on the distributional hypothesis. This hypothesis essentially states that similar words appear in similar contexts, hence its distribution reflects its meaning (Boleda, 2020). Distributional semantics represents the meaning of words as vectors, that is points in a multidimensional space, so words with similar vectors, i.e. words that occupy a similar region in that space, will also be semantically similar. In the semantic space created by a distributional semantics 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). The higher the cosine similarity, the closer the vectors are to each other, meaning a higher semantic similarity (Boleda, 2020; Chandrasekaran & Mago, 2021). If we assume that derived lexemes stray further from the meaning of the base than inflected ones, while in inflection the core meaning stays the same, then distributional semantics is a great way to explore this criterion. Consequently, inflectional processes should maintain more similarity in meaning than derivational processes in a vector space, i.e. the relation between words related by an inflectional process should be closer than the relation between words related by a derivational process.

There are three recent studies that have advanced our understanding of how distributional semantics can contribute to the inflection-derivation debate. Bonami and Paperno (2018) conduct a distributional study in order to assess if inflection is semantically more regular than derivation. They conclude that a categorical boundary between inflection and derivation cannot

be found, although inflectional relations are more semantically stable on average. For their study they use a French lexicon and their own French embeddings model, a continuous bag of words (CBOW) model with negative sampling, subsampling, a window size of 5 and a vector size of 400, 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, measuring 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. Rosa and Žabokrtský (2019) use a FastText Czech model (although they puposefully 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 broadly find that inflectional forms are more similar to each other than derivational forms. Taking into account only the measure of cosine similarity, they find it separates inflection and derivation adequately, although negation (using the prefix *-ne*) and grade inflection possess a high cosine similarity, closer to derivation, than widely accepted inflectional processes like case marking or number. Combining both measures, derivation continues to be separated from inflection, although negation shows lower similarity than derivation. Exploring derivations in detail, they find high cosine similarity in aspectual morphology regarding the change of a perfective verb to its imperfective counterpart, a traditionally considered (in Czech) category-preserving derivational process, i.e. a process that does not change the category or part of speech. In some category-changing derivation, like between N and V they find lower cosine similarity but higher in ADJ to ADV or N to POSS ADJ. Finally, Haley et al. (2024) conducted a study trying to classify constructions into inflection or derivation in 26 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. Like the previous study, they used FastText models for all of the languages. What they find is that their best model can cor-

rectly predict the morphological process most of the time (89% +-1) but many constructions lie between both categories. Their results indicate that there is no strict boundary between inflection and derivation, but that they belong to a gradient

One factor these studies do not explore in detail is the change of category in derivation, it is only shortly mentioned in Rosa and Žabokrtský (2019). As it has been pointed out before, derivation may or may not change the category of the lexeme it creates. This is a key element that it is worth exploring, specially due to the fact that it is one of the mainly referenced criteria to distinguish between inflection and derivation. For this reason, this thesis aims to explore the difference between category-preserving and category-changing derivational processes using distributional semantics models.

Additionally, we are interested in exploring a linguistic question between inflectional processes and category-preserving derivational processes, thus the consideration to examine this in more detail. Similarly as it was pointed out in Rosa and Žabokrtský (2019) regarding Czech, some processes involving aspect in Polish are not agreed upon. Specifically, a way of forming perfective verbs from imperfectives and a way of forming imperfective verbs from derived perfective verbs. Verbs in Polish have traditionally been described as having a counterpart in the opposite aspect, i.e. an imperfective verb such as *pisać* (write.IPFV) has a perfective pair, in this case *napisać* (write.PFV). These perfective pairs are formed by so-called empty prefixes, prefixes that only change the aspect of the verb and do not add any meaning, hence this process is usually considered inflectional. In some cases, aspectual pairs are not created with an empty prefix, but through suffixation from a perfective verb, like in *kupić* (buy.PFV) – *kupować* (buy.IPFV). Deriving new verbs from imperfective ones is done through prefixation and the resulting ones are generally perfective, e.g. *podpisać* (sign.PFV) derived from the previous verb *pisać* (write.IPFV). Since the resulting derived verb is perfective, it needs an imperfective counterpart, in which case it is commonly obtained through suffixation, e.g. *podpisywać* (sign.IPFV). These imperfective verbs are called secondary imperfectives, and this process is also generally considered an inflectional one. In a nutshell, the points of contention are many. Some authors consider that empty prefixes do add some change in meaning, therefore this process should be considered derivational, while others maintain there is no change.

