Slovenščina 2.0

Jezikovne tehnologije in digitalna humanistika

Language technologies and digital humanities

Let. 9 (2021), št. 1

Univerza v Ljubljani
FILOZOFSKA FAKULTETA

Slovenščina 2.0
Letnik/Volume 9, Številka/Issue 1, 2021

ISSN: 2335-2736

Založila/Published by
Znanstvena založba Filozofske fakultete Univerze v Ljubljani

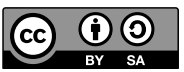
Izdal/Issued by
Center za jezikovne vire in tehnologije Univerze v Ljubljani

Za založbo/For the publisher
Roman Kuhar, dekan Filozofske fakultete

Publikacija je brezplačna./Publication is free of charge.

Publikacija je dostopna na/Avaliable at: dostopna na: https://revije.ff.uni-lj.si/slovenscina2/index

Revija izhaja s podporo Javne agencije za raziskovalno dejavnost Republike Slovenije./
This journal is published with the support of the Slovenian Research Agency (ARRS).



Kataložni zapis o publikaciji (CIP) pripravili v Narodni in univerzitetni knjižnici v Ljubljani
COBISS.SI-ID=68235779
ISBN 978-961-06-0500-3 (PDF)

KAZALO

SLOVENŠČINA 2.0: LANGUAGE TECHNOLOGIES AND DIGITAL HUMANITIES

Darja FIŠER
Faculty of Arts, University of Ljubljana; Institute of Contemporary History;
Jožef Stefan Institute

Tomaž ERJAVEC
Jožef Stefan Institute

Ajda PRETNAR
Institute of Contemporary History

Fišer, D., Erjavec, T., Pretnar, A. (2021): Slovenščina 2.0: Language technologies and digital humanities. Slovenščina 2.0, 9(1): i–vi.

DOI: https://doi.org/10.4312/slo2.0.2021.1.i-vi

The current special issue of the journal *Slovenščina 2.0* revisits a topic that has been one of the major focal points of its editorial tradition from the start. In fact, an entire issue was devoted to Language Technologies already in the first year of the journal's existence. With this collection of papers, which arrives nearly a decade later, we take stock of the current state of affairs in the field of development of resources, tools and methods for analyzing written, spoken and multimodal communication as well as their application in Digital Humanities, which has recently become a growing area of research in Slovenia.

The special issue presents eight extended papers from Slovenian as well as international authors that were originally presented at the 2020 *Language technologies and digital humanities conference* as well as a short student paper. They comprise work in language and speech technologies, language resources, digital linguistics, and digital humanities for Slovenian as well as several other languages. The special issue was reviewed by: Špela Arhar, Marko Bajec, Václav Cvrček, Simon Dobrišek, Helena Dobrovoljc, Polona Gantar, Vojko Gorjanc, Jurij Hadalin, Mateja Jemec Tomazin, Iztok Kosem, Cvetana Krstev, Nikola Ljubešić, Nataša Logar, Maja Miličević Petrović, Igor Mozetič,

Tanja Samardžić, Miha Seručnik, Mojca Stritar Kučuk, Janez Štebe, Simon Šuster, Darinka Verdonik, Špela Vintar, Jerneja Žganec Gros and Slavko Žitnik. The editors of the special issue would like to thank the authors and the reviewers for their dedicated work.

On the topic of language and speech technologies, **Marko Robnik-Šikonja**, **Kristjan Reba** and **Igor Mozetič** use cross-lingual word embeddings to transfer classification models for a Twitter sentiment classifier between 13 languages. **Matej Ulčar, Anka Supej**, **Marko Robnik-Šikonja** and **Senja Pollak** evaluate Slovenian and Croatian word embeddings in terms of gender bias using word analogy calculations. **Lucija Gril**, **Mirjam Sepesy Maučec**, **Gregor Donaj** and **Andrej Žgank** present the development of an automatic recognizer of Slovenian speech for the domain of daily news broadcasts using the the UBM BNSI Broadcast News and IETK-TV databases to train the speech recognizer using deep neural networks.

With a focus on language resources and digital linguistics, **Lucia Vlášková** and **Hana Strachoňová** present the challenges and solutions for creating an online dictionary of the Czech sign language. **Dolores Lemmenmeier-Batinić** gives an account of building a corpus of spoken Serbian and discusses current challenges in the processing of spoken data, and the implications of data re-use regarding transcriptions of speech. **Jakob Lenardič** and **Darja Fišer** perform a comparative corpus analysis of modal adverbs in Slovenian academic texts from different disciplines and study levels. **Darinka Verdonik**, **Simona Majhenič**, **Špela Antloga**, **Sandi Majninger**, **Marko Ferme**, **Kaja Dobrovoljc**, **Simona Pulko**, **Mira Krajnc Ivič** and **Natalija Ulčnik** present the development of an e-learning environment for improving writing and communication skills of Slovenian pupils.

From the digital humanities perspective, **Katja Meden** and **Ana Cvek** give an account of a major rehaul of the Historiography Citation Index that will improve the indexing of citations of scientific publications for historiographers. **Rok Mrvič** and **Špela Zupančič** survey and demonstrate the functionality of Slovenian online dialectological resources and tools.

Compared to the first special issue on Language Technologies published in this journal in 2013 where the focus of research was on the development of

basic resources and tools for Slovenian and related languages, we can observe a shift to the implementation of state-of-the-art machine learning methods, multilingual approaches, critical evaluation of technologies, and development of services for the end user. This, along with a much longer list of co-authors who come from many more institutions and countries, and work on many more languages than in the original special issue, suggests that the field has advanced significantly in the past decade and will continue to thrive, so we are already looking forward to the next special issue with a similar focus in the future.

SLOVENŠČINA 2.0: JEZIKOVNE TEHNOLOGIJE IN DIGITALNA HUMANISTIKA

Pričujoča posebna številka revije Slovenščina 2.0 se vrača k temi, ki je bila ena od osrednjih uredniških izhodišč vse od nastanka revije, saj je bila jezikovnim tehnologijam posvečena že njena prva tematska številka. Skoraj desetletje kasneje z naborom prispevkov predstavimo trenutno stanje razvoja virov, orodij in metod za analizo pisne, govorne in multimodalne komunikacije, hkrati pa se posvetimo tudi njihovi praktični uporabi v digitalni humanistiki, ki postaja vse bolj razširjeno raziskovalno področje tudi v Sloveniji.

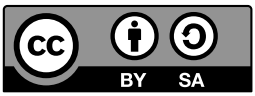
Posebna številka predstavlja osem razširjenih prispevkov slovenskih in tujih avtorjev, ki so bili izvorno predstavljeni na konferenci *Jezikovne tehnologije in digitalna humanistika* leta 2020. Prispevki vključujejo raziskave in nadgradnje jezikovnih in govornih tehnologij, jezikovnih virov, digitalnega jezikoslovja ter digitalnohumanistične raziskave tako za slovenščino kot za nekatere druge jezike. Posebno številko so recenzirali Špela Arhar, Marko Bajec, Václav Cvrček, Simon Dobrišek, Helena Dobrovoljc, Polona Gantar, Vojko Gorjanc, Jurij Hadalin, Mateja Jemec Tomazin, Iztok Kosem, Cvetana Krstev, Nikola Ljubešić, Nataša Logar, Maja Miličević Petrović, Igor Mozetič, Tanja Samardžić, Miha Seručnik, Mojca Stritar Kučuk, Janez Štebe, Simon Šuster, Darinka Verdonik, Špela Vintar, Jerneja Žganec Gros in Slavko Žitnik. Uredniki posebne številke se iskreno zahvaljujemo avtorjem in recenzentom za njihovo predano delo.

Na področju jezikovnih in govornih tehnologij **Marko Robnik-Šikonja**, **Kristjan Reba** in **Igor Mozetič** predstavijo uporabo medjezikovnih vložitev besed za prenos napovednih modelov strojnega učenja za klasifikacijo sentimenta na Twitterju med trinajstimi jeziki. **Matej Ulčar, Anka Supej**, **Marko Robnik-Šikonja** in **Senja Pollak** poročajo o evalvaciji spolne pristranskosti slovenskih in hrvaških besednih vložitev s pomočjo besednih analogij. **Lucija Gril**, **Mirjam Sepesy Maučec**, **Gregor Donaj** in **Andrej Žgank** pa predstavijo razvoj avtomatskega razpoznavalnika slovenskega govora za dnevna poročila, pri čemer razpoznavalnik govora z globokimi nevronskimi mrežami naučijo na podatkih UBM BNSI Broadcast News in IETK-TV.

Na področju jezikovnih virov in digitalnega jezikoslovja **Lucia Vlášková** in **Hana Strachoňová** obravnavata izzive in rešitve pri snovanju spletnega slovarja češkega znakovnega jezika. **Dolores Lemmenmeier-Batinić** opiše postopek oblikovanja korpusa govorjene srbščine in obravnava posledice ponovne rabe transkripcij govora. **Jakob Lenardič** in **Darja Fišer** izvedeta primerjalno analizo rabe modalnih prislovov v slovenskih akademskih besedilih med različnimi področji in ravnmi izobrazbe. **Darinka Verdonik**, **Simona Majhenič**, **Špela Antloga**, **Sandi Majninger**, **Marko Ferme**, **Kaja Dobrovoljc**, **Simona Pulko**, **Mira Krajnc Ivič** in **Natalija Ulčnik** predstavijo razvoj učnega spletnega okolja za razvoj pisnih in govornih veščin slovenskih učencev.

Z vidika digitalne humanistike **Katja Meden** in **Ana Cvek** opišeta pomembno prenovitev Zgodovinarskega indeksa citiranosti, ki bo zgodovinarjem v pomoč pri indeksiranju znanstvenih objav. **Rok Mrvič** in **Špela Zupančič** pregledata in prikažeta uporabnost spletnih orodij in virov za slovenska narečja.

V primerjavi s prvo tematsko številko na temo jezikovnih tehnologij iz leta 2013, kjer je bil poudarek na razvoju osnovnih virov in orodij za slovenščino in sorodne jezike, je v tokratni izdaji opazen premik k uvajanju naprednih tehnik in metod strojnega učenja, večjezikovnim pristopom, kritičnemu ocenjevanju obstoječih tehnologij ter razvoju storitev za končnega uporabnika. Ta premik, hkrati z daljšim seznamom avtorjev z različnih institucij in držav, ki se ukvarjajo z veliko širšim naborom jezikov kot v prvi številki, nakazuje na izjemen razmah področja v zadnjem desetletju. Digitalna humanistika in jezikoslovne tehnologije se bodo očitno uspešno razvijale še naprej, zato se že veselimo prihodnje številke na podobno temo.



CROSS-LINGUAL TRANSFER OF SENTIMENT CLASSIFIERS

Marko ROBNIK-ŠIKONJA
Faculty of Computer and Information Science, University of Ljubljana

Kristjan REBA
Faculty of Computer and Information Science, University of Ljubljana

Igor MOZETIČ
Jožef Stefan Institute

Robnik-Šikonja, M., Reba, K., Mozetič, I. (2021): Cross-lingual transfer of sentiment classifiers. Slovenščina 2.0, 9(1): 1–25.

DOI: https://doi.org/10.4312/slo2.0.2021.1.1-25

Word embeddings represent words in a numeric space so that semantic relations between words are represented as distances and directions in the vector space. Cross-lingual word embeddings transform vector spaces of different languages so that similar words are aligned. This is done by mapping one language's vector space to the vector space of another language or by construction of a joint vector space for multiple languages. Cross-lingual embeddings can be used to transfer machine learning models between languages, thereby compensating for insufficient data in less-resourced languages. We use cross-lingual word embeddings to transfer machine learning prediction models for Twitter sentiment between 13 languages. We focus on two transfer mechanisms that recently show superior transfer performance. The first mechanism uses the trained models whose input is the joint numerical space for many languages as implemented in the LASER library. The second mechanism uses large pretrained multilingual BERT language models. Our experiments show that the transfer of models between similar languages is sensible, even with no target language data. The performance of cross-lingual models obtained with the multilingual BERT and LASER library is comparable, and the differences are language-dependent. The transfer with CroSloEngual BERT, pretrained on only three languages, is superior on these and some closely related languages.

Keywords: natural language processing, machine learning, text embeddings, sentiment analysis, BERT models

1 INTRODUCTION

Word embeddings are representations of words in numerical form, as vectors of typically several hundred dimensions. The vectors are used as input to machine learning models; for complex language processing tasks, these generally are deep neural networks. The embedding vectors are obtained from specialised neural network-based embedding algorithms, e.g., fastText (Bojanowski et al., 2017) for morphologically-rich languages. Word embedding spaces exhibit similar structures across languages, even when considering distant language pairs like English and Vietnamese (Mikolov et al., 2013). This means that embeddings independently produced from monolingual text resources can be aligned, resulting in a common cross-lingual representation, called cross-lingual embeddings, which allows for fast and effective integration of information in different languages.

There exist several approaches to cross-lingual embeddings. The first group of approaches uses monolingual embeddings with an optional help from a bilingual dictionary to align the pairs of embeddings (Artetxe et al., 2018a). The second group of approaches uses bilingually aligned (comparable or even parallel) corpora to construct joint embeddings (Artetxe and Schwenk, 2019). This approach is implemented in the LASER library[1] and is available for 93 languages. The third type of approaches is based on large pretrained multilingual masked language models such as BERT (Devlin et al., 2019). In this work, we focus on the second and third group of approaches. In particular, from the third group, we apply two variants of BERT models, the original multilingual BERT model (mBERT), trained on 104 languages, and trilingual CroSloEngual BERT (Ulčar and Robnik-Šikonja, 2020) trained on Croatian, Slovene, and English (CSE BERT).

Sentiment annotation is a costly and lengthy operation, with a relatively low inter-annotator agreement (Mozetič et al., 2016). Large annotated sentiment datasets are, therefore, rare, especially for low-resourced languages. The transfer of already trained models or datasets from other languages would increase the ability to study sentiment-related phenomena for many more languages than possible today.

1 https://github.com/facebookresearch/LASER

Our study aims to analyse the abilities of modern cross-lingual approaches for the transfer of trained models between languages. We study two cross-lingual transfer technologies, using a joint vector space computed from parallel corpora with the LASER library and multilingual BERT models. The advantage of our study is sizeable comparable classification datasets in 13 different languages, which gives credibility and general validity to our findings. Further, due to the datasets' size, we can reliably test different transfer modes: direct transfer between languages (called a zero-shot transfer) and transfer with enough fine-tuning data in the target language. In the experiments, we study two cross-lingual transfer modes based on projections of sentences into a joint vector space. The first mode transfers trained models from source to target languages. A model is trained on the source language(s) and used for classification in the target language(s). This model transfer is possible because texts in all processed languages are embedded into the common vector space. The second mode expands the training set with instances from other languages, and then all instances are mapped into the common vector space during neural network training. Besides the cross-lingual transfer, we analyse the quality of representations for the Twitter sentiment classification and compare the common vector space for several languages constructed by the LASER library, multilingual BERT models, and the traditional bag-of-words approach. The results show a relatively low decrease in predictive performance when transferring trained sentiment prediction models between similar languages and superior performance of multilingual BERT models covering only three languages.

The paper is divided into four more sections. In Section 2, we present background on different types of cross-lingual embeddings: alignment of monolingual embeddings, building a common explicit vector space for several languages, and large pretrained multilingual contextual models. We also discuss related work on Twitter sentiment analysis and cross-lingual transfer of classification models. In Section 3, we present a large collection of tweets from 13 languages used in our empirical evaluation, the implementation details of our deep neural network prediction models, and the evaluation metrics used. Section 4 contains four series of experiments. We first evaluate different representation spaces and compare the LASER common vector space with

multilingual BERT models and convential bag-of-ngrams. We then analyse the transfer of trained models between languages from the same language group and from a different language group, followed by expanding datasets with instances from other languages. In Section 5, we summarise the results and present ideas for further work.

2 BACKGROUND AND RELATED WORK

Word embeddings represent each word in a language as a vector in a high dimensional vector space so that the relations between words in a language are reflected in their corresponding embeddings. Cross-lingual embeddings attempt to map words represented as vectors from one vector space to another so that the vectors representing words with the same meaning in both languages are as close as possible. Søgaard et al. (2019) present a detailed overview and classification of cross-lingual methods.

Cross-lingual approaches can be sorted into three groups, described in the following three subsections. The first group of methods uses monolingual embeddings with (an optional) help from bilingual dictionaries to align the embeddings. The second group of approaches uses bilingually aligned (comparable or even parallel) corpora for joint construction of embeddings in all handled languages. The third type of approaches is based on large pretrained multilingual masked language models such as BERT (Devlin et al., 2019). In contrast to the first two types of approaches, the multilingual BERT models are typically used as starting models, which are fine-tuned for a particular task without explicitly extracting embedding vectors.

In Section 2.1, we first present background information on the alignment of individual monolingual embeddings. We describe the projections of many languages into a joint vector space in Section 2.2, and in Section 2.3, we present variants of multilingual BERT models. In Section 2.4, we describe related work on Twitter sentiment classification. Finally, in Section 2.5, we outline the related work on cross-lingual transfer of classification models.

2.1 Alignment of monolingual embeddings

Cross-lingual alignment methods take precomputed word embeddings for each language and align them with the optional use of bilingual dictionaries.

Two types of monolingual embedding alignment methods exist. The first type of approaches map vectors representing words in one of the languages into the vector space of the other language (and vice-versa). The second type of approaches maps embeddings from both languages into a joint vector space. The goal of both types of alignments is the same: the embeddings for words with the same meaning must be as close as possible in the final vector space. A comprehensive summary of existing approaches can be found in (Artetxe et al., 2018a). The open-source *vecmap*[2] library contains implementations of methods described in (Artetxe et al., 2018a), and can align monolingual embeddings using a supervised, semi-supervised, or unsupervised approach.

The supervised approach requires the use of a bilingual dictionary, which is used to match embeddings of equivalent words. The embeddings are aligned using the Moore-Penrose pseudo-inverse, which minimises the sum of squared Euclidean distances. The algorithm always converges but can be caught in a local maximum. Several methods (e.g., stochastic dictionary introduction or frequency-based vocabulary cut-off) are used to help the algorithm climb out of local maxima. A more detailed description of the algorithm is given in (Artetxe et al., 2018b).

The semi-supervised approach uses a small initial seeding dictionary, while the unsupervised approach is run without any bilingual information. The latter uses similarity matrices of both embeddings to build an initial dictionary. This initial dictionary is usually of low but sufficient quality for later processing. After the initial dictionary (either by seeding dictionary or using similarity matrices) is built, an iterative algorithm is applied. The algorithm first computes optimal mapping using the pseudo-inverse approach for the given initial dictionary. The optimal dictionary for the given embeddings is then computed, and the procedure iterates with the new dictionary.

When constructing mappings between embedding spaces, a bilingual dictionary can help as its entries are used as anchors for the alignment map for supervised and semi-supervised approaches. However, lately, researchers have proposed methods that do not require a bilingual dictionary but rely on the

2 https://github.com/artetxem/vecmap

adversarial approach (Conneau et al., 2018) or use the words' frequencies (Artetxe et al., 2018b) to find a required transformation. These are called unsupervised approaches.

2.2 Projecting into a joint vector space

To construct a common vector space for all the processed languages, one requires a large aligned bilingual or multilingual parallel corpus. The constructed embeddings must map the same words in different languages as close as possible in the common vector space. The availability and quality of alignments in the training set corpus may present an obstacle. While Wikipedia, subtitles, and translation memories are good sources of aligned texts for large languages, less-resourced languages are not well-presented and building embeddings for such languages is a challenge.

LASER (Language-Agnostic SEntence Representations) is a Facebook research project focusing on joint sentence representation for many languages (Artetxe and Schwenk, 2019). Strictly speaking, LASER is not a word but sentence embedding method. Similarly to machine translation architectures, LASER uses an encoder-decoder architecture. The encoder is trained on a large parallel corpus, translating a sentence in any language or script to a parallel sentence in either English or Spanish (whichever exists in the parallel corpus), thereby forming a joint representation of entire sentences in many languages in a shared vector space. The project focused on scaling to many languages; currently, the encoder supports 93 different languages. Using LASER, one can train a classifier on data from just one language and use it on any language supported by LASER. A vector representation in the joint embedding space can be transformed back into a sentence using a decoder for the specific language.

2.3 Multilingual BERT and CroSloEngual BERT

BERT (Bidirectional Encoder Representations from Transformers) embedding (Devlin et al., 2019) generalises the idea of a language model (LM) to masked LMs, inspired by the cloze test, which checks understanding of a text by removing a few words, which the participant is asked to replace. The masked LM randomly masks some of the tokens from the input, and

the task is to predict the missing token based on its neighbourhood. BERT uses transformer neural networks (Vaswani et al., 2017) in a bidirectional sense and further introduces the task of predicting whether two sentences appear in a sequence. The input representation of BERT are sequences of tokens representing sub-word units. The input is constructed by summing the embeddings of corresponding tokens, segments, and positions. Some widespread words are kept as single tokens; others are split into sub-words (e.g., frequent stems, prefixes, suffixes—if needed down to single letter tokens). The original BERT project offers pre-trained English, Chinese, and multilingual model. The latter, called mBERT, is trained on 104 languages simultaneously.

To use BERT in classification tasks only requires adding connections between its last hidden layer and new neurons corresponding to the number of classes in the intended task. The fine-tuning process is applied to the whole network, and all the parameters of BERT and new class-specific weights are fine-tuned jointly to maximise the log-probability of correct labels.

Recently, a new type of multilingual BERT models emerged that reduce the number of languages in multilingual models. For example, CSE BERT (Ulčar and Robnik-Šikonja, 2020) uses Croatian, Slovene (two similar less-resourced languages from the same language family), and English. The main reasons for this choice are to represent each language better and keep sensible sub-word vocabulary, as shown by Virtanen et al. (2019). This model is built with the cross-lingual transfer of prediction models in mind. As CSE BERT includes English, we expect that it will enable a better transfer of existing prediction models from English to Croatian and Slovene.

2.4 Twitter sentiment classification

We present a brief overview of the related work on automated sentiment classification of Twitter posts. We summarise the published labelled sets used for training the classification models and the machine learning methods applied for training. Most of the related work is limited to only English texts.

To train a sentiment classifier, one needs a reasonably large training dataset of tweets already labelled with the sentiment. One can rely on a proxy, e.g.,

emoticons used in the tweets, to determine the intended sentiment; however, high-quality labelling requires the engagement of human annotators. There exist several publicly available and manually labelled Twitter datasets. They vary in the number of examples from several hundred to several thousand, but to the best of our knowledge, so far, none exceeds 20,000 entries. Saif et al. (2013) describe eight Twitter sentiment datasets and introduce a new one that contains separate sentiment labels for tweets and entities. Rosenthal et al. (2015) provide statistics for several of the 2013–2015 SemEval datasets.

There are several supervised machine learning algorithms suitable to train sentiment classifiers from sentiment labelled tweets. For example, in the SemEval-2015 competition, before the rise of deep neural networks, the most often used algorithms for the sentiment analysis on Twitter (Rosenthal et al., 2015) were support vector machines (SVM), maximum entropy, conditional random fields, and linear regression. In other cases, frequently used classifiers were naive Bayes, k-nearest neighbours, and even decision trees. Often, SVM was shown as the best performing classifier for the Twitter sentiment. However, only recently, when researchers started to apply deep learning for the Twitter sentiment classification, considerable improvements in classification performance were observed (Wehrmann et al., 2017; Jianqiang et al., 2018; Naseem et al., 2020). Similarly to our approach, recent approaches use contextual embeddings such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019), but in a monolingual setting.

2.5 Transfer of trained models

Cross-lingual word embeddings can be used directly as inputs in natural language processing models. The main idea is to train a model on data from one language and then apply it to another, relying on shared cross-lingual representation. Several tasks have been attempted in testing cross-lingual transfe. Søgaard et al. (2019) survey the transfer in the following tasks: document classification, dependency parsing, POS tagging, named entity recognition, super-sense tagging, semantic parsing, discourse parsing, dialogue state tracking, entity linking (wikification), sentiment analysis, machine translation, natural language interference, etc. For example, Ranasinghe

and Zampieri (2020) apply large pretrained models in a similar way as we but use offensive language domain and only four languages from different families (English, Spanish, Bengali, and Hindu). In sentiment analysis, which is of particular interest in this work, Mogadala and Rettinger (2016) evaluate their embeddings on the multilingual Amazon product review dataset. In the Twitter sentiment analysis, Wehrmann et al. (2017) use LSTM networks but first learn a joint representation for four languages (English, German, Portuguese, and Spanish) with character-based convolutional neural networks.

3 DATASETS AND EXPERIMENTAL SETTINGS

This section presents the evaluation metrics, experimental data, and implementation details of the used neural prediction models.

3.1 Evaluation metrics

Following Mozetič et al. (2016), we report the $\overline{F}_1$ score and classification accuracy (*CA)*. The $F_1(c)$ score for class value c is the harmonic mean of precision p and recall r for the given class c, where the precision is defined as the proportion of correctly classified instances from the instances predicted to be from the class c, and the recall is the proportion of correctly classified instances actually from the class c:

$$F_1(c) = \frac{2p_c r_c}{p_c + r_c}.$$

The F_1 score returns values from the [0,1] interval, where 1 means perfect classification, and 0 indicates that either precision or recall for class c is 0. We use an instance of the F_1 score specifically designed to evaluate the 3-class sentiment models (Kiritchenko et al., 2014). $\overline{F}_1$ is defined as the average over the positive (+) and negative (−) sentiment class:

$$\overline{F}_1 = \frac{F_1(+) + F_1(-)}{2}.$$

$\overline{F}_1$ implicitly considers the ordering of sentiment values by considering only the extreme labels, positive (+) and negative (-). The middle, neutral, is taken

into account indirectly. $\overline{F}_1 = 1$ implies that all negative and positive tweets were correctly classified, and as a consequence, all neutrals as well. $\overline{F}_1 = 0$ indicates that all tweets were classified as neutral, and consequently, all negative and positive tweets were incorrectly classified.

$\overline{F}_1$ is not the best performance measure. First, taking the arithmetic average of the F_1 scores over different classes (called macro F_1) is methodologically misguided (Flach and Kull, 2015). It is justified only when the class distribution is approximately even, as in our case. Second, $\overline{F}_1$ does not account for correct classifications by chance. A more appropriate measure that allows for class ordering, classification by chance, and class labelling with disagreements is Krippendorff's alpha-reliability (Krippendorff, 2013). However, since $\overline{F}_1$ is commonly used in the sentiment classification community, and the results are typically well-correlated with the alpha-reliability, we decided to report our experimental results in terms of $\overline{F}_1$.

The second score we report is the classification accuracy CA, defined as the ratio of correctly predicted tweets N_c to all the tweets N:

$$CA = \frac{N_c}{N}.$$

3.2 Datasets

We use a corpus of Twitter sentiment datasets (Mozetič et al., 2016), consisting of 15 languages, with over 1.6 million annotated tweets. The languages covered are Albanian, Bosnian, Bulgarian, Croatian, English, German, Hungarian, Polish, Portuguese, Russian, Serbian, Slovak, Slovene, Spanish, and Swedish. The authors studied the annotators' agreement on the labelled tweets. They discovered that the SVM classifier achieves significantly lower score for some languages (English, Russian, Slovak) than the annotators. This hints that there might be room for improvement for these languages using a better classification model or a larger training set.

We cleaned the above datasets by removing the duplicated tweets, weblinks, and hashtags. Due to the low quality of sentiment annotations indicated by low self-agreement and low inter-annotator agreement, we removed Albanian and Spanish datasets. For these two languages, the self-agreement expressed with $\overline{F}_1$ score is 0.60 and 0.49, respectively; the inter-annotator agreement is

0.41 and 0.42. As defined above, $\overline{F}_1$ is the arithmetic average of F_1 scores for the positive and negative tweets, where $F_1(c)$ is the fraction of equally labelled tweets out of all the tweets with the label c.

In the paper where the datasets were introduced (Mozetič et al., 2016), Serbian, Croatian, and Bosnian tweets were merged into a single dataset. The three languages are very similar and difficult to distinguish in short Twitter posts. However, it turned out that this merge resulted in a poor classification performance due to a very different quality of annotations. In particular, Serbian (71,721 tweets) was annotated by 11 annotators, where two of them accounted for over 40% of the annotations. All the inter-annotator agreement measures come from the Serbian only (1,880 tweets annotated twice by different annotators, $\overline{F}_1$ is 0.51), and there are very few tweets annotated twice by the same annotator (182 tweets only, $\overline{F}_1$ for the self-agreement is 0.46). In contrast, all the Croatian and Bosnian tweets were annotated by a single annotator, and we have reliable self-agreement estimates. There are 84,001 Croatian tweets, 13,290 annotated twice, and the self-agreement $\overline{F}_1$ is 0.83. There are 38,105 Bosnian tweets, 6,519 annotated twice, and the self-agreement $\overline{F}_1$ is 0.78. The authors concluded that the annotation quality of the Croatian and Bosnian tweets is considerably higher than that of the Serbian. If one constructs separate sentiment classifiers for each language, one observes a very different performance than reported originally. The individual classifiers are better and "well-behaved" compared to the joint Serbian/Croatian/Bosnian model. In this paper, we follow the authors' suggestion that datasets with no overlapping annotations and different annotation quality are better not merged. As a consequence, the Serbian, Croatian, and Bosnian datasets are analysed separately. The characteristics of all the 13 datasets are presented in Table 1.

Table 1: *The characteristics of datasets*

	Number of tweets				Agreement ($\bar{F}_1$)	
Language	**Negative**	**Neutral**	**Positive**	**All**	**Self-**	**Inter-**
Bosnian	12,868	11,526	13,711	38,105	0.78	-
Bulgarian	15,140	31,214	20,815	67,169	0.77	0.50
Croatian	21,068	19,039	43,894	84,001	0.83	-
English	26,674	46,972	29,388	103,034	0.79	0.67
German	20,617	60,061	28,452	109,130	0.73	0.42
Hungarian	10,770	22,359	35,376	68,505	0.76	-
Polish	67,083	60,486	96,005	223,574	0.84	0.67
Portuguese	58,592	53,820	44,981	157,393	0.74	-
Russian	34,252	44,044	29,477	107,773	0.82	-
Serbian	24,860	30,700	16,161	71,721	0.46	0.51
Slovak	18,716	14,917	36,792	70,425	0.77	-
Slovene	38,975	60,679	34,281	133,935	0.73	0.54
Swedish	25,319	17,857	15,371	58,547	0.76	-

Note. The left-hand side reports the number of tweets from each category and the overall number of instances for individual languages. The right-hand side contains self-agreement of annotators and inter-annotator agreement for tried languages where more than one annotator was involved.

3.3 Implementation details

In our experiments, we use three different types of prediction models, BiLSTM neural networks using joint vector space embeddings constructed with the LASER library, and two variants of BERT, mBERT, and CSE BERT. The original mBERT (bert-multi-cased) is pretrained on 104 languages, has 12 transformer layers, and 110 million parameters. The CSE BERT uses the same architecture but is pretrained only on Croatian, Slovene, and English. In the construction of sentiment classification models, we fine-tune the whole network, using the batch size of 32, 2 epochs, and Adam optimiser. We also tested larger numbers of epochs and larger batch sizes in preliminary experiments, but this did not improve the performance.

The cross-lingual embeddings from the LASER library are pretrained on 93 languages, using BiLSTM networks, and are stored as 1024 dimensional embedding vectors. Our classification models contain an embedding layer, followed by a multilayer perceptron hidden layer of size 8, and an output layer

with three neurons (corresponding to three output classes, negative, neutral, and positive sentiment) using the softmax. We use the ReLU activation function and Adam optimiser. The fine-tuning uses a batch size of 32 and 10 epochs.

Further technical details are available in the freely available source code.

4 EXPERIMENTS AND RESULTS

Our experimental work focuses on model transfer with cross-lingual embeddings. However, to first establish the suitability of different embedding spaces for Twitter sentiment classification, we start with their comparison in a monolingual setting in Section 4.1. We compare the three neural approaches presented in Section 3.3 (common vector space of LASER, mBERT, and CSE BERT). As a baseline, we use the classical approach using bag-of-ngram representation with the SVM classifier. In the cross-lingual experiments, we focus on the two most-successful types of model transfer, described in Sections 2.2 and 2.3: the common vector space of the LASER library and the variants of the multilingual BERT model (mBERT and CSE BERT). We conducted several cross-lingual transfer experiments: transfer of models between languages from the same (Section 4.2) and different language family (Section 4.3), as well as the expansion of training sets with varying amounts of data from other languages (Section 4.4). In the experiments, we did not systematically test all possible combinations of languages and language groups as this would require an excessive amount of computational time and reporting space, and would not contribute to the clarity of the paper. Instead, we arbitrarily selected a representative set of language combinations in advance. We leave a comprehensive systematic approach based on informative features (Lin et al., 2019) for further work.

4.1 Comparing embedding spaces

To establish the appropriateness of different embedding approaches for our Twitter sentiment classification task, we start with experiments in a monolingual setting. We compare embeddings into a joint vector space obtained with the LASER library with mBERT and CSE BERT. Note that there is no transfer between different languages in this experiment but only a test of

the suitability of the representation, i.e. embeddings. To make the results comparable with previous work on these datasets, we report results obtained with 10-fold blocked cross-validation. There is no randomisation of training examples in the blocked cross-validation, and each fold is a block of consecutive tweets. It turns out that standard cross-validation with a random selection of examples yields unrealistic estimates of classifier performance and should not be used to evaluate classifiers in time-ordered data scenarios (Mozetič et al., 2018).

As a baseline, we report the results of SVM models without neural embeddings that use Delta TF-IDF weighted bag-of-ngrams representation with substantial preprocessing of tweets (Mozetič et al., 2016). As the datasets for the Bosnian, Croatian, and Serbian languages were merged in (Mozetič et al., 2016) due to the similarity of these languages, we report the performance on the merged dataset for the SVM classifier. Results are presented in Table 2.

Table 2: *Comparison of different representations: supervised mapping into a joint vector space with the LASER library, mBERT, CSE BERT, and bag-of-ngrams with the SVM classifier*

	LASER		mBERT		CSE BERT		SVM	
Language	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**
Bosnian	**0.68**	0.64	0.65	0.60	**0.68**	**0.65**	(0.61	0.56)
Bulgarian	0.53	**0.59**	**0.58**	**0.59**	0.00	0.45	0.52	0.54
Croatian	0.72	0.68	0.64	0.66	**0.76**	**0.71**	(0.61	0.56)
English	0.62	0.65	**0.68**	**0.68**	0.67	0.66	0.63	0.64
German	0.52	0.64	**0.66**	**0.66**	0.31	0.59	0.54	0.61
Hungarian	0.63	0.67	**0.65**	**0.69**	0.57	0.65	0.64	0.67
Polish	**0.70**	0.66	**0.70**	**0.70**	0.56	0.57	0.68	0.63
Portuguese	0.48	0.47	0.50	0.49	0.12	0.22	**0.55**	**0.51**
Russian	**0.70**	**0.70**	0.64	0.64	0.07	0.43	0.61	0.60
Serbian	0.50	0.54	0.50	0.52	0.30	0.50	**(0.61**	**0.56)**
Slovak	**0.72**	**0.72**	0.67	0.66	0.69	0.71	0.68	0.68
Slovene	0.57	0.58	0.58	0.58	**0.60**	**0.61**	0.55	0.54
Swedish	**0.67**	0.64	**0.67**	**0.65**	0.54	0.56	0.66	0.62
#Best	5	3	6	6	3	3	2	2

Note. The best score for each language and metric is in bold. In the last row, we count the number of best scores for each model. The SVM results for Bosnian, Croatian, and Serbian were obtained with the model trained on the merged dataset of these languages model and are therefore not directly compatible with the language-specific results for the other representations.

