PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 115 OCTOBER 2020 105–127

Morphological Networks for Persian and Turkish: What Can Be Induced from Morpheme Segmentation?

Hamid Haghdoost,[a] Ebrahim Ansari,[ab] Zdeněk Žabokrtský,[b]
Mahshid Nikravesh,[a] Mohammad Mahmoudi[a]

[a] Department of Computer Science and Information Technology,
Institute for Advanced Studies in Basic Sciences
[b] Institute of Formal and Applied Linguistics,
Faculty of Mathematics and Physics, Charles University

Abstract

In this work, we propose an algorithm that induces morphological networks for Persian and Turkish. The algorithm uses morpheme-segmented lexicons for the two languages. The resulting networks capture both derivational and inflectional relations. The network induction algorithm can use either manually annotated lists of roots and affixes, or simple heuristics to distinguish roots from affixes. We evaluate both variants empirically. We use our large hand-segmented set of word forms in the experiments with Persian, which is contrasted with employing only a very limited manually segmented lexicon for Turkish that existed previously. The network-induction algorithm uses gold segmentation data for initializing the networks, which are subsequently extended with additional corpus-attested word forms that were unseen in the segmented data. For this purpose, we use existing morpheme-segmentation tools, namely supervised and unsupervised version of Morfessor, and (unsupervised) MorphSyn. The experimental results show that the accuracy of segmented initial data influences derivational network quality.

1. Introduction

Even though the Natural Language community put more focus on inflectional morphology in the past, one can observe a growing interest in research on derivational morphology (and other aspects of word formation) recently, leading to the existence of various morphological data resources. One relatively novel type of resource

 Corresponding author: ansari@iasbs.ac.ir
Cite as: Hamid Haghdoost, Ebrahim Ansari, Zdeněk Žabokrtský, Mahshid Nikravesh, Mohammad Mahmoudi. Morphological Networks for Persian and Turkish: What Can Be Induced from Morpheme Segmentation?. The Prague Bulletin of Mathematical Linguistics No. 115, 2020, pp. 105–127. doi: 10.14712/00326585.007.

is word formation networks, some of which represent information about derivational morphology in the shape of a rooted tree. In such networks, the derivational relations are represented as directed edges between nodes that represent lexemes (Lango et al., 2018).

In our work, we present a procedure that builds a morphological network for Persian and Turkish using a word segmentation lexicon. The resulting network (a directed graph) represents each cluster of morphologically related word forms as a tree-shaped component of the overall graph. Thus, the specific feature of our network is that it captures both derivational and inflectional relations in a single structure (at this moment, the two types of relations are not distinguished at all). Figure 1 shows an example of such a tree for the Persian language, which represents a base morpheme meaning "to know" and all its derived and inflected descendants. In this example, the path from the root to one of the deepest leaves corresponds to the following meanings: (1) "to know", (2) "knowledge"/"science", (3) "university", (4) "a person from a university", (5) "some people from a university".

What we use as a primary source of morphological information for Persian is our manually annotated morpheme-segmented lexicon of Persian word forms, which is the only segmented lexicon for this language. At the same time, to the best of our knowledge, this lexicon containing 45,300 words could be considered as the biggest publicly available manually segmented lexicon at all (for any language). For Turkish, we use a previously existing morpheme-segmented dataset published in the Morpho Challenge 2010 Shared Task[1]. It has about 600K unsegmented words and 1000 gold standard segmented words.

Additional corpus-attested words that are not stored in the manually annotated lexicon are added into the network using automatic morpheme-segmentation methods. In order to segment new words, we used both supervised and unsupervised versions of Morfessor (Creutz et al., 2007; Grönroos et al., 2014), a popular automatic segmentation toolkit, and the MIT Arabic Segmenter (Lee et al., 2011). After performing the segmentation of unseen word forms, the process of inducing morphological relations is the same as for hand-segmented words.

The paper is organized as follows: Section 2 addresses related work on derivational morphology networks and morphological segmentation. Section 3 describes our morpheme-segmented Persian lexicon, including details on technical preprocessing and manual annotation. Section 4 describes our network construction approach. Section 5 presents experimental results and a discussion of various experiment configurations. Finally, Section 6 concludes.

[1] http://morpho.aalto.fi/events/morphochallenge2010/datasets.shtml

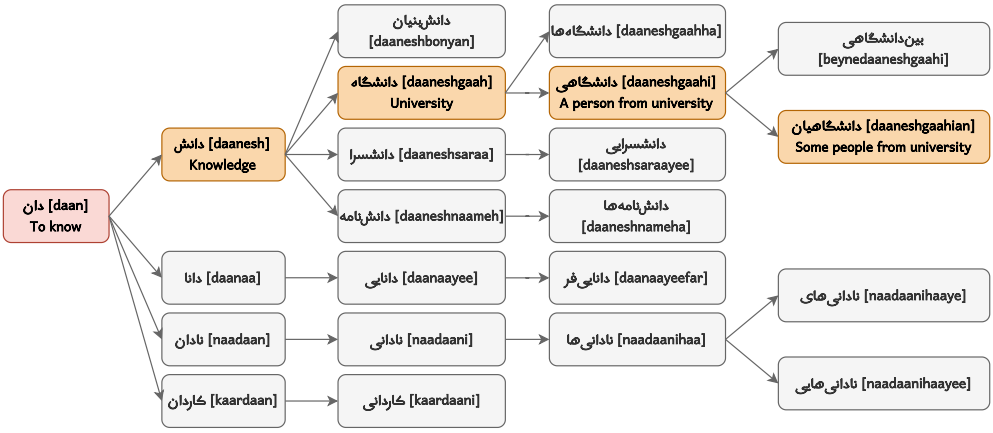

Figure 1. A sample of a Persian morphological tree for root دان [dan] which means "to know". The path from the root to one of the deepest leaf corresponds to the following meanings: (1) "to know", (2) "knowledge"/"science", (3) "university", (4) "a person from university", (5) "some people from university".

