Automating derivational morphology for Slovenian

Tomaž Erjavec[1], Marko Pranjić[1,4], Andraž Pelicon[1], Boris Kern[2], Irena Stramljič Breznik[3], Senja Pollak[1]

[1]Jožef Stefan Institute, Jamova cesta 39, Ljubljana, Slovenia
[2]ZRC SAZU, Fran Ramovš Institute of the Slovenian Language and University of Nova Gorica, School of Humanities, Novi Trg 4, Ljubljana, Slovenia and University of Nova Gorica
[3]University of Maribor, Faculty of Arts, Koroška cesta 160, Maribor, Slovenia
[4]Jožef Stefan International Postgraduate School, Jamova cesta 39, Ljubljana, Slovenia
E-mail: {tomaz.erjavec,marko.pranjic,andraz.pelicon,senja.pollak}@ijs.si, boris.kern@zrc-sazu.si

Abstract

In this paper, we focus on computational approaches for supporting derivational word formation analysis in Slovenian. The main contributions are two-fold: first, we derive word formation rules and chains from given examples of the trail volume of a derivational dictionary and apply them to larger lexicons from two Slovenian resources; and second, we propose the first morphological segmenter for Slovenian. More specifically, based on the digitised trail volume (words starting with *b*) of the derivational dictionary of Slovenian, we extracted suffixal word-formation rules, and applied them to two lexicons of Slovenian, Sloleks and the one extracted from the metaFida corpus, to acquire new word formation instances for each chaining rule. The study of word-formation chains is relevant because it gives us an insight into word-formation mechanisms and productivity. The results show that when the derived chaining rules were applied to Sloleks, 21.95% to 31.58% of derivational chains are correct. In contrast, when the chaining rules were applied to the metaFida lexicon, the results are very noisy, with an extremely low percentage of correct chains. Next, motivated by the fact that morphological segmentation is a prerequisite for determining the structure of word formation chains and the need for more general analysis on the level of morphemes, we implemented the first automated morphological segmentation models for Slovenian. The supervised model is based on BiLSTM-CRF and achieves F1-Score of 83.98%, which is significantly higher than the two implemented unsupervised baselines, Morfessor and MorphoChain, to which we the model is compared.

Keywords: derivational morphology; word formation; automated morphological segmentation; derivational dictionary; morphological chains

1. Introduction

Word formation is a branch of linguistics which helps to analyse the lexical vitality of a given language and also shows trends of language development. Slovenian is characterised by an extremely rich morphemic structure of words, a result of multistage formation: e.g. in the first stage, the adjective *mlad/young* yields the noun *mladost/youth*, which in turn yields the adjective *mladosten/youthful* in the second stage, which yields the noun *mladostnik/adolescent*, yielding the possessive adjective *mladostnikov/adolescent's* in the fourth stage. The example shows the compatibility of four suffixal formants: *ost + -en + -ik + -ov*. The compatibility of formants is considered as the ability of different word-formational formants to co-exist within the multistage formation, taking into account the semantic extension aspect. Our paper contributes to the goal of better understanding the

characteristics of word-formation and semantic extension mechanisms in the contemporary Slovenian language, by determining the systemic predictability of formation in terms of compatibility of formants, with a focus on suffixal formants.

While there were some linguistic descriptions of Slovenian word formation (Vidovič Muha, 1988; Toporišič, 2000), including the description of formation of words in several stages that enables the linguistic investigation of multistage word-formation in Slovenian (see Breznik (2004); Kern (2010, 2020)), there is a lack of corpus-based grounding of theoretical findings. In the field of natural language processing, several researchers (Ruokolainen et al. (2013); Cotterell et al. (2015, 2019); Zundi & Avaajargal (2022); Peters & Martins (2022)) addressed the problem of morphological analysis, but there is no such study for Slovenian.

The main contributions are two-fold: first, we derive word formation rules and chains from given examples of the trail volume of the derivational dictionary BBSJB, and apply them to larger lexicons from two Slovenian resources; and second, we propose the first morphological segmenter for Slovenian. While the tasks are of different nature, they both contribute to the final goal of analysing word formation processes and their combinatorics in Slovenian. In the first case, we applying the rules derived from the existing database, and in the second one, we do not get specific rules, but get more general segmentation rules, which are less sensitive to the noisy corpora and are an underlying component of various systems for analysing word formation processes. In the derivation of word formation rules, we currently concentrated on suffix-adding rules only, as they are by far the most common in Slovenian, while in the segmentation task, the approach is more general and also other affixes are considered.

The basis for our study was the already existing Trail volume (headwords starting with the letter *b*) of the derivational dictionary of Slovenian (BBSJB) (Breznik, 2004). The dictionary gathers words in word families centred around a root, and inside those presents sequences of derivations, also split into constituent morphemes and giving the part-of-speech of the source and derived words. We leveraged BBSJB constructing morphological rules and chains (e.g., for *boj/a 'fight'* $\rightarrow$ *bojevati* $\rightarrow$ *bojevanje*). The derived rules can then be applied to infer examples from novel corpora, with the goal of comprehensive and corpus-grounded linguistic description of derivational processes, beyond the currently available dataset consisting of letter *b* headwords only. Moreover, BBSJB was leveraged for constructing a dataset for training and evaluation of morphological word segmentation, which is a prerequisite for determining the structure of word formation chains, also beyond the ones described in the rules derived from the BBSJB data. The work was performed in the scope of the project Formant Combinatorics in Slovenian.

The paper is structured as follows. After presenting related work in Section 2, Section 3 describes the resources used in our study (BBSJB, Sloleks and metaFida). Next, we present the methodology of rule-based chain extraction (Section 4.1) and morphological segmentation (Section 4.2), including two unsupervised and one supervised model. Section 5 contains the results of the rule-based chain evaluation and compares different morphological segmentation approaches and is followed by the conclusion and plans for future work (Section 6).

