Abstract

This project investigates to what extent morphological inflection can be automatically distinguished from derivation based solely on word forms. The debate on inflection and derivation remains highly contentious in linguistic literature, with some viewing them as fundamentally similar or existing on a gradient, while others argue for a clear distinction. Despite extensive theoretical discussions, empirical evidence remains limited. One proposed distinction is semantic regularity: inflection is expected to be stable in its syntactic and semantic effects across lexemes (e.g., cinema is to cinemas as cat is to cats), whereas derivation is less so (e.g., delegation is not to delegate as election is to elect). However, this criterion has yet to be systematically tested large scale cross-linguistically. Building on previous work (e.g., Bonami and Paperno (2018); Rosa and Žabokrtský (2019)), this project uses distributional semantics and word embeddings to assess semantic regularity as a potential proxy for differentiating inflection and derivation. It focuses on two morphologically rich languages, Spanish and Polish. Additionally, the project examines differences in semantic regularity also among various types of inflection, such as case, number, gender, and person.

Keywords

morphology, inflection, derivation

Contents

1. Índice provisional

1. Introduction

   Distinction between inf/der (1 pag? media?)

   Distributional semantics (media pagina)

   Word vectors (media pagina)

   UniMorph (media pagina)

   Why Polish and Spanish? :D

2. Methodology

2. Introduction

The distinction between morphological inflection and morphological derivation is still an unresolved question. In order to tell them apart, authors provide different criteria (see for instance Booij (2006), Aronoff and Fudeman (2011), Booij (2012), Haspelmath and Sims (2013), Stump (2005) or Stump (2017), among many others). Some authors are of the opinion that they are essentially the same thing (Haspelmath, 2024) or at least that they exist on different ends of a scale or gradient (see Bybee (1985) or 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 often provide examples of exceptions to the very same criteria, so there is no infallible one.

Among the long list of criteria there is one that has been used before in research articles in order to provide empirical results on the distinction debate, there is the semantic regularity criterion: "Inflection is semantically more regular than derivation." (Stump, 2005). Regarding these academic articles that have explored this debate before there is Bonami and Paperno (2018), Rosa and Žabokrtský (2019) and Haley et al. (2024), which follow the same criterion even though they take different metrics.

In order to automatically test whether this criterium holds or not, we need to make use of the distributional hypothesis which can be explained as similar words appear in similar contexts.

3. Background and rationale

3.1 Inflection and derivation

3.2 Distributional semantics

3.3 Word vectors

3.4 UniMorph

3.5 Related studies

4. Methodology

This study explores the distinction between inflection and derivation in Polish and Spanish using various word embedding models, both static and contextual.

To achieve this, several models were implemented. For static embeddings, used Word2Vec and FastText were used, and for contextual embeddings, the Multilingual BERT model was used. The Word2Vec model for Spanish was trained on the Spanish Billion Words (SBW) corpus. For Polish, the IPIPAN Word2Vec model (nkjp+wiki-forms-all-300-skipg-ns) was used, which is comparable in quality to SBW. It was trained on the National Corpus of Polish (NKJP) and Wikipedia, includes all parts of speech and word forms, and produces 300-dimensional vectors using the skip-gram algorithm with negative sampling. In addition, FastText embeddings for both Spanish and Polish were applied to incorporate subword-level information.

In order to conduct an initial analysis two separate datasets were constructed, one for inflection and another one for derivation, extracting the data from UniMorph. For the inflection analysis, a Pivot/Inflection dataset was constructed.

The data was filtered to include the following verb tenses in Spanish:

- Present Indicative. UniMorph category: V;IND;PRS.

- Past Imperfect. UniMorph category: V;IND;PST;IPFV.

- Future Indicative. UniMorph category: V;IND;FUT.

This resulted in a dataset of 148,051 rows, each consisting of a base form, its inflected form, and the morphological category. Additional forms such as participles and gerunds are planned for future inclusion.

In Polish the filtering included:

- Present. UniMorph category: V;PRS.
- Past. UniMorph category:V;PST.
- Future. UniMorph category:V;FUT.

The resulting dataset contains 23,615 rows. For the derivation analysis the data provided by UniMorph was used without changes.

The initial analysis revealed some errors in both datasets. In the derivations dataset the label U (unspecified or unknown) presented some issues. The goal is to eliminate all instances of unknown categories, to get rid of this noise and have cleaner results and means.

1. In Spanish there are 20 rows that contain a derivation that results in U (i.e. N:U or V:U) and 107 in Polish. 2. On the other hand, there are even more derivations in which the pivot is tagged with U (U:N, U:ADJ...), 36 in Spanish and 253 in Polish.

