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, distributional semantics

Contents

1. Introduction

2. Background and rationale

2.1 Inflection and derivation

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.

2.2 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 2.3) 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 a Word2Vec French 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 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 find that inflectional forms are more similar to each other than derivational forms and that some derivational processess 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.

2.3 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), *fastText* (Bojanowski et al., 2017) (see section 3 for a detailed explanation of the specific models). 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 3.

2.4 UniMorph

UniMorph is a database that provides annotated morphological inflection tables and derivational morphology of numerous languages (Batsuren et al., 2022).

3. Methodology

This study explores the distinction between inflection and derivation in Polish and Spanish using two static word embedding models *word2vec* and *fastText*. The Spanish *word2vec* model was trained on the Spanish Billion Words (SBW) corpus (Cardellino, 2016). For Polish, the IP-IPAN *word2vec* model *nkjp+wiki-forms-all-300-skipg-ns* was used. This one was specifically chosen because it resembles the SBW model. It was trained on the National Corpus of Polish (NKJP) (enlace?) and Wikipedia, includes all parts of speech and word forms, and produces 300-dimensional vectors using the skip-gram algorithm with negative sampling. Regarding the *fastText* embeddings, the ones used are the newest version available in the *fastText* site [1] (Grave et al., 2018).

3.1 Constructing the datasets

In order to conduct the analysis two separate datasets were constructed, one for inflection and another one for derivation, extracting the necessary data from UniMorph. For the inflection analysis, a pivot/inflection 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 *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. This resulted in a dataframe of 148,051 rows (CHECK NUM OF ROWS) , each consisting of a base form (pivot), its inflected form (inflection), and its morphological category. 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). The resulting dataset contains 23,615 rows.

In the derivation data, the label U (unspecified or unknown) was either fixed or removed. In Spanish there are 20 rows that contain a derivation that results in U (X:U) and 107 in Polish. On the other hand, there are even more derivations in which the pivot is tagged with U (U:X),

[1] https://fasttext.cc/

Pivot	Inflection	Category
honrar	honro	V;IND;PRS;1;SG
honrar	honras	V;IND;PRS;2;SG;INFM
honrar	honra	V;IND;PRS;3;SG
honrar	honramos	V;IND;PRS;1;PL
honrar	honráis	V;IND;PRS;2;PL
honrar	honran	V;IND;PRS;3;PL

Table 1. Inflections dataframe sample.

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 affixes 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 assigned (V:X) to 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. Fixing the category assigned to the derivation position (X:U) proved trickier. There are numerals and words that are not nouns, adjectives, adverbs or verbs (all the UniMorph categories). The latter can simply be dropped, and since numerals could be tagged as either N or ADJ, they can also be dropped. The total number of wrongly tagged rows in this case is low (36), so it will have no major influence on the results.

Regarding the fixes in the Polish data, some rows incorrectly labelled U in the derivation position were fixed looking at the affixes. The affixes *-any*, *-ony*, *-ty*, *-y*, or *-ący*, *-ęty* should take the label ADJ, because they are all endings that participles take. There also are some formatting issues. Some rows under the same condition contain the pivot and the derived form joined together in the pivot cell (i.e. *mylićpomylić pomylić*). This was fixed as well just removing the derivation from the pivot column. Some rows that contained verbs in both columns but were not correctly labeled in the derivation form were fixed as well. Three specific rows incorrectly labeled U:U were changed to ADJ:ADJ since they contained adjectives. The rest can be removed as well as they do not contain any nouns, adjectives, adverbs or verbs.

There are more incorrectly labelled rows in the Polish data. There are many rows (163) that contain infinitives in both the pivot and the derivation column, but are incorrectly labelled U:V. In Polish, infinitives end in *-ć* or *-c*, so the label correction was easily done. Since almost

200 rows have been fixed, the resulting ones labeled U:X are not many (21). They also contain mistakes, such as words that are not nouns, adjectives,adverbs or verbs so they can just be dropped. Both the Spanish and Polish dataframes do not contain any row labeled with U anymore.

3.2 Obtaining the mean similarity

3.3 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.

3.4 Implementing a random baseline

In order to evaluate the effectiveness of the methodology and provide a point of comparison, a baseline was implemented. It consisted of filtering the data, that is separately filtering inflections by tense (present, past or future) and derivations by category, then each filter was shuffled 100 times while extracting the mean similarity of the whole filtered data. That is, 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.