2. Related work

Work on automatic induction of rules for Slovenian lemmatisation has already been researched Slovenian a while ago Erjavec & Džeroski (2004), where Inductive Logic Programming was used to derived rules that compute the lemma of a word given its word-form and part-of-speech tag. This work was then followed up with approaching the same task but using so called Ripple Down Rules (Juršič et al., 2010). But while at first glance the two approaches could be also used to predict derivational rules, there is a considerable difference between inflectional and derivational morphology, as a word-form will always have a lemma, while a word will not necessarily yield a derivation, nor will a potentially derived word necessary be such, i.e. both the source and target words in a derivational process must be attested in a lexicon. It should also be noticed that there also exists an automatically derived but manually checked set of morphological rules (Arhar Holdt et al., 2020) that relate entries in the Sloleks morphological lexicon (Dobrovoljc et al., 2022). While we also also use this lexicon in our experiments, the rules themselves, again, cover only inflection, and are therefore not useful for work on derivational morphology. Rules for morphologically related words have been designed and applied to Sloleks in Čibej et al. (2020). The resource contains only word pairs, not entire chains, and automatic segmentation was performed without evaluation of the method.

Beyond the Slovenian natural language processing landscape, there are several directions. For Croatian, a closely related language, CroatianCroDeriV (Filko et al., 2019; Šojat et al., 2014) was developped, a language resource that contains data about morphological structure and derivational relatedness of verbs. Focusing on derivational processes from computational methods' perspective (see e.g. (Vylomova et al., 2017)). Evaluation of word embeddings by Gladkova et al. (2016) evaluates the processes in the scope of analogy tasks, and shows that derivational morphology is significantly more difficult to model than inflectional. Works by Lazaridou et al. (2013); Cotterell & Schütze (2018); Hofmann et al. (2020a) for example, attempt to predict a derived form given a corresponding base form. In recent research, Hofmann et al. (2020b) leverage pre-trained Neural Network Language Models and propose DagoBERT (Derivationally and generatively optimized BERT) for generation of derivationally complex words.

Morphological segmentation is a task closely related to the analysis of derivational morphology. Although the resulting segmentation does not provide explicit rules for word formation, the output of automatic morphological segmentation is a chain of morphemes. The task of morphological segmentation has generated considerable scientific attention, with several shared tasks (e.g. SIGMORPHON Batsuren et al. (2022), MorphoChallenge Kurimo et al. (2010)) being organized. For baselines in our work, we selected Morfessor and MorphoChain methods. Morfessor is a family of probabilistic machine learning methods for morphological segmentation from text data, and Morfessor 2.0 (Smit et al. (2014)), while MorphoChain Narasimhan et al. (2015) is an unsupervised model used for morphological segmentation that integrates orthographic and semantic views of words. In one of the earlier studies, Cotterell et al. (2015) designed a machine learning system for joint morphological segmentation and morpheme tagging which directly models morphotactics. Ruokolainen et al. (2013), which is also the foundation of our supervised model, addressed the task of morphological segmentation as a character-based sequence labelling task. The authors modelled the sequence labelling task with a Conditional Random Field (CRF) model. The joint BiLSTM-CRF models, introduced by Huang et al. (2015), were later

successfully used for a number of sequence tagging tasks such as part-of-speech tagging and named entity recognition. Several recent studies have achieved state-of-the-art results by using Transformer-based encoder-decoder models (Zundi & Avaajargal (2022); Peters & Martins (2022)). However, these models usually require relatively large amount of labeled data to properly converge. Further, due to large number of training parameters, such models are prohibitively expensive for training and inference in terms of computational power.

3. Resources

In this section we overview the data used in our experiments, starting with trial volume of the Derivational Dictionary of Slovenian, which was the essential resource for the study, and following with the two subsidiary resources, namely the morphological dictionary of Slovenian called Sloleks, and the metaFida corpus of Slovenian.

3.1 The Derivational Dictionary of Slovenian

The basis for our study was the already digitised Word-family dictionary of Slovenian, Trial volume for headwords beginning with letter *b*, or Besednodružinski slovar slovenskega jezika, Poskusni zvezek za iztočnice na B Stramljič Breznik (2005), BSSJB in short. The dictionary gathers words in word families centred around a root, and inside those presents sequences of derivations, also split into constituent morphemes, together with their type (e.g. suffix, prefix, compound, etc.) and giving the part-of-speech of the source and derived words. The trail volume contains 666 word families and 11,194 derivations. The trial volume was first converted from its source encoding into TEI Lex0 Romary & Tasovac (2018)[1], which is a TEI-based XML schema and guidelines for encoding dictionaries and other lexical resources, developed in the scope of the DARIAH research infrastructure.

The trial dictionary in Lex0 contains all the information from the source, and, additionally, a conversion of the part-of-speech and lexical properties of the entry from the source (Slovenian) labels to Universal Dependencies morphological features (de Marneffe et al., 2021) and to its MULTEXT-East morphosyntactic description (MSD) (Erjavec, 2012), which makes the resource better compatible with other Slovenian lexical and corpus resources. It also introduces a taxonomy of morpheme types, and links the morphemes to it.

Figure 1 gives an example of the encoding and content of a typical word family (here for the word *baba*, in this case giving only its first derived word (*babaj*, which can also be further nested, to give higher order derived words. It is thus possible to make sequences (chains) of the derivations, also giving the order number (level) of each derived word. The root word of a word-formation family (*baba* in the example) will always be level 0, while *babaj* will be 1.

To simplify processing for our experiments, we converted the TEI Lex0 format into a TSV file, which contains all the information relevant to our experiments, in particular the ID of the word family, the ID of the entry, the lemma, its level, the chain of words, of lexical properties, and morpheme types, all starting from the family root. The dictionary in this format was then used as the starting point for all further experiments.

