Slovene Morphological and Word Formation Segmentation: A Novel Dataset and Evaluation

Marko Pranjić[1,2], **Boris Kern**[3,5], **Ines Voršič**[4], **Senja Pollak**[1]
[1]Jožef Stefan Institute, Ljubljana, Slovenia.
[2]Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
[3]ZRC SAZU, Fran Ramovš Institute of the Slovenian Language, Ljubljana, Slovenia
[4]Faculty of Education, University of Maribor, Maribor, Slovenia
[5]School of Humanities, University of Nova Gorica, Nova Gorica, Slovenia

Abstract

We introduce the first publicly available manually annotated dataset for morphological segmentation and word-formation analysis for Slovene, containing 1,935 words annotated by two domain experts. The dataset provides three types of linguistic information: morphological and word-formation segments with zero-morpheme and simplex annotations. We present a four-stage annotation approach achieving inter-annotator agreement of 86.80% Krippendorff's Alpha for morphological segmentation and 85.16% for word-formation segments. Computational validation using a morphological segmentation model achieves 87.78% BPR F_1 on morphological segmentation and 83.05% on word-formation segments. Despite being smaller than previous datasets derived from non-public resources, our dataset enables high performance and supports reproducible research for morphological analysis tools for Slovene.

Keywords: morphological segmentation, word formation, Slovene, dataset annotation, language resources

1. Introduction

Slovene, a morphologically rich South Slavic language spoken by approximately 2.5 million people, presents challenges for Natural Language Processing (NLP) due to its complex inflectional morphology and derivational processes. Despite advances in NLP driven by large-scale datasets, Slovene and other morphologically rich languages with limited labeled training data remain underserved, limiting progress on computational modeling and linguistic analysis.

Slovene language features six grammatical cases, three numbers (including dual), and extensive derivational morphology that produces long words. Like other Slavic languages, Slovene is characterized by a rich morphemic structure, which is a result of multistage word formation (Kern, 2017). For example, 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 in turn yields the noun `mladostnik` (adolescent) in the third stage, which in turn yields the possessive adjective `mladostnikov` (adolescent's) in the fourth stage (Kern, 2024). These characteristics make morphological analysis important for downstream NLP tasks, while providing a case study for developing methods that generalize to other morphologically complex languages.

Advances in morphological segmentation methods, from unsupervised statistical approaches (Creutz and Lagus, 2002, 2005) to supervised models (Huang et al., 2015) and pretrained language models (Pranjić et al., 2024), demonstrate the improved performance when training data is available. Evaluation on Slovene (Pranjić and Pollak, 2024) relied on automatically generated segments from the word-formation chains present in *Slovenian Word Family Dictionary, A Test Volume for Headwords beginning with B* (BSSJB; Stramljič Breznik, 2004). Access constraints on this resource present challenges for independent reproduction of results.

Morphological segmentation information also improves representation in large language models, as research shows that incorporating morphological information during pretraining improves convergence and downstream performance (Hou et al., 2023). For languages like Slovene with limited labeled training data, such datasets and methods provide linguistic information and help bridge the performance gap with high-resource languages.

The development of morphological analysis tools for Slovene is constrained by the lack of publicly available annotated datasets. The dataset introduced with this work simultaneously captures three dimensions of morphological analysis: (i) surface-level morpheme segmentation, (ii) word-formation segments reflecting derivational processes, and (iii) simplex, corresponding base word that has not been derived through a word-formation process and cannot be divided into two or more word-formation morphemes. Both morphological and word-formation segments include ∅-morpheme (zero-morpheme) annotations, which represent morphemes without phonetic form that mark grammatical distinctions not explicitly realized in speech, such as the absence of an inflectional suffix. A ∅-morpheme poses a particular modelling challenge because models must predict the absence of a

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1781–1793
11–16 May 2026.

segment rather than its presence.

For datasets in other languages, even much more resourced than Slovene, it is rare to have both morphological and word-formation segments with ∅-morpheme annotations, as well as simplex forms. Previous research on automated morphological segmentation in Slovene (Erjavec et al., 2023; Pranjić and Pollak, 2024) relied on the restricted BSSJB dictionary (Stramljič Breznik, 2004), a word formation resource containing derivational chains (e.g., *badminton* → *badminton-ist* ('badminton player') → *badmintonist-ka* ('female badminton player')); previous work automatically generated morphological segments from these without clearly separating word-formation from inflectional morphology. These automatically generated segments lack manual validation, making their reliability uncertain for morphological analysis tasks. This limitation, coupled with the restricted access to the resource, prevents reproducible research in morphological analysis and hinders computational linguistic research and automation for Slovene.

To address this gap, we introduce a publicly available dataset[1] annotated by two domain experts, providing word-level multidimensional morphological annotations for Slovene, enabling computational modeling and linguistic analysis. Morphological segmentation identifies all morphemes (including inflectional endings), whereas word-formation segmentation focuses only on derivational formants used to create new words. Our dataset provides both perspectives along with simplex annotations, enabling rich morphological analysis of Slovene language.

This paper makes three contributions: (1) We introduce the first publicly available Slovene dataset for morphological analysis, with multidimensional morphological annotations, containing morphological segments and word formation segments, both containing ∅-morphemes, and simplex annotations for 1,935 words. (2) We present a four-stage annotation approach achieving inter-annotator agreement of 86.80% Krippendorff's Alpha for morphological segmentation and 85.16% for word formation segments. (3) Through computational validation of our annotations by training a morphological segmentation model to probe dataset consistency and learnability, we achieve 87.78% BPR F_1 on morphological segmentation and 83.05% BPR F_1 on predicting word-formation segments, improving on the previously reported score for morphological segmentation in Slovene (Pranjić and Pollak, 2024) despite using a smaller dataset.

We organize the rest of the paper as follows. Section 2 surveys existing resources and methods for morphological segmentation. Section 3 describes our dataset construction methodology, annotation procedure, and inter-annotator agreement evaluation. Section 4 validates dataset consistency and learnability through computational modeling using a morphological segmentation model. We conclude in Section 5 with discussion of findings and directions for future work.

