Marko Pranjić
Senja Pollak

ADVANCEMENTS IN AUTOMATIC MORPHOLOGICAL SEGMENTATION FOR SLOVENIAN[1]

1 Introduction

The computational task of morphological word segmentation aims to break–up a given word into its morphemes (roots and affixes). For example, a word *'stopped'* is composed from two morphemes, *'stop'* and *'ed'*. There is a distinction between canonical and surface-level morphological segmentation where canonical segmentation aims to recover morphemes *'stop'* and *'ed'*, as above, and surface level segmentation aims to recover their orthographic representation from the target word, *'stopp'* and *'ed'* in the above case. Although canonical morphemes are very interesting from the linguistic point of view, the focus of this paper is on the latter, surface level (concatenative) morphemes, which has high relevance for computational processing and representations of textual data, and is related to tasks such as tokenization and stemming.

The initial step in any computational task involving text is representing textual data with numbers. Each part of the text represented as a single number is called a token, and the process to encode the text to tokens is called tokenization. The simplest approach is to represent each textual character (grapheme)

[1] We thank Irena Stramljič Breznik for allowing us access to the Word-family dictionary of Slovenian, Trial volume for headwords beginning with a letter B (BSSJB; Besednodružinski slovar slovenskega jezika, Poskusni zvezek za iztočnice na B), used as a source of the data in this study. The work was supported by the Slovenian Research and Innovation Agency core research programme Knowledge Technologies (P2-0103), as well as the project Formant combinatorics in Slovenian (J6-3131).

with a number. This approach is rarely used as the performance of the computational task is lower compared to using other approaches. Modern language models involving deep-learning neural networks replace a sequence of characters with a single number. Each sequence is chosen such that the characters from the sequence are commonly found together. For example, a word *discard* is in some models split into *disc* and *ard* where each of those parts is statistically common to be encountered as part of a word. In the process of training the model, with the help of a large amount of text data, the model assigns information to each of those tokens. The above described approach does not use information of the morphemes, but recent research indicates that including morphological information to the tokenization process has the potential to improve the fundamental abilities of large language models to represent text (Knigawka 2022, Hofmann et al. 2021, El-Kishky et al. 2019), improvements already demonstrated in machine translation (Dyer et al. 2008, Huck et al. 2017), dependency parsing (Seeker et al. 2015), and speech recognition (Creutz et al. 2007).

The MorphoChallenge (Kurimo et al. 2010) competition (2005–2010) supported the development of several surface-level morphological segmenters (Creutz et al. 2002, Creutz et al. 2005). Those systems leverage large amounts of textual data in order to determine statistical patterns within words and implicitly deduce morphemes. When explicit information on morphemes is available, a standard approach to morphological segmentation is the BiLSTM–CRF (Huang et al. 2015) model. Large language models (LLMs), pretrained on large amounts of textual data, have been successfully used in most NLP tasks reaching superior results to previous approaches, yet their application to morphological segmentation has been limited. To our knowledge, the first such application is the LLMSegm (Pranjić et al. 2024) that views morphological segmentation as a task of predicting whether a single position within the word corresponds to the boundary between morphemes. The input to the model includes a hypothesis for such a boundary and the model is trained to distinguish between correct and incorrect boundaries. The LLMSegm was evaluated in multiple languages and in most cases has been shown to improve over competing approaches. As the Slovenian language is considered a less–resourced language, with a limited amount of data available to train large pretrained language models, LLMSegm is especially important as it promises improved performance in this setting. Until now, the performance of LLMSegm was evaluated on English, Finnish,

Turkish, as well as in low–resource setting on South African languages. While in the low–resource setting, each language has more than 15,000 morphologically annotated words, in the Slovenian language, such data is even more scarce with the only dataset reported being only 9,883 words (Erjavec et al. 2023).