2. Related work

For some languages, intensive research exists with a focus on the construction of resources specializing in derivation. For instance, DerivBase (Zeller et al., 2013) describes a rule-based framework for inducing derivational families in German, and DErivCelex (Shafaei et al., 2017) presents an algorithm that extracts derivationally related lexicons for this language too. Hathout and Namer (2014) proposed Démonette that offers derivational morpho-semantic information for French. Šnajder (2014) presented DerivBase.Hr as a high-coverage derivational morphology resource for Croatian. Another derivational resource for Croatian is CroDeriV presented by Šojat et al. (2014) that contains data about the morphological structure and derivational relatedness of verbs. The Derinet network for Czech (Ševčíková and Žabokrtský, 2014; Žabokrtský et al., 2016) is a large linguistic resource containing over 1 million lexemes. Rafea and Shaalan (1993) presented a lexical analyzer for inflected Arabic words. For the English language, Habash and Dorr (2003) constructed and evaluated a large-scale database called CatVar, which contains categorical variations of English lexemes. Other relevant resources are (Vilares et al., 2001; Baranes and Sagot, 2014; Lango et al., 2018) for Spanish, Word Formation Latin (Litta et al., 2016), and (Piasecki et al., 2012; Kaleta, 2017; Lango et al., 2018) for Polish. Cross-linguistic research into morphological derivations is described in Kőrtvélyessy (2019). Kyjánek et al. (2019) presented

an attempt at collecting the existing derivational resources for eleven languages and at harmonizing them under a unified annotation scheme.

However, for many other languages, the data resources which provide information about derived words are scarce or lacking.

Our study is focused on Persian and Turkish. Both languages are morphologically rich languages with powerful and versatile word formation processes.

Persian is a Western Iranian language belonging to the Indo-European languages, predominantly spoken within Iran, Afghanistan and Tajikistan. Having many affixes to form new words (a few hundred), the Persian language uses derivational agglutination to form new words from nouns, adjectives, and verb stems.

Turkish is the major member of the Turkic language family, which is a subfamily of the Altaic languages. Turkish is spoken in Turkey, Cyprus, and elsewhere in Europe and the Middle East. Extensive agglutination is a prominent feature of both the Turkish language and the Persian language.

To our knowledge, research on Persian morphology is very limited. Rasooli et al. (2013) claimed that performing morphological segmentation in the pre-processing phase of statistical machine translation could improve the quality of translations for morphologically rich and complex languages. Although they segmented only an extremely limited and non-representative sample of Persian words (tens of Persian verbs), the quality of their machine translation system increases by 1.9 points of BLEU score. Arabsorkhi and Shamsfard (2006) proposed an algorithm based on Minimum Description Length with certain improvements for discovering the morphemes of the Persian language through automatic analysis of corpora. However, since no Persian segmentation lexicon was made publicly available, we decided to create a manually segmented lexicon for Persian that contains 45K words now.

For our approach, we also need automatic morpheme segmentation. The discussion about this task can be traced back to Harris (1955). Recent research on morpheme segmentation has been usually focused on unsupervised learning (Goldsmith, 2001; Creutz and Lagus, 2002; Poon et al., 2009; Narasimhan et al., 2015; Cao and Rei, 2016), whose goal is to find the segmentation boundaries using an unlabeled set of word forms (or possibly a corpus too). Probably the most popular unsupervised systems are LINGUISTICA (Goldsmith, 2001) and Morfessor, with a number of variants (Creutz and Lagus, 2002; Creutz et al., 2007; Grönroos et al., 2014); a semi-supervised extension of Morfessor was introduced by Kohonen et al. (2010). Poon et al. (2009) presented a log-linear model that uses overlapping features for unsupervised morphological segmentation. Lee et al. (2011) describe the MIT Arabic Segmenter, which uses the syntactic context of words and utilizes connections between part-of-speech categories and morphological segmentation of words. Narasimhan et al. (2015) proposed Morphochain, which is an unsupervised morphological analysis model integrating orthographic and semantic perspectives.

In our study, we use a combination of hand-annotated segmentation with segmentation generated by the supervised version of Morfessor, whose performance for our

purposes is superior to the performance of the unsupervised version, as described in Section 4.3.

3. Data

In this section we introduce the data which we used in our work. Section 3.1 describes the Persian data and Section 3.2 is a brief description of the Turkish corpus. We do not provide details about morphology of Persian and Turkish; they can be found in Jones (1807), and in Underhill (1976), respectively.

3.1. Data for Persian

We have introduced a hand-annotated segmentation data for Persian formerly (Haghdoost et al., 2019). In the following text, we describe the procedure of data creation in more detailed specification against the previous paper.

3.1.1. Corpus Collection

For compiling a set of word forms to be covered by our network, we use three Persian corpora focused on different domains.

The first source is the Persian Wikipedia (Karimi et al., 2018). The data is extracted from the Wikipedia archive that is available from the Linguatools website.[2] The files provided in the Wikipedia dataset are stored in an XML file format containing all the documents in Wikipedia for many languages, out of which we use only the Persian part. We removed XML markup and used only plain texts from the corpus.

The second source is the Bijankhan corpus (Bijankhan et al., 2011), which is a popular Persian monolingual corpus. The corpus collects daily news and other texts. The Bijankhan collection contains about 2.6 million words manually tagged with a tag set that contains 40 Persian POS tags. Again, we used only plain texts from this corpus.

The third language resource that we used is Persian-NER[3] (Poostchi et al., 2018), developed for the task of Persian named entity recognition. The resource recognizes named entities such as persons, places, and organizations.

3.1.2. Preprocessing and Tokenization

We extracted and normalized Persian sentences from all three corpora using the **Hazm** toolkit.[4] Hazm is a Python library for processing Persian text, including tokenization and lemmatization.

[2] https://linguatools.org/tools/corpora/wikipedia-monolingual-corpora/

[3] https://github.com/HaniehP/PersianNER

[4] https://github.com/sobhe/hazm

For the stemming task, we used the stemmer tool presented by Taghizadeh (Taghi-Zadeh et al., 2015). This stemmer combines cues related to orthography, corpus frequency, and syntactic distributions to induce stemming rules. It processes data in two steps. In the first step, all words of the annotated text corpus are used to automatically induce stemming rules. In the second part, the rule-based stemmer uses those stemming rules to induce words' stems. For lemmatization, we used a Persian lemma collection and the mentioned tool.

An important feature of the written form of Persian and Arabic languages is the existence of semi-space. A semi-space separates neighboring parts of a word and the separated character is narrower than a normal space. It prevents sticking morphemes. For example, word "کتابها" (books) is a combination of the word "کتاب" and "ها", in which the former is Persian translation of word "book" and the latter is the morpheme for a plural form. We can say these semi-space signs segment words into smaller morphemes. However, in formal writing and in all normal Persian corpora, this space is neglected frequently and it could make a lot of problems in Persian and Arabic morphological segmentation tasks. For example both forms for the previous example, "کتاب ها" and "کتابها", are considered correct in Persian texts and have the same meaning. In this work, all missing semi-spaces are automatically detected and corresponding words are updated accordingly.