[1]The dataset is available at: http://hdl.handle.net/11356/2060

2. Related Work

Baxi and Bhatt (2024) highlight the evolution from rule-based to neural approaches in computational morphology. Neural models achieve substantially higher performance than traditional methods but require large annotated datasets, which creates challenges for languages with limited labeled training data. This tension motivates specialized resources for underserved languages.

Major multilingual resources include UniMorph (Batsuren et al., 2022a), Universal Dependencies (Nivre et al., 2020), and MorphyNet (Batsuren et al., 2021), covering 169, 148, and 15 languages respectively. These resources focus on universal patterns rather than language-specific complexities. For Slovene, UniMorph provides only basic morphological features (grammatical case, gender, number, and part of speech), while Universal Dependencies focuses on syntactic annotation without morphological segmentation.

Hämäläinen et al. (2021) present a methodology for creating large-scale morphological datasets and training neural models for 22 languages. Their work covers morphological analysis, generation, and lemmatization tasks but excludes Slovene and does not address morphological segmentation or word formation.

Specialized word-formation resources demonstrate the value of dedicated derivational analysis. DeriNet (Sevcíková and Žabokrtský, 2014; Vidra et al., 2019) contains over 1 million Czech lexemes with 810 thousand derivational relations using semi-automated segmentation. By explicitly representing how complex words are formed from simpler bases, DeriNet enhances computational lexicons (Horenovská, 2019), aids in developing more sophisticated language technologies for Czech (Žabokrtský et al., 2016; Musil et al., 2019), and serves as a benchmark for evaluating morphological and derivational models.

Evaluation frameworks such as the SIGMORPHON 2022 Shared Task (Batsuren et al., 2022b) and Morpho Challenge (Kurimo et al., 2010) establish methodologies for morphological analysis but exclude Slovene. SIGMORPHON 2022 Shared Task focuses on canonical morpheme segmentation, while Morpho Challenge addresses surface-level segmentation. Both frameworks lack

∅-morpheme annotation and word-formation information.

Previous work on Slovene morphology has been constrained by resource availability. The BSSJB derivational dictionary (Stramljič Breznik, 2004) was previously used for surface-level morphological segmentation (Erjavec et al., 2023; Pranjić and Pollak, 2024). However, BSSJB is primarily a word-formation resource without information on morphemes. Related work derived this information using a rule-based process without manual verification.

The most comprehensive Slovene morphological resource is Sloleks (Čibej et al., 2022), a reference morphological lexicon containing 365,000 entries with inflected word forms and grammatical descriptions. Sloleks 3.0 includes morphological patterns, accentuated word forms, and phonetic transcriptions. Despite providing extensive inflectional and grammatical features, Sloleks does not include morpheme-level segmentation information.

Slovene resources specifically for morphological segmentation remain limited. The only publicly available resource we are aware of is derived from Wiktionary and covers 68 languages (Metheniti and Neumann, 2020). Slovene data contains information on word lemmas, prefixes, and suffixes for only 113 words, making it insufficient for robust morphological analysis.

A more recent resource, ArboSloleks (Čibej, 2024), provides word-formation trees with pairs of morphologically related lexemes organized around root lexeme. However, ArboSloleks is automatically generated from word relations data (Čibej et al., 2024) and lacks manual verification.

Several annotation aspects remain underdeveloped across existing resources. Explicit ∅-morpheme annotation is rare, with most datasets lacking this information despite its importance for linguistic analysis (Marşan et al., 2022).

Baxi and Bhatt (2024) identify three relevant research gaps: (1) focus on resource-rich languages, (2) lack of language-specific depth for morphologically complex languages, and (3) rare datasets combining multiple morphological dimensions. No widely adopted dataset provides explicit surface morpheme segmentation with ∅-morphemes, word-formation information, and simplex annotations in a single resource.

Our work addresses these limitations by providing the first publicly available Slovene dataset with manually annotated morphological segments, word-formation segments, and simplex annotations. Unlike semi-automated approaches like DeriNet, which does not measure annotator agreement, our dataset creation process includes inter-annotator agreement evaluation. In contrast to the restricted BSSJB dictionary, our resource is publicly available and enables reproducible research. While existing resources cover subsets of these dimensions, none combine all three annotation types.

Recent work in automated morphological segmentation has explored pretrained language model-based approaches. Pranjić et al. (2024) proposed LLMSegm, a word segmentation through binary classification of morpheme boundaries using a pretrained language model. The original work evaluated the approach on high-resource languages like English, Finnish, and Turkish in a low-data setting (1,000 training examples), and on low-resource African languages in a high-data setting (over 10,000 annotated training examples).

For Slovene specifically, Pranjić and Pollak (2024) evaluated LLMSegm using automatically derived morpheme segments on a dataset closer to the high-data setting (9,883 examples). While LLMSegm was introduced for morphological segmentation, the approach can be applied to other word segmentation tasks, making it suitable for both morphological and word-formation segmentation.

3. Dataset construction

3.1. Data Source and Selection

We sourced our target words from Sloleks 3.0 (Čibej et al., 2022), the reference morphological lexicon of Slovene. From this lexicon, we randomly selected words for manual annotation using the following constraints:

- **Removed numbers.** Numerical expressions follow morphological patterns that are not representative of word formation processes in Slovene, making them unsuitable for evaluating morphological segmentation algorithms.
- **Manually verified.** Only entries that underwent human validation were included to avoid errors from automated lexicon generation.
- **Removed proper nouns.** Proper nouns and acronyms exhibit morphological behaviors that deviate from standard Slovene word-formation patterns; proper nouns follow irregular declension patterns while acronyms resist morphological modification.

3.2. Annotation Task Definition

The dataset introduced with this work simultaneously captures three dimensions of word-level morphological analysis: (i) surface-level morpheme segmentation, (ii) word-formation segments reflecting derivational processes, and (iii) simplex, corresponding base word that has not been derived through a word-formation process and cannot be divided into two or more word-formation morphemes.

Word formation is the broad, overarching process for creating new words with a derivational morphology as a major type of word formation. Inflectional morphology is a separate process that modifies existing words to fit their grammatical role and is therefore not considered a type of word formation.

