Attempting to separate inflection and derivation using vector space representations

Rudolf Rosa **Zdeněk Žabokrtský**
Charles University, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Prague 1, Czech Republic
{rosa,zabokrtsky}@ufal.mff.cuni.cz

Abstract

We investigate to what extent inflection can be automatically separated from derivation, just based on the word forms. We expect pairs of inflected forms of the same lemma to be closer to each other than pairs of inflected forms of two different lemmas (still derived from a same root, though), given a proper distance measure. We estimate distances of word forms using edit distance, which represents character-based similarity, and word embedding similarity, which serves as a proxy to meaning similarity. Specifically, we explore Levenshtein and Jaro-Winkler edit distances, and cosine similarity of FastText word embeddings. We evaluate the separability of inflection and derivation on a sample from DeriNet, a database of word formation relations in Czech. We investigate the word distance measures directly, as well as embedded in a clustering setup. Best results are achieved by using a combination of Jaro-Winkler edit distance and word embedding cosine similarity, outperforming each of the individual measures. Further analysis shows that the method works better for some classes of inflections and derivations than for others, revealing some limitations of the method, but also supporting the idea of replacing a binary inflection-derivation dichotomy with a continous scale.

1 Motivation

The distinction between inflection and derivation is a traditional linguistic dichotomy, with a range of criteria to tell them apart (Stump, 1998; Haspelmath and Sims, 2013). However, the criteria are typically not easily testable in an automated way; rather, they are designed for a manual investigation carried out by a linguist.

In this work, we attempt to distinguish inflection from derivation automatically, based solely on the word forms, without using any annotated resources and any human decision-making. For each pair of morphologically related word forms, we want to automatically decide whether they are inflected forms (inflections, for short) of the same lemma, or not. We specifically focus on the lexical meaning change criterion by Stump (1998), as listed by Bonami and Paperno (2018): "if two morphologically related words have distinct lexical meaning, they must be related by derivation".

Obviously, if there is a lemmatizer available for the language under study (or a corpus annotated with lemmas which can be used to train the lemmatizer), the task could be trivially solved by lemmatizing the two word forms and checking whether the lemmas are identical or distinct. However, we are not interested in such a solution, as the necessary resources are only available for a small number of languages. The vast majority of the world's languages are under-resourced, lacking such datasets or tools, which gravely limits any research on such languages. The ability to perform the inflection-derivation distinction automatically, assuming only the availability of a plain text corpus of the language, would thus be of great value. Admittedly, for many languages, no plain text corpus of a considerable size is available; in such cases, we are out of luck. Nevertheless, medium-size plain text corpora exist for hundreds of languages – Wikipedia[1] covers 300 languages (Rosa, 2018), JW300 (Agić and Vulić, 2019) features texts

[1] https://www.wikipedia.org/

from Watchtower[2] for 300 languages (around 100k sentences each), and the text of the whole or a part of the Bible is available for as many as 1,400 languages (Mayer and Cysouw, 2014).

Still, in this work, our goal is not (yet) practical, i.e. devising a tool applicable to under-resourced languages, but rather exploratory, investigating the mere feasibility of such an approach. Therefore, we only use a single resource-rich language for the investigation, so that we can reliably analyze the performance of our approach, for which we need annotated datasets.

Moreover, as an outlook to future work, we are also interested in empirically exploring the boundary between derivation and inflection, which is notoriously vague. We hope that empirical computational methods could provide some solid ground in this respect, revealing to which extent the boundary can be observed, and possibly even providing empirical means of estimating the inflectionality and derivationality of individual phenomena, e.g. in the form of a scalar value.

Thus, while we hope the presented work to have some practical applications, our primary motivation is sheer curiosity. Can we automatically distinguish inflection from derivation, without using annotated data? How clear does the boundary seem to be? Can we estimate the position of a morphological operation on the inflection-derivation scale? Which operations, traditionally annotated as derivations, seem to behave more like inflections, and vice versa? This work is just a starting point on our journey to empirically explore such questions. Nevertheless, it already allows us to peek at what really seems to be going on in language (Czech language, at this stage) in terms of inflection and derivation.