The SVM baseline using bag-of-ngrams representation mostly achieves lower predictive performance than the two neural embedding approaches. We speculate that the main reason is more information about the language structure contained in precomputed dense embeddings used by the neural approaches. Together with the fact that standard feature-based machine learning approaches require much more preprocessing effort, it seems that there are no good reasons why to bother with this approach in text classification; we, therefore, omit this method from further experiments. The mBERT model is the best of the tested methods, achieving the best $\overline{F}_1$ and CA scores in six languages (in bold), closely followed by the LASER approach, which achieves the best $\overline{F}_1$ score in five languages and the best CA score in three languages. The CSE BERT is specialised for only three languages, and it achieves the best scores in languages where it is trained (except in English, where it is close behind mBERT), and in Bosnian, which is similar to Croatian. Overall, it seems that large pretrained transformer models (mBERT and CSE BERT) are dominating in the Twitter sentiment prediction. The downside of these models is that their training, fine-tuning, and execution require more computational time than precomputed fixed embeddings. Nevertheless, with progress in optimisation techniques for neural network learning and advent of computationally more efficient BERT variants, e.g., (You et al., 2020), this obstacle might disappear in the future.

4.2 Transfer to the same language family

The transfer of prediction models between similar languages from the same language family is the most likely to be successful. We test several combinations of source and target languages from Slavic and Germanic language families. We report the results in Table 3.

In each experiment, we use the entire dataset(s) of the source language as the training set and the whole dataset of the target language as the testing set, i.e. we do a zero-shot transfer. We compare the results with the LASER embeddings with BiLSTM network using training and testing set from the target language, where 70% of the dataset is used for training and 30% for testing. As we use large datasets, the latter results can be taken as an upper bound of what cross-lingual transfer models could achieve in ideal conditions.

The results from Table 3 (bottom line) show that there is a gap in the performance of transfer learning models and native models. On average, the gap in $\overline{F}_1$ is 5% for the LASER approach, 6% for mBERT, and 8% for CSE BERT. For CA, the average gap is 7% for both LASER and mBERT and 8% for CSE BERT. However, there are significant differences between languages, and we advise to test both LASER and mBERT for a specific new language, as the models are highly competitive. The CSE BERT is slightly less successful measured with the average performance gap over all languages as the gap is 8% in both $\overline{F}_1$ and CA. However, if we take only the three languages used in the training of CSE BERT (Croatian, Slovene, and English) as shown in

Table 3: *The transfer of trained models between languages from the same language family using LASER common vector space, mBERT, and CSE BERT*

		LASER		**mBERT**		**CSE BERT**		**Both target**	
Source	**Target**	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**
German	English	0.55	0.59	**0.63**	**0.64**	0.42	0.42	0.62	0.65
English	German	0.55	0.60	**0.66**	**0.70**	0.50	0.58	0.53	0.65
Polish	Russian	**0.64**	**0.59**	0.57	0.57	0.50	0.40	0.70	0.70
Polish	Slovak	**0.63**	0.59	0.58	0.59	**0.63**	**0.65**	0.72	0.72
German	Swedish	0.58	0.57	**0.59**	**0.59**	0.58	0.56	0.67	0.65
German Swedish	English	**0.58**	**0.60**	0.55	0.56	0.41	0.42	0.62	0.65
Slovene Serbian	Russian	0.53	0.55	0.57	**0.57**	**0.58**	0.48	0.70	0.70
Slovene Serbian	Slovak	**0.59**	0.52	0.57	0.59	0.48	**0.60**	0.72	0.72
Serbian	Slovene	0.54	**0.57**	0.54	0.54	**0.56**	0.55	0.60	0.60
Serbian	Croatian	**0.67**	0.64	0.65	0.62	0.65	**0.70**	0.73	0.68
Serbian	Bosnian	**0.65**	0.61	0.61	0.60	0.59	**0.62**	0.67	0.64
Polish	Slovene	0.51	0.48	**0.55**	**0.54**	0.50	0.53	0.60	0.60
Slovak	Slovene	0.52	0.51	0.54	0.54	**0.58**	**0.58**	0.60	0.60
Croatian	Slovene	0.53	0.53	0.53	0.54	**0.61**	**0.60**	0.60	0.60
Croatian	Serbian	**0.54**	**0.52**	0.52	0.51	0.52	0.49	0.48	0.54
Croatian	Bosnian	0.66	0.61	0.57	0.56	**0.67**	**0.62**	0.67	0.64
Slovene	Croatian	0.70	0.65	0.64	0.63	**0.73**	**0.69**	0.73	0.68
Slovene	Serbian	**0.52**	**0.55**	0.46	0.49	0.47	0.50	0.48	0.54
Slovene	Bosnian	**0.66**	0.61	0.58	0.56	**0.66**	**0.62**	0.67	0.64
Average performance gap		0.05	0.07	0.06	0.07	0.08	0.08		

Note. We compare the results with both training and testing set from the target language using the LASER approach (the right-most two columns).

Table 4, conclusions are entirely different. The average performance gap is 0% in $\overline{F}_1$ and 1% in the classification accuracy, meaning that we get almost a perfect cross-lingual transfer for these languages on the Twitter sentiment prediction task.

We also tried more than one input language at once, for example, German and Swedish as source languages and English as the target language, as shown in Table 3. The success of the tested combinations is mixed: for some models and some languages, we slightly improve the scores, while for others, we slightly decrease them. We hypothesise that our datasets for individual languages are large enough so that adding additional training data does not help.

Table 4: *The transfer of sentiment models between all combinations of languages on which CSE BERT was trained (Croatian, Slovene, and English)*

		LASER		mBERT		CSE BERT		Both target	
Source	**Target**	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**
Croatian	Slovene	0.53	0.53	0.53	0.54	**0.61**	**0.60**	0.60	0.60
Croatian	English	**0.63**	0.63	**0.63**	**0.66**	0.62	0.64	0.62	0.65
English	Slovene	0.54	0.57	0.50	0.53	**0.59**	**0.57**	0.60	0.60
English	Croatian	0.62	**0.67**	0.67	0.63	**0.73**	**0.67**	0.73	0.68
Slovene	English	0.63	0.64	**0.65**	**0.67**	0.63	0.64	0.62	0.65
Slovene	Croatian	0.70	0.65	0.64	0.63	**0.73**	**0.69**	0.73	0.68
Croatian English	Slovene	0.54	0.54	0.53	0.54	**0.60**	**0.58**	0.60	0.60
Croatian Slovene	English	0.62	0.61	**0.65**	**0.67**	0.63	0.65	0.62	0.65
English Slovene	Croatian	0.64	0.68	0.63	0.63	**0.68**	**0.70**	0.73	0.68
Average performance gap		0.04	0.03	0.04	0.03	0.00	0.01		

4.3 Transfer to a different language family

The transfer of prediction models between languages from different language families is less likely to be successful. Nevertheless, to observe the difference, we test several combinations of source and target languages from different language families (one from Slavic, the other from Germanic, and vice-versa). We compare the LASER approach with mBERT models; the CSE BERT is not constructed for this setting, and we skip it in this experiment. We report the results in Table 5.

The results show that with the LASER approach, there is an average decrease of performance for transfer learning models of 11% (both $\overline{F}_1$ and CA), and for mBERT, the gap is 9%. This gap is significant and makes the resulting transferred models less useful in the target languages, though there are considerable differences between the languages.

Table 5: *The transfer of trained models between languages from different language families using LASER common vector space and mBERT*

		LASER		mBERT		Both target	
Source	**Target**	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**
Russian	English	**0.52**	0.56	**0.52**	**0.57**	0.62	0.65
English	Russian	**0.57**	**0.58**	0.55	0.57	0.70	0.70
English	Slovak	0.46	0.44	**0.57**	**0.58**	0.72	0.72
Polish, Slovene	English	0.58	0.57	**0.60**	**0.60**	0.62	0.65
German, Swedish	Russian	0.61	**0.61**	**0.62**	0.59	0.70	0.70
English, German	Slovak	0.50	0.47	**0.56**	**0.54**	0.72	0.72
German	Slovene	**0.54**	**0.56**	0.53	0.54	0.60	0.60
English	Slovene	**0.54**	**0.57**	0.50	0.53	0.60	0.60
Swedish	Slovene	**0.54**	**0.56**	0.52	0.54	0.60	0.60
Hungarian	Slovene	0.52	0.52	**0.53**	**0.54**	0.60	0.60
Portuguese	Slovene	0.51	0.49	**0.54**	**0.54**	0.60	0.60
Average performance gap		0.11	0.11	0.09	0.09		

Note. We compare the results with both training and testing set from the target language using the LASER approach (the right-most two columns).

4.4 Increasing datasets with several languages

Another type of cross-lingual transfer is possible if we increase the training sets with instances from several related and unrelated languages. We conduct two sets of experiments in this scenario. In the first setting, reported in Table 6, we constructed the training set in each experiment with instances from several languages and 70% of the target language dataset. The remaining 30% of target language instances are used as the testing set. In the second setting, reported in Table 7, we merge *all* other languages and 70% of the target language into a joint training set. We compare the LASER approach, mBERT, and also CSE BERT, as Slovene and Croatian are involved in some combinations.

Table 6 shows a gap between learning models using the expanded datasets and models with only target language data. The decrease is more extensive for both BERT models (on average around 10%) than for the LASER approach (the decrease is on average 3% for $\overline{F}_1$ and 5% for CA). These results indicate that the tested expansion of datasets was unsuccessful, i.e. the provided amount of training instances in the target language was already sufficient for successful learning. The additional instances from other languages in the transformed space are likely to be of lower quality than the native instances and therefore decrease the performance.

Table 6: *The expansion of training sets with instances from several languages*

		LASER		mBERT		CSEBERT		Target only	
Source	**Target**	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**
English, Croatian, Slovene	Slovene	0.58	0.53	0.46	0.45	**0.60**	**0.58**	0.60	0.60
English, Croatian, Serbian, Slovak	Slovak	**0.67**	**0.65**	0.57	0.54	0.27	0.37	0.72	0.72
Hungarian, Slovak, English, Croatian, Russian	Russian	**0.67**	**0.65**	0.61	0.59	0.63	0.61	0.70	0.70
Russian, Swedish, English	English	0.60	0.61	**0.62**	0.60	0.59	**0.62**	0.62	0.65
Croatian, Serbian, Bosnian, Slovene	Slovene	0.54	**0.58**	0.44	0.45	**0.57**	0.56	0.60	0.60
English, Swedish, German	German	0.55	0.60	**0.60**	**0.64**	0.47	0.58	0.53	0.65
Average performance gap		0.03	0.05	0.08	0.11	0.11	0.10		

Note. We compare the LASER approach, mBERT, and CSE BERT. As the upper bound, we give results of the LASER approach trained on only the target language.

The results in Table 7, where we test the expansion of the training set (consisting of 70% of the dataset in the target language) with all other languages, show that using many languages and significant enlargement of datasets is also not successful. The two improvements in the LASER approach over using only target language are limited to a single metric (F_1 in case of Bulgarian and Serbian), which indicates that true positives are favoured at the expense of true negatives. For all the other languages, the tried expansions of training sets are unsuccessful for the LASER approach; the difference to native models

is on average 3.5% for the $\overline{F}_1$ score and 6% for CA. The mBERT models are in almost all cases more successful in this massive transfer than LASER models, and they sometimes marginally beat the reference mBERT approach trained only on the target language.

Table 7: *The expansion of training sets with instances from all other languages (+70% of the target language instances) to train the LASER approach and mBERT*

	LASER				mBERT			
	All & Target		Only Target		All &Target		Only Target	
Target	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**	$\overline{F}_1$	**CA**
Bosnian	0.64	0.59	0.67	0.64	0.63	0.60	0.65	0.60
Bulgarian	**0.54**	0.56	0.50	0.59	**0.60**	**0.60**	0.58	0.59
Croatian	0.63	0.57	0.73	0.68	**0.65**	0.63	0.64	0.66
English	0.58	0.60	0.62	0.65	0.64	**0.69**	0.68	0.68
German	0.52	0.59	0.53	0.65	0.61	0.66	0.66	0.66
Hungarian	0.59	0.61	0.60	0.67	0.65	0.69	0.65	0.69
Polish	0.67	0.63	0.70	0.66	**0.71**	**0.71**	0.70	0.70
Portuguese	0.44	0.39	0.52	0.51	**0.52**	**0.52**	0.50	0.49
Russian	0.66	0.64	0.70	0.70	**0.67**	**0.66**	0.64	0.64
Serbian	**0.52**	0.49	0.48	0.54	**0.53**	0.51	0.50	0.52
Slovak	0.64	0.61	0.72	0.72	0.67	0.65	0.67	0.66
Slovene	0.54	0.50	0.60	0.60	0.56	0.54	0.58	0.58
Swedish	0.63	0.59	0.67	0.65	0.67	0.64	0.67	0.65
Avg. gap	0.03	0.06			0.00	0.00		

Note. We compare the results with the training on only the target language. The scores where models with the expanded training sets beat their respective reference scores are in bold.

5 CONCLUSIONS

We studied state-of-the-art approaches to the cross-lingual transfer of Twitter sentiment prediction models: mappings of words into the common vector space using the LASER library and two multilingual BERT variants (mBERT and trilingual CSE BERT). Our empirical evaluation is based on relatively large datasets of labelled tweets from 13 European languages. We first tested the success of these text representations in a monolingual setting. The results show that BERT variants are the most successful, closely followed by the LASER approach, while the classical bag-of-ngrams coupled with the SVM

classifier is no longer competitive with neural approaches. In the cross-lingual experiments, the results show that there is a significant transfer potential using the models trained on similar languages; compared to training and testing on the same language, with LASER, we get on average 5% lower $\overline{F}_1$ score and with mBERT 6% lower $\overline{F}_1$ score. The transfer of models with CSE BERT is even more successful in the three languages covered by this model, where we get no performance gap compared to the LASER approach trained and tested on the target language. Using models trained on languages from different language families produces larger differences (on average around 10% for $\overline{F}_1$ and CA). Our attempt to expand training sets with instances from different languages was unsuccessful using either additional instances from a small group of languages or instances from all other languages. The source code of our analyses is freely available[3].

We plan to expand BERT models with additional emotional and subjectivity information in future work on sentiment classification. Given the favourable results in cross-lingual transfer, we will expand the work to other relevant tasks.

Acknowledgments

The research was supported by the Slovene Research Agency through research core funding no. P6-0411 and P2-103, as well as project no. J6-2581. This paper is supported by European Union's Horizon 2020 Programme project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media, grant no. 825153), and Rights, Equality and Citizenship Programme project IMSyPP (Innovative Monitoring Systems and Prevention Policies of Online Hate Speech, grant no. 875263). The results of this publication reflect only the authors' view, and the Commission is not responsible for any use that may be made of the information it contains.

3 https://github.com/kristjanreba/cross-lingual-classification-of-tweet-sentiment

REFERENCES

Artetxe, M., Labaka, G., & Agirre, E. (2018a). Generalising and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Artetxe, M., Labaka, G., & Agirre, E. (2018b). A robust self-learning method for fully unsupervised crosslingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics:Vol 1 (Long Papers)* (pp. 789–798).

Artetxe, M., & Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot crosslingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, *7*, 597–610.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Conneau, A., Lample, G., Ranzato, M.A., Denoyer, L., & J'egou, H. (2018). Word' translation without parallel data. In *6th Proceedings of International Conference on Learning Representation (ICLR)*. Retrieved from https://openreview.net/pdf?id=H196sainb

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers)* (pp. 4171–4186).

Flach, P., & Kull, M. (2015). Precision-recall-gain curves: PR analysis done right. In *Advances in Neural Information Processing Systems (NIPS)* (pp. 838–846).

Jianqiang, Z., Xiaolin, G., and Xuejun, Z. (2018). Deep convolution neural networks for Twitter sentiment analysis. *IEEE Access*, *6*, 23253–23260.

Kiritchenko, S., Zhu, X., Mohammad, S. M. (2014). Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, *50*, 723–762.

Krippendorff, K. (2013). *Content Analysis, An Introduction to Its Methodology* (3rd ed.) Thousand Oaks, CA, USA: Sage Publications.

Lin, Y. H., Chen, C. Y., Lee, J., Li, Z., Zhang, Y., Xia, M., Rijhwani, S., et al. (2019). Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 3125–3135).

Mikolov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting similarities among languages for machine translation. *arXiv preprint* 1309.4168.

Mogadala, A., & Rettinger, A. (2016). Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification. In *Proceedings of NAACL-HLT* (pp. 692–702).

Mozetič, I., Grčar, M., & Smailović, J. (2016). Multilingual Twitter sentiment classification: The role of human annotators. *PLOS ONE*, *11*(5). doi: 10.1371/journal.pone.0155036

Mozetič, I., Torgo, L., Cerqueira, V., & Smailović, J. (2018). How to evaluate sentiment classifiers for Twitter time-ordered data? *PLoS ONE 13*(3).

Naseem, U., Razzak, I., Musial, K., & Imran, M. (2020). Transformer based deep intelligent contextual embedding for Twitter sentiment analysis. *Future Generation Computer Systems*, *113*, 58–69.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualised word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long Papers)* (pp. 2227–2237).

Ranasinghe, T., & Zampieri, M. (2020). Multilingual Offensive Language Identification with Cross-lingual Embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 5838–5844).

Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S. M., Ritter, A., & Stoyanov, V. (2015). SemEval-2015 task 10: Sentiment Analysis in Twitter. In *Proceedings of 9th International Workshop on Semantic Evaluation (SemEval)* (pp. 451–463).

Saif, H., Fernández, M., He, Y., Alani, H.(2013). Evaluation datasets for Twitter sentiment analysis: A survey and a new dataset, the STS-Gold. In *1st Intl. Workshop on Emotion and Sentiment in Social and Expressive Media: Approaches and Perspectives from AI (ESSEM)*.

Søgaard, A., Vulić, I., Ruder, S., & Faruqui, M. (2019). *Cross-Lingual Word Embeddings*. Morgan & Claypool Publishers.

Ulčar, M., & Robnik-Šikonja, M. (2020). FinEst BERT and CroSloEngual BERT. In *International Conference on Text, Speech, and Dialogue (TSD)* (pp. 104–111).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)* (pp. 5998–6008).

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luoto-lahti, J., Salakoski, T., Ginter, F., & Pyysalo, S. (2019). Multilingual is not enough: BERT for Finnish. *arXiv preprint* 1912.07076.

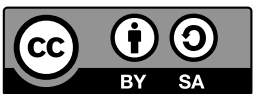
Wehrmann, J., Becker, W., Cagnini, H. E., & Barros, R. C. (2017). A character-based convolutional neural network for language-agnostic Twitter sentiment analysis. In 2017 *International Joint Conference on Neural Networks (IJCNN)* (pp. 2384–2391).

You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., Song, X., et al. (2020). Large batch optimization for deep learning: Training BERT in 76 minutes. In *8th International Conference on Learning Representations (ICLR), 26-30 April, 2020, Addis Ababa, Ethiopia*.

MEDJEZIKOVNI PRENOS KLASIFIKATORJEV SENTIMENTA

Vektorske vložitve predstavijo besede v številski obliki tako, da so semantične relacije med besedami zapisane kot razdalje in smeri v vektorskem prostoru. Medjezikovne vložitve poravnajo vektorske prostore različnih jezikov, kar podobne besede v različnih jezikih postavi blizu skupaj. Medjezikovna poravnava lahko deluje na parih jezikov ali s konstrukcijo skupnega vektorskega prostora več jezikov. Medjezikovne vektorske vložitve lahko uporabimo za prenos modelov strojnega učenja med jeziki in s tem razrešimo težavo premajhnih ali neobstoječih učnih množic v jezikih z manj viri. V delu uporabljamo medjezikovne vložitve za prenos napovednih modelov strojnega učenja za napovedovanje sentimenta tvitov med trinajstimi jeziki. Osredotočeni smo na dva, v zadnjem času najuspešnejša, načina prenosa modelov. Prvi način uporablja modele naučene na skupnem vektorskem prostoru za mnoge jezike, izdelanem s knjižnico LASER. Drugi način uporablja velike, na mnogih jezikih vnaprej naučene, jezikovne modele tipa BERT. Naši poskusi kažejo, da je prenos modelov med podobnimi jeziki smiseln tudi povsem brez učnih podatkov v ciljnem jeziku. Uspešnost večjezikovnih modelov BERT in LASER je primerljiva, razlike so odvisne od jezika. Medjezikovni prenos z modelom CroSloEngual BERT, predhodno naučenim na le treh jezikih, je v teh in nekaterih sorodnih jezikih še precej boljši.

Ključne besede: obdelava naravnega jezika, strojno učenje, vektorske vložitve besedil, analiza sentimenta, modeli BERT



Slovenščina 2.0, 2021 (1)

SLOVENE AND CROATIAN WORD EMBEDDINGS IN TERMS OF GENDER OCCUPATIONAL ANALOGIES

Matej ULČAR
Faculty of Computer and Information Science, University of Ljubljana

Anka SUPEJ
Jožef Stefan Institute

Marko ROBNIK-ŠIKONJA
Faculty of Computer and Information Science, University of Ljubljana

Senja POLLAK
Jožef Stefan Institute

Ulčar, M., Supej, A., Robnik-Šikonja, M., Pollak, S. (2021): Slovene and Croatian word embeddings in terms of gender occupational analogies. Slovenščina 2.0, 9(1): 26–59.

DOI: https://doi.org/10.4312/slo2.0.2021.1.26-59

In recent years, the use of deep neural networks and dense vector embeddings for text representation have led to excellent results in the field of computational understanding of natural language. It has also been shown that word embeddings often capture gender, racial and other types of bias. The article focuses on evaluating Slovene and Croatian word embeddings in terms of gender bias using word analogy calculations. We compiled a list of masculine and feminine nouns for occupations in Slovene and evaluated the gender bias of fastText, word2vec and ELMo embeddings with different configurations and different approaches to analogy calculations. The lowest occupational gender bias was observed with the fastText embeddings. Similarly, we compared different fastText embeddings on Croatian occupational analogies.

Keywords: word embeddings, gender bias, word analogy task, occupations, natural language processing

1 INTRODUCTION

Gender biases in language are studied from many different perspectives. Sociolinguistic studies report how language use differs between men and women (e.g., women tend to have a richer vocabulary, use typical grammatical structures, and express themselves more moderately) (Lakoff, 1973; Tannen, 1990; Argamon et al., 2003). Observations that language use varies between the genders inspired author profiling studies on texts in different languages and of different genres (Koolen and van Cranenburgh, 2017; Pardo et al., 2015; Martinc et al., 2017), also in Slovene (Verhoeven et al., 2017; Škrjanec et al., 2018).[1]

The gender dimension is present as a linguistic variation in corpora and in the form of multi-layered bias, both in individual texts and in larger corpora. Research suggests that:

- The bias is manifested as lack of mentions of women: corpora often used in research contain significantly fewer female pronouns (Zhao et al., 2018) or other references to women (Caldas-Coulhard and Moon, 2010; Baker, 2010).
- Women are less often authors or editors (Hill and Shaw, 2013): only 16% of Wikipedia editors are female.
- Corpora capture stereotypical collocations (Pearce, 2008), which refer to women primarily through their reproductive function (Gorjanc, 2007) and do not associate them with (social) power (Baker, 2010).

Recent rapid developments in natural language processing (NLP) are primarily associated with the use of deep neural networks. Their use requires a representation of text in the form of numeric vectors, called word embeddings. The relations between words are expressed in the geometry of the embedded vector space: semantically related embeddings lie close in the vector space and are arranged in similar directions. This enables the study of relations beyond superficial similarities between words, e.g. through analogies such as the

1 Note that in these studies non-binary identities are not considered. Male or female gender is assigned based on, for example, author's username on social media platforms or based on other grammatical markers.

relationship *Madrid:Spain* being analogous to the relationship *Paris:France* (Mikolov et al., 2013b).

As it turns out, word embeddings often contain bias, be it gender, race, or other types. Biases in word embeddings manifest through semantic associations and consequent proximities in the vector space (Mikolov et al., 2013b). Biases can be numerically evaluated by, for example, calculating cosine similarity between embeddings that describe a specific concept (e.g. gender) and potentially biased concepts. For example, Caliskan et al. (2017) show that word embeddings associate women with arts and men with science. Utilizing the aforementioned cosine similarity, a powerful approach to demonstrate potential bias in word embeddings is through a calculation of occupational analogies (Bolukbasi et al., 2016). Denoting a vector of word *w* with *v(w)*, this approach checks the existence of the following relationships between male and female word vectors: *v(man) - v(male occupation) ≈ v(woman) - v(female occupation)*. An example for Slovene is *v(moški) - v(učitelj) ≈ v(ženska) - v(učiteljica)*, where *učitelj* and *učiteljica* correspond to the masculine and feminine form of the noun for the concept (occupation) *teacher*, while *moški* and *ženska* denote *man* and *woman (the gender concept),* respectively. In case of no gender bias, the relationship between vectors for man and the masculine form of occupation and between the vector for woman and the feminine form of the same occupation would be approximately the same, as illustrated in Figure 1. However, being derived from naturally occurring text, it is not unexpected that human biases and social positions are captured in embeddings.

The illustration shows a simplified depiction of a few examples with 2-dimensional vectors. The arrows represent the difference between vectors *v(f)* and *v(m)*. The end points of arrows originating in masculine nouns for occupations represent the expected positions of equivalent feminine nouns if there were no bias.

In addition to studies that have shown the bias in word embeddings, different biases can be transferred onto algorithms for different NLP tasks, from machine translation (Prates et al., 2020; Vanmassenhove et al., 2018) to sentiment analysis (Kiritchenko and Mohammad, 2018). On the other hand, some authors (Nissim et al., 2019) warn that the analogy task's design may excessively emphasise biases.

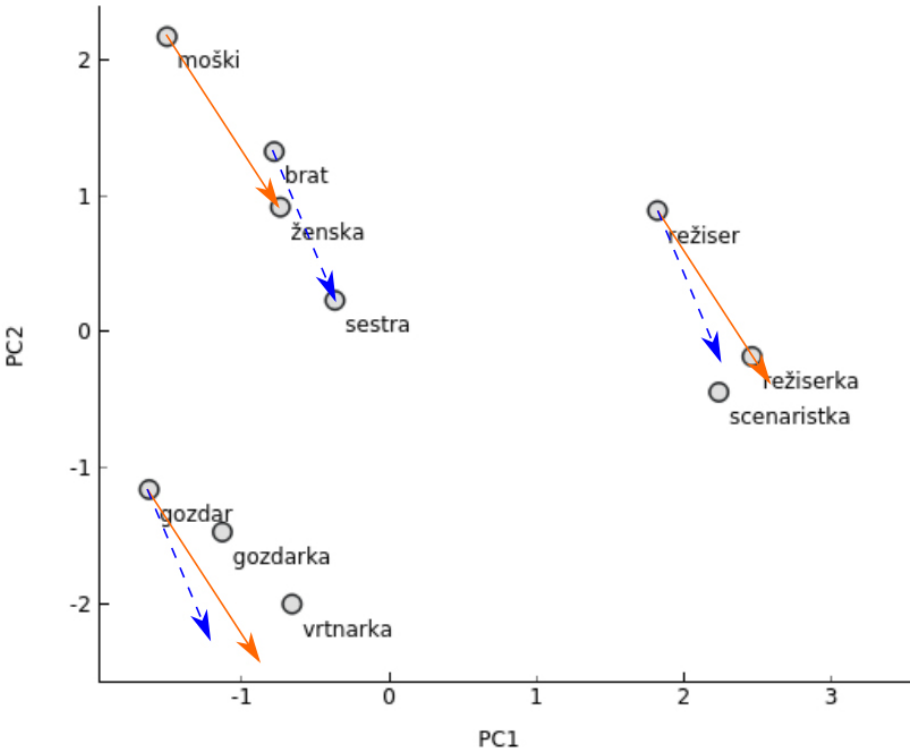


Figure 1: A simplified depiction of word vectors. The orange full arrow represents the difference between vectors for ženska [woman] and moški [man]. The blue dashed arrow represents the difference between vectors for sestra [sister] and brat [brother]. These two arrows indicate the expected (non-biased) gender difference vectors. For two male occupations, režiser [film director$_M$] and gozdar [forester$_M$], we add the gender difference vectors, and depict the resulting nearest female occupations (analogies), i.e. (gozdarka [forester$_F$] and vrtnarka [gardener$_F$]; režiserka [film director$_F$] and scenaristka [scriptwriter$_F$]). The difference to the expected non-biased point is larger for the gozdar - gozdarka pair.

Our study makes certain simplifications. First, we are not paying attention to non-binary expressions of gender, for example we do not specifically address the references such as *on/ona* or a newly proposed form introduced to be more inclusive of nonbinary gender identities *on_a* (Kern and Dobrovoljc, 2017) or noun writings of type *učitelj/učiteljica* (and *učitelj_ica*). Next, for many professions, the male form can be used as a general reference for a profession regardless of gender and we do not make any distinction between mentions of occupations when relating to a male representative or using a general mention (note also that unmarkedness of the masculine form in terms of gender is not anymore universally accepted (Kern and Dobrovoljc, 2017; Popič and

Gorjanc, 2018)). As we analyse and compare the gender bias between different embedding models, these are not severe limitations, as all the embedding models are treated equally. Moreover, similar studies on languages where the gender of a noun is not expressed morphologically can run into more serious problems (see the warnings by Nissim et al. (2019)).

The main contribution of the paper is the evaluation of Slovene and Croatian word embedding models in terms of gender, which has not yet been sufficiently researched (the exception being the analysis of the Slovene w2v model in Supej et al. (2019) and Croatian evaluation of embeddings in Svoboda and Beliga (2018)). The paper extends our work (Supej et al., 2020), where we focused on quantitative evaluation and comparison of a wide range of Slovene models and different approaches to evaluation, while in this paper, we extend the work and also compare Croatian word embeddings models. The focus of the paper is to draw the attention of the developers of linguistic and technological tools (which are based on word embeddings) to the implications the usage of biased embeddings might have. Despite indirectly problematising language bias and pointing out several stereotypical associations, a detailed critical interpretation falls out of this paper's scope.

The paper is divided into further six sections. We first present related work (Section 2). Section 3 describes Slovene and Croatian lists of male and female occupations and specifies the word embedding models used. In Sections 4 and 5, methodology and results are addressed, followed by a discussion in Section 6, and conclusions with plans for further work in Section 7.

2 RELATED WORK

Language corpora and datasets reflect linguistic variations (including different types of bias) in relation to social factors. NLP tools are trained on these data and can inherit the contained variations and biases. The bias in corpora can negatively impact NLP tools (Sun et al., 2019) and can perpetuate biases held towards certain groups. Word embeddings are trained on large corpora to capture syntactic and semantic relations between words and capture the expressed biases.

For instance, it has been shown that standard training data sets for part-of-speech perform better on older people's language (Hovy and Søgaard, 2015).

Garimella et al. (2019) show that a part-of-speech tagger and a dependency parser perform successfully on texts written by women, regardless of what data they had been trained on initially. On the other hand, male authors' texts are better tagged/parsed when the training data contained enough texts written by men. The success of tools such as parsers on male authors' texts may be due to the imbalances in the training data favouring male authorship. It has also been shown that NLP tools are more effective when demographic variations are considered (Volkova et al., 2013; Hovy, 2015). Hovy (2015) shows that including the information on the age and gender of authors improves the performance of three tasks in five different languages.

Biases can have negative consequences in the coreference resolution task (Zhao et al., 2018) and can perpetuate biases held towards certain groups (see examples in Zhao et al., 2017). In the context of texts on mental illness, Hutchinson et al. (2020) note that topics such as gun violence, homelessness, and addiction are over-represented, leading to disability topics receiving particularly negative scores in sentiment analysis tasks. Besides the aspects above, some authors call the attention to the effect biases can have on detection tools. For example, misogyny detection models may attribute high scores to non-misogynous texts simply because the latter contain the so-called identity terms, i.e. terms associated with misogyny (Nozza et al., 2019). In sum, the interplay of bias and NLP is an important and interesting field receiving increasing attention, notably regarding word embeddings, as explained next.

In terms of word embeddings, researchers have studied bias by investigating the proximity of gender-related words to other words in the vector space. For example, Garg et al. (2018) show that the adjective *honourable* lies closer to the word *man* than to the word *woman*. Second, biases are reflected in analogies, e.g. Bolukbasi et al. (2016) show that the embedding space solution of the analogy *man:computer programmer* ≈ *woman:x* is *x* = *homemaker*. Nissim et al. (2019) warn that such analogies overemphasise the practical impact of the biases.

As already mentioned, gender bias in word embeddings is often studied on analogies of occupations, which is also our study's case. In morphologically rich languages, such as Slovene and Croatian, the gender of words is expressed morphologically. Therefore, the result of the gender analogy is expected to be

the female form of the male variant of the occupation (and vice versa). Svoboda and Beliga (2018) included masculine and feminine versions of job positions in Croatian as one of the evaluation aspects of Croatian word2vec and fastText word embeddings. Preliminary research on word2vec embeddings in Slovene (Supej et al., 2019) showed that the analogy task's accuracy is reasonably high both when attempting to find the female and the male equivalent of an occupation. Results nevertheless reflect gender biases: the first result of the analogy *woman:secretary* ≈ *man:x* is *x* = *boss*, while the first ten results of different analogies indicate other gender inequalities: the association of women with house chores and men with occupations of a higher status etc. In the work of Supej et al. (2020) that we extend in this paper, different word2vec, fastText and ELMo embeddings are compared on Slovene pairs of male and female occupations.

As tools based on biased word embeddings may reinforce biases (Zhao et al., 2017), many research groups focused on *debiasing* word embeddings: the main goal of such algorithms is to prevent language models from reproducing racist, sexist or in other ways harmful content. Debiasing also has other advantages – it has been shown that debiasing contributes to correct coreference resolution (Zhao et al., 2018). Some examples of these methods are equalising the distances between gender-specific words and occupations (Bolukbasi et al., 2016; Bordia and Bowman, 2019), inserting additional restrictions into the training corpus (e.g. ensuring equal representation of occupational activities between the genders in the training data) (Zhao et al., 2017), removing texts that cause bias (Brunet et al., 2019), and training gender-neutral word embeddings (Zhao et al., 2018). Schick et al. (2021) recently proposed a self-diagnosis and self-debiasing model where large language models examine their outputs regarding the potential presence of undesirable attributes. They introduced a debiasing algorithm that reduces the likelihood of a model producing biased text. Moreover, researchers recently also focused on methods for debiasing sentence representations, addressing the difficulty of retraining models that are often proposed in debiasing research (retraining models like BERT and ELMo often proves infeasible in practice) (Liang et al., 2020). Gonen and Goldberg (2019) caution that many debiasing methods only conceal bias, which continues to be present in the embeddings, and that many metrics used in the debiasing

research have only positive predictive ability (i.e. they can detect the presence of bias but not its absence). On the other hand, studies such as Hirasawa and Komachi (2019) show that debiasing improves multimodal machine translation, thereby underlining the promising future of this research field. In our study, we do not aim to debias embeddings but only compare different embedding approaches in Slovene and Croatian concerning their gender bias.

