Models for Automatic Morphological Inflection of Serbian and Croatian Based on the srLex and hrLex Morphological Lexicons

Scientific paper DOI: 10.18485/judig.2025.1.ch2

Jaka Čibej[1] *0000-0002-3037-6848*

Abstract

Open-source machine-readable morphological lexicons are useful for morphosyntactic tagging of corpora and represent a crucial step toward compiling modern digital dictionary databases. In the paper, we present the first step toward extending the functionalities of Pregibalnik, a custom developed open-source tool for Slovene lexicon expansion, to cover Serbian and Croatian and help automatically expand the lexicons with new entries. We describe the process of extraction of morphological patterns from the hrLex and srLex inflectional lexicons of Croatian and Serbian, as well as a robust process of feature selection based on ending word parts. The features are used to develop a series of machine-learning models to predict morphological patterns for Croatian and Serbian lexemes, achieving an average F1-micro score of 0.85 (depending on lexeme type). This also helps identify potential inconsistencies within the current versions of the lexicons. The extracted patterns and models are available under a Creative Commons CC-BY 4.0 license.

Keywords: lexicon, morphology, inflection, lexicon expansion, Croatian, Serbian

1. Introduction

Open-source machine-readable morphological lexicons are not only helpful for human users (particularly for those studying highly inflectional languages as a second language) but are a useful resource for a wide range of tasks in natural language processing and computational linguistics. They can be

[1] Centre for Language Resources and Technologies, University of Ljubljana, Faculty of Arts, University of Ljubljana, jaka.cibej@ff.uni-lj.si

used to improve morphosyntactic tagging of corpora and represent a crucial step toward compiling modern digital dictionary databases. An example is the *Digital Dictionary Database of Slovene* (DDDS; Kosem et al., 2021), an open-access lexicographic relational database that is being developed at the Centre for Language Resources and Technologies of the University of Ljubljana. The morphological basis for DDDS is the *Sloleks Morphological Lexicon of Slovene* (Čibej et al., 2022). In the RSDO (Development of Slovene in a Digital Environment)[2] project, version 2.0 with approximately 100,800 lexemes was updated to version 3.0 by adding approximately 265,000 new lexemes from the *Gigafida 2.0 Corpus of Written Standard Slovene* (Krek et al., 2020), along with their inflected forms, accentuated forms, and IPA/SAMPA pronunciations. All were automatically generated using *Pregibalnik*[3] ("Inflector" in English; from the Slovene verb *pregibati* 'to inflect'), a custom-developed open-source tool for Slovene lexicon expansion (more on this in Section 2). Sloleks is also used in the development of the Slovene CLASSLA-Stanza models for lemmatization (Terčon et al., 2023) and morphosyntactic tagging (Ljubešić et al., 2023).

Two open-source lexicons similar to Sloleks have been published for Serbian and Croatian – srLex 1.3 (Ljubešić 2019a) and hrLex 1.3 (Ljubešić 2019b), compiled from srWaC (Ljubešić & Klubička, 2016a) and hrWaC (Ljubešić & Klubička, 2016b) web corpora, respectively. Similar to Sloleks, srLex and hrLex are also used in the Serbian and Croatian CLASSLA-Stanza models, which is why it is important to keep the lexicons up-to-date and extend them with new lexemes. Because Croatian and Serbian are structurally similar to Slovene[4] and because they share a similar infrastructural framework, the same method applied to Slovene data can be used (with some minor adjustments) to extend the functionalities of *Pregibalnik* to also cover Croatian and Serbian. However, while machine-learning methods for lexicon expansion have already been used to predict paradigms for Croatian and Serbian, the results are either not available under an open-access license (see Šnajder, 2013) or are not directly compatible with the infrastructure of *Pregibalnik*: for instance, the machine-readable paradigms used by Ljubešić et al., 2016 were only available in the

[2] RSDO Project Site: https://rsdo.slovenscina.eu/

[3] The code for Pregibalnik is available on Github: https://github.com/clarinsi/SloInflector
Pregibalnik is also available as an API service:
https://orodja.cjvt.si/pregibalnik/redoc
https://orodja.cjvt.si/pregibalnik/docs
https://orodja.cjvt.si/pregibalnik/form-generator/docs
https://orodja.cjvt.si/pregibalnik/form-generator/redoc

[4] In this paper, we treat Serbian and Croatian as completely separate because we use different resources (srLex and hrLex, respectively) to develop their inflectional models. This is a purely pragmatic decision made in accordance with the infrastructure of *Pregibalnik* and is not intended as a reflection of the linguistic continuum in actual language use.

