A Machine-Learning Approach to Predicting the Pronunciation of Pre-Consonant *l* in Standard Slovene

Jaka Čibej
jaka.cibej@ff.uni-lj.si
Centre for Language Resources and Technologies & Faculty of Arts, University of Ljubljana
Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

The pronunciation of pre-consonant *l* in Slovene words (e.g. *alge*, *polž*, *gledalka*) is not easily predictable (/l/, /u̯/, or both) and poses a problem for the otherwise effective rule-based grapheme-to-phoneme conversion. We present a method to discriminate between the various pronunciations of pre-consonant *l* using machine-learning models trained on vectors of character-level *n*-gram features from approximately 153,500 manually annotated Slovene words with pre-consonant *l* from the *ILS 1.0* dataset. We achieve an accuracy of 86% (over a majority baseline of 76.53%) and conclude the paper with potential steps for future work.

Keywords

pronunciation, grapheme-to-phoneme conversion, pre-consonant *l*, pronunciation ambiguity, Slovene

1 Introduction

In languages that are characterized by greater orthographic depth (i.e., a greater discrepancy between the written form and its pronunciation), such as English or French, grapheme-to-phoneme (G2P) conversion requires more sophisticated methods such as neural networks (see e.g. [10] for French and [14] for English). Slovene, on the other hand, features a much more transparent orthography ([15]; [17]). Phonetic transcriptions of Slovene words – with some exceptions, such as acronyms, symbols, numerals, and certain words of foreign origin (e.g. *sommelier*), including proper nouns (e.g., *Johnson*; more on this in [3]) – can be very reliably generated using a rule-based approach, especially if taking the accentuated form (e.g., *drevó* instead of the unaccentuated *drevo*) as the starting point, as the diacritic disambiguates the position of the accent and the manner of pronunciation of the accentuated vowel grapheme. The *Slovene IPA/X-SAMPA G2P Converter*[1] achieves an accuracy of approximately 98% (based on an evaluation on a stratified sample of words; see [2]).

However, there are several exceptions (in addition to the ones already mentioned) in which the pronunciation of certain graphemes is much more difficult to predict with rules. We focus on one such problem in this paper: the pronunciation of pre-consonant *l* in Slovene words. The grapheme *l*, when preceding a consonant grapheme, can be pronounced as either /l/ or /u̯/. In some cases, both variants are acceptable. Examples include words such as *alge* ('algae', IPA: /ˈaːlgɛ/, but never */ˈaːu̯gɛ/), *polž* ('snail', IPA: /ˈpɔːu̯ʃ/, but never */ˈpɔːlʃ/), *gledalka* ('spectator (female)', IPA: /glɛˈdaːu̯ka/ or /glɛˈdaːlka/), and *decimalka* ('decimal number', IPA: /dɛciˈmaːlka/, but never */dɛciˈmaːu̯ka/). The reasons for these different pronunciations are historic and etymological in some cases, while in others, the difference cannot be easily explained and has more to do with conventions in language use. The issue of pre-consonant *l* has been tackled by Slovene linguistics for more than a century (see [4] for a brief overview). Perception tests and small-scale surveys ([16]; [11]) have recently been conducted to collect data for lexicographic resources (such as the *Slovenian Normative Guide 8.0*),[2] but empirical data remains scarce: relevant language resources are not machine-readable or openly accessible (as is the case of the *Dictionary of Slovenian Literary Language*[3]) or contain inconsistent data (e.g., *OptiLeX* [19]). In this paper, we use the recently published *ILS 1.0* dataset ([1]; described in Section 2).

Because the *Slovene IPA/X-SAMPA G2P Converter* is currently entirely rule-based, all pre-consonant *l* graphemes are transcribed as /l/, resulting in errors that need manual corrections when compiling language resources. Our goal is to implement a machine-learning approach[4] to disambiguate between different pronunciations. Increasing the accuracy of the converter is important in the context of the automatic compilation of modern lexicographic resources that can also be used as machine-readable databases for training models (including large language models) and improving speech recognition and speech synthesis for Slovene. We describe the dataset (Section 2), the statistical analysis used for feature selection (Section 3), the results (Section 4), and several steps for future work (Section 5).

