Predicting Pronunciation Types in the Sloleks Morphological Lexicon of Slovene

Jaka Čibej[1,2]
jaka.cibej@ff.uni-lj.si
jaka.cibej@ijs.si
[1]Faculty of Arts, University of Ljubljana
[2]Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

We present an experiment dealing with the automatic prediction of pronunciation types for lemmas in the *Sloleks Morphological Lexicon of Slovene*. We perform a statistical analysis on a number of mostly *n*-gram-based features and use a set of statistically significant features to train and test several machine learning models to discriminate between lemmas for which a phonetic transcription can be generated automatically using Slovene grapheme-to-phoneme (G2P) conversion rules (e.g. *Novak*), and lemmas with pronunciation that follows other G2P rules (e.g. *Shakespeare*).

Keywords

grapheme-to-phoneme conversion, pronunciation types, morphological lexicon, proper nouns, Slovene

1 Introduction

The *Sloleks Morphological Lexicon of Slovene* [2] is the largest open-access database containing machine-readable information on the morphological properties of Slovene lemmas (e.g. *miza* 'table', noun, common, feminine) and their inflected forms (e.g. *mize*, singular, genitive; *mizo*, singular, accusative). Since version 2.0 [3], each lemma and inflected form also contains accentuated forms (e.g. *míza*) and phonetic transcriptions in the International Phonetic Alphabet (IPA) and its equivalent X-SAMPA (e.g. IPA: /ˈmiːza/, X-SAMPA: /"mi:za/). Both transcriptions were generated automatically from accentuated forms, first in version 2.0 using a rudimentary rule-based system, then again in 3.0 with a greatly improved and linguistically informed rule-based grapheme-to-phoneme (G2P) conversion tool for Slovene.[1]

Rule-based G2P conversion for Slovene (particularly from accentuated forms) yields very good results and leaves only a minority of issues to be resolved manually because in terms of its orthographic depth, Slovene features a shallow orthography ([9]) in which each grapheme in the alphabet generally corresponds to one phoneme (see e.g. [4]) and the spelling-sound correspondence is relatively direct ([1]; [11]): the pronunciation rules allow for words to be pronounced correctly based on their graphemic representation, with some exceptions and several predictable phoneme assimiliations (such as the assimilation of voiceless consonant phonemes to their voiced equivalents *glasba* 'music', IPA: /ˈglaːzba/, or vice-versa, voiced-to-voiceless, *podpreti* 'to support', IPA: /pɔtˈpreːti/).

However, not all entries in Sloleks follow Slovene G2P principles. For a number of words, particularly proper nouns denoting people (*Shakespeare*, *Sharon*), locations (*Sydney*, *Birmingham*), inhabitants (*Newyorčan* 'New Yorker'), etc.; as well as adjectives derived from proper nouns (*aachenski* 'pertaining to Aachen', *Acronijev* 'belonging to Acroni'), the phonetic transcription cannot be generated using Slovene G2P rules. In such cases with foreign orthographic elements that indicate relations between graphemes and phonemes that are unusual for Slovene, Slovene linguistic and lexicographic practice (see e.g. [5]) first requires a transliteration into the closest equivalent using Slovene graphemes, which can then be used to generate the phonetic transcription using Slovene G2P rules (e.g. *Newyorčan* → *njújórčan* → IPA: /ˈnjuːˈjoːrtʃan/).

Because of this, it is necessary to discriminate between different *pronunciation types*: categories of words that follow Slovene G2P rules (*Slovene G2P*) and those that do not (e.g. *Other G2P*; more on this in Section 2). Pronunciation types denote the manner in which the phonetic transcription of the word can be generated. In some cases, assigning the pronunciation type to a lemma is trivial – if the lemma contains a grapheme that is not part of the Slovene alphabet[2] (e.g. *x*, *y*, *w*, *q*), it belongs into the *Other G2P* category (e.g. *Byron*, *Oxford*). There are, however, many exceptions that belong in the *Other G2P* category despite being comprised entirely of Slovene graphemes (e.g. *Matt*, *Sharon*).

In Sloleks 3.0, the first cca. 100,000 lemmas that had been part of version 2.0 were manually annotated with pronunciation types, whereas the 264,000 new entries (added automatically from the *Gigafida 2.0 Corpus of Modern Standard Slovene* [6]) still lack this information. Because manual annotation from scratch is time-consuming, we performed an experiment to determine to what degree the pronunciation type can be predicted automatically by relying on the scarce linguistic and morphosyntactic information that can be extracted from an individual lemma.

