Identification of root morphs in morphologically segmented data

Vojtěch John and **Magda Ševčíková** and **Zdeněk Žabokrtský**
Charles University, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Praha, Czech Republic

Abstract

As a result of the ongoing push for unification, extension and integration of morphological resources, need arises for reliable low-resource morph classification, especially root identification. The paper reports on our experiments with multiple root identification methods with various degrees of supervision, tested on several Indo-European languages, showing, among others, that given morphological segmentation, surprisingly good root identification can be achieved using simple unsupervised statistical methods, the main bottlenecks being compounding and homomorphy resolution.

1 Introduction

The recent push for cross-lingual unification of morphological resources has, among others, brought about the unification of various resources devoted to morphological segmentation (Batsuren et al., 2022b; Žabokrtský et al., 2022), i.e. the task of dividing words into the smallest meaning-bearing units (morphemes or morphs), as well as the closely connected task of morphological classification – dividing the morphemes to classes (of various granularity). Nevertheless, in the available resources, the overall quality and/or completeness of morphological segmentation tends to be higher than that of the morphological classification, or the classification is even missing completely. This is reinforced by the fact that the state-of-the-art morphological segmentation approaches (as witnessed by the 2022 SIGMORPHON shared task; Batsuren et al. 2022a) are based on neural networks and neither include morphological classification nor can be straightforwardly used for obtaining it (even though there are some promising exceptions; e.g. Bolshakova and Sapin 2021, who use neural networks for both morphological segmentation and classification with word-level accuracy of over 90 %). As a result, morphological segmentation of reasonable quality is often easier to obtain than the corresponding morphological classification.

Furthermore, as the tasks of morphological segmentation and classification are closely connected to derivational morphology and as the derivational resources for a given language often contain quite different lexical material than that of the segmentation resources, the degree of their mutual transferability poses an interesting problem. There have been attempts to use derivational trees for obtaining morphological segmentation together with very coarse-grained classification (Bodnár et al., 2020). For an approach using segmented words to build derivational trees, the natural first step seems to be automated morph classification, especially root identification on the pre-segmented data (intuitively, it seems that we could build derivation trees from segmented words using morph classification combined with homonymy and allomorphy resolution); the methods used for root identification would be preferably as little supervised as possible, to minimize requirements on the resources.

The present paper starts with a brief introduction of basic terminology (Section 2) and with an overview of data sources and experiments related to our task (Section 3). Section 4 reports on our experiments with multiple root identification methods with various degrees of supervision, tested on several Indo-European languages. The results analysed in Section 5 document that surprisingly good root identification can be achieved using simple unsupervised statistical methods. Concluding remarks and some ideas for future work are sketched in Section 6.

Proceedings of the 4th Int. Workshop on Resources and Tools for Derivational Morphology (DeriMo 2023), pages 23–32,
Dubrovnik, Croatia, 5-6 October 2023.

2 Theoretical background

Although *morpheme* and *morph* are traditional notions belonging to the core linguistic terminology, their definitions vary in the literature. In the present paper, along the lines of Haspelmath (2020, p. 117), *morph* is understood as "a minimal pairing of syntacticosemantic content and a string of phonological segments" and considered as the basic unit of morphological analysis. Morphs are smaller than words (cf. three morphs in *play+er+s*), or identical with them (e.g. *chair* consisting solely of a root morph). Morphs repeat across sets of words, with certain (so-called, *cranberry*) morphs being the exception (Aronoff, 1976). As morphs are the basic building blocks in inflection and in word-formation processes, they may appear in multiple formal variants in different contexts (allomorphy); cf. the root allomorphs *sheep* and *shep* in the nouns *sheep* and *shep+herd*. *Vice versa*, a particular form can convey different meanings; cf. homonymy of both the root and the inflectional marker in the noun *bear+s* and the verb *bear+s*. In general, words are expected to be fully decomposable into morphs. In the present paper, this task is called *morphological segmentation*, but alternative names are also used (morphemic segmentation, morphemic analysis, etc.).