Morphological and word-formation segments. We distinguish between morphological segments and word-formation segments. Consider the adjective *nepozidan* ('not built-up'):

- **Morphological:** `ne-po-zid-a-n-∅`
- **Word-formation:** `ne-po-zida-n`

While morphological segments describe inflection, the word-formation process follows this derivational chain:

zidati → *pozidati* → *pozidan* → *nepozidan*
'to build' → 'to build up' → 'built-up' → 'not built-up'

This illustrates how prefixation and participial derivation interact: *zidati* ('to build') receives the perfective prefix *po-* to indicate completion (to finish building, to build up), becomes the participle *pozidan* ('built-up') expressing a resultant state, and finally receives the negative prefix *ne-* to form the antonym *nepozidan* ('not built-up'). This creates a semantic chain:

process → completion → state → negation

Zero-morpheme. The ∅-morpheme (zero-morpheme) represents a morpheme without phonetic form, used to represent grammatical distinctions that are not explicitly marked. It can function as both morphological and word-formation morpheme. For example, *korak-∅* ('a step'). The zero morpheme here serves a dual role: (i) as an inflectional morpheme, nominative masculine nouns in Slovene typically have the ∅-morpheme (contrast with the genitive singular form *koraka*, where suffix *-a* appears); and (ii) as a word-formational morpheme, because *korak* is derived from the verb *korakati* ('to walk in a steady manner', 'to march'), and the ∅-morpheme is necessary in forming the noun from the verb.

Simplex. A simplex represents a word that has not been formed through a word-formation process and therefore cannot be divided into two or more word-formation morphemes. For example, a participle *leteč* ('flying') has a simplex *leteti* ('to fly') rather than *let* ('the act of flying', 'flight').

3.3. Annotation Methodology

Two domain experts performed all annotations. Determining simplex words and identifying morphemes in Slavic derivatives is a highly complex linguistic task. Word formation analysis requires identifying shared semantic components between motivating and motivated forms, often necessitating consultation of multiple dictionaries and etymological resources. This process demands specialized linguistic expertise to distinguish genuine derivational relationships from apparent morphological similarities.

Given the complexity of simultaneously annotating morphological segments, word-formation boundaries, ∅-morphemes, and word simplexes, this approach prioritized consistency over efficiency. As this was our first experience with multidimensional morphological annotation, the iterative structure allowed us to develop shared understanding while learning the task. The core principles governing the annotation process are detailed in Appendix A.

Stage 1: Calibration. We collaboratively discussed annotation of seven words during a meeting to establish guidelines for compound words and ∅-morphemes. This step established common understanding of annotation goals.

Stage 2: Refinement. Two annotators jointly annotated 50 words using a collaborative online tool. The small scope enabled guideline refinement, edge-case discussion, and collaborative error reduction.

Stage 3: Annotation. Each annotator independently annotated 975 words (half of 1,950 total). This step is intended for annotation efficiency while building on the shared understanding from previous stages. Collaboration was still allowed, and annotators even consulted other experts when word origins were unclear.

Stage 4: Validation. Both annotators independently annotated an additional set of 50 words (different from the Stage 2 refinement set) without collaboration to enable inter-annotator agreement calculation.

Our dataset combines multiple annotation types. To provide comprehensive evaluation, we validate annotation quality through inter-annotator agreement and probe dataset learnability through computational modeling. First, we evaluate annotation reliability through inter-annotator agreement using Krippendorff's Alpha (Krippendorff, 1970), which provides a measure of consistency across annotations. This step addresses the combination of morphological and word-formation segments, which require different linguistic expertise and annotation decisions. Second, we validate dataset usability by training LLMSegm as a probe to demonstrate that the annotations are sufficiently consistent for learning the underlying segmentation information. This computational validation confirms that our dataset contains reliable annotations and serves as a valuable resource.

3.4. Dataset Characteristics

During annotation of the last two stages, approximately 5% of words were skipped due to challenges with the sourced data, most commonly adjectives derived from proper nouns. Other reasons included abbreviations, mistyped words, unknown words, and words of unclear origin. We removed obvious annotation errors like mistyped segments where the original word couldn't be reconstructed from the annotated segments, resulting in a final dataset of 1,935 annotated words. The domain experts noted that annotation required significant effort and consultation with specialists for complex cases.

The final dataset contains 1,935 annotated words with an average length of 9.1 characters ($SD = 2.7$). The dataset includes 6,457 total morphological segments (average 3.3 morphemes per word) and 5,292 total word-formation segments (average 2.7 segments per word).

The dataset covers several word types: nouns (54.4%), adjectives (24.7%), verbs (11.8%), adverbs (8.8%), pronouns (0.2%), particles (0.1%), and prepositions (0.1%). Discounting the removed categories (numbers and proper nouns), this distribution mostly reflects frequency patterns in Sloleks 3.0. We analyze morphological complexity through affixation patterns and compound formation. Of the annotated words, 33.8% contain prefixes, 88.0% contain suffixes, and 8.8% are compound words.[2] The prevalence of suffixation (88.0%) reflects Slovene's inflectional morphology, where suffixes encode grammatical information including case, gender, and number.

The most frequent prefixes are *po-* (12.1%), *o-* (9.8%), *ne-* (8.4%), *pre-* (7.9%), and *za-* (6.2%). These prefixes serve verbal functions, indicating direction, aspect, completion, intensity, or negation. This distribution aligns with Slovene's prefixal verb system, which is central to aspect marking and meaning derivation.

The suffix analysis shows inflectional and derivational patterns. High-frequency inflectional suffixes include grammatical endings (*-a*, *-o*, *-e*, *-i*, *-n*, *-ti*) that encode case, gender, number, and grammatical form. Common derivational suffixes include *-en* and *-ost* (e.g., in *nerešen* 'unsolved' and *mladost* 'youth'). The average of 2.7 suffixes per word explains why the mean morpheme count exceeds 3.0, as most Slovene words derive from a single root modified by affixes.

This distribution provides coverage of Slovene morphological patterns while maintaining annotation complexity across linguistic phenomena. In Table 1, we provide a sample of the data with a target word, its segmentation to morphemes and word-formation segments together with a ∅-morpheme, and a simplex of the word.

[2]Number of prefixes and suffixes is the number of segments before, and after, the root. The root was algorithmically determined by the best match with a simplex of the word.

3.5. Annotation Quality Evaluation

We evaluate annotation quality through inter-annotator agreement using multiple complementary metrics. We report Krippendorff's Alpha (K_α) (Krippendorff, 1970) as our primary reliability measure, alongside BPR F_1 scores and accuracy to provide task-specific evaluation.