[1] https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html

```
<entry xml:lang="sl" type="mainEntry" xml:id="bssj36337">
  <form type="lemma">
    <orth type="headword">baba</orth>
    <pron>bába</pron>
  </form>
  <gramGrp type="orig">
    <gram type="pos">sam.</gram> <gram type="gender">ž</gram></gramGrp>
  <gramGrp type="UD" xml:lang="en">
    <gram type="pos">NOUN</gram> <gram type="other">Gender=Fem</gram></gramGrp>
  <gramGrp type="MTE" xml:lang="en"><gram type="other">Nc</gram></gramGrp>
  <form type="morphemes">
    <orth><seg n="1" ana="#root1Morph">baba</seg></orth>
    <pron><seg n="1" ana="#root1Morph">bába</seg></pron>
  </form>
  <entry xml:lang="sl" type="mainEntry" xml:id="bssj45627">
    <form type="lemma">
      <orth type="headword">babaj</orth>
      <pron>babáj</pron>
    </form>
    <gramGrp type="orig">
      <gram type="pos">sam.</gram> <gram type="gender">m</gram></gramGrp>
    <gramGrp type="UD" xml:lang="en">
      <gram type="pos">NOUN</gram> <gram type="other">Gender=Masc</gram></gramGrp>
    <gramGrp type="MTE" xml:lang="en">
      <gram type="other">Ncm</gram>
    </gramGrp>
    <form type="morphemes">
      <orth><seg n="1" ana="#root1Morph">bab</seg>
      <seg n="2" ana="#suffix1Morph">aj</seg></orth>
      <pron><seg n="1" ana="#root1Morph">bab</seg>
      <seg n="2" ana="#suffix1Morph">áj</seg></pron>
    </form>
  </entry>
  ...
</entry>
```

Figure 1: Start of an example entry form BSSJB.

3.2 The Sloleks lexicon

Sloleks 2.0 Dobrovoljc et al. (2022) is a Slovenian morphological lexicon, which gives for each entry, inter alia, the lemma of the word, its complete inflectional paradigm (the word-forms paired with their MULTEXT-East morphosyntactic description and Universal Dependencies morphological features), and the frequency of occurrence of each triplet in the Gigafida reference corpus of Slovenian Krek et al. (2019). As the lexicon contains over 100,000 hand-verified lemmas, it is a large and clean resource for finding pairs of lemmas, where one is a derivation of the other.

As a preprocessing step we extracted all the lemmas from Sloleks, together with their lexical features (e.g. *Ncf* for *Noun, type="common", gender="feminine"*) and the frequency of occurrence in Gigafida. This lexicon contains 85,398 lemmas, somewhat less than Sloleks, as derivationally non-productive entries, such as numerals, abbreviations, pronouns etc. were not included.

3.3 The metaFida corpus

The metaFida 0.1 corpus (Erjavec, 2022) is the prototype edition of the collection of 34 Slovenian corpora, which are available on the concordancers of the CLARIN.SI research infrastructure, and was made so that linguists can analyse the Slovenian language using a single resource. The corpus is, by definition, the largest corpus of Slovenian, with over 3,6 billion words and so seemed as a good candidate to collect ever more lemmas than are available in Sloleks.

As each word in metaFida is marked up with its MULTEXT-East MSD, we extracted from it a lexicon identical in format to the Sloleks lemma lexicon, i.e. a list of lemmas, accompanied by their lexical features and the frequency of occurrence in metaFida; the lexicon was also filtered the in the same manner as the Sloleks one. This gave us a lexicon with 1,229,345 lemmas.

While the metaFida lexicon most likely covers the lexis of Slovenian very well, certainly much better than Sloleks, it also, as the huge number of lemmas makes obvious, contains a large portion of noise, from encoding errors and typos, to errors in automatic lemmatisation and PoS tagging. This noise is not very noticeable at the token level, but in a lexicon each mistake can produce a new lexical entry.

4. Methodology

4.1 Rule-based Chain Extraction

In our first experiment the goal was to explore the possibility of using the existing information in the trail volume of Derivational Dictionary of Slovenian (BSSJB) to induce, on the basis of the Sloleks and metaFida lexicons, entries for word families not present in BSSJB.

The method relies on the headwords, morpheme segmentation and Universal Dependencies part-of-speech labels present for each entry in BSSJB. We take pairs of entries (source and derived word), and construct pairs or rules (”deep” and ”surface” rules) that map the source word to the derived word, the former formulated as a sequence of morphemes, and the latter as regular expressions. For example, if we take the entry *boj-evati/to fight (VERB)* from which the entry *bojev-anje/figthing (NOUN)* is derived, we construct the deep rule *VERB:X-evati → NOUN:X-anje* and pair it with the surface rule describing the derivation as a minimal transformation on surface forms, in this case *”VERB:X+ti → NOUN:X+nje”*. It should be noted that we concentrate on rules that operate on suffixes only.

Such rules are then also gathered into chains, as presented in the dictionary (e.g. for *boj/a fight” (NOUN) → boj-evati → bojev-anje).* We currently concentrated on suffix rules only, with BSSJB yielding 1,641 distinct rules and 1,649 chains.

We next applied the constructed surface rules to part-of-speech / lemmas pairs from the Sloleks and metaFida lexicons. For each entry in the lexicons we try to apply the left-hand part of the regular expression of all surface rules to it, also taking into account the part-of-speech, and, if successful, construct the target word. If the target word with the correct part-of-speech also exists in the lexicon, we have found a potential derivational

pair, and can assign to it the the deep derivational rule. Once the complete lexicon has been processed, we also connect, as far as possible, the found pairs into chains.

With this, the initial set of morphological rules and chains from BSSJB and consisting only of roots starting with the letter *b* is extended to words starting with all the others letter of the alphabet (e.g. *izklic* → *izklicevati* → *izklicevanje*). Of course, not all the found pairs and chains - esp. for the metaFida lexicon - are valid derivations or derivation chains, but the derived resource could offer a good starting point for manual verification. Using the described method, we gathered 117,769 potential pairs and 32,823 potential chains from From Sloleks, while from the metaFida lexicon we get 1,549,644 potential pairs and 496,486 potential chains.