With this paper, we extend our previous work performed on morphological segmentation for Slovenian (Erjavec et al. 2023) and combine it with our recently proposed LLMSegm method (Pranjić et al. 2024) for morphological segmentation. We use a derivational dictionary of Slovenian, Trial volume for headwords beginning with a letter B (BSSJB; *Besednodružinski slovar slovenskega jezika, Poskusni zvezek za iztočnice na B*) (Stramljič Breznik, 2004) for training and evaluation, and compare how different pretrained language models (monolingual and multilingual) influence the LLMSegm's performance.

The paper is organized into five sections. In Section 1 we introduce the paper and briefly describe related work. Section 2 introduces the data used to evaluate the morphological segmentation in Slovenian. Section 3 describes the LLMSegm approach and model variants used to evaluate the morphological segmentation in Slovenian. We present the experimental results on Slovenian in Section 4. Conclusions and ideas for further work are discussed in Section 5.

2 Slovene morphological segmentation datasets

Morphological segmentation approaches can be divided into supervised and unsupervised methods. Supervised methods require annotated data, in this case, words together with their morphological segmentation in order to induce a model to segment new unseen words. In contrast, unsupervised methods work with any text corpus by leveraging statistical patterns found in the text. In Slovenian, several resources can be used for unsupervised models, like metaFida (Erjavec 2022) corpus or SloLex dictionary (Dobrovoljc et al. 2022). However, this line of work does not achieve good results on the morphological segmentation task in Slovenian. The best performing model, using the MorphoChain approach together with the SloLex lexicon, correctly segments only one in five words (Erjavec et al. 2023).

For supervised approaches, availability of the adequate data is the main concern due to high effort required to annotate the dataset. Datasets for

morphological segmentation published for the SIGMORPHON 2022 Shared Task on Morpheme Segmentation (Batsuren et al. 2022) contain over 500,000 words and their morphemes and are publicly available for many of the languages. Unfortunately, not for Slovenian where the situation is much different as there is, to our knowledge, not a single publicly available dataset that can be used for this task. Nonetheless, some experiments with access to adequate data studied and evaluated the morphological segmentation in Slovenian (Erjavec et al. 2023). This dataset was based on the BSSJB dictionary (Stramljič Breznik 2004) that lists words with their root and derived words together with prefixes and suffixes for a total of 11,159 words. For example, the Slovene word *babeževanje* (flirting) has the corresponding entry *baba* → *bab–ež* → *babež–evati* → *babežev–anje* where each step of the chain presents an addition of a morpheme to the root word. Although this resource reflects the word–formation of the target word instead of morphology, it was still used due to similarities to the morphological segmentation task.

In order to prepare this data for the supervised task of morphological segmentation, some simple rule–based preprocessing was performed. In the above example, the resulting morphological segmentation would be *bab–ež–ev–anje*. Together with skipping reflexive pronouns, some words were left out due to limits of such rule–based processing, bringing the size of the final dataset down to 9,883 words (Erjavec et al. 2023). We will use this data to evaluate the LLMSegm approach on the task of morphological segmentation in Slovenian.[2]

3 LLMSegm morphological segmentation approach

The LLMSegm is an approach with publicly available implementation that models the morphological segmentation task as a binary classification for morpheme boundary prediction. A potential morpheme boundary refers to the position between characters in the word, which may or may not correspond to the actual morpheme boundary. By "actual boundary" we refer to the surface level boundary, where one morpheme in the word ends and the following one begins. Based on the training data, the model learns to predict if the potential

[2] We thank Irena Stramljič Breznik, author of BSSJB, for allowing us access to the data.

boundary corresponds to the actual morpheme boundary. The input to the model includes the target word with a separator token placed inside the word at some position, and is required to guess the label of ✓ or ✗ for actual morpheme boundary value. The model then learns to predict the value for each potential morpheme boundary.