4.1 Cleaning the derivational dataset

Taking a quick look through this data one can see many mistakes such as formatting issues or verbs, adjectives or nouns being labeled U. When it comes to Spanish, the number is not too high, so it is something that is worth fixing in order to get rid of this label. Fixing the first group seems fairly easy since a category can be assigned based on the affix.

In order to clean the Spanish derivations dataset a new category according to the affix was assigned. All the affixes that end in *-ero*, *-ez*, *-ismo*,*-í* and *-illa* were changed to N. V was also assigned to those that contain the affixes *-ar* and *-ear*. As a result 6 rows were obtained. They can be eliminated from the final dataset because of all the mistakes they contain.

The second group of Spanish derivations (:U) cannot be easily fixed with a Python script, it contains many numerals and words that are not N, ADJ, ADV or V. Those rows that do not contain any of such categories can be dropped and the rest probably needs to be fixed manually. It contais some verbs, nouns and adjectives labeled with U, for instance *cuarenta cuanrentóń U:N -ón*. For some reason *cuarenta* is labeled in other rows as N but not in this one. Since numerals can be N or ADJ, alongside all the other issues with this group all these rows (35) can be dropped. It is a low number that will not affect the results.

Polish data seems to need more work as there are more incorrect labels, but it can be fixed more easily. Affixes such as *-any*, *-ony*, *-ty*, *-y*, or *-ący*, *-ęty* take the label ADJ, because they are all endings that participles take. There also are some formatting issues. Some rows under the same condition (X:U) contain the pivot and the derived form joined together in the pivot cell (i.e. *mylićpomylić pomylić*). This can be fixed as well just removing the form from the pivot column and assigning to row the correct categories. Some rows that contained verbs in both columns but were not correctly labeled, so they were fixed as well. This was fairly easy since in Polish verbs in the infinitive form end in *-ć* (most of them) or *-c*. Three rows incorrectly labeled U:U were changed to ADJ:ADJ since they contained adjectives.

The table above represents the rest, which can be removed as well as they do not contain any nouns, adjectives, adverbs or verbs. Take for instance the appearance of *co*, *kto* or *jaki*. which are relative pronouns. Everything ending in *-ś* and *-ż* or *-że* are not nouns nor adjectives nor adverbs nor verbs, but other types of pronouns or particles, so they can be removed. Regarding the data labeled as U:X, it can be done much better than in Spanish because there are many rows (163) that contain verbs ending in *-ć* in both the pivot and the derivation column but are incorrectly labeled as U:V, for example *kręcić skręcic U:V s-* or *paść przepaść U:V prze-*. The label change to V:V can easily be done. There are also 27 rows that contain verbs ending in *-c* in both the pivot and the derivation cells, which can be fixed just like previously done on the other group of verbs.

Finally some pivots that are verbs but are not labeled as such can be changed just by looking at the ending, although this needs to be done carefully as some nouns can also end in

-ć or *-c*, for this reason it will only be done on the mislabeled ones (the ones labelled as U), which are all verbs.

Since almost 200 rows have been fixed, the resulting ones labeled as U:X contain only 21 rows, with some mistakes or words that are not N, ADJ, ADV or V so they can just be dropped. Both resulting datasets do not contain any row labeled with U anymore.

Inflection

Model	Language	Mean Similarity
FastText	Spanish	0.51
	Polish	0.54
Word2Vec	Spanish	0.50
	Polish	0.40
Mult BERT	Spanish	0.92
	Polish	0.93

Derivation

Model	Language	Mean Similarity
FastText	Spanish	0.51
	Polish	0.54
Word2Vec	Spanish	0.50
	Polish	0.40
Mult BERT	Spanish	0.92
	Polish	0.93

Table 1. Mean similarity between pivot and form in inflection and derivation by model and language

4.2 Cleaning the inflectional dataset

The only thing to clean in the inflections dataset are the *vos* and *usted* forms.

4.3 Subsetting the most frequent lemmas and affixes

For this task in Spanish the 10000 most frequent lemmas in CREA was used. The verbs that appear in both datasets, UniMorph and CREA, were extracted and a subset of 1568 lemmas was obtained from a total of 6695.

For the Polish data sgjp.pl was used. Using the site's implemented filter the 8500 most common lexemes were extracted and after filtering the verbs from that list, 1832 verbs were obtained. Afterwards that list of verbs was compared to the UniMorph data and those that appear in both datasets were extracted resulting in 455 lemmas from a total of 844 that appear in the UniMorph data.

To create the subset of affixes, the most common affixes in the UniMorph data itself were taken.