4.2 Morphological Segmentation

In this section we describe the unsupervised and supervised models used for morphological segmentation. We also present the dataset we constructed for the task of morphological segmentation based on BSSJB. We used this dataset to train and evaluate our supervised model as well as for the evaluation of unsupervised models. While in general supervised models tend to perform better, this is sensitive on the size of the training data, especially for deep learning models. Therefore, we are interested in whether unsupervised models trained on large amount of data outperform supervised models with relatively small labeled training dataset (around 10,000 examples).

4.2.1 Datasets

We have generated a gold standard dataset for morphological segmentation based on the Derivational Dictionary of Slovenian (BSSJB), more specifically on the morphological sequence chains with which we enriched the original version of the dictionary (see Section 3.1). As described in Section 4.1, the morphological chains contain only the information on the latest derivational suffix at each level. For example, the word *babeževanje* has the corresponding morphological chain *baba* → *bab-ež* → *babež-evati* → *babežev-anje*. In order to train a supervised automatic morphological segmenter, we had to preprocess the morhpological chains to obtain a gold label segmentation of the word for all the levels (e.g. *bab-ež-ev-anje*). The preprocessing was done programatically using simple rule-based approach. Since Slovene is morphoglogically complex and words frequently omit certain phonemes as they are derived, the rule-based approach produced a small amount of faulty segmentations. In order to limit the amount of noise in the training set, such examples were removed from the data, resulting in 210 words being omitted from the dataset. Some of the words present in the dictionary were reflexive verbs and were therefore recorded with a reflexive pronoun (e.g. *babiti se*). Since the reflexive pronouns themselves are not derivational morphemes we decided to remove them from the dataset during preprocessing. The resulting dataset contains 9,883 words and their gold standard morphological segmentations. This dataset was used for training the supervised approaches, and for evaluation of supervised and unsupervised segmentation models.

In addition, for unsupervised methods, we also used Sloleks and metaFida, which are described in Sections 3.2 and 3.3. The metaFida corpus was additionally cleaned up by removing all words containing characters not found in Slovene alphabet (namely, *x, y, w,*

q), all words that contain a sequence of a single character repeated successively 3 or more times, all words shorter than 3 letters, and all words occurring less than 4 times in the corpus. Due to entries in the Sloleks lexicon being manually verified, we did not do any data cleanup or preprocessing.

4.2.2 Morfessor

Morfessor is a family of probabilistic machine learning methods for morphological segmentation from text data. The underlying model is trained such that it optimizes for maximum a posteriori (MAP) estimate of models parameters Θ given the training data D:

$$\Theta_{MAP} = \arg\max_{\Theta} p(\Theta)p(D|\Theta) \tag{1}$$

During training, for each word all possible two segment combinations are evaluated. The segmentation that produces the lowest cost is selected and the same procedure is recursively applied to the resulting segments. During inference, a variation of Viterbi algorithm is used to produce the segmentation with the lowest cost. In this work, a *Morfessor 2.0* Smit et al. (2014) variant of the model was used.

We induce two Morfessor models, one using the words from Sloleks lexicon and one using words from metaFida corpus. Due to entries in the Sloleks lexicon being manually verified, we train the Morfessor model on this data in a type-based training regime that assigns equal frequency for each word in the corpus. In metaFida, we have actual count of occurrences for each word in corpus. In contrast to the approach taken with Sloleks, we train the model in log-token based training regime where number of occurrences of words are modified to use logarithm of the raw count instead. While frequency-based weighting in metaFida serves as a regulariser for the noise inherent to the dataset, we did not opt for this strategy for Sloleks, as the resource is clean and manually verified.

4.2.3 MorphoChain

Introduced in Narasimhan et al. (2015), MorphoChain is an unsupervised model used for morphological segmentation that integrates orthographic and semantic views of words. On the orthographic level, several features are used to estimate how the affixes are reused, how the words are changed when new morphemes are added to the chain and whether a sequence of morphemes exists in the corpus. Semantic comparison between words uses an additional list of word vector representations, like those produced by deep-learning models.

The model was configured by specifying letters from Slovene alphabet, by lowering the minimum morpheme size to 2, and specifying word vectors to be used for the semantic features. We use existing, publicly available, Slovene fastText word vectors described in Ljubešić & Erjavec (2018).

As in previous section, we induce one model using words from Sloleks lexicon and one using words from metaFida corpus.

4.2.4 BiLSTM-CRF - Tagging of morphological segments

Following Ruokolainen et al. (2013), we model the problem of morphological segmentation as a sequence labelling problem on the character level.[2] Each character c in a target word w is labeled with a label from the label set $y \in \{START,\ B,\ M,\ E,\ S,\ STOP\}$. From this set, a character labeled with B represents a character at the beginning of the morpheme, M represents character in the middle, and E represents the character at the end of the morpheme. A label S is used for characters that are morphemes by themselves. For example, a word *bankomat* with a ground-truth segementation *bank-o-mat* will be transformed to labels as *[B,M,M,E,S,B,M,E]*. The special labels *START* and *STOP* are added to the beginning of each word in order to constrain the model further.