3 DATA

In this section, we first present the lists of occupations in Slovene and Croatian we used to analyse gender biases, followed by the embedding models.

3.1 List of occupations

We first describe the list of occupations we collected for Slovene, followed by its equivalent in Croatian. Our selection of occupations in Slovene is based on the Standard Classification of Occupations (Vlada RS, 1997), based on the *International Standard Classification of Occupations*. Most occupations in this classification are multi-word expressions (e.g. *upravljalec/upravljalka metalurškega žerjava* [en. *metallurgical crane operator*]), which are less suitable for computation with embeddings due to their specificity and length. To calculate analogies, we limit our approach to single-word occupations. The complete list of single-word occupations in Slovene includes 422 male/female occupation pairs, further reduced in line with the following criteria:

1. An occupation has to exist both in female and male grammatical gender (gender-neutral words such as *pismonoša* [en. *postman*] are not included in the list).

2. An occupation as a common noun occurs at least 500 times in the Corpus of Written Standard Slovene *Gigafida 2.0* (2020).

3. When a more established version of the occupation exists, we manually add a synonym with the same root (e.g. in the case of *fotografka,* an arguably more established *fotografinja* was added [en. *photographer*]). When calculating analogies, the form more frequent in the corpora is inserted at the input, but all synonyms (if they appear among the results) are considered a correctly solved analogy.

4. If the standard classification does not include the female (e.g. *dramatik* [en. *playwright*]) or male variant (e.g. *prostitutka* [en. *prostitute*]) of the occupation, the missing version is manually added if it exists and appears in the Gigafida corpus (e.g. there are no established words for female and male versions of *postrešček* [en. *porter*] and *hostesa* [en. *hostess*], respectively).

5. Occupations where either the female or the male occupation variant is a homograph (e.g. *detektivka* [en. *detective*] also denotes a detective novel) or where an occupation could be associated with a context unrelated to occupations (e.g. *čarovnik/čarovnica* [en. *wizard/witch*]), were excluded from the final set of occupations. Likewise, we filtered out occupations that are also proper names, such as *kovač* [en. *blacksmith*]; for differentiating between common nouns and proper names Sloleks 2.0 (Dobrovoljc et al., 2019) was used. The final list contains 234 occupation pairs and is freely accessible in the CLARIN repository[2].

For Croatian, we compiled a list of occupations from two existing sources. The first source contains occupations from the word analogy dataset by Svoboda and Beliga (2018). It consists of 109 pairs of single-word occupations. The second source is ESCO (European Skills, Competences, Qualifications and Occupations)[3] and lists 2942 occupations in male and female form. Similar to the Slovene list of occupations, most of the classifications from ESCO are multi-word expressions, e.g. *špediterski službenik / špediterska službenica za uvoz i izvoz riba, rakova i mekušaca* [en. *import-export specialist in fish, crustaceans and molluscs*]. After removing all multi-word occupations, the ESCO source contains 309 pairs of single-word occupations. The final, combined list from both sources, filtered to remove duplicates, contains 375 occupation pairs.

3.2 Word embedding models

Different configurations of word embeddings for Slovenian and Croatian were used in the experimental phase. We first list the Slovene embedding models followed by the Croatian ones.

2 http://hdl.handle.net/11356/1347

3 https://ec.europa.eu/esco/portal

3.2.1 Slovene word embedding models

We analyse two non-contextual embedding models, fastText and word2vec, and the ELMo contextual model.

- fastText (Bojanowski et al., 2017):
  - 100-dimensional vectors, trained on Gigafida 2.0 in the EU EMBEDDIA[4] project,
  - 300-dimensional vectors, trained as above,
  - 100-dimensional word vectors from the Sketch Engine portal (*word*),
  - 100-dimensional word vectors from the Sketch Engine portal, where vectors are embeddings of word lemmas,
  - 100-dimensional CLARIN.SI-embed.sl vectors (Ljubešić and Erjavec, 2018), and
  - 300-dimensional vectors from the fastText.cc portal;
- word2vec (Mikolov et al., 2013a): 256-dimensional vectors, trained for the needs of the Kontekst.io portal (Plahuta, 2020); available at request[5];
- ELMo (Peters et al., 2018): 1024-dimensional vectors, contextual embeddings built in the EU EMBEDDIA project, trained on Gigafida (Ulčar, 2019). Contextual embeddings produce a different vector for each occurrence of the word based on its context. We computed word vectors from sentences in Slovene Wikipedia. To get a single representation for each word, comparable to other embeddings, for each of the 200,000 most common words, we calculated the centroid vector of all word occurrences. Several different types of vectors were used:
  - vectors from the output of the first (CNN) layer of the network that is context-independent (i.e. *layer 0*),

4 http://embeddia.eu/

5 https://kontekst.io/kontakt

- vectors from the output of the second (first LSTM) layer of the network that is context-dependent (i.e. *layer 1*),
- vectors from the output of the third (second LSTM) layer of the network that is context-dependent (i.e. *layer 2*).

3.2.2 Croatian word embedding model

For the Croatian language, we analyse several non-contextual embedding models:

- fastText (Bojanowski et al., 2017):
  - 100-dimensional vectors, trained in the EU EMBEDDIA project,
  - 300-dimensional vectors, trained as above,
  - 100-dimensional CLARIN.SI-embed.hr vectors of words and lemmas (Ljubešić, 2018),
  - 300-dimensional vectors from the fastText.cc portal.

4 EVALUATION METHODOLOGY

To assess the gender bias for each of the embedding models and each occupation, we calculated occupational analogies in four ways. However, the core analogy computation is the same in all cases: for every occupation of a masculine grammatical gender O_m, we search for a feminine noun equivalent O_f. The following vector is calculated:

$$v(d) = v(O_m) - v(m) + v(f),$$

where $v(m)$ is the male vector, and $v(f)$ is the female vector. If there were no gender biases, $v(d)$ would be equal or very similar to $v(O_f)$. For every vector $v(d)$, we find N closest word vectors according to the cosine similarity (we use $N = 1, 5,$ or 10). When searching for closest words, all words appearing in the embeddings are considered, except for the words *man*, *woman*, the word O_m, and the words containing non-alphabetic characters (numbers, hyphens, punctuation etc.). If the word O_f is located among the N-closest words, we consider the analogy correct; else it is marked as incorrect. We convert all letters to lowercase: e.g. the words *Zdravnik*, *zdravnik* and *ZDRAVNIK* are

all converted to *zdravnik* and thus considered the same word. The process is repeated for each female variant of an occupation O_f where we look for the male equivalent O_m. Here, the vector $v(d)$ is calculated as:

$$v(d) = v(O_f) - v(f) + v(m).$$

When looking for closest words, O_f is omitted from the set of words, just as O_m was ignored before. The final result represents the proportion of correctly determined cases. The metric is called *precision at N* (*P@N*). A higher *N* allows for finding additional closest hits in the vector space.

Two approaches were used to determine the baseline male vector $v(m)$ and female vector $v(f)$:

- The first approach defines *m* simply as the word *man* and *f* as *woman* (in Slovene corresponding to *moški* and *ženska* and in Croatian to *muškarac* and *žena).*
- In the second approach, similarly to Bolukbasi et al. (2016), the difference $v(f) - v(m)$ or $v(m) - v(f)$ is defined as the average difference of vectors of word pairs which refer specifically to a woman or man (Table 1).

Table 1: *Inherently male-female word pairs in Slovene (left) and Croatian (right)*

Slovene male-female word pairs		**Croatian male-female word pairs**	
m	***f***	***m***	***f***
moški [man]	ženska [woman]	muškarac [man]	žena [woman]
gospod [sir]	gospa [madam]	gospodin [sir]	gosopođa [madam]
fant [boy]	dekle [girl]	momak [boy]	djevojka [girl]
deček [boy]	deklica [girl]	dječak [boy]	djevojčica [girl]
brat [brother]	sestra [sister]	brat [brother]	sestra [sister]
oče [father]	mati [mother]	otac [father]	majka [mother]
sin [son]	hči [daughter]	sin [son]	kći [daughter]
dedek [grandfather]	babica [grandmother]	djed [grandfather]	baka [grandmother]
mož [husband]	žena [wife]	suprug [husband]	supruga [wife]
on [he]	ona [she]	on [he]	ona [she]
fant [boy]	punca [girl]	tata [dad]	mama [mum]
stric [uncle]	teta [aunt]		

When searching for the N closest words, we also tested lemmatisation's influence: in this case, all words in word embeddings were lemmatised using the LemmaGen[6] tool. By doing so, the effect of different word forms stemming from, e.g. conjugation and declination, was offset: for example, word forms *zdravnico* and *zdravnice* are considered a single near word since they share the same lemma *zdravnica* [doctor_{F}].

5 RESULTS

We present the results showing biases in all embeddings described in Section 3. We use the *P@N* measure, where *N* equals 1, 5, or 10. Some of the occupations from our list are not covered by all word embeddings, i.e. there is no word vector for them. Any example where the searched-for word is not among the top *N* closest words is counted as incorrect, even if the searched-for word does not appear in the embeddings. In cases where the embeddings do not cover the input occupation, and we cannot calculate the vector *v(d),* we dismiss all such examples so that they do not affect the final result. The reader, interested in the results where non-covered examples are also considered, is referred to our conference paper (Supej et al., 2020).

The results for Slovene analogies are presented in Table 2 and for the Croatian analogies in Table 3. Results for experiments where we have a masculine expression for the occupation O_m as the input, and we search for the equivalent feminine expression of the same occupation O_f, are shown in the rightmost columns (*m* input) for each language. Results, where we have O_f as the input and search for O_m, are shown in leftmost columns (*f* input) for each language. As explained in Section 4, we tested different approaches. The approaches where we lemmatised all the words or used the average difference of vectors of pairs of words from Table 1 generally perform better (i.e. they express lower gender bias). These two options have the suffixes *lem* and *avg* appended in the tables, respectively. In this section, we only show the results for applying both of these options (we do not apply lemmatisation to fastText (lemma) embeddings as they are already lemmatised). Full results are presented in Appendix A in Table 8 for Slovenian and in Table 9 for Croatian.

6 https://github.com/vpodpecan/lemmagen3/

Table 2: *Results for all Slovenian embeddings*

Slovene word embeddings	dimensions and approach	*f* input			*m* input		
		P@1	P@5	P@10	P@1	P@5	P@10
ELMo Embeddia	1024D l0 lem avg	0.907	0.933	0.947	0.370	0.398	0.403
	1024D l1 lem avg	0.907	0.947	0.947	0.381	0.392	0.398
	1024D l2 lem avg	0.880	0.933	0.933	0.376	0.398	0.398
fastText.cc	300D lem avg	0.613	0.884	0.948	0.655	0.755	0.764
fastText Embeddia	100D lem avg	0.906	0.971	0.976	0.677	0.720	0.724
	300D lem avg	**0.947**	**0.976**	**0.982**	0.685	0.720	0.724
fastText CLARIN.SI-embed.sl	100D lem avg	0.839	0.940	0.950	**0.761**	**0.880**	**0.902**
fastText Sketch Engine (word)	100D lem avg	0.930	0.962	0.973	0.725	0.781	0.785
fastText Sketch Engine (lemma)	100D avg	0.673	0.931	0.960	0.598	0.786	0.821
word2vec Kontekst.io	256D lem avg	0.679	0.853	0.872	0.407	0.550	0.593

Note. Results for each approach, where we have a feminine word for occupation on the input (*f* input), and we search for the equivalent masculine term, and where we have a masculine word for occupation on the input (*m* input), and we search for the equivalent feminine term. The examples where the embeddings do not cover the input occupation were dismissed. The best result in each column is in bold.

Table 3: *Results for all Croatian embeddings*

Croatian word embeddings	dimensions and approach	*f* input			*m* input		
		P@1	P@5	P@10	P@1	P@5	P@10
fastText.cc	300D lem avg	0.731	0.939	0.954	0.546	0.637	0.644
fastText Embeddia	100D lem avg	0.905	0.941	0.968	0.625	0.666	0.672
	300D lem avg	**0.923**	**0.982**	**0.986**	0.631	0.675	0.678
fastText CLARIN.SI-embed.hr (word)	100D lem avg	0.907	0.930	0.944	**0.673**	**0.746**	**0.754**
fastText CLARIN.SI-embed.hr (lemma)	100D avg	0.244	0.678	0.826	0.266	0.521	0.588

Note. For each approach, where we have a feminine word for occupation on the input (*f* input) and we search for the equivalent masculine term, and where we have a masculine word for occupation on the input (*m* input) and we search for the equivalent feminine term. The examples where the embeddings do not cover the input occupation were dismissed. The best result in each column is in bold.

The results show that both lemmatisation of the words and using the average of several inherently male or female words for male and female vectors improve the reported scores. Applying both approaches gives the best results in most cases. For finding the closest *N* words, we have also tried the CSLS

measure (Cross-Domain Similarity Local Scaling) (Conneau et al., 2018) instead of the cosine similarity. This measure avoids the problem of hubness in the search for nearest neighbours. Namely, some words (called hubs in the nearest neighbour graph representation) may be nearest neighbours of many other words, while others are nearest neighbours of no other word (outliers). CSLS computes nearest neighbours in both directions and largely avoids the problem of hubness. For the experiments with O_f on the input and searching for O_m, there is no significant difference in results between the cosine similarity and CSLS. For the experiments with O_m on the input and searching for O_f, using CSLS gives lower precision than the cosine similarity. This is especially the case where we used the words "man" and "woman" for vectors *v(m)* and *v(f)*. When using averages of several inherently male and female words for vectors *v(m)* and *v(f)*, the difference in precision between the cosine similarity and CSLS is smaller, but the cosine similarity still outperforms CSLS.

We give a more detailed discussion of the results for each approach in the next section. We only present the results of the cosine similarity measure.

6 DISCUSSION

In the case of Slovene word embeddings, the fastText CLARIN.SI-embed.sl embeddings reach the highest precision in the analogy task for male versions of occupations at the input (Table 2). When there are female versions of occupations at the input, the embedding model reaching the highest precision is fastText Embeddia. Similar results are observed for Croatian embeddings (Table 3). Lemmatisation of the output and averaging several inherently male and female words for vectors *v(m)* and *v(f)* (instead of using only the embeddings for *woman* or *man*) improves the precision in the analogy task for different models and different input data. As described in Section 5, we dismiss the examples where the embeddings do not cover the input occupation. If we do not dismiss these examples but instead count them as incorrect, the share of occupations covered by the embeddings has the largest effect on the score. The results for Slovene can be found in our paper (Supej et al., 2020). The fastText CLARIN.SI embeddings would then score the best, as these embeddings cover the occupations best. This is especially important for the female occupations since they have much lower coverage than male occupations.

Results in Table 2 and Table 3 have been filtered, so that the words *man, woman* and the occupation on the input are removed from the list of analogy results, as explained in Section 4. With unfiltered results, the input occupation is often the result of the analogy task (Table 4). For more detailed results (not only with lemmatisation and using several inherently male and female words for *v(m)* and *v(f)*) see Table 10 in Appendix A.

With the fastText Embeddia model, we reach similar results using 100- and 300-dimensional vectors (see Table 2 and Table 3). Other embeddings are not directly comparable with regards to dimensionality as they were trained on different resources. However, corpora used to train the embeddings play a more important role than the number of dimensions. The FastText Embeddia model in Table 4 shows that dimensionality plays a role in determining how often the input occupation is the result of the analogy. In a different setup, when considering the occupations that are not covered in the embeddings, dimensionality strongly influences the results (Supej et al., 2020).

Table 4: *Share of cases where the result of the analogy with the highest cosine similarity is the input occupation itself - before filtering is done to produce the results in Table 2 and Table 3 (both male to female and female to male analogies)*

Slovene word embeddings	Dimensions and approach	Share of outputs equal to inputs	Croatian word embeddings	Dimensions and approach	Share of outputs equal to inputs
ELMo Embeddia	1024D l0 lem avg	0.547			
	1024D l1 lem avg	0.423			
	1024D l2 lem avg	**0.064**			
fT fastText.cc	300D lem avg	0.831	fT fastText.cc	300D lem avg	0.672
fT Embeddia	100D lem avg	0.143	fT Embeddia	100D lem avg	**0.094**
	300D lem avg	0.419		300D lem avg	0.352
fT CLARIN.SI-embed.sl (word)	100D lem avg	0.316	fT CLARIN.SI-embed. hr (word)	100D lem avg	0.103
fT Sketch Engine (word)	100D lem avg	0.096			
fT Sketch Engine (lemma)	100D avg	0.803	fT CLARIN.SI-embed.hr (lemma)	100D avg	0.837
w2v Kontekst.io	256D lem avg	0.483			

Note. The number of all cases is 468 (from 234 occupation pairs) for Slovene and 750 (from 375 occupation pairs) for Croatian.

The coverage of masculine occupations is higher than that of feminine occupations in all word embedding models (Table 5). FastText CLARIN.SI-embed.sl word embeddings achieve the highest coverage of female occupations, while ELMo word embeddings contained only 75 of the 234 female occupations. As explained in Section 3.2.1, ELMo embeddings are limited to only 200,000 most common words in Wikipedia; therefore, we have significantly lower coverage of occupations for ELMo. For comparison, other word embedding models cover around 1 million words. Masculine occupations that do not appear in the embeddings are typically occupations associated with women (e.g. male variants of *seamstress* and *cosmetician*, in Slovene *šiviljec* and *kozmetik*, respectively). Likewise, feminine occupations not present in the embeddings are traditionally male occupations (e.g. embedding models do not contain female variants of occupations like *auto mechanic* and *carpenter* (in Slovene *avtomehaničarka* and *tesarka*, respectively), or occupations that have been culturally taken up exclusively by men, e.g., *nadškof* (en. *archbishop*). Poor representation of female occupations can also be attributed to other factors – Zhao et al. (2018) report that the mentions referring to men are more likely to contain a job title compared to female mentions.

Table 5: *Coverage of male (m) and female (f) occupations from the list in different embeddings as a ratio between covered occupations and all occupations*

Slovene embeddings	*m*	*f*	Croatian embeddings	*m*	*f*
ELMo	0.774	0.321			
fastText cc	0.979	0.739	fastText cc	0.848	0.527
fastText Embeddia	0.991	0.726	fastText Embeddia	0.856	0.594
fastText CLARIN.SI-embedd.sl	**1.000**	**0.932**	fastText CLARIN.SI-embedd.hr (word)	0.914	**0.722**
fastText Sketch Engine (word)	0.996	0.791	fastText CLARIN.si-embedd.hr (lemma)	**0.955**	**0.722**
fastText Sketch Engine (lemma)	**1.000**	0.863			
word2vec Kontekst.io	0.987	0.667			

Nissim et al. (2019) claim that most studies exaggerate biases pointed out by analogy tasks. The design of these studies excludes the input occupation from the possible results, even if the calculations could lead to this exact occupation to have the highest cosine similarity and hence appear in the results. This criticism is more relevant for English studies as in Slovene the gender in

occupations is for the most part expressed by word morphology. Even though we omitted the input occupations from the results, which is a standard practice when calculating analogies, we analysed the results before this filtering. Analysis of the results showed that the input occupation is indeed often the result with the highest cosine similarity (Table 4), varying significantly between different models.

When manually comparing the results of different models from Tables 2 and 3, we also notice several differences between the models. In the case of ELMo and word2vec models, the outputs are largely occupations. The results of the analogy task in the case of fastText Embeddia, CLARIN.SI-embed.sl and Sketch Engine (word) are occupations, as well as words related to the occupation on the input, or words that share the same root as the input occupation. Results of the fastText.cc and Sketch Engine (lemma) models are typically words sharing the root with the input occupation.

Analogy results are interesting from a semantic point of view. The first results of the analogy task (Slovene "fastText Embeddia 100D lem avg") *ženska:krojačica :: moški:x* being *x=krojač* [en. *woman:tailor*$_F$ *:: man:tailor*$_M$] and *ženska:šivilja :: moški:x* being *x=krojač* [en. *woman:seamstress :: man:tailor*] are interesting. For example, while word embedding of *šiviljec* [en. *seamster*] is not available, *krojač* [en. *tailor*], a semantically linked one, from another morphological word family is. Another interesting element is illustrated by one of the results of the analogy: *ženska:manekenka :: moški:x* where *x=nogometaš* [en. *woman:model :: man:footballer*] (Croatian "fastText Embeddia 100D lem avg"). While *model* and *footballer* are not corresponding to the same professions, this result is an indication that female models and male footballers appear in similar textual contexts. It would be interesting to investigate those contexts further (e.g. both occupations represent desirable identities, such as being beautiful, rich, famous, successful).

There are indeed more examples where results of certain analogies (especially in the case of "word2vec Kontekst.io lem avg model") are not linked to the input occupation or are stereotypical. For example, the results of the analogy *moški:rudar :: ženska:x* in the aforementioned w2v model are, e.g. *barbika* [en. *barbie*], *klovnesa* [en. *clown*$_F$], *čarovnica* [en. *witch*], *lutka* [en. *doll*], *prostitutka* [en. *prostitute*$_F$], *akrobatka* [en. *acrobat*$_F$], *najstnica* [en.

teenager$_F$], *opica* [en. *monkey*], *princeska* [en. *princess*], *striptizeta* [en. *stripper*$_F$]. The case of stereotypical analogies in the w2v model is pointed out by Supej et al. (2019).

As part of the analysis, a frequency list of analogy results for female and male input occupations was compiled for each word embedding model (only the *lem avg* configuration of the models was taken into account) (see Table 6 for Slovene and Table 7 for Croatian).

The most frequently occurring words mostly follow the pattern that for a male occupation on the input, a female occupation is expected on the output. Presented Slovene embedding models follow this pattern; in the case of the Croatian embeddings, there are several examples among the frequently occurring words that do not follow the pattern: in the "fastText cc lem avg" with a female occupation on the input, there are several frequently occurring female occupation variants also on the output, e.g. *ethicist*, *biologist* (*etičarka*, *biologinja*, respectively). For *etičarka,* it is possible that this result is influenced by other similar words (e.g. *kozmetičarka*), as fastText models consider subword information. The most frequently occurring words are primarily occupations but not always – for example, female Scottish national (*Škotkinja*) and *father* (*otac*) frequently appear in the Croatian "fastText cc lem avg" model while one of the frequent words in the Slovene "word2vec Kontekst.io lem avg" is *korenjak* (denoting a brave man).

In Slovene word embeddings, we notice a pattern of the most frequently occurring feminine occupations/words appearing more often than the most frequently occurring male occupations in the "ELMo l2 lem avg" and "w2v Kontekst.io lem avg" models. Similar is observed for Croatian models presented in Table 7; however, the most frequently occurring words appear less often than in the Slovene embeddings. One possible explanation is that the models mentioned above contain fewer word embeddings than some other models (200,000 or approximately 600,000 for each model). Both models exhibit a lower representation of the female versions of occupations in the embeddings. Occupations that nevertheless appear in the embeddings, therefore, reappear more often. There are overall more male occupations in the embeddings, possibly causing individual male occupations to come up less frequently than female ones.

Table 6: *Most common words that appear among the top 10 results of the analogy task (that is, among the 10 closest words to the searched-for term, based on the cosine similarity measure) for selected Slovene embedding models*

ELMo Embeddia l2 lem avg				fastText CLARIN.SI lem avg				word2vec Kontekst.io lem avg			
***m* input**		***f* input**		***m* input**		***f* input**		***m* input**		***f* input**	
Result	***n***	**Result**	***n***	**Result**	***n***	**Result**	***n***	**Result**	***n***	**Result**	***n***
bolničarka [nurse]	47	geograf [geographer$_M$]	9	šivilja [seamstress]	15	mizar [carpenter$_M$]	11	kuharica [cook$_F$]	44	ortoped [orthopedist$_M$]	14
biokemičarka [biochemist$_F$]	39	politolog [political scientist$_M$]	8	ključavničarka [locksmith$_F$]	11	biology [biologist$_M$]	10	gospodinja [homemaker$_F$]	38	pisatelj [writer$_M$]	14
frizerka [hairdresser$_F$]	39	biolog [biologist$_M$]	7	inštalaterka [installer$_F$]	9	ključavničar [locksmith$_M$]	9	šivilja [seamstress]	33	kardiolog [cardiologist$_M$]	13
trgovka [salesperson$_F$]	39	dramaturg [playwright$_M$]	7	keramičarka [ceramist$_F$]	9	zgodovinar [historian$_M$]	9	frizerka [hairdresser$_F$]	32	nevrolog [neurologist$_M$]	13
čistilka [cleaner$_F$]	34	književnik [writer$_M$]	7	filologinja [philologist$_F$]	8	internist [internist$_M$]	8	kozmetičarka [cosmetician$_F$]	30	urolog [urologist$_M$]	13
znanstvenica [scientist$_F$]	34	scenarist [screenwriter$_M$]	7	oftalmologinja [ophthalmologist$_F$]	8	režiser [director$_M$]	8	čistilka [cleaner$_F$]	29	psihiater [psychiatrist$_M$]	12
kuharica [cook$_F$]	33	animator [animator$_M$]	6	filozofinja [philosopher$_F$]	7	arheolog [archeologist$_M$]	7	fotografinja [photographer$_F$]	29	ekolog [ecologist$_M$]	11
geologinja [geologist$_F$]	30	esejist [essayist$_M$]	6	geofizičarka [geophysicist$_F$]	7	natakar [waiter$_M$]	7	zdravnica [doctor$_F$]	29	hišnik [janitor$_M$]	11
perica [laundress]	28	etnolog [ethnologist$_M$]	6	kmetica [farmer$_F$]	7	pisatelj [writer$_M$]	7	služkinja [maid]	26	biolog [biologist$_M$]	10
služkinja [maid]	28	fotograf [photographer$_M$]	6	nevrokirurginja [neurosurgeon$_F$]	7	primarij [senior doctor$_M$]	7	trgovka [salesperson$_F$]	26	korenjak [brave man]	10
biologinja [biologist$_F$]	27	illustrator [illustrator$_M$]	6	strugarka [worker using a planer machine$_F$]	7	stomatolog [stomatologist$_M$]	7	slikarka [painter$_F$]	25	maneken [model$_M$]	10
gospodinja [homemaker$_F$]	26	lutkar [puppeteer$_M$]	6	geologinja [geologist$_F$]	6	tesar [carpenter$_M$]	7	tajnica [secretary$_F$]	25	režiser [director$_M$]	10
matematičarka [mathematician$_F$]	26	paleontolog [paleontologist$_M$]	6	hematologinja [hematologist$_F$]	6	fotoreporter [photojournalist$_M$]	6	veterinarka [veterinarian$_F$]	25	akademik [academic$_M$]	9
mikrobiologinja [microbiologist$_F$]	26	pravnik [jurist$_M$]	6	kardiologinja [cardiologist$_F$]	6	gostilničar [innkeeper$_M$]	6	znanstvenica [scientist$_F$]	25	akademski slikar [academic painter$_M$]	9
arheologinja [archeologist$_F$]	25	režiser [director$_M$]	6	paleontologinja [paleontologist$_F$]	6	kardiolog [cardiologist$_M$]	6	socialna delavka [social worker$_F$]	24	glasbenik [musician$_M$]	9

Table 7: *15 most common words that appear among the top 10 results of the analogy task (that is, among the 10 closest words to the searched-for term, based on the cosine similarity measure) for selected Croatian embedding models*

ELMo Embeddia l2 lem avg				fastText cc lem avg				fastText CLARIN.SI-embedd.hr (word) lem avg			
***m* input**		***f* input**		***m* input**		***f* input**		***m* input**		***f* input**	
Result	***n***	**Result**	***n***	**Result**	***n***	**Result**	***n***	**Result**	***n***	**Result**	***n***
krojačica [tailor$_F$]	34	povjesničar [historian$_M$]	10	kemičarka [chemist$_F$]	12	etičarka [ethicist$_F$]	8	krojačica [tailor$_F$]	31	znanstvenik [scientist$_M$]	16
automehaničarka [auto mechanic$_F$]	29	konobar [waiter$_M$]	10	vještakinja [expert$_F$]	11	otfamologinja [ophthalmologist$_F$]	7	automehaničarka [auto mechanic$_F$]	23	biology [biologist$_M$]	16
zavarivačica [welder$_F$]	20	biolog [biologist$_M$]	9	fizičarka [physicist$_F$]	10	redatelj [director$_M$]	6	zavarivačica [welder$_F$]	22	profesor [professor$_M$]	9
keramičarka [ceramist$_F$]	16	umjetnik [artist$_M$]	8	biokemičarka [biochemist$_F$]	10	glumac [actor$_M$]	6	šivačica [seamstress]	18	povjesničar [historian$_M$]	9
kemičarka [chemist$_F$]	15	sociolog [sociologist$_M$]	8	vozačica [driver$_F$]	9	biologinja [biologist$_F$]	6	keramičarka [ceramist$_F$]	18	konobar [waiter$_M$]	9
biokemičarka [biochemist$_F$]	15	fizioterapeut [physiotherapist$_M$]	8	pravnica [jurist$_F$]	9	paleografkinja [paleographer$_F$]	5	soboslikarica [painter-decorator$_F$]	17	genetičar [geneticist$_M$]	9
šivačica [seamstress]	14	redatelj [director$_M$]	7	frizerka [hairdresser$_F$]	9	ihtiologinja [ichthyologist$_F$]	5	biokemičarka [biochemist$_F$]	16	redatelj [director$_M$]	8
spremačica [maid]	14	poslovođa [manager$_{F/M}$]	7	masažerka [massage therapist$_F$]	8	suscenarist [co-screenwriter$_M$]	4	kemičarka [chemist$_F$]	15	poslovođa [manager$_{F/M}$]	8
čistačica [cleaner$_F$]	13	paleontolog [paleontologist$_M$]	7	tehničarka [technician$_F$]	7	scenografkinja [scenographer$_F$]	4	genetičarka [geneticist$_F$]	13	policajac [police officer$_M$]	8
genetičarka [geneticist$_F$]	13	književnik [writer$_M$]	7	političarka [politician$_F$]	7	otac [father]	4	cvjećarka [florist$_F$]	12	zaposlenik [employee$_M$]	7
fizičarka [physicist$_F$]	13	geologinja [geologist$_F$]	7	matematičarka [mathematician$_F$]	7	književnik [writer$_M$]	4	biofizičarka [biophysicist$_F$]	12	umjetnik [artist$_M$]	7
astrofizičarka [astrophysicist$_F$]	13	dramaturg [playwright$_M$]	7	lutkarica [puppeteer$_F$]	7	dopukovnik [lieutenant colonel$_M$]	4	znanstvenica [scientist$_F$]	11	sociolog [sociologist$_M$]	7
šnajderica [seamstress]	12	znanstvenik [scientist$_M$]	6	glumica [actor$_F$]	7	daktilografkinja [typist$_F$]	4	geologinja [geologist$_F$]	11	snimatelj [cameraman]	7
mehaničarka [mechanic$_F$]	12	zaštitar [security guard$_M$]	6	trgovkinja [salesperson$_F$]	6	astrobiologinja [astrobiologist$_F$]	4	tehničarka [technician$_F$]	10	satnik [captain$_M$]	7
informatičarka [computer scientist$_F$]	12	sociologinja [sociologist$_F$]	6	terapeutkinja [therapist$_F$]	6	škotkinja [Scottish national$_F$]	3	mehaničarka [mechanic$_F$]	10	porter [doorkeeper$_M$]	7

In the case of the Slovene "ELMo l2 lem avg" and "w2v Kontekst.io lem avg" models, occupations of a lower social class (*čistilka* [en. *cleaner*$_F$], *perica* [en. *laundress*], *gospodinja* [en. *homemaker*$_F$]), as well as archaic occupations with women in inferior roles (*služkinja* [en. *maid*]) are observed among the frequent analogy results of female grammatical gender. Socially inferior occupations are rare among the most frequent male analogies. There are less socially inferior occupations observed among the Croatian results (exceptions being, e. g., the female variants of *cleaner* and *maid* (*čistačica* and *spremačica,* respectively) in the "ELMo Embeddia l2 lem avg" model).

We observed that certain words (especially female occupations) appear among the results despite being semantically unrelated to the input occupation. Several analogy results (especially in the case of a typical male occupation on the input) are unrelated to the input occupation (e.g. *bolničarka* [en. nurse$_F$] is the first result of the analogy *moški:rudar :: ženska:x* [en. *man:miner :: woman:x*] and *šivilja* [en. *seamstress*] the first result of the analogy *moški:avtomehanik :: ženska:x* [en. *man:auto mechanic :: woman:x*] in the Slovene model "fastText Embeddia 100D lem avg"). One explanation is that certain word embeddings are more "central" than the others and, therefore, the closest neighbour of many other words. To check if this explanation is true, instead of the cosine similarity measure, we used the CSLS measure (Conneau et al., 2018) that considers the shared distances of *N* closest neighbours. We observed that the precision is worse when using the CSLS measure than the cosine similarity (Section 5), and therefore we do not report these results. However, when observing the most common words, returned as the analogy task results (Table 6 and Table 7), the distribution of the most common words is more uniform when using the CSLS measure.

Direct comparison of models between Croatian and Slovene is not possible, as the embeddings are trained on different text corpora, and the professions used for analogy calculations are not the same. However, we can notice that in Croatian the occupational gender bias in tested embeddings is slightly higher. Interestingly, the statistical data shows that the employment gap and the pay gap between women and men are lower in Slovenia compared to Croatia (Eurostat, 2021). In future, it would be interesting to study if the female employment rate and gap, as well as the gap in salaries for the same professions between countries,

is correlated with the gender bias in embeddings models trained on the corresponding national languages and the changes of this correlation through time.