2 Dataset

ILS 1.0 ([1]; described in more detail in [4]) is a dataset of approx. 173,400 inflected Slovene word forms (of approx. 6,000 Slovene lexemes) containing a single pre-consonant *l* grapheme. Each occurrence of pre-consonant *l* was annotated for its pronunciation by 5 linguists (2 annotations per occurrence). The word forms were extracted from the manually validated lexemes of *Sloleks 3.0* [5], the largest open-access dataset with machine-readable morphosyntactic information on Slovene words. Table 1 shows the distribution of word forms by agreement: in 89% of word

[1] The *Slovene IPA/X-SAMPA G2P Converter* is part of *Pregibalnik*, a custom tool that was developed for the expansion of the *Sloleks Morphological Lexicon of Slovene* [5], which is the morphological basis for the *Digital Dictionary Database of Slovene* [8]. *Pregibalnik* is available as open-access code at https://github.com/clarinsi/SloInflector and as an API service at https://orodja.cjvt.si/pregibalnik/docs; the *Slovene IPA/X-SAMPA G2P Converter* is also available as an API at https://orodja.cjvt.si/pregibalnik/g2p/docs.

Information Society 2025, Ljubljana, Slovenia

https://doi.org/10.70314/is.2025.sikdd.1

[2] *Pravopis 8.0* (*Slovenian Normative Guide 8.0*): https://pravopis8.fran.si/
[3] The *Dictionary of Slovenian Literary Language* (SSKJ) is available at https://fran.si/.
[4] An attempt at using machine learning for Slovene phonetic transcriptions was made by [9]; however, the method was evaluated on the *Sloleks Morphological Lexicon of Slovene 3.0* [5], where the issue of pre-consonant *l* is still unresolved.

Table 1: Word forms in *ILS 1.0* by agreement.

Pronunciation	Number of Forms	%
/l/	117,459	67.73
/u̯/	23,884	13.77
Both	12,160	7.01
Both | /l/	11,205	6.46
Both | /u̯/	7,051	4.07
/l/ | /u̯	1,660	0.96
Total	**173,419**	**100.00**

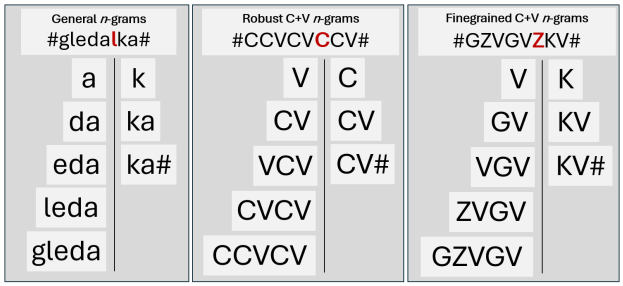

Figure 1: Extraction of character-level *n*-gram features for the pre-consonant *l* in the word *gledalka*.

forms (highlighted in gray), the annotators agree on the pronunciation of pre-consonant *l*. They disagree in 11% of the examples, with one annotator allowing for both pronunciation variants and the other allowing for only one pronunciation. Complete disagreement is present only in less than 1% of the examples.

We use the 153,503 forms with complete agreement as training data for machine-learning models as described in the following sections. It should be noted, however, that while *ILS 1.0* is the largest open-access dataset on pre-consonant *l* pronunciations, it is not completely representative of language use in general (with annotations by only 5 linguists with a background in translation and Slovene studies; these can be biased towards linguistic rules that might not reflect real language use). Despite this, the dataset is robust enough to help disambiguate the more obvious examples (such as *alge*, IPA: /ˈaːlgɛ/, and *polž*, IPA: /ˈpɔːu̯ʃ/).

3 Feature Selection

To some extent, the pronunciation of pre-consonant *l* depends on the preceding and subsequent graphemes,[5] so we use character-level *n*-grams as features for prediction. For each pre-consonant *l* in each word form, we identify the *n*-grams ($1 \leq n \leq 5$) in its direct left/right surroundings as shown in Figure 1 (see footnote 6). We include word boundary markers (#) to discriminate between word-initial and word-final *n*-grams. We also perform the same extraction on robust and finegrained C+V representations of each word form.[6]

[5]The *Slovenian Normative Guide 8.0* (*Pravopis 8.0*, see https://pravopis8.fran.si), for instance, states that a pre-consonant *l* preceded by the grapheme *o* is often characterized by the /u̯/ pronunciation; this is true of words that historically used the syllabic *l* (e.g. *polh* IPA: /ˈpɔːu̯x/ 'dormouse'; *volk* IPA: /ˈvɔːu̯k/ 'wolf'). However, there are exceptions as not all *ol* *n*-grams originate from the syllabic *l* (e.g., *polkovnik* IPA: /pɔlˈkɔːu̯nik/ 'colonel'; *voltaža* IPA: /vɔlˈtaːʒa/ 'voltage').