Similarly to the original work, we model the sequence tagging problem with a Conditional Random Field (CRF) model and use it to train a morphological segmenter for Slovenian language. The main advantage of the CRF model is that it models the output sequence by considering dependencies between output variables. The CRF models the conditional probability of a sequence of labels $\hat{y}$ with respect to the input sequence $\hat{x}$ as follows:

$$P(\hat{y}|\hat{x}, w) = \frac{exp(\sum_t^T \hat{w} * F(y_{t-1}, y_t, \hat{x}, t))}{\sum_{y' \in Y} exp(\sum_t^T \hat{w} * F(y'_{t-1}, y'_t, \hat{x}, t))} \tag{2}$$

where t represents the position of the character in the sequence, T denotes the total length of the sequence, w denotes the parameter vector and F represents the feature function. During inference, we find the sequence of labels that maximizes the conditional probability from Equation 2 using Viterbi algorithm. We modify the original work however by employing a BiLSTM network as a feature function. The advantage of this approach is that the feature functions are not a preset mapping from words to features but are trained jointly with the CRF model. Furthermore, the use of the BiLSTM network allows us to effectively use feature information from the past n sequence steps when assigning the tag to the $n+1$-th character.

The input to our model is a word to be segmented, split into separate characters with two special *START* and *STOP* characters added to the beginning and the end of the character sequence. Each character in the sequence is then embedded into a shared embedded space $\mathbb{R}^e$ where e denotes the dimensionality of the embeddings. The embedded sequence of characters is then modelled by a BiLSTM network serving as a feature extractor which transforms the input at each step to $\mathbb{R}^h$ where h represents the hidden size of the BiLSTM. The output from the last step of the BiLSTM network is then linearly transformed to $\mathbb{R}^l$ where l is the dimensionality of the label set. Dropout is applied on the input to the linear transformation as a form of regularization. This output is then used as the emission scores for the Conditional Random Field which outputs the tag at the next step.

In our experiments, we set the embedding size e as 50 and the hidden size h of the BiLSTM as 25 while the dropout probability is set to $p = 0.2$. Training of the model is performed in batches with the batch size set to 32. For efficient computation, the lengths of the

[2] While we do not have permission for sharing the segmented data, the code for the segmentation method is public: https://gitlab.com/Andrazp/automating_derivational_morphology_for_slovenian

sequences were padded with padding tokens to the same length. The maximum size of the sequence was set to 30 characters which corresponds to the longest word in the dataset. For training, the training fold of the dataset is split into training and evaluation sets in 90%–10% proportion. During initial experiments, we have observed slow convergence of the model especially in the early stages of training. For this reason, we let the model train for 100 epochs. After each epoch, the performance of the model was evaluated on the evaluation set to prevent overfitting.

For final evaluation, we use 5-fold cross-validation, where we repeat the training procedure five times, each time evaluating on different fold of the data. We construct 5-fold cross-validation data by arranging words into folds such that all entries sharing the same root of the word are in the same fold. This is achieved by first collecting words into groups according to their root. We form a multiway number partitioning optimization problem Graham (1969) such that word groups are assigned to 5 bins in a way that minimizes differences between number of words between each bin. This optimization problem is solved using a greedy Longest-processing-time-first (LPT) algorithm[3]. In this way we ensure two important properties of our training data. One, each fold contains approximately equal number of words[4]. Two, closely related words that are derived from the same root are always assigned to the same fold. Using such constructed folds, the model is always evaluated on the words containing roots unseen during training and this enables us to test the performance of the model when applied on new words.

5. Evaluation

In this section we present the evaluation and results of rule-based chain extraction and of the machine learning-based models for morphological segmentation. Section 5.1 presents the results of chain extraction Sloleks and metaFida corpora and analyses the most common manually extracted chains. Section 5.2 presents results of the three approaches for automatic morphological segmentation of words.

5.1 Rule-based Chain Evaluation

In Table 1, we present frequencies of chain lengths on each dataset. The columns for the BSSJB dataset contains lengths from gold standard data, while Sloleks and metaFida columns contain statistics for inferred chains on each corpus. Words in the BSSJB have chains with length from 0 (root words) to 6, with the most common length being 1, ie. words composed of just a root and a single additional morpheme. Regarding Sloleks and metaFida, if the chain extraction method returned a chain with less than two rules, the resulting chain was discarded to reduce the amount of noise. For this reason, some values in the table are missing. Even if this is taken into account, there is a large discrepancy between statistics of the inferred chains and chains found in the gold standard BSSJB dictionary.

Some chains occur more often than others. In Table 2 we see ten most frequent rule chains inferred on Sloleks and metaFida, with a relative frequency of words in the corpus

[3] The implementation is available in the PRTPY library: https://github.com/erelsgl/prtpy

[4] A perfect division that assigns each fold the same number of words is not possible due to a total amount of entries in the dataset not being divisible by 5.

Length	BSSJB freq.	Sloleks freq.	metaFida freq.
0	5.92%	-	-
1	38.23%	-	-
2	34.15%	87.14%	85.75%
3	15.31%	12.49%	13.86%
4	5.23%	0.36%	0.39%
5	1.07%	0.01%	<0.01%
6	0.10%	0.00%	0.00%

Table 1: Comparison between distributions of chain lengthts. Column for the BSSJB dataset shows distribution of chain lengths on manually annotated data, while Sloleks and metaFida show distributions of inferred chains. Due to noise reduction, inferred chains with the length less than 2 were discarded, therefore the statics are not directly comparable with BSSJB.

explained by these rules. All chains show a combination of two morphemes. The most frequent chain in both corpora is *NOUN → ADJECTIVE (-en) → ADVERB (-o)*, see the following examples:

- *abeceda* 'alphabet' → *abeceden* 'alphabetical' → *abecedno* 'alphabetically'
- *čast* 'honour' → *časten* 'honourable' → *častnost* 'honourability'
- *didaktika* 'didactics' → *didaktičen* 'didactic' → *didaktično* 'didactically'

Among the most productive rules are the first-stage adjectives on *-en*, which form the base for the second-stage nouns on *-ost*, *-a* and *-ik* and the verbs on *-eti*. For example:

- *absurd* 'absurd (noun)' → *absurden* 'absurd (adjective)' → *absurdnost* 'absurdity'
- *žito* 'cereal (noun)' → *žiten* 'cereal (adjective)' → *žitnica* 'a grain silo'
- *dež* 'rain' → *dežen* 'rainy' → *dežnik* 'umbrella'
- *led* 'ice' → *leden* 'icy' → *ledeneti* 'to freeze'

In the Sloleks corpus, three chains with a non-noun simplex (non-derivative from) stand out:

- VERB:X → ADJ:X-en → NOUN:X-ost (e.g., *ganiti* 'to move (emotionally)' → *ganjen* 'moved' → *ganjenost* 'emotions from being moved'),
- ADJ:X → VERB:X-ati → NOUN:X-anje,
- ADJ:X → NOUN:X-ik → NOUN:X-ica

On the other side, in the metaFida corpus there is only one chain with a non-noun simplex:

- VERB:X → NOUN:X-0 → ADJ:X-en

In order to evaluate the accuracy of the rule chain derivations we tasked a single expert in linguistics and word formation to manually verify the correctness of the entire inferred rule

Most Common Rules	Frequency
SloLeks	
NOUN:X → ADJ:X-en → ADV:X-o	3.49%
NOUN:X → ADJ:X-en → NOUN:X-ost	2.82%
VERB:X → ADJ:X-en → NOUN:X-ost	2.70%
NOUN:X → ADJ:X-ski → ADV:X-o	2.25%
NOUN:X → VERB:X-ati → NOUN:X-anje	2.06%
ADJ:X → VERB:X-ati → NOUN:X-anje	1.86%
NOUN:X → NOUN:X-a → NOUN:X-ica	1.79%
NOUN:X → ADJ:X-en → NOUN:X-ica	1.76%
ADJ:X → NOUN:X-ik → NOUN:X-ica	1.75%
NOUN:X → ADJ:X-en → NOUN:X-ik	1.60%
metaFida	
NOUN:X → ADJ:X-en → ADV:X-o	3.37%
NOUN:X → ADJ:X-en → NOUN:X-ost	2.00%
NOUN:X → NOUN:X-0 → ADJ:X-en	1.82%
NOUN:X → ADJ:X-en → VERB:X-eti	1.72%
NOUN:X → ADJ:X-ski → ADV:X-o	1.69%
VERB:X → NOUN:X-0 → ADJ:X-en	1.61%
NOUN:X → ADJ:X-en → NOUN:X-a	1.56%
NOUN:X → NOUN:X-a → NOUN:X-ica	1.52%
NOUN:X → ADJ:X-en → NOUN:X-ik	1.38%
NOUN:X → VERB:X-ati → NOUN:X-anje	1.38%

Table 2: Ten most common rule chains inferred on SloLeks and metaFida with a relative frequency of words in the corpus explained with this rules.

chain (in future work, we plan to extend this part by conducting inter-annotator agreement experiments). For each corpus, we randomly select words and rule chain that explains the formation of selected word. The words are selected such that all words from a particular corpus in the verification dataset have distinct rule chains. The rule chains were randomly chosen with probability of being selected proportional to the logarithm of frequency of this chain occurring in the vocabulary. Due to very small number of words exhibiting longer rules, as per Table 1, we selected 100 examples for rules of size 2, 100 examples for rules of size 3, and all available examples for each rule of sizes 4 and 5. In total, this procedure selected 233 words from SloLeks and 264 words from metaFida. Next, we exclude words starting with *b*, as the examples could be identical to the ones in the BBSJB gold standard. The results of manual evaluation are presented in Table 3. The results are relatively low, especially on metaFida which is a noisy data. There are several sources of mistakes, including: semantically unrelated words (e.g., diva 'diva'→ divji 'wild'), incorrect order in word-formation chain (e.g., krsten 'baptismal' → krst 'baptism'), incorrect simplex, i.e. non-derivative form (e.g., zobati 'to nibble' (instead of zob 'tooth') → zoben 'dental').

Corpus	Chain length	Sample size	Correct	Accuracy
Sloleks	2	94	25	26.60%
Sloleks	3	82	18	21.95%
Sloleks	4	19	6	31.58%
metaFida	2	98	5	5.10%
metaFida	3	92	3	3.26%
metaFida	4	42	0	0.00%
metaFida	5	1	0	0.00%

Table 3: Results of the manual verification of rule chains inferred on Sloleks and metaFida.

5.2 Morphological Segmentation Evaluation

5.3 Evaluation metrics

In this section we present the results achieved by the inferred models on the task of morphological segmentation. For all models, we report precision, recall, F_1 score, and accuracy. We define F_1 score analogous to Ruokolainen et al. (2013). Each correctly predicted split between two morphemes in a word adds to the true positives (TP), under-splitting and over-splitting count towards false negatives (FN) and false positives (FP), respectively. As an example, for a ground truth segmentation *bank-o-mat* and a prediction *ban-ko-mat*, we have one false negative prediction (*bank•omat*, this split is not detected), one false positive (*ban•komat*, the split is added by the model but not present in the gold standard), and one true positive prediction (*banko•mat*).

The F_1 score is then defined as follows:

$$F_1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{3}$$

We define accuracy of the model as the fraction of words with completely correct segmentation, or alternatively, as a probability estimate of the model returning a correct segmentation. Although this metric is usually not used in semantic segmentation task (cf. Batsuren et al. 2022, Narasimhan et al. 2015), we consider it highly relevant and intuitive for model comparison.

The results for unsupervised methods are provided on the entire dataset (see Section 4.2.1), while for the supervised method, where 5-fold cross-validation was used, the results are presented as an average score across all training runs, together with a standard deviation between them.

5.4 Evaluation results

We present the results in the Table 4.