7 CONCLUSIONS AND FURTHER WORK

We evaluated different Slovene and Croatian word embeddings on analogies of male and female occupations (using different configurations and approaches to calculate analogies). Our focus is on the quantitative evaluation, and the results may be informative for developers of NLP tools. The lowest gender bias was obtained using the fastText embeddings. In finding female analogies (male occupation on the input), the best performing models proved to be fastText CLARIN.SI-embed.sl and fastText CLARIN.SI-embed.hr for Slovene and Croatian, respectively, while the best performing models for finding male analogies (female occupation on the input) were the respective fastText Embeddia models. The approach where averages of several inherently male and female words were used instead of using only the embeddings for woman or man improved the results. Lemmatization likewise improves the precision. With female occupations at the input, the best results (P@10) of 0.982 and 0.986 are achieved using the "fastText Embeddia 300D lem avg" models for Slovene and Croatian, respectively (the examples where the embeddings do not cover the input occupation were dismissed). With male occupations on the input, the best results of 0.902 and 0.754 are produced by the "fastText CLARIN.SI-embed.sl 100D lem avg" and "fastText CLARIN.SI-embed.hr 100D (lem) avg" (cases where the input occupation is not present among the embeddings were likewise dismissed). Lowest results for male input reflect lower coverage of female occupation equivalents in the embeddings model. The "fastText CLARIN.SI-embed.sl" and "fastText CLARIN.si-embedd.hr (lemma)" models contain the highest ratio of searched-for female and male occupations. The qualitative analysis identifies the word2vec Kontekst.io model as the model with the highest degree of gender bias in the results (stereotypically male/female occupations appearing among the results regardless of the grammatical gender of the input occupation).

In future work, we will focus on a detailed qualitative analysis and the relationship between word embeddings, language, and social power. Moreover, we will align occupations in Slovene and Croatian. Further work will also encompass an evaluation of BERT contextual embeddings and experiments in

other languages. The impact of the gender bias will be tested in predictive models on practical tasks such as the sentiment analysis.

Acknowledgments

The research was supported by the Slovene Research Agency through research core funding no. P6-0411 and P2-103, as well as project no. J6-2581. This paper is supported by European Union's Horizon 2020 Programme project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media, grant no. 825153). The results of this paper reflect only the author's view and the Commission is not responsible for any use that may be made of the information it contains.

REFERENCES

Argamon, S., Koppel, M., Fine, J., & Shimoni, A. R. (2003). Gender, genre, and writing style in formal written texts. *TEXT*, *23*, 321–346.

Baker, P. (2010). Will Ms ever be as frequent as Mr? A corpus-based comparison of gendered terms across four diachronic corpora of British English. *Gender & Language*, *4*(1), 125–149.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS'16)* (pp. 4356–4364).

Bordia, S., & Bowman, S. (2019). Identifying and Reducing Gender Bias in Word-Level Language Models. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop,* (pp. 7–15).

Brunet, M. E., Alkalay-Houlihan, C., Anderson, A., & Zemel, R. S. (2019). Understanding the Origins of Bias in Word Embeddings. *Proceedings of International Conference on Machine Learning (ICML 2019).*

Caldas-Coulhard, C. R., & Moon, R. (2010). 'Curvy, hunky, kinky': Using corpora as tools for critical analysis. *Discourse & Society*, *21*(2), 99–133.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora necessarily contain human biases. *Science*, *356*(6334), 183–186.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., & Jegou, H. (2018). Word translation without parallel data. *Proceedings of the International Conference on Learning Representation (ICLR)*.

Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T., Romih, T., Arhar Holdt, Š., Čibej, J., Krsnik L., & Robnik-Šikonja, M. (2019). Morphological lexicon Sloleks 2.0. CLARIN.SI. http://hdl.handle.net/11356/1230

Eurostat (2021). Gender statistics. Retrieved from https://ec.europa.eu/eurostat/statistics-explained/index.php/Gender_statistics#Labour_market

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *PNAS*, *115*(16).

Garimella, A., Banea, C., Hovy, D., & Mihalcea, R. (2019). Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. *Proceedings of the 57th Annual Meeting of the ACL* (pp. 3493–3498).

Gigafida 2.0. Retrieved from https://viri.cjvt.si/gigafida

Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *Proceedings of NAACL-HLT 2019* (pp. 609–614).

Gorjanc, V. (2007). Kontekstualizacija oseb ženskega in moškega spola v slovenskih tiskanih medijih. In I. Novak-Popov (Ed.), *Stereotipi v slovenskem jeziku, literaturi in kulturi: zbornik predavanj 43. seminarja slovenskega jezika, literature in culture* (pp. 173–180). Ljubljana: Center za slovenščino kot drugi/tuji jezik.

Hill, B., & Shaw, A. (2013). The Wikipedia gender gap revisited: Characterising survey response bias with propensity score estimation. *PloS One*, 8.

Hirasawa, T., & Komachi, M. (2019). Debiasing Word Embeddings Improves Multimodal Machine Translation. *Proceedings of Machine Translation Summit XVII, Vol. 1* (pp. 32–42).

Hovy, D., & Søgaard, A. (2015). Tagging performance correlates with author age. *Proceedings of the 53rd Annual Meeting of the ACL and the 7th IJCNLP* (pp. 483–488).

Hovy, D. (2015). Demographic factors improve classification performance. *Proceedings of the 53rd Annual Meeting of the ACL and the 7th IJCNLP* (pp. 752–762).

Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., & Denuyl, S. (2020). Social Biases in NLP Models as Barriers for Persons with Disabilities. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5491–5501).

Kern, B., & Dobrovoljc, H. (2017). Pisanje moških in ženskih oblik in uporaba podčrtaja za izražanje »spolne nebinarnosti«. Jezikovna svetovalnica. Retrieved from https://svetovalnica.zrc-sazu.si/topic/2247/pisanje-mo%C5%A1kih-in-%C5%BEenskih-oblik-in-uporaba-pod%C4%8Drtaja-za-izra%C5%BEanje-spolne-nebinarnosti

Kiritchenko, S., & Mohammad, S., (2018). Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics* (pp. 43–53).

Koolen, C., & van Cranenburgh, A. (2017). These are not the stereotypes you are looking for: Bias and fairness in authorial gender attribution. *Proceedings of the First Ethics in NLP workshop* (pp. 12–22).

Lakoff, R. (1973). Language and woman's place. *Language in Society*, *2*(1), 45–80.

Liang, P. P, Li, I. M., Zheng, E., Lim, Y. C., Salakhutdinov, R., & Morency, L. (2020). Towards Debiasing Sentence Representations. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5502–5515).

Ljubešić, N., & Erjavec, T. (2018). Word embeddings CLARIN.SI-embed.sl 1.0. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1204

Ljubešić, N. (2018). Word embeddings CLARIN.SI-embed.hr 1.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1205

Martinc, M., Škrjanec, I., Zupan, K., & Pollak, S. (2017). PAN 2017: Author profiling - gender and language variety prediction: notebook for PAN at CLEF 2017. *Proceedings of the Conference and Labs of the Evaluation Forum*.

Mikolov, T., Corrado, G. S., Chen, K., & Dean, J. (2013a). Efficient estimation of word representations in vector space. *Proceedings of the International Conference on Learning Representations* (pp. 1–12).

Mikolov, T., Yih, W-t., & Zweig, G. (2013b). Linguistic regularities in continuous space word representations. *Proceedings of the 2013 Conference of the North American Chapter of the ACL: Human Language Technologies* (pp. 746–751).

Nozza, D., Volpetti, C., & Fersini, E. (2019). Unintended Bias in Misogyny Detection. *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence* (pp. 149–155).

Nissim, M., van Noord, R., & van der Goot, R. (2019). Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, *46*(3), 487–497.

Pearce, M. (2008). Investigating the collocational behaviour of man and woman in the BNC using Sketch Engine. *Corpora*, *3*(1), 1–29.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualised word representations. *Proceedings of NAACL-HLT 2018* (pp. 2227–2237).

Plahuta, M. (2020). O slovarju. Retrieved from https://kontekst.io/oslovarju

Popič, D., & Gorjanc, V. (2018). Challenges of adopting gender-inclusive language in Slovene. *Suvremena lingvistika, 44*(86), 329–350.

Prates, M. O. R., Avelar, P. H., & Lamb, L. C. (2020). Assessing gender bias in machine translation: A case study with Google Translate. *Neural Computing and Applications*, *32*, 6363–6381.

Rangel, F., Celli, F., Rosso, P., Potthast, M., Stein, B., & Daelemans, W. (2015). Overview of the 3rd author profiling task at PAN 2015. In L. Cappellato, N. Ferro, G. J. F. Jones in E. SanJuan (Eds.), *CLEF 2015 Labs and Workshops, Notebook Papers*.

Schick, T., Udupa, S., & Schütze, H. (2021). Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. *arXiv preprint arXiv:2103.00453*.

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K-W., & Wang, W. Y. (2019). Mitigating gender bias in

natural language processing: Literature review. *Proceedings of the 57th Annual Meeting of the ACL* (pp. 1630–1640).

Supej, A., Plahuta, M., Purver, M., Mathioudakis, M., & Pollak, S. (2019). Gender, language, and society: Word embeddings as a reflection of social inequalities in linguistic corpora. *Proceedings of the Slovensko sociološko srečanje 2019 – Znanost in družbe prihodnosti* (pp. 75–83).

Supej, A., Ulčar, M., Robnik-Šikonja, M., & Pollak, S. (2020). Primerjava slovenskih besednih vektorskih vložitev z vidika spola na analogijah poklicev. *Proceedings of the Conference on Language Technologies & Digital Humanities 2020* (pp. 93–100).

Svoboda, L., & Beliga, S. (2018). Evaluation of Croatian Word Embeddings. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (pp. 1512–1518).

Škrjanec, I., Lavrač, N., & Pollak, S. (2018). Napovedovanje spola slovenskih blogerk in blogerjev. In D. Fišer (Ed.), *Viri, orodja in metode za analizo spletne slovenščine* (pp. 356–373). Ljubljana: Znanstvena založba FF.

Tannen, D. (1990). *You Just Don't Understand: Women and Men in Conversation*. New York: Ballantine Books.

Ulčar, M. (2019). ELMo embeddings model, Slovenian. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1257

Vanmassenhove, E., Hardmeier, C., & Way, A. (2018). Getting gender right in neural machine translation. *Proceedings of the EMNLP* (pp. 3003–3008).

Verhoeven, B., Škrjanec, I., & Pollak, S. (2017). Gender profiling for Slovene Twitter communication: The influence of gender marking, content and style. *Proceedings of the 6th BSNLP Workshop* (pp. 119–125).

Vlada RS (1997). 1641. uredba o uvedbi in uporabi standardne klasifikacije poklicev. *Uradni list RS*, *28*, 2217. Retrieved from https://www.uradni-list.si/glasilo-uradni-listrs/vsebina?urlid=199728&stevilka=1641

Volkova, S., Wilson, T., & Yarowsky, D. (2013). Exploring demographic language variations to improve multilingual sentiment analysis in social media. *Proceedings of the EMNLP* (pp. 1815–1827).

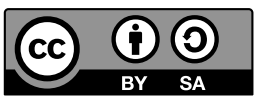
Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K-W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *Proceedings of the EMNLP* (pp. 2979–2989).

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K-W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. *Proceedings of the NAACL-HLT* (pp. 15–20).

PRIMERJAVA SLOVENSKIH IN HRVAŠKIH BESEDNIH VEKTORSKIH VLOŽITEV Z VIDIKA SPOLA NA ANALOGIJAH POKLICEV

V zadnjih letih je uporaba globokih nevronskih mrež in gostih vektorskih vložitev za predstavitve besedil privedla do vrste odličnih rezultatov na področju računalniškega razumevanja naravnega jezika. Prav tako se je pokazalo, da vektorske vložitve besed pogosto zajemajo pristranosti z vidika spola, rase ipd. Prispevek se osredotoča na evalvacijo vektorskih vložitev besed v slovenščini in hrvaščini z vidika spola z uporabo besednih analogij. Sestavili smo seznam moških in ženskih samostalnikov za poklice v slovenščini in ovrednotili spolno pristranost modelov vložitev fastText, word2vec in ELMo z različnimi konfiguracijami in pristopi k računanju analogij. Izkazalo se je, da najmanjšo poklicno spolno pristranost vsebujejo vložitve fastText. Tudi za hrvaško evalvacijo smo uporabili sezname poklicev in primerjali različne fastText vložitve.

Ključne besede: besedne vložitve, spolna pristranost, besedne analogije, poklici, obdelava naravnega jezika



APPENDIX 1

We present the results, comparing different approaches described in Section 4 and Section 5. The approach where we lemmatised all the words has the suffix *lem* appended in the tables. The approach where we used the average difference of vectors of pairs of words from Table 1 has the suffix *avg* appended in the tables. The results for Slovene word embeddings are shown in Table 8, the results for Croatian word embeddings in Table 9 and the share of cases, where the input occupation is the result of the analogy task, in Table 10.

Table 8: *Results for Slovenian embeddings*

Slovene word embeddings	dimensions and approach	*f* input			*m* input		
		P@1	P@5	P@10	P@1	P@5	P@10
ELMo Embeddia	1024D l0 avg	0.707	0.933	0.947	0.166	0.359	0.387
	1024D l0	0.427	0.920	0.947	0.210	0.376	0.398
	1024D l0 lem avg	0.907	0.933	0.947	0.370	0.398	0.403
	1024D l0 lem	0.893	0.947	0.947	0.376	0.392	0.403
	1024D l1 avg	0.907	0.947	0.947	0.381	0.392	0.398
	1024D l1	0.880	0.947	0.947	0.376	0.392	0.392
	1024D l1 lem avg	0.907	0.947	0.947	0.381	0.392	0.398
	1024D l1 lem	0.907	0.947	0.947	0.376	0.392	0.392
	1024D l2 avg	0.880	0.933	0.933	0.376	0.398	0.398
	1024D l2	0.853	0.920	0.933	0.370	0.398	0.398
	1024D l2 lem avg	0.880	0.933	0.933	0.376	0.398	0.398
	1024D l2 lem	0.853	0.920	0.933	0.370	0.398	0.398
fastText.cc	300D avg	0.393	0.798	0.913	0.607	0.738	0.751
	300D	0.150	0.561	0.792	0.445	0.703	0.734
	300D lem avg	0.613	0.884	0.948	0.655	0.755	0.764
	300D lem	0.457	0.861	0.919	0.498	0.725	0.751
fastText Embeddia	100D avg	0.900	0.971	0.976	0.672	0.716	0.720
	100D	0.471	0.871	0.906	0.638	0.716	0.720
	100D lem avg	0.906	0.971	0.976	0.677	0.720	0.724
	100D lem	0.735	0.924	0.941	0.638	0.716	0.720
	300D avg	0.835	0.971	0.976	0.668	0.716	0.724
	300D	0.329	0.859	0.959	0.685	0.720	0.720
	300D lem avg	**0.947**	**0.976**	**0.982**	0.685	0.720	0.724
	300D lem	0.818	0.971	0.976	0.685	0.720	0.720

Slovene word embeddings	dimensions and approach	*f* input			*m* input		
		P@1	P@5	P@10	P@1	P@5	P@10
fastText CLARIN.SI-embed.sl	100D avg	0.784	0.913	0.940	**0.761**	0.868	0.880
	100D	0.083	0.587	0.780	0.705	0.855	0.885
	100D lem avg	0.839	0.940	0.950	**0.761**	**0.880**	**0.902**
	100D lem	0.651	0.881	0.917	0.709	0.859	0.885
fastText Sketch Engine (word)	100D avg	0.886	0.962	0.973	0.717	0.768	0.777
	100D	0.211	0.757	0.908	0.691	0.768	0.777
	100D lem avg	0.930	0.962	0.973	0.725	0.781	0.785
	100D lem	0.811	0.951	0.962	0.691	0.768	0.781
fastText Sketch Engine (lemma)	100D avg	0.673	0.931	0.960	0.598	0.786	0.821
	100D	0.510	0.812	0.891	0.380	0.658	0.756
word2vec Kontekst.io	256D avg	0.679	0.853	0.872	0.407	0.550	0.593
	256D	0.365	0.590	0.718	0.251	0.489	0.515
	256D lem avg	0.679	0.853	0.872	0.407	0.550	0.593
	256D lem	0.513	0.686	0.795	0.251	0.489	0.519

Note. For each approach, where we have a feminine word for occupation on the input (*f* input) and we search for the equivalent masculine term, and where we have a masculine word for occupation on the input (*m* input) and we search for the equivalent feminine term. The examples where the embeddings do not cover the input occupation were dismissed. The best result in each column is in bold.

Table 9: *Results for Croatian embeddings*

Croatian word embeddings	dimensions and approach	*f* input			*m* input		
		P@1	P@5	P@10	P@1	P@5	P@10
fastText.cc	300D avg	0.604	0.883	0.944	0.536	0.603	0.609
	300D	0.452	0.838	0.914	0.429	0.599	0.606
	300D lem avg	0.731	0.939	0.954	0.546	0.637	0.644
	300D lem	0.660	0.924	0.954	0.508	0.618	0.634
fastText Embeddia	100D avg	0.896	0.941	0.959	0.625	0.669	0.672
	100D	0.797	0.928	0.937	0.459	0.634	0.656
	100D lem avg	0.905	0.941	0.968	0.625	0.666	0.672
	100D lem	0.833	0.932	0.941	0.503	0.641	0.662
	300D avg	0.829	0.937	0.973	0.616	0.675	0.675
	300D	0.703	0.914	0.950	0.431	0.662	0.672
	300D lem avg	**0.923**	**0.982**	**0.986**	0.631	0.675	0.678
	300D lem	0.865	0.950	0.964	0.578	0.672	0.675

Croatian word embeddings	dimensions and approach	*f* input			*m* input		
		P@1	P@5	P@10	P@1	P@5	P@10
fastText CLARIN.SI-embed.hr (word)	100D avg	0.896	0.933	0.941	0.670	**0.749**	**0.754**
	100D	0.778	0.904	0.919	0.491	0.699	0.740
	100D lem avg	0.907	0.930	0.944	**0.673**	0.746	**0.754**
	100D lem	0.815	0.904	0.915	0.550	0.711	0.746
fastText CLARIN.SI-embed.hr (lemma)	100D avg	0.244	0.678	0.826	0.266	0.521	0.588
	100D	0.278	0.593	0.693	0.126	0.336	0.406

Note. For each approach, where we have a feminine word for occupation on the input (*f* input) and we search for the equivalent masculine term, and where we have a masculine word for occupation on the input (*m* input) and we search for the equivalent feminine term. The examples where the embeddings do not cover the input occupation were dismissed. The best result in each column is in bold.

Table 10: *Share of cases where the result of the analogy with the highest cosine similarity is the input occupation itself - before filtering is done to produce the results of Tables 2 and 3 (both male to female and female to male analogies)*

Slovene word embeddings	Dimensions and approach	Share of outputs equal to inputs	Croatian word embeddings	Dimensions and approach	Share of outputs equal to inputs
ELMo Embeddia	1024D l0 avg	0.547			
	1024D l0	0.547			
	1024D l0 lem avg	0.547			
	1024D l0 lem	0.547			
	1024D l1 avg	0.423			
	1024D l1	0.483			
	1024D l1 lem avg	0.423			
	1024D l1 lem	0.483			
	1024D l2 avg	**0.064**			
	1024D l2	0.088			
	1024D l2 lem avg	**0.064**			
	1024D l2 lem	0.088			
fT fastText.cc	300D avg	0.831	fT fastText.cc	300D avg	0.672
	300D	0.825		300D	0.664
	300D lem avg	0.831		300D lem avg	0.672
	300D lem	0.825		300D lem	0.664

Slovene word embeddings	Dimensions and approach	Share of outputs equal to inputs	Croatian word embeddings	Dimensions and approach	Share of outputs equal to inputs
fT Embeddia	100D avg	0.143	ft Embeddia	100D avg	**0.094**
	100D	0.141		100D	**0.094**
	100D lem avg	0.143		100D lem avg	**0.094**
	100D lem	0.141		100D lem	**0.094**
	300D avg	0.419		300D avg	0.352
	300D	0.513		300D	0.441
	300D lem avg	0.419		300D lem avg	0.352
	300D lem	0.513		300D lem	0.441
fT CLARIN.SI-embed.sl (word)	100D avg	0.316	fT CLARIN.SI-embed.hr (word)	100D avg	0.103
	100D	0.310		100D	0.114
	100D lem avg	0.316		100D lem avg	0.103
	100D lem	0.310		100D lem	0.114
fT Sketch Engine (word)	100D avg	0.096			
	100D	0.135			
	100D lem avg	0.096			
	100D lem	0.135			
fT Sketch Engine (lemma)	100D avg	0.803	fT CLARIN. SI-embed.hr (lemma)	100D avg	0.837
	100D	0.927		100D	0.771
w2v Kontekst.io	256D avg	0.483			
	256D	0.718			
	256D lem avg	0.483			
	256D lem	0.718			

Note. The number of all cases is 468 (from 234 occupation pairs) for Slovene and 750 (from 375 occupation pairs) for Croatian.

Slovenščina 2.0, 2021 (1)

AVTOMATSKO RAZPOZNAVANJE SLOVENSKEGA GOVORA ZA DNEVNOINFORMATIVNE ODDAJE

Lucija GRIL, Mirjam SEPESY MAUČEC, Gregor DONAJ, Andrej ŽGANK

Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru

Gril, L., Sepesy Maučec, M., Donaj, G., Žgank, A. (2021): Avtomatsko razpoznavanje slovenskega govora za dnevnoinformativne oddaje. Slovenščina 2.0, 9(1): 60–89.

DOI: https://doi.org/10.4312/slo2.0.2021.1.60-89

Na področju govornih in jezikovnih tehnologij predstavlja avtomatsko razpoznavanje govora enega izmed ključnih gradnikov. V prispevku bomo predstavili razvoj avtomatskega razpoznavalnika slovenskega govora za domeno dnevnoinformativnih oddaj. Arhitektura sistema je zasnovana na globokih nevronskih mrežah. Pri tem smo ob upoštevanju razpoložljivih govornih virov izvedli modeliranje z različnimi aktivacijskimi funkcijami. V postopku razvoja razpoznavalnika govora smo preverili tudi, kakšen je vpliv izgubnih govornih kodekov na rezultate razpoznavanja govora. Za učenje razpoznavalnika govora smo uporabili bazi UMB BNSI Broadcast News in IETK-TV. Skupni obseg govornih posnetkov je znašal 66 ur. Vzporedno z globokimi nevronskimi mrežami smo povečali slovar razpoznavanja govora, ki je tako znašal 250.000 besed. Na ta način smo znižali delež besed izven slovarja na 1,33 %. Z razpoznavanjem govora na testni množici smo dosegli najboljšo stopnjo napačno razpoznanih besed (WER) 15,17 %. Med procesom vrednotenja rezultatov smo izvedli tudi podrobnejšo analizo napak razpoznavanja govora na osnovi lem in F-razredov, ki v določeni meri pokažejo na zahtevnost slovenskega jezika za takšne scenarije uporabe tehnologije.

Ključne besede: avtomatsko razpoznavanje slovenskega govora, lastnosti slovenskega jezika, dnevnoinformativne oddaje, globoke nevronske mreže, izgubni govorni kodeki.

1 UVOD

V zadnjem desetletju spremljamo izredno hiter razvoj področja umetne inteligence, ki mu botruje predvsem tehnološki napredek na področju velikih podatkov in algoritmov za globoko učenje. To je pripeljalo tudi do izboljšanja metod na področju govornih in jezikovnih tehnologij. Strateški cilji države se lahko tako učinkovito osredotočajo na vključujočo družbo, ki uspešno uporablja tehnologije digitalizacije. Naravna interakcija med človekom in napravami v inteligentnem okolju je eden izmed ključnih vidikov sprejemljivosti tehnologije.

Splošna razširjenost pametnih naprav, kot so mobilni telefoni, je prispevala k povečanju količin različnega zvočnega (in slikovnega) gradiva, ki je na voljo uporabniku. V želji zagotoviti učinkovit dostop do informacij, ki jih vsebuje takšna množica zvočnega gradiva, je neobhodno potrebna uporaba tehnoloških rešitev.

Ena izmed jedrnih tehnologij, ki omogočajo ustrezno podporo za zajemanje informacij, tako iz uporabniškega ali medijskega zvočnega toka kot tudi iz uporabniškega vmesnika naprav v inteligentnem okolju, je avtomatsko razpoznavanje govora (ASR). Deluje lahko v zelo različnih scenarijih, od preprostega ukaznega krmiljenja do zahtevnih sistemov za razpoznavanje spontanega govora več govorcev. S kompleksnostjo scenarija je praviloma obratno sorazmerna uspešnost razpoznavanja govora. Na področju avtomatskega razpoznavanja govora je do pomembnih korakov v razvoju prišlo na točki, ko je bilo možno za to nalogo učinkovito uporabiti globoke nevronske mreže. Te so zamenjale prejšnjo arhitekturo, ki je temeljila na prikritih modelih Markova in zadnja leta ni več prinašala bistvenega napredka. Metode globokega učenja danes predstavljajo privzeto arhitekturo na praktično vseh področjih govornih in jezikovnih tehnologij.

Pomemben vidik predstavlja tudi računska zahtevnost, ki lahko pogosto trči ob vprašanja zagotavljanja zasebnosti govorcev, kadar je v uporabi procesiranje v oblaku. Ta vidik je lahko izrednega pomena, kadar govorimo o tehnologijah za vključujočo družbo, ki pogosto pokrivajo zelo osebne vidike komunikacije uporabnikov.

Področje avtomatskega razpoznavanja govora je neločljivo povezano z razpoložljivostjo govornih virov za posamezni jezik. Tukaj nastopi težava pri jezikih,

za katere obstaja manjši (komercialni) interes za implementacijo ASR. To se lahko še dodatno potencira s posebnostmi določenih jezikov, ki otežijo avtomatsko razpoznavanje govora. V kategorijo za procesiranje zahtevnih jezikov sodi tudi slovenščina. Zanjo je značilna visoka pregibnost besed in relativno prost vrstni red besed v stavku. Obe lastnosti pomembno vplivata na rezultate razpoznavanja govora, saj prvič povečata akustično zamenljivost besed in iskalni prostor razpoznavalnika, drugič pa zmanjšata predikcijsko zmožnost statističnih jezikovnih modelov.

Razvoj prvih sistemov govornih tehnologij za slovenščino se je začel že pred 30 leti, vendar finančno in časovno zahteven razvoj govornih virov v zadnjem desetletju ni uspel slediti intenzivnemu razvoju v svetu. Postopki globokega učenja razpoznavalnikov govora namreč za učinkovito delovanje potrebujejo govorne baze v obsegu več 100 oz. 1000 ur transkribiranih posnetkov. Za področje slovenskega jezika pričakujemo razpoložljivost tako obsežnih govornih virov kot enega od rezultatov projekta Razvoj slovenščine v digitalnem okolju (RSDO, b.d.), ki bo potekal do leta 2023.

Cilj pričujočega raziskovalnega dela je predstaviti razvoj sistema za avtomatsko razpoznavanje slovenskega govora z globokimi nevronskimi mrežami, ki deluje za domeno dnevnoinformativnih oddaj. Takšen avtomatski razpoznavalnik govora je lahko zelo pomembno govornotehnološko orodje za različne scenarije uporabe, kot so na primer avtomatsko indeksiranje govorne vsebine, avtomatsko podnaslavljanje ali avtomatsko prevajanje govora v govor. Za učinkovito doseganje teh ciljev je treba uporabljati razpoznavalnike govora z metodami globokega učenja. Dosedanji sistemi za avtomatsko razpoznavanje slovenskega govora za domeno dnevnoinformativnih oddaj (Žgank in Sepesy Maučec, 2010; Žgank idr., 2014) so temeljili na predhodni arhitekturi prikritih modelov Markova.

V prispevku želimo podati oceno, kakšen primanjkljaj pri prehodu na novo arhitekturo globokih nevronskih mrež predstavljajo omejene govorne baze za slovenski jezik. Pri izgradnji modelov smo se odločili za uporabo različnih aktivacijskih funkcij nevronskih mrež ter na ta način izvedli primerjavo arhitektur. Podoben potek eksperimenta razvoja razpoznavalnika govora so uporabili za španski jezik (Zorrilla idr., 2016), kjer so bile izhodišče obstoječe metode, ki so jih nato preverili na že predhodno uporabljenih govornih

bazah za španski jezik. Hkrati nas v okviru raziskave zanima tudi, kakšen vpliv ima uporaba izgubnih kodekov na rezultate avtomatskega razpoznavanja govora. Izgubni kodeki so postali pomembni že z razmahom različnih internetnih pretočnih storitev. Še posebej velik pomen pa so dobili v času epidemije covida-19, ko se je večina komuniciranja in funkcioniranja družbe preselila v oddaljen način. Podobno primerjavo vpliva izgubnega kodiranja na rezultate razpoznavanja govora sta za drugo domeno in jezik izvedla Pollak in Behunek (2011). V zadnjem delu prispevka bomo izvedli tudi analizo napak razpoznavanja govora in na ta način poskušali ugotoviti vpliv visoke pregibnosti na rezultate razpoznavanja govora. Raziskovalno delo smo zasnovali na slovenski bazi televizijskih dnevnoinformativnih oddaj UMB BNSI Broadcast News (Žgank idr., 2004) in IETK-TV (Žgank idr., 2014), saj ti govorni bazi trenutno še vedno predstavljata najprimernejši vir za takšno analizo, hkrati pa omogočata tudi primerljivost rezultatov s starejšimi sistemi avtomatskega razpoznavanja govora.

V nadaljevanju članka bomo najprej predstavili trenutno stanje na področju razpoznavanja govora za slovenski jezik. V tretjem poglavju bo sledila kratka predstavitev teoretičnega ozadja metod, ki se danes uporabljajo pri gradnji avtomatskih razpoznavalnikov govora. Opisali bomo tudi področje govornih kodekov. V četrtem poglavju bomo predstavili uporabljene govorne in jezikovne vire. Postopek izdelave akustičnih in jezikovnih modelov eksperimentalnega sistema bomo opisali v petem poglavju. Rezultate in analizo vrednotenja razpoznavanja govora bomo predstavili v šestem poglavju. V zadnjem poglavju bomo podali zaključne misli.

2 PREGLED PODROČJA AVTOMATSKEGA RAZPOZNAVANJA GOVORA ZA SLOVENSKI JEZIK

2.1 Govorni viri za slovenski jezik

Že v uvodu smo zapisali, da predstavljajo govorni viri ključno komponento za razvoj avtomatskega razpoznavalnika govora. Pomembno je, da s svojimi značilnostmi in obsegom materiala vplivajo tudi na to, katero arhitekturo nevronskih mrež, ki so danes najbolj aktualna tehnologija pri razvoju razpoznavalnikov, bo možno uspešno naučiti.

Dosedanji razvoj govornih virov za slovenski jezik lahko razdelimo na dve obdobji. V prvem obdobju, ki se je začelo v devetdesetih letih prejšnjega stoletja, je bil poudarek na razvoju govornih baz za omejene scenarije izoliranih ali vezanih besed. Snemalni kanal je bil ali studio ali telefon, obseg govornega material pa praviloma med 10 in 15 ur. V to skupino lahko uvrstimo govorne baze: FDB 1000 Slovenian SpeechDat(II) (Kaiser in Kačič, 1997), Polidat (Žgank idr., 2002), Gopolis (Dobrišek idr., 1998), VNTV/VNRAD (Žibert idr., 2003) in SNABI. Delni sklopi naštetih baz že vsebujejo tudi tekoči govor, vendar je zaradi omejene količine govornega materiala praktičen razvoj splošnega razpoznavalnika govora še nemogoč.

V drugem obdobju razvoja govornih baz za slovenski jezik, ki se je začelo okoli leta 2004, se aktivnosti osredotočijo na tekoči govor. Bistveno se razširi domena vključenega materiala, kot snemalni kanal pa se dodatno pojavi televizija oziroma druge oblike javnega govora, kot so npr. predavanja. Obseg govornih baz se poveča na nekaj 10 ur posnetkov. Sem lahko prištejemo sledeče televizijske baze: UMB BNSI Broadcast News (36 ur) (Žgank idr., 2004), SiBN Broadcast News (36 ur) (Žibert in Mihelič, 2004), IETK-TV (30 ur) in GOS javni podkorpus (42 ur) (Verdonik idr., 2013). Predavanja najdemo v bazi SI TEDx-UM (54 ur, avtomatske transkripcije) (Žgank idr., 2016) in bazi GOS-VideoLectures (22 ur) (Verdonik, 2018). Baza SloParl (Žgank idr., 2006) vsebuje 100 ur posnetkov in magnetogramov parlamentarnih razprav iz DZ RS, baza SOFES (Dobrišek idr., 2017) pa 10 ur posnetkov s poizvedbami po letalskih informacijah.

Dostopnost predstavljenih govornih baz pokriva skoraj celotni spekter možnosti. Nekatere so prosto dostopne preko iniciative Clarin oz. na spletnih straneh avtorjev. Druge baze so dostopne proti plačilu preko organizacije ELRA. Del baz pa je namenjen izključno interni uporabi in tako nedostopen širši raziskovalni skupnosti. Z vidika razvoja področja avtomatskega razpoznavanja govora za slovenski jezik predstavlja takšna razdrobljena dostopnost velik izziv.

Skupna dolžina transkribiranih posnetkov v predstavljenih govornih bazah je približno 250 ur. Dodatnih 150 ur posnetkov je transkribiranih samo avtomatsko ali v obliki magnetogramov. Tudi če bi kljub različnim omejitvam v dostopnosti uspeli združiti vse govorne baze, prihaja med njimi v zasnovi do

tako velikih razlik, da bi bilo učenje razpoznavalnika govora na takšen način neizvedljivo. Ob upoštevanju kriterija sorodnosti in dostopnosti govornih baz je trenutno praktično možno za učenje slovenskega razpoznavalnika govora uporabiti med 50 in 100 urami posnetkov. Takšen obseg učnega materiala je premajhen za uporabo naprednejših arhitektur globokega učenja.

To dejstvo lepo kaže na nujno potrebo po tretjem obdobju v razvoju govornih baz za slovenski jezik, kjer je cilj pridobiti nekaj 100 do 1.000 ur posnetkov, ki so prosto dostopni in omogočajo potencialno kombiniranje virov v prihodnosti. V to kategorijo bo sodila govorna baza, ki nastaja v okviru projekta RSDO.

2.2 Avtomatsko razpoznavanje govora za slovenski jezik

V nadaljevanju bomo podali še kratek pregled ključnih aktivnosti na področju avtomatskega razpoznavanja slovenskega govora. Raziskave so začele potekati okoli leta 1990. Prvi sistemi razpoznavanja govora so delovali za preprostejše scenarije, kot so: krmiljenje preprostih aplikacij (Kačič idr., 1988), klasifikacija fonemov (Mihelič idr., 1992) ali razpoznavanje števk (Imperl idr., 1996). V naslednjem koraku so sledili zahtevnejši scenariji, ki temeljijo na vezanih besedah – dialog za poizvedovanje o letalskih informacijah (Ipšić idr., 1999) ter poizvedovanje o telefonskih številkah (Imperl in Kačič, 1999). Prehod na scenarije razpoznavanja tekočega govora z velikim slovarjem besed (Kaiser idr., 2000) prvič pokaže na izzive, povezane s kompleksnostjo visokopregibnega slovenskega jezika, ter težave zaradi ne dovolj razvitih govornih virov. Delno je to možno izničiti z omejitvijo na ozko domeno, kot so na primer vremenske napovedi (Žibert idr., 2000). Pomembnejši pa je bil korak v smeri razvoja novih govornih virov s področja dnevnoinformativnih oddaj (Žgank idr., 2004; Žibert in Mihalič, 2004), ki so potem služile za razvoj kompleksnejših razpoznavalnikov tekočega govora (Žgank idr., 2006; Dobrišek in Mihelič, 2010; Žgank in Sepesy Maučec, 2010; Žgank idr., 2014).