2 Related Work

In morphology, derivation and inflection are traditionally distinguished. The former one deals with creating word forms from the same lexeme, while the latter one captures processes for the creation of new lexemes. Like with many other linguistic dichotomies, there is a critical debate about the existence of a real divide between inflection and derivation, ranging from approaches trying to define precise operational criteria to distinguish the two, through those that assume rather a gradual scale, to those that reject this opposition as such. The arguments used in the debate were summarized e.g. by Booij (2006) and by ten Hacken (2014).[3]

Originally, the criteria for distinguishing inflection from derivation were formulated mostly using high-level linguistic notions (for instance, inflected forms of lexemes are supposed to preserve lexical meaning), which makes it difficult to evaluate in an objective way. More recently (roughly in the last two decades), there are attempts to find the boundary using also psycholinguistic or even brain-imaging methods, see e.g. (Julínková, 2012) and (Bozic and Marslen-Wilson, 2010), respectively. Typically, the experimental results are mixed, indicating that some such assumed opposition partially correlates with measurements, but without offering any clear-cut divide either. In addition, all such experiments are naturally hard to scale to bigger data and/or more languages.

In our study, we take the existence of a crisp inflection-derivation boundary as an assumption, and we try to get close to the boundary in a fully unsupervised way, using only unlabelled corpus data.

For evaluation purposes, we accept the boundary as technically defined in existing morphological NLP resources for Czech. More specifically, we use MorfFlex CZ (Hajič and Hlaváčová, 2016) to bind inflected word forms with their lemmas (more exactly, we use only corpus-attested word forms), and the word-formation database DeriNet (Ševčíková and Žabokrtský, 2014), in which relations between derivationally related lexemes are represented in the form of rooted trees (one tree per a derivational family).

To the best of our knowledge, the only work to investigate a similar question is the recent research of Bonami and Paperno (2018). Similar to us, the authors are interested in a way to turn the human-centered criteria of distinguishing inflection from derivation into something empirically testable. The authors investigated the semantic regularity criterion ("inflection is semantically more regular than derivation"),

[2] https://www.jw.org/

[3] The debate seems not much heated for the Czech language nowadays, however, there are linguistic phenomena in Czech which are considered inflection by some scholars and derivation by others. For instance, the category of comparative is handled as inflection in modern NLP tools for Czech, but was considered word formation e.g. by Trávníček (1951).

while we selected the lexical meaning change criterion in this work (distinct lexical meanings indicate derivation).

Both Bonami and Paperno (2018) and us are interested in the meanings of the individual words, and both works make the usual choice of using word embeddings as a proxy to word meanings. As the criterion that we test is simpler, our method is also simpler: we directly measure the difference of word embeddings to estimate the distance of meanings. To estimate the *regularity* of meaning change, Bonami and Paperno (2018) take a further step of estimating an embedding vector shift corresponding to a particular morphological operation, and observe that the vector shift tends to be more regular for inflectional operations than for derivational operations.

A partially related work is that of Musil et al. (2019), showing that there is some regularity in the vector shift corresponding to individual derivational operations. However, the authors do not contrast this with inflectional operations. We utilize their work to provide categories of derivational operations, which are not yet annotated in DeriNet and have to be estimated heuristically.

While the methods used by us and previously mentioned authors are rather simple, we are unaware of any other substantial research in this direction. There is research on unsupervised morphology induction, represented by the well-known Morfessor system of Creutz and Lagus (2007), the interesting ParaMor system (Monson et al., 2008) which attempts to find inflectional paradigms, as well as the earlier minimum description length-based system of Goldsmith (2001). While some ideas behind these systems are related to our interests and may potentially be useful to us, their goal is to perform morphological segmentation, which is a related but different task. Another related area is stemming (Lovins, 1968; Porter, 2001), which can be thought of as simple lemmatization. However, stemmers tend to be too coarse, often assigning the same stem to both inflections and derivations. Moreover, they are typically rule-based and thus language-specific, which is not in line with our goals.