Some words in the included corpora cannot be considered correct Persian words. To reduce the number of such words, we decided to remove words with low frequency. Words with more than 10 occurrences in the corpora were selected for manual annotation and those having less than ten occurrences were ignored in our experiments. Selected words were stored in a spreadsheet table, as illustrated in Figure 2.

3.1.3. Manual Annotation

We distributed 80K words resulting from the previous phase among our sixteen annotators in such a way that each word was annotated by two independent annotators. Annotators decided about the lemma of a word under question, segmentation points, plurality, ambiguity (whether a word has more than one meaning), being Named Entity, or they might mark the word for deletion if they think it is not a proper Persian word. Denoting the segmentation points was sufficient for generating derivational network. However, we decided to extract more information about words, because denoting the other pieces of information was not so time consuming, and they could be useful in future work. For example, denoting Named Entities could be utilized in Named Entity Recognition tasks. Our automatic segmenter tool is based on the work of Taghi-Zadeh et al. (2015), in which suffixes are stripped using rules automatically induced from a corpus. The segmenter offered a pre-segmentation (i.e., some very simple suggestions) to our annotators if it finds a word's segmentation reaching a high confidence score.

50165	X	شیدا	شیدایی	0		126	ش	ی	د	ا	ی	ی		
50166	X	شیده	شیده	0	N	17	ش	ی	د	ه				
50167	X	شیدور	شیدور	0		15	ش	ی	د	و	ر			
50168	X	شیدی	شیدی	0	N	43	ش	ی	د	ی				
50169	X	شیر	شیر	1	N	4694	ش	ی	ر					
50170	X	شیرآباد	شیرآباد	0		21	ش	ی	ر	آ	ب	ا	د	
50171	X	شیرآبه	شیرآبه	0		12	ش	ی	ر	آ	ب	ه		
50172	X	شیرآلات	شیرآلات	0		27	ش	ی	ر	آ	ل	ا	ت	
50173	X	شیرا	شیرا	0	N	23	ش	ی	ر	ا				
50174	X	شیرابه	شیرابه	0		32	ش	ی	ر	ا	ب	ه		
50175	X	شیراز	شیراز	0	N	6953	ش	ی	ر	ا	ز			
50176	X	شیرازه	شیرازه	0		56	ش	ی	ر	ا	ز	ه		

Figure 2. A snapshot of extracted data stored in a spreadsheet editor. Column 1: row ID. Column 2: distinguishing proper Persian words (“X”) from words to be deleted (“D”). Column 3: the stem of the word. Column 4: the original word form. Column 5: marking ambiguous words, 0 means “non-ambiguous” and 1 denotes “ambiguous”. Column 6: The word is a Named Entity (N) or not (empty). Column 7: frequency of the word in the source corpus. Rest: individual characters in the word form.

word	lemma	form	ambiguity	segmentation			
آزمایش (experiment)	آزمایش	X	0	آزما	یش		
آزمایشات (experiments)	آزمایش	X	0	آزما	یش	ات	
آزمایشاتی (some experiments)	آزمایش	X	0	آزما	یش	ات	ی
آسیه (Asieh (NE))	آسیه	E	0	آسیه			
دام (trap - livestock)	دام	X	1	دام			

Figure 3. A sample of hand-annotated dataset.

We removed almost words that were marked to be deleted by both annotators. The remaining 50K words (including around 12K words, for which the annotator delivered completely identical annotation) were sent for the inter-annotation disagreement resolution. In this phase, all disagreements were resolved. Finally, all words were quickly reviewed by two Persian linguists. The whole process took almost six weeks and the total number of words stored in the resulting lexicon is about 45K. Lemmas and some extra information about those words are also included.

In the final released dataset, every word is formatted as follows: Words are separated by "\n" and in each line (for each word) we have this information:

$word\ lemma\ form\ ambiguity\ segment_1\ segment_2\ ...\ segment_n$

Where "form" could be one of these:
- V: Verb
- E: Named entity word
- I: Irregular plural
- X: None of the above

The "ambiguity" field could be 0 which means the word has only one meaning and is 1 when the word has more than one meaning. Figure 3 shows a sample of final annotated data.

The resulting data resource (Ansari et al., 2019a) is publicly available under a permissive license (CC BY-NC-SA) for other researchers interested in the morpheme segmentation of Persian. Recently, we used the data for supervised morpheme segmentation task (Ansari et al., 2019b).

3.2. Data for Turkish

We have used a text corpus for Turkish that is publicly available from the Morpho Challenge 2010 event, whose aim was to find the morpheme analysis of the word forms in the data. There was a small set of gold-standard segmented data provided for semi-supervised learning of morpheme analysis, and we have used it in our supervised segmentations. In the mentioned dataset there is a list of word forms which is extracted from a text corpus and each word in the list is preceded by its frequency in the corpus used. The corpora have been preprocessed for the Morpho Challenge (tokenized, lower-cased, some conversion of character encodings).

The format of gold segmented data as well as the output of the mentioned input data for Morpho Challenge is like this: Each line of the file contains a word (e.g., "kontrole") separated from its analysis (e.g., "kontrol +DAT") by one TAB character. Morpheme labels in the analysis are separated from each other by a space character. For some words there are multiple correct analyses. These alternative analyses are separated by a comma (,). The Turkish gold-standard analyses have been obtained

from a morphological parser developed at Boğaziçi University (Sak et al., 2008). It is based on Oflazer's finite-state machines (Oflazer, 1994), with a number of changes.

4. Morphological Network Construction

In this section, the network induction based on a set of morpheme-segmented word forms is described. Subsection 4.1 introduces our algorithm developed for this task, while Subsection 4.2 describes an extension employing automatic segmentation. Subsection 4.3 describes an automatic network expansion procedure using a morphological segmenter named Morfessor and Finally in Subsection 4.4, the effect of segmentation algorithm is examined on two languages.

4.1. Automatic Network Construction

The core idea of this work is to construct a morphological network using a morpheme-segmented lexicon, be the segmentation loaded from a human-annotated lexicon, or automatically in a fully unsupervised or semi-supervised fashion.