Prvi slovenski razpoznavalnik govora z globokimi nevronskimi mrežami je bil razvit v okviru večjezičnega razpoznavanja za južnoslovanske jezike (Nouza idr., 2016). V zadnjem desetletju postaja na področju razpoznavanja govora poleg domene dnevnoinformativnih oddaj pomembna tudi domena predavanj. K temu je v veliki meri pripomogel razvoj multimedijske tehnologije in priljubljenost masovnih spletnih predavanj (MOOC). Tako pride tudi

v slovenskem prostoru do izgradnje ustreznih govornih baz s tega področja (Zwitter Vitez idr., 2013; Verdonik idr., 2017). Avtomatski razpoznavalnik govora z globokimi nevronskimi mrežami, ki deluje za to domeno, je predstavil Ulčar s sodelavci (2019) in vključuje sledeče govorne vire: GOS 1.0 (Zwitter Vitez idr., 2013), Gos VideoLectures 2.0 (Verdonik idr., 2017) in Sofes 1.0 (Dobrišek idr., 2017).

3 ARHITEKTURE ZA AVTOMATSKO RAZPOZNAVANJE GOVORA

Na področju arhitekture avtomatskih razpoznavalnikov govora obstajata dve glavni skupini. Prvo predstavljajo sistemi s prikritimi modeli Markova, ki so bili glavni gradnik akustičnega modeliranja v preteklosti. Drugo skupino, ki je danes standardna, pa predstavljajo sistemi na osnovi nevronskih mrež.

3.1 Prikriti modeli Markova

Prikriti modeli Markova predstavljajo metodo statističnega modeliranja, kjer na osnovi vhodnih vektorjev značilk ocenjujemo verjetnost hipoteze izgovorjenega besedila. Običajno se uporabljajo večstanjski levo-desni prikriti modeli, kjer je porazdelitvena funkcija gostote verjetnosti modelirana s skupino uteženih multivariantnih Gaussovih porazdelitvenih funkcij. Z vidika računske kompleksnosti in količine zahtevanega učnega materiala gre praviloma za manj zahtevne sisteme v primerjavi z globokimi nevronskimi mrežami.

3.2 Globoke nevronske mreže

Nevronske mreže predstavljajo metodo na področju strojnega učenja, ki deloma posnema dogajanje v nevronskem sistemu. Mreže so sestavljene iz nevronov, ki so razporejeni v plasti – vhodno plast, notranje plasti in izhodno plast. Kadar je arhitektura nevronske mreže načrtovana tako, da vsebuje dve ali več plasti, govorimo o globoki nevronski mreži. Število globokih plasti, ki jih uporabimo v postopku strojnega učenja, je v veliki meri odvisno od količine učnega gradiva.

Vsak nevron izvaja matematično operacijo, kjer najprej izračuna uteženo vsoto vrednosti na svojih vhodih, nato pa to vsoto uporabi v aktivacijski funkciji, da izračuna izhodno vrednost nevrona. Izhodi nevronov so potem povezani na vhode drugih nevronov.

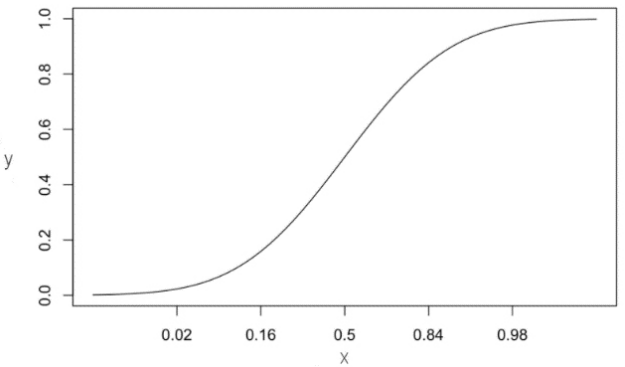
Aktivacijske funkcije so lahko različnih tipov – stopničaste, linearne ali nelinearne. Stopničasta aktivacijska funkcija temelji na pragovni vrednosti (angl. threshold). Če je vhodna vrednost nad ali pod določenim pragom, se nevron aktivira in pošlje naslednji plasti povsem enako vrednost. Linearna funkcija vzame vhodno vrednost nevrona, jo pomnoži z utežjo in generira izhodni signal. Nelinearne aktivacijske funkcije omogočajo kompleksnejše preslikave vhodnih vrednosti v izhodne.

Tanh je hiperbolična tangenta funkcija, ki jo uporabljamo kot aktivacijsko funkcijo pri globokih nevronskih mrežah. Zaloga vrednosti funkcije je med –1 in 1, zaradi česar je povprečje skrite plasti 0 ali blizu te vrednosti. To pomeni, da je učenje na naslednji plasti veliko lažje.

P-norm je nelinearna aktivacijska funkcija, katere izhod se izračuna kot:

$$y = (\sum_i |x_i|^p)^{1/p}, \tag{1}$$

kjer so vektorji x majhna skupina vhodnih vrednosti. Vrednost p je spremenljiva in zanjo je bilo pokazano (Zhang idr., 2014), da s p = 2 pridobimo najboljše rezultate.



Slika 1: Graf aktivacijske funkcije p-norm.

Pri načrtovanju arhitekture globokih nevronskih mrež lahko dodamo ozka podatkovna grla, ki jih bomo v nadaljevanju navajali kar kot ozka grla. Ozko grlo je plast, ki ima manj nevronov kot plast pred ali za njo. Takšne plasti spodbudijo, da se značilke bolje prilagodijo razpoložljivemu prostoru parametrov, ki ga omejimo z velikostjo ozkega grla. Z ozkim grlom tako dosežemo predstavitev vhoda z manjšo dimenzijo.

Prav tako se pri načrtovanju arhitekture uporabljajo razne oblike ansambla. Ideja ansambla je, da namesto enega klasifikatorja zgradimo več klasifikatorjev, ki na koncu glasujejo o končni odločitvi. Učenje poteka na enakih učnih podatkih za vsako iteracijo. Po vsaki iteraciji se doda vrednost, ki je zmnožek vrednosti β in križne entropije izhoda trenutne mreže ter geometrijsko povprečnih zadnjih vrednosti izhoda ansambla mrež. Vrednost β eksponentno narašča glede na začetno in končno vrednost β, ki jo izberemo.

V zadnjih letih so nevronske mreže postale popularne na raznih področjih strojnega učenja, tudi pri razpoznavanju govora (Nassif idr., 2019). Ker pa gre pri razpoznavanju govora za razpoznavanje časovne vrste, vse arhitekture nevronskih mrež niso primerne.

Med korakom učenja se nevronska mreža prilagaja na učne podatke tako, da spreminja uteži. Pri uporabi pa nato dajemo nove podatke na vhodno plast omrežja ter opazujemo rezultate na izhodni plasti. V nadaljevanju bomo preizkusili, kako dobro delujejo glede na našo učno množico različne nelinearne aktivacijske funkcije, ki so uporabljene pri izgradnji razpoznavalnikov govora. Pogosto uporabljeni sta p-norm in tanh, ki smo ju kombinirali še z ansamblom in ozkim grlom, saj smo želeli preveriti, ali bodo dodatni koraki doprinesli k izboljšanju rezultatov.

3.3 Zvočni kodeki

Za stiskanje podatkov uporabljamo kodiranje, ki nam omogoča, da lahko informacijo zapišemo z manj biti kakor na začetku. Pri zapisu zvoka lahko na takšen način zmanjšamo pasovno širino in velikost stisnjene zvočne datoteke.

Kodiranje je lahko brezizgubno ali izgubno. Brezizgubni zvočni kodeki zmanjšajo obseg podatkov, vendar ohranijo vso informacijo, ki jo lahko ponovno pridobimo po dekodiranju. Pri izgubnih kodekih se odstranjujejo informacije v časovnem in/ali frekvenčnem prostoru, ki jih človek ne more zaznati zaradi psihoakustičnih značilnosti slušne zaznave. Z uporabo izgubnih kodekov se zmanjša bitna ločljivost zvoka, zaradi česar po dekodiranju nikoli ne pridobimo prvotne informacije v celoti. Vpliv popačenj izgubnih kodekov želimo ohraniti na tako nizki ravni, da ne vplivajo bistveno na subjektivno zaznavo kakovosti zvoka.

Izgubni zvočni kodeki so pomembno pridobili na veljavi z razmahom internetnih storitev, še posebej v obliki pretočnega dostopa do vsebin in različnih oblik dela na daljavo v času epidemije covida-19. Posledično moramo upoštevati njihov vpliv tudi na področju avtomatskega razpoznavanja govora.

4 UPORABLJENI GOVORNI IN JEZIKOVNI VIRI

Osrednji vir podatkov, ki smo jih uporabili za akustično modeliranje, je predstavljala govorna baza UMB BNSI Broadcast News (Žgank idr., 2004), ki jo distribuira organizacija ELRA (2015). Govorna baza vsebuje posnetke dnevnoinformativnih televizijskih oddaj RTV Slovenija v obsegu 36 ur. Od tega je 30 ur namenjenih učenju akustičnih modelov. Oddaje so nastale v letih 1999–2003, tako da je bila z vidika naprav, uporabljenih v produkciji, tehnologija delno drugačna, kot jo srečamo danes (npr.: snemalne naprave z izgubnimi kodeki, povezave VoIP, spletne komunikacijske platforme). V bazi je skupaj 1.565 govorcev, od tega 1.069 moških in 477 žensk. Za 19 govorcev spola ni bilo možno nedvoumno določiti.

Posnetki so bili ročno segmentirani in transkribirani. Hkrati je bilo označeno tudi akustično ozadje in negovorni akustični dogodki. To je posledica produkcije oddaj, saj je pogosto v ozadje zvočnega posnetka glavnega govorca montiran zvočni posnetek iz videa ali pa drugo zvočno ozadje, kot je na primer glasba. Pri avtomatskem razpoznavanju govora je pomemben vidik tudi, ali gre za bran, načrtovan ali spontan govor, saj ta značilnost pomembno vpliva na dosežene rezultate.

V predhodnem odstavku naštete parametre v domeni razpoznavanja govora televizijskih oddaj karakterizirajo F-razredi (Schwartz idr., 1997). Ti so definirani na sledeč način:

- *F0: bran govor v studijskem okolju,*
- *F1: spontan govor v studijskem okolju,*
- *F2: bran/spontan govor preko telefona,*
- *F3: bran/spontan govor z glasbo v ozadju,*
- *F4: bran/spontan govor z drugim zvočnim ozadjem,*
- *F5: govorci, katerih materni jezik ni slovenščina,*
- *FX: preostalo.*

Predstavljene F-razrede bomo uporabili pri podrobnejši analizi rezultatov v šestem poglavju, saj bodo služili za oceno težavnosti testnega scenarija. Pomembno namreč odražajo akustično ozadje in s tem nakazujejo na potencialni vpliv degradacij na rezultat razpoznavanja govora. F-razredi so v govorni bazi zastopani v različnih deležih. Ker predstavlja testni nabor v dolžini 3 ur manj kot eno desetino baze, se to odraža tudi v zastopanosti F-razredov. Tako v testni množici v celoti manjka razred F5 z govorci, katerih materni jezik ni slovenščina. Po obsegu pa je najmanjši razred F2, ki vsebuje govor, posnet preko telefona. Ta kategorija vsebuje samo osem segmentov treh govorcev, ki skupaj izgovorijo nekaj več kot 100 besed.

Nabor učne množice za akustično modeliranje avtomatskega razpoznavalnika govora smo razširili še z govorno bazo IETK-TV, ki pa zaradi omejitev avtorskih pravic ni širše dostopna. Ta baza predstavlja nadgradnjo baze UMB BNSI Broadcast News in je nastala na osnovi istih specifikacij. Obsega 29 ur transkribiranih posnetkov 784 govorcev, ki so v celoti namenjeni akustičnemu modeliranju. Nabor različnih televizijskih oddaj je v bazi IETK-TV razširjen v primerjavi z bazo UMB BNSI, saj so vključeni tudi intervjuji in okrogle mize. Posledično je delež spontanega govora v bazi IETK-TV več kot enkrat večji kot v bazi UMB BNSI Broadcast News.

Za gradnjo jezikovnega modela učnega korpusa nismo razširjali. Uporabili smo sledeče korpuse: BNSI-Speech (573 tisoč besed), BNSI-Text (11 milijonov besed) in FidaPLUS (621 milijonov besed) (Arhar in Gorjanc, 2007). Korpus Večer smo iz učenja izločili, saj so njegovi članki vsebovani v korpusu FidaPLUS.

5 EKSPERIMENTALNI SISTEM

Osnovna zasnova eksperimentalnega sistema za avtomatsko razpoznavanje govora, uporabljena v teh eksperimentih, je enaka za pristopa HMM in DNN. Zajet govorni signal je najprej treba predprocesirati in pretvoriti v vektorje značilk. Nato lahko izvedemo razpoznavanje govora, kjer uporabimo akustične in jezikovne modele ter fonetični slovar. Akustične modele smo s pristopi strojnega učenja predhodno naučili na transkribirani učni govorni bazi, jezikovne modele pa na učnem besedilnem korpusu.

5.1 Akustično modeliranje

Za izgradnjo avtomatskega razpoznavalnika govora smo uporabili odprtokodno orodje Kaldi (Povey idr., 2011), ki omogoča izgradnjo sistema z metodami globokega učenja.

Za začetek učenja akustičnih modelov potrebujemo transkribirane posnetke v formatu WAV. Za učno kot tudi testno množico je treba pripraviti vse spremljajoče datoteke. Za učni postopek smo kot osnovo vzeli Kaldijev postopek učenja z bazo Mini LibriSpeech, ki smo ga ustrezno nadgradili. Uporabljeni postopek učenja je po dosedanjih izkušnjah dajal dobre rezultate, hkrati sta velikosti obeh baz primerljivi.

V naslednjem koraku s pomočjo že pripravljenih skript v orodju Kaldi pripravimo še ostale datoteke, ki so potrebne za učenje akustičnih modelov. Izvorni signal oknimo in nato tvorimo značilke v obliki mel-frekvenčnih kepstralnih koeficientov (MFCC). Posamezni vektor značilk je imel 13 elementov, ki smo jim dodali še prvi in drugi odvod. Sledil je postopek akustičnega modeliranja, kjer zaporedoma izvajamo učenje modelov in njihove poravnave pred ponovnim učenjem novega modela. V primeru orodja Kaldi gre za hibridno metodo, kjer v prvem koraku uči prikrite modele Markova, v drugem koraku pa globoko nevronsko mrežo. Kot osnovno enoto za akustično modeliranje smo uporabili slovenske grafeme.

Prikriti modeli Markova, uporabljeni v akustičnem modeliranju, imajo tristanjsko levo-desno topologijo. Izgradnja akustičnih modelov poteka postopoma, kjer se koraki učenja parametrov modela z Baum-Welchovo reestimacijo izmenjujejo s koraki prisilne poravnave izboljšanih različic učnih transkripcij. Za monofonske akustične modele smo uporabili 40 iteracij, za kontekstno odvisne trifonske modele pa 35 iteracij učenja.

Sledilo je učenje globokih nevronskih mrež. Pri tem smo kot arhitekturo uporabili navadno usmerjeno globoko nevronsko mrežo. V okviru akustičnega modeliranja smo uporabili različne aktivacijske funkcije. Tako smo preverili možen vpliv arhitekture nevronskih mrež na avtomatsko razpoznavanje slovenskega govora. Prva je bila aktivacijska funkcija p-norm (Zhang idr., 2014). Vrednost parametra p smo nastavili na 2, saj je bilo v preteklosti pokazano (Zhang idr., 2014), da lahko pri tej vrednosti pričakujemo najboljše rezultate.

Učenje nevronske mreže je potekalo v 15 regularnih epohah in 5 dodatnih, kar sta prevzeta parametra za takšen potek. Inicialno stopnjo učenja smo nastavili na 0,02 in končno stopnjo učenja na 0,004. Vhodno število nevronov smo nastavili na 2000 in izhodno število na 400. Nastavljene vrednosti parametrov ustrezajo predlaganim v okolju Kaldi za razpoložljivo količino učnega materiala. V eksperimentu smo implementirali 2 skriti plasti in 4 skrite plasti, saj smo na takšen način prilagajali arhitekturo glede na velikost učnega seta.

Aktivacijsko funkcijo p-norm smo v naslednjem poskusu združili z uporabo ozkega grla. Pri tej kombinaciji s pomočjo nelinearnih vrednosti ustvarjamo značilke ozkega grla. Dimenzijo ozkega grla smo nastavili na 42. Vrednost *p* smo ohranili na 2. Prav tako smo ohranili število epoh in stopnji učenja. Implementirali smo 4 skrite plasti, saj se je iz predhodnega preizkusa izkazalo, da se je model z dvema skritima plastema slabše izkazal. Preizkusili pa smo tudi arhitekturo z nekoliko manj nevroni, in sicer smo vhodno število nevronov nastavili na 1000 in izhodno število nevronov na 200. Tudi v tem primeru smo uporabili 4 skrite plasti.

P-norm smo kombinirali tudi z metodo ansambla. Parametre za p-norm smo za število epoh in vrednost p smo nastavili enako kot v prejšnjih dveh primerih. Prav tako smo tudi tukaj uporabili arhitekturo s 4 skritimi plastmi. Število vhodnih in izhodnih nevronov smo prilagajali tako kot v prejšnjem primeru. V prvem primeru smo uporabili 1000 vhodnih in 200 izhodnih, v drugem pa 2000 vhodnih in 400 izhodnih. Dodali smo parameter velikosti ansambla, ki smo ga nastavili na 4, ter inicialno in končno vrednost β. Inicialno vrednost β smo nastavili na 0,1, končno pa na 5. Te vrednosti so bile nastavljene glede na izhodiščne parametre v okolju Kaldi.

V naslednjem poskusu smo uporabili aktivacijsko funkcijo tanh. Pri tej arhitekturi smo uporabili 20 regularnih epoh in 5 dodatnih. Arhitektura vsebuje dve skriti plasti s 375 nevroni. Tukaj smo enako kot pri aktivacijski funkciji p-norm nastavili inicialno stopnjo učenja na 0,02 in končno stopnjo učenja na 0,004. Tako smo sledili primerljivosti arhitektur.

V kombinaciji aktivacijske funkcije tanh in ozkega grla smo se odločili, da uporabimo enake parametre kot pri globoki nevronski mreži z aktivacijsko funkcijo tanh. Dimenzijo ozkega grla smo nastavili na 42.

Predstavljeni parametri so v veliki meri odvisni tako od količine učnega materiala kot tudi od njegove raznolikosti. Posledično jih je treba ustrezno prilagoditi za vsak govorni vir. Parametre, ki jih nismo vključili v primerjavo, smo nastavili empirično oziroma s pomočjo informacij o sistemih drugih avtorjev. Cilj je doseči dobre rezultate razpoznavanja govora, hkrati pa ohraniti zmožnost posplošitve na nove testne vzorce. V nasprotnem primeru dosežemo prekomerno prileganje globoke nevronske mreže. V takšnem primeru sicer lahko dosežemo izvrsten rezultat razpoznavanja govora na zelo sorodnem testnem gradivu. Kakor hitro pa je testno gradivo raznolikejše, pride do drastičnega poslabšanja rezultatov razpoznavanja govora. Zato je takšno prekomerno prileganje učinek, ki se mu želimo izogniti. Omejena količina razpoložljivega učnega govornega materiala je bila tudi razlog, da nismo uporabili kompleksnejših metod globokega učenja, kot so na primer »end-to-end« globoke nevronske mreže.

5.2 Jezikovno modeliranje

V eksperimentih smo uporabili dva slovarja, prvi je vseboval 64.000 besed, drugi pa 250.000. Pripadajoča slovarja izgovorjav smo tvorili na osnovi grafemskih akustičnih enot, ki smo jim dodali model tišine in pa ločen model različnih negovornih zvokov, ki jih je tvoril govorec. Prvi slovar, ki smo ga naredili z enakim postopkom kot avtorji v prispevkih Žgank in Sepesy Maučec (2010) ter Žgank idr. (2014), obsega 64.000 besed. Vsebuje vse besede korpusov BNSI-Speech in BNSI-Text. Do velikosti 64.000 smo ga dopolnili z najpogostejšimi besedami iz korpusa Večer. Drugi slovar izhaja iz prvega. Do velikosti 250.000 smo ga dopolnili z najpogostejšimi besedami iz korpusa FidaPLUS. Korpus FidaPLUS smo za razširitev slovarja uporabili zato, ker je to obsežen in reprezentativen korpus splošnega slovenskega jezika. Z razširitvijo slovarja smo želeli zmanjšati delež besed izven slovarja (OOV), ki je v primeru prvega slovarja znašal 4,22 %, drugega pa 1,33 %. Ker oba slovarja vsebujeta besede iz korpusa BNSI-Speech, so med običajnimi besedami tudi različna mašila in onomatopeje, ki smo jih modelirali kar na osnovi njihove zvočne pojave, in ne kot posebne, ločene, akustične modele.

Z orodjem SRI Language Modeling Toolkit (Stolcke, 2002) smo zgradili trigramske modele s prvim slovarjem. Uporabili smo enak potek kot avtorji

v Žgank in Sepesy Maučec (2010) ter Žgank idr. (2014). Tudi z drugim slovarjem smo zgradili interpoliran trigramski model. V vseh treh komponentah smo uporabili Good-Turingovo glajenje in sestopanje po Katzu. V komponenti BNSI-text smo izločili trigrame s frekvenco 1, v komponenti FidaPLUS pa bigrame s frekvenco 1 in trigrame s frekvencama 1 in 2. Na ta način smo dobili trigramski model, ki je bil primerljive velikosti kot trigramski model s prvim slovarjem. Perpleksnost modela na testni množici je bila 284.

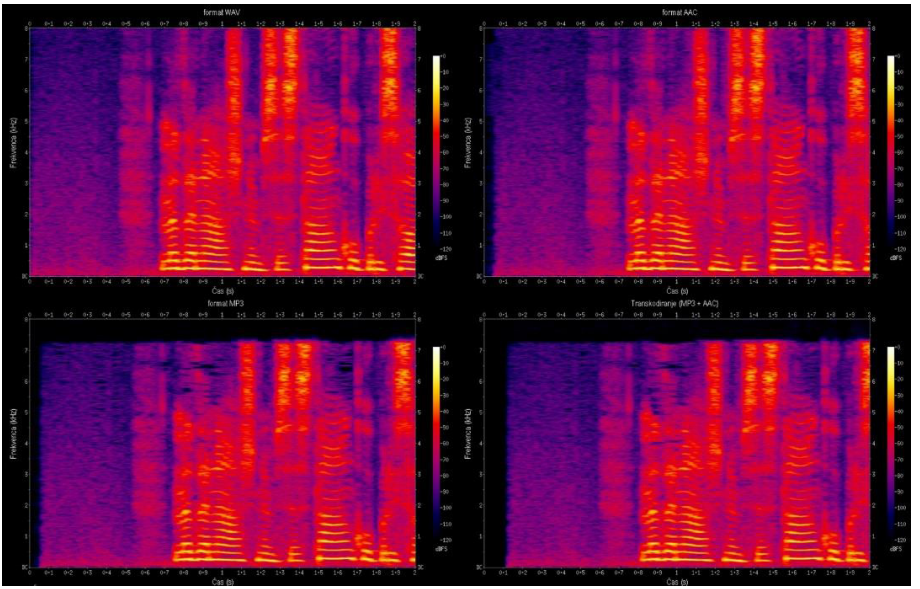
5.3 Izgubno stiskanje govora

Datoteke govorne baze UMB BNSI Broadcast News so v formatu WAV, ki ne uporablja stiskanja zvoka. Zanima nas, kakšno vlogo imajo izgubni kodeki pri avtomatskem razpoznavanju govora. V ta namen smo pripravili nove testne sete zvočnih datotek, ki smo jih najprej pretvorili v format z izgubnim kodekom in potem nazaj v izvorni format, potreben za razpoznavanje govora. V tem delu eksperimenta smo uporabili izgubna kodeka MPEG-1 Audio Layer III (MP3) in njegovega naslednika Advanced Audio Coded (AAC), ki je del skupine kodekov MPEG-2 Part 7. Kodek MP3 je definiran v standardih ISO/IEC 11172-3:1993 in ISO/IEC 13818-3:1995, kodek AAC pa v standardu ISO/IEC 13818-7:1997. Ključna razlika med njima je, da AAC omogoča še bolj učinkovito izgubno stiskanje zvoka pri enakem nivoju človeku zaznavnih degradacij.

Z orodjem SoX smo pretvorili izvorno testno množico datotek iz formata WAV v AAC pri bitni hitrosti 64 kbit/s in 128 kbit/s. Bitna hitrost originalnih datotek v formatu WAV je bila 256 kbit/s. Nato smo ponovili postopek še v obratni smeri in nove stisnjene datoteke pretvorili nazaj v format WAV.

Z orodjem FFmpeg smo pretvorili izvorno testno množico iz formata WAV v MP3 z bitno hitrostjo 64 kbit/s in 128 kbit/s. Postopek smo ponovili še v obratni smeri, da smo iz MP3 pretvorili posnetke nazaj v format WAV.

V naslednjem koraku smo želeli preveriti še, kakšen je vpliv transkodiranja na avtomatsko razpoznavanje govora. V tem primeru gre za večkratno zaporedno kodiranje z izgubnimi kodeki. Vzeli smo testne posnetke v formatu WAV, ki so že bili pretvorjeni v format MP3 z bitno hitrostjo 128 kbit/s, in jih ponovno pretvorili v format AAC z bitno hitrostjo 128 kbit/s in nazaj v format WAV.



Slika 2: Primerjava spektrogramov zvočnega zapisa dolžine 2 sekund v različnih zvočnih formatih.

Na Sliki 2 lahko opazimo, da pride pri formatu MP3 do rezanja frekvenc, višjih od 7,5 kHz, kar je značilno za pretvarjanje v format MP3 pri nizkih bitnih hitrostih. Glede na spektrogram, ki ga dobimo z zvočnim posnetkom formata WAV, lahko na ostalih treh spektrogramih opazimo razlike v deležih spektralne energije v različnih pasovih. Te razlike so nekoliko bolj vidne pri formatu MP3 kakor pri formatu AAC.

Za analizo vpliva izgubnih kodekov na avtomatsko razpoznavanje govora smo uporabili jezikovni model velikosti 64.000 in globoke nevronske akustične modele z aktivacijsko funkcijo tanh, ki so dosegli najboljše rezultate pri testiranju brez izgubne kompresije.

6 REZULTATI RAZPOZNAVANJA GOVORA

Vrednotenje različnih sistemov avtomatskega razpoznavanja govora smo izvedli na testni množici baze UMB BNSI Broadcast News (BNSI-eval), ki vsebuje 4 televizijske oddaje v obsegu 3 ur. Za metriko vrednotenja uspešnosti razpoznavanja govora smo uporabili delež napačno razpoznanih besed (Word Error Rate – WER), ki je definiran kot:

$$WER(\%) = \frac{(I+D+S)}{N} \cdot 100, \quad (2)$$

kjer je I število vrinjenih besed, D število izbrisanih besed in S število zamenjanih besed. N predstavlja število vseh besed v testni množici. V delu analize rezultatov smo kot metriko uporabili tudi delež napačno razpoznanih lem (Lemma Error Rate – LER), ki je definiran kot:

$$LER(\%) = \frac{(i+d+s)}{n} \cdot 100, \quad (3)$$

kjer je i število vrinjenih lem, d število izbrisanih lem in s število zamenjanih lem. n je skupno število vseh lem v testni množici in je enako številu besed N.

V prvem koraku evalvacije smo primerjali, kako je spreminjanje parametrov modelov vplivalo v koraku učenja z nevronsko mrežo, ko smo uporabili aktivacijsko funkcijo p-norm. V Preglednici 1 lahko vidimo rezultate WER, ki smo jih dosegli pri razpoznavanju testnega nabora.

Preglednica 1: *Primerjava rezultatov WER glede na različne nastavitve parametrov*

Aktivacijska funkcija	Število skritih plasti	Število vhodnih nevronov	Število izhodnih nevronov	WER [%]
p-norm	2	1000	200	19,85
p-norm	4	1000	200	19,22
p-norm z ozkim grlom	2	1000	200	19,73
p-norm z ozkim grlom	4	1000	200	19,04
p-norm z ozkim grlom	4	2000	400	19,36
p-norm z ansamblom	4	1000	200	19,54
p-norm z ansamblom	4	2000	400	19,59

Osnovna aktivacijska funkcija p-norm doseže najboljši rezultat, ko uporabimo 1000 nevronov na vhodu in 200 na izhodu s štirimi plastmi. Sistem, ki ima samo dve skriti plasti, doseže za 0,63 % slabši WER. Rezultat nekoliko izboljšamo v kombinaciji z ozkim grlom, kjer uporabimo 1000 nevronov na vhodu, 200 na izhodu, implementirane pa so bile 4 skrite plasti. V kombinaciji z ozkim grlom dosežemo nato tretji najboljši WER 19,36 %, ki je zgolj za 0,14% slabši od arhitekture s samo p-norm aktivacijsko funkcijo in 0,32 % slabši od najboljšega rezultata. Najslabši rezultat dobimo v kombinaciji aktivacijske

funkcije p-norm in ozkega grla z dvema skritima plastema, 1000 vhodnimi in 200 izhodnimi nevroni. V primerjavi z najboljšim rezultatom, doseženim zgolj s p-norm aktivacijsko funkcijo razpoznave govora, je za 0,51 % slabša in za 0,69 % slabša v primerjavi z najboljšim rezultatom aktivacijske funkcije p-norm v kombinaciji z ozkim grlom. Pri aktivacijski funkciji p-norm z ansamblom dosežemo boljši rezultat, če izberemo manj nevronov, in sicer 1000 na vhodu in 200 na izhodu. Dobljeni WER je 19,54 % in je za samo 0,05 % slabši v primerjavi z enako kombinacijo z več nevroni na vhodu in izhodu. Od najboljšega rezultata s samo p-norm aktivacijsko funkcijo se razlikuje za 0,32 % in 0,50 % od najboljšega dosežena rezultata.

V drugem koraku evalvacije smo izvedli primerjavo med avtomatskim razpoznavalnikom govora s prikritimi modeli Markova in globokimi nevronskimi mrežami. Pri tem je sistem s prikritimi modeli Markova služil za primerjavo z rezultati sistema, ki so ga Žgank in sodelavci objavili leta 2014 in je dosegel najboljši WER 26,81 %. Rezultati napake razpoznavanja besed s trigramskim jezikovnim modelom in slovarjem besed z velikostjo 64.000 so predstavljeni v Preglednici 2.

Preglednica 2: *Rezultati razpoznavanja govora s trigramskim 64.000 jezikovnim modelom*

Sistem	**WER [%]**
netransformiran HMM	26,48
transformiran HMM	24,28
DNN s p-norm	19,22
DNN s p-norm in z ozkim grlom	19,04
DNN s p-norm ansamblom	19,54
DNN s tanh	18,76
DNN s tanh in z ozkim grlom	23,33

Izhodiščna primerjava akustičnih modelov HMM kaže, da je prehod na novo ogrodje za avtomatsko razpoznavanje govora potekal brez težav, saj smo dosegli zelo primerljiv WER (s 26,81 % na 26,48 %). Osnovne akustične modele HMM je možno dodatno nadgraditi z metodama od govorca neodvisne transformacije značilk z uporabo LDA (angl. Linear Discriminant Analysis) in MLLT (angl. Maximum Likelihood Linear Transform) (Gales, 1999), kar izboljša rezultat s 26,48 % na 24,28 %. Vendar je to izboljšanje relativno

omejeno v primerjavi z možnostmi, ki jih v ustreznih pogojih omogoča globoko učenje. Najboljši rezultat dobimo z aktivacijsko funkcijo tanh, kjer WER znaša 18,76 %. V kombinaciji z ozkim grlom se razpoznavanje govora poslabša za 4,57 %. Kombinacija z ozkim grlom je nekoliko doprinesla pri razpoznavanju z aktivacijo funkcijo p-norm, kjer je rezultat s samo aktivacijsko funkcijo izboljšala za 0,18 %. Aktivacijska funkcija p-norm v kombinaciji z ansamblom ne prinese izboljšanja, saj je rezultat za 0,32 % slabši v primerjavi s samo aktivacijsko funkcijo p-norm. Prehod na globoke nevronske mreže za akustično modeliranje izboljša napako razpoznavanja besed na 18,76 %, kar predstavlja statistično pomembno razliko. Pri tem je treba posebej izpostaviti, da je količina govornega učnega materiala relativno omejena z vidika metod globokega učenja. Za učenje akustičnega modela z aktivacijsko funkcijo tanh smo na grafični kartici z NVIDIA grafičnim procesorjem V100 potrebovali 15,5 ur. Čas dekodiranja testnega nabora pa je trajal 22 minut, tako da je bil faktor realnega časa xRT približno 0,12.

V drugem koraku smo izvedli vrednotenje, kako vpliva na rezultate izboljšani jezikovni model z bistveno večjim slovarjem besed. Prehod s 64.000 besed na 250.000 besed namreč izdatno zniža delež besed izven slovarja in ga približa deležu, ki ga najdemo v tipičnih jezikovnih modelih za angleški jezik pri velikosti slovarja 64.000. Se pa poveča perpleksnost takšnega jezikovnega modela. Rezultati razpoznavanja govora z akustičnimi modeli DNN in obema trigramskima jezikovnima modeloma so predstavljeni v Preglednici 3.

Preglednica 3: *Rezultati razpoznavanja govora z akustičnimi modeli DNN z različnima trigramskima jezikovnima modeloma*

Jezikovni model	WER [%]
64.000-3g	19,22
250.000-3g	15,17

Tudi v scenariju razpoznavanja govora s slovarjem besed z velikostjo 250.000 je prišlo do znatnega zmanjšanja napake razpoznavanja besed, saj je WER znašal 15,17 %. S povečanjem slovarja razpoznavalnika govora smo tako izboljšali delovanje za 4,05 %, kar je primerljivo z zmanjšanjem deleža OOV. Pri tem smo ohranili kompleksnost sistema na primerljivi ravni, za kar smo poskrbeli med procesom izdelave jezikovnega modela. Razpoznavanje

slovenščine z nevronskimi mrežami so predstavili tudi Ulčar idr., 2019. Dosegli so WER 27,16 % na bazi GOS VideoLectures 2.0. Pri gradnji akustičnega modela so dodali tudi učenje s prilagajanjem govorcu (angl. speaker adaptive training), ki smo ga mi v gradnji izpustili. Modele GMM-HMM so nato uporabili kot osnovo za učenje modela DNN-HMM. Uporabili so arhitekturi TDNN in LSTM, preizkušali pa so več različnih konfiguracij mrež, kjer so različno povezovali plasti in spreminjali število skritih plasti. Zaradi uporabe različnih govornih in jezikovnih virov doseženi rezultati sicer niso neposredno primerljivi med seboj.