3 Approach

Our central hypothesis is that word forms that are inflections of the same lemma tend to be *more similar* than inflections of different lemmas. To measure the similarity of word forms, we investigate two somewhat orthogonal simple approaches.

Our first method is to use string edit distances, which measure how much the word forms differ on the character level. In our work, we use the Jaro-Winkler (JW) edit distance (Winkler, 1990) and the Levenshtein edit distance (Levenshtein, 1966).

As the second method, we propose to measure similarity of word embeddings (Mikolov et al., 2013; Grave et al., 2018). It has been shown that cosine similarity of word embeddings tends to capture various kinds of word similarities, including morphological, syntactic, and semantic similarities, and can be thought of as a proxy to meaning similarity.

We then apply the methods to sets of corpus-attested words belonging to one derivational family, i.e. a set of words that are, according to a database of word formation relations, all derived from a common root, together with their inflections extracted from a lemmatized corpus. Some words in the set are thus inflections of a common lemma, while others are inflections of different lemmas derived from a common root. We evaluate the accuracy with which the methods separate inflections from derivations, both independently for each pair of word forms as well as in an unsupervised clustering setup.

4 Word form distance measures

4.1 String similarity

For string similarity, we use the Levenshtein (LD) edit distance (Levenshtein, 1966) and the Jaro-Winkler (JW) edit distance (Winkler, 1990).

A potential advantage of JW over LD is that it gives more importance to the beginnings of the strings than to their ends. We find this to be advantageous, as most of the inflection usually happens at the end of the word, i.e. suffixing is more common than prefixing. Specifically, as shown by Table 1, adapted from the WALS database by Dryer and Haspelmath (2013), half of the studied languages showing a non-trivial amount of inflectional morphology are predominantly suffixing, and further 15% show a

preference for suffixing; moreover, nearly all Eurasian languages, which one is most likely to encounter in practice, fall into this category. For the languages with no clear prefixing-suffixing preference (18% of studied inflectional languages), we expect LD to be more appropriate than JW; however, these are mostly low-resource indigenous languages found in central Africa and the Americas, not frequently encountered in practice.[4]

Value	Languages	% of all	% of inflectional
Little or no inflectional morphology	141	15%	–
Predominantly suffixing	406	42%	49%
Moderate preference for suffixing	123	13%	15%
Approximately equal amounts of suffixing and prefixing	147	15%	18%
Moderate preference for prefixing	94	10%	11%
Predominantly prefixing	58	6%	7%

Table 1: Values of Map 26A, Prefixing vs. Suffixing in Inflectional Morphology, showing the number and proportion of languages with various prefixing/suffixing preferences. Adapted from WALS (Dryer and Haspelmath, 2013).

Moreover, JW is in the $[0, 1]$ range, making it easily comparable and combinable, while LD returns a natural number of edit operations. For practical reasons, we transform LD into the $[0, 1]$ range by dividing it with the total length of the pair of word forms:[5]

$$LD_{rel}(w1, w2) = \frac{LD_{abs}(w1, w2)}{|w1| + |w2|} \tag{1}$$

While the edit distances treat all distinct characters as equally distant, some types of changes to the word form tend to be more common during inflection, and thus should presumably have a lower weight in the distance measure. To compute the edit distance of a pair of strings, we thus optionally average their edit distance with edit distance of their *simplified variants*; the simplification consists of lowercasing, transliteration to ASCII using the Unidecode library,[6] and deletion of non-initial vowels (a e i o u y).