A *root* morph conveys lexical meaning. Other morphs, if present in the word's structure, are classified with respect to the root: the root is preceded by one or more *prefixes* (*re-* in *re+play*) and followed by one or more *suffixes* (*-er* in *play+er*); a final suffix that expresses inflectional categories (*-s* in *play+er+s*) can be distinguished by the term *ending*. In words with multiple roots (compounds), *interfixes* are often used to link the roots (*-s-* in the German noun *Arbeit+s+amt* 'employment office'). In this paper, the task of morph classification is limited to the identification of roots.

The experiments are carried out on seven languages for which morphologically segmented and annotated data are available. Despite the high quality of the data, it should be kept in mind that the segmentation recorded in the data is not always uncontroversial. It depends on the granularity of the analysis, the inclusion of etymological aspects, and other criteria. Similarly, the classification as available in the sources documents that the categories distinguished in theory are sometimes difficult to apply to authentic data. There are always cases in the data that do not fully fit either category and require a decision to be made. One such example is neoclassical formations, which are debated either as multi-root words (compounds), or single-root words where the root is preceded by a prefix(oid) or followed by a suffix(oid). Consistent decision-making is a challenge when annotating individual sources, even more so across sources from different languages. See the classification of morphemes in German verbs and other examples in the error analysis in Section 5.2.

3 Related work

3.1 Data resources

There are several relevant types of data resources, both mono- and multilingual. Instead of enumerating the resources for all the included languages individually, in the following survey we will concentrate on the unified multilingual databases. The corresponding papers usually provide a useful guide to the monolingual resources included in the given project.

First of all, there are morphological segmentation databases. These vary in quality. Some of them, like the multilingual derivational and inflectional database MorphyNet (Batsuren et al., 2021), are automatically or semi-automatically generated, so they cannot be used as gold data (at least once the accuracy of the classification methods is close to the accuracy of the provided segmentation). Universal Segmentations (UniSegments; Žabokrtský et al. 2022) is a multilingual collection of language resources containing morphological segmentation. The resources differ in several important respects, including origin (manually or automatically annotated) as well as the presence and granularity of morphological annotation.

Closely connected to (or even overlapping with) these are multilingual morphological lexicons. The largest unification effort to date, the UniMorph project (Batsuren et al., 2022b), contains in its latest release both morphological segmentation and morphological classification for at least 16 languages. Nevertheless, the segmentation is sometimes dubious or incompatible with our approach to morphological

classification. Thus, for instance, in the Czech data, lemmas of unmotivated words (represented as root nodes in derivational trees) are used instead of root morphs,[1] while in the German data, the words are segmented to morphemes (in the canonical form), not to morphs.

Finally, derivational networks, grouping words that come from the same derivational root, can be used for distinguishing root morphs and derivational affixes. Furthermore, several of these already contain morphological segmentation and classification. Universal Derivations (UDer; Kyjánek et al. 2020) is a multilingual collection of derivational resources, unified to the form of collections of derivational trees. That is, the words are organized in rooted tree structures with the edges representing the derivational relation (*child node* was derived from *parent node*). There is a relevant overlap between resources included in UDer and UniSegments, as some of the derivational resources also contain information relevant to morphological segmentation and classification.

3.2 Morphological segmentation and classification

The methods used for morphological classificaton vary according to both the quantity and quality of required data; as a rule, the more information is included in the data, the less data is needed. For languages with large and rich resources like Russian, both morphological segmentation and morphological classification can be approached using neural networks (Bolshakova and Sapin, 2021). Even for morphological classification of underresourced languages like Uspaneko (Ginn and Palmer, 2023) or Lezgi (Moeller and Hulden, 2018), neural models have been used with considerable success (around 80 or 90 % accuracy), especially given the very fine-grained tagset. It is to be noted, however, that in the case of Lezgi, where the authors performed both segmentation and classification using both neural network and CRF classifier, the CRF classifier proved to be more successful than the employed neural seq2seq network.