K_α quantifies overall agreement between annotators, with values ranging from -1 (systematic disagreement) to 1 (perfect agreement). An $K_\alpha \geq 0.8$ indicates reliable data with good agreement, while $K_\alpha = 0$ represents chance-level agreement. For segmentation tasks, we complement this with BPR F_1 scores, which directly measure how well annotators' segment boundaries align. BPR F_1 is the standard evaluation metric for morphological segmentation models, making it a suitable point of comparison with inter-annotator evaluation.

The ∅-morpheme annotation agreement can be viewed either as a task of determining whether the ∅-morpheme is present at each possible position (comparable to segmentation evaluation), or as a binary task that evaluates whether both annotators placed ∅-morpheme to matching positions. While the former approach yields a 97.78% F_1-score, the latter is consistent with the simplex evaluation. We therefore report K_α and Accuracy for ∅-morpheme placement and do not include F_1-score in the summary of agreement results presentend in Table 2.

All annotation types achieve K_α values above 0.85, indicating high reliability. Given the small sample size (50 words, 2.5% of the dataset), we performed bootstrap resampling with 10,000 iterations to determine the stability of our IAA estimates. The resulting narrow 95% confidence intervals (95% CI) for K_α confirm that our inter-annotator agreement is statistically reliable: morphological segments, 86.80% with 95% CI $[81.39, 91.66]$; word formation, 85.16% $[78.85, 90.90]$; and simplex, 85.28% $[74.30, 93.58]$. The narrow confidence intervals for segmentation tasks reflect that agreement is calculated on boundary positions across all characters, with the 50 words providing close to 450 individual decision points.

Our results compare favorably with inter-annotator agreement reported in prior work. The only comparable study we identified is (Volodina et al., 2021), which reports K_α in the range 87–93% for word-formation segment annotation in Swedish. Our K_α values (85.16% for word-formation segments) demonstrate comparable annotation quality for Slovene, a morphologically richer language.

Word	Morpheme segments	Word formation segments	Simplex
pisati (‘to write’)	pis-a-ti	pisati	pisati (‘to write’)
pisatelj (‘writer’)	pis-a-telj-∅	pisa-telj	pisati (‘to write’)
napis (‘inscription’)	na-pis-∅	na-pis-∅	pisati (‘to write’)
gledalec (‘watcher’)	gled-a-lec-∅	gled-alec	gledati (‘to watch’)
mizarka (‘female carpenter’)	miz-ar-k-a	miz-ar-ka	miza (‘table’)
mizarski (‘carpenter’s’)	miz-ar-sk-i	miz-ar-ski	miza (‘table’)
golfigrišče (‘golf course’)	golf-∅-igr-išč-e	golf-∅-igr-išče	golf, igrati (‘golf’, ‘to play’)

Table 1: A sample of the annotated data showing target word, morphological and word formation segments, and word simplex.

	BPR F_1	Accuracy	K_α
Morphological segments	92.36%	64.58%	86.80%
Word formation segments	91.63%	66.67%	85.16%
Simplex	–	85.42%	85.28%
∅–morpheme	–	97.92%	95.86%

Table 2: Inter-annotator agreement across annotation types.

4. Dataset Validation via Automatic Segmentation

We validate dataset consistency and learnability by training LLMSegm (Pranjić et al., 2024), a method that adapts pretrained language models for word segmentation tasks. LLMSegm is well-suited for our low-data regime, as it demonstrates strong performance even with limited training examples by leveraging knowledge from pretrained language models. In this work, LLMSegm serves as a probe to determine whether our annotations are sufficiently consistent for computational learning. We evaluate on both morphological and word-formation segmentation using 1,935 annotated examples, comparable to the low-data setting in prior work.

4.1. Evaluation Metrics

We evaluate morphological and word-formation segmentation using two complementary approaches. Boundary Precision, Recall, and F_1-score (BPR) measures the correctness of segment boundaries, while Accuracy measures exact matches of complete word segmentations. These metrics align with standard evaluation practices in related work, enabling direct comparison with prior work.

Our dataset includes morphological and word-formation segments, both containing ∅-morpheme information, which makes direct comparison with existing morphological segmentation resources difficult. We report evaluation metrics on segmentation modelling without accounting for ∅-morphemes to enable comparison with prior work.

For morphological and word-formation segmentation, we use both BPR F_1 and Accuracy metrics to evaluate computational performance. For word simplex identification and ∅-morpheme annotation (for both segmentation types), we evaluate annotation quality only through inter-annotator agreement using K_α and Accuracy (see Table 2), as these annotations are primarily used for linguistic analysis rather than as downstream task targets.

Given a predicted segmentation $S_{pred} = \{p_1, p_2, \ldots, p_n\}$ and a gold segmentation $S_{gold} = \{g_1, g_2, \ldots, g_m\}$, where each p_i and g_j represents a segment, we define boundary positions ($b_{k,k+1}$) as the position between subsequent segments. Let B_{pred} be the set of boundary positions in the predicted segmentation and B_{gold} be the set of boundary positions in the gold segmentation. We calculate the metrics as:

$$\text{Precision} = \frac{|B_{pred} \cap B_{gold}|}{|B_{pred}|} \quad (1)$$

$$\text{Recall} = \frac{|B_{pred} \cap B_{gold}|}{|B_{gold}|} \quad (2)$$

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

For morphological and word-formation segmentation, we also evaluate word-level accuracy as exact match between predicted and gold segmentations. Given a set of W words, accuracy is defined as the proportion of words where the predicted segmentation S_{pred} exactly matches the gold segmentation S_{gold} (all segments are identical):

$$\text{Accuracy} = \frac{|\{i \in \{1, \ldots, W\} : S^{(i)}_{pred} = S^{(i)}_{gold}\}|}{W} \quad (4)$$

4.2. Experimental Setup

Our approach follows the LLMSegm approach (Pranjić et al., 2024), adapting the binary classification formulation for morphological and word-formation segmentation tasks. For each word, the model predicts whether a boundary exists for each possible boundary position and outputs binary decisions that are reconstructed to form the complete word segmentation. Following LLMSegm, we use cross-entropy loss with class weighting to address class imbalance. We modify LLMSegm approach by incorporating label smoothing (Szegedy et al., 2015) during loss calculation, tuned via hyperparameter optimization. Label smoothing serves as a regularization technique that prevents overfitting of the model on our small dataset by encouraging less confident predictions.