4.2 Word embedding similarity

We use the cosine similarity of pretrained FastText word embeddings (Grave et al., 2018), downloaded from the FastText website.[7] Compared to the classical Word2vec (Mikolov et al., 2013), FastText embeddings have the benefit of employing subword embeddings. This means that they seamlessly handle out-of-vocabulary word forms, and also that they implicitly capture string similarity to some extent.[8]

The cosine similarity is computed as the inner product of the normalized FastText vectors of the pair of word forms; for practical reasons, we also shift it from the $[-1, 1]$ interval to the $[0, 1]$ interval, and reverse it to turn the similarity measure into a distance measure:

$$COS(w1, w2) = \frac{1 - vec(w1) \cdot vec(w2)}{2 \cdot |vec(w1)| \cdot |vec(w2)|} \tag{2}$$

4.3 Combined distance measure

We also combine the edit distance with the embedding distance via multiplication of the similarities.[9] As will be shown later, JW achieves better results than LD; therefore, we only use JW in the combination:

$$CD(w1, w2) = 1 - (1 - JW(w1, w2)) \cdot (1 - COS(w1, w2)) \tag{3}$$

[4]When dealing with a language with a preference for prefixing inflectional morphology (18% of studied inflectional languages), one can simply reverse the word forms before applying JW.

[5]Another option would be to divide the distance only by the length of the longer word. In our case, we chose a normalization that implicitly incorporates length similarity of the words.

[6]https://pypi.org/project/Unidecode/

[7]https://fasttext.cc/

[8]In brief preliminary experiments, FastText achieved significantly better results than Word2vec.

[9]The multiplication works like a logical "and": close word forms should be *similar* both string-wise *and* in meaning.

5 Evaluation

5.1 Evaluation methods

For the main evaluation, we use two methods, both evaluating to which extent the distance measures are able to separate inflections of the same lemma from inflections of different lemmas on a set of words belonging to a common derivational family.

5.1.1 Pairwise evaluation

In the pairwise evaluation method, we find a distance threshold that optimally separates inflection pairs from non-inflection pairs. We define W_{infl} as the set of all pairs of word forms that are inflections of the same lemma, and W_T as the set of all pairs of word forms whose distance is lower than a threshold T:

$$W_{infl} = \{w1, w2 | lemma(w_1) = lemma(w_2)\}; W_T = \{w1, w2 | dist(w_1, w_2) < T\} \tag{4}$$

We then compute the precision, recall, and F1 score of inflection pairs closer than T:

$$P_T = \frac{|W_{infl} \cap W_T|}{|W_T|}; R_T = \frac{|W_{infl} \cap W_T|}{|W_{infl}|}; F_T = \frac{2 \cdot P_T \cdot R_T}{P_T + R_T} \tag{5}$$

And finally, we find a threshold T that maximizes the F_T score:

$$F_{pairwise} = argmax_{T \in [0,1]} F_T \tag{6}$$

The resulting F1 score is a kind of an upper bound accuracy for the method, as the optimal separating threshold is selected in an oracle manner.

5.1.2 Clustering-based evaluation

We also perform a clustering of the word forms, and then evaluate the resulting clusters. We apply agglomerative clustering[10] from Scikit-learn (Pedregosa et al., 2011) with average linkage. The algorithm starts by assigning each word form to a separate cluster. In each step, it then merges the pair of clusters with the lowest average distance of their elements. We stop the algorithm once the number of clusters reaches the oracle number of lemmas in the derivational family.

We then evaluate the clustering in a similar way as in the pairwise method, with the objective that inflections should fall into common clusters and non-inflections should fall into different clusters. We define W_{infl} as in (4), and W_{clust} as the set of all pairs of word forms that fell into the same cluster:

$$W_{clust} = \{w1, w2 | clust(w_1) = clust(w_2)\} \tag{7}$$

We then compute the precision, recall, and F1 score of inflection pairs clustered together:

$$P_{clust} = \frac{|W_{infl} \cap W_{clust}|}{|W_{clust}|}; R_{clust} = \frac{|W_{infl} \cap W_{clust}|}{|W_{infl}|}; F_{clust} = \frac{2 \cdot P_{clust} \cdot R_{clust}}{P_{clust} + R_{clust}} \tag{8}$$

5.2 Experiment setting

We extract derivational families from DeriNet v1.7 (Žabokrtský et al., 2016),[11] a database of Czech word formation relations. As the database only contains word lemmas, we enrich the extracted lemma sets with inflections of the lemmas found in the Czech National Corpus, subcorpus SYN v4 (Křen et al., 2016), a large corpus of Czech lemmatized automatically using morphological analyzer MorfFlex CZ (Hajič and Hlaváčová, 2016).[12] We lowercase all the word forms.