The paper is structured as follows: we describe the dataset that was used for the statistical analysis and machine learning experiment (Section 2), as well as the process of feature selection (Section 3). We train several machine-learning models and evaluate their performance using 10-fold cross-validation (Section 4). Finally, we manually evaluate a sample of automatically annotated entries (Section 5) and conclude the paper with our plans for future work (Section 6).

[1]The Slovene G2P tool is part of *Pregibalnik*, a piece of software used for the automatic expansion of the *Sloleks Morphological Lexicon of Slovene*: https://github.com/clarinsi/SloInflector It was developed within the *Development of Slovene in the Digital Environment* project. The Slovene G2P converter is also available as an API-service: https://orodja.cjvt.si/pregibalnik/g2p/docs

[2]Although *ć* and *đ* are not part of the Slovene alphabet, they are phonemically transparent and frequently occur in names of Slovene citizens, so they are not counted as foreign characters for the purposes of this task.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/https://doi.org/10.70314/is.2024.sikdd.2

Table 1: Lemmas in Sloleks 3.0 by Pronunciation Type

Pronunciation Type	Frequency	%
-	264,538	72.41%
Slovene G2P	94,750	25.93%
Other G2P	3,066	0.84%
Numeral	1,840	0.50%
Acronym	845	0.23%
Slovene G2P with minor deviation	113	0.03%
Abbreviation	70	0.02%
Ambiguous G2P	69	0.02%
Symbol	49	0.01%
Total	**365,340**	**100.00%**

Table 2: Lemmas in Sloleks 3.0 with *Other G2P* pronunciation type by Morphosyntactic Properties

Morphosyntactic Properties	Frequency	%
Adjective, possessive	1,092	35.62%
Noun, proper, masculine	958	31.25%
Noun, proper, feminine	713	23.26%
Adjective, general	142	4.63%
Noun, common, masculine	127	4.14%
Noun, common, feminine	20	0.65%
Adverb, general	10	0.33%
Noun, common, neuter	2	0.07%
Verb, main, imperfective	2	0.07%
Total	**3,066**	**100.00%**

2 Dataset

Sloleks 3.0 contains a total of 365,340 entries, but only approximately 28% have been manually assigned one of 8 pronunciation types[3] (as shown in Table 1). For the classification task, we focus only on the two most frequent pronunciation types (*Other G2P* and *Slovene G2P*). [4]

In terms of their morphosyntactic features, the *Other G2P* lemmas mostly consist of possessive adjectives and proper nouns, collectively accounting for cca. 90% of the category (as shown in Table 2), but only 15% of the portion of Sloleks annotated with pronunciation types.

The final dataset for statistical analysis and machine learning consisted of 94,863 *Slovene G2P* lemmas (e.g. *dekadentnost*, *Košak*, *prefiltriran*) and 3,066 *Other G2P* lemmas (e.g. *Elizabeth*, *Presley*, *Sinclaire*).

3 Statistical Analysis and Feature Selection

From each lemma, we extracted a series of features that could help discriminate between the two classes: (a) percentage of *Slovene G2P* graphemes within the lemma (i.e. graphemes of the Slovene alphabet as well as *ć* and *đ*); (b) morphosyntactic features (e.g. *noun*, *proper*, *masculine*); (c) relative frequencies[5] of character-level uni-, bi-, and trigrams within the lower-cased lemma (e.g. *Matt* $\rightarrow f_r(m)$, $f_r(a)$, ..., $f_r(ma)$, $f_r(at)$, ..., $f_r(mat)$, ...); (d) relative frequencies of character-level uni-, bi-, and trigrams from a robust CVC-conversion of the lemma, substituting consonant graphemes with *C* and vowel graphemes with *V* (e.g. *Matt* $\rightarrow$ *CVCC* $\rightarrow f_r(C)$, $f_r(V)$, ..., $f_r(CV)$, $f_r(VC)$, ..., $f_r(CVC)$, ...); (e) relative frequencies of character-level uni-, bi-, and trigrams from a finegrained CVC-conversion of the lemma[6] (e.g. *Matt* $\rightarrow$ *ZVKK* $\rightarrow f_r(Z)$, $f_r(V)$, ..., $f_r(ZV)$, $f_r(VK)$, ..., $f_r(ZVK)$, ...)