The ones that reject empty prefixes only consider suffixed imperfective forms as true aspectual pairs, but others argue that there is also a change in meaning in secondary imperfectivization, excluding them from the true aspectual pairs. In this case the only true aspectual pairs would be non-prefixed perfective verbs and their suffixed imperfective counterpart, i.e *kupić* (buy.PFV) – *kupować* (buy.IPFV). This is described and investigated in more detail in the second study of this thesis (see subsection 3.1).

In this fashion, the objectives of this thesis are two-fold: exploring the inflection-derivation debate emphasizing category-preserving and caregory-changing derivation by means of distributional semantics, and, using the same methods, trying to shed light on a theoretical debate regarding aspectual morphology in Polish.

2. First study

2.1 Methodology

This study explores the distinction between inflection and derivation in Polish and Spanish using two static word embedding models: Word2Vec (Mikolov et al., 2013) and FastText (Bojanowski et al., 2017). The main difference between both is that Word2Vec operates on a word level and FastText on a subword level. This means a Word2Vec model has vector representations for whole words but FastText for subwords. In case the models encounter words that were missing from its training data, only a FastText model can compute word representations for such words (Bojanowski et al., 2017). Specifically, the chosen Spanish Word2Vec embeddings model was trained on the Spanish Billion Words (SBW)[1] corpus (Cardellino, 2016). For Polish, the IPIPAN Word2Vec embeddings model *nkjp+wiki-forms-all-300-skipg-ns*[2] 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)[3] and Wikipedia, includes all parts of speech and word forms, and produces 300-dimensional vectors. Regarding the FastText em-

[1] https://crscardellino.net/SBWCE/
[2] https://dsmodels.nlp.ipipan.waw.pl/
[3] https://nkjp.pl/

beddings, the ones used are the newest version available in the FastText[4] site for both languages (Grave et al., 2018).

The data that was used was taken from UniMorph (Universal Morphology), 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 3). With the release of UniMorph 4.0 they included derivational information and an annotation schema to represent derivational processes (Table 4).

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
zadzwonić	zadzwoniliście	V;PST;2;PL;MASC;HUM
umówić się	umówiło	V;PST;3;SG;NEUT
zaniechać	zaniechałyście	V;PST;2;PL

Table 3. 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	żywczanka	N:N	-ka
wykrwawić	wykrwawiać	V:V	-ać
hejnał	hejnałowy	N:ADJ	-owy
zadrzeć	zadarcie	V:N	-cie

Table 4. Sample from the Spanish (top) and Polish (bottom) UniMorph derivational dataset.

In order to conduct the analysis two separate datasets were built for each language, one for inflection and another one for derivation, filtering the necessary data from UniMorph and cleaning it afterwards. For the inflectional morphology analysis, a lemma-inflection-category dataframe was constructed. The data was filtered to include the following verb tenses

[4]https://fasttext.cc/

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 *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 derivational morphology analysis, the label U (unknown) in UniMorph was either fixed or removed in the final dataframe. In the Spanish data, there were 20 rows that contained a derivational process that results in U (i.e. X:U, where the lemma is to the left of the colon and the derived word is on the right) and 107 in Polish. On the other hand, there were even more rows in which the lemma was tagged 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 of the row was assigned to the lemma position. All the lemmas 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 could 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 derivational data, some rows incorrectly labelled U in the derivation position (X:U) were fixed by looking at the affix column. The rows containing affixes *-any*, *-ony*, *-ty*, *-y*, or *-ący*, *-ęty* were changed to X:ADJ as these are endings taken by participles. Some formatting issues were also fixed, like rows in the lemma position which contained both the lemma and the derived form (e.g. *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 could be removed as well as they did not contain any nouns, adjectives, adverbs or verbs. Tackling the lemma position, there were many rows

(163) that contained infinitives in both columns but were incorrectly labelled U:V instead of V:V. Since infinitives in Polish end in *-ć* or *-c* this was a straightforward fix. After that, there were few rows (21) with issues. Just like in previous instances, these rows contained words that are not nouns, adjectives, adverbs or verbs and they were dropped as a result. In the end, all four dataframes (inflection and derivation in Spanish and Polish) do not contain any row labeled with U anymore.