We can illustrate the model's prediction with a previously discussed example of a Slovene word *babeževanje*, see Table 1. For example, the potential morpheme boundary after the third character (shown with the • symbol) would be represented as model's input *bab•eževanje*. The model should then predict either ✓ for the morpheme boundary at that place or ✗ in case of disagreeing with the actual morpheme boundary at this position. In this case, the correct prediction would be ✓ as *babeževanje* is correctly segmented to *bab-ež-ev-anje* and the third position represents one of the three correct morpheme boundaries. In LLMSegm, the model is applied multiple times[3], i.e. for every position in the target word the potential boundary is classified as an actual boundary or not. As the final step, all of the morpheme boundaries are joined to form word segments as the final output. For additional details taken into account in the LLMSegm approach, see (Pranjić et al. 2024).

Word	Input	Prediction		Boundaries	Output
babeževanje	b•abeževanje	✗			bab•ež•ev•anje
	ba•beževanje	✗			
	bab•eževanje	✓	→	bab•eževanje	
	babe•ževanje	✗			
	babež•evanje	✓	→	babež•evanje	
	babeže•vanje	✗			
	babežev•anje	✓	→	babežev•anje	
	babeževa•nje	✗			
	babeževan•je	✗			
	babeževanj•e	✗			

Table 1: The steps performed by the LLMSegm approach in morphological segmentation of words

[3] Although this raises the computational requirements several–fold, the model can be run on the GPU and all such positions queried within a single batch thus amortizing the cost to a single query of the model.

3.1 Models

In this section, we introduce several models that are used in conjunction with the LLMSegm in our empirical evaluation. Each pretrained model requires its own specific tokenization that was used to prepare the data during training. The performance of the model can be significantly influenced by the tokenization as tokenized input that is aligned with morphology already introduces a lot of information regarding the segmentation and the task of the model is simpler, however the research shows (Hou et al. 2023) that the statistical tokenizers do not correspond to morphological segmentation. Nevertheless, in order to measure the impact of the model and its training, we also report the metrics on the tokenizer output itself.

We consider three pretrained language models (which are compared in LLMSegm training, as well as baseline tokenizers) as well as a BiLSTM– classifier, which was the current state-of-the-art method for Slovene morpheme segmentation:

- Glot500 (ImaniGooghari et al. 2023): The work that introduced LLMSegm primarily used *Glot500–m* large language model for the segmentation task. This is a multilingual model supporting over 500 languages.
- CSEBert (Ulčar et al. 2020): A trilingual model trained on Croatian, Slovenian, and English corpora. Focusing on three languages, the model performs better than multilingual BERT, while still offering an option for cross-lingual knowledge transfer.
- SloBERTa (Ulčar et al. 2021): The monolingual Slovene RoBERTa model is the state–of–the–art model for context dependent text modelling in Slovenian.
- BiLSTM–CRF (Huang et al. 2015): For comparison with previously reported results on Slovenian, we will include the best results reported in (Erjavec et al. 2023). Those results were achieved with the bidirectional Long Short–Term Memory (BiLSTM) neural network coupled with the conditional random fields (CRF) model.

For training LLMSegm, we fine–tune each model by using the hyperparameters outlined in (Pranjić et al. 2024): AdamW optimizer (Loshchilov et al. 2017) using batches of 256 examples, linear learning rate schedule with 20 warm-up iterations to a learning rate of 2e-5, and a dropout of 0.01, for a total of 30 epochs.

3.2 Evaluation measures

The standard metric to evaluate performance of the morphological segmentation models is boundary precision-recall (BPR) metric (Ruokolainen et al. 2016), consisting of precision (*P*), recall (*R*) and F_1 score (a harmonic mean of precision and recall). They can be formally defined by comparing predicted morpheme boundaries with ground truth morpheme boundaries such that instances of correctly predicted boundaries (TP), instances where the model wrongly predicts a boundary (FP), and instances of model failing to predict a boundary (FN) are counted and used in the expressions below:

$$P = \frac{TP}{TP + FP} \quad R = \frac{TP}{TP + FN} \quad F_1 = 2 \cdot \frac{P \cdot R}{P + R}$$

The metrics defined above measure the performance of the model on individual morphemes disregarding the ability to correctly segment the complete word. For this reason, we will also include accuracy of the whole word segmentation defined as the fraction of words where the model correctly segments the word to morphemes.