[6]In the robust C+V form, all consonant graphemes are substituted with *C* and all vowel graphemes with V. In the finegrained C+V form, 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.

Table 2: Contingency table for the general *n*-gram *c* when following a pre-consonant *l*.

Pronunciation → ↓ Presence	/l/	/u̯/	/l/+/u̯/
Yes	2,653	1,847	5,980
No	114,898	22,045	6,180

Table 3: A sample of statistically significant general character-level *n*-grams.

n-Gram	χ^2	p	V	$r_{\|max\|}$	Category
c	38,199.59	****	0.499	178.81, /l/, No	post-*l*
n	29,081.52	****	0.435	79.27, /l/, No	post-*l*
ce	16,003.46	****	0.323	118.30, /l/, No	post-*l*
o	77,025.17	****	0.708	227.83, /l/, No	pre-*l*
po	48,241.29	****	0.560	193.98, /l/, No	pre-*l*
a	16,592.50	****	0.329	-79.85, /l/, No	pre-*l*

We extract a total of 8,082 different general *n*-grams (consisting of actual graphemes; 3,041 in pre-*l* position, 5,541 in post-*l* position), 116 different robust C+V *n*-grams (65 pre- and 51 post-*l*), and 603 different finegrained C+V *n*-grams (262 pre- and 341 post-*l*). For each *n*-gram, we compile a contingency table. For instance, Table 2 shows the occurrences of the general *n*-gram *c* in the position directly following a pre-consonant *l* (e.g., *morilca*, 'murderer', masculine common noun, genitive singular form) depending on the pronunciation of the pre-consonant *l*.

In order to determine statistically significant features that help discriminate between different pronunciations, we performed a series of Pearson's χ^2 tests [12] and corrected for family-wise error rate with the Holm-Bonferroni method [7]. We calculated Cramér's V [6] as the measure of effect size.[7] This resulted in a total of 4,263 statistically significant features (1,856 pre-*l* general and 1,794 post-*l* general *n*-grams; 60 pre-*l* and 40 post-*l* robust C+V *n*-grams; 242 pre-*l* and 271 post-*l* finegrained C+V *n*-grams). Several statistically significant pre-*l* general *n*-grams are shown in Table 3.[8] The table shows the values of the χ^2 statistic and Cramér's V, the p-value representations, the maximum absolute value of Pearson's residuals (and its position in the contingency table), and the category of the *n*-gram (post-*l* or pre-*l*). With the exception of the *a* *n*-gram, which is more indicative of the /l/ pronunciation, the others indicate one of the other two options (/u̯/; or /l/+/u̯/). The results also confirm the statement found in the *Slovenian Normative Guide 8.0* that the *o* grapheme in pre-*l* position is strongly indicative of the /u̯/ pronunciation.

4 Prediction and Evaluation

We compiled a custom vectorizer based on the identified features. The vectorizer scans each input word form (along with its Multext-East v6 morphosyntactic tag[9]) for all occurrences of

[7]We calculate Cramér's V as $\sqrt{\frac{\chi^2}{N*d_{min}}}$, where χ^2 is the Pearson's χ^2 statistic, N is the total sample size, and d_{min} is the minimum dimension of the contingency table.

[8]For all tests, the degrees of freedom (df) were equal to 2 and the total sample size (N) was equal to 153,603. The p-values should be interpreted in the following manner: **** → $p \leq 0.0001$; *** → $p \leq 0.001$; ** → $p \leq 0.01$; * → $p < 0.05$

[9]Multext-East v6 Morphosyntactic specifications: https://nl.ijs.si/ME/V6/msd/html/msd-sl.html

Table 4: Model performance based on 10-fold cross-validation.

Model	A	BA	P	R	F1
LinearSVC	**86.08**	72.39	**69.26**	55.39	61.54
Multin. NB	77.29	69.54	33.33	**81.84**	47.36
kNN (k=5)	85.91	**73.30**	64.11	62.98	**63.53**
Majority	76.53	-	-	-	-

Table 5: Confusion matrix for the Linear Support Vector classifier.