Even though the morph classification as such has not been much concentrated upon, it often appears as a subtask or byproduct of other tasks. Thus Goldsmith (2001) combines minimum description length with several heuristics to get candidate stems and suffixes, while Schone and Jurafsky (2001), or more recently Soricut and Och (2015) induce morphological rules using automatically extracted affixes. Strongly related to morphological classification is interlinear glossing. This task consists in finding morphological glosses (i.e. lexical meaning in the target language and/or morphological categories expressed by the morph), given a morphologically segmented text in a source language and its translation in a target language. Although the current approaches dealing with low-resource languages (Zhao et al., 2020) or CRF (McMillan-Major, 2020) yield interesting results, even there a significant amount of input data with very fine-grained annotation is needed to achieve reasonable accuracy.

4 Experiments

4.1 Data

In our choice of test languages, we were limited primarily by the quality and accessibility of morphological resources for individual languages. The quality of the segmentation resources is very important for the reliability of our results as we will obtain our test data from them. We have therefore selected the languages for which there exist manually segmented and annotated resources included in the UniSegments 1.0 project (Žabokrtský et al., 2022), and we added Czech, for which we have our own manually annotated data. Further, in some of our semi-supervised methods, we use derivational trees. As they are used more or less as a basis for heuristics, there is no need to shun automatically generated data. We have therefore used the derivational resources available in the Universal Derivations project (Kyjánek et al., 2020). The resources are listed in Table 1.

4.2 Methods

As baselines, we have used three simple statistical heuristics. Firstly, we take as roots all the longest morphs of the words (*MaxLen*). In the following methods, if not explicitly described otherwise, if two morphs gain the same score (which should happen very rarely), we pick the first of them. Secondly,

[1] Czech lemmas are rarely simplex, monomorphemic words, because even unmotativated words can contain mandatory inflectional affixes.

resource	included in	language	size	lemmas x words	morphs	root tokens	root morphs
CroDeriV	UDer, USeg	Croatian	15 657	Lemmas	65 455	15 819	4 569
MorphoLex	USeg	English	68 624	Words	151 960	77 308	20 153
Morpholex-FR	USeg	French	15 954	Words	29 087	16 290	11 085
CELEX	USeg	German	51 728	Lemmas	118 920	69 457	16 749
KuznetsEfremDict	USeg	Russian	73 447	Lemmas	318 647	86 726	6 912
DerIvaTario	UDer, USeg	Italian	10 991	Lemmas	31 246	10 991	5 566
Czech	–	Czech	10 438	Lemmas	40 155	10 438	1 985
Démonette	UDer	French	22 060				
CatVar	UDer	English	82 675				
DeriNetRU	UDer	Russian	337 632				
DeriNet	UDer	Czech	1 027 665				

Table 1: Morphological resources for segmentation and derivation used in our experiments. The Czech data we use are included neither in UDer nor in USeg; information about the structure of the data is included only for the gold segmentation data, not for the derivational tree databases

we label as root morph the morph with the fewest occurrences in the dictionary of segmented lemmas or words (*MinFreq*). This is motivated by the hypothesis that in most of the languages homomorphy between root and non-root morphs is unusual and there is only a limited number of affixes but a large number of root morphs. Thus, in the dictionary, roots will appear in conjunction with the affixes (which are few), and therefore not as often as the affixes, which will appear in conjunction with (many) roots. As our third baseline solution (*MinNeighborEntropy*), following a similar observation, namely that the root morphs predict their neighbouring morphs much better than the affixes, we compute for each morph in the dictionary the entropies of distributions of left and right neighbouring morphs. We then mark as root the morphs with the smallest maximum of the two entropies. It should be noted, however, that the last observation is not self-evident; it would not hold in cases when there is more than one compulsory suffix (or prefix) and when some of the affixes are always surrounded by other affixes. Fourthly (*UnweightedMix*), we combine the first three heuristics in an unweighted way (using the inverse value when required) and use the resulting score. Almost all the above-mentioned methods (except for *MaxLen*) are severely limited by the fact that they can select at most one root morph per word. We have therefore in our last fully unsupervised solution *ProbabMix* used normalized morph scores from the last heuristics *UnweightedMix*, obtaining a probability distribution, subsequently averaging the probabilities (of given morph being root) across the data. Then, we select as root morphs all the morphs achieving at least 5 % probability.[2]