Apertium format,[5] which for instance sometimes does not clearly distinguish between morphological patterns for masculine, feminine, and neuter nouns, which according to the MULTEXT-East Morphosyntactic Specifications (MTE)[6] used by *Pregibalnik* are lexeme-level features that clearly discriminate between morphological patterns.

In this paper, we present the first step toward extending the functionalities of *Pregibalnik* to cover Serbian and Croatian and help automatically expand the lexicons with new lexemes using an easily accessible API service. The paper is structured as follows: we first present the structure of *Pregibalnik* focusing on the form generator component (Section 2), then describe the process of extracting morphological patterns from srLex and hrLex (Section 3) and the features used in predictions (Section 4). We evaluate the developed models (Section 5) and provide a brief qualitative analysis of some of the most frequent misclassifications (Section 6), then conclude the paper with some suggestions for future work (Section 7).

2. Lexicon Expansion with Pregibalnik

Pregibalnik currently consists of three components which can be used separately or as part of a single process: the form generator, the accentuator, and the IPA/SAMPA grapheme-to-phoneme converter. The workflow is shown in Figure 1.

The tool takes a lemma and its MTE lexeme-level morphosyntactic features (e.g. the Slovene word *omikron* 'omicron' noun, common, masculine) as input and first generates a complete paradigm of forms inflected by case, number, tense, etc. (nominative singular *omikron*, genitive singular *omikrona*, dative singular *omikronu*, and so on). This is then forwarded to the other two components to add accentuated forms (*ómikron*) and pronunciations (IPA: /ˈoːmikɾɔn/) as well. In this paper, we focus on form generation for Croatian and Serbian as srLex and hrLex currently only include inflected forms.

The first component of *Pregibalnik* generates the set of forms by first extracting a set of features from the input lemma in the form of a numeric vector. For the Slovene form generation models, the features are mostly based on a linguistically informed list of ending word parts (mostly suffixes used in word formation, e.g. *'acija'* in *liofilizacija* 'lyophilization') as well as several other features, such as the ratio of upper-case and lower-case characters (e.g. to help detect acronyms such as *ZN* (*Združeni narodi* 'United Nations'), which

[5] Croatian-Bosnian-Serbian paradigms are available at: https://sourceforge.net/p/apertium/svn/HEAD/tree/languages/apertium-hbs/apertium-hbs.hbs.metadix

[6] MULTEXT-East Morphosyntactic Specifications for Slovene: https://nl.ijs.si/ME/V6/msd/html/msd-sl.html

are inflected differently compared to other nouns in the same category. The numeric vector is then fed into one of several models (based on the part-of-speech of the relevant lexeme) that predicts the code of the morphological pattern, a blueprint consisting of pairs of MTE morphosyntactic tags and their ending word parts. The pattern is then used to generate the entire paradigm. The workflow is presented in Figure 2.