True → ↓ Predicted	/l/	/u̯/	/l/+/u̯/	∑
/l/	**22,006**	1,495	729	24,230
/u̯/	1,071	**2,764**	31	3,866
/l/+/u̯/	434	519	**1,672**	2,625
∑	23,511	4,778	2,432	-

pre-consonant *l*, extracts the surrounding *n*-grams, converts the morphosyntactic tag into 146 morphosyntactic features, and represents the occurrence as a 4,409-dimensional vector of {0,1} values (with 0 and 1 indicating the absence or presence, respectively, of the *n*-gram in the direct surroundings of the pre-consonant /l/). We compile a total of 153,503 vectors in this way and use the *scikit-learn* Python library [13] to train several models for a classification task with three classes: the goal is to correctly predict whether a pre-consonant *l* is pronounced as /l/, /u̯/, or both.

4.1 Automatic Evaluation

We trained three different models: a Linear Support Vector Classifier (LinearSVC), a Multinomial Naïve Bayes Classifier (Multin. NB), and a *k* Nearest Neighbors Classifier (kNN) and evaluate their performance with a 10-fold cross-validation (with a stratified random test set of word forms). The results are shown in Table 4.[10] The worst performing model is the Multin. NB classifier, which barely achieves an above-baseline accuracy and a very low F1-score compared to the other two classifiers, although its recall is much higher. In terms of balanced accuracy and F1-score, the best model is the kNN classifier. However, it seems that the algorithm is not the most suited for this type of data. It performs similarly to the LinearSVC classifier, but if we compare the sizes of the resulting models, it becomes apparent that the LinearSVC model is much more efficient (with a size of approximately 100 kB) compared to the kNN model, which is overly inflated (with a size of more than 2 GB), possibly indicating overfitting.[11]

Because the LinearSVC model is the most viable, we analyze its performance in more detail. Table 5 shows the confusion matrix for the classifications of the LinearSVC model on a stratified test set (20% of the total 153,503 dataset instances). The model seems to lean more towards the most frequent category (/l/) in its predictions, with approximately 30% of /u̯/ and /l/+/u̯/ instances being misclassified as /l/, whereas 94% of the /l/ instances are classified correctly. It seems that instances allowing both pronunciations are very rarely misclassified as /u̯/ (only 1%). It should also be noted that the instances of /l/+/u̯/ misclassified as either /u̯/ or /l/ are not entirely incorrect, just incomplete. Compared to the rule-based approach (which classifies everything as /l/), the model performs quite well in terms of /l/+/u̯/ and /u̯/ instances and sacrifices only 6% of its accuracy for /l/ instances. In order to determine any future improvements to the model, we analyze some of the misclassified examples in more detail in Section 4.2.

[10] A, BA, P, R, and F1 refer to accuracy, balanced accuracy, macro-precision, macro-recall and macro-F1, respectively.

[11] We also ran a 10-fold cross-validation using only *n*-gram features (no morphosyntax). The performance of the models was slightly worse, e.g. for LinearSVC: A = 85.05, BA = 69.14, P = 68.94, R = 46.85, F1 = 55.76.

4.2 Manual Evaluation

We performed a manual analysis of the misclassified examples to determine whether there are any patterns to the errors that could be help further improve the model with additional features. Due to space limitations, we only focus on the most obvious problems in this paper.

In the examples where the /l/ pronunciation was misclassified as /u̯/, many words contain a pre-consonant *l* followed by the grapheme *d* (*kaldera* 'caldera', *buldožerski* 'pertaining to a bulldozer', *heraldičen* 'heraldic', *bodibilder* 'bodybuilder'). The majority of these examples are pronounced with /l/, with the exception of words like *dopoldne* 'late morning', *popoldanski* 'pertaining to the afternoon', where the pre-consonant *l* is preceded by an *o* grapheme. This could indicate that an additional *n*-gram feature should be added (the *l* along with its preceding and subsequent graphemes: *old*, *ald*, etc.). This could resolve some other misclassifications, such as *impulziven* 'impulsive' and *pulzirajoč* 'pulsating', where words with the *ulz* combination are never pronounced as /u̯/, but words with *olz* are (e.g., *polzeti* 'to slip'). The emergence of such patterns in the misclassifications is a good sign that the classifiers might benefit from a joint pre-*l*/post-*l* feature. This will be explored in future versions.

Many of the instances in which the /u̯/ was misclassified as /l/ contain compound words with the element *pol* 'semi, half': *polnag* 'half-naked', *polfinale* 'semi-final', *polpuščava* 'semi-desert'. Because the element *pol* is always pronounced with /u̯/, this is also true of derived compound words. However, the *n*-gram features used offer no indication of morpheme boundaries, so these misclassifications can be expected.