[10] https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering

[11] http://ufal.mff.cuni.cz/derinet

[12] MorfFlex CZ (Hajič and Hlaváčová, 2016) offers complete inflectional paradigms, which leads to generating many extremely rare or even unused word forms. Thus we prefer to use only corpus-attested word forms in our study.

Distance	Embeddings (COS)	Levenshtein (LD)		Jaro-Winkler (JW)		Combination (CD)	
Simplified	–	no	yes	no	yes	no	yes
Pairwise	35.13%	37.71%	38.49%	38.73%	38.47%	41.27%	41.86%
Clustering	30.36%	31.75%	32.04%	34.39%	34.75%	36.44%	37.13%

Table 2: F1 scores of inflection identification using pairwise or clustering-based evaluation, with various word form distance measures. Edit distances optionally additionally use simplified word forms.

For the evaluation presented here, we randomly sampled 42 out of the 561 derivational families which contain at least 50 lemmas.[13,14] The derivational families range from 51 to 751 lemmas, totalling 4,514 lemmas, which are expanded through the corpus to 69,743 word forms.

We perform both the pairwise evaluation and the clustering-based evaluation on each of the derivational families separately, and report macro-averages[15] of F1 scores.

5.3 Results

The results in Table 2 show that the proposed method can separate inflection from derivation to some extent, reaching F1 scores around 40%.

This number is somewhat hard to interpret, as there is no clear baseline to compare it to. On one hand, current supervised lemmatizers typically reach accuracies well over 90%, but in a quite different setting. On the other hand, inflection pairs form only around 2% of our dataset, as the vast majority of the word form pairs are various inflections of rather distant derivations, so a trivial random baseline would achieve a score around 2%. The proposed distance measures thus manage to separate a small set of close inflections and derivations from a large set consisting of most of the non-inflections and some of the inflections.

Interestingly, both the edit distances and the embedding distance achieve accuracies in a similar range (with the embedding distance being slightly weaker), despite the methods being quite different.[16] Their combination then achieves even better results in both evaluation measures.

The JW distance achieves slightly better performance than LD, presumably due to the fact that it gives more weight to prefixes than suffixes, while inflection mostly happens at the suffix, as was already discussed. We can also see that the word form simplification generally slightly improves the results.

6 Further analysis

To get a better understanding on how the suggested distance measures perform on the task, we perform several further pairwise analyses.

In the main evaluation, we used pairs of all word forms belonging to the same derivational family. In such a setting, most pairs consist of rather distant word forms, which are clearly non-inflectional and thus rather boring. Therefore, we now focus only on the closest pairs of word forms, linked by a single inflectional or derivational operation:

- lemmas linked by a derivational edge; e.g. “dýchat” (breathe) – “dýchatelný” (breathable)
- forms of one lemma, differing only in one feature;[17] e.g. “písně” (song$_{sg,gen}$) – “písní” (songs$_{pl,gen}$)
- forms of two lemmas linked by a derivational edge, not differing in any morphological feature; e.g. “barvám” (colours$_{pl,dat}$) – “barvičkám” (crayons$_{pl,dat}$)