After the data processing was done, a Python script was written and used on the four dataframes (inflection and derivation in Spanish and Polish). In this script, the vector of the lemma and the vector of the second item (either the inflected form or the derived word) were obtained row by row. Afterwards, the cosine similarity of the two vectors was calculated using the `cosine` function of the `scipy` library and it was added to the dataframe. This data was exported afterwards for future an analysis in R (R Core Team, 2024). This automatic process was run using both word embedding models and both languages in inflection and derivation.

2.2 Results

We fit two Bayesian Beta regression hierarchical models to assess differences in cosine similarity across three types of affixation: category-changing derivation, category-preserving derivation, and inflection. In both models, affixation type was included as the only fixed effect. As random effects, we included intercepts for model (word2vec and FastText), language (Polish and Spanish), and their interaction. We set the same Student-t prior (df = 5 , μ = 0 , σ = 3) on all fixed effects, including the intercept. For the random effects, we used a half Student-t prior (df = 3 , μ = 0 , σ = 2.5). The precision parameter φ was assigned a weakly informative Gamma prior (α = 0.01 , β = 0.01).

In the first model, the fixed effect is Helmert contrast-coded. We compare the mean cosine similarity of category-preserving derivation (type1) to category-changing derivation, and the mean cosine similarity of inflection (type2) to the average of the previous two. In the second model, the fixed effect is treatment-coded, with the reference level set to category-preserving derivation. We thus compare category-changing derivation and inflection separately to category-preserving derivation.

The Beta model with Helmert coding reveals a strong effect of affixation type on cosine similarity. Specifically, category-preserving derivation shows higher similarity than category-changing derivation (estimate = 0.235, 90% CI [0.136, 0.335], posterior probability > 0 = 1.000). There is also moderate evidence that inflectional forms are more similar than the average of the two derivational types (estimate = 0.041, 90% CI [0.000, 0.083], posterior probability > 0 = 0.949).

In the Beta model with treatment coding (reference level: category-preserving derivation), cosine similarity varied across affixation types. Category-changing derivation showed substantially lower similarity than category-preserving derivation (estimate = -.471, 90% CI [-0.672, -0.269], posterior probability < 0 = 1.000). In contrast, there was little evidence for a difference between inflection and category-preserving derivation (estimate = -0.113, 90% CI [-0.298, 0.074], posterior probability < 0 = 0.838). The intercept, representing the mean similarity for category-preserving derivation, was estimated at 0.157 (90% CI [-0.934, 1.298]), with low certainty that it differed from zero (posterior probability > 0 = 0.639).

Inflection

Model	Language	Mean similarity
FastText	Spanish	0.55
	Polish	0.48
Word2Vec	Spanish	0.53
	Polish	0.51

Derivation

Model	Language	Mean similarity
FastText	Spanish	0.51
	Polish	0.54
Word2Vec	Spanish	0.50
	Polish	0.40

Table 5. Mean similarity between lemma and inflectional or derivational element by model and language.

In order to evaluate the effectiveness of the methods and provide a point of comparison, the mean similarity of the target pairs in the data was compared to a random baseline. This baseline 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 2.