4 Results

In this section, we compare the performance of different approaches to the morphological segmentation task. All results (precision, recall, F_1–score and accuracy) are an average of the model's performance on the validation fold from 5–fold cross–validation evaluation strategy, and are shown in Table 2.

Model	Precision (%)	Recall (%)	F1 Score (%)	Accuracy (%)
BiLSTM-CRF	83.45	84.58	83.98	47.73
Glot500 tokenizer	34.32	36.19	35.23	9.25
SloBERTa tokenizer	20.57	23.91	22.12	1.94
CSEBert tokenizer	26.57	28.11	27.32	7.33
LLMSegm–Glot500	86.16	83.89	84.96	**55.82**
LLMSegm–SloBERTa	85.80	**85.36**	**85.42**	55.34
LLMSegm–CSEBert	**86.64**	83.53	85.03	55.32

Table 2: Comparing performance of different models on the task of morphological segmentation.

The worst results are, as expected, achieved by tokenizer baselines. It is interesting to note that the tokenizer built specifically for Slovenian (SloBERTa tokenizer) performs much worse on all metrics compared to the Glot500 tokenizer built to support over 500 languages. The real performance of the morphological segmentation task is visible in rows for three LLMSegm approaches, and all of those reach very similar performance regardless of the use of monolingual, trilingual, or a multilingual model. And LLMSegm models show improved performance on all metrics compared to the BiLSTM–CRF model. It is interesting to compare some predicted segmentations with the ground truth in order to gain a better understanding of the model behavior. Such comparison is shown in Table 3.

Prediction	Ground Truth	Match
astr•o•biolog•ija	astr•o•biologija	✗
barv•ar•stvo	barv•ar•stvo	✓
borba	bor•ba	✗
brod•n•ar•stvo	brod•nar•stvo	✗
de•blok•irati	de•blok•irati	✓
ka•mn•o•brus•ec	kamn•o•brus•ec	✗
mikro•bio•log•ija	mikr•o•biologija	✗
ne•besed•o•tvor•en	ne•besed•o•tvor•en	✓
ne•borb•en•ost	ne•bor•b•en•ost	✗
pre•buš•iti	pre•bušiti	✗

Table 3: Predictions of the LLMSegm–Glot500 model compared to the ground truth labels.

In most examples, there is only a single incorrect boundary prediction either splitting a valid morpheme into two, or joining two morphemes together. The model gets incorrect results in most cases containing an interfix morpheme (astr–o–biologija, kamn–o–brus–ec). An interesting mistake is splitting the word *biologija* (biology) (in words *astrobiologija*, *mikrobiologija*) into *bio•log•ija* or to *biolog•ija*, which is a reasonable mistake to make. In the ground truth, the word *biologija* is not split into any morphemes, which reflects the nature of the dataset as a word–formation resource that was converted to the morphological segmentation dataset.

5 Conclusion and Future work

In this paper we have extended our previous work on morphological segmentation and achieved state-of-the-art results for Slovenian, a language with a particular challenge of being both a less–resourced language and having very scarce resources for morphological segmentation. Our previous work (Erjavec et al. 2023) performed the first study on automatic morphological segmentation in Slovenian, and we attribute this to the lack of available data. Even the data used in the aforementioned work is a resource for study of word formation that was adapted for morphological segmentation. Although the BiLSTM–CRF model (Huang et al. 2015) trained on this data achieved a promising result (Erjavec et al. 2023), we should investigate if the results generalize to other examples not having word roots beginning only with the letter 'B' as in the used datasets.