Additional *n*-gram features could be extracted from the accentuated forms of words. In some examples, the accentuation diacritic can disambiguate the pronunciation of the subsequent pre-consonant *l*. For instance, *dólnji* 'pertaining to something that is downwards or downstream' and *prestólničen* are pronounced with /l/, whereas *tôlšča* 'blubber' and *pôlhográjski* 'pertaining to the town of Polhov Gradec' are pronounced with /u̯/. However, accentuation is rarely written in Slovene and is much more difficult to assign automatically compared to morphosyntactic features. Relying on too many features that are not easily extractable would make the model less robust (more on this in Section 5).

5 Conclusion

We presented a machine-learning approach to improve the accuracy of phonetic transcriptions of Slovene words that contain the ambiguous pre-consonant *l*. While the method does improve accuracy (86% over a majority baseline of cca. 76%) by using very simple character-level *n*-gram and morphosyntactic features, it does not resolve the problem entirely. Aside from several exceptions in language use which are difficult to predict (e.g. *gasilci*,

čistilka; both pronounced with /l/ even though the majority of words ending with *-ilec* and *-ilka* in the dataset can be pronounced with either /l/ or /u̯/), the analysis of misclassified examples has shown several potential future steps that can be implemented to further improve the performance of the models. First, several additional features should be tested. Some of the features are simple, such as word length or number of syllables in word (which could potentially help to correctly classify words such as *volk* and *polh*; short words where the pre-consonant *l* is pronounced as /u̯/). The relative position of the pre-consonant *l* in the word could also potentially be helpful. Several more complex features could also be added, such as word formation relations and morpheme boundaries to help disambiguate, for instance, *decimal-ka* 'decimal number', which is derived from the adjective *decimalen* 'pertaining to decimal numbers' and is pronounced with /*l*/; and *mor-ilka* 'murderer (feminine)', which is derived from the verb *moriti* 'to murder' and can be pronounced as either /l/ or /u̯/). Taking into account the accentuated form of the word could also help: for instance, the *ôl* accentuation – *vôlk* 'wolf', *pôlh* 'dormouse' – indicates the /u̯/ pronunciation, while the *ól* accentuation is indicative of the /l/ pronunciation, e.g. *pólka* 'polka'). However, more complex features cannot be extracted from the word form itself, so making the model too heavily reliant on external linguistic knowledge would sacrifice its robustness and usefulness for unseen words. We will explore these options in our future work but we will first focus on the simplest features to determine the upper boundary of accuracy that can be achieved based solely on the word form and its morphosyntactic features. We will perform additional statistical analyses on *n*-grams containing the pre-consonant *l* as well, and once the optimal model is achieved, it will also be evaluated on previously unseen words containing the pre-consonant *l* that have not been included in the *ILS 1.0* dataset. The results will hopefully also provide more interesting material for further linguistic analyses (such as exceptions to the rules).

As already mentioned, the *ILS 1.0* dataset does not necessarily accurately reflect the linguistic landscape of pre-consonant *l* pronunciation in Slovene words, and more annotations along with perceptive tests and surveys are required. The pronunciations will be manually validated as part of the work on the *Digital Dictionary Database of Slovene* [8], the largest machine-readable open-access database of Slovene linguistic and lexicographic data. The pronunciations will also be cross-referenced with the recordings from the *GOS Corpus of Spoken Slovene* [18], which contains real recordings of Slovene speech and can contribute towards a more accurate distribution of different pronunciations for individual lexemes (e.g., how many occurrences of /glɛˈdaːu̯ka/ or /glɛˈdaːlka/), along with any potential relevant metadata (for instance, whether the pronunciation depends on the region the speaker originates from). The models can then be re-trained on new data and further improved to better reflect real language use.

The models will be implemented into the *Slovene IPA/X-SAMPA Grapheme-to-Phoneme Converter* as part of the *Pregibalnik* tool for automatic Slovene lexicon expansion, which is available under a Creative Commons BY-SA 4.0 license.[12]

[12]The best-performing LinearSVC model (and the accompanying code) for the prediction of pre-consonant *l* pronunciation is available on Github: https://github.com/jakacibej/sikdd2025_predicting_preconsonant_l

Acknowledgements

The research presented in this paper was carried out 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* (I0-E004), all funded by the Slovenian Research and Innovation Agency (ARIS). The author also thanks the anonymous reviewers for their constructive comments.