In our proposed algorithm, first, we partition the set of word forms into subsets sharing the same root morphemes, and thus the root morpheme must be recognized among all morphemes in a given word form. We approximate the distinction between root morphemes and affixes using the number of occurrences of individual morphemes in the lexicon. After gathering the frequency counts, the m most frequent segments (we used 100 and 200 for m in our experiments) are removed from the set of potential root morphemes; all the remaining morphemes are stored in a set named $roots$. The underlying intuition is that affixes tend to repeat across many derivational clusters, and thus tend to be more frequent than root morphemes.[5] Table 1 shows an example of the most frequent segments based on our Persian segmented lexicon; all of them are classified correctly using this heuristics (i.e., all of them serve as affixes in Persian).

In the second phase, we add nodes to our morphological graph (i.e., the network contains morphological trees) based on the assembled set of root morphemes. For each r_i from the $roots$ set, we create a set of words that contain r_i. We name this set $words_i$. Now, we add r_i as a new node to our derivational graph. In the next step, we find and connect all the words in $words_i$ in the network. We divide all the words in $words_i$ into n smaller sets $words_{i,2}, words_{i,3}, ..., words_{i,n}$ based on the number of their segments. The set $words_{i,j}$ includes all words containing r_i and their number of segments is equal to j. First, we check all w in $words_{i,2}$ and if it contains a node in the tree that includes r_i, we add it to the network graph, otherwise we add w to the $remaining$ set. Then, for the next group, $words_{i,3}$, we follow a similar procedure;

[5] For simplicity of the model, we assume the boundary between root morphemes and affixes to be sharp. We do not introduce any borderline category such as affixoids.

however, we add all w in $words_{i,3}$ when it contains a node existing in $words_{i,2}$ (i.e., set of words with two segments). Then we add them to $remaining$ if there is not any subset in our current graph. We iterate this procedure until we pass all sets. Now, for each w in $remaining$ set, we check all added nodes and add w as a child of any node with the maximum number of segments. It means it would be connected to the root if there is no other option available.

Algorithm 1 shows a simple pseudocode of the segmentation graph generating procedure. The $generate$ function is recursive and gets $root$, current $tree$, remaining $words$ and current $step$ as the input parameters and returns a new $tree$ and remaining $words$. The $overlap$ function gets two words as the input and checks the left and right overlap count of the morphemes and returns the maximum of them.

For example, consider two words "understanding" with segments "understand+ing" and "misunderstanding" with segments "mis+understand+ing". The left overlap number of these words is 0 because there are not equal segments from the starting point of words but the right overlap number of them is 2 because two segments (understand and ing) are equal when we are browsing segments from the reverse side, from end to beginning. Finally, the algorithm returns the maximum number of them as a return value.

Algorithm 1 pseudocode of generating derivational graphs.

```
1:  function GENERATE(root, tree, words, n)          ▷ recursive network generation
2:      tree[root] ← root
3:      for all words do
4:          for all leaves(tree[root]) do
5:              if OVERLAP(leaf, word) > n then
6:                  setChildToLeaf(tree, leaf, word)
7:              else
8:                  appendTo(remains, word)
9:              for all leaves do
10:                 tree, remains ← GENERATE(leaf, tree, remains, n + 1)
11:     return tree, remains
12: function OVERLAP(x, y)
13:     return max(leftOverlap(x, y), rightOverlap(x, y))
14: for all segmentationSets do
15:     tree, remains ← GENERATE(root, {}, set, 1)
```

4.2. Semi-automatic Network Construction

As expected, our frequency-based identification of root morphemes vs. affixes is only an approximation; there are frequent morphemes such as [shah] "king" (clearly

not an affix) among the first 200 frequent segments. In order to quantify the influence of such wrongly classified affixes, we performed a modified version of the above-described experiment. This time, after frequency counting, we selected the m most frequent morphemes, and one annotator decided whether the morphemes are root morphemes or not (such annotation is not a time-consuming task for a human at all). The rest of the experiment remained the same. Again, we set m equal to 100 or 200.

i	seg.	freq.	i	seg.	freq.	i	seg.	freq.	i	seg.	freq.
1	ی [y]	9118	11	ایی [ee]	583	21	هم [ham]	278	31	اَست [ast]	216
2	ها [haa]	4819	12	اَل [al]	561	22	ید [id]	274	32	اَش [ash]	206
3	ه [h]	2898	13	تر [tar]	746	23	آ [aa]	274	33	دَان [daan]	198
4	آن [aan]	1708	14	آت [aat]	425	24	م [m]	267	34	شان [shaan]	193
5	می [mi]	1112	15	ب [b]	422	25	در [dar]	260	35	گاه [gaah]	192
6	یّ [yee]	941	16	ین [een]	396	26	کار [kaar]	258	36	کن [kon]	189
7	ش [sh]	891	17	ده [deh]	383	27	ساز [saaz]	254	37	پر [por]	187
8	ن [n]	864	18	شد [shod]	359	28	دو [do]	241	38	نَا [naa]	178
9	ند [nd]	782	19	دار [daar]	337	29	بر [bar]	239	39	ت [t]	173
10	د [d]	658	20	و [oo]	308	30	گر [gar]	232	40	شاه [shaah]	164

Table 1. 40 most frequent morphemes in the Persian hand-segmented lexicon, most of which are non-roots. For example, the first morpheme is the indefinite article in Persian, the second and fourth morphemes are two different plural suffixes, the third one is a suffix for female form of names, and finally, the last one is used to create the present continuous form.

4.3. Automatic Network Expansion Using Morpheme Segmentation Generated by Morfessor

Relying on the availability of manually annotated morpheme boundaries for each word in the network is clearly a bottleneck. Thus we propose an automatic procedure to expand the existing derivational network by adding selected unseen words into the graph. In other words, once the primary network based on golden annotations is ready, we try to add new words into it using the core algorithm explained in Section 4.1, just that the morpheme segmentation is produced by an automatic tool such as Morfessor. Figure 4 shows the workflow of the segmentation process.