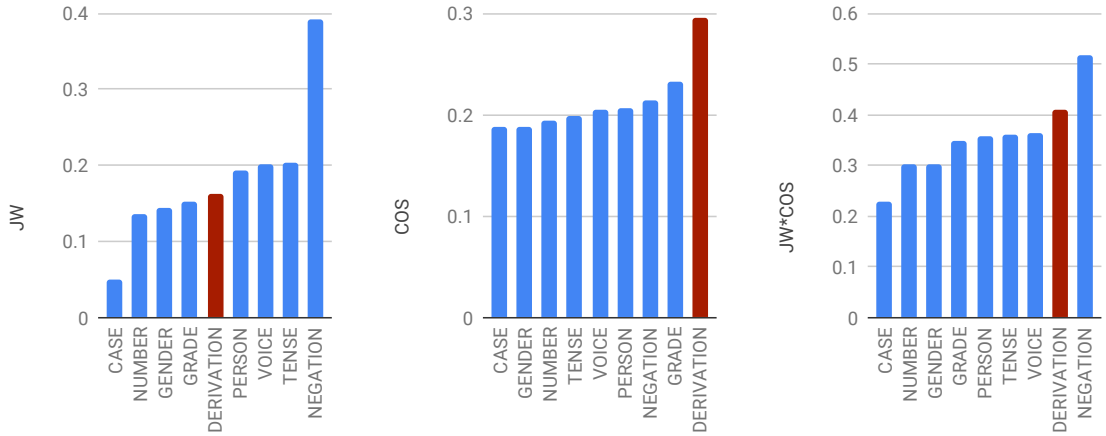
[13]While this may bias the research, we found that small derivational families, when filtered against corpus-attested forms, typically provide too small and sparse data for a meaningful analysis of derivational relations.

[14]Specifically, we use the following derivational roots: barva, báseň, bavit, bílý, bloknout, bloudit, budovat, bydlet, část, cena, cesta, chránit, dýchat, hádat, hospodář, hrát, hvězda, kód, kouř, křest, kult, malovat, norma, pět, politika, prach, produkt, program, rada, rodit, rovný, spět, střelit, tělo, typ, um, vědět, vinout, vládat, voda, zeď, žena.

[15]As the number of word form pairs grows quadratically with the number of word forms, we need to prevent a few largest derivational families from dominating the results.

[16]It is worth noting that FastText operates on character n-grams as well as full words, thus implicitly also capturing some string-based similarity.

[17]If a feature is set for only one of the forms (e.g. gender which is marked on verbs in the past tense but not in the present tense), we treat it as not differing; thus e.g. a change of verb tense is treated as a change in the tense feature only.



(a) JW distance (b) COS distance (c) JW*COS distance

Figure 1: Count-weighted average distances based on inflection type (or derivation).

Even then, although we are filtering the data for only corpus-attested forms, the corpus is large enough to contain many weird uncommon word forms, which we are not particularly interested in analyzing, as we rather want to see how the distance measures perform in typical situations. Therefore, we perform a token-based rather than type-based evaluation, simulating repeated evaluation of each word form based on its count in the corpus; as we are performing pairwise evaluation, we simply take the product of the counts of the individual word forms for the pair count.

6.1 Inflections

Figure 1 shows the count-weighted average distances of inflections that differ in the individual morphological features; average distance of derivations is also shown.

JW distance (Figure 1a) is very high for negation; this is expected since it modifies the beginning of the word (using the 'ne-' prefix), to which JW is particularly sensitive. The average JW distance for case inflection is very low; this is due to the token-based evaluation, as most frequent cases typically differ only in one or two word-final characters. We also note that the JW distance for grade inflection is surprisingly low, given the fact that superlatives in Czech are formed by prefixation ('nej-'); however, comparatives are much more common in the corpus (in type-based evaluation, grade inflection would rank much higher). JW clearly does not separate derivation from inflection well, as derivations exhibit a medium JW distance on average. Obviously, JW is also quite unsuitable for irregular inflections and suppletives, such as "jde" – "šel" (goes – went).

Figure 1b shows that COS distance of word embeddings separates inflection from derivation very well. Otherwise, highest COS distances are observed for negation and grade inflections, which are morphological operations on the boundary between inflection and derivation. While word embeddings are known to often perform poorly at distinguishing synonyms from antonyms, we did not observe this issue.