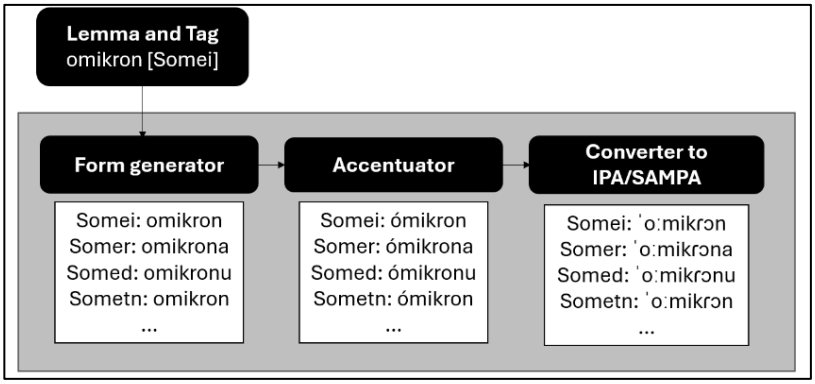

Figure 1: Representation of the *Pregibalnik* workflow for the Slovene masculine common noun *omikron* 'omicron'.

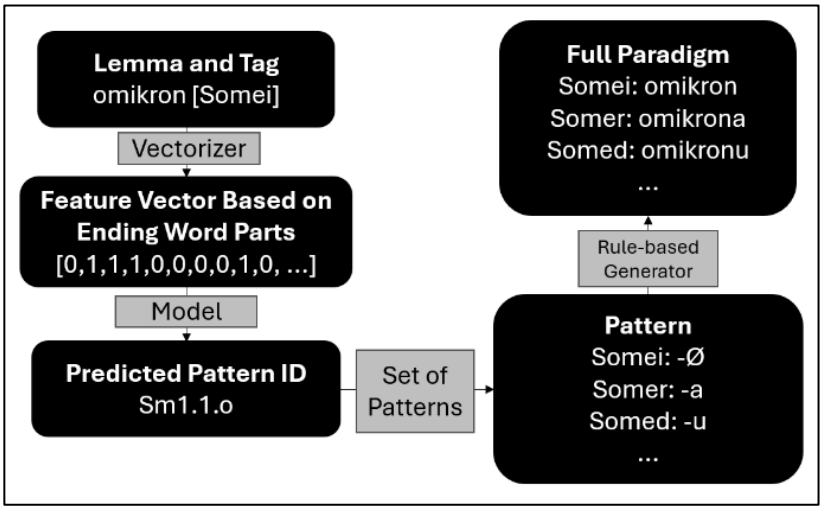

Figure 2: Form generation worfklow in *Pregibalnik*.

The set of machine-readable Slovene morphological patterns were automatically extracted from Sloleks using a robust approach (see Section 3 for a more detailed description on the example of hrLex and srLex). The

patterns were then manually validated and hierarchically sorted (Arhar Holdt & Čibej 2018; Arhar Holdt 2021) before being used in machine-learning predictions.

We performed the same bottom-up process of morphological pattern extraction on hrLex and srLex. However, we test a more agnostic approach with no pre-defined list of word ending parts for pattern predictions and no hierarchization, which requires some additional manual work and linguistic expertise.

3. Morphological Pattern Extraction

The hrLex 1.3 and srLex 1.3 lexicons consist of approximately 164,000 and 169,000 lexemes,[7] respectively, and contain the following data: word forms, their lemmas, morphosyntactic tags and features according to the Serbo-Croatian MULTEXT-East (MTE) Morphosyntactic Specifications[8], morphosyntactic tags and features according to the Universal Dependencies annotation scheme, and the absolute and relative frequencies of the form-lemma-tag combination from the corpus (hrWaC and srWaC for hrLex and srLex, respectively). An excerpt from hrLex is shown in Table 1.

Table 1: Excerpt from hrLex.

Form	Lemma	MTE Tag	MTE Features	UD Tag	UD Features	f_A	f_R
hljeba	hljeb	Ncmsg	Type=common Gender=masculine Number=singular Case=genitive	NOUN	Case=Gen Gender=Masc Number=Sing	588	0.000421
hljeb	hljeb	Ncmsn	Type=common Gender=masculine Number=singular Case=nominative	NOUN	Case=Nom Gender=Masc Number=Sing	269	0.000192
hljebu	hljeb	Ncmsd	Type=common Gender=masculine Number=singular Case=dative	NOUN	Case=Dat Gender=Masc Number=Sing	2	0.000001

The process of extracting morphological patterns was based on a simple algorithm that first searches for all forms pertaining to a lexeme,

[7] A lexeme is an entry in the lexicon, consisting of the lemma form, its inflectional forms, and their morphosyntactic features.