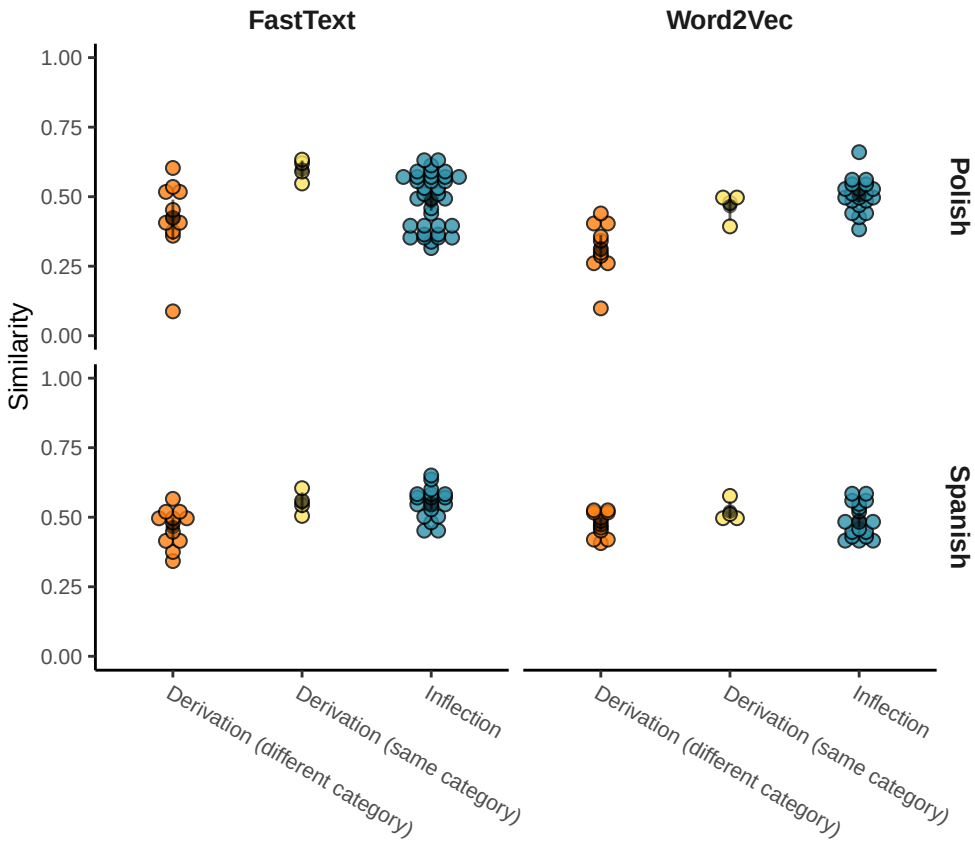

Figure 1. Similarity between lemma and inflectional or derivational element by model and language. Each coloured dot represents a category in derivation and a different conjugation form (number, person and tense) in inflection. The black dot represents the mean similarity of the whole morphological process, with the standard deviation around it.