In the combined distance measure (Figure 1c), derivation remains quite well separated from inflection, apart from the boundary operation of negation. In all measures, the same three operations show the lowest distances: inflection for case, number, and gender.[18] These are quite typical pure inflections, extremely productive, governed by clear rules, mostly determined by syntactic rules and agreement, with a small and regular effect on the meaning of the inflected words.

[18] A morphological change of gender is considered to be an inflectional operation in Czech on verbs and adjectives, where it is governed by agreement, but not on nouns.

6.2 Derivations

In Figure 1, we grouped all derivational operations together, as the version of DeriNet we used does not contain any labels of derivational edges. However, with the help of the heuristic labelling by Musil et al. (2019), we performed a manual inspection of the results and gathered a number of observations.

We observed very low distances in all measures for the change from a perfective verb to its imperfective counterpart. While traditionally treated as derivation in Czech, this is a very regular and productive suffixation, and the change in meaning is also quite small and regular, and could thus be also treated as inflection. Interestingly, this is only partially true for the inverse of forming a perfective from a naturally imperfective verb, where a range of prefixes can be used, and there are often multiple options in use with varying meanings. This is correspondingly manifested by large JW distances of the forms, but COS distances remain low (although higher than for the perfectivisation).

Other low-distance operations that we observed are the transition from an adjective to an adverb (low JW, medium COS), formation of a diminuitive noun (low JW and COS), and formation of a possessive adjective from a noun (medium to low JW and COS), all of which are highly regular and productive, associated with a regular small change of meaning. All of these can be regarded as somewhere between a derivation and an inflection, motivating the idea of using a continous inflection-derivation scale rather than a strict binary categorization.

Derivations that radically change the part of speech, such as a transition between a verb and a noun, typically have higher COS distances, as the shift of the meaning is usually large; JW distances are medium to high, as there is often a large change of the suffix, sometimes accompanied by changes in the root. This is in line with these being quite prototypical derivations.

We observed a rather high contrast of a medium to low JW but a high COS distance for the change of a masculine noun to its feminine variant. This is typically performed by a small semi-regular suffix change, but considerably changes the meaning, albeit in a very regular way.

7 Conclusion

In this work, we attempted to automatically distinguish inflection from derivation without learning from annotated data, mainly based on the assumption that derivations tend to shift the meaning more than inflections. We tried several word distance measures, based both on the characters in the word forms, as well as distributional vector space representations of the word forms. We found a multiplication of Jaro-Winkler distance with cosine distance of FastText word embeddings to achieve the best results.

We used two evaluation setups, either directly separating the word form pairs based on the optimal distance threshold found in an oracle way, or clustering the word forms with an agglomerative clustering algorithm.

We conducted experiments on a subset of a Czech word formation database, observing F1 accuracy of inflection separation around 40%. Further analysis of the results showed that different classes of inflections and derivations are typically associated with different word form distances. To some extent, this corresponds to the inflectionality of some derivations and the derivationality of some inflections; however, to some extent, this is simply an artifact of the properties of the methods.

In future, we would like to employ multiple inflection-derivation distinction criteria described in the literature to improve the methods. From a research point of view, we are interested in arriving at an empirical measure of inflectionality versus derivationality of morphological operations, as this seems to be a more adequate view than a strict binary separation of inflection from derivation.

We also intend to extend this method to a wider range of languages. Our preliminary experiments on a set of 23 languages (Rosa and Žabokrtský, 2019) indicate that this should be feasible, obtaining some promising results for 20 of the languages (Arabic, Estonian, and 18 Indo-European languages).

We make all our code available on GitHub.[19]

[19] https://github.com/ptakopysk/lemata

Acknowledgments

This work was supported by the grants No. GA19-14534S and GA18-02196S of the Czech Science Foundation and the project No. DG16P02B048 of the Ministry of Culture of the Czech Republic. It has been using language resources developed, stored, and distributed by the LINDAT/CLARIAH CZ project (LM2015071, LM2018101).