As the second section of our experiment, we use the information contained in the UDer derivational databases. We have experimented with two approaches: Firstly, we computed the edit distance between each morph and the root of the derivational tree of the current word and all its child nodes, either by itself *DerivRoot* or in combination with the previous three unsupervised heuristics *DerivRoot* + *UnweightedMix*. Secondly, in the *LongestInDerivTree* method, for some of the languages, we used all the words in the tree to get a rough approximation of the root by finding the longest common part of the words (including a "?" wildcard to partially handle allomorphy).

Finally, for comparison with the supervised methods, we have trained a CRF tagger as implemented in the nltk package (Bird, 2006), on training data from UniSegments; that is, we treated the segmented words as sentences and the morphs as tagged words (with only roots and non-roots being distinguished as the tagset categories).

5 Evaluation

5.1 Evaluation methods

For our experiments, we have used data in Croatian (Table 2), German (Table 3), English (Table 4), Italian (Table 5), Russian (Table 6), French (Table 7), and Czech (Table 8). We have run our experiments on 5 000 randomly selected segmented words from each of the languages; for the only supervised method,

[2]The hyperparemeters were selected arbitrarily and could probably be improved, given large-enough development data; that would, nevertheless, change the setting from unsupervised to (semi-)supervised

method	word-level accuracy	average precision	average recall	average F-measure
OracleOneRoot	98.6 %	100 %	99.3 %	99.5 %
MaxLen	86.0 %	91.9 %	97.2 %	93.4 %
MinFreq	97.3 %	98.7 %	98.0 %	98.3 %
MinNeighborEntropy	97.1 %	98.5 %	97.8 %	98.0 %
UnweightedMix	96.7 %	98.1 %	97.4 %	97.7 %
ProbabMix	91.9 %	95.8 %	99.0 %	96.8 %
DerivTree	95.8 %	97.2 %	96.5 %	96.7 %
DerivTree + UnweightedMix	97.1 %	98.5 %	97.8 %	98.1 %
LongestInDerivTree	97.3 %	**98.7 %**	98.0 %	98.2 %
CRF tagger	**98.3 %**	**98.7 %**	**99.1 %**	**98.8 %**

Table 2: Croatian

method	word-level accuracy	average precision	average recall	average F-measure
OracleOneRoot	57.5 %	100 %	78.2 %	85.3 %
MaxLen	59.6 %	94.5 %	80.3 %	84.1 %
MinFreq	55.7 %	97.6 %	76.1 %	83.2 %
MaxNeighborEntropy	55.7 %	97.6 %	76.1 %	83.1 %
UnweightedMix	55.8 %	97.7 %	76.3 %	83.3 %
ProbabMix	83.4 %	97.0 %	92.7 %	93.7 %
DerivTree	55.7 %	97.7 %	76.2 %	83.2 %
DerivTree + UnweightedMix	55.9 %	**97.8 %**	73.4 %	83.4 %
CRF tagger	**92.2 %**	97.3 %	**98.0 %**	**97.1 %**

Table 3: German

method	word-level accuracy	average precision	average recall	average F-measure
OracleOneRoot	87.8 %	100 %	93.9 %	95.9 %
MaxLen	84.2 %	95.2 %	93.5 %	93.2 %
MinFreq	85.3 %	97.3 %	91.3 %	93.3 %
MinNeighborEntropy	85.4 %	97.4 %	91.4 %	93.4 %
UnweightedMix	85.6 %	97.7 %	91.6 %	93.6 %
ProbabMix	91.0 %	97.3 %	96.5 %	96.3 %
DerivTree	85.2 %	97.2 %	91.2 %	93.2 %
DerivTree + UnweightedMix	85.5 %	97.5 %	91.5 %	93.5 %
CRF tagger	**94.0 %**	**97.7 %**	**97.6 %**	**97.2 %**

Table 4: English

the CRF tagger, we have additionally selected another set of 5 000 words as training data. The sizes of the train and test sets were selected so that all the methods can be tested on the same data and (for the supervised method) the size of training data is the same for all the methods (as for the unsupervised methods, the test set is the train set). Since many of our methods only select the best candidate for the root (all apart from the *CRF Tagger*, *MaxLen* and *ProbabMix*), we have also run an oracle experiment (*OracleOneRoot*), selecting at most one root morph for each word.