As is shown in Figure 4, Morfessor is used in two different phases. First, in the initial data segmentation to create the primary morphological network. The second

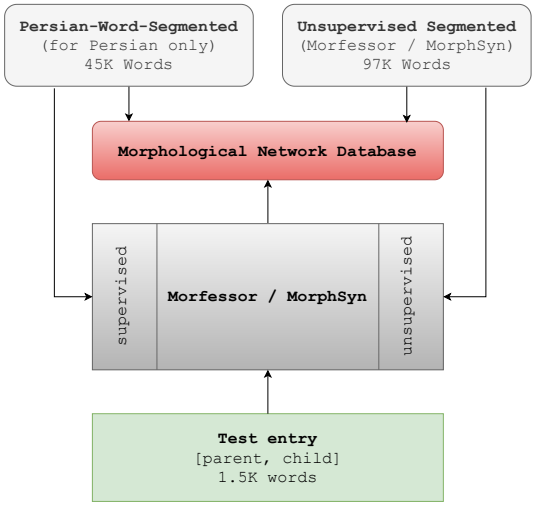

Figure 4. Morphological Network Database construction flowchart which shows the primary network construction and the expansion procedure.

part of adopting Morfessor is when we have some new words (i.e. test words) and we want to add them into our existing network and we can use Morfessor to segment them in an automatic way. In other words, in the testing phase, we have words that do not exist in our hand-annotated dataset and for creating the derivational network of morphemes we need segmentation for them too. We selected Morfessor to segment this new unseen words. Morfessor works in two ways; supervised and unsupervised: we created two models of Morfessor and in the testing phase when a new word is under question, we segment it and add it to our existing tree based on that segmentation.

In this experiment, the unsupervised model is created based on 97K words we collected from the raw text and the supervised Morfessor is trained using the 45K hand-annotated dataset. Experimental results in Section 5 show that the supervised model has better performance in comparison with the unsupervised one in the final tree accuracy.

4.4. The Effect of Segmentation Algorithm

The assumption that the accuracy of derivational networks depends substantially on morpheme segmentation quality is confirmed by another experiment in which we compared derivational networks created using outputs of two different segmentation

Data Language	Segmentation Method	Segmentation Accuracy	Network Accuracy
Persian	Morfessor	0.41	0.856
Persian	MorphSyn	0.37	0.307
Turkish	Morfessor	0.21	0.499
Turkish	MorphSyn	0.12	0.289

Table 2. Better segmentation improves derivational network accuracy.

algorithms. In order to do this, we used Morfessor[6] and MorphSyn[7]. Morfessor is a family of machine-learning methods that segment words into morphemes. More specifically, we used methods named Morfessor Baseline and Categories-MAP; both of which are based on probabilistic generative models. MorphSyn (MIT Arabic Segmenter) is a segmentation tool that uses a connection between part-of-speech categories and morphological properties. Our primary data for both Persian and Turkish was not a text corpus and the context that words occurred in was not denoted in the data. This reason led us to use the first model of MorphSyn. This model is basic and does not model the relationship between words and POS tags.

To compare our selected segmentation algorithms, we trained unsupervised models with 97,000 words of unlabeled data. The derivational network was created using these forms and then the accuracy of the derivational network was reported. For more clarity, we calculated the accuracy of morpheme segmentation and reported it. Table 2 shows the accuracy of unsupervised morphological segmentation and derivational network accuracy for both methods. The reported segmentation accuracy is based on the total word accuracy, i.e. if there was a wrong segmentation boundary on the segmented word, the whole segmentation of word considered as wrong. The segmentation accuracy calculated on a 1000 gold-standard segmented data for both languages and the results are reported. For network accuracy, we selected 400 random parent-child in the trees, then the accuracy is calculated by dividing the count of the correct parent-child in the network, by the total number of selected pairs (400). Morfessor had higher accuracy than MorphSyn as well as in derivational network accuracy that is a result of correct segmentation cases.

5. Experiments and Discussion

In order to estimate the quality of the resulting network, we randomly selected 400 nodes and checked manually if their parent nodes are identified correctly (or if the nodes are correctly marked are derivational tree roots, i.e., they are parentless). We

[6] https://Morfessor.readthedocs.io/en/latest/

[7] http://groups.csail.mit.edu/rbg/code/morphsyn/

Figure 5. Samples of generated trees using our procedures. For example in the left tree, the root of the tree is the word "Square" and the words like "Track and Field" and "Squares" are its first level children.

ran our automatic and semi-automatic versions of the algorithm using two thresholds for skipped root morphemes, 100 and 200. Table 3 summarizes the results for the individual experiment configurations. In all cases, the number of nodes in the generated graphs is 45K, which is equal to the total number of words in our manually segmented lexicon. Finally, Figure 5 shows three sample sub-graphs extracted by our algorithms.

non-root selection	number of non-roots	accuracy
automatic	100	89.5%
automatic	200	86.3%
semi-automatic	100	91.0%
semi-automatic	200	92.8%

Table 3. Accuracy for both automatic and semi-automatic methods using different numbers of non-roots in primary phase on 400 randomly selected nodes (i.e., words).

In the next experiment, we evaluated the expansion of unseen words. Table 4 shows the results of eight configurations of our experiments using Morfessor as the automatic morpheme segmentation tool. In the first half of the table, we used all available words to create out initial network and the unsupervised version of Morfessor is used for the initial segmentation. In the bottom half of Table 4, all rows show the results when the hand-annotated segmented data is used. Similarly to the previous experiment, we removed and cleaned most frequent non-root morphemes in two ways:

in automatic removing during which we ignore all first 200 frequent morphemes, and in manual removing during which the selection and removing is done by an annotator. In other words, the first two columns of this table represent the configuration of the initial tree creation. The third column of Table 4 represents the method we used for segmenting the new words and in this column. Caption "Supervised" declares we used supervised Morfessor, which is trained using 45K hand-annotated data and "Unsupervised" indicates that the segmentation is done by using a fully unsupervised version of Morfessor. For all tests in this experiment, we provided a hand-annotated morphological network with 1500 words.

init. network creation	non-root selection	test words seg.	Accuracy
97K/Segmented by Morfessor	automatic	sup. Morfessor	0.893
97K/Segmented by Morfessor	automatic	uns. Morfessor	0.777
97K/Segmented by Morfessor	manual	sup. Morfessor	0.893
97K/Segmented by Morfessor	manual	uns. Morfessor	0.777
45K Persian-Word-Segmented	automatic	sup. Morfessor	0.919
45K Persian-Word-Segmented	automatic	uns. Morfessor	0.846
45K Persian-Word-Segmented	manual	sup. Morfessor	0.934
45K Persian-Word-Segmented	manual	uns. Morfessor	0.866

Table 4. Accuracy for tree structures on 1.5K unseen words. "test word seg." column indicates the selected algorithm for unseen word segmentation

5.1. Derivation Network for the Turkish Language

Our algorithm relies fundamentally on morpheme segmentation, and the derivational network accuracy is thus directly related to the accuracy of segmentation of data. For more investigation on this claim, we ran our network generating procedure on the Turkish language.