For (c), (d), and (e), the initial and final uni-, bi-, and trigrams of the lemma were extracted separately as well, as in some cases the position of the *n*-gram in the word can be indicative of one class over another.

For general character-level *n*-grams, the first 1,498 with a frequency of at least 500 across all Sloleks 3.0 lemmas were analyzed; these cover cca. 88.34% of all *n*-gram occurrences. For robust CVC and finegrained CVC *n*-grams, all were analyzed. We performed the Kruskal–Wallis H test [7] (k=2, n=97,056) on a total of 6,148 features, out of which 2,490 (40%) were statistically significant.[7] Statistically significant features by categories are shown in Table 3. 1,146 features are more indicative of *Slovene G2P* and 1,344 are more indicative of *Other G2P*. As shown in Table 4, only three of the top 10 general *n*-grams indicative of *Other G2P* actually contain non-*Slovene G2P* characters, confirming that detecting lemmas from the *Other G2P* category is more complex and requires more than simply taking into account non-*Slovene G2P* graphemes.

Table 3: Statistically Significant Features by Category

Feature Category	Number
Percentage of *Slovene G2P* characters	1
Morphosyntactic features	3
General character-level *n*-grams	1,119
Initial character-level *n*-grams	398
Final character-level *n*-grams	468
General robust CVC *n*-grams	66
Initial robust CVC *n*-grams	44
Final robust CVC *n*-grams	39
General finegrained CVC *n*-grams	157
Initial finegrained CVC *n*-grams	102
Final finegrained CVC *n*-grams	93
Total	**2,490**

4 Pronunciation Type Prediction

The identified features (along with several placeholder *n*-grams to take into account any graphemes not covered in the initial dataset) were taken into account to develop a custom vectorizer that converts a given lemma and its lexical features based on the MulText-East v6 (MTE-6) Morphosyntactic Specifications for

[3]It should be noted that all the inflected forms within the entry effectively inherit the pronunciation type.

[4]Symbols in Sloleks are rare, along with entries within the *Ambiguous G2P* category (where an entry can either follow Slovene G2P rules or not, depending on the context – e.g. *Amanda* as a Slovene name: /amˈaːnda/; or as an English name with a pronunciation adjusted to the Slovene set of phonemes: /əmˈɛːnda/). Abbreviations and numerals are easily identifiable, and while acronyms have a separate manner of generating phonetic transcriptions which also depends on their morphological patterns, they are also mostly identifiable with rules. Because of its rarity and similarity to *Slovene G2P*, the *Slovene G2P with minor deviation* category was merged into *Slovene G2P* for the classification task.

[5]Relative frequencies were calculated as $f_r(x_n) = f_a(x_n)/\sum f_a(y_n)$, e.g. the absolute frequency of *n*-gram *x* of length *n* within the lemma divided by the sum of absolute frequencies of each *n*-gram *y* of length *n* within the lemma.

[6]In the finegrained CVC-conversion, consonant graphemes were generalized into more finegrained categories, e.g. graphemes denoting Slovene sonorants (M), voiced (G) and voiceless obstruents (K), foreign consonants (X), etc.

[7]Effect size was calculated as $\eta^2 = (H - k + 1)/(n - k)$, as reported in [10].

Table 4: Top 10 Statistically Significant General Character-Level *n*-Grams by Effect Size (η^2)

n-Gram	H	p	η^2	Means
y	11509.36	$p \leq 0.0001$	0.1186	$\mu_S < \mu_O$
w	9595.25	$p \leq 0.0001$	0.0989	$\mu_S < \mu_O$
ch	7558.60	$p \leq 0.0001$	0.0778	$\mu_S < \mu_O$
ll	6295.96	$p \leq 0.0001$	0.0649	$\mu_S < \mu_O$
ss	3804.26	$p \leq 0.0001$	0.0392	$\mu_S < \mu_O$
nn	3220.65	$p \leq 0.0001$	0.0332	$\mu_S < \mu_O$
th	2973.89	$p \leq 0.0001$	0.0306	$\mu_S < \mu_O$
wa	2761.53	$p \leq 0.0001$	0.0284	$\mu_S < \mu_O$
tt	2745.10	$p \leq 0.0001$	0.0283	$\mu_S < \mu_O$
co	2571.20	$p \leq 0.0001$	0.0265	$\mu_S < \mu_O$