We use four evaluation metrics, one on the word-level (accuracy) and three on the morph level (resp. root-level): precision, recall, and F-measure, averaged over the words (so that every word has the same weight). For the morph-level metrics, we formulate the task rather as root identification than morph classification to gain a rough error analysis. Thus, for most of the languages, for instance, precision significantly higher than recall would mean that most of the errors were false negatives; i.e. a root was identified incorrectly as a non-root.

5.2 Error analysis

In the evaluation, we take the test data at the face value. Nevertheless, it should be noted that some of the measured errors might be actually due to errors in the data. Firstly, the provided segmentation might be incorrect. For example, the German data contain the word *übersichtlich* segmented erroneously as *über+sich+tlich*; this caused wrong classification of the morph *-tlich* as root by the *MinFreq* baseline, as the erroneous morph appears very infrequently in the data. Second, the errors might be caused by (seemingly) arbitrary decisions in the morph classification in the data. For example, the German data,

method	word-level accuracy	average precision	average recall	average F-measure
OracleOneRoot	100 %	100 %	100 %	100 %
MaxLen	67.7 %	75.8 %	84.2 %	78.5 %
MinFreq	**97.5 %**	**97.5 %**	97.5 %	**97.5 %**
MinNeighborEntbropy	96.8 %	96.8 %	96.8 %	96.8 %
UnweighedMix	96.6 %	96.7 %	96.7 %	96.7 %
ProbabMix	90.8 %	94.4 %	**98.0 %**	95.6 %
DerivTree	87.7 %	87.7 %	87.7 %	87.7 %
DerivTree + UnweighedMix	96.1 %	96.1 %	96.1 %	96.1 %
CRF tagger	96.2 %	97.1 %	97.9 %	97.3 %

Table 5: Italian

Russian	Word-level accuracy	average precision	average recall	average F-measure
OracleOneRoot	82.5 %	100 %	91.1 %	94.1 %
MaxLen	60.6 %	81.2 %	85.6 %	80.4 %
MinFreq	76.4 %	93.2 %	84.7 %	87.5 %
MinNeighborEntropy	74.9 %	91.7 %	83.2 %	86.0 %
UnweightedMix	76.9 %	93.8 %	85.3 %	88.1 %
ProbabMix	80.1 %	92.0 %	94.8 %	92.0 %
DerivTree	72.3 %	88.8 %	80.5 %	83.2 %
DerivTree + UnweightedMix	78.1 %	94.9 %	86.4 %	89.2 %
CRF tagger	**90.2 %**	**96.0 %**	**95.2 %**	**95.0 %**

Table 6: Russian

method	word-level accuracy	average precision	average recall	average F-measure
OracleOneRoot	97.8 %	100 %	98.9 %	99.2 %
MaxLen	87.2 %	92.0 %	94.0 %	92.4 %
MinFreq	94.6 %	96.8 %	95.7 %	96.1 %
MinNeighborEntropy	94.7 %	96.9 %	95.8 %	96.2%
UnweightedMix	94.7 %	96.9 %	95.8 %	96.1 %
ProbabMix	92.9 %	96.5 %	**97.8 %**	**96.7 %**
DerivTree	94.6 %	96.8 %	95.7 %	96.0 %
DerivTree + UnweightedMix	**94.8 %**	**97.0 %**	95.9 %	96.2 %
LongestInDerivTree	**94.8 %**	**97.0 %**	95.9 %	96.3 %
CRF tagger	94.4 %	**97.0 %**	96.8 %	**96.7 %**