In this experiment, we generated a derivational network from the Turkish part of Morpho Challenge 2010 (Virpioja et al., 2011) dataset.[8] It has about 600K Turkish word forms (i.e., inflected word forms), which are enough for the unsupervised segmentation tasks, and there are 1000 gold-standard segmented words. We used both supervised and unsupervised Morfessor for word segmentation and generated Turkish trees using our network creation algorithm. Table 5 shows the most frequent segments after running unsupervised Morfessor on the data; all of them are suffixes. While the majority of the first 500 most frequent words were suffixes, we decided

[8] http://morpho.aalto.fi/events/morphochallenge2010/datasets.shtml

rank	segment	freq.	rank	segment	freq.	rank	segment	freq.
1	lar	9830	11	larI	5433	21	dan	3865
2	ler	9056	12	da	5330	22	’in	3856
3	n	8779	13	li	5148	23	den	3723
4	i	8729	14	in	5126	24	la	3696
5	a	7906	15	m	5087	25	le	3652
6	e	7749	16	dir	4383	26	I	3631
7	k	6644	17	s	4165	27	larIn	3495
8	de	5982	18	lerin	4024	28	ye	3421
9	leri	5931	19	nin	4009	29	lI	3178
10	si	5456	20	ya	3981	30	lere	2919

Table 5. 30 most frequent morphemes in the Turkish segmented lexicon by unsupervised Morfessor that all of them are suffixes. For example for the 5 first ranked morphemes: “lar” and “ler” are plural suffixes, “n” is possessive suffix for 2nd person singular, “i” is possessive suffix for 3rd person singular, and “a” is equates to “to” or “towards” in English.

to ignore them in the root selection phase using two approaches: automatically and manually.

For evaluation, we created a set of 400 parent-child derivational pairs. In our experiments, we obtained 49.4% accuracy in unsupervised primary data and 47.6% correct parent-child prediction in the best configuration of supervised primary data. The complete evaluation of the Turkish network is presented in Table 6. As is shown in this table, the accuracies shown are not as good as the Persian results; we assume that the most limiting factor is the amount of hand-segmented words for Turkish, which is much smaller than that for Persian.

Primary Data	Remove Count	Network Accuracy
supervised	300	0.476
supervised	500	0.387
unsupervised	300	0.491
unsupervised	500	0.494

Table 6. Accuracy of 400 randomly selected words in the Turkish derivational networks created by supervised (trained on 1000 gold segmented words) and unsupervised (trained by 600k raw words) segmentation algorithms on primary segmented data using two thresholds for non-root removing phase.

Figure 6 shows four samples from the Turkish network created. From left to right they are eteg (Skirt), garib (Strange), kamCi (Whip) and bagaj (Luggage).

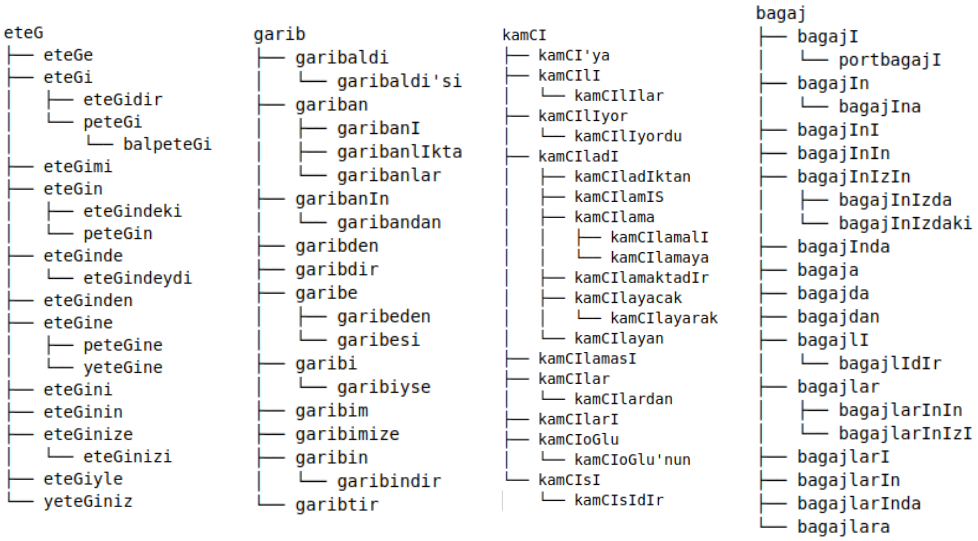

Figure 6. Examples of running our algorithm on 600K Turkish data segmented by supervised and unsupervised Morfessor. In the created trees, nodes are the words and every edge between the nodes represents morphologically relation between words. For example, in the left tree, the root word means "Skirt" and the compounds like "his / her skirt" and "in the his / her skirt" are it's children.

5.2. Error Analysis

In this section, we present an error analysis based on our observations. In the first experiment, when we created a morphological network using the hand-segmented lexicon and the whole procedure was automatic (Section 4.1), we explored two different error types. The first one appeared when we wrongly labeled a root morpheme as an affix (if it was ranked among top-frequent morphemes). For example, as can be seen in Table 1, the word "شَاه [shaah]", which means "king" and ranked 40, is a root morpheme, but we automatically labeled it as a non-root. The second common

type of error happened when our method classified a non-root morpheme as a root morpheme. For example, morpheme "ون [oon] (plural suffix)" was classified wrongly as a root morpheme by our algorithm.

In the second experiment (Section 4.2), we solved the first problem by checking the frequent morphemes manually, and as we expected, the accuracy of the result was better compared with automatic non-root selection. However, the second problem (false roots) still existed. The main reason of this problem is that there are not enough words in our segmented lexicon, and thus our algorithm is not able to identify correct parts of rare words as their root morphemes.

In our third experiment (i.e. expanding the existing graph by adding unseen words), which is described in Section 4.3, the main reason of observed errors was the wrong segmentation for some of the newly added words. It means in some cases, Morfessor offered an incorrect segmentation, which consequently led to wrong morpheme detection and wrong parent-child identification. Table 7 shows five examples of wrong segmentation of supervised and unsupervised Morfessor for our test words. Moreover, in some cases, there was not any child and parent word for test words and consequently, our algorithm could not expand the graph correctly based on them. However, this error happened very few times, while our primary graph was big enough.