References

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, pages 3204–3210. https://www.aclweb.org/anthology/P19-1310.

Olivier Bonami and Denis Paperno. 2018. Inflection vs. derivation in a distributional vector space. *Lingue e linguaggio* 17(2):173–196.

Geert Booij. 2006. Inflection and derivation. In Keith Brown, editor, *Encyclopedia of Language and Linguistics, Second Edition*, Elsevier, pages 654–661.

Mirjana Bozic and William Marslen-Wilson. 2010. Neurocognitive contexts for morphological complexity: Dissociating inflection and derivation. *Language and Linguistics Compass* 4(11):1063–1073.

Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing (TSLP)* 4(1):3.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig. http://wals.info/.

John Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics* 27(2):153–198. https://doi.org/10.1162/089120101750300490.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Jan Hajič and Jaroslava Hlaváčová. 2016. MorfFlex CZ 160310. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-1673.

Martin Haspelmath and Andrea Sims. 2013. *Understanding morphology*. Routledge.

Radka Julínková. 2012. *„Split Morphology Hypothesis" na materiálu češtiny*. Master's thesis, Univerzita Palackého v Olomouci, Filozofická fakulta.

Michal Křen, Václav Cvrček, Tomáš Čapka, Anna Čermáková, Milena Hnátková, Lucie Chlumská, Tomáš Jelínek, Dominika Kováříková, Vladimír Petkevič, Pavel Procházka, Hana Skoumalová, Michal Škrabal, Petr Truneček, Pavel Vondřička, and Adrian Zasina. 2016. SYN v4: large corpus of written Czech. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-1846.

Vladimir Iosifovich Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*. volume 10, pages 707–710.

Julie Beth Lovins. 1968. Development of a stemming algorithm. *Mech. Translat. & Comp. Linguistics* 11(1-2):22–31.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Languages Resources Association (ELRA), Reykjavik, Iceland, pages 3158–3163. http://www.lrec-conf.org/proceedings/lrec2014/pdf/220_Paper.pdf.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119.

Christian Monson, Jaime Carbonell, Alon Lavie, and Lori Levin. 2008. Paramor: Finding paradigms across morphology. In Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Douglas W. Oard, Anselmo Peñas, Vivien Petras, and Diana Santos, editors, *Advances in Multilingual and Multimodal Information Retrieval*. Springer Berlin Heidelberg, Berlin, Heidelberg, pages 900–907.

Tomáš Musil, Jonáš Vidra, and David Mareček. 2019. Derivational morphological relations in word embeddings. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Florence, Italy, pages 173–180. https://www.aclweb.org/anthology/W19-4818.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Martin F Porter. 2001. Snowball: A language for stemming algorithms.

Rudolf Rosa. 2018. Plaintext Wikipedia dump 2018. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-2735.

Rudolf Rosa and Zdeněk Žabokrtský. 2019. Unsupervised lemmatization as embeddings-based word clustering. *CoRR* abs/1908.08528. http://arxiv.org/abs/1908.08528.

Gregory T Stump. 1998. Inflection. In Andrew Spencer and Arnold M. Zwicky, editors, *The Handbook of Morphology*, London: Blackwell, pages 13–43.

Pius ten Hacken. 2014. Delineating derivation and inflection. In *The Oxford handbook of derivational morphology*, Oxford University Press.

František Trávníček. 1951. *Mluvnice spisovné češtiny*, volume 1. Slovanské nakl.

Magda Ševčíková and Zdeněk Žabokrtský. 2014. Word-Formation Network for Czech. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*. ELRA, Reykjavik, Iceland, pages 1087–1093.

William E. Winkler. 1990. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods (American Statistical Association)*. pages 354–359. http://www.amstat.org/sections/srms/Proceedings/papers/1990_056.pdf.

Zdeněk Žabokrtský, Magda Ševčíková, Milan Straka, Jonáš Vidra, and Adéla Limburská. 2016. Merging data resources for inflectional and derivational morphology in Czech. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. pages 1307–1314.