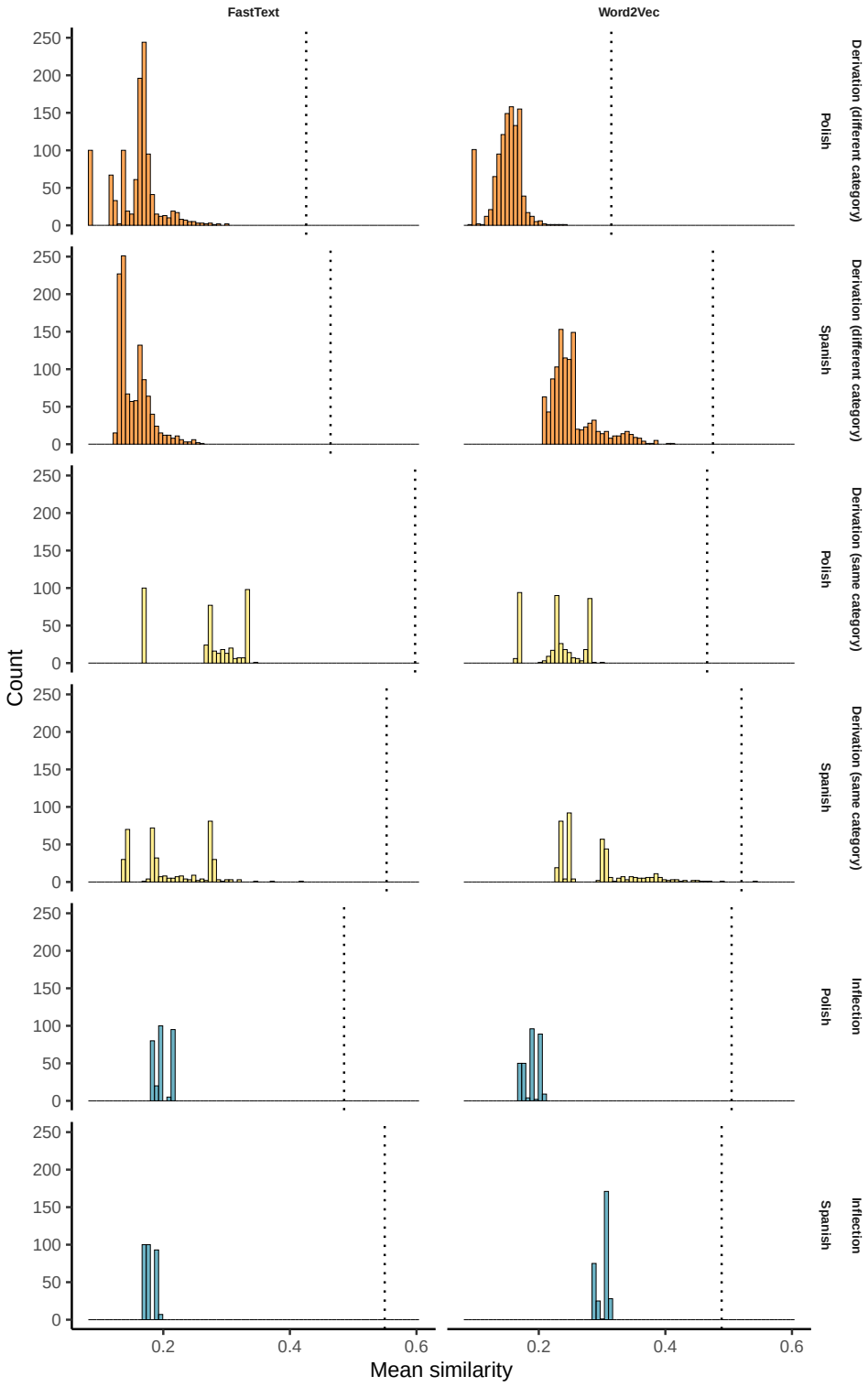

Figure 2. Distribution of the random baseline. Each datapoint represents an instance of the mean similarity of the type of morphological process, tense/category, model and language. The dotted line represents the mean similarity value by category/tense on the data.

2.3 Discussion

We find evidence for a difference between inflection and derivation when category-preserving and category-changing derivation are not distinguished, and inflection is compared to their average (Helmert-coded model). However, when using treatment coding and directly comparing inflection to category-preserving derivation, no strong difference emerges. Across both models, there is strong evidence that category-preserving derivation differs from category-changing derivation, with the latter showing substantially lower cosine similarity.

3. Second study

3.1 The morphology of aspect in Polish

The traditional view in Polish is that verbs come in pairs, one being imperfective and the other being perfective, which are generally distinguished by prefixes or suffixes (Nagórko, 2010). 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 6. 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).

Imperfective	**Perfectives**
pisać 'write'	*dopisać* 'add'
	popisać 'scribble'
	wpisać 'write in'
	podpisać 'sign'
	zapisać 'fill'
	prepisać 'copy'
	opisać 'describe'
	odpisać 'write back'

Table 6. Examples of derived verbs of the verb *pisać* (write.IPFV).

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 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 *-nąć*, 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, i.e. prefixed verbs that have a different lexical meaning, shown in Table 6, 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. We can see examples of some secondary imperfectives provided by Willim (2006) in Table 7.

Under this simple overview, it seems prefixed forms are formed by derivational pro-

Imperfective	Derived perfective	Secondary impf.
kupić 'buy'	*przekupić* 'bribe'	*przekupywać*
bić 'hit'	*przebić* 'pierce'	*przebijać*
służyć 'serve'	*zasłużyć* 'deserve'	*zasługiwać*
brać 'take'	*obrać* 'peel'	*obierać*

Table 7. Examples of secondary imperfectives.

cesses and suffixed forms by inflectional processes. Some authors (Grzegorczykowa 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 Grzegorczykowa 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 6), there is no true aspectual pair formed by empty prefixes since this type of pairs have a slightly different meaning, they argue, although Grzegorczykowa 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, Grzegorczykowa 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 supletism, *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 (Mlynarczyk, 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.

3.2 Methodology

This study explores whether empty prefixes and suffixation in Polish behave more like inflection or derivation. For this, a dataset consisting of three types of pairs was designed: (1) an imperfective verb and the perfective pair with an empty prefix, (2) an imperfective verb and a derived prefixed perfective verb and (3) a derived prefixed perfective verb, and a suffixed form or forms with stem/vowel alternation (so called secondary imperfectives). The derived verbs are a point of consensus among all the authors in that they involve a derivational process, but the other two types of pairs are not.