[8] MULTEXT-East Morphosyntactic Specifications for Serbo-Croatian (v6): https://nl.ijs.si/ME/V6/msd/html/msd-hbs.html

then first identifies the immutable part, i.e. the part that is common to all the forms (Table 2).

Table 2: Tags and forms with immutable parts (in bold) for the lexeme abakus (noun, common, masculine) from hrLex 1.3.

Singular Forms	Plural Forms
Ncmsn: **abakus**	Ncmpn: **abakus**i
Ncmsg: **abakus**a	Ncmpg: **abakus**a | **abakus**â
Ncmsd: **abakus**u	Ncmpd: **abakus**ima
Ncmsan: **abakus**	Ncmpa: **abakus**e
Ncmsv: **abakus**e	Ncmpv: **abakus**i
Ncmsl: **abakus**u	Ncmpl: **abakus**ima
Ncmsi: **abakus**om	Ncmpi: **abakus**ima

The immutable part is then removed from the forms to determine the mutable parts for each tag and create the blueprint for the morphological pattern pertaining to the lexeme (Table 3). Each unique morphological pattern is assigned an ID (formatted as P_{lexicon}_{lexeme-level features}_{sequential_number}, e.g. P_hrLex_Ncm_1).

Table 3: Morphological pattern extracted from the lexeme abakus (noun, common, masculine) from hrLex 1.3.

Singular Forms	Plural Forms
Ncmsn: -Ø	Ncmpn: -i
Ncmsg: -a	Ncmpg: -a | -â
Ncmsd: -u	Ncmpd: -ima
Ncmsan: -Ø	Ncmpa: -e
Ncmsv: -e	Ncmpv: -i
Ncmsl: -u	Ncmpl: -ima
Ncmsi: -om	Ncmpi: -ima

The results of the pattern extraction are shown in Table 4. The difference in the number of extracted patterns between hrLex and srLex is immediately apparent, with srLex accounting for three times the number of patterns extracted from hrLex. The reason for this discrepancy lies in the treatment of the combinations of Ekavian and Ijekavian forms in srLex, where both spelling variants are included as part of the same lexeme (e.g. the lexeme *cenovnik* 'price list' contains like *cenovnik* and *cjenovnik*),

which results in a great number (up to 60 %) of single-occurrence patterns (for instance, the immutable part in the lexeme containing both *cenovnik* and *cjenovnik* is *c-*, while the mutable parts are *-enovnik*/*-jenovnik*, which do not fit any other morphological pattern). This raises the question of the manner of including Ekavian and Ijekavian forms in the lexicons. They should arguably be treated as separate lexemes since Ekavian and Ijekavian phenomena are not part of inflectional morphology, but rather variants of lexemes with the same morphological patterns.

Table 4: Extracted morphological patterns from hrLex and srLex.

Lexeme type	Patterns in hrLex 1.3	Patterns in srLex 1.3
Noun, common, masculine (Ncm)	284	552
Noun, common, feminine (Ncf)	81	406
Noun, common, neuter (Ncn)	44	272
Noun, proper, masculine (Npm)	178	178
Noun, proper, feminine (Npf)	47	48
Noun, proper, neuter (Npn)	11	11
Verb, main (Vm)	254	466
Adjective, general (Ag)	173	656
Adjective, possessive (As)	3	361
Adjective, participial (Ap)	24	140
Adverb, general (Rg)	136	616
Adverb, participial (Rr)	44	239
Total	**1,279**	**3,945**