Najboljši doseženi rezultat razpoznavanja govora z jezikovnim modelom 250.000 3g je že primerljiv oziroma se je zelo približal rezultatom razpoznavanja govora v domeni televizijskih oddaj v nekaterih drugih jezikih. Avtorji v (Lleida idr., 2019) poročajo, da je na tekmovanju Albayzin RTVE 2018 Challenge za španščino najboljši sistem dosegel WER 16,45 %. Pri tem so uporabljali učni nabor posnetkov v dolžini več kot 200 ur.

V naslednjem koraku smo primerjali rezultate, ki smo jih pridobili s testnimi množicami, kjer smo uporabili dodatno izgubno kodiranje zvočnih zapisov.

Preglednica 4: *Rezultati razpoznavanja govora z vplivom izgubnih kodekov*

Kodek	WER [%]
MP3-64 kbit/s	19,21
MP3-128 kbit/s	19,10
AAC-64 kbit/s	18,96
AAC-128 kbit	18,84
MP3+AAC- 128kbit/s	19,47

Najboljši rezultat dobimo z izgubnim kodekom AAC, ki prinaša 0,08 % slabši rezultat glede na rezultat, ki smo ga dobili s posnetki v formatu WAV. Slabše se odreže kodek MP3, ki ima za 0,45 % slabši rezultat pri bitni hitrosti 64 kbit/s in za 0,34 % slabši rezultat pri 128 kbit/s. Manjša bitna hitrost poslabša rezultat za približno 0,1 %. Najslabše rezultate prinese transkodiranje, rezultat se poslabša za 0,71 %. Kodiranje z izgubnimi kodeki ne prinaša velikega poslabšanja rezultatov razpoznavanja govora. Na njihovi podlagi lahko predpostavimo, da bi takšen razpoznavalnik govora učinkovito deloval tudi

s posnetki, ki uporabljajo izgubne kodeke. Podobno kakor je bilo prikazano v članku (Pollak in Behunek, 2011), kjer so primerjali razpoznavanje govora z izgubnim kodekom MP3 pri različnih bitnih hitrostih, lahko opazimo, da je razpoznavalnik govora sposoben učinkoviteje razpoznavati posnetke, kadar je na voljo govor, kodiran z višjo bitno hitrostjo.

V nadaljevanju poglavja bomo podrobneje predstavili analizo doseženih rezultatov razpoznavanja govora. Tukaj smo uporabili akustične modele brez dodatne nadgradnje v obliki transformacije značilk. Odgovoriti poskušamo na vprašanje, kako različni faktorji vplivajo na WER. V to skupino sodijo delež besed izven slovarja, pregibna oblika besed, akustično ozadje in način govora.

Referenčne transkripcije in rezultate razpoznavanja smo oblikoslovno označili ter lematizirali z označevalnikom slovenskega jezika Obeliks (Grčar idr., 2012). Oznake besedne vrste in leme so nam koristile pri podrobnejši analizi rezultatov.

S primerjavo lematizirane referenčne transkripcije ter lematiziranih rezultatov razpoznavanja govora smo določili delež napačno razpoznanih lem ter izluščili napake, kjer je lema pravilno razpoznana, besedna oblika pa ne. Na takšen način smo lahko delno analizirali vpliv pregibnosti slovenskega jezika na rezultate razpoznavanja govora. Za lematizacijo smo se odločili, ker pravilno razpoznana lema poda več informacij kot pa pravilno razpoznani koren besede ali uporaba deleža napačno razpoznanih znakov. V primeru pravilno razpoznane leme lahko predvidevamo, da se v večji meri ohrani pomen kot pa v primeru pravilno razpoznanega korena besede. S tem želimo pridobiti boljšo oceno, ali bi bralec avtomatske transkripcije lahko pravilno razumel pomen stavka in opazil le slovnično napako, medtem ko bi se pri pravilno razpoznanem korenu besede spremenil pomen stavka. Ta razlika je še bolj očitna v primeru uporabe deleža napačno razpoznanih znakov, saj lahko en napačni znak spremeni pomen stavka. S pomočjo oblikoslovnih oznak pa smo nato še napake v besedni obliki razdelili po besednih vrstah.

V Preglednici 5 so predstavljeni podrobnejši rezultati. Pri izhodiščnih rezultatih za HMM in DNN s 64.000 besedami v slovarju je prišlo do manjšega odstopanje v WER v primerjavi s prejšnjimi rezultati. Razlog za to odstopanje je uporaba drugega orodja za analizo rezultatov, ki nekoliko drugače poravna

rezultate razpoznavanja govora z referenčnimi transkripcijami. Razdeljeni rezultati po F-razredih in po spolu kažejo večinoma podobna izboljšanja pri prehodih med sistemi. Opazna je razlika med rezultati za moške in ženske govorce, ki znaša 4,29 %. To razliko bo v prihodnosti treba še podrobneje analizirati. Večja izboljšanja vidimo v razredih F1, F3 in FX pri prehodu s sistema HMM na DNN ter pri razredu F2 pri prehodu na večji slovar, ki pa predstavlja le zelo majhen del testne množice. Medtem ko za bran studijski govor dosegamo WER 7,83 %, sprememba na spontani govor ali dodajanje akustičnega ozadja poslabša rezultate v rangu 10 do 21 %. Pri tem je pričakovano poslabšanje večje, če je v ozadju dodana glasba.

Preglednica 5: *Podrobnejša predstavitev rezultatov razpoznavanja po F-razredih in spolu ter rezultati pravilnosti razpoznave lem*

Sistem	HMM 64.000-3g	DNN 64.000-3g	DNN 250.000-3g
WER [%]	24,33	18,82	15,17
WER – F0 [%]	14,63	11,46	7,83
WER – F1 [%]	31,57	24,27	20,99
WER – F2 [%]	58,47	39,83	38,14
WER – F3 [%]	33,43	25,83	21,01
WER – F4 [%]	27,14	21,10	17,28
WER – FX [%]	31,95	24,15	21,45
WER – Moški [%]	26,16	20,74	17,06
WER – Ženske [%]	21,95	16,43	12,77
LER [%]	23,33	17,70	14,06
WER – LER	1,00	1,12	1,11

Rezultati deleža napačno razpoznanih lem LER so po pričakovanjih nižji od rezultatov WER. Te razlike nakazujejo napake v razpoznavanju, kjer je sistem napačno razpoznal besedno obliko, vendar imata tako razpoznana kot pravilna beseda enako lemo. Vidimo, da je razlika manjša pri sistemu z večjim slovarjem, kar nakazuje, da je za del napačno razpoznanih besednih oblik odgovoren omejen slovar.

Treba je dodati, da je bila v nekaterih primerih razpoznana pravilna besedna oblika, vendar je lematizator označil različni lemi med hipotezo in referenco. Ti primeri so se šteli kot napake v vrednotenju LER. To se dogaja predvsem

pri primerih, kjer se zaradi drugih napak (izbrisanih ali vrinjenih kratkih besed) spremeni kontekst besede. Na primer, besedna oblika *ukrepa* je lahko označena z lemo *ukrep* (samostalnik) ali pa *ukrepati* (glagol). Ocenjujemo pa, da je delež teh primerov le majhen. Iz tega sklepamo, da je delež napak, ki so posledica pregibnosti jezika, nekoliko višji kot pa razlika med WER in LER, namreč okoli 1 %.

V nadaljevanju smo pregledali napake v besedni obliki pri isti lemi glede na besedno vrsto. Rezultati so podani v Preglednici 6. Podali smo le pregibne besedne vrste (brez zaimkov). Primerjamo sistema HMM 64.000-3g in DNN 250000-3g. Vidimo, da je le relativno izboljšanje pri napačno razpoznanih oblikah številnikov primerljivo z relativnim izboljšanjem skupnega rezultata, ki je 34,9 %. Najmanjše relativno izboljšanje pa vidimo pri glagolih. Skupno relativno izboljšanje napak v besedni obliki je približno dvakrat manjše od relativnega izboljšanja skupnega rezultata.

Preglednica 6: *Napačno razpoznane besedne oblike glede na besedno vrsto*

Besedna vrsta	Št. napak v HMM 64.000-3g	Št. napak v DNN 250.000-3g	Relativna izboljšava [%]
Samostalnik	309	265	14,2
Pridevnik	155	112	27,7
Glagol	148	135	8,8
Številnik	16	10	37,5
Prislov	0	1	-
SKUPAJ	628	522	16,9

Rezultati kažejo na to, da sistem s povečanim slovarjem in uporabo globokih nevronskih mrež pomembno zmanjša skupni delež napak razpoznavanja. Vidimo lahko, da je relativno zmanjšanje napak zaradi pregibnosti besed manjše glede na skupno zmanjšanje. V sistemu DNN 250.000-3g je tako delež napak zaradi pregibnosti 13,3 %, kar je več kot pri sistemu HMM, kjer je ta delež 10,5 %.

Pregled posameznih najpogostejših parov zamenjav ne kaže zanimivih rezultatov glede pregibnih besed. Večinoma se v pogostih parih zamenjav pojavljajo kratke besede (npr. zamenjave so – se, na – no ipd.). Najpogostejši par zamenjave, kjer je prišlo do napake v besedni obliki polnopomenske

pregibne besede, je par stališče – stališča, ki se pojavi štirikrat v sistemu DNN 250.000-3g.

Doseženi rezultati kažejo, da je z obstoječimi slovenskimi govornimi viri možno učinkovito graditi razpoznavalnike govora za domeno dnevnoinformativnih oddaj, če govorimo o preprostejših akustičnih pogojih. Kakor hitro pa dodamo zahtevnejše akustične pogoje, se rezultati poslabšajo. S tega vidika je pomembno delo na povečevanju razpoložljivih govornih virov za slovenski jezik. Z vidika visoke pregibnosti slovenskega jezika se je pokazalo, da lahko to lastnost učinkovito naslovimo z zniževanjem deleža besed izven slovarja. Na takšen način lahko modeliramo večino besed, težavna kategorija pa ostajajo kratke besede, ki so si akustično podobne. Za izboljšano akustično modeliranje v takšnih primerih pa je ponovno neobhodno potrebno več učnega govornega materiala. Pristop z zmanjševanjem deleža besed izven slovarja kaže, da je za doseganje primerljivih rezultatov razpoznavanja govora z jeziki, kot je angleščina, potreben za 3- do 5-krat večji slovar razpoznavalnika govora.

7 SKLEP

V članku smo predstavili sistem za avtomatsko razpoznavanje slovenskega govora v domeni televizijskih oddaj. Najboljši doseženi rezultat deleža napake razpoznavanja besed je znašal 15,17 %. Takšen sistem je po svojih rezultatih razpoznavanja govora že primerljiv z nekaterimi rezultati, doseženimi za druge jezike. Izboljšanje je v pretežni meri rezultat uporabe akustičnih modelov z globokimi nevronskimi mrežami in vpliva zmanjšanja deleža besed izven slovarja. Z večanjem slovarja smo uspešno zmanjšali vpliv pregibnosti slovenskega jezika.

Podrobnejša analiza po F-razredih in lemah je pokazala, da je nadaljnje izboljšanje rezultatov možno doseči predvsem na račun izboljšanja akustičnega modeliranja v primeru kratkih besed in govora v zahtevnejših pogojih. V prihodnjem delu se je tako smiselno osredotočiti na povečanje gradiva za učenje akustičnih modelov in s tem povezane spremembe v arhitekturi takšnih modelov.

Zahvala

Zahvaljujemo se avtorjem besedilnega korpusa FidaPLUS, ki so nam omogočili njegovo uporabo za jezikovno modeliranje avtomatskega razpoznavalnika govora.

Raziskovalno delo je bilo delno sofinancirano s strani ARRS po pogodbi št. P2-0069. Raziskovalno delo je bilo delno opravljeno v okviru projekta RSDO – Razvoj slovenščine v digitalnem okolju. Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

LITERATURA

Arhar, Š., & Gorjanc, V. (2007). Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo, (52)*2, 95–110.

Dobrišek, S., Gros, J., Mihelič, F., & Pavešić, N. (1998). Recording and labelling of the GOPOLIS Slovenian speech database. V *First International Conference on language resources & evaluation*: Granada, Spain, 28–30 May 1998 (str. 1089–1096). European Language Resources Association.

Dobrišek, S., & Mihelič, F. (2010). Zmanjševanje odvečnosti končnih pretvornikov za učinkovito gradnjo razpoznavalnikov slovenskega govora z velikim besednjakom. V *Jezikovne tehnologije: zbornik 13. mednarodne multikonference, Informacijska družba IS* (str. 24–27).

Dobrišek, S., Žganec Gros, J., Žibert, J., Mihelič, F., & Pavešić, N. (2017). Speech Database of Spoken Flight Information Enquiries SOFES 1.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1125

ELRA. (2015). Pridobljeno s http://www.elra.info

Gales, M. J. (1999). Semi-tied covariance matrices for hidden Markov models. *IEEE transactions on speech and audio processing, 7*(3), 272–281.

Grčar, M., Krek, S., & Dobrovoljc, K. (2012). Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. V T. Erjavec in J. Žganec Gros (ur.), *Zbornik Osme konference Jezikovne tehnologije*,

Ljubljana, Slovenija (str. 89–94). Ljubljana: Institut Jožef Stefan. Pridobljeno s http://nl.ijs.si/isjt12/JezikovneTehnologije2012.pdf

Imperl, B., Kačič, Z., & Horvat, B. (1996). Razpoznavanje osamljenih besed s polzveznimi Prikritimi modeli Markova. V *Zbornik pete Elektrotehniške in računalniške konference ERK* (str. B/231–234).

Imperl, B., & Kačič, Z (1999). Connected digits and natural numbers recognition for the telephone multilingual speech dialog systems. *V Proceedings of the 4th international workshop on Electronics, control, measurement and signals ECMS* (str. 164–167).

Ipšić, I., Mihelič, F., Dobrišek, S., Žganec Gros, J., & Pavešić, N. (1999). A Slovenian spoken dialog system for air flight inquiries. V *Eurospeech '99: proceedings, 6th European Conference on Speech Communication and Technology* (str. 2659–2662).

Kačič, Z., Horvat, B., & Greif, Š. (1988). Man-machine communication: speaker-independent speech recognition. *Informatica: an international journal of computing and informatics, (12)*1, 6–12.

Kaiser, J., & Kačič, Z. (1997). SpeechDat (II) Slovenian Database for the Fixed Telephone Network. Maribor, Slovenia: University of Maribor.

Kaiser, J., Sepesy Maučec, M., Kačič, Z., & Horvat, B. (2000). Razpoznavanje tekočega slovenskega govora z velikim slovarjem. V T. Erjavec in J. Gros (ur.), *Jezikovne tehnologije* (str. 39–44). Ljubljana: Institut Jožef Stefan. Pridobljeno s http://nl.ijs.si/isjt00/zbornik/sdjt00-Kaiser06.pdf

Lleida, E., Ortega, A., Miguel, A., Bazán-Gil, V., Pérez, C., Gómez, M., & De Prada, A. (2019). Albayzin 2018 evaluation: the iberspeech-RTVE challenge on speech technologies for spanish broadcast media. *Applied Sciences*, *9*(24), 5412.

Mihelič, F., Ipšić, I., Dobrišek, S., & Pavešić, N. (1992). Feature representations and classification procedures for Slovene phoneme recognition. *Pattern recognition letters*, *13*(12), 879–891.

Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., & Shaalan, K. (2019). Speech recognition using deep neural networks: A 463 systematic review. *IEEE Access* 2019, 7, 19143–19165.

Nouza, J., Safarik, R., & Cerva, P. (2016). ASR for South Slavic Languages Developed in Almost Automated Way. V *Interspeech* (str. 3868–3872).

Pollak, P., & Behunek, M. (2011). Accuracy of MP3 speech recognition under real-word conditions: Experimental study. V *Proceedings of the International Conference on Signal Processing and Multimedia Applications* (str. 1–6). IEEE.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N.,..., Silovsky, J. (2011). The Kaldi speech recognition toolkit. V *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

RSDO. (b. d.). Pridobljeno s https://www.cjvt.si/rsdo/

Schwartz, R., Jin, H., Kubala, F., & Matsoukas, S. (1997). Modeling Those F-Conditions – or not. V *Proc. DARPA Speech Recognition Workshop*, Chantilly, ZDA.

Stolcke, A. (2002). SRILM – an extensible language modeling toolkit. SRILM – an extensible language modeling toolkit. V *International Conference on Speech and Language Processing* (str. 901–904).

Ulčar, M., Dobrišek, S., & Robnik-Šikonja, M. (2019). Razpoznavanje slovenskega govora z metodami globokih nevronskih mrež. *Uporabna informatika. 27,* 3.

Verdonik, D., Kosem, I., Vitez, A., Krek, S., & Stabej, M. (2013). Compilation, transcription and usage of a reference speech corpus: The case of the Slovene corpus GOS. *Language resources and evaluation*, *47*(4), 1031–1048.

Verdonik, D., Potočnik, T., Sepesy Maučec, M., & Erjavec T. (2017). Spoken corpus Gos VideoLectures 2.0 (transcription). Maribor: Fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru. Pridobljeno s http://hdl.handle.net/11356/1222

Verdonik, D. (2018). Korpus in baza Gos Videolectures. V D. Fišer in A. Pančur (ur.), *Zbornik 11. konference Jezikovne tehnologije in digitalna humanistika* (str. 265–268). Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani. Pridobljeno s http://nl.ijs.si/jtdh18/JTDH-2018-Proceedings.pdf

Zhang X., Trmal, J., Povey, D., & Khudanpur, S. (2014). Improving deep neural network acoustic models using generalized maxout networks. V 2014 *IEEE international conference on acoustics, speech and signal processing (ICASSP)* (str. 215–219). IEEE.

Zorrilla, A. L., Dugan, N., Torres, M. I., Glackin, C., Chollet, G., & Cannings, N. (2016). Some asr experiments using deep neural networks on spanish databases. *Advances in Speech and Language Technologies for Iberian Languages*. IberSPEECH.

Zwitter Vitez, A., Zemljarič Miklavčič, J., Krek, S., Stabej, M., & Erjavec, T. (2013). Spoken corpus Gos 1.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1040

Žgank, A., Kačič, Z., & Horvat, B. (2002). Preliminary evaluation of Slovenian mobile database PoliDat. V *Proceedings of the Third International Conference on Language Resources and Evaluation* (LREC'02).

Žgank, A., Rotovnik, T., Sepesy Maučec, M., Verdonik, D., Kitak, J., Vlaj, D., Hozjan, V., ..., Horvat, B. (2004). Acquisition and annotation of Slovenian broadcast news database. V *Fourth international conference on language resources and evaluation*, *LREC 2004* (str. 2103–2106). Lizbona, Portugalska. Pridobljeno s http://www.lrec-conf.org/proceedings/lrec2004/pdf/123.pdf

Žgank, A., Rotovnik, T., Grašič, M., Kos, M., Vlaj, D., & Kačič, Z. (2006). Sloparl-Slovenian parliamentary speech and text corpus for large vocabulary continuous speech recognition. V *Ninth International Conference on Spoken Language Processing*. Pridobljeno s http://dblp.uni-trier.de/db/conf/interspeech/interspeech2006.html#ZgankRGKVK06

Žgank, A., Rotovnik, T., Sepesy Maučec, M., & Kačič, Z. (2006). Osnovna zgradba razpoznavalnika slovenskega tekočega govora UMB Broadcast News. V T. Erjavec in J. Žganec Gros (ur.), *Jezikovne tehnologije: zbornik 9. mednarodne multikonference Informacijska družba IS* (str. 99–118). Ljubljana: Institut Jožef Stefan. Pridobljeno s http://nl.ijs.si/is-ltc06/proc/

Žgank, A., & Sepesy Maučec, M. (2010). Razpoznavalnik tekočega govora UMB Broadcast News 2010: nadgradnja akustičnih in jezikovnih modelov. V T. Erjavec in J. Žganec Gros (ur.), *Jezikovne tehnologije 2010* (28–31). Ljubljana: Institut Jožef Stefan. Pridobljeno s http://nl.ijs.si/isjt10/JezikovneTehnologije2010.pdf

Žgank, A., Donaj, G., & Sepesy Maučec, M. (2014). Razpoznavalnik tekočega govora UMB Broadcast News 2014: kakšno vlogo igra velikost učnih virov. V V T. Erjavec in J. Žganec Gros (ur.) *Zbornik 9. konference Jezikovne tehnologije, Informacijska družba IS* (str. 147–150). Ljubljana: Institut

Jožef Stefan. Pridobljeno s http://library.ijs.si/Stacks/Proceedings/InformationSociety/2014/2014_IS_CP_Volume-G_(LT).pdf

Žgank, A., Sepesy Maučec, M., & Verdonik, D. (2016). The SI TEDx-UM speech database: A new Slovenian spoken language resource. V *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (str. 4670–4673).

Žibert, J., Mihelič, F., & Dobrišek, S. (2000). Avtomatično podnaslavljanje vremenskih napovedi. V B. Zajc (ur.), *Zbornik devete Elektrotehniške in računalniške konference, Portorož, Slovenija, 21.–23. september 2000* (str. 165–168).

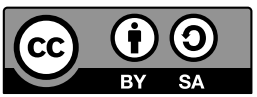
Žibert, J., Martinčić-Ipšić, S., Ipšić, I., & Mihelič, F. (2003). Bilingual speech recognition of Slovenian and Croatian weather forecasts. V Proceedings EC-VIP-MC 2003. *4th EURASIP Conference focused on Video/Image Processing and Multimedia Communications* (IEEE Cat. No. 03EX667) (Vol. 2, str. 637–642). IEEE.

Žibert, J., & Mihelič, F. (2004). Development of Slovenian broadcast news speech database. V *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)* (str. 2095–2098). Pridobljeno s http://www.lrec-conf.org/proceedings/lrec2004/pdf/98.pdf

SLOVENIAN AUTOMATIC SPEECH RECOGNITION FOR BROADCAST NEWS

In speech and language technologies, automatic speech recognition is one of the key building blocks. In this article, we will explain the development of an automatic recognizer of Slovenian speech for the domain of daily news broadcasts. The architecture of the system is based on a deep neural net. Considering the available speech sources, we performed modeling with various activation functions. In the development of speech recognition, we also checked the impact of lossy speech codecs on speech recognition results. We used the UBM BNSI Broadcast News and IETK-TV databases to train the speech recognizer. The total amount of voice recordings was 66 hours. In parallel with the deep neural networks, we increased the speech recognition dictionary, which amounted to 250,000 words. In this way, we reduced the out-of-vocabulary rate to 1.33%. Speech recognition on the test set achieved the best WER of 15.17%. While evaluating the results, we also performed a more detailed analysis of speech recognition errors based on lemmas and F-conditions, which to some extent show the complexity of the Slovenian language for such scenarios of technology use.

Keywords: automatic speech recognition, characteristics of Slovenian language, broadcast news, deep neural networks, lossy speech codecs



Slovenščina 2.0, 2021 (1)

SIGN LANGUAGE LEXICOGRAPHY: A CASE STUDY OF AN ONLINE DICTIONARY

Lucia VLÁŠKOVÁ
Support Centre for Students with Special Needs (Teiresiás), Masaryk University

Hana STRACHOŇOVÁ
Faculty of Arts, Masaryk University

Vlášková, L., Strachoňová, H. (2021): Sign language lexicography: a case study of an online dictionary. Slovenščina 2.0, 9(1): 90–122.

DOI: https://doi.org/10.4312/slo2.0.2021.1.90-122

As a growing field of study within sign language linguistics, sign language lexicography faces many challenges that have already been answered for audio-oral language material. In this paper, we present some of these challenges and methods developed to help navigate the complex lexical classification field. The described methods and strategies are implemented in the first Czech sign language (ČZJ) online dictionary, a part of the platform *Dictio*, developed at Masaryk University in Brno. We cover the topic of lemmatisation and how to decide what constitutes a lexeme in sign language. We introduce four types of expressions that qualify for a dictionary entry: a simple lexeme, a compound, a derivative, and a set phrase. We address the question of the place of classifier constructions and shape and size specifiers in a dictionary, given their peculiar semantic status. We maintain the standard classification of classifiers (whole entity and holding classifiers) and size and shape specifiers (SASSes; static and tracing specifiers). We provide arguments for separating the category of specifiers from the category of classifiers. We discuss the proper treatment of mouthings and mouth gestures concerning citation forms, derivation and translation. We show why it is difficult in sign language to distinguish synonyms from variants and how our proposed phonological criteria can help. We explain how to construct a semantic definition in a sign language and what is the solution for multiple meanings of one form. We offer simple guidelines for forming proper examples of use in a sign language. And finally, we briefly comment on the process of the translation between sign and spoken languages. We conclude the paper with a summary of roles that *Dictio* plays in the ČZJ-signing community.

Keywords: sign language, lexicography, dictionary, methodology

1 INTRODUCTION

Dictio is a multilingual online dictionary that includes multiple languages, both sign and spoken. This ongoing project is being realised at Masaryk University in Brno, Czech Republic. Currently, it includes entries for the following languages (ordered by the approximate number of entries): Czech (120 thousand), Czech Sign Language – ČZJ (13 thousand), Slovak Sign Language (5 thousand), Slovak (5,5 thousand), English (5,5 thousand), Austrian German (5,5 thousand), Austrian Sign Language (3,5 thousand), International Sign (170), and American Sign Language (120). Only a section of the entries has been published, the rest is still the subject of editing work of multiple working groups, including international teams of Deaf university employees. At the time of writing (January 2021), the number of the sign language published entries are as follows: Czech Sign Language – 3075, Slovak Sign Language – 35, International Sign – 12, American Sign Language – 20.

The field of sign language lexicography has been growing rapidly. Considering Stokoe's (1960/2005) description of the lexical units in American Sign Language as the pioneering work which respects the established linguistic principles, sixty years later, we make use of systematised databases for a whole range of sign languages in the form of printed books or offline and online databases (see the overview in McKee and Vale, 2017 or Fenlon et al., 2015). Since the seminal work of Johnston and Schembri (1999) on lemmatisation of the Australian Sign Language corpus (and closely connected Australian Sign Language lexical database), several researchers have published their experiences in the form of applicable universal guidelines for the lexicographic work on any sign language. Recently, many topics concerning mainly the electronic lexical databases have been addressed in the literature: e.g., history and options of the sign description and search (Zwitserlood, 2010, focusing on dictionaries of Dutch Sign Language), lexicographic specifics of sign languages compared to spoken languages (Kristoffersen and Troelsgård, 2012, with particular focus on the lexical database of Danish Sign Language), phonological and morphological variation in the process of lemmatisation (Fenlon et al., 2015, on the material of British Sign Language), and others. At the beginning (around 2009), our project was inspired mainly by the work of Johnston and Schembri (1999) and online public dictionaries of Italian Sign Language (e-LIS) and French Sign Language (Elix).

The choice of our sources of inspiration arose from the ambition of our project: to create an up-to-date sign language dictionary comparable to standard spoken language dictionaries. Firstly, we were interested in providing linguistic metadata like the sign's lexical category, its region of use, or its grammatical modifications (hence Johnston and Schembri's work). Secondly, we aimed to create semantic definitions and examples of use for each meaning directly in ČZJ. Even today, that is not obvious for a sign language dictionary. We can still find several sign language dictionaries that explain the meaning of a sign using the surrounding spoken language (in some cases, that also applies to the examples of use). From this perspective, we consider the editors of e-LIS and Elix to be pioneers who we wanted to emulate.

In the absence of a representative ČZJ corpus, the linguistic material for the ČZJ part of the dictionary comes from two primary sources: previously published dictionaries and ČZJ informants. *Dictio* has the ambition to collect all the published ČZJ dictionaries and make them available in one database. That covers printed books (mainly Potměšil, 2002, 2004, 2004a), CDs (Langer, 2005, 2005a, 2008, a.o.), and other individual projects (e.g., diploma theses focusing on specific semantic fields, teaching materials for ČZJ commercial or university courses). The collection of previously published material is being edited, annotated and completed by a team of native signers of ČZJ, ČZJ interpreters and linguists. A substantial part of the team's work is to discuss synonyms and variants for the published entries. This way, plenty of new material is being elicited for the *Dictio* database.

In this paper, we introduce selected topics from sign language lexicography. The idea is to describe some linguistic issues we have encountered while working on the ČZJ part of the dictionary and propose guidelines applicable to the field of sign language lexicography in general. ČZJ was the first language introduced into the dictionary. Creating the linguistic methodology has been especially challenging since the original vision of the entire project was to construct the first monolingual dictionary, in this case, a dictionary of ČZJ, where the meaning and the use of the signs are explained and illustrated solely in ČZJ. As *Dictio* was becoming multilingual, links to the parts containing other languages (translations) were added to the entries. That is why proper semantic definitions were crucial, which will also be discussed below.

2 LEMMATISATION AND TYPES OF DICTIONARY ENTRIES

The most fundamental question when compiling a sign language dictionary is what kind of signs to include, i.e., what constitutes an entry in a dictionary. The following strategy has been developed to answer this question: first, we take all the possible kinds of signs occurring in natural speech (lexeme, deixis, description, compound, collocation, set phrase) and divide them into two groups according to their complexity: the ones that do not consist of multiple semantic units (lexeme, deixis) and the ones that do (description, collocation, compound, set phrase). The first group is illustrated with the signs BLACK and IX-a, the latter with DEFECT, FEBRUARY, VETERINARY and 25^{TH}.[1] DEFECT contains two lexical roots: FAULT and BREAK-DOWN. In FEBRUARY, a native signer can distinguish the roots of MASK and DANCE. VETERINARY is formed by a sequence of DOCTOR, FOCUS and ANIMAL. And finally, 25^{TH} simply linearizes the numerals 20 and 5^{TH}. Among the group of simple expressions, we set aside the expression, the meaning of which changes according to the referent (deixis: IX-a) and select the expression with a conventionally established meaning (lexeme: BLACK). We single out the expressions with a non-compositional meaning from the group of complex expressions, i.e., the set phrase (DEFECT) and the compound (FEBRUARY). Similarly to the spoken language dictionaries, collocations (25^{TH}) and descriptions (VETERINARY) are not listed as dictionary entries. Language users combine them regularly using the established lexicon and grammar of the language. However, they found their place in the example section of the entry (see Section 7 of this paper).

The above-described strategy leaves us with only three candidates for a dictionary entry: a traditional lexeme (BLACK), a compound (FEBRUARY), and a set phrase (DEFECT), with conventionally established meanings. In *Dictio*, however, we make another distinction, i.e. we divide the group of traditional lexemes into a group of motivated/derived signs and a group of simple unmotivated signs. Therefore, we classify signs into four types of entries: simple signs, compounds, set phrases, and derivatives. Let us briefly comment on each type.

1 We use the gloss IX-a for an index pointing at a location *a*, as is common. A possible translation could be *that*.

Simple signs are monomorphemic. In our diagnostics of a sign language morpheme (namely the root), we follow Sandler (2006) and her two criteria that must be met to classify the sign as monomorphemic: The Selected Finger Constraint and The Place Constraint. The Selected Finger Constraint (originally in Mandel, 1981; revisited by Sandler, 1989) says that only one set of fingers can be selected within a morpheme. Note that this requirement allows the internal movement of the fingers.[2] Compare a monomorphemic sign LAMP which displays one selection of fingers (changing their position from closed to open) with the sign RECOMMEND, which contains two selection of fingers (one open finger in the initial position, all open fingers in the final position), and is thus analysed as multimorphemic (a compound).

The second criterion we consider is The Place Constraint (originally in Battison, 1978; revisited by Sandler, 1989). It states that a morpheme can contain only one place of articulation. There are four main places of articulation: the neutral space, the head, the trunk, and the non-dominant hand. A movement from one location to another within the same main area is not considered a change of the place. The logic of the constraint is applied as follows: the sign POST-OFFICE is multimorphemic (a compound) because the dominant hand moves from the head to the non-dominant hand. In contrast, the sign NAME is compliant with the constraint: the hand moves from the contralateral to the ipsilateral side of the forehead. Both locations are a part of just one place of articulation (the head), and that is why the sign is classified as monomorphemic (simple).

Compounds are morphologically complex signs that originated by merging two independent signs, i.e., two free morphemes. From the semantic point of view, compounds are not bound to introduce a new meaning, as seen in the ČZJ example of SUN^GLASSES ‘sunglasses’. Nevertheless, it is possible, e.g., FLOWER^SPRING ‘May’ (Mladová, 2009). It is often difficult to distinguish compounds from set phrases, another type of entries in our dictionary. Set phrases also consist of two (or more) free morphemes, but their meaning is not compositional, e.g., in ČZJ sign UNIVERSITY, which consists of HIGH

2 Selected fingers are fingers that constitute the handshape. The fingers may be open (like in SUGAR with selected thumb and index finger) or closed (like in POST-OFFICE with all the fingers selected). The internal movement is defined as a change of the orientation of the dominant hand or a change of the position of its fingers (open/closed).

and SCHOOL. However, in the case of compounds it is not the semantic shift that classifies them as such but the phonological reduction/assimilation, as defined by Zeshan (2004): the first sign is shortened and loses stress, any repetitions and internal movements are deleted, handshape and location can be assimilated, and the passive hand can function as a place of articulation.[3] On the other hand, no such modification can be found in set phrases, where all constituting signs are fully realised.

The last type represented in our dictionary are the derivatives, defined as forms that have been derived from their respective motivating signs through adding or changing a non-manual component, which we will discuss in more detail in Section 4. Typically, this process occurs while deriving a technical or more specific term from a general vocabulary sign. Sandler (2006) affirms that mouthing is of a significant lexical role. Take an example from ČZJ where SACCHARIDE is derived from SUGAR. These two signs have the same manual component but differ in mouthing. SUGAR is standardly articulated without mouthing, and SACCHARIDE contains the mouthing of the Czech word for *saccharide*.[4]

Another critical question is the choice of a citation form (headword) of each entry. Following Johnston and Schembri (1999), only the unmodified signs in their basic forms are present in the lexicon (and, therefore, the dictionary), inflexion and modification are part of the grammar. Modification can take several forms, as defined in Zeshan (2002, 2004): (i) modified movement expresses the change in aspect, number, degree or directionality (verbal inflexion encoding the subject and/or the object of the given verb like $_{1}$RETURN$_{2}$ 'I return (sth) to you' vs $_{2}$RETURN$_{1}$ 'you return (sth) to me'; or intensification like in RAIN vs RAIN-A-LOT); (ii) modified handshape signals classifier constructions and numeral incorporation (e.g., HOUR can incorporate numerals up to 10, as seen in FOUR-HOUR with an incorporated numeral *four*); (iii) modified facial expressions distinguish between clause types, such as indicative, interrogative, negative (e.g., LIKE and NOT-LIKE) and others. In *Dictio*,

3 At least one reduction/assimilation pattern must be present to classify the item as a compound.

4 More precisely, the sign for SUGAR may be accompanied by the mouthing of the Czech word for *sugar*, but the sign for SACCHARIDE must be articulated with the mouthing of the Czech word for *saccharid*e.

the information whether a sign can incorporate numerals, (classifiers for) subject and/or object, and other modifiers is given in the grammatical part of the dictionary entry. The lexeme is presented in its basic form, i.e. singular, non-modified and non-intensified sign, such as the above-mentioned HOUR. The basic form for signs that incorporate a numeral is the one with incorporated ONE. For directional signs, it is the form directed from the speaker to the addressee.