The main resource was the Grammatical Dictionary of Polish (SGJP)[5]. This dictionary does not contain definitions but detailed grammatical information of more than 450,000 entries of nouns, verbs, adjectives, prefixes, numerals and so on. The information is presented mainly as inflectional tables, as well as other grammatical information relevant for every lexeme. Other than that, this dictionary has a powerful search engine with a great variety of filters. Thanks to this, the most frequent lexemes, among other things, can be easily accessed. All the verbs from the most common 8,000 lexemes were extracted, which consisted of 1,819 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, the imperfectives were filtered and their perfective counterpart was extracted from the same site, obtaining 987 rows of imperfective-perfective pairs. They were tagged automatically, based on their prefix and their ending, as empty prefixed pairs or secondary imperfectives and verbs that contain a stem/vowel alternation. Afterwards, the whole dataset was manually checked, in order to fix possible inaccuracies that the script might have introduced. All the derived verbs of the imperfective ones were then extracted, and tagged, using Wikisłownik[6] (Polish Wiktionary), which are stored under *wyrazy pokrewne* 'related expressions'. Once again, the

[5]http://sgjp.pl/
[6]https://pl.wiktionary.org/

data was manually checked, mainly because the script cannot account for all the formatting inconsistencies of the website. An example of the final dataset can be seen in Table 8. In the end, the whole dataset consisted of 274 rows of derivational relations, 228 rows of empty prefixation and 485 rows of suffixation/stem alternation.

Verb	**Aspect**	**Pair**	**Pair aspect**	**Pair type**
pisać	IPFV	*napisać*	PFV	empty
pisać	IPFV	*podpisać*	PFV	derived
pisać	IPFV	*odpisać*	PFV	derived
pisać	IPFV	*wpisać*	PFV	derived
podpisywać	IPFV	*podpisać*	PFV	suffixed
oceniać	IPFV	*ocenić*	PFV	suffixed
odmawiać	IPFV	*odmówić*	PFV	suffixed

Table 8. Example rows of the Polish aspectual dataset.

3.3 Results

We fit two Bayesian Beta regression hierarchical models to assess differences in cosine similarity across three types of affixation in Polish aspect morphology: inflectional suffixes, empty prefixes, and derivational prefixes. In both models, affixation type was included as the only fixed effect. As random effects, we included intercepts for model (word2vec and FastText). We set the same priors as those in the models fitted in Study 1. In the first model, the fixed effect is Helmert contrast-coded. We compare the mean cosine similarity of empty prefixes (`pair_type1`) to inflecitonal suffixes, and the mean cosine similarity of derivational prefixes (`pair_type2`) to the average of the previous two. In the second model, the fixed effect is treatment-coded, with the reference level set to empty prefixes. We thus compare separately derivational prefixes and inflectional suffixes to empty prefixes.

In the Beta regression model with Helmert coding (where the first contrast compares empty prefixes and suffixes, and the second compares derivational prefixes to the average of the previous two) cosine similarity varied across affixation types. There is strong evidence that derivational prefixes have lower similarity compared to the average of empty prefixes, and suffixes (estimate = -0.247, 90% CI [-0.261, -0.233], posterior probability < 0 = 1.000). In contrast, the difference between empty prefixes and suffixes forms is negligible (estimate =

0.008, 90% CI [-0.017, 0.033], posterior probability > 0 = 0.715).

The results from the Beta regression model with treatment coding (reference level: empty prefixes) are consistent with those of the previous model. Derived forms show substantially lower similarity compared to forms with empty prefixes (estimate = -0.750, 90% CI [-0.803, -0.697], posterior probability < 0 = 1.000). In contrast, suffixal forms do not differ meaningfully from forms with empty prefixes (estimate = -0.017, 90% CI [-0.067, 0.031], posterior probability < 0 = 0.720).