4. Prediction Features Based on Typical Ending Word Parts

To construct the set of features for predicting morphological patterns, we first export frequency lists of ending word parts (1-5-grams) from both lexicons for each lexeme type (common masculine nouns, general adverbs, etc.). We compare the frequency (f_A) of each ending word part (a) within each morphological pattern (P) to its frequency outside the morphological pattern (f_B) to obtain the pattern typicality score (S), which indicates how typical the ending word part is for pattern P:

$$S(a, P) = \frac{f_A + 0.01}{f_B + 0.01}$$

We then calculate the global typicality score (G) of each ending word part by comparing the maximum and average pattern typicalities across morphological patterns:

$$G(a) = \frac{S_{max} + 0.01}{S_{avg} + 0.01}$$

For each lexeme type, we thus obtain a list of ending word parts along with their absolute frequencies and global typicality scores. The higher the ratio between the maximum and average is, the more typical the ending word part is for a specific morphological pattern, which indicates that the ending word part can contribute toward discriminating between different patterns.

Table 5 shows the top 10 ending word parts for common feminine nouns in hrLex 1.3 sorted by frequency. Ending word parts such as *-a*, *-ca*, and *-ica* are less useful for discriminating between patterns, whereas *-ja*, *-ija*, *-cija* on the one hand and *-t*, *-st*, *-ost*, and *-nost* on the other feature higher typicality scores.

Table 5: Top 10 ending word parts for common feminine nouns in hrLex 1.3.

Ending word part	Absolute frequency	Global Pattern Typicality
a	10,203	43.78
t	2,952	79.88
st	2,924	80.51
ost	2,872	80.57
ja	2,805	76.81
ca	2,483	46.83
ica	2,364	50.03
ija	2,323	79.50
nost	2,104	80.53
cija	1,182	80.83

We made a selection of ending word parts for each of the 12 lexeme categories. We removed ending word parts that occur in less than 10 lemmas and kept the first 500 word parts sorted by typicality (or all of the relevant ending word parts if the list contained less than 500 word parts). We compiled two separate vectorizers (one for each language) that use the lists of ending word parts from the relevant lexicon (hrLex or srLex) to construct a numeric vector from the input lemma. The vector of each

relevant lexeme is then paired with the morphological pattern code to compile the training data for machine learning models, which we present in more detail in the following section.

5. Model Training and Quantitative Evaluation

We trained separate models for each lexeme type to avoid any unnecessary misclassification errors on the level of parts-of-speech – a single model trained on all morphological patterns regardless of their lexeme-level features could potentially assign e.g. an adverbial pattern to a verb or vice-versa.

Four model architectures[9] were considered, as shown in Table 6. For each model type and each language, 12 models were trained for each lexeme type, and evaluation scores were aggregated across different patterns. which lists F1-micro scores over all morphological patterns. We list F1-micro scores here to present the overall model performance on the lexicon, not an average across different morphological patterns as some classes are very infrequent and are likely the results of errors in the lexicon rather than linguistic idiosyncrasies that need to be accurately predicted.

Table 6: F1-micro scores for morphological pattern classification in hrLex 1.3 and srLex 1.3.

Model	F1-micro (hrLex 1.3)	F1-micro (srLex 1.3)
k Neighbors Classifier (k=5)	0.8366	0.8317
Linear Support Vector Classifier	0.8534	**0.8553**
Logistic Regression	**0.8607**	0.8507
Multinomial Naïve Bayes Classifier	0.8467	0.8352

The scores were obtained through a 10-fold cross-validation using 80% of the data for training and 20% for testing. Both the training and testing datasets were stratified by morphological patterns. Not all morphological patterns were included as classification classes, however – as previously mentioned (see Table 4 in Section 3), the extraction from the lexicons (particularly srLex) resulted in many patterns that only occur once (approx. 62% of patterns in srLex and 49% of patterns in hrLex). These could not be part of a stratified sample, so they were excluded from the classification process.