4.2.1. Evaluation Strategy

We use holdout test set approach to ensure unbiased performance evaluation. We extract 10% of the data as a holdout test set using a fixed random seed (42) for reproducibility. The remaining 90% constitutes the training and validation data, which is further split into 90% training and 10% validation portions for hyperparameter optimization. The holdout test set is used exactly once for final evaluation after hyperparameter optimization is complete.

4.2.2. Hyperparameter Optimization

We use Optuna (Akiba et al., 2019) with a Tree-structured Parzen Estimator (TPE) sampler for hyperparameter tuning, which efficiently explores the hyperparameter space by modeling the relationship between hyperparameters and BPR F_1-score on validation data. The search space includes model selection among three pretrained language models: `SloBERTa` (Ulčar and Robnik-Šikonja, 2021) – a monolingual Slovene BERT, CroSloEngualBERT (Ulčar and Robnik-Šikonja, 2020) – a Croatian-Slovene-English trilingual BERT, and Glot500 (ImaniGooghari et al., 2023) – a multilingual model supporting over 500 languages including Slovene. Additionally, we tune learning rate (1×10^{-5} to 5×10^{-5}), batch size ($\{64, 128, 256\}$), weight decay (1×10^{-4} to 0.1), and label smoothing (0.0 to 0.2) parameter used in loss calculation.

The training uses the AdamW optimizer (Loshchilov and Hutter, 2017) with weight decay regularization and a learning rate schedule with linear warmup for 300 steps. We use early stopping with patience of 10 epochs, monitoring BPR F_1 score on validation data. The maximum number of epochs is fixed at 40, with gradient norm clipping at 1.0 for training stability. We conduct 100 trials per task with the optimization objective to maximize BPR F_1 score on validation data. Table 3 summarizes the optimal hyperparameters found for each task.

4.3. Results

Table 4 presents the results of LLMSegm using the monolingual SloBERTa model, which achieved the best performance among the evaluated models.

The results demonstrate that our dataset supports effective learning of both morphological and word-formation segmentation. LLMSegm achieves 87.78% BPR F_1 on morphological segmentation and 83.05% on word-formation segments. The higher label smoothing requirement for word-formation segmentation (0.086) compared to morphological segmentation (0.023) may indicate greater annotation noise or task difficulty, which aligns with the lower inter-annotator agreement observed for word-formation segments (85.16% vs. 86.80% K_α), though the difference is small and potentially insignificant. Although learning rate is comparable between two tasks, word-formation segmentation task has much higher weight decay (2.77×10^{-3} vs. 3.8×10^{-4}) supporting the hypothesis that word formation requires more regularization during training.

Among published results on Slovene morphological segmentation, Pranjić and Pollak (2024) reported the highest score with a BPR F_1-score of 85.42%. Although those results were evaluated on private data, the resource used to construct the data contains word-formation chains rather than true morphological segmentation annotations, and the dataset was automatically generated with a rule-based system without manual verification. Our dataset is smaller (1,935 words vs. 9,883 in BSSJB) yet reaches 87.78% BPR F_1 score, demonstrating the quality of our manually annotated dataset. The dataset's public availability also ensures reproducibility and enables result refinement.

5. Conclusion and Future work

In this study, we present a dataset for morphological segmentation and word formation in Slovene, addressing a gap in Slovene NLP resources. The complexity of Slovene's morphology presents challenges in creating datasets that capture this information, making this dataset a valuable resource for NLP research. We report annotation consistency with Krippendorff's Alpha inter-annotator agreement of 86.80% for morphological segmentation and 85.16% for word-formation segments. This measure indicates high consistency and quality of annotations, which form the foundation for training models in morphological and word-formation segmentation.

Hyperparameter	Morphological segments	Word formation segments
Model	SloBERTa	SloBERTa
Batch size	128	128
Learning rate	4.36×10^{-5}	3.05×10^{-5}
Weight decay	3.80×10^{-4}	2.77×10^{-3}
Label smoothing	0.023	0.086

Table 3: Optimal hyperparameters for morphological and word-formation segmentation tasks.

	Precision	Recall	BPR F_1	Accuracy
Morphological segments	87.24%	88.32%	87.78%	53.61%
Word formation segments	82.80%	83.29%	83.05%	52.58%

Table 4: Performance of LLMSegm on our dataset using a two-phase holdout evaluation with a monolingual SloBERTa model.

Our four-stage annotation approach may appear cautious due to prioritizing consistency over efficiency. Each stage served as a quality assurance measure and a learning opportunity, allowing us to develop shared understanding. The high inter-annotator agreement demonstrates that this approach ensured dataset quality despite our initial learning curve.

The K_α calculated on our data is a lower bound of inter-annotator agreement due annotators collaborating on examples where they were unsure how to segment the target word. Examples used for K_α calculation were annotated without collaboration, making annotation disagreement more likely.

Our LLMSegm experiments validate data consistency and learnability, reaching 87.78% and 83.05% BPR F_1-score on morphological segmentation and word-formation segments, respectively, using held-out data. This demonstrates that our dataset is sufficiently consistently annotated for computational learning and provides a base for achieving a good performance in NLP tasks.

We argue that a corpus of 1,935 words provides a sufficient basis for the current task. While models like Chipmunk (Cotterell et al., 2015) achieve results on Finnish (88.46%) and Turkish (82.17%) (see Pranjić et al., 2024) by incorporating external linguistic features such as spellchecker data and prefix lists, our approach achieves a comparable accuracy of 87.78% using only the provided dataset. This performance also aligns with or exceeds results reported by LLMSegm (Pranjić et al., 2024) on Finnish (84.44%) and Turkish (87.69%), which use training sets roughly half the size of ours. These results suggest that our corpus is large enough to capture the necessary morphological generalizations without requiring supplementary external resources.

While the results are promising, there are areas for improvement. The performance indicates that certain word types or morphological structures pose challenges, warranting exploration. As proposed by Pranjić et al. (2024), the trained model could be probed for likelihood of segmenting different affixes to inform this exploration.