Although modern language models pretrained on a large corpus of text show superior performance to alternative approaches, they were not previously applied to the task of morphological segmentation. With the recently proposed LLMSegm (Pranjić et al. 2024), the idea to leverage the information from the pretrained models became possible. In this paper, we apply this approach on Slovenian and improve on the results reported in related work (Erjavec et al. 2023). The LLMSegm approach was, in related work (Pranjić et al. 2024), used with the Glot-500 model (ImaniGooghari et al. 2023). We additionally evaluate the LLMSegm with a monolingual Slovene model and trilingual Croatian–Slovenian–English model. All those approaches converge to the similar score and improve over the previous state-of-the-art BiLSTM–CRF model.

Very small differences in performance of LLMSegm with different pretrained models indicate a bottleneck with either the model or the data. As results of LLMSegm on other languages (English, Turkish, Finnish, as well as South African languages, see Pranjić et al. 2024) still outperform the results presented here, we believe that future work to improve the morphological segmentation for Slovenian should be focused on creating of high–quality publicly available dataset for morphological segmentation. We have performed only a superficial analysis of results and more detailed look into errors made by the model could provide relevant insight. In particular, LLMSegm outputs a

likelihood for each prediction and taking it into account would inform us on the uncertainty of the model on the prediction. In this paper we used a trilingual CSEBert model (Ulčar et al. 2020) that offers an option for cross–lingual knowledge transfer by supporting Croatian, a language with high level of similarity to Slovenian. Future work can quantify the ability for such knowledge transfer between languages.

References

Batsuren, Khuyagbaatar, et al. "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, July 2022, pp. 103–116. doi: 10.18653/v1/2022.sigmorphon-1.11.

Creutz, Mathias, and Lagus, Krista. "Unsupervised Discovery of Morphemes". In: *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*. 2002, pp. 21–30. doi: 10.3115/1118647.1118650. url: https://aclanthology.org/W02-0603.

Creutz, Mathias, et al. "Morph-based speech recognition and modeling of out-of-vocabulary words across languages". In: *ACM Transactions on Speech and Language Processing (TSLP)* 5 (Dec. 2007), p. 3. doi: 10.1145/1322391.1322394.

Creutz, Mathias, Philip, Johan, and Lagus, Hannele Krista. "Inducing the Morphological Lexicon of a Natural Language from Unannotated Text". English. In: *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05)*. 2005, pp. 106–113.

Dobrovoljc, Kaja, et al. *Morphological lexicon Sloleks 2.0*. Slovenian language resource repository *CLARIN.SI*, 2022.

Dyer, Christopher, Muresan, Smaranda, and Resnik, Philip. "Generalizing Word Lattice Translation". In: *Proceedings of ACL-08: HLT*. 2008, pp. 1012–1020.

El-Kishky, Ahmed, et al. "Parsimonious Morpheme Segmentation with an Application to Enriching Word Embeddings". In: *2019 IEEE International Conference on Big Data (Big Data)* (2019), pp. 64–73.

Erjavec, Tomaž. *Corpus of combined Slovenian corpora metaFida 0.1*. Slovenian language resource repository *CLARIN.SI*, 2022.

Erjavec, Tomaž, et al. "Automating derivational morphology for Slovenian". In: *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference*. Lexical Computing CZ, 2023, pp. 449–465.

Hofmann, Valentin, Pierrehumbert, Janet, and Schütze, Hinrich. "Superbizarre Is Not Superb: Derivational Morphology Improves BERT's Interpretation of Complex Words". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2021, pp. 3594–3608. doi: 10.18653/v1/2021.acl-long.279.

Hou, Jue, et al. "Effects of sub-word segmentation on performance of transformer language models". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Bouamor, Houda, Pino, Juan, and Bali, Kalika. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 7413–7425. doi: 10.18653/v1/2023.emnlp-main.459.