It should also be noted that we evaluated the performance of the models based on their ability to correctly predict morphological pattern codes, not individual inflected forms. The scores could potentially be higher if taking

[9] The models were trained using the scikit-learn library in Python (Pedregosa et al., 2011).

into account individual inflected forms – two morphological patterns with completely different pattern codes might in fact share a large number of inflected forms (e.g. patterns for animate or inanimate masculine common nouns, which only differ in the accusative singular form). It can also be difficult to predict the form of vocative singular of masculine nouns from the -a- declension (with the -a ending in the genitive singular form. For instance, unlike the hrLex example *abakuse* (see Table 2), similar nouns also exhibit vocative forms ending with -u: *dinosaurusu*, *glasu*, *fizikusu*, etc. (see Nikolić 2017). Another caveat is that hrLex and srLex are not gold-standard lexicons and were automatically generated, so the evaluations are not to be interpreted as comparisons to a manually annotated dataset, but rather how well the models represent the current state of the lexicons (described in more detail by Ljubešić et al., 2016).

Although the evaluation showed that the Linear Support Vector Classifier performed slightly better on srLex, we opted for Logistic Regression models in the end as that is also the same architecture used by the Slovene form generation models in *Pregibalnik*. In total, 24 final Logistic Regression models were trained in total (on the entire dataset). Their evaluations are shown in Table 7. It should also be noted that models were not developed for certain lexeme types that are not inflected and can be assigned a morphological pattern using a simple rule-based approach (e.g. interjections, conjunctions, abbreviations). The same rationale is applied to the Slovene form generator in *Pregibalnik*.

Table 7: Evaluation scores for Logistic Regression models for different lexeme types.

Lexeme type	hrLex 1.3			srLex 1.3		
	Accuracy	Baseline	F1-micro	Accuracy	Baseline	F1-micro
Ncm	0.65	0.28	0.85	0.64	0.27	0.85
Ncf	0.85	0.45	0.94	0.83	0.45	0.93
Ncn	0.97	0.81	0.99	0.93	0.77	0.95
Npm	0.86	0.43	0.99	0.86	0.43	0.99
Npf	0.93	0.88	0.93	0.92	0.88	0.93
Npn	0.57	0.33	0.85	0.56	0.33	0.85
Ag	0.58	0.33	0.85	0.57	0.33	0.84
Ap	0.99	0.99	0.99	0.93	0.92	0.93
As	0.99	0.99	0.99	0.99	0.99	0.99
Vm	0.70	0.21	0.96	0.64	0.19	0.97
Rg	0.72	0.73	0.72	0.71	0.73	0.72
Rr	0.95	0.86	0.96	0.87	0.81	0.96

All models achieve an above-baseline (majority classifier) accuracy with the exception of participial adjectives and general adverbs. A more detailed qualitative analysis is required to identify the exact root of this issue. However, it appears that in the current versions of both lexicons, many adverbs and adjectives seem to be lemmatized as infinitives of verbs (e.g. the participal adverb *abdicirajući* is tagged as an adverb, but lemmatized as the infinitive *abdicirati* 'to abdicate'; the same with *detonirajući – detonirati* 'to detonate' and *fermentirajući – fermentirati* 'to ferment'). This poses a problem because the lemma form is not present among the actual inflected forms, so the models probably do not learn much from lemma ending word parts. This lemmatization principle is also arguably counter-intuitive for users and introduces unnecessary ambiguities in the lexicon, which might cause more tagging errors if the tagger needs to decide between e.g. *abdicirati* as an adverb, adjective, or verb. This is something that can be addressed in future versions of the lexicons.

6. Preliminary Qualitative Evaluation

Due to space limitations, we only provide a brief preliminary manual evaluation of the performance of the models in this paper and leave a more detailed pattern-by-pattern analysis for future work.

Some classification errors can be attributed to inconsistencies in the lexicons. For instance, the proper masculine noun *Tomislavko* in hrLex features a morphological pattern with only singular forms, whereas the proper masculine noun *Žeško* features both singular and plural forms. The model correctly predicts the full morphological pattern in both cases. In some cases, the morphological pattern extraction revealed that several adjectives and adverbs in both lexicons feature incomplete patterns with only superlative forms, as is the case of *prevaran* 'deceitful'. The model correctly predicts the full pattern, so the classification can be partially used to identify inconsistencies and help with manual corrections.