Table 5: Model Performance Based on 10-Fold Cross-Validation

Model	A	BA	P	R	F1	ROC AUC
LinearSVC	**99.08**	**87.87**	**96.36**	**87.87**	**91.64**	**98.89**
Multin. NB	97.38	79.17	78.12	79.17	78.62	96.55
kNN (k=5)	98.25	75.17	93.67	75.17	81.74	91.63
Majority	96.87	-	-	-	-	-

Slovene[8] into a 2,500-dimensional numerical vector. The entire dataset was converted into vectors and split into a training set (80%) and a test set (20%), both stratified by class. Three models[9] (Linear Support Vector Classifier (LinearSVC), Multinomial Naive Bayes Classifier (Multin. NB), and k Nearest Neighbors Classifier (kNN)) were trained and evaluated with 10-fold cross-validation. The results are listed in Table 5[10] and show that LinearSVC outperforms the other two models. All three exhibit above-baseline accuracy compared to the majority classifier, but Multinomial NB and kNN perform much worse in terms of balanced accuracy as well as precision and, in case of kNN, recall. Recall is also somewhat lower with LinearSVC, which is to be expected – some *Other G2P* lemmas might contain no indicative *n*-grams and are thus hard to detect; on the other hand, once identified, the model is very precise in its prediction.

Table 6 shows the confusion matrix for the LinearSVC model tested on the 20% stratified test dataset. The model very rarely misclassifies *Slovene G2P* lemmas, and more frequently errs with *Other G2P* lemmas. A closer inspection of the misclassified *Slovene G2P* examples reveals several errors in the original dataset: *Beethoven*, *Ratzinger*, *Rotterdam*, *Franco*, *Oberstdorf*, and *Keller* were in fact correctly classified as *Other G2P*, but they are miscategorized as *Slovene G2P* in the original dataset. Other misclassifications include examples of foreign proper nouns and possessive adjectives that contain unusual grapheme combinations for Slovene (e.g. *Andreas*, *Aurelio*, *Hilton*, *Simpsonov*), but their pronunciation can still be derived from their graphemic representation (e.g. *Andreas* → IPA: /and'reːas/).

On the other hand, *Other G2P* lemmas misclassified as *Slovene G2P* include *Andersonov*, *Atkinsov*, *Batmanov*, in which the grapheme

[8]MTE-6: https://nl.ijs.si/ME/V6/msd/html/msd-sl.html The vectorizer uses Slovene morphosyntactic tags, e.g. *Slz* (S – noun, l – proper, z – feminine).
[9]All models were trained using the Python library *scikit-learn*. [8]
[10]A, BA, P, R, and F1 refer to accuracy, balanced accuracy, macro-precision, macro-recall and macro-F1, respectively.

Table 6: Confusion Matrix for Linear Support Vector Classifier

True → ↓ Predicted	Slovene G2P	Other G2P	∑
Slovene G2P	**18,939**	140	19,079
Other G2P	34	**473**	507
∑	18,973	613	-

Table 7: Confusion Matrix for Manual Evaluation

True → ↓ Predicted	Slovene G2P	Other G2P	∑
Slovene G2P	**86**	9	95
Other G2P	14	**91**	105
∑	100	100	-

'a' is pronounced as /ɛ/, but this cannot be discerned from the graphemic representation itself. Other misclassified examples are more obviously pertaining to *Other G2P*, e.g. *Dorfmeister*, *Faulknerjev*, *Flaubertov*, *Heisenbergov*, *Balfourjev*. This might indicate that not all indicative *n*-grams have been included as features (e.g. 'ei', 'ou'), possibly for lack of evidence in the original dataset or because they are less frequent and have not been included in the initial batch of statistical tests. As the lexicon expands with new entries, the model will be updated with new examples and new features to potentially improve performance.

5 Manual Evaluation

We trained a new instance of the LinearSVC model on the entire dataset and used it to annotate the remaining cca. 264,000 lemmas from Sloleks 3.0 with no pronunciation type, resulting in 86,730 lemmas with *Other G2P* and 177,808 lemmas with *Slovene G2P*.