Spanish

Affix	Count
-mente	2997
-dor	1316
-ar	1310
-ero	1123
-miento	913
-ico	870
des-	836
-ción	831
-ear	676
-ista	642
-ito	638
-ismo	549
-ón	533
-idad	499
-al	491

Polish

Affix	Count
-owy	5804
-ka	5487
-anie	3421
-ość	3287
-ny	2414
-ie	2161
-enie	1669
-ek	1521
-ować	1517
-o	1393
-ik	1249
-ski	1212
-ąć	1158
-stwo	770
za-	742

Table 2. Top 15 affixes in Spanish and Polish in UniMorph data.

Inflection

Model	Language	Mean Similarity
FastText	Spanish	0.49
	Polish	0.49
Word2Vec	Spanish	0.50
	Polish	0.53
Mult BERT	Spanish	0.93
	Polish	0.90

Derivation

Model	Language	Mean Similarity
FastText	Spanish	0.51
	Polish	0.54
Word2Vec	Spanish	0.50
	Polish	0.40
Mult BERT	Spanish	0.92
	Polish	0.93

Table 3. Mean similarity between pivot and form in inflection and derivation by model and language on the subset data.

4.4 Implementing a random baseline

The previous baseline shuffled the whole data 10 times. In this case a better baseline is implemented. It consists of filtering the data, that is separately filtering inflections by tense (present, past or future) and derivations by category, then shuffling each filter 100 times while extracting the mean similarity of the whole filtered data. That means we are extracting the mean similarity a hundred times by type of morphological process (inflection or derivation), tense/category, model and language and plotting its distribution.

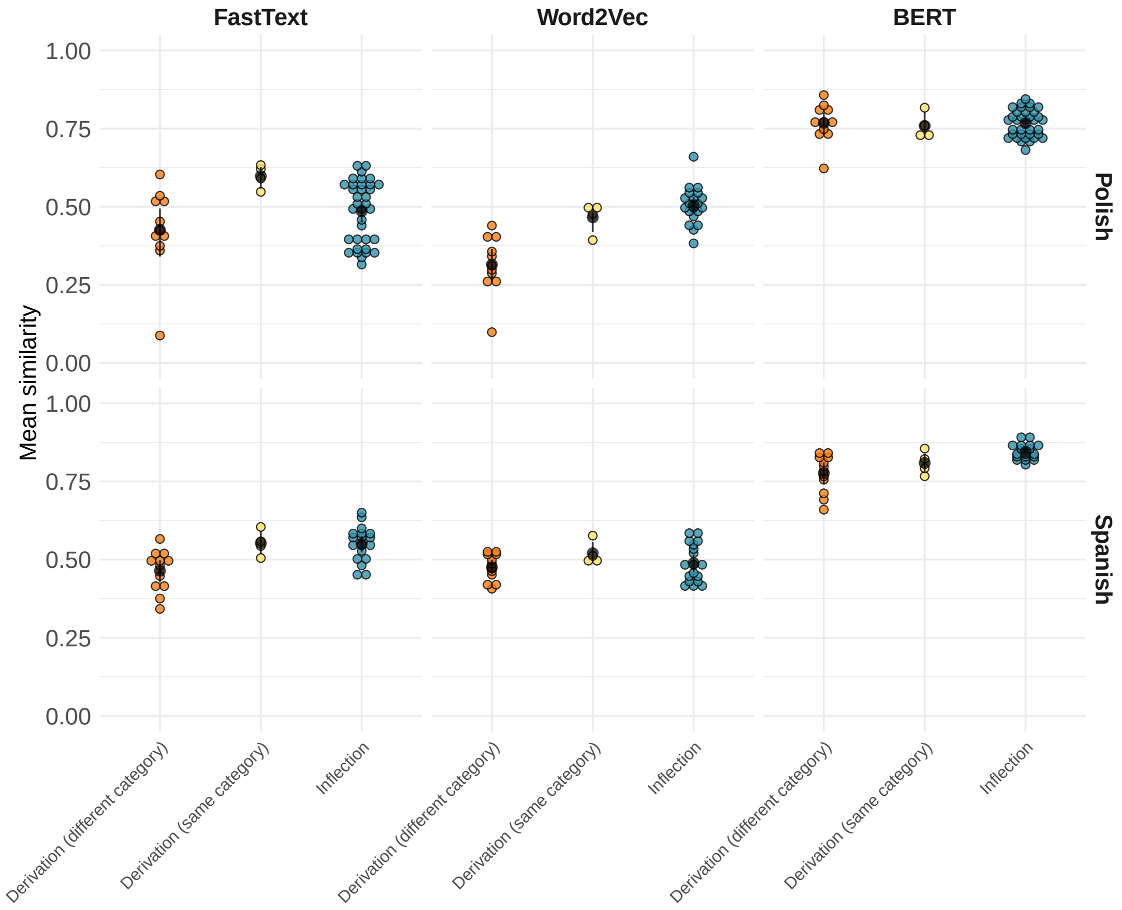

Figure 1. Mean similarity between pivot and form in inflection and derivation by model and language.