On the other hand, there are several errors that can be expected due to inherent linguistic ambiguities. As in Slovene, Serbian and Croatian also have the distinction between animate and inanimate masculine nouns. Animacy is hard to predict for a simple model based simply on lemma-based features, so animate nouns are frequently misclassified as inanimate and vice versa. A similar issue occurs with adjectives and adverbs, for which the model has difficulties determining whether the pattern should feature gradation (with comparative and superlative forms) or not. These problems have also been encountered in Slovene models. In the future,

these issues will be addressed using post-processing methods that confirm morphological patterns with data in corpora, or large language models that can potentially fill the gaps of simple machine-learning models.

7. Conclusion and Future Work

In the paper, we presented the extraction of morphological patterns from the srLex and hrLex inflectional lexicons of Serbian and Croatian, and the development of open-access models for the automatic generation of inflected forms for Serbian and Croatian lexemes based on the MULTEXT-East morphosyntactic specifications. Both the extracted morphological patterns and the models are available on Github[10] under the Creative Commons BY-SA 4.0 license.

The models can be used to expand the lexicon with additional lexemes from corpora. In our future work, we will implement the models into *Pregibalnik* to make them available as an API service. The extracted morphological patterns provide a good basis for a more thorough linguistic analysis, and the patterns can be further hierarchized (similar to Arhar Holdt & Čibej, 2018) and finally included as additional metadata into srLex and hrLex. Before manual validation, the patterns can be compared to the Apertium format patterns provided by Ljubešić et al. (2016) to identify similarities and discrepancies. Overall, the methodology to extract patterns and develop models is relatively language-independent and can also be applied to other languages (South Slavic or otherwise).

As a side-product, the analysis has also provided a list of potential inconsistencies in the existing version of the lexicons (e.g. the list of patterns occurring only once), which can be used in future manual validation campaigns to prioritize the most problematic lexemes.

Acknowledgment

The work presented in the paper was supported by the *COST Action CA21167 – Universality, Diversity, and Idiosyncrasy in Language Technology* (UniDive). The author also acknowledges the financial support from the Slovenian Research and Innovation Agency (research core funding No. P6-0411 – *Language Resources and Technologies for Slovene*). A sincere word of gratitude also goes to the anonymous reviewers for their constructive comments.

[10] Github repository:
https://github.com/jakacibej/judig2024_morphological_inflection_srlex_hrlex

References

[1] Arhar Holdt, Špela & Jaka Čibej. "Oblikoslovni vzorci v leksikonu Sloleks: izhodiščni nabor za samostalnike." *Slovenščina 2.0: Empirične, Aplikativne in Interdisciplinarne Raziskave, 6(2)*. (2018): 33–66. https://doi.org/10.4312/slo2.0.2018.2.33-66

[2] Arhar Holdt, Špela. "Oblikoslovni vzorci za strojno procesiranje slovenščine." Arhar Holdt, Špela (ed.): *Nova slovnica sodobne standardne slovenščine: viri in metode*. Založba Univerze v Ljubljani. (2021): 87–124. https://doi.org/10.4312/9789610605478

[3] Čibej, Jaka, Kaja Gantar, Kaja Dobrovoljc, Simon Krek, Peter Holozan, Tomaž Erjavec, Miro Romih, Špela Arhar Holdt, Luka Krsnik, Marko Robnik-Šikonja. "*Morphological lexicon Sloleks 3.0*." Slovenian language resource repository CLARIN.SI, ISSN 2820-4042. (2022). http://hdl.handle.net/11356/1745.

[4] Kosem, Iztok, Simon Krek, Polona Gantar. "Semantic data should no longer exist in isolation: the Digital Dictionary Database of Slovenian." *EURALEX XIX, Congress of the European Association for Lexicography, Lexicography for inclusion*. (2021): 81–83.