References

[1] Jaka Čibej. 2024. Dataset of annotated slovene words with pre-consonant l ILS 1.0. Slovenian language resource repository CLARIN.SI. (2024). http://hdl.handle.net/11356/2025.

[2] Jaka Čibej. 2023. Leksikon besednih oblik sloleks. poročilo projekta razvoj slovenščine v digitalnem okolju aktivnost ds1.3. Development of Slovene in a Digital Environment. (2023). https://www.cjvt.si/rsdo/wp-content/uploads/sites/18/2023/06/RSDO_Kazalnik_Sloleks_v2.pdf.

[3] Jaka Čibej. 2024. Predicting pronunciation types in the sloleks morphological lexicon of slovene. In *Data mining and data warehouses (SiKDD): Information Society (IS) 2024 - proceedings of the 27th International Multiconference: volume C*. Institut „Jožef Stefan", 23–26. https://is.ijs.si/wp-content/uploads/2024/11/IS2024_Volume-C.pdf.

[4] Jaka Čibej. 2025. Statistična analiza izgovora črke l v slovenskem oblikoslovnem leksikonu sloleks. *Jezikoslovni zapiski*, 31, 1, (maj 2025), 37–54. doi:10.3986/JZ.31.1.03.

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

[6] Harald Cramér. 1946. *Mathematical Methods of Statistics. Princeton Mathematical Series*. Vol. 9. Princeton University Press.

[7] Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6, 2, 65–70.

[8] Iztok Kosem, Simon Krek, and Polona Gantar. 2021. Semantic data should no longer exist in isolation: the digital dictionary database of slovenian. In *9th EURALEX International Congress "Lexicography for Inclusion"*, 81–83. https://elex.is/wp-content/uploads/2021/09/Semantic-Data-should-no-longer-exist-in-isolation-the-Digital-Dictionary-Database-of-Slovenian_Kosem-Krek-Gantar_EURALEX2020.pdf.

[9] Janez Križaj, Simon Dobrišek, Aleš Mihelič, and Jerneja Žganec Gros. 2022. Uporaba postopkov strojnega učenja pri samodejni slovenski grafemsko-fonemski pretvorbi. In *Jezikovne tehnologije in digitalna humanistika: zbornik konference 2022*. Inštitut za novejšo zgodovino, 248–251. https://nl.ijs.si/jtdh22/pdf/JTDH2022_Proceedings.pdf.

[10] Xavier Marjou. 2021. Gipfa: generating ipa pronunciation from audio. In *eLex 2021 Conference Proceedings*, 588–597. https://elex.link/elex2021/wp-content/uploads/2021/08/eLex_2021_38_pp588-597.pdf.

[11] Tanja Mirtič. 2019. Glasoslovne raziskave pri pripravi splošnega razlagalnega slovarja. In *Slovenski javni govor in jezikovno-kulturna (samo)zavest*. Znanstvena založba Filozofske fakultete, 81–90. https://centerslo.si/wp-content/uploads/2019/10/Obdobja-38_Mirtic.pdf.

[12] Karl Pearson. 1900. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50, 302, 157–175. eprint: https://doi.org/10.1080/14786440009463897. doi:10.1080/14786440009463897.

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

[14] Uwe Reichel, Hartmut R. Pfitzinger, and Horst-Udo Hain. 2008. English grapheme-to-phoneme conversion and evaluation. In *Speech and Language Technology 11*, 159–166. https://www.phonetik.uni-muenchen.de/~reichelu/publications/ReichelPfitzingerHainSASR2008.pdf.

[15] 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.

[16] Hotimir Tivadar. 2004. Priprava, izvedba in pomen perceptivnih testov za fonetično-fonološke raziskave (na primeru analize fonoloških parov). *Jezik in slovstvo*, 49.2, 2, 17–36. https://ojs.zrc-sazu.si/jz/article/view/14222.

[17] 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*.

[18] Darinka Verdonik et al. 2023. Spoken corpus gos 2.1 (transcriptions). Slovenian language resource repository CLARIN.SI. (2023). http://hdl.handle.net/11356/1863.

[19] Jerneja Žganec Gros, Tanja Mirtič, Miroslav Romih, and Kozma Ahačič. 2022. *Slovar izgovarjav OptiLEX*. (1. e-izd. ed.). Založba ZRC. ISBN: 978-961-05-0672-0. https://doi.org/10.3986/9789610506720.