Among the unsupervised baseline models, there is a consistent difference in F_1 score when model is inferred on metaFida vs. on the Sloleks corpus. Although the metaFida corpus

Model	Precision	Recall	F_1 score	Accuracy
Morfessor 2.0 (mFida)	63.99%	22.64%	33.45%	15.97%
Morfessor 2.0 (Sloleks)	40.53%	34.19%	37.09%	13.90%
MorphoChain (mFida)	62.42%	23.88%	34.54%	15.33%
MorphoChain (Sloleks)	63.33%	34.86%	44.97%	20.90%
BiLSTM-CRF	**83.45%** (±0.9)	**84.58%** (±2.7)	**83.98%** (±1.2)	**47.73%** (±1.7)

Table 4: Results on the inferred models on the task of morphological segmentation.

used in training is significantly larger (x6.5), the F_1 score is consistently improved with Sloleks corpus, especially the recall component of the metric. The MorphoChain model consistently outperforms the Morfessor 2.0 model on F_1 score. This is to be expected as MorphoChain model also includes semantic information when resolving the morphemes of the word.

The supervised BiLSTM-CRF approach shows the strongest performance on our dataset. All metrics show consistent performance over the folds as indicated by very low standard deviations which shows the model is not sensitive to the variability in the training data. An advantage of this model is that even though it is supervised it can be effectively trained on smaller amounts of labeled data.

6. Conclusion and future work

With this work we tackle the problem of automating the derivational morphology for Slovenian language with two complementary approaches. With one approach, we induce a model on annotated data of derivational dictionary and produce rules that explain transformation from a base word to a derived one. With the other approach, we induce a model for morphological segmentation and evaluate it on the derivational dictionary.

Although the extraction of rule chains provides a richer information about word formation, the accuracy of our approach is not satisfactory when evaluated on a random selection of words from Sloleks lexicon and metaFida corpus. Results on the metaFida corpus are significantly worse than those inferred on the Sloleks lexicon. One explanation for this is the amount of noise present in each dataset. Entries in the Sloleks lexicon were manually verified, which is not the case for metaFida corpus. This opens up a topic to be explored in future work, how to improve the rule-based chain extraction by incorporating the probabilistic estimates derived from the word frequencies, or even the semantic similarity of words as used in MorphoChain (Narasimhan et al. (2015)).

Morphological segmentation was explored by evaluating both unsupervised and supervised models, and evaluated on a dataset constructed from the derivational dictionary. Unsupervised models were induced on both Sloleks lexicon and metaFida corpus, while the supervised model was induced and evaluated on the constructed dataset using the 5-fold cross-validation. All unsupervised approaches have very low values of F_1 score and accuracy, but those results are comparable with results reported in related work (cf. Batsuren et al. (2022)).

The supervised approach based on the BiLSTM-CRF model achieves higher scores compared to the unsupervised approaches which is to be expected as it is trained on the BSSJB dataset with supervision. While care has been taken to prevent the model from overfitting on the root of the word and capitalizing on this during evaluation, the model is able to learn better patterns as the training and test set come from the same data distribution.

For future work, we will evaluate the BiLSTM-CRF model on other out-of-distribution datasets to fully gauge its performance in a practical setting. Furthermore, the current training and test data contain only words starting on letter *b*. While we assume the rules for morphological derivation are general across the vocabulary of a language, we would like to test the model on a more varied vocabulary to gauge the impact of this particular bias of our dataset. We also plan to leverage automated morphological segmentation for deriving novel rules from the actual corpora, which will enable to analyse word formation processes and formant combinatorics beyond the rules described in the BBSJB trial data. The developed methods have high potential for faster and corpus driven approaches to creation of contemporary derivational dictionaries.

7. Acknowledgements

This article was written in the framework of the project Formant Combinatorics in Slovenian (J6-3134) funded by the Slovenian Research Agency (ARRS). We also acknowledge the ARRS funding through the core programmes Knowledge Technologies (P2-0103) and The Slovenian Language in Synchronic and Diachronic Development (P6-0038).

8. References

Arhar Holdt, Š., Čibej, J., Laskowski, C. & Krek, S. (2020). *Morphological patterns from the Sloleks 2.0 lexicon 1.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1411.

Batsuren, K., Bella, G., Arora, A., Martinovic, V., Gorman, K., Žabokrtský, Z., Ganbold, A., Dohnalová, Š., Ševčíková, M., Pelegrinová, K., Giunchiglia, F., Cotterell, R. & Vylomova, E. (2022). The SIGMORPHON 2022 Shared Task on Morpheme Segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Seattle, Washington: Association for Computational Linguistics, pp. 103–116. URL https://aclanthology.org/2022.sigmorphon-1.11.

Breznik, I.S. (2004). *Besednodružinski slovar slovenskega jezika, Poskusni zvezek za iztočnice na B (Word-family dictionary of Slovenian, Trial volume for headwords beginning with letter B).* Maribor: Slavistično društvo.

Čibej, J., Arhar Holdt, Š. & Krek, S. (2020). List of word relations from the Sloleks 2.0 lexicon 1.0. URL http://hdl.handle.net/11356/1386. Slovenian language resource repository CLARIN.SI.

Cotterell, R., Kirov, C., Hulden, M. & Eisner, J. (2019). On the Complexity and Typology of Inflectional Morphological Systems. *Transactions of the Association for Computational Linguistics*, 7, pp. 327–342. URL https://doi.org/10.1162/tacl_a_00271. https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00271/1923163/tacl_a_00271.pdf.

Cotterell, R., Müller, T., Fraser, A. & Schütze, H. (2015). Labeled Morphological Segmentation with Semi-Markov Models. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*. Beijing, China: Association for Computational Linguistics, pp. 164–174. URL https://aclanthology.org/K15-1017.

Cotterell, R. & Schütze, H. (2018). Joint semantic synthesis and morphological analysis of the derived word. *Transactions of the Association for Computational Linguistics*, 6, pp. 33–48.

de Marneffe, M.C., Manning, C.D., Nivre, J. & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2), pp. 255–308. URL https://doi.org/10.1162/coli_a_00402. https://direct.mit.edu/coli/article-pdf/47/2/255/1938138/coli_a_00402.pdf.

Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T., Romih, M., Arhar Holdt, Š., Čibej, J., Krsnik, L. & Robnik-Šikonja, M. (2022). *Morphological lexicon Sloleks 2.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1230.

Erjavec, T. (2012). MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages. *Language Resources and Evaluation*, 46(1), pp. 131–142. URL https://doi.org/10.1007/s10579-011-9174-8.

Erjavec, T. (2022). *Corpus of combined Slovenian corpora metaFida 0.1*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1746.

Erjavec, T. & Džeroski, S. (2004). Machine learning of morphosyntactic structure: lemmatizing unknown Slovene words. *Applied artificial intelligence*, 18, p. 17–41.

Filko, M., Šojat, K. & Štefanec, V. (2019). Redesign of the Croatian derivational lexicon. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*. Prague, Czechia: Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, pp. 71–80. URL https://aclanthology.org/W19-8509.

Gladkova, A., Drozd, A. & Matsuoka, S. (2016). Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*. pp. 8–15.

Graham, R.L. (1969). Bounds on Multiprocessing Timing Anomalies. *SIAM Journal on Applied Mathematics*, 17(2), pp. 416–429. URL https://doi.org/10.1137/0117039. https://doi.org/10.1137/0117039.

Hofmann, V., Pierrehumbert, J. & Schütze, H. (2020a). Predicting the growth of morphological families from social and linguistic factors. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. ACL.

Hofmann, V., Pierrehumbert, J.B. & Schütze, H. (2020b). DagoBERT: Generating derivational morphology with a pretrained language model. *arXiv preprint arXiv:2005.00672*.

Huang, Z., Xu, W. & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Juršič, M., Mozetič, I., Erjavec, T. & Lavrač, N. (2010). LemmaGen: multilingual lemmatisation with induced Ripple-Down rules. *Journal of universal computer science*, 16(9), p. 1190–1214.

Kern, B. (2010). Stopenjsko besedotvorje. *Slavistična revija*, 58, p. 35–348.

Kern, B. (2020). Obrazilna kombinatorika v besedotvornih sestavih glagolov čutnega zaznavanja, p. 67–79.

Krek, S., Erjavec, T., Repar, A., Čibej, J., Arhar Holdt, Š., Gantar, P., Kosem, I., Robnik-Šikonja, M., Ljubešić, N., Dobrovoljc, K., Laskowski, C., Grčar, M., Holozan, P., Šuster, S., Gorjanc, V., Stabej, M. & Logar, N. (2019). *Corpus of Written Standard Slovene Gigafida 2.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1320.

Kurimo, M., Virpioja, S., Turunen, V. & Lagus, K. (2010). Morpho Challenge 2005-2010: Evaluations and Results. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*. Uppsala, Sweden: Association for Computational Linguistics, pp. 87–95. URL https://aclanthology.org/W10-2211.

Lazaridou, A., Marelli, M., Zamparelli, R. & Baroni, M. (2013). Compositional-ly derived representations of morphologically complex words in distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pp. 1517–1526.

Ljubešić, N. & Erjavec, T. (2018). Word embeddings CLARIN.SI-embed.sl 1.0. URL http://hdl.handle.net/11356/1204. Slovenian language resource repository CLARIN.SI.

Narasimhan, K., Barzilay, R. & Jaakkola, T. (2015). An Unsupervised Method for Uncovering Morphological Chains. *Transactions of the Association for Computational Linguistics*, 3, pp. 157–167. URL https://aclanthology.org/Q15-1012.

Peters, B. & Martins, A.F.T. (2022). Beyond Characters: Subword-level Morpheme Segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Seattle, Washington: Association for Computational Linguistics, pp. 131–138. URL https://aclanthology.org/2022.sigmorphon-1.14.

Romary, L. & Tasovac, T. (2018). TEI Lex-0: A Target Format for TEI-Encoded Dictionaries and Lexical Resources. In *Conference of the Text Encoding Initiative: TEI as a Global Language*. Tokyo. https://doi.org/10.5281/zenodo.2613594.

Ruokolainen, T., Kohonen, O., Virpioja, S. & Kurimo, M. (2013). Supervised Morphological Segmentation in a Low-Resource Learning Setting using Conditional Random Fields. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 29–37. URL https://aclanthology.org/W13-3504.

Smit, P., Virpioja, S., Grönroos, S.A. & Kurimo, M. (2014). Morfessor 2.0: Toolkit for statistical morphological segmentation. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden: Association for Computational Linguistics, pp. 21–24. URL https://aclanthology.org/E14-2006.

Šojat, K., Srebačić, M., Tadić, M. & Pavelić, T. (2014). CroDeriV: a new resource for processing Croatian morphology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 3366–3370. URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/1074_Paper.pdf.

Stramljič Breznik, I. (2005). Kvantitativne lastnosti slovenskega tvorjenega besedja v poskusnem besednodružinskem slovarju za črko B. *Slavistična revija*, 53(4), p. 505–520. URL http://www.dlib.si/details/URN:NBN:SI:DOC-EWKNYRGH. Bibliografija: str. 518-520 Summary.

Toporišič, J. (2000). *Slovenska slovnica*. Maribor: Obzorja.

Vidovič Muha, A. (1988). *Slovensko skladenjsko besedotvorje ob primerih zloženk*. Ljubljana: Znanstvena založba Filozofske fakultete, Partizanska knjiga.

Vylomova, E., Cotterell, R., Baldwin, T. & Cohn, T. (2017). Context-aware prediction of derivational word-forms. *arXiv preprint arXiv:1702.06675*.

Zundi, T. & Avaajargal, C. (2022). Word-level Morpheme segmentation using Transformer neural network. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Seattle, Washington: Association for Computational Linguistics, pp. 139–143. URL https://aclanthology.org/2022.sigmorphon-1.15.