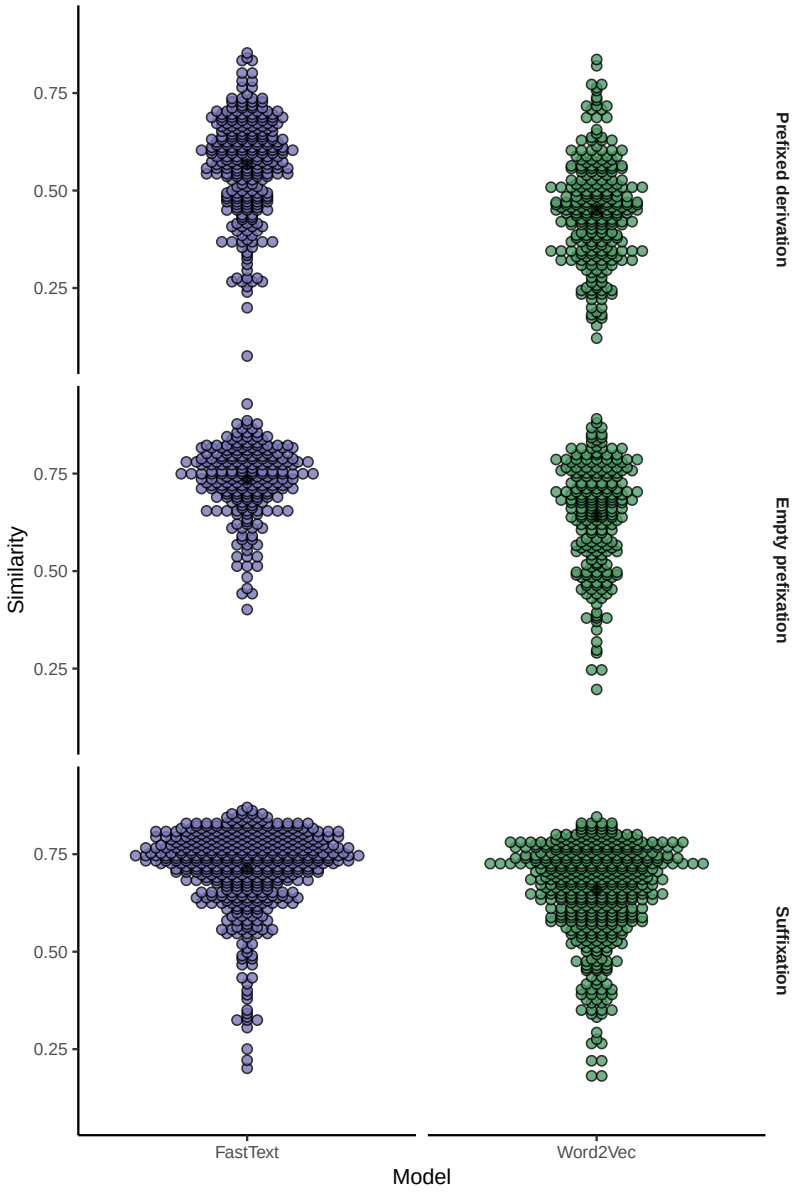

Figure 3. Similarity between an imperfective verb and three different ways of forming perfectives from it: prefixed derivation, empty prefixation and suffixation/stem alternation. Each coloured dot represents an imperfective-perfective pair. The black dot represents the mean similarity of all pairs.

Model	Derived	Empty prefix	Suffixed
FastText	0.57	0.73	0.71
Word2Vec	0.45	0.64	0.66

Table 9. Mean similarity in Polish imperfective-perfective pairs.

3.4 Discussion

Across both models, there is strong evidence that cosine similarity in forms with empty prefixes is comparable to that of inflectional suffixes, and both are markedly higher than the similarity observed in derivational prefixes. Regardless of whether the derivational forms are category-preserving, we find a consistent and substantial difference between inflectional and derivational morphology. These results suggest that empty prefixes pattern with inflectional morphology in terms of semantic similarity.

4. General discussion

This thesis has shed light on two different but related topics, the morphological inflection-derivation debate and the Polish aspectual marking debate. Using the methods developed in the first study, we tried to answer a theoretical question in the morphology of Polish.

How is it possible that we find contradictory evidence in the studies?

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.

Cappellaro, C., & Meinschaefer, J. (2022). Inflexion, Derivation, Compounding. In A. Ledgeway & M. Maiden (Eds.), *The Cambridge Handbook of Romance Linguistics* (pp. 400–433). Cambridge University Press. https://doi.org/10.1017/9781108580410.016

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

Mlynarczyk, 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.

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

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

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.