Huang, Zhiheng, Xu, Wei, and Yu, Kai. "Bidirectional LSTM-CRF models for sequence tagging". In: *arXiv preprint* arXiv:1508.01991 (2015).

Huck, Matthias, et al. "Producing Unseen Morphological Variants in Statistical Machine Translation". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. 2017, pp. 369–375.

ImaniGooghari, Ayyoob, et al. "Glot500: Scaling Multilingual Corpora and Language Models to 500 Languages". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 1082–1117.

Knigawka, Łukasz. "Constructing a Derivational Morphology Resource with Transformer Morpheme Segmentation". In: *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*. Dec. 2022, pp. 104–109.

Kurimo, Mikko, et al. "Morpho Challenge 2005–2010: Evaluations and Results". In: *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*. 2010, pp. 87–95. url: https://aclanthology.org/W10-2211.

Loshchilov, Ilya, and Hutter, Frank. "Decoupled Weight Decay Regularization". In: *International Conference on Learning Representations*. 2017.

Pranjić, Marko, Robnik-Šikonja, Marko, and Pollak, Senja. "LLMSegm: Surface-level Morphological Segmentation Using Large Language Model". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by Calzolari, Nicoletta, et al. Torino, Italia: ELRA and ICCL, May 2024, pp. 10665–10674.

Ruokolainen, Teemu, et al. "A Comparative Study of Minimally Supervised Morphological Segmentation". In: *Computational Linguistics* 42.1 (Mar. 2016), pp. 91–120. doi: 10.1162/COLI_a_00243.

Seeker, Wolfgang, and Çetinoğlu, Özlem. "A Graph-based Lattice Dependency Parser for Joint Morphological Segmentation and Syntactic Analysis". In: *Transactions of the Association for Computational Linguistics* 3 (June 2015), pp. 359–373. issn: 2307-387X. doi: 10.1162/tacl_a_00144.

Stramljič Breznik, Irena. *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, 2004.

Ulčar, M., and Robnik-Šikonja, M. "FinEst BERT and CroSloEngual BERT: less is more in multilingual models". In: *Text, Speech, and Dialogue (TSD 2020)*. Ed. by Sojka, P., et al. Vol. 12284. *Lecture Notes in Computer Science*. Springer, 2020.

Ulčar, Matej, and Robnik-Šikonja, Marko. "SloBERTa: Slovene monolingual large pretrained masked language model". In: *24th international multiconference Information Society 2021*. Vol. C. *Data Mining and Data Warehouses*. 2021.

ADVANCEMENTS IN AUTOMATIC MORPHOLOGICAL SEGMENTATION FOR SLOVENIAN

This research paper focuses on automatic surface level segmentation for Slovenian, a less–resourced language with scarce resources for this task. Related work already covered the case of unsupervised morphological segmentation methods with unsatisfactory results and significant results were achieved only when a dataset with explicit information on word morphemes was used. Large language models pretrained on large amount of textual data have been successfully used in many NLP tasks reaching superior results to previous approaches, yet their application on the morphological segmentation task has been successfully applied only recently. By leveraging LLMSegm, the first such approach using pretrained language model, together with the digitized derivational dictionary of Slovenian, we improve on the previously published morphological segmentation results for Slovenian that used BiLSTM–CRF model. We evaluate LLMSegm in three variants, using monolingual Slovenian, trilingual Croatian–Slovenian–English, and multilingual Glot-500 model. All experiments leveraging LLMSegm achieve relatively similar scores and outperform baseline approaches and previously published results. Small differences in performance of three LLMSegm variants point to the limited amount of data and we suggest creating a high–quality, publicly available dataset for Slovene morphological segmentation to further improve results. Finally, we analyze several errors made by the trained model and suggest future work that can analyze uncertainty of the model prediction and apply the model in a cross–lingual setting.

KEYWORDS: morphological segmentation, computational linguistics, Slovenian, large language models