We performed a preliminary manual evaluation consisting of a random sample of 100 examples from each class. The results are shown in the confusion matrix in Table 7. Although the sample is too small to be representative of the whole, it indicates that the model performs well even on unseen data, achieving an accuracy of 88.50% (P=0.91, R=0.87, F1=0.89) over a majority baseline accuracy of 50.00%.

The misclassifications of *Other G2P* as *Slovene G2P* include examples such as *Mukhamedov*, *Beatli*, *Livenza*, and *Preidler*, with limited indicators that the words belong to the *Other G2P* category. Most graphemes in these examples are pronounced according to *Slovene G2P* criteria, with the exception of individual *n*-grams ('nz', 'ei', 'kh'), some of which have not been included in the set of features. In other examples, only one or two vowel graphemes are indicative of *Other G2P* pronunciation (e.g. *Trendlina*, which is also a lemmatization error; the correct lemma is *Trendline*; and *Sanberg*), and the pronunciation of single vowel graphemes appears harder to predict than consonant graphemes or combinations thereof.

Similarly, the misclassifications of *Slovene G2P* lemmas as *Other G2P* lemmas include examples such as *Doneck*, *Barson*, *Bronson*, *Piersanti*, and *Faustini*. While these are proper nouns of foreign origin, their Slovene pronunciation can either be fully discerned from their graphemic representation (e.g. *Doneck* → IPA: /dɔ'neːtsk/), or it only differs slightly from what Slovene

grapheme-to-phoneme conversion would produce (e.g. *Faustini* → automatically converted IPA: /faus'ti:ni/; correct IPA: /fa͡us'ti:ni/).

6 Conclusion

In the paper, we presented the results of an attempt to automatize the assignment of pronunciation types to lemmas in the *Sloleks Morphological Lexicon of Slovene*. The results show that a model based on a series of mostly *n*-gram features can provide good results when discriminating between *Slovene G2P* and *Other G2P* categories, with the best performance achieved by the Linear Support Vector Classifier. However, there is still room for improvement, particularly in terms of recall – a number of *Other G2P* lemmas from the test set were misclassified as *Slovene G2P*, while those classified correctly were classified with a relatively high precision score. *n*-grams that are statistically significant as indicative of one class have proven to be useful features for model development, but because they are not evenly distributed and occur sporadically in different lemmas, it would make sense to further improve the model by performing the same statistical analysis (as described in Section 3) on the long tail of less frequent *n*-grams to prepare a more comprehensive list of indicative *n*-grams. The current version of the model is very light-weight and additional features should not cause the model to become overencumbered.

There are several possibilities for further development of the model. Firstly, instead of using relative frequencies of *n*-grams as features, it would be useful to test how different measures such as TF–IDF, absolute frequencies, or even Boolean values influence the performance of the model, and potentially also test several other machine learning algorithms (e.g. Random Forest Classifier). Secondly, while the other pronunciation types from Sloleks 3.0 (acronyms, abbreviations, etc.) are relatively easily identifiable (but much less frequent), in the next step, it would be informative to include them in the training set and test out the model's performance on the full set of categories. Thirdly, a statistical analysis should be performed on the probabilities with which the model makes decisions and to what degree they correlate with the percentage of graphemes that differ from the shallow orthographical Slovene G2P rules (e.g. *Anderson*, with arguably only 'a' not following Slovene G2P rules; vs. *Châteaux*, where the majority of graphemes are pronounced completely differently compared to Slovene G2P rules). This would require the preparation of a separate dataset in which graphemes are manually aligned to either the graphemes of their transliterated Slovene graphemic forms (*Newyorčan* → *njújórčan*) or their Slovene IPA transcriptions. By assigning scores that reflect the degree of orthography depth for the individual lemma, it would be possible to use the dataset to train a regression model.

Similarly, *Other G2P* lemmas from Sloleks 3.0 can be manually annotated with their language of origin and transliterated according to the recently published transliteration rules of *Pravopis 8.0*,[11] the new orthographic manual of Slovene, which at the time of writing this paper is still in development. Such a dataset would enable the development of a model for language identification for individual lemmas, and, ultimately, a model for automatizing transliteration of lemmas of foreign origin into their Slovene equivalents. As of now, no such tool yet exists for Slovene, and even the new orthographic manual anticipates that all transliteration should be done manually, which begs the question whether at least part of the work can be automatized. This would be an important step in the development of a modern, digital infrastructure for Slovene orthography, and would facilitate the automatic expansion of modern digital dictionary databases and datasets for automatic speech recognition.