Regarding practical deployment of models trained on this data in real-world applications, tokens excluded from our analysis (proper nouns, acronyms, typos) require specific handling strategies. We recommend a pipeline approach where such tokens are detected via Named Entity Recognition (NER) or normalization tools (such as spellcheckers) and processed separately, allowing the segmentation model to focus on productive morphological and word-formation patterns.

We measured inter-annotator agreement on $\emptyset$-morpheme placement and word simplex, without computational evaluation of machine learning tasks using this information, leaving this for future work.

In future work we plan to explore several research directions. First, we will use the dataset with generative large language models in in-context learning settings to evaluate the performance of morphological segmentation with minimal examples. We also plan to investigate the potential of morphological segmentation for improving LLM training by replacing default tokenizers with morphologically-aware alternatives, following the work by Hou et al. (2023).

Second, we will leverage our annotated resource for detection of word-formation chains, which could enable computational analysis of derivational relationships in Slovene. This could support applications in etymology, lexical semantics, and language learning.

Third, we will explore applications in diachronic semantic change research. Current approaches represent word meaning as averages of subtoken embeddings (Giulianelli et al., 2020), but we hypothesize that morphologically-aware subtokens could improve temporal semantic analysis by preserving meaningful morphological boundaries.

Finally, we aim to expand the dataset to include

more diverse morphological patterns, as well as develop integrated models that simultaneously handle all annotation types including simplex identification and ∅-morpheme prediction.

6. Limitations

Our work has several limitations. First, the dataset contains 1,935 words, which is smaller than typical morphological segmentation datasets. While sufficient for initial evaluation, this size limits the effectiveness of data-intensive approaches like deep neural networks and does not capture the full diversity of Slovene morphological patterns.

Second, our dataset coverage has specific gaps. We focus on entries from Sloleks, which introduces frequency bias toward more common words and word types, under-representing rare morphological patterns. Additionally, we exclude proper nouns and numerical expressions.

Third, our computational experiments focus on morphological and word-formation segmentation, but do not incorporate simplex and ∅-morpheme prediction for both segmentation types into the modeling framework. The evaluation framework also excludes ∅-morphemes from segmentation metrics to enable comparison with prior work.

7. Ethics Statement

The research involves minimal ethical risk. We source all words from the publicly available lexical database, containing standard Slovene vocabulary. Our dataset supports research on Slovene, a morphologically rich language with limited labeled training data for morphological segmentation, contributing to linguistic diversity in NLP.

8. Acknowledgments

This work was supported by the Slovenian Research and Innovation Agency (ARIS) through the core research program Knowledge Technologies (P2-0103), the project Formant combinatorics in Slovenian (J6-3131), and the project Large Language Models for Digital Humanities (GC-0002).

9. Bibliographical References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

K. Batsuren, O. Goldman, S. Khalifa, N. Habash, W. Kieraś, G. Bella, B. Leonard, G. Nicolai, K. Gorman, Y. G. Ate, M. Ryskina, S. Mielke, E. Budianskaya, C. El-Khaissi, T. Pimentel, M. Gasser, W. A. Lane, M. Raj, M. Coler, J. R. M. Samame, D. S. Camaiteri, E. Z. Rojas, D. López Francis, A. Oncevay, J. López Bautista, G. C. S. Villegas, L. T. Hennigen, A. Ek, D. Guriel, P. Dirix, J. Bernardy, A. Scherbakov, A. Bayyr-ool, A. Anastasopoulos, R. Zariquiey, K. Sheifer, S. Ganieva, H. Cruz, R. Karahóǧa, S. Markantonatou, G. Pavlidis, M. Plugaryov, E. Klyachko, A. Salehi, C. Angulo, J. Baxi, A. Krizhanovsky, N. Krizhanovskaya, E. Salesky, C. Vania, S. Ivanova, J. White, R. H. Maudslay, J. Valvoda, R. Zmigrod, P. Czarnowska, I. Nikkarinen, A. Salchak, B. Bhatt, C. Straughn, Z. Liu, J. N. Washington, Y. Pinter, D. Ataman, M. Wolinski, T. Suhardijanto, A. Yablonskaya, N. Stoehr, H. Dolatian, Z. Nuriah, S. Ratan, F. M. Tyers, E. M. Ponti, G. Aiton, A. Arora, R. J. Hatcher, R. Kumar, J. Young, D. Rodionova, A. Yemelina, T. Andrushko, I. Marchenko, P. Mashkovtseva, A. Serova, E. Prud'hommeaux, M. Nepomniashchaya, F. Giunchiglia, E. Chodroff, M. Hulden, M. Silfverberg, A. D. McCarthy, D. Yarowsky, R. Cotterell, R. Tsarfaty, and E. Vylomova. 2022a. UniMorph 4.0: Universal Morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855, Marseille, France. European Language Resources Association.

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

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

Jatayu Baxi and Brijesh Bhatt. 2024. Recent advancements in computational morphology : A comprehensive survey.

Ryan Cotterell, Thomas Müller, Alexander Fraser,

and Hinrich Schütze. 2015. Labeled morphological segmentation with semi-Markov models. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 164–174, Beijing, China. Association for Computational Linguistics.

Mathias Creutz and Krista Lagus. 2002. Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*, pages 21–30.

Mathias Johan Philip Creutz and Krista Hannele Lagus. 2005. Inducing the morphological lexicon of a natural language from unannotated text. In *Proc. International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05)*, pages 106–113.

Tomaž Erjavec, Marko Pranjić, Andraž Pelicon, Boris Kern, Irena Stramljič Breznik, and Senja Pollak. 2023. Automating derivational morphology for slovenian. In *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference*, pages 449—-465. Lexical Computing CZ.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

Miika Hämäläinen, Niko Partanen, Jack Rueter, and Kimmo Kohonen. 2021. Neural morphology dataset and models for multiple languages, from the large to the endangered. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 166–177, Reykjavik, Iceland (Online). Linköping University Electronic Press.

Karolína Horenovská. 2019. Extending czech thesauri using word-formation network. In *Conference on Theory and Practice of Information Technologies*.