Spanish		Polish	
Affix	**Count**	**Affix**	**Count**
-mente	2997	-owy	5804
-dor	1316	-ka	5487
-ar	1310	-anie	3421
-ero	1123	-ość	3287
-miento	913	-ny	2414
-ico	870	-ie	2161
des-	836	-enie	1669
-ción	831	-ek	1521
-ear	676	-ować	1517
-ista	642	-o	1393
-ito	638	-ik	1249
-ismo	549	-ski	1212
-ón	533	-ąć	1158
-idad	499	-stwo	770
-al	491	za-	742

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

Inflection			Derivation		
Model	**Language**	**Mean Similarity**	**Model**	**Language**	**Mean Similarity**
FastText	Spanish	0.49	**FastText**	Spanish	0.51
	Polish	0.49		Polish	0.54
Word2Vec	Spanish	0.50	**Word2Vec**	Spanish	0.50
	Polish	0.53		Polish	0.40
BERT	Spanish	0.82	**BERT**	Spanish	0.82
	Polish	0.73		Polish	0.83

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

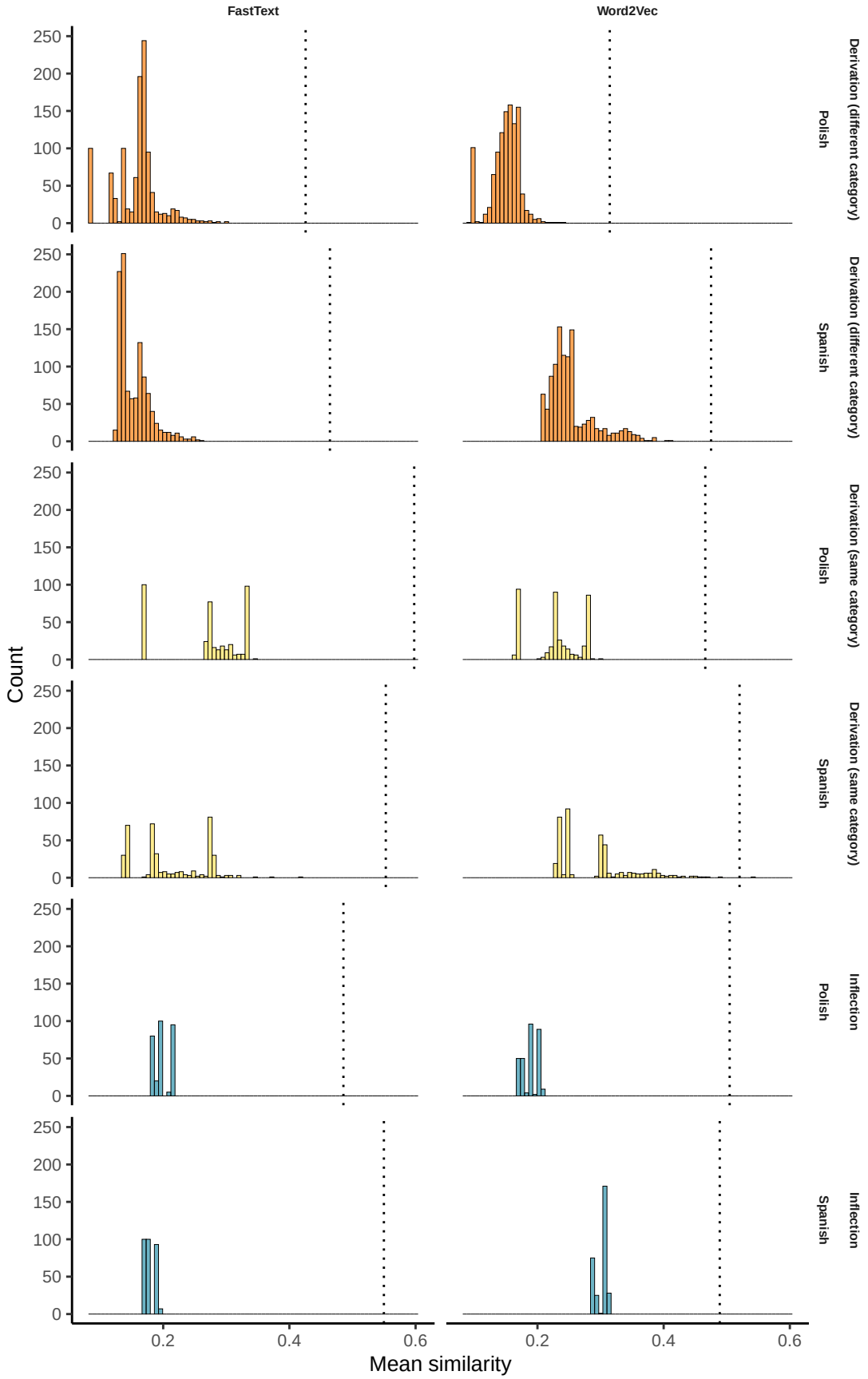