However, there are exceptional cases when the dictionary also covers other than basic forms of signs. Such instances include deixis with fixed hand position, e.g., the pronouns I and MY that are always signed facing the speaker, and, correspondingly, YOU and YOUR, always facing the addressee. Furthermore, lexicalised forms of different types have their place in the dictionary, e.g., lexicalised deixis. Take the ČZJ verb HEAR, which is realised by pointing to the speaker's ear with a crooked index finger. As deixis, the pointing sign would be interpreted as *that* (consequently, as *ear*). The lexicalisation process is observed at two levels: formal and semantic. The formal change consists in the movement modification (the hand moves from the ear). During the semantic shift, the meaning no longer corresponds to the object that is being pointed at. It shifted to the activity realized by the object. Other forms of lexicalisation include lexicalised classifier constructions, which we will discuss in the following section, or lexicalised fingerspelling, as the sign for *engineer* – I-N-G, fingerspelled with the letters of the ČZJ alphabet.

3 CLASSIFIERS, SPECIFIERS AND LEXICALISED CONSTRUCTIONS

Classifiers have repeatedly proven to be an exciting research topic among sign linguists. This section will focus on different classifiers, a closely related group of specifiers, and the ways of properly incorporating them into a dictionary.

Sign language classifiers are considered a special kind of morphemes, the meaning of which is not precisely specified. They represent nominals and denote relevant properties of the respective entities via different configurations of the manual articulator (Zwitserlood, 2012), specify shapes and dimensions of objects, and denote spatial relations and motion events (Sandler and Lillo-Martin, 2006). Such entities are then categorised according to their

properties into groups, e.g., flat objects, long and thin objects, two-legged beings, etc. Classifiers have been attested in all known sign languages (Sandler and Lillo-Martin, 2006), thus constituting a stable class with common general attributes, although the inventory of the particular classifiers differs from one language to another (Zwitserlood, 2012).

The categorisation of different types of classifiers has been a subject of much discussion. Earlier literature (Supalla, 1986, a.o.) had divided them into multiple classes based on various characteristics (e.g., semantics, shape, function, animacy) before currently stabilizing on two main types: whole entity classifiers and handling classifiers, based more on their function in grammar rather than their semantic properties (Zwitserlood, 2012). This internal classification is used in *Dictio* as well, and we will briefly comment on each group in the following passage.

Whole entity classifiers denote their referents in their entirety. They are more abstract and 'refer to general semantic classes rather than to visually perceived physical properties' (Sandler and Lillo-Martin, 2006, p. 77). However, various classifiers can denote a single entity, each highlighting a different relevant aspect (Zwitserlood, 2012).

An example from ČZJ is the representation of a person in a hypothetical story describing various activities of the person. We can talk, e.g., about a teacher who at first comes in the classroom (using the classifier for a person; CL:person), and later sits down at the table (represented by the classifier for two legs; CL:two-legs). The referent remains the same (the teacher), while two different classifiers describe his/her actions. Whole entity classifiers play a syntactic role of a subject. They combine with intransitive verbs that express the movement or localization of the referent in space.

On the other hand, handling classifiers utilize iconicity on a larger scale; they indicate the entity's shape as it is being held or manipulated with. The manual articulator represents itself – a hand holding the entity. This strategy gives the speaker much more room to choose among different classifiers according to the situation in the actual world (Zwitserlood, 2012). Handling classifiers play a syntactic role of an object. They combine with transitive verbs that express the manipulation with the object in space (e.g., CL:round-object).

From the morphological point of view, classifiers are bound morphemes. They must occur jointly with other expressions within so-called classifier constructions, within which they are incorporated mostly into classifier verbs, i.e., verbs denoting movement, position or existence of a referent in space or some kind of manipulation (Zwitserlood, 2012). Classifier constructions represent a very productive strategy in sign languages, and this unstable semantic and morphological status prevents them from being documented in a dictionary.

However, classifiers outside of classifier constructions (so-called classifier handshapes) can be documented. In our dictionary, classifier handshapes are registered in individual lexical entries if there is a (relatively neutral) stabilised representative form with (at least roughly) delimited meaning (e.g., via extensional definition by listing possible referents, see Section 5).

An example of such a classifier handshape from ČZJ is one of the most common, basic handshapes – an open palm with all fingers stretched out (CL:flat-object). In the grammar part of this entry, the sign is categorised into its classifier group, whole entity classifiers. Two meanings are listed: a denotation of either flat objects or four-tired vehicles. Consequently, definitions and examples of use are listed for each meaning separately; in this case a sentence where the classifier denotes a book in the former, and a car in the latter meaning.

Let us turn now to the lexical category of the size and shape specifiers (SASSes). Like classifiers, SASSes are highly iconic and describe the visual characteristics of entities. While some researchers understand the SASSes as a classifier type, we follow Zwitserlood (2012) by placing them apart. Without doubt, there are some morphological, syntactic and semantic properties shared by the domain of classifiers and SASSes, e.g., some common handshapes, a postposition to the noun and their interpretation fully dependent on the preceding noun. However, we argue for an independent lexical category of SASSes building on the following differences: firstly, SASSes carry out different syntactic functions than classifiers. Typically, they behave like modifiers (noted, e.g., in Sandler and Lillo-Martin 2006, p. 77). They specify the preceding noun's properties, unlike classifiers, which substitute the noun and have a role

resembling more that of pronouns.[5] From the morphological point of view, SASSes are independent, meaning that they are not incorporated into any verbal predicates like classifiers are. The movement in classifier constructions is always a parameter of the verb. The classifier is just the handshape. On the other hand, the movement present during the articulation of the specifier represents a proper phonological parameter of the specifier, alongside, of course, its handshape. Following the standard classification, we distinguish two types of SASSes in *Dictio*: static and tracing SASSes (e.g., Quer et al., 2019). Static SASSes do not contain the parameter of movement. Their interpretation is based on the handshape (single-handed signs; e.g., SASS:dot) or the hands' respective positions (two-handed signs; e.g., SASS:size). On the other hand, tracing SASSes do contain movement, which is crucial for their interpretation. A good example is SASS:rectangle. The resulting meaning is composed of the handshape (the distance between the open fingers), the hands' position, and the imaginary trace that the fingers leave behind while moving. We can also find several examples in which the interpretation derives merely from the movement alone (SASS:circle, a.o.).

For a specifier to be registered as a separate entry in our dictionary, the same criteria apply as those for classifiers; a stabilised representative form with a roughly delimited meaning has to be attested. That is the case of SASS:three-rows that covers two general meanings: *three scratches* or *three lines*.

As we mentioned above, there are cases of handshapes common both to the domain of classifiers and SASSes alike. Among the numerous examples in ČZJ, we note the following two: CL:flat-object is used, as was mentioned before, as a whole entity classifier for flat objects or motorized vehicles with four wheels in combination with verbs of movement and localization. The same handshape can also be used in the SASS describing an object's surface or a border of an area. Similarly, CL:thin-object is a handling classifier that represents a thin held object. The same handshape is used as a parameter of a SASS describing a long cylindrical shape of an object. Since *Dictio* organizes the entries on the basis of the formal criteria of the signs, a shared handshape between the classifiers and the SASSes constitutes one single entry. Take for

5 Although, Zwitserlood (2012) also notes the nominal and adverbial function for SASSes in American Sign Language.

example the handshape mentioned above – an open palm with all fingers stretched out (CL:flat-object): the dictionary entry with the default variant of the handshape in the headword contains five semantic fields, each of which represents a separate meaning (with their own semantic definition and examples of use). The first four explain the meaning and the use of the handshape within different classifier constructions, whereas the last field describes and exemplifies its use as a SASS.

Sometimes classifiers and specifiers undergo the process of lexicalisation. In that case, they are included in the dictionary and treated as lexemes. In these structures, the otherwise productive forms become 'frozen'. Their features (handshape, movement, place) no longer contribute morphological content to the given expression but bear only a phonological status (Sandler and Lillo-Martin, 2006). In ČZJ, we have, e.g., signs BOW (≈ ARCHERY) and TREE, which originated by lexicalising a classifier; or YOGHURT and OMELETTE, in which the motivating specifier can be recognised.

We are using a few additional criteria for distinguishing a productive classifier/SASS from a lexicalised form (other than the intuitions of native signers). First of all, we check for the meaning shift. The productive classifiers/SASSes are forms with an interpretation that is highly dependent on the preceding noun. After lexicalisation, the meaning of the form is fixed. That fact manifests itself in the redundancy of the nominal antecedent (which is obligatory for a productive classifier/SASS). And finally, the lexicalised forms originating from classifiers/SASSes acquire a mouthing that reflects the corresponding Czech translation. In contrast, a mouthing of Czech words is absent in productive classifiers/SASSes.

4 MOUTH PATTERNS ACCOMPANYING SIGNS

Non-manual components of signs defined as 'all linguistically significant elements that are not expressed by the hands' (Pfau and Quer, 2010) are equally as important for speech comprehension and production as the manual articulators. These components can take the form of head and body movements, facial expressions, or mouth patterns. In this section, we will focus on the last type and assess which mouth patterns should and should not be documented in a dictionary.

Mouth patterns are commonly divided into mouth gestures and mouthings, differing in their relationship to the surrounding spoken language. Mouthings (or spoken components) are either influenced or directly derived from the corresponding word in the surrounding spoken language; they are silent articulations of the whole word or a part of it, usually its first syllable (Pfau and Quer, 2010). Mouthings are understood as cross-modal borrowings (Sandler and Lillo-Martin, 2006; Mareš, 2011). It is possible to observe a gradual change and adaptation to the 'host' language, a process typical for borrowings observed among spoken languages as well.

In our ČZJ data, we found two situations: (i) mouthings that are a conventional part of the sign and have no apparent effect on the interpretation; (ii) mouthings that distinguish among lexemes with otherwise identical manual components. The examples of the first type are the signs NAME, COUNT or WORK. These three examples illustrate that this type of mouthing is quite variable in its form. It varies among the silent articulation of the Czech equivalent, first syllables of the Czech equivalent, or a word semantically related to it: the manual articulation of NAME is accompanied by the mouthing of the Czech equivalent for *name*. COUNT appears with two initial syllables of the Czech equivalent for the verb *to count* and the non-manual part of WORK 'to work' is formed by the mouthing of the Czech word for the noun *work*, and not the verb. Moreover, the signers' preferences vary: some signers are more precise in mouthing of the Czech words than others. Hence, several variants mentioned above are acceptable for one lexeme, depending on the speaker.

The latter type of mouthing (mouthing that changes the meaning) can be found in the field of terminology. It represents one of the ČZJ strategies for expressing expert or technical terms. Remember, e.g., SUGAR and SACCHARIDE mentioned above in Section 2 – these signs share the manual part and differ by mouthing. From the semantic point of view, we understand these examples as a specification (or narrowing) of a general meaning. We observed that this strategy is not limited to the field of science, technology or other kinds of expertise. Consider the classifier construction for pouring little particles (CL:pour), articulated without mouthing, and the signs SALT, PEPPER and SPICE. All four share the same manual part, and the interpretation of the last three is determined by the mouthing of the Czech words for *salt, pepper* and *spice*.

Let us now turn to the second type of mouth patterns. Mouth gestures (or oral components) are defined as 'all motions/positions of the mouth that are not derived from a spoken language and contribute to the speech structure' (e.g., Mareš, 2011, p. 8). They are therefore considered a native component of the given sign language.

Unlike mouthings (or at least the first type mentioned above), their form is relatively stable. Similarly to mouthings, we found two possible situations that contain the use of mouth gestures: (i) as an obligatory part of the sign (potentially a phoneme); or (ii) modifying the meaning of the sign. The first situation is exemplified by the signs HAVE/BE and WIND. Both of them are considered ungrammatical when pronounced without the mouth gesture. However, the mouth gesture does not associate with any particular semantics. On the other hand, cases of mouth gestures modifying the sign's meaning are visible in SMALL and RAIN-A-LOT. Morphologically speaking, the manual part of SMALL is the same as the manual part of the size and shape specifier expressing the size in general (SASS:size). The mouth gesture realized by the tip of the tongue coming out of the mouth modifies the sign's meaning by adding the semantic feature 'small'. Similarly, the manual part of RAIN-A-LOT shares the manual part with RAIN. The mouth gesture formed mainly by the puffy cheeks adds the aspectual modification (intensification).[6]

In order for mouth patterns to be included in *Dictio*, they need to satisfy two conditions: (i) they are obligatory for the given sign; and (ii) they do not introduce additional meaning in the sense that they do not modify the sign in terms of intensification, adjectival or adverbial modification, nor do they express the speaker's attitude (Mareš, 2011, p. 24; Pfau and Quer, 2010, p. 385). As a result, *Dictio* registers cases like NAME, COUNT, WORK, SUGAR, SACCHARIDE, HAVE/BE and WIND in separate dictionary entries. Examples like SMALL and RAIN-A-LOT are analysed as complex morphological structures (simultaneously articulated phrases) and do not appear in the headword of a dictionary entry.

Any obligatory mouth patterns are given in the grammatical description for each meaning of the lexical entry (a corresponding Czech word for mouthings

6 In fact, the mouth gesture is just a part of the complex grammatical marker of intensification. The other obligatory component is the modification of the movement (fast repetition).

and specialised symbols for different mouth gestures). In the case of a single sign (conveying a single meaning) with variable mouth patterns available, the headword is accompanied by the most neutral one. The other options are classified as variants of that sign and (in the optimal case) displayed on videos within the grammatical part of the entry.

5 STRATEGIES OF SEMANTIC DEFINITIONS

So far, we have discussed what kinds of lexemes are eligible to be listed in a dictionary, but let us now turn to each lexical entry structure with a particular focus on their definitions. The definition of a lexical entry is a crucial part of any monolingual dictionary. Thus, it is important to develop a firmly established method before beginning any lexicographic work and adhere to it throughout compiling a dictionary. This can be especially challenging in sign language dictionaries, where there is very little prior work to build on, and one may encounter several unprecedented issues. In *Dictio*, we face these challenges with the help of precisely outlined processes for forming each definition.

The Oxford Handbook of Lexicography contains an extensive chapter on the history and philosophical foundations of the concept of a dictionary definition (Hanks, 2016). However, with the lexicographic task at hand, we turned to the manuals describing current practice (e.g., Filipec, 1995) and we found two main strategies for defining the meaning – intensional and extensional definition. To define a lexeme intensionally means to specify necessary and sufficient conditions for using a given lexeme. Such intensional definition has the following structure: first, the closest general term, a hypernym, is posited to categorise the lexeme into a broader semantic class; the next step is to list necessary distinguishing properties in order to differentiate the lexeme from other elements of the same semantic class. This way, we delimit all potential occurrences while ruling out other cases.[7] A nice example of the application of this general lexicographic strategy is the definition of the sign CD-ROM,

7 Since the key to the intensional definition is to capture the internal hierarchy of a given semantic area, the work of Půlpánová (2007) on ČZJ becomes useful. In her thesis, she investigated the signs used for categorisation in ČZJ. Under *categorisation*, she understands the expression of hyper-hyponymic relations in the lexicon. Such functional signs are, e.g., TYPE and GROUP in her elicited ČZJ expression ANIMAL TYPE GROUP HOME (in the meaning of *pet*).

which is given here in glosses and can be seen under the link: CD-ROM$_a$ IX-a CL:round-object SASS:thin$_b$ IX-b SAVE DATA HOW CL:draw-circles$_a$ HAVE/BE$_a$ SASS:little-hills$_a$ 0 1 0 1.

Extensional definitions employ a different strategy. They specify an extension of a given lexeme, e.g., by naming a typical representative or several objects that are members of a specific set, requiring the reader to extract the properties common to all listed examples and compile the meaning of the lexeme from them. Such a definition can be accompanied by qualitative or circumstantial properties of a concept, e.g., size, colour, or application. An example is the semantic definition of the sign BLACK, which is given here in glosses and can be seen under the link: COLOUR IX-a LOOK-LIKE SUN GO-DOWN GET-DARK IX-b.

Between the two strategies, it is always preferred in our dictionary to use the intensional definition. However, in sporadic cases, the meaning can be determined extensionally or by combining the two, i.e., by specifying a superordinate concept followed by several examples of referents.

6 MULTIPLE MEANINGS AND SEMANTIC RELATIONS

In each lexical entry, the field of semantic relations includes both the intra-language relations (synonyms, antonyms), and the inter-language relations (translations). We will comment in detail on the first type, leaving the latter for Section 8. However, let us first consider the cases of polysemy.

In our dictionary, we follow the traditional practice of listing every meaning of a polysemous word under one lexical entry. These individual meanings differ, and therefore separate definitions, examples (and translations) are needed for them.[8]

In principle, we have encountered three types of situations: (i) a general term with multiple meanings (e.g., GERMAN, which may stand for the country or a citizen of the country); (ii) a technical term with different meanings for their respective semantic fields of use (e.g., the sign BASIS with three different

8 Currently, we are not able to differentiate between polysemy and homonymy. In the absence of an etymological dictionary of ČZJ, we register as polysemous all lexical units with more than one semantic definition.

meanings – for the field of informatics, mathematics, and chemistry); and (iii) a sign with general and technical use. If the two forms are entirely identical – including the non-manual component – two meanings can be defined with the general one listed as first. However, more often, new mouthing is added during the creation of the technical term. In this case, we understand the non-manual component as a phoneme, and we register each sign under a separate entry.[9]

6.1 Synonym-variant distinction

In *Dictio*, we register synonyms (expressions with identical or nearly identical meanings) and variants (expressions with identical meanings wholly interchangeable with the headword). A question closely tied to both is how to distinguish them and classify them according to their formal and semantic relationship to a given lexical entry.

For audio-oral languages, a dictionary entry standardly contains the citation form of a lexeme and all the variants (Čermák, 1995), e.g., the gender variants in Czech: *brambor* 'potato-masculine' vs *brambor-a* 'potato-feminine'. However, two (or more) expressions of a different word-forming nature are not considered variants but synonyms (Filipec, 1995), e.g., the Czech pair: *jazykověda* 'linguistics' (Czech origin) vs *lingvistika* 'linguistics' (foreign origin).

What seems like a simple task for spoken languages (basically, common root signals variants, different roots – synonyms) becomes a challenge for sign languages because the discussion about the definition of morphemes and lexical roots is still open-ended (Zwitserlood, 2012). The lexicographic processing of the variants in sign languages has been addressed in Johnston and Schembri's (1999) canonical work for Australian Sign Language. However, the topic of synonyms is not elaborated.

In *Dictio*, a method has been developed (and is now being applied) to distinguish variants from synonyms in ČZJ (with possible extension to other sign languages). Our approach builds on the Sandler's (2006) phonological Hand-Tier model and contributes a set of clear criteria for distinguishing variants from synonyms.

9 See Section 4 above, namely examples SUGAR and SACCHARIDE.

The Hand-Tier model (depicted in Fig. 1) groups the phonological features of a given sign into categories (parameters) and subcategories, which are hierarchically organised and partly dependent on each other. The three main parameters are (i) handshape (or hand configuration); (ii) place of articulation; and (iii) movement. The handshape parameter can be further divided into smaller sets, e.g., orientation with features like [palm] and [wrist], which helps us record, simply put, which direction the signer's hand is facing. Within the handshape parameter, a subcategory registers the features of the non-dominant hand in symmetrical signs. The non-dominant hand either copies the dominant hand in its configuration or has one of the unmarked handshapes depicted in Fig. 2. Sandler (2006, p. 161) defines such handshapes as maximally distinct, the easiest to produce, the first to be acquired by children and the most frequent in sign language production. Note that the very same phonological subcategory (the non-dominant hand) can also be found in the place parameter. It is assigned in the case of two-handed non-symmetrical signs, within which the non-dominant hand fulfils the role of a place of articulation. Moving on to the next parameter, the place of articulation is defined by features conveying the main signing areas such as [head], [trunk] or the above-mentioned non-dominant hand. However, these can be in turn combined with the features from a subcategory called setting, e.g., [high], [low] or [proximal]. Moreover, the place category features can be divided into two sets corresponding to two locations of a sign (if applicable): an initial and a final position. In this case, it is also possible to link a certain position to a certain set of handshape features that describe the sign's form in that particular position. We have seen it, e.g., in the sign RECOMMEND, where the initial position is linked to a place of articulation on the cheek with the handshape of one extended finger, and the final position is articulated on the non-dominant hand with all the fingers extended. Finishing the description of the Hand-Tier model with the last main category of movement, we can see that it is unique with respect to its complexity and partition because there is no further division into subcategories within, there are only particular phonological features like [arc], [convex] or [rep] (= repetition).

Let us now turn back to the lexicographic task at hand: distinguishing variants from synonyms in ČZJ. Researchers have marked that a pair of signs is likely

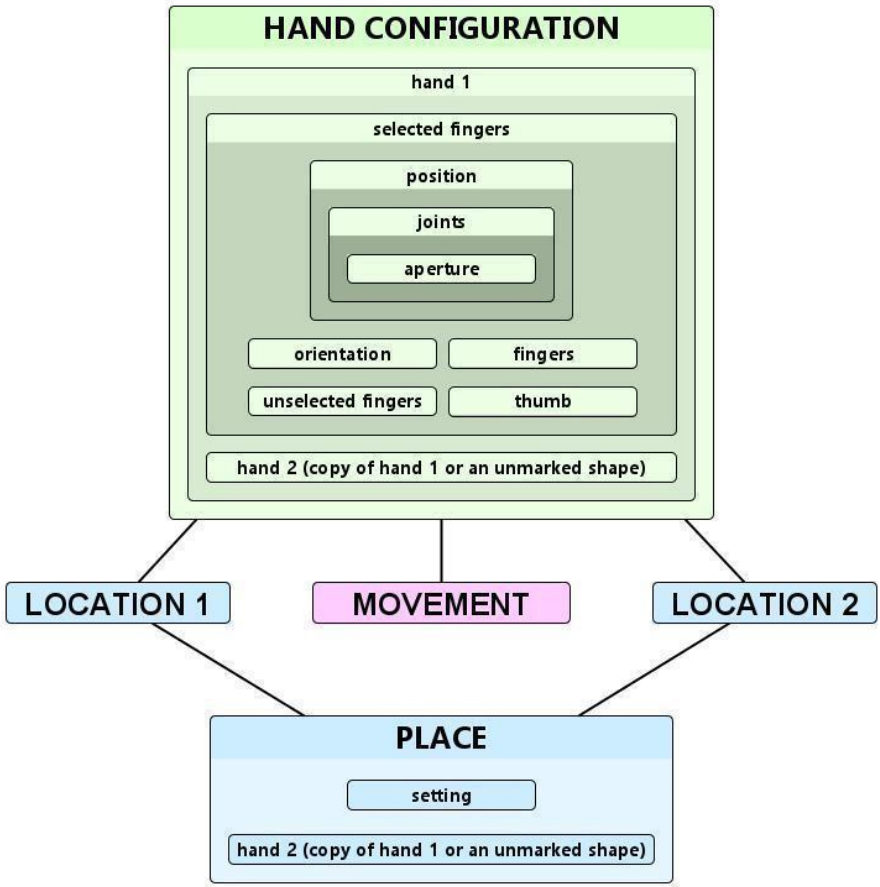


Figure 1: The Hand-Tier model.

Figure 2: Unmarked handshapes.

to be variants if they differ in just one parameter (Fenlon et al., 2015). However, the exact nature and characterization of the notion of one parameter was not specified and remained a subject of debate. This is where the Hand-Tier

model can help determine what should be understood as a difference in one or more parameters, how to account for minimal pairs of signs and, consequently, which signs should be labelled as variants and which as synonyms. With this in mind, we propose to classify a pair of lexemes as variants in case their (possibly multiple) differing phonological features fall within only one of the three main parameters described above: handshape, place of articulation or movement. In other cases, we propose to classify them as synonyms. Let us look more closely at some specific classification issues and their possible solutions based on the Hand-Tier model.

Firstly, there are pairs with only a simple difference within one parameter. Variants altering within the handshape are exemplified by PRAGUE#1 and PRAGUE#2, whereas WHY#1 and WHY#2 demonstrate variants with a different movement. BROTHER-IN-LAW#1 and BROTHER-IN-LAW#2 differ in the place of articulation, but seemingly also in orientation. However, the orientation of the dominant hand is relative. It is always evaluated with respect to the place of articulation (in our example-pair, the upper part of the trunk and the non-dominant hand). Since the dominant hand and the place of articulation are in the same configuration in both signs (contact with the ulnar side of the hand), we analyze them as having the same features for orientation and differing only in the place of articulation.

Secondly, there are slightly more complicated cases to label, namely the pairs of signs with more than one difference in their respective phonological features. It still holds that as long as those differing features belong to a single main category, the signs are analyzed as variants. Take the ČZJ signs FOURTEEN#1 and FOURTEEN#2 as examples. At first glance, they differ in the orientation of the dominant hand (towards the addressee vs the signer), i.e. a feature within the main category of handshape, and in three aspects belonging to the main category of the place of articulation: (i) the handshape of the non-dominant hand, i.e. all vs one selected finger (in other words, a fist vs an extended thumb); (ii) the orientation of the non-dominant hand, i.e. the palm towards the addressee vs facing down; and (iii) the location, i.e. where exactly does the dominant hand touch the non-dominant one. If the two signs differed in their handshapes and their places of articulation, they would be classified as synonyms. Nevertheless, as we have seen before, the orientation is relative,

so the seemingly different handshape features are predictable and follow from the location (iii). Therefore, at the phonological level, these two signs differ only within the features that belong to the one main category of the place of articulation, and as such are classified as variants.

Moving on to the higher level of contrast between two signs – from variants to synonyms – a straightforward example of synonymy is presented with the ČZJ signs KITCHEN#1 and KITCHEN#2. The lexemes differ in all three main categories, and there is no doubt that they do not share a morphological root. However, not all synonyms are so clear-cut. Examples similar to MAY#1 and MAY#2 (which represent two forms from several variants and synonyms for *May*) are challenging, since they present two morphologically related forms. Nonetheless, given that they differ in two of the three main categories, namely handshape and movement, we conclude that they should be classified as synonyms. More complicated cases, such as MAY#1 and MAY#2, show that we are working with a scale rather than a binary distinction.

Building up from the least differences to the most, we have covered which sign pairs are considered variants and which ones are classified as synonyms. We will now focus on variants and present their different types. The primary distinction lies in their phonological status: a variant can be either phonetic or phonological. A phonetic variant in a sign language is produced slightly differently from the usual, conventional manner by an individual speaker. On the other hand, a difference found in a phonological variant is rooted more deeply, and the differing parameter can even play a role in a minimal pair. However, at this level of ČZJ exploration, there is no concrete methodology of distinguishing phonetic and phonological variants that could be used systematically in the dictionary. Therefore, we consult native signers of ČZJ and their intuitions to determine which differences between two signs are considered insignificant (= phonetic variants) and which ones are treated as using a different parameter within the sign (= phonological variants). Let us demonstrate with the following example. When it comes to the various number of repeating movements within a pair of signs, the pairs with several movements each (e.g., 2 and 3 repetitions, respectively, in signs CHRISTMAS#1 and CHRISTMAS#2) were not judged as having a different phonological parameter, and are therefore registered as phonetic variants. On the other hand, when the contrast is

between a single movement and several repeated ones (e.g., in signs WHY#1 and WHY#2), it is judged as a difference in the movement parameter of the sign, and as such it is a basis for classifying the two signs as phonological variants. This conclusion is also supported by other occurrences of this contrast and its undeniable phonological merit, e.g., in the minimal pair of MORNING and CLOTHES, where it is the only differing feature. Thus, we analyse the difference between one and several movements as the phonological feature [rep] and place it in the movement category.[10]

Once we have distinguished phonetic and phonological variants, let us look more closely at the latter ones. Phonological variants can be further divided into grammatical and stylistic ones. A grammatical variant is a lexeme that is freely interchangeable with the headword and does not add any extra information about the speaker. On the other hand, a stylistic variant adds such information about, e.g., social status, regional categorisation or a generation the speaker belongs to. Thus, grammatical and stylistic variants relate to the given lexeme in all its meanings, as opposed to synonyms, as was noted above, which are linked to the individual meanings within the entry.

7 EXAMPLES OF USE

In this section, we discuss examples, namely what kinds of expressions are appropriate for an example and what guidelines need to be followed when adding an example to an entry. In the absence of a ČZJ representative corpus, the examples of use are not elicited but created by the team of native signers, forming a small corpus by itself.

It is desirable to include at least one, but ideally, several examples are listed in each lexical entry, demonstrating the use of a given lemma in different communicative situations. An example could be an expression (two or more signs), a sentence, or an utterance (several sentences) illustrating the use of the lemma and/or its variants.

The fundamental idea of examples is to portray how lexemes are used in natural language. Therefore, it is not unusual to exemplify modification where

10 The feature of [rep] is mentioned in Sandler (2006), but its exact definition and place in the model have remained unclear.

possible, such as numeral and classifier incorporation, the inflexion of directional verbs, aspectual modification, and plural and negated forms.

As an illustration of the strategy described above, consider two examples for MONTH. The first example contains a simple citation form, the second one a pluralised form with an incorporated numeral: (i) TOMORROW MONTH MAY (video under the link), (ii) SUMMER IN-THAT YEAR PERIOD$_a$ HAVE$_a$ FOUR THREE-MONTH++$_a$ 2ND$_a$ SEGMENT$_a$ IN-THAT JUNE 21TH UNTIL SEPTEMBER 22TH (video under the link).

8 TRANSLATIONS

The final section focuses on the bilingual part of our dictionary and notes some specific processes inherent to the bimodal character of *Dictio*. As was mentioned previously, *Dictio* was initially designed as a monolingual dictionary. However, as the project grew in size, more languages (spoken and sign) were added to the interface. Therefore, it became increasingly important to establish a coherent method of managing the ties among the languages and the specific entries with a translational counterpart. However, this effort still focused mostly on Czech and ČZJ, which retain their positions of the most documented languages within *Dictio*.

With a project of this size, naturally, there are many different translators among the contributors, each assigned their own respective (pair of) languages depending on their language training. Due to this dictionary's specific bimodal character, we are faced with several types of translation techniques based on the particular combination of languages in question – they can be both signed, both spoken, or it is a signed-spoken pair. In this paper, we will examine some specifics of the last type.

First let us outline two general principles concerning the translation process, which have been applied throughout the dictionary. Firstly, when linking two corresponding lexemes from different languages via translation, it is essential to target the specific meanings (if there are several to choose from) and not equate the two dictionary entries. It is a common practice that ensures, e.g., that the English polysemous word *bed* is linked to the Czech lexeme *postel* only in the meaning of 'a piece of furniture for sleeping' and not 'the bottom

of the sea, lake or river', which is conveyed by the Czech lexeme *dno*. Secondly, while finding the corresponding equivalent (sign or spoken), the translators never rely only on their knowledge of the languages they work with. That means, when they look, e.g., for the Czech translation of the English lexeme *bed*, they never work only with the headword in the dictionary entry. They are always guided by the semantic definition(s) and assign the translation that corresponds to the definition. That is why the definitions need to be construed clearly and unambiguously (and when a certain definition lacks these qualities, it needs to be revised). However, even clear and unambiguous definitions can have different translations, which are often linked among each other as synonyms.

Let us now focus in more detail on the translation process employed between a signed and a spoken language, demonstrated by some tricky examples from Czech and ČZJ. It proved useful to provide the editors with the following guidelines concerning the use of mouthing. In ČZJ, there are several situations where only the mouth pattern differentiates between several signs with identical manual components. It is important to be guided by the mouth pattern while translating these signs into a spoken language. As we have shown before (in Section 4), this is useful especially when linking a set of morphologically and semantically related ČZJ signs like SALT, PEPPER and SPICE to their respective Czech translations. Translators tend to understand such sets as one sign language lexeme with several options of mouthing. However, in *Dictio*, each mouthing determines one dictionary entry. Hence the Czech translations should be distributed accordingly.

At the same time, relying solely on the non-manual component of the sign will not suffice and can be misleading. In some cases, the mouthing and the sign translation differ, although they can be related. Take BECAUSE in ČZJ as an example: the sign has a mandatory mouthing of the Czech word *důvod* 'a reason'. However, the entry contains two meanings, one of them is translated into Czech as *důvod* 'a reason' and the other as *protože* 'because'. Note that even in the second meaning, the sign is still accompanied by the silent articulation of the Czech word *důvod* 'a reason'.

Until now, we talked about cases that represent linking two dictionary entries, although at the level of individual meaning: for example, the first meaning

of ČZJ SALT is translated as Czech *sůl* in its first meaning ('white material, in powder or chunks, used to prepare dishes'). However, some entries need a translation that does not qualify as a dictionary entry. Below, we describe two types of situations with one thing in common: the ČZJ lexeme fulfils the requirement for a dictionary entry (see Section 2 above), but the corresponding Czech translation does not.

The first type of examples can be illustrated by the signs with numeral incorporation, like LAST-WEEK. Morphologically speaking, the sign consists of a handshape for the numeral SEVEN, and a movement of the sign PAST. Compositionally, we could read the meaning as 'seven days ago'. However, the Czech translation (*minulý týden* 'last week') is a common noun phrase with an adjective modifier (a collocation, from the lexicographic point of view). In general, those are the situations, in which the signed member of the pair is a single lexical unit (and as such is recorded in the dictionary), while the translation into the spoken language is a common syntactic phrase (which is not recorded in the dictionary). Apart from numeral incorporation, we might name examples like CHAINSAW (*motorová pila* in Czech) or AT-NOON (*v poledne* in Czech).

The second type of examples is represented with the ČZJ sign NOT-HAVE/BE, a suppletive negative form for HAVE/BE. While the Czech translation for the latter is listed as a dictionary entry (*mít* 'to have', *být* 'to be'), the irregular ČZJ form is translated by a regular Czech form (*nemít* 'not to have' and *nebýt* 'not to be'). Naturally, the regular negative forms of verbs are not listed as dictionary entries. They are produced by a regular word-forming process of adding a negative prefix *ne-* 'not'. The technical solution in *Dictio* is to provide the Czech translation in the form of a plain text, that means, without an interactive link to a corresponding semantic equivalent in the Czech part of the dictionary.

9 CONCLUSION

Dictio is a work in progress, similar to any other dictionary trying to capture and describe natural language. However, even now, in its developmental stages, it already serves multiple functions. *Dictio* has been used in ČZJ courses, linguistic education, and by translators, providing valuable examples of signs

and their categorisation. Moreover, it represents the most extensive ČZJ material collection to date, containing both the individual signs and the utterances elicited from native signers.