Jue Hou, Anisia Katinskaia, Anh-Duc Vu, and Roman Yangarber. 2023. Effects of sub-word segmentation on performance of transformer language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7413–7425, Singapore. Association for Computational Linguistics.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. 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)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Boris Kern. 2017. *Stopenjsko besedotvorje. Na primeru glagolov čutnega zaznavanja*. Number 11 in Lingua Slovenica. Založba ZRC, Ljubljana.

Boris Kern. 2024. Considering word formation in compiling dictionaries. In *Lexicography and Semantics: proceedings of the XXI EURALEX International Congress, 8-12 October 2024, Cavtat, Croatia*, pages 438–448. Institut za hrvatski jezik.

Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70.

Mikko Kurimo, Sami Virpioja, Ville Turunen, and Krista Lagus. 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*, pages 87–95, Uppsala, Sweden. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Büşra Marşan, Salih Akkurt, Muhammet Şen, Merve Gürbüz, Onur Güngör, Şaziye Özateş, Suzan Uskudarli, Arzucan Özgür, Tunga Gungor, and Balkiz Ozturk. 2022. Enhancements to the boun treebank reflecting the agglutinative nature of turkish. In *Proceedings of the 14th International Workshop on Turkish Natural Language Processing (TurNLP 2022)*, volume 3315, pages 73–84, Marseille, France. CEUR-WS.org.

Eleni Metheniti and Guenter Neumann. 2020. Wikinflection corpus: A (better) multilingual, morpheme-annotated inflectional corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3905–3912, Marseille, France. European Language Resources Association.

Tomáš Musil, Jonáš Vidra, and David Mareček. 2019. Derivational morphological relations in word embeddings. In *Proceedings of the 2019*

ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 173–180, Florence, Italy. Association for Computational Linguistics.

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

Marko Pranjić and Senja Pollak. 2024. Razvoj automatske morfološke segmentacije za slovenski jezik = Advancements in automatic morphological segmentation for slovene. In *Stopenjsko besedotvorje: 23. mednarodna znanstvena konferenca Komisije za besedotvorje pri Mednarodnem slavističnem komiteju*, page 99–100. Besedilo v hrv. in angl.

Marko Pranjić, Marko Robnik-Šikonja, and Senja Pollak. 2024. 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)*, pages 10665–10674, Torino, Italia. ELRA and ICCL.

Magda Sevcíková and Z. Žabokrtský. 2014. Word-formation network for czech. In *International Conference on Language Resources and Evaluation*.

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

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the inception architecture for computer vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826.

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

M. Ulčar and M. Robnik-Šikonja. 2020. FinEst BERT and CroSloEngual BERT: less is more in multilingual models. In *Text, Speech, and Dialogue TSD 2020*, volume 12284 of *Lecture Notes in Computer Science*. Springer.

Jonáš Vidra, Zdeněk Žabokrtský, Magda Ševčíková, and Lukáš Kyjánek. 2019. DeriNet 2.0: Towards an all-in-one word-formation resource. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*, pages 81–89, Prague, Czechia. Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics.

Elena Volodina, Yousuf Ali Mohammed, and Therese Lindström Tiedemann. 2021. CoDeRooMor: A new dataset for non-inflectional morphology studies of Swedish. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 178–189, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Zdeněk Žabokrtský, Magda Ševčíková, Milan Straka, Jonáš Vidra, and Adéla Limburská. 2016. Merging data resources for inflectional and derivational morphology in Czech. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1307–1314, Portorož, Slovenia. European Language Resources Association (ELRA).

10. Language Resource References

A. Bajec, K. Ahačič, M. Bojc, M. Borin, S. Bunc, V. Cestnik, F. Dajnko, M. Furlan, A. Glinšek, A. Gložančev, M. Gradečak, K. Grašič, M. Hajdín, M. Hajnšek-Holz, T. Jarc, F. Jakopin, A. Janežič, M. Kmecl, T. Korošec, Z. Krassnig, L. Legiša, A. Lukan, M. Mejak, M. Miklavčič, J. Moder, M. Orožen, M. Petelin, B. Pogačnik, B. Rebož, T. Rjavec, E. Sicherl, P. Simoniti, D. Slavič, M. Smolnikar, M. Snoj, J. Snoj, S. Suhadolnik, H. Tuma, E. Šuštarič, J. Čampa, S. Šifrer, and M. Žmuc. 2019. Dictionary of the Slovenian standard language (elexis) 1.0. Slovenian language resource repository CLARIN.SI.

Čibej, Jaka. 2024. *Dataset of Slovene word formation trees ArboSloleks 1.0*. Slovenian language resource repository CLARIN.SI.

Čibej, Jaka and Arhar Holdt, Špela and Krek, Simon. 2024. *List of word relations from the Sloleks 2.0 lexicon 1.1*. Slovenian language resource repository CLARIN.SI.

Čibej, Jaka and Gantar, Kaja and Dobrovoljc, Kaja and Krek, Simon and Holozan, Peter and Erjavec, Tomaž and Romih, Miro and Arhar Holdt, Špela and Krsnik, Luka and Robnik-Šikonja, Marko.

2022. *Morphological lexicon Sloleks 3.0*. Slovenian language resource repository CLARIN.SI.

Janez Dular, Milena Hajnšek-Holz, Franc Jakopin, Janko Moder, Jože Toporišič, Martin Ahlin, Ljudmila Bokal, Alenka Gložančev, Janez Keber, Branka Lazar, Zvonka Praznik, Jerica Snoj, Nastja Vojnovič, Stane Suhadolnik, Peter Weiss, and Vlado Nartnik. 2019. Dictionary of the slovenian normative guide - SP2001 (ELEXIS). Slovenian language resource repository CLARIN.SI.

Ivanka Šircelj-Žnidaršič, Milena Hajnšek-Holz, Polona Kostanjevec, Andreja Žele, Marjeta Humar, Vlado Nartnik, Janez Keber, Borislava Košmrlj-Levačič, and Primož Jakopin. 1998. *Besedišče slovenskega jezika z oblikoslovnimi podatki: A - Ž: po gradivu za slovar sodobnega knjižnega jezika zbrane besede, ki niso bile sprejete v Slovar slovenskega knjižnega jezika*. Zbirka Slovarji. Znanstvenoraziskovalni center SAZU, Založba ZRC, Ljubljana.