Table 7: French

method	word-level accuracy	average precision	average recall	average F-measure
OracleOneRoot	100 %	100 %	100 %	100 %
MaxLen	76.1 %	86.1 %	97.4 %	89.7 %
MinFreq	96.1 %	96.1 %	96.1 %	96.1 %
MinNeighborEntropy	96.1 %	96.1 %	96.1 %	96.1 %
UnweightedMix	97.2 %	97.2 %	97.2 %	97.2 %
ProbabMix	95.4 %	97.6 %	**99.9 %**	98.4 %
DerivTree	97.7 %	97.7 %	97.7 %	97.7 %
DerivTree + UnweightedMix	**98.6 %**	**98.6 %**	98.6 %	98.6 %
CRF tagger	97.6 %	98.5 %	99.5 %	**98.8 %**

Table 8: Czech

containing annotations like *aus+(führ)+en*, but also *(unter)+(führ)+en*,[3] do not seem to draw any clear borderline between prefixes and roots. Some of the undesirable features of the data might however also favor the systems. One of these is undersegmentation; in some cases, the words are not segmented at all, making root identification trivial. Thus, for instance, the English data contain clearly undersegmented words like *(bishopric)*, *(salsify)* or *(wringing)*.

One of the main limitations of most of the baseline solutions is the inability of the heuristics to recognize multiple root morphs in the same word; this, while not an issue for languages and word categories where compounds are scarce (like Croatian verbs) did significantly decrease the accuracy of the algorithm in languages where compounds are common (e.g. German; compare the average precision and recall; compare also with results of the oracle experiment). For example, on Czech, the best performance was

[3]The morphs labeled as roots are in brackets.

language	avg morphs per word	compounds	root-affix homomorphy	avg word-level precision	best word-level precision
Czech	3.84	0.0 %	0.1 %	94.3 %	98.6 %
German	2.48	42.5 %	1.6 %	64.3 %	92.2 %
English	2.23	12.2 %	0.6 %	87.0 %	94.0 %
French	1.83	2.2 %	0.4 %	93.5 %	94.8 %
Croatian	4.18	1.4 %	0.2 %	95.0 %	98.3 %
Italian	2.82	0.0 %	0.5 %	91.2 %	97.5 %
Russian	4.33	17.4 %	1.5 %	76.2 %	90.2 %

Table 9: Morphological complexity

method	Czech	German	English	French	Croatian	Italian	Russian
MaxLen	76.1 %	88.5 %	92.0 %	88.8 %	86.8 %	67.7 %	68.9 %
MinFreq	96.1 %	96.8 %	97.1 %	96.8 %	**98.7** %	**97.5** %	92.4 %
MinNeighborEntropy	96.1 %	96.7 %	97.2 %	96.8 %	98.5 %	96.8 %	90.6 %
UnweightedMix	97.2 %	96.9 %	**97.4** %	96.8 %	98.1 %	96.6 %	93.1 %
ProbabMix	95.4 %	93.7 %	95.4 %	94.4 %	92.3 %	90.8 %	85.2 %
DerivTree	97.7 %	96.9 %	96.9 %	96.7 %	97.2 %	87.7 %	87.5 %
DerivTree + UnweightedMix	**98.6 %**	**97.1 %**	97.3 %	**96.9** %	98.5 %	96.1 %	**94.5 %**
CRF tagger	97.6 %	93.0 %	96.3 %	96.4 %	98.4 %	96.2 %	94.3 %

Table 10: Word-level accuracy on data without compounding

achieved by *DerivTree + UnweightedMix*, which selects only one root, while for German, even the simplest baseline (*MaxLen*) able to select more than one root performed better than the oracle. Approximately the same effect, although on a much smaller scale, can be observed for English and Russian. Furthermore, for languages rich in compounding, the *ProbabMix* method performed significantly better than all the remaining non-CRF heuristics. However, if we remove the compounds from the test data, the word-level accuracy changes significantly (see Table 10). In such a setting, both the CRF tagger and *ProbabMix* are outperformed by other methods for all the languages; the best methods are then either the simple statistics (*MinFreq* or *UnweightedMix*) or *DerivTree + UnweightedMix*.