5. Conclusion

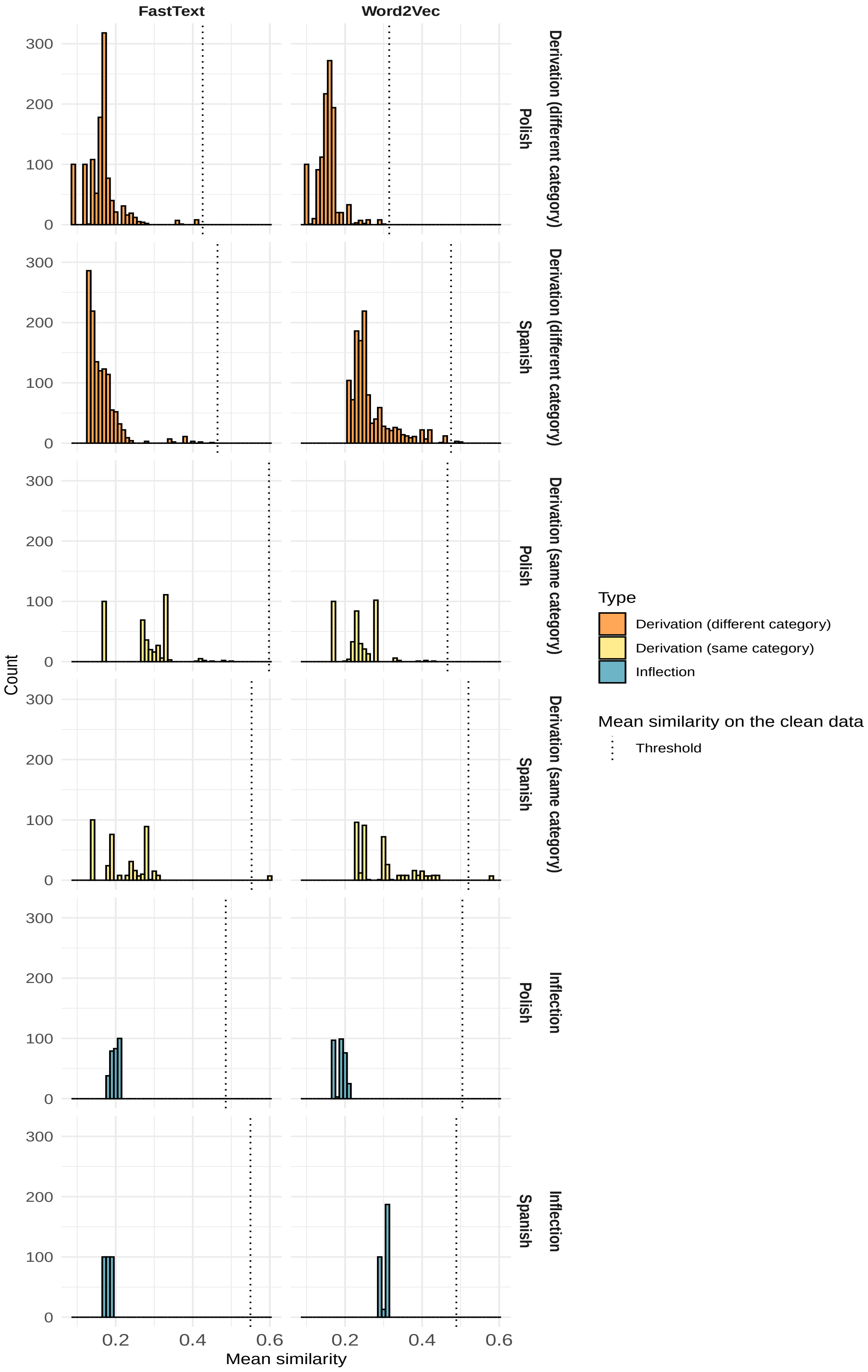

Figure 2. Random baseline

References

Aronoff, M., & Fudeman, K. A. (2011). *What is morphology?* Wiley-Blackwell. https://thuvienso.hoasen.edu.vn/handle/123456789/8789

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.

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

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