Marko Snoj. 2016. *Slovenski etimološki slovar*, 3. izd. edition. Založba ZRC.

A. Core Annotation Principles

Given the morphological complexity of Slovene, the annotation process relied on specialized linguistic expertise in Slavic word formation. The core principles governing the annotation process are outlined below.

A.1. Verification Process

Determining simplex words and identifying morphemes in Slavic derivatives requires fine-grained linguistic judgement. To resolve ambiguities, verify semantic components, and confirm etymological roots, we systematically consulted established standard dictionaries, primarily the Dictionary of the Standard Slovenian Language (SSKJ; Bajec et al., 2019), the Slovenian Normative Guide (SP 2001; Dular et al., 2019), and the Lexicon of the Slovenian Language (BSJ; Šircelj-Žnidaršič et al., 1998). For deeper historical context, we cross-referenced a specialized etymological dictionary (Snoj, 2016).

To establish robust structural rules and resolve opaque etymological edge cases (such as historical loan-translations), we also consulted specialized historical Slovenian dictionaries alongside established morphemic, word-family, and comparative Slavic frameworks for other relevant languages.

A.2. Word-Formation and Simplex Identification

The fundamental organizing principle of the dataset is the synchronic relationship between the motivating (base) word and the motivated (derived) word. The identification of word-formation segments and simplex forms followed these rules:

- **Simplex:** The derivational chain strictly begins with a simplex (a non-derived base word) that serves as the root for subsequent derivatives.
- **Semantic Criteria:** When the derivational relationship is ambiguous, semantic criteria determine the segmentation. Boundaries are placed only where a clear semantic relationship and derivational process can be established between the motivating and motivated words.
- **Multi-Stage Derivation:** Derivatives are analyzed through their intermediate motivating forms. For example, a word like *nepozidan* ('not built-up') is analyzed through its derivational stages (*zidati* 'to build' → *pozidati* 'to build up' → *pozidan* 'built-up').
- **Segmentation Granularity:** Word-formation segmentation groups affixes into functional units based on how the word was derived. Grammatical elements added in a single derivational step are kept together as one segment. For example, in the noun *izpuhtevanje* ('evaporation'), the entire suffix *-anje* is treated as a single derivational unit (i.e., *iz-puh-t-ev-anje*).

A.3. Morphological Segmentation

While word-formation segmentation isolates derivational formants, morphological segmentation decomposes words into their smallest grammatical units, including inflectional morphology.

- **Inflectional Boundaries:** After establishing the derivational stem, we separate it from all inflectional affixes (such as suffixes encoding case, gender, number, or verbal aspect).
- **Verbal Paradigms:** The verbal theme vowel and infinitive marker are segmented from the root (e.g., *zid-a-ti* 'to build').
- **Segmentation Granularity:** In contrast to word-formation segmentation, morphological segmentation isolates the smallest possible grammatical and inflectional units. Complex suffixes are fully decomposed into their constituent parts, regardless of the derivational process. For example, in the noun *izpuhtevanje* ('evaporation'), word-formation treats the suffix *-anje* as a single unit, but morphological segmentation breaks it down into the thematic vowel *-a-*, the noun-forming suffix *-nj-*, and the neuter nominative ending *-e* (resulting in *iz-puh-t-ev-a-nj-e*).

A.4. The ∅-Morpheme (Zero-Morpheme)

Because the ∅-morpheme represents a morpheme without phonetic form, it requires specific annotation rules across both segmentation tasks:

- **Inflectional ∅-Morphemes (Morphological Segmentation):** We insert a ∅-morpheme to denote an empty morphological slot within an inflectional paradigm. A common example is the nominative singular form of masculine nouns (e.g., *korak*-∅ 'a step', compared to the genitive singular *korak-a* 'of a step').
- **Derivational ∅-Morphemes (Word-Formation):** The ∅-morpheme also acts as a word-formational marker indicating zero-derivation, where a new word is formed from a base without an overt affix. For example, the noun *korak* ('a step') is derived from the verb *korakati* ('to step') via a ∅-morpheme.
- **Structural ∅-Morphemes (Compounding):** When two roots are joined in a compound word without an overt linking vowel (an interfix like *-o-*), we insert a ∅-morpheme to structurally represent the required compounding boundary (e.g., *trikolesen* 'three-wheeled', from *tri* 'three' + *kolo* 'wheel', segmented as *tri*-∅-*kol-es-en*-∅).

A.5. Specific Relational Rules

To ensure consistency across the dataset, we applied specific relational rules to common morphological patterns and edge cases:

- **Stem Extensions and Allomorphy:** When derivational processes force a root to take on an oblique or extended form (an allomorph), we link the derivative back to the base simplex but segment the extended stem. For example, words derived from *dan* ('day') use the oblique stem *dnev-* (from the genitive *dneva*); thus, *osemdneven* ('eight-day') is mapped to the simplex *dan* but segmented structurally as *osem*-∅-*dnev-en*-∅. Similarly, *kolesarski* ('cycling', adjective) maps to the simplex *kolo* ('bicycle') but is segmented using the extended stem *koles-ar-ski*.
- **Sciences and Experts:** When establishing the relation between a field of study and an expert, the expert is consistently morphemized as a derivative of the science (e.g., *ornitologija* 'ornithology' → *ornitolog*-∅ 'ornithologist').
- **Feminatives:** Terms for female persons or animals are structurally treated as derivatives formed from their male counterparts (e.g., *genealog* 'genealogist' → *genealog-inj-a* 'female genealogist').
- **Morphemic Nodes:** When two elements merge with phonological alternations, the morphemic boundary is generally placed to preserve the underlying root (e.g., *siromak-* 'poor person' + *-stvo* becomes *siromaš-tvo* 'poverty', recognizing the consonant shift).
- **Hypothetical Words:** If a motivating word is not found in standard dictionaries but a derivative exists, a hypothetical intermediate derivative is assumed to maintain the logical structural link. For example, the adverb *neprenehoma* ('incessantly') is segmented as derived from a hypothetical positive intermediate *(prenehoma)*, maintaining the derivational chain back to the simplex *nehati* ('to stop').
- **Compound Words:** Multi-root words (compounds) are treated as having multiple motivating simplexes and are segmented accordingly.