The main reason for most faults in our fourth experiment was the wrong segmentation, which is a consequence of having too limited training data. Especially for supervised segmentation of Turkish data, there is simply not enough segmented data available. Also, our observations show that the segmentation boundary count in an average Turkish word is higher than in the case of Persian. Based on this observation we can say Turkish agglutination is more extensive than Persian. Besides, the observations show that the derivational trees for Turkish are in average much deeper than the Persian ones. It also could be a result of higher agglutination.

In Section 4.4, the goal was to compare two methods on two different languages. We hypothesize that the higher segmentation accuracy as well as the higher derivational network accuracy for Persian is a consequence of less extensive agglutination compared to Turkish.

6. Conclusions and future work

In this work, we developed and empirically evaluated an algorithm for creating a morphological (derivational and inflectional) network using a morpheme-segmented lexicon. Our algorithm tries to find all root candidates automatically and creates connections for all words of the lexicon. In addition, we evaluated a modification of our procedure based on a hand-validated set of non-root morphemes.

In the second part of this work, we tried to expand the morphological network by adding 1500 new words into the existing network. While this procedure is automatic, we tried to segment new test words using both supervised and unsupervised versions

word	correct segmentation	unsup. Morfessor	sup. Morfessor
آبزی [aabzi]	آب ـ زی	آبزی	آب ـ ز ـ ی
آبشُشها [aabshoshha]	آب ـ شُش ـ ها	آبشُش ـ ها	آب ـ ش ـ ش ـ ها
تعهدنامه [taahodnameh]	تعهد ـ نامه	ت ـ عهدنامه	ت ـ عهد ـ نامه
بی‌اِجازه [biejaazeh]	بی ـ اِجازه	ب ـ ی ـ اِجازه	ب ـ ی ـ اِجازه
حاکمیت [haakemiat]	حاکم ـ یت	حاکمیت	ح ـ آک ـ میت

Table 7. Sample segmentation of supervised and unsupervised Morfessor for test words in the Persian language.

of Morfessor, an automatic segmentation toolkit. These segmented morphemes are used as the input of our proposed algorithm to find the parents of new words.

We experimented both with Persian and Turkish; the derivational networks for Persian had better final accuracy, which could be a result of lower agglutination compared the Turkish language.

In addition, we evaluated and compared the usage of two unsupervised segmentation algorithms (i.e., Morfessor and MorphSyn) and experimental results showed the better segmentation leads to a more accurate network.

Acknowledgments

The research was supported by OP RDE project No. CZ.02.2.69/0.0/0.0/16_027/0008495, International Mobility of Researchers at Charles University, and by grant No. 19-14534S of the Grant Agency of the Czech Republic. It has been using language resources developed, stored, and distributed by the LINDAT/CLARIAH CZ project (LM2015071, LM2018101). The authors would like to thank all people who listed below who helped us to collect and create the dataset:

- Alireza Abdi
- Sahar Badri
- Abbas Beygi
- Shoeila Behrouznia
- Aysan Chehreh
- Matin Ebrahimkhani
- Aryan Fallah
- Fatemeh Fallah
- Seyed Amirhossein Hosseini
- Amirhossein Mafi
- Zohreh Kazemi
- Nazanin Pakdan
- Seyed Ahmad Sharifi

Bibliography

Ansari, Ebrahim, Zdeněk Žabokrtský, Hamid Haghdoost, and Mahshid Nikravesh. Persian Morphologically Segmented Lexicon 0.5, 2019a. URL https://hdl.handle.net/11234/1-3011. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Ansari, Ebrahim, Zdeněk Žabokrtský, Mohammad Mahmoudi, Hamid Haghdoost, and Jonáš Vidra. Supervised Morphological Segmentation Using Rich Annotated Lexicon. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 52–61, Varna, Bulgaria, September 2019b. INCOMA Ltd. doi: 10.26615/978-954-452-056-4_007. URL https://www.aclweb.org/anthology/R19-1007.

Arabsorkhi, Mohsen and Mehrnoush Shamsfard. Unsupervised Discovery of Persian Morphemes. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations*, EACL '06, pages 175–178, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. doi: 10.3115/1608974.1609002. URL http://dl.acm.org/citation.cfm?id=1608974.1609002.

Baranes, Marion and Benoît Sagot. A Language-independent Approach to Extracting Derivational Relations from an Inflectional Lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2793–2799, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

Bijankhan, Mahmood, Javad Sheykhzadegan, Mohammad Bahrani, and Masood Ghayoomi. Lessons from building a Persian written corpus: Peykare. *Language Resources and Evaluation*, 45(2):143–164, 2011.

Cao, Kris and Marek Rei. A Joint Model for Word Embedding and Word Morphology. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 18–26, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-1603. URL https://www.aclweb.org/anthology/W16-1603.

Creutz, Mathias and Krista Lagus. Unsupervised Discovery of Morphemes. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*, pages 21–30. Association for Computational Linguistics, July 2002. doi: 10.3115/1118647.1118650. URL https://www.aclweb.org/anthology/W02-0603.

Creutz, Mathias, Teemu Hirsimäki, Mikko Kurimo, Antti Puurula, Janne Pylkkönen, Vesa Siivola, Matti Varjokallio, Ebru Arisoy, Murat Saraçlar, and Andreas Stolcke. Morph-based Speech Recognition and Modeling of Out-of-vocabulary Words Across Languages. *ACM Trans. Speech Lang. Process.*, 5(1):3:1–3:29, December 2007. ISSN 1550-4875. doi: 10.1145/1322391.1322394. URL http://doi.acm.org/10.1145/1322391.1322394.

Goldsmith, John. Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics*, 27(2):153–198, June 2001. ISSN 0891-2017. doi: 10.1162/089120101750300490. URL https://doi.org/10.1162/089120101750300490.

Grönroos, Stig-Arne, Sami Virpioja, Peter Smit, and Mikko Kurimo. Morfessor FlatCat: An HMM-Based Method for Unsupervised and Semi-Supervised Learning of Morphology. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1177–1185, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C14-1111.

Habash, Nizar and Bonnie Dorr. A categorial variation database for English. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 17–23. Association for Computational Linguistics, 2003. doi: 10.3115/1073445.1073458.

Haghdoost, Hamid, Ebrahim Ansari, Zdeněk Žabokrtskỳ, and Mahshid Nikravesh. Building a Morphological Network for Persian on Top of a Morpheme-Segmented Lexicon. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*, pages 91–100, 2019.

Harris, Zellig. From phoneme to morpheme. *Language*, 31:209–221, 1955. doi: 10.1007/978-94-017-6059-1_2.