In addition, although our preliminary experiments with LLMs (ChatGPT 3.5 and 4.0) classifying *Slovene G2P* and *Other G2P* lemmas have yielded much worse results than the best performing LinearSVC model, more systematic experiments are warranted.

As part of our future work, we intend to implement the model into *Pregibalnik*,[12] which is used for automatically extending the lexicon and currently assigns no pronunciation type. The model itself is available under the Apache 2.0 license on Github[13], while the pronunciation type annotations will be included in future versions of Sloleks and, eventually, manually validated.

Acknowledgements

The research presented in this paper was conducted within the research project titled *Basic Research for the Development of Spoken Language Resources and Speech Technologies for the Slovenian Language* (J7-4642), the research programme *Language Resources and Technologies for Slovene* (P6-0411), and the CLARIN.SI research infrastructure, all funded by the Slovenian Research and Innovation Agency (ARIS). The author also thanks the anonymous reviewers for their constructive comments.

References

[1] Derek Besner and Marilyn Chapnik Smith. 1992. Chapter 3 basic processes in reading: is the orthographic depth hypothesis sinking? In *Orthography, Phonology, Morphology, and Meaning*. Advances in Psychology. Vol. 94. Ram Frost and Leonard Katz, editors. North-Holland, 45–66. DOI: https://doi.org/10.1016/S0166-4115(08)62788-0.

[2] Jaka Čibej et al. 2022. Morphological lexicon sloleks 3.0. Slovenian language resource repository CLARIN.SI. (2022). http://hdl.handle.net/11356/1745.

[3] Kaja Dobrovoljc, Simon Krek, Peter Holozan, Tomaž Erjavec, Miro Romih, Špela Arhar Holdt, Jaka Čibej, Luka Krsnik, and Marko Robnik-Šikonja. 2019. Morphological lexicon sloleks 2.0. Slovenian language resource repository CLARIN.SI. (2019). http://hdl.handle.net/11356/1230.

[4] Florina Erbeli and Karmen Pižorn. 2012. Reading ability, reading fluency and orthographic skills: the case of l1 slovene english as a foreign language students. English. *Center for Educational Policy Studies Journal*, 2(3), 119–139. https://files.eric.ed.gov/fulltext/EJ1130208.pdf.

[5] Nataša Gliha Komac et al. 2015. Koncept novega razlagalnega slovarja slovenskega knjižnega jezika. Inštitut za slovenski jezik Frana Ramovša ZRC SAZU. (2015). https://fran.si/179/novi-slovar-slovenskega-knjiznega-jezika/datoteke/Potrjeni_koncept_NoviSSKJ.pdf.

[6] Simon Krek et al. 2019. Corpus of written standard slovene gigafida 2.0. Slovenian language resource repository CLARIN.SI. (2019). http://hdl.handle.net/11356/1320.

[7] William H. Kruskal and W. Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47, 260, 583–621. eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1952.10483441. DOI: 10.1080/01621459.1952.10483441.

[8] F. Pedregosa et al. 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

[9] Anja Schüppert, Wilbert Heeringa, Jelena Golubovic, and Charlotte Gooskens. 2017. Write as you speak? a cross-linguistic investigation of orthographic transparency in 16 germanic, romance and slavic languages. English. *From semantics to dialectometry*, 32, 303–313. ISBN: 9781848902305.

[10] Maciej Tomczak and Ewa Tomczak. 2014. The need to report effect size estimates revisited. an overview of some recommended measures of effect size. *Trends in Sport Sciences*, 1(21), 19–25.

[11] Antal van den Bosch, Alain Content, Walter Daelemans, and Beatrice de Gelder. 1994. Analysing orthographic depth of different languages using data-oriented algorithms. In *Proceedings of the 2nd International Conference on Quantitative Linguistics*.

[11] *Pravopis 8.0: Pravila novega slovenskega pravopisa za javno razpravo*. https://pravopis8.fran.si/, 9 August 2024

[12] *Pregibalnik*: https://github.com/clarinsi/SloInflector; the entire tool is also available as an API-service: https://orodja.cjvt.si/pregibalnik/docs

[13] GitHub: https://github.com/jakacibej/sikdd2024_predicting_pronunciation_types