Figure 1. Distribution of the random baseline results. The dotted line represents the mean similarity value on the clean data.

3.5 Statistical analysis

4. Results

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

Table 4. Mean similarity between pivot and form in inflection and derivation by model and language

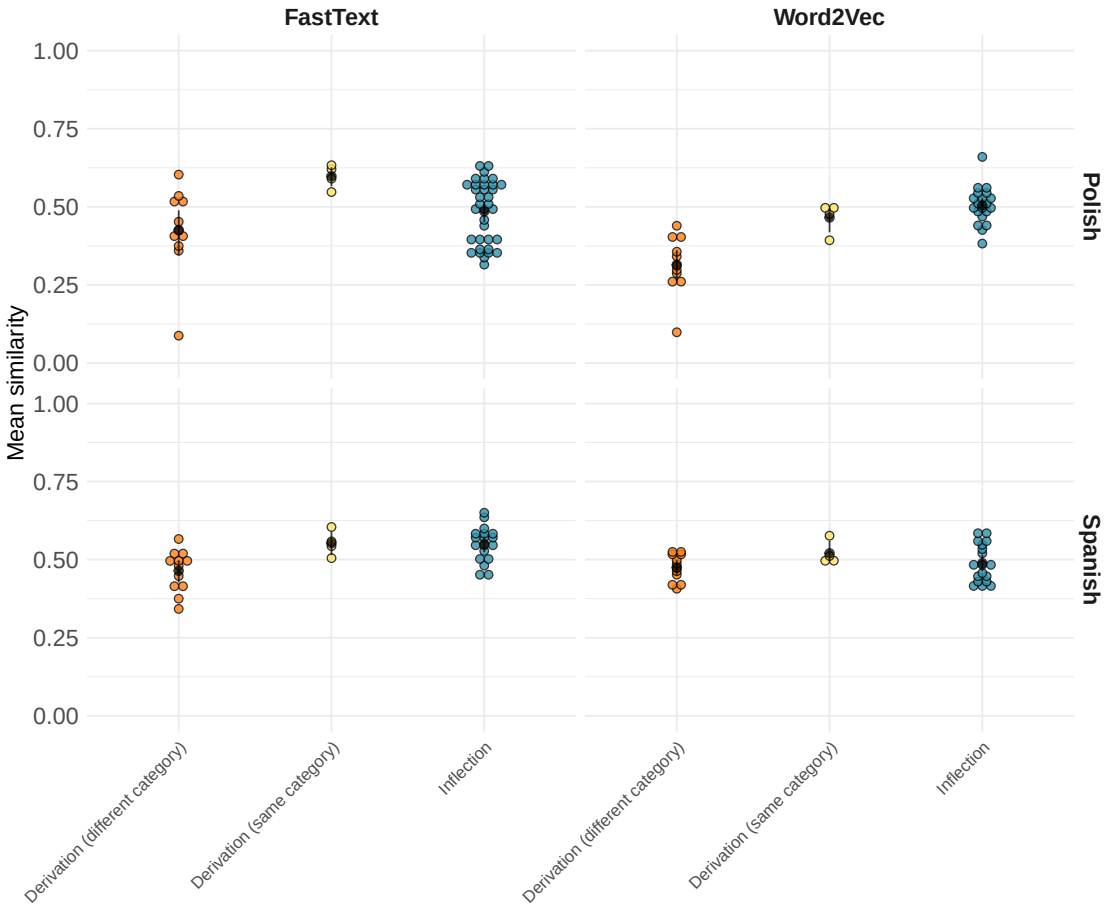

Figure 2. Mean similarity between pivot and form in inflection and derivation by model and language. Each dot represents a category. The black dot represents the mean similarity of all the categories and the standard deviation around it.

5. Discussion

6. Conclusion

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

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

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

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

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