Although in most of the metrics and most of the languages, the CRF tagger yields the best results, in all but two of the languages (Czech and French) some unsupervised method is more accurate than those using derivational trees. Furthermore, the difference in performance between the heuristics and the CRF classifier is often almost negligible. Interestingly, the results do not seem to be affected by the morphological complexity of the languages, as documented by Table 9.

Another interesting question is the influence of homomorphy and allomorphy resolution. Homomorphy might affect the performance either indirectly (in the computation of the heuristics) or directly, as is the case for the ProbabMix method, which presupposes no homomorphy between roots and affixes. Allomorphy might cause errors especially for the methods using derivational trees, where the edit distance between the morphs and the root word is used. It should be noted, however, that allomorphy might be irrelevant or even work in favour of some of the methods (e.g. MinFreq). Both homomorphy and allomorphy are very hard to detect in a completely unsupervised setting, although some approaches could possibly be adapted from the comparable task of word sense disambiguation.

While we do not possess any reliable method to detect allomorphy-related errors even with the gold data, homomorphy of root and non-root morphs is easily detectable in the gold data. As listed in Table 9, the languages vary in the percentage of instances of root-affix homomorphy in the test data. A comparison of these with the percentage of homomorph misclassification (in Table 11) shows that even the indirect influence of homomorphy is considerable in the statistics – both the *UnweightedMix* and the *DerivTree + UnweightedMix* err disproportionately often in homomorph classificaton, although the disproportion is not so marked as for the *ProbabMix* and the CRF tagger. It is also noteworthy that the CRF tagger is in some cases more prone to homomorphy-related errors than the simple *ProbabMix* method (German, Russian).

method	Czech	German	English	French	Croatian	Italian	Russian
UnweightedMix	6 %	8 %	8 %	6 %	14 %	17 %	22 %
ProbabMix	16 %	23 %	18 %	15 %	49 %	24 %	39 %
DerivTree + UnweightedMix	4 %	8 %	8 %	6 %	12 %	16 %	22 %
CRF tagger	12 %	40 %	19 %	11 %	32 %	23 %	45 %

Table 11: Homomorphy-related errors

6 Conclusion

We have compared several root identification methods on seven Indo-European languages, using simple unsupervised heuristics, derivational-tree-based heuristics, and a CRF tagger. The experiments show that simple unsupervised statistical methods are sufficient for cross-linguistically highly precise root identification. While the results can be slightly improved using derivational trees, the CRF taggers, trained on a small dataset, usually achieved further improvement. The main bottlenecks of the current methods seem to be compounding, homomorphy resolution (for the CRF tagger), and potentially allomorphy resolution (for the derivational trees).

In the future, as the unsupervised heuristics proved to provide unexpectedly good results, we would like to further probe their possible combination with other methods, possibly as sources for generating data, on which a neural classifier could be pre-trained. We would also like to concentrate on (preferably, low-resource) homomorphy and allomorphy resolution, drawing inspiration from the approach by Harsha et al. (2022).

Secondly, we would like to concentrate on increasing the granularity of the classification. Given morphological lexicons for the respective languages, derivational databases could then be used similarly as in John and Žabokrtský (2023). We would also like to mine other available high-quality multilingual resources containing morphological information, notably the Universal Dependencies (Nivre et al., 2020), which contain rich morphological annotation in the form of so-called Universal Features.

7 Acknowledgements

This work has been supported by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2018101). It has been using data, tools and services provided by the LINDAT/CLARIAH-CZ Research Infrastructure.

References

Mark Aronoff. 1976. *Word Formation in Generative Grammar*. The MIT Press, Cambridge.

Khuyagbaatar Batsuren, Gábor Bella, Aryaman Arora, Viktor Martinovic, Kyle Gorman, Zdeněk Žabokrtský, Amarsanaa Ganbold, Šárka Dohnalová, Magda Ševčíková, Kateřina Pelegrinová, Fausto Giunchiglia, Ryan Cotterell, and Ekaterina Vylomova. 2022a. The SIGMORPHON 2022 Shared Task on Morpheme Segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Association for Computational Linguistics, Seattle, Washington, pages 103–116.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2021. MorphyNet: a large multilingual database of derivational and inflectional morphology. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Association for Computational Linguistics, Online, pages 39–48. https://doi.org/10.18653/v1/2021.sigmorphon-1.5.