[5] Krek, Simon, Špela Arhar Holdt, Tomaž Erjavec, Jaka Čibej, Andraž Repar, Polona Gantar, Nikola Ljubešić, Iztok Kosem, and Kaja Dobrovoljc. "Gigafida 2.0: The Reference Corpus of Written Standard Slovene." *Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France. European Language Resources Association*. (2020): 3340–3345.

[6] Ljubešić, Nikola, Filip Klubička, Željko Agić, Ivo-Pavao Jazbec. "New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian." *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia. European Language Resources Association (ELRA). (2016): 4264–4270. https://aclanthology.org/L16-1676/

[7] Ljubešić, Nikola, Filip Klubička. "Croatian web corpus hrWaC 2.1." Slovenian language resource repository CLARIN.SI, ISSN 2820-4042. (2016b). http://hdl.handle.net/11356/1064.

[8] Ljubešić, Nikola, Filip Klubička. "Serbian web corpus srWaC 1.1." Slovenian language resource repository CLARIN.SI, ISSN 2820-4042. (2016a). http://hdl.handle.net/11356/1063.

[9] Ljubešić, Nikola, Luka Terčon, Jaka Čibej. "The CLASSLA-Stanza model for morphosyntactic annotation of standard Slovenian 2.0." Slovenian language resource repository CLARIN.SI, ISSN 2820-4042. (2023). http://hdl.handle.net/11356/1767.

[10] Ljubešić, Nikola. "*Inflectional lexicon hrLex 1.3*." Slovenian language resource repository CLARIN.SI, ISSN 2820-4042. (2019b). http://hdl.handle.net/11356/1232

[11] Ljubešić, Nikola. "*Inflectional lexicon srLex 1.3*." Slovenian language resource repository CLARIN.SI, ISSN 2820-4042. (2019a). http://hdl.handle.net/11356/1233

[12] Nikolić, Miroslav B. "Облици вокатива једнине именица мушког рода I врсте у српском књижевном језику". *Српски језик: студије српске и словенске. - Vol. 22, No. 1* (2017): 5–34.

[13] Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, vol. 12*. (2011): 2825–2830. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

[14] Šnajder, Jan. "Models for predicting the inflectional paradigm of Croatian words." *Slovenščina 2.0, 1 (2)*, (2013): 1–34. http://www.trojina.org/slovenscina2.0/arhiv/2013/2/Slo2.0_2013_2_02.pdf.

[15] Terčon, Luka, Jaka Čibej, Nikola Ljubešić. "The CLASSLA-Stanza model for lemmatisation of standard Slovenian 2.0." Slovenian language resource repository CLARIN.SI, ISSN (2023): 2820–4042. http://hdl.handle.net/11356/1768.

Modeli za automatsku morfološku fleksiju srpskog i hrvatskog jezika na osnovu morfoloških leksikona srLex i hrLex

Jaka Čibej

Sažetak

Mašinski čitljivi morfološki leksikoni otvorenog koda korisni su za morfosintaksičko označavanje korpusa i predstavljaju ključni korak ka sastavljanju savremenih baza podataka digitalnih rečnika. U radu predstavljamo prvi korak ka proširenju funkcionalnosti *Pregibalnika*, prilagođenog alata otvorenog koda za proširenje slovenačkog leksikona, tako da pokrije srpski i hrvatski jezik i pomoći će automatskom proširenju leksikona novim unosima. Opisujemo proces izdvajanja morfoloških obrazaca iz hrLex i srLex morfoloških leksikona hrvatskog i srpskog jezika, kao i robustan proces selekcije atributa na osnovu završnih delova reči. Atributi se koriste za razvoj serije modela mašinskog učenja za predviđanje morfoloških obrazaca za hrvatske i srpske lekseme, postižući prosečan F1-mikro rezultat od 0,85 (u zavisnosti od tipa lekseme). Ovo takođe pomaže da se identifikuju potencijalne nedoslednosti unutar trenutnih verzija leksikona. Izvučeni obrasci i modeli dostupni su pod licencom Creative Commons CC-BY 4.0.

Ključne reči: leksikon, morfologija, fleksija, proširenje leksikona, hrvatski, srpski