Hathout, Nabil and Fiammetta Namer. Démonette, a French derivational morpho-semantic network. *Linguistic Issues in Language Technology*, 11(5):125–168, 2014.

Jones, William. *A grammar of the Persian language*, volume 5. John Stockdale, 1807.

Kaleta, Zbigniew. Automatic Pairing of Perfective and Imperfective Verbs in Polish. In *Proceedings of the 8th Language and Technology Conference*, 11 2017.

Karimi, Akbar, Ebrahim Ansari, and Bahram Sadeghi Bigham. Extracting an English-Persian Parallel Corpus from Comparable Corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018.*, 2018.

Kohonen, Oskar, Sami Virpioja, Laura Leppänen, and Krista Lagus. Semi-supervised extensions to Morfessor baseline. In *Proceedings of the Morpho Challenge 2010 Workshop*, pages 30–34, 2010.

Kőrtvélyessy, Lívia. Cross-linguistic research into derivational networks. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*, pages 1–4, 2019.

Kyjánek, Lukáš, Zdeněk Žabokrtskỳ, Magda Ševčíková, and Jonáš Vidra. Universal Derivations Kickoff: A Collection of Harmonized Derivational Resources for Eleven Languages. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*, pages 101–110, 2019.

Lango, Mateusz, Magda Ševčíková, and Zdeněk Žabokrtský. Semi-Automatic Construction of Word-Formation Networks (for Polish and Spanish). In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan, May 2018. European Language Resource Association. URL `https://www.aclweb.org/anthology/L18-1291`.

Lee, Yoong Keok, Aria Haghighi, and Regina Barzilay. Modeling syntactic context improves morphological segmentation. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 1–9. Association for Computational Linguistics, 2011.

Litta, Eleonora, Marco Passarotti, and Chris Culy. Formatio formosa est. Building a Word Formation Lexicon for Latin. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016.*, 2016. URL `http://ceur-ws.org/Vol-1749/paper32.pdf`.

Narasimhan, Karthik, Regina Barzilay, and Tommi Jaakkola. An unsupervised method for uncovering morphological chains. *Transactions of the Association for Computational Linguistics*, 3:157–167, 2015. doi: 10.1162/tacl_a_00130.

Oflazer, Kemal. Two-level description of Turkish morphology. *Literary and linguistic computing*, 9(2):137–148, 1994. doi: 10.1093/llc/9.2.137.

Piasecki, Maciej, Radoslaw Ramocki, and Marek Maziarz. Recognition of Polish Derivational Relations Based on Supervised Learning Scheme. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 916–922, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA).

Poon, Hoifung, Colin Cherry, and Kristina Toutanova. Unsupervised Morphological Segmentation with Log-linear Models. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 209–217, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-41-1. doi: 10.3115/1620754.1620785. URL http://dl.acm.org/citation.cfm?id=1620754.1620785.

Poostchi, Hanieh, Ehsan Zare Borzeshi, and Massimo Piccardi. BiLSTM-CRF for Persian Named-Entity Recognition ArmanPersoNERCorpus: the First Entity-Annotated Persian Dataset. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018.*, 2018.

Rafea, Ahmed A and Khaled F Shaalan. Lexical analysis of inflected Arabic words using exhaustive search of an augmented transition network. *Software: Practice and Experience*, 23 (6):567–588, 1993. doi: 10.1002/spe.4380230602.

Rasooli, Mohammad Sadegh, Ahmed El Kholy, and Nizar Habash. Orthographic and Morphological Processing for Persian-to-English Statistical Machine Translation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1047–1051, Nagoya, Japan, October 2013. Asian Federation of Natural Language Processing. URL https://www.aclweb.org/anthology/I13-1144.

Sak, Haşim, Tunga Güngör, and Murat Saraçlar. Turkish language resources: Morphological parser, morphological disambiguator and web corpus. In *International Conference on Natural Language Processing*, pages 417–427. Springer, 2008. doi: 10.1007/978-3-540-85287-2_40.

Ševčíková, Magda and Zdeněk Žabokrtský. Word-Formation Network for Czech. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1087–1093, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

Shafaei, Elnaz, Diego Frassinelli, Gabriella Lapesa, and Sebastian Padó. DErivCELEX: Development and Evaluation of a German Derivational Morphology Lexicon based on CELEX. In *Proceedings of the DeriMo workshop*, 2017.

Šnajder, Jan. DerivBase.hr: A High-Coverage Derivational Morphology Resource for Croatian. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3371–3377, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

Šojat, Krešimir, Matea Srebačić, Tin Pavelić, and Marko Tadić. CroDeriV: a new resource for processing Croatian morphology. *Proceedings of the Language Resources and Evaluation-LREC*, 14:3366–3370, 2014.

Taghi-Zadeh, Hossein, Mohammad Hadi Sadreddini, Mohammad Hasan Diyanati, and Amir Hossein Rasekh. A new hybrid stemming method for Persian language. *Digital Scholarship in the Humanities*, 32(1):209–221, 11 2015. ISSN 2055-7671. doi: 10.1093/llc/fqv053. URL https://doi.org/10.1093/llc/fqv053.

Underhill, Robert. *Turkish grammar*. MIT press Cambridge, MA, 1976.

Vilares, Jesús, David Cabrero, and Miguel A. Alonso. Applying Productive Derivational Morphology to Term Indexing of Spanish Texts. In Gelbukh, Alexander, editor, *Computational Linguistics and Intelligent Text Processing*, pages 336–348, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg. ISBN 978-3-540-44686-6. doi: 10.1007/3-540-44686-9_34.

Virpioja, Sami, Ville T. Turunen, Sebastian Spiegler, Oskar Kohonen, and Mikko Kurimo. Empirical Comparison of Evaluation Methods for Unsupervised Learning of Morphology. *TRAITEMENT AUTOMATIQUE DES LANGUES*, 52(2):45–90, 2011. ISSN 1248-9433.

Žabokrtský, Zdeněk, Magda Ševčíková, Milan Straka, Jonáš Vidra, and Adéla Limbursk á. 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, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://www.aclweb.org/anthology/L16-1208.

Zeller, Britta, Jan Šnajder, and Sebastian Padó. DErivBase: Inducing and Evaluating a Derivational Morphology Resource for German. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1201–1211, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P13-1118.

Address for correspondence:
Ebrahim Ansari
ansari@iasbs.ac.ir
Malostranské náměstí 25, 118 00 Praha 1, Czech Republic