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóǧa, Stella Markantonatou, George

Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud'hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova. 2022b. UniMorph 4.0: Universal Morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, pages 840–855. https://aclanthology.org/2022.lrec-1.89.

Steven Bird. 2006. NLTK: The Natural Language Toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*. Association for Computational Linguistics, Sydney, Australia, pages 69–72.

Jan Bodnár, Zdeněk Žabokrtský, and Magda Ševčíková. 2020. Semi-supervised induction of morpheme boundaries in Czech using a word-formation network. In *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings 23*. Springer, pages 189–196.

Elena I. Bolshakova and Alexander S. Sapin. 2021. Building a Combined Morphological Model for Russian Word Forms. In *Analysis of Images, Social Networks and Texts: 10th International Conference, AIST 2021, Tbilisi, Georgia, December 16–18, 2021, Revised Selected Papers*. Springer, Berlin, Heidelberg, page 45–55.

Michael Ginn and Alexis Palmer. 2023. Taxonomic loss for morphological glossing of low-resource languages. *arXiv preprint arXiv:2308.15055* .

John Goldsmith. 2001. Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics* 27(2):153–198. https://doi.org/10.1162/089120101750300490.

N. Sree Harsha, Ch. Nageswar Kumar, Vijaya Krishna Sonthi, and K. Amarendra. 2022. Lexical ambiguity in natural language processing applications. In *2022 International Conference on Electronics and Renewable Systems (ICEARS)*. pages 1550–1555. https://doi.org/10.1109/ICEARS53579.2022.9752297.

Martin Haspelmath. 2020. The morph as a minimal linguistic form. *Morphology* 30(2):117–134.

Vojtěch John and Zdeněk Žabokrtský. 2023. The unbearable lightness of morph classification. In Kamil Ekštein, František Pártl, and Miloslav Konopík, editors, *Text, Speech, and Dialogue*. Springer Nature Switzerland, Cham, pages 105–115.

Lukáš Kyjánek, Zdeněk Žabokrtský, Magda Ševčíková, and Jonáš Vidra. 2020. Universal Derivations 1.0, A Growing Collection of Harmonised Word-Formation Resources. *The Prague Bulletin of Mathematical Linguistics* 115:5–30.

Angelina McMillan-Major. 2020. Automating gloss generation in interlinear glossed text. In *Proceedings of the Society for Computation in Linguistics 2020*. Association for Computational Linguistics, New York, New York, pages 355–366. https://aclanthology.org/2020.scil-1.42.

Sarah Moeller and Mans Hulden. 2018. Automatic glossing in a low-resource setting for language documentation. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, pages 84–93. https://aclanthology.org/W18-4809.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, pages 4034–4043. https://aclanthology.org/2020.lrec-1.497.

Patrick Schone and Daniel Jurafsky. 2001. Knowledge-Free Induction of Inflectional Morphologies. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*. https://aclanthology.org/N01-1024.

Radu Soricut and Franz Och. 2015. Unsupervised Morphology Induction Using Word Embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Denver, Colorado, pages 1627–1637. https://doi.org/10.3115/v1/N15-1186.

Xingyuan Zhao, Satoru Ozaki, Antonios Anastasopoulos, Graham Neubig, and Lori Levin. 2020. Automatic interlinear glossing for under-resourced languages leveraging translations. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), pages 5397–5408. https://doi.org/10.18653/v1/2020.coling-main.471.

Zdeněk Žabokrtský, Niyati Bafna, Jan Bodnár, Lukáš Kyjánek, Emil Svoboda, Magda Ševčíková, and Jonáš Vidra. 2022. Towards Universal Segmentations: UniSegments 1.0. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*. ELRA, Marseille, France, pages 1137–1149.