This paper presented several methods implemented during the creation of the first Czech Sign Language online dictionary. We introduced the formal and semantic criteria for lemmatisation and classified the headwords into four groups: a simple lexeme, a compound, a derivative, and a set phrase. We established the place of the classifiers and the size and shape specifiers in the dictionary by applying our criteria consistently: once a stable form can be associated with a conventional meaning, it qualifies for a dictionary entry. We argued for an independent category of size and shape specifiers, apart from the classifiers, by showing their different grammatical properties. We explored several functions of mouthing and mouth gestures and proposed the criteria for this type of non-manuals in the headword: obligatoriness and absence of a grammatical or pragmatic modification function. We introduced the two types of semantic definitions (intensional and extensional) and specified the appropriate use for each of them. We discussed multiple meanings and semantic relations and showed the complexity of variant-synonym classification in sign languages. We elaborated the minimal difference requirement for the variant pairs using the phonological Hand-Tier model. We offered a guideline to create sound examples of use by highlighting the variability of the headword. Finally, we commented on translating between spoken and sign languages and discussed various types of sign-spoken lexeme pairs resulting from this process.

Dictio poses many lexicographic challenges, and solving them brings us closer to understanding the nature of Czech Sign Language (among others) and its phenomena. One of the most challenging topics that will be addressed in the near future is the assignment of lexical categories to the signs.

Acknowledgments

We would like to acknowledge *Dictio* for providing us with all video examples given in the text and Appendix.

REFERENCES

Battison, R. (1978). *Lexical borrowing in American sign language*. Linstok Press, Silver Spring.

Čermák, F. (1995). Paradigmatika a syntagmatika slovníku: možnosti a výhledy. In F. Čermák & R. Blatná (Eds.), *Manuál lexikografie. Jinočany: H&H, 1,* 90–115.

Dictio: Multilingual Online Dictionary. (2020). Brno: Masaryk University. Retrieved from https://www.dictio.info

Le Dico Elix – Le dictionnaire vivant en langue des signes française (LSF). (2020). Retrieved from https://dico.elix-lsf.fr/

e-LIS: Electronic bilingual dictionary Italian Sign Language – Italian. (2020). Retrieved from http://elis.eurac.edu/index_en.html

Fenlon, J., Cormier, K., & Schembri, A. (2015). Building BSL SignBank: The lemma dilemma revisited. *International Journal of Lexicography*, *28*(2), 169–206. Retrieved from https://www.researchgate.net/publication/276164152_Building_BSL_SignBank_The_lemma_dilemma_revisited

Filipec, J. (1995). Teorie a praxe jednojazyčného slovníku výkladového. In: F. Čermák in R. Blatná (Eds.), *Manuál lexikografie. Jinočany: H&H*, *1,* 14–49.

Hanks, P. (2016). Definition. In P. Durkin (Ed.), *The Oxford handbook of lexicography*. doi: 10.1093/oxfordhb/9780199691630.001.0001

Johnston, T., & Schembri, A. C. (1999). On defining lexeme in a signed language. *Sign language & linguistics, 2(2)*, 115–185.

Kristoffersen, J. H., & Troelsgård, T. (2012). The electronic lexicographical treatment of sign languages: The Danish Sign Language Dictionary. In S. Granger in M. Paquot (Eds.), *Electronic Lexicography*. Oxford University Press.

Langer, J., Ptáček, V., & Dvořák, K. (2005). *Znaková zásoba českého znakového jazyka k rozšiřujícímu studiu surdopedie se zaměřením na znakový jazyk (I, II)*. Olomouc: Palacký University.

Langer, J., Ptáček, V., & Dvořák, K. (2005a). *Znaková zásoba českého znakového jazyka k rozšiřujícímu studiu surdopedie se zaměřením na znakový jazyk (III, IV)*. Olomouc: Palacký University.

Langer, J., & Kukolová, P. (2008). *Slovník vybraných pojmů znakového jazyka pro oblast biologie člověka a zdravovědy*. Praha: Fortuna.

Mandel, M. (1981). *Phonotactics and morphophonology in American Sign Language*. PhD dissertation, University of California. Retrieved from https://escholarship.org/content/qt90v1j5kx/qt90v1j5kx.pdf

Mareš, J. (2011). Orální komponenty v českém znakovém jazyce. Bc. thesis, Charles University. Retrieved from http://hdl.handle.net/20.500.11956/50230

McKee, R., Vale, M., Hanks, P., & de Schryver, G. M. (2017). Sign language lexicography. *International Handbook of Modern Lexis and Lexicography*. Berlin/Heidelberg: Springer. Retrieved from https://www.researchgate.net/publication/319881867

Mladová, P. (2009). Kompozita v českém znakovém jazyce. Bc. thesis, Charles University.

Pfau, R., & Quer, J. (2010). Nonmanuals: their grammatical and prosodic roles. In D. Brentari (Ed.), *Sign Languages* (pp. 381–402). New York: Cambridge University Press.

Potměšil, M. (2002). *Všeobecný slovník českého znakového jazyka, A–N. Praha: Fortuna.*

Potměšil, M. (2004). *Všeobecný slovník českého znakového jazyka, O–Ž. Praha: Fortuna.*

Potměšil, M. (2004a). *Všeobecný slovník českého znakového jazyka, O–Ž – doplněk. Praha: Fortuna.*

Půlpánová, L. (2007). *Kategorizace v českém znakovém jazyce*. Mgr. thesis, Charles University. Retrieved from http://hdl.handle.net/20.500.11956/13566

Quer, J., Cecchetto, C., & Donati, C. (2017). *SignGram Blueprint: A guide to sign language grammar writing* (p. 896). Berlin: De Gruyter. Retrieved from https://www.researchgate.net/publication/321962244_SignGram_Blueprint_A_Guide_to_Sign_Language_Grammar_Writing

Sandler, W. (1989). Phonological Representation of the Sign: Linearity and Non-linearity in American Sign Language. Dordrecht: Foris.

Sandler, W. (2006). Phonology. In W. Sandler & D. Lillo-Martin (Eds.), *Sign Language and Linguistic Universals*, *1,* 111–278. New York: Cambridge University Press.

Sandler, W. (2006). Entering the lexicon: lexicalization, backformation and cross-modal borrowing. In W. Sandler & D. Lillo-Martin (Eds.), *Sign Language and Linguistic Universals*, *1*, 94–107. New York: Cambridge University Press.

Sandler, W., & Lillo-Martin, D. (2006). Classifier constructions. In W. Sandler & D. Lillo-Martin (Eds.), *Sign Language and Linguistic Universals*, *1*, 76–93. New York: Cambridge University Press.

Stokoe, W. C. (1960/2005). Sign language structure: An outline of the visual communication systems of the American deaf. *Journal of Deaf Studies and Deaf Education*, *10*(1), 3–37.

Supalla, T. (1986). The classifier system in American sign language. *Noun classes and categorization, 7*, 181–214.

Zeshan, U. (2002). *Towards a notion of 'word' in sign languages* (pp. 153–179). Cambridge: Cambridge University Press.

Zeshan, U. (2004). Interrogative and Negative Construction in Sign Languages. *Language*, *80*(1), 7–39.

Zwitserlood, I. (2010). Sign Language Lexicography in the Early 21st Century and a Recently Published Dictionary of Sign Language of the Netherlands. *International Journal of Lexicography*, *23*(4), 443–476.

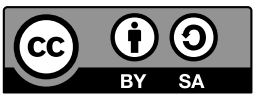
Zwitserlood, I. (2012). Classifiers. In R. Pfau, M. Steinbach & B. Woll (Eds.), *Sign Language. An International Handbook* (pp. 158–186). Berlin/ Boston: De Gruyter Mouton. Retrieved from https://www.researchgate.net/publication/291214641_Classifiers

LEKSIKOGRAFIJA ZNAKOVNEGA JEZIKA: ŠTUDIJA PRIMERA SPLETNEGA SLOVARJA

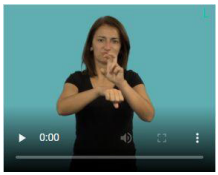
V prispevku so predstavljeni tako nekateri izzivi leksikografije znakovnih jezikov kot rešitve za te izzive, ki so bile rabljene v prvem spletnem slovarju češkega znakovnega jezika (ČZJ), ki je del platforme *Dictio*, razvite na Masarykovi univerzi v Brnu na Češkem. V prvem razdelku prispevka je predstavljena platformo *Dictio*, govorjeni in znakovni jeziki, ki so vključeni v to bazo podatkov, število javnih vnosov in temelji te baze. Kratko je povzeto metodološko ozadje projekta, izpostavljena pa je edinstvena lastnost slovarja – pomenske definicije in primeri rabe v češkem znakovnem jeziku. V drugem razdelku so kriteriji lematizacije aplicirani na gradivo iz znakovnega jezika, definirani pa so tudi jezikoslovni kriteriji za slovarska gesla. Predstavljena je tipologija kandidatov za slovarski vnos, te tudi kratko komentiramo. Gre za preproste lekseme, zloženke, izpeljanke, zveze, deiktične izraze, opise in kolokacije. S pomočjo množice pomenskih in morfoloških kriterijev identificiramo prve štiri kot izraze, ki so lahko vključeni v slovar.V tretjem razdelku pojasnimo leksikografski proces dveh prominentnih leksikalnih kategorij znakovnega jezika, tj. klasifikatorjev in določil velikosti in oblike. Ohranimo standardni klasifikaciji klasifikatorjev (celotna entiteta ali klasifikator držanja) ter določil velikosti in oblike (statična in pomična določila) ter podamo argumente za ločevanje kategorij klasifikatorjev od kategorij določil. V četrtem razdelku opišemo dva tipa prvin, ki morata biti poleg kretenj odražena v slovarju: oralizacija in premikanje ust. S pomočjo primerov pojasnimo njuno funkcijo ter pokažemo, da so v slovarju zabeležene le tiste prvine, ki so obvezne in ne delujejo kot modifikatorji. V petem razdelku pojasnimo koncept dveh tipov pomenskih definicij: intenzijske in ekstenzijske definicije. Podamo primere obeh in prikažemo argumente, ki govorijo v prid prvemu tipu definicij. V razdelku 6 podamo prve primere večpomenskosti. Predstavimo tipologijo večpomenskih leksemov v ČZJ in pojasnimo njihovo organizacijo v slovarskem geslu. Nato se posvetimo k sopomenskosti. Pojasnimo razliko med sopomenko in različico v znakovnem jeziku ter predstavimo natančno metodo za razlikovanje med tema skupinama, pri čemer gradimo na modelu »hand-tier« (Sandler, 2006). V sedmem razdelku podamo preprosta navodila za oblikovanje pravih primerov rabe v znakovnem jeziku. Razdelek 8 je namenjen procesu prevajanja, in sicer prevajanja iz znakovnega v govorjeni jezik. Razpravljamo o pomenu pomenskih definicij in prvin, ki niso kretnje. Kratko komentiramo tehnične rešitve za asimetrične pare, v katerih eden od delov prevoda ni naveden kot

slovarsko geslo. Prispevek zaključimos povzetkom vlog, ki jih v skupnosti uporabnikov češkega znakovnega jezika igra platforma *Dictio*.

Ključne besede: znakovni jezik, leksikografija, slovar, metodologija



APPENDIX 1: LIST OF MENTIONED *DICTIO* ENTRIES

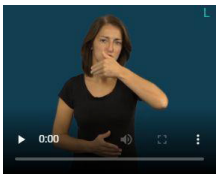


ANIMAL

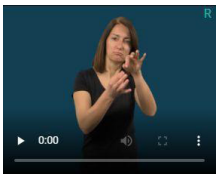
AT-NOON

BASIS

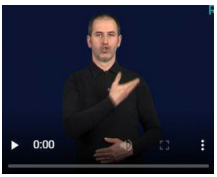
BECAUSE



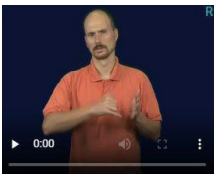
BLACK



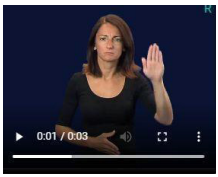
BOW/ARCHERY



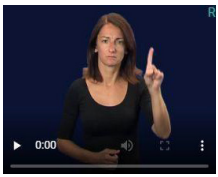
BROTHER-IN-LAW#1



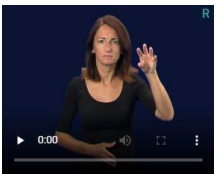
CD-ROM



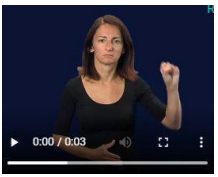
CL:flat-object



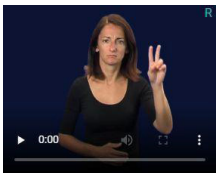
CL:person



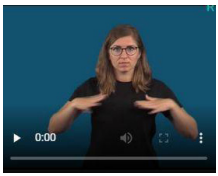
CL:round-object



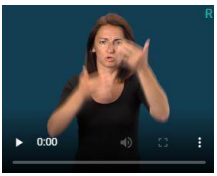
CL:thin-object



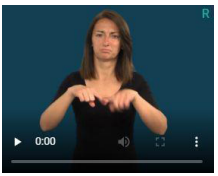
CL:two-legs



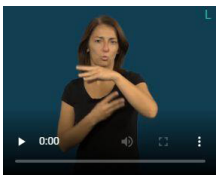
CLOTHES



COUNT

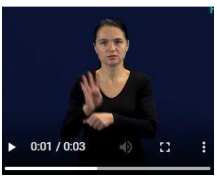


DEFECT



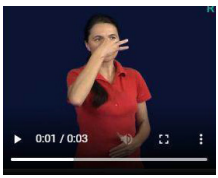
FEBRUARY

FLOWER^SPRING

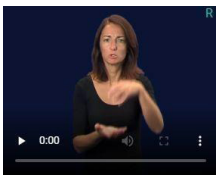


FOURTEEN#1

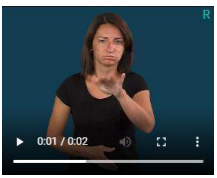
FOURTEEN#2



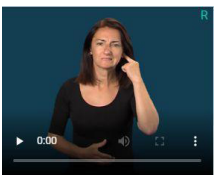
GERMAN/GERMANY



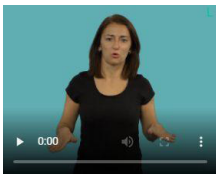
GROUP



HAVE/BE

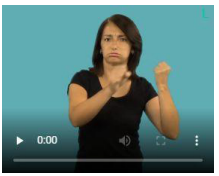


HEAR



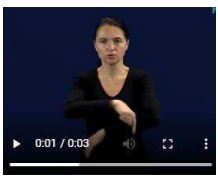
HOME

HOUR

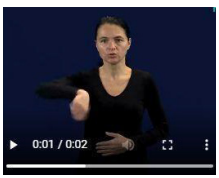


CHAINSAW

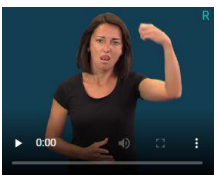
I



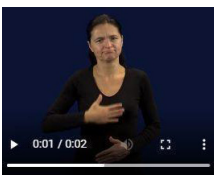
KITCHEN#1



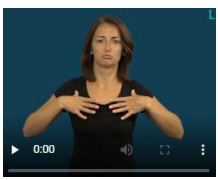
KITCHEN#2



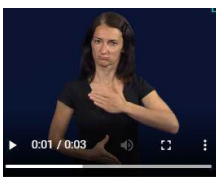
LAMP



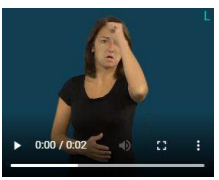
LIKE



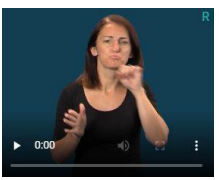
MORNING



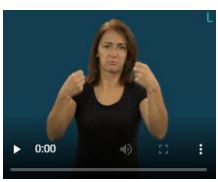
MY



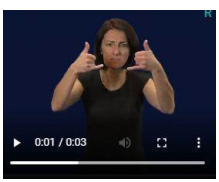
NAME



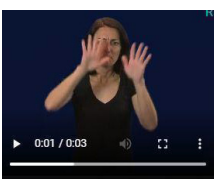
POST-OFFICE



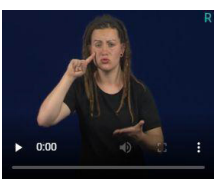
PRAGUE#1



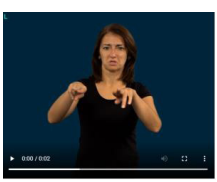
PRAGUE#2

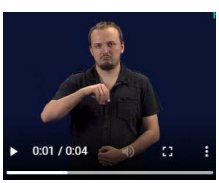


RAIN

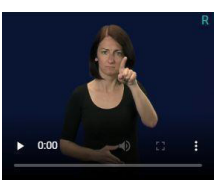


RECOMMEND

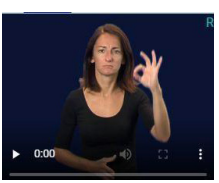

RETURN



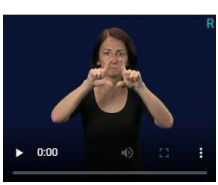
SALT/PEPPER/SPICE /CL:pour



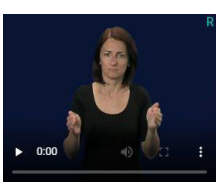
SASS:circle



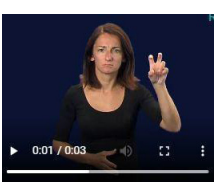
SASS:dot



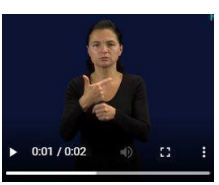
SASS:rectangle



SASS:size



SASS:three-rows

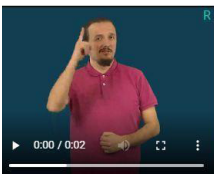


SUGAR/SACCHARIDE

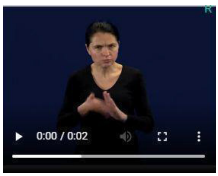
SUN^GLASSES

TREE

TYPE

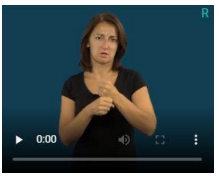


UNIVERSITY

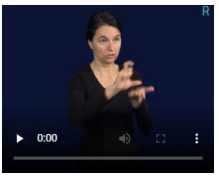


WHY#1

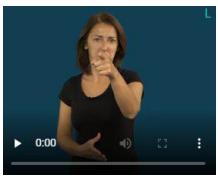
WIND



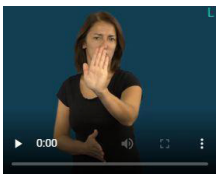
WORK



YOGHURT



YOU



YOUR

APPENDIX 2: NOTATIONAL CONVENTIONS

SIGN	A gloss of a lexical sign is given in small caps.
$SIGN_a$	A letter subscript indicates the expression is signed in locus *a* (= a position in the signing space). Locus names (*a, b, c*...) are assigned from the signer's right to left.
$_aSIGN_b$	Two letter subscripts indicate a sign signed from locus *a* to locus *b*. Loci 1 and 2 correspond to the position of the signer and addressee, respectively.
INDEX-a/IX-a	A pointing sign towards the locus *a*.
SIGN-SIGN	Two hyphenated expressions indicate that more than one word is required to gloss a single sign.
S-I-G-N	Small caps letters separated by hyphens indicate fingerspelled words.
SIGN^SIGN	Two signs joined by a caret indicate compounding or a sign plus affix combination.
SIGN++	Two pluses indicate sign reduplication.
SIGN#1	A number after a hashtag indicates a variant of a sign.
CL:c 'x'	A classifier is indicated using CL, followed by its specification/description, and its meaning in single quotes.
SASS:sass 'x'	A shape and size specifier is indicated using SASS, followed by its specification/description, and its meaning in single quotes.

CONVERTING RAW TRANSCRIPTS INTO AN ANNOTATED AND TURN-ALIGNED TEI-XML CORPUS: THE EXAMPLE OF THE CORPUS OF SERBIAN FORMS OF ADDRESS

Dolores LEMMENMEIER-BATINIĆ
Department of Slavonic Languages and Literatures, University of Zurich

Lemmenmeier-Batinić, D. (2021): Converting raw transcripts into an annotated and turn-aligned TEI-XML corpus: the example of the Corpus of Serbian Forms of Address. Slovenščina 2.0, 9(1): 123–144.

DOI: https://doi.org/10.4312/slo2.0.2021.1.123-144

This paper describes the procedure of building a TEI-XML corpus of spoken Serbian starting from raw transcripts. The corpus consists of semi–structured interviews, which were gathered with the aim of investigating forms of address in Serbian. The interviews were thoroughly transcribed according to GAT transcribing conventions. However, the transcription was carried out without tools that would control the validity of the GAT syntax, or align the transcript with the audio records. In order to offer this resource to a broader audience, we resolved the inconsistencies in the original transcripts, normalised the semi-orthographic transcriptions and converted the corpus into a TEI-format for transcriptions of speech. Further, we enriched the corpus by tagging and lemmatising the data. Lastly, we aligned the corpus turns to the corresponding audio segments by using a force-alignment tool. In addition to presenting the main steps involved in converting the corpus to the XML-format, this paper also discusses current challenges in the processing of spoken data, and the implications of data re-use regarding transcriptions of speech. This corpus can be used for studying Serbian from the perspective of interactional linguistics, for investigating morphosyntax, grammar, lexicon and phonetics of spoken Serbian, for studying disfluencies, as well as for testing models for automatic speech recognition and forced alignment. The corpus is freely available for research purposes.

Keywords: spoken Serbian, language biographical interviews, forms of address, data re-usability

1 INTRODUCTION

Serbian has long been an under-resourced language despite the long tradition of work on language corpora in the "West Balkans" (see Dobrić, 2012). Up until the past decade, there have been only two notable corpora of Serbian: Corpus of Serbian Language (Kostić, 2003) and SrpKor Corpus of Contemporary Serbian Language (Krstev and Vitas, 2005; Popović, 2010; Utvić, 2011). In the past decade, several corpora have been created in order to amend the lack of resources regarding the written data (Ljubešić and Klubička, 2014; Ljubešić et al., 2016; Miličević and Ljubešić, 2016; Batanović et al., 2018). However, although there has been a global increase in popularity of spoken language resources and tools (see Batinić et al., to appear), Serbian still lacks spoken language corpora. Considerable advances have been made regarding the Torlak dialect (Vuković, 2021), resources for automatic speech recognition and synthesis (Delić et al., 2013; Suzić et al., 2014), and specialised spoken corpora, such as the SCECL[1] corpus on early child language (Anđelković et al., 2001) and SrMaCo[2] corpus on language of Serbian minority in Hungary.

Creating corpora of spoken language demands not only field access in order to obtain recordings of spoken language data, but also intensive manual work to transcribe them. These two steps are usually the most time-consuming in the corpus creation, and prevent spoken corpora from growing at the same pace as written corpora (see Schmidt, 2016, pp. 127–128). Therefore, in order to address the lack of spoken language resources, it is convenient to start compiling spoken corpora from existing recordings and transcriptions. This paper presents a compilation of a corpus of Serbian forms of address, which has been created from an existing collection of interviews gathered for investigating Serbian forms of address (Ulrich, 2018). The interviewees were asked about forms (expressions) they use to address their relatives, friends, colleagues, neighbours, etc. The corpus contains 19 transcriptions of interviews amounting to a total of 171,552 tokens (19,5 hours of speech).

1 *Serbian Corpus of Early Child Language (SCECL)*. Available at: https://sla.talkbank.org/TBB/childes/Slavic/Serbian/SCECL.

2 *Spoken corpus of the Serbian minority in Hungary (SrMaCo)*. Available at: http://spokencorpus.eu/cms/bosco-2/.

While the first steps of the corpus compilation have been presented in Lemmenmeier-Batinić et al. (2020), this paper discusses them in more detail, and shows some additional steps that have been made since, such as evaluation of linguistic annotations, and integration of forced alignment. It also discusses the implications of data re-use for linguistic research, and encourages further sharing of high-quality transcripts of speech, while at the same time stressing the importance of using current transcription tools for facilitating not only one's own work, but also the future usability of collected material.

2 CORPUS OF SERBIAN FORMS OF ADDRESS

2.1 Recordings and metadata

The source data consists of transcriptions and audio-files of interviews with 19 participants (9 female, 10 male). The topic of the interviews are Serbian expressions that are used to address other people. The interview guidelines have four main parts: in the first part, the interviewer asks questions about forms of address interviewees use to address family members, friends, neighbors, colleagues, etc. In the second part, questions are asked about forms of address for people that have some particular profession or function. In the third part, the interviewer lists certain forms of address, and asks if participants use them. In the fourth part of the questionnaire, interviewees have the opportunity to elaborate on the topic of their attitudes and assessments about particular forms of address.[3] The interviews were recorded during 2008 and 2009. The interviewer (female) was aged 27 at the time of recording. With the exception of the interviewer, who acquired Serbian as a foreign language, all the interviewees are native speakers of Serbian. At the time of recording, participants were aged 27 to 64 years. Most of them resided in Belgrade and Niš, and had a university degree (see Table 1).

Most interviews were held in private homes. However, some of them were recorded in bars, restaurants or shopping malls, which often resulted in lower quality of audio-recordings. The interviews last about 61 minutes in average, and contain 171,552 tokens (10,045 types).[4] An overview over the size of each transcript in tokens and minutes is given in Table 2.

3 See Ulrich (2018, pp. 338–341) for detailed interview guidelines.

4 The token count includes full and truncated words.

Table 1: *Speaker metadata*

Id	Sex	Age	Origin	Residency	Education
S	f	27	CH	Zurich	university
F1	f	28	Belgrade	Belgrade	technical college
F2	f	27	Belgrade	Zurich	university student
F3	f	27	Niš	Niš, Kotor	university
F4	f	44	Lazarevo	Belgrade	university
F5	f	58	Belgrade	Belgrade	university
F6	f	55	Niš	Niš	university
F7	f	55	Skopje	Niš	high school
F8	f	64	Leskovac	Niš	high school
F9	f	60	Pirot	Niš	technical college
M1	m	28	Niš	Niš	university
M2	m	27	Niš	Niš, Kotor	university
M3	m	29	Niš	Niš	university
M4	m	27	Užice	Belgrade	university student
M5	m	33	Belgrade	Belgrade	university
M6	m	27	Belgrade	Belgrade	high school
M7	m	38	Belgrade	Belgrade	university
M8	m	44	Belgrade	Belgrade	high school
M9	m	54	Niš	Niš	university
M10	m	61	Belgrade	Belgrade	university

Table 2: *Transcript length and duration*

Transcript	Token count	Duration
F1	12,784	01:24:53
F2	8,463	01:12:12
F3	9,135	00:55:25
F4	5,995	00:38:26
F5	9,159	00:55:12
F6	7,365	00:40:40
F7	6,693	00:48:19
F8	5,408	00:44:21
F9	13,681	01:29:55
M1	9,140	00:58:33
M2	11,653	01:20:21

Transcript	Token count	Duration
M3	7,283	00:51:08
M4	10,445	01:11:46
M5	11,762	01:07:43
M6	9,836	01:18:54
M7	9,774	01:05:27
M8	6,485	00:45:44
M9	5,260	00:36:59
M10	11,231	01:29:12
Total	171,552	19:35:10

The participants originally agreed to their data being used for the project of investigating Serbian forms of address by Ulrich (2018). For securing the possibility of data re-use for other research projects as well, interviewees were retraced in 2020/2021 and they were asked to sign a data privacy agreement stating that their interviews can be used for research purposes.[5] The audio files were cut in order to match exactly with the start and the end of the corresponding transcripts prior to any other processing.

2.2 Transcripts

Although the aim of the data collection was a content analysis (see Ulrich, 2018), all the interviews were thoroughly transcribed following the GAT transcribing conventions (Selting et al., 1998, 2009), which were originally developed for purposes of conversation analysis and interactional linguistics. GAT differentiates between three levels of transcription granularity: *minimal* (Selting et al., 2009), *basic* and *fine-grained* (Selting et al., 1998, 2009). Ulrich's (2018) transcripts contain most features of basic transcripts (annotation of pauses, breathing, incidents, overlaps, vocal length, etc.), while some other features are omitted (such as segmenting turns in intonational phrases, and annotation of pitch movement) or sporadically applied (like focus accent annotation). Some features of fine-grained transcription conventions were used, out of which some were consistently applied in all transcripts, such as the annotation of pace and loudness (<<p>...>), and other were used only

5 Three participants could not be retraced and two of them had passed away. We do not share the audio interviews of these participants.

occasionally, such as the annotation of pitch jumps (↑). Overlaps were marked with square brackets, as proposed in GAT, but they were not vertically aligned, so it is not always possible to reconstruct which segments overlap with which. An excerpt from one of the transcripts is given in Example 1.

Example 1: *Excerpt from an original transcript (transcript id: F8)*[6]

S: i: e: i samo (--) kako (--) e:: (.) kako VAs oslovljavaju na pijaci (-) kad vi: kupujete

K: ko kako (.) ko gospođo (-) ko (--) e: seko ko: (-) ženo (-) ko kako (.) kom kako <<lachend> padne napamet> ((lacht))

S: ↑e: da: (-) <<p> pa da (.) za= (-) primetila sam na pijaci (.) ima naj ((lacht)) zanimljivije [((lacht))]

K: [da (--) pa] pa pijaca je uopšte najzanimljivija

S: jeste

K: najzanimljivija i: (-) .h i ovo= ove (-) emisije kad gledamo preko televizije kad

S: aha

K: uglavnom se posećuju PIjace jer je tu nešto najinteresantnije [((lacht))]

6 For reasons of clarity, some annotations are omitted in the English translation:
S: and e: and just (--) how (-) e:: (.) how do people address you at the market (-) when you are buying
K: it depends who (.) some say misses (-) some (--) e: sister some (-) women (-) it depends who (.) it depends how <laughing> it occurs to them> ((laughs))
S: oh yes (-) <<p> well yes (.) I noticed it's most ((laughs)) interesting at the market ((laughs))
K: yes (--) well the market is the most interesting of it all
S: yes it is
K: the most interesting and: (-) .h and this= those (-) shows we watch on television when
S: aha
K: they mostly visit the markets because there is something most interesting there ((laughs))
S: e (-) ((laughs)) yes (-) ((laughing)) exactly
[…]
S: mhm mhm (<<p> mhm) good> .hh e: so how would you (-) e: address a taxi driver for example
K: (2.5s) m exclusively with the polite form
S. mhm (--) mhm
K: exclusively with the polite form (-) .h I don't' use <<rall> sir> to address

S: [e (-) ((lacht)) da] (-) ((lächelnd)) baš tako

[...]

S: mhm mhm (<<p> mhm) dobr↑o> .hh e: onda kako biste (-) e: oslovljavali vozača taksija naprimer

K: (2.5s) m isključivo sa vi

S: mhm (--) mhm

K: isključivo sa vi (-) .h <<rall> ne oslov>ljava= o=oslovljavam <<rall> gospodine>

The transcripts are very consistent, despite the fact that all interviews were transcribed without using any transcription software that would control the GAT syntax, and that the transcripts were originally not meant for re-distribution to a larger audience. However, with such a large amount of manual work, inconsistencies and typing errors are inevitable. For instance, different types of parenthesis ("(", "((", and "{") were occasionally used to annotate same information. Metalinguistic annotations were mostly written in German ("lacht" 'laughs'), but sometimes also in Serbian ("smeje se"). Rarely, symbols that are not proposed in GAT were used (* - <). The symbol "=" was, amongst other uses, frequently used for marking truncated (incomplete) words, which differs from its description in GAT, where it is proposed for marking fast continuation of new segments ("*latching*", Selting et al., 2009, p. 392; Selting et al., 1998, p. 31), or for marking contractions ("und=äh") and two syllabic reception signals such as "hm=hm" (only in the first GAT version, see Selting et al., 1998, p. 31). However, the frequent annotation of truncated words with "=" provided very valuable information, and was kept for further processing. Despite some inconsistencies, the transcriptions were accurate enough to permit a conversion into a standardised format such as XML, while including (most) annotations in the markup. Interviews were originally transcribed in Microsoft Word, and were converted to plain text files in order to allow for further data processing. The original files had a simple structure (one line for each speaker turn) and transporting them to plain text required no additional editing.

3 CONVERTING THE CORPUS TO TEI-XML

3.1 Preprocessing

Prior to XML-conversion, annotations of incidents, gaps, comments, pace, loudness, ambiguous segments ("je/i" 'it is/and') and occurrences of annotations with the equals sign ("=") were extracted, corrected, and made consistent. For instance, since the use of parentheses was not always consistent, all the parentheses were checked and marked with the corresponding label in the intermediate step (see Table 3).

Table 3: *Categorising comments in the preprocessing step (excerpt)*

Original annotation	Changes (intermediate step)
{Auslassung 14:58-15:53} *omission 14:58-15:53*	((gap:extent: 55s))
{Telefon klingelt} *the phone is ringing*	((incident: zvoni telefon))
((klopft auf den Tisch)) *knocks on the table*	((incident: kuca o sto))

In total, 707 unique annotations were checked, out of which 665 have been changed, and stored into intermediate (clean) transcript text files. Most corrections were related to the use of the equals sign, metalinguistic comments, and annotations of pace and speed that were set in the middle of words, which had to be reconstructed (for instance: "mla<<lachend> đi>" was changed to "mlađi" 'younger'; "po<imenu" was changed to "po imenu" 'by name'). The metalinguistic comments were translated into Serbian (see Table 3). Although they had to be adjusted in the preprocessing step, features of fine-grained transcription were not considered in further processing, because they were either seldom used in the transcripts (annotation of pitch jumps and focus accents) or because their conversion to TEI required prioritisation of overlapping annotations (in cases like "<<rall/p>…>"), and annotation of shifts on a sub-word level (like in "mla<<lachend> đi>" 'younger'). As shown in Section 3.3, we opted to keep the segmentation at word-level, and to provide a structure that makes XML-search and parsing of words as basic entities an undemanding task.

3.2 Normalisation

The interviews were transcribed based on their phonetic realisation, hence not always according to orthographic rules. In order to provide a corpus with normalised (standard) variants as well, tokens that did not occur in the Serbian lexicon srLex[7] (Ljubešić et al., 2016) were extracted and manually checked. Out of 387 types that were not present in srLex, 119 were correct (mostly rare words, proper names, or colloquialisms). The remaining 268 had to be normalised. Two types of normalised tokens were stored for further processing: corrections of transcriber's orthographic or typing errors (ex."označavaju" for "osnačavaju" 'they mark'), and standard variants of spoken forms (ex. "hoćete" for "oćete" 'you want'). The normalisation affected 4,055 tokens (2.4%) and 972 types (9.7%) in the corpus.

3.3 Marking up the corpus with TEI-annotations

Preprocessed transcripts have been converted into XML format following TEI conventions for transcriptions of speech.[8] Transcripts were segmented in speaker turns (<u>), and each turn was further segmented into full words: <w>, truncated words: <del>, unclear segments: <unclear>, gaps: <gap>, incidents: <incident>, vocalised non-lexical elements: <vocal>, and pauses: <pause>. Words that have been normalised to standard forms are stored in the @norm attribute. The original orthographic or transcription mistakes are stored as @orig. In addition to lemmatised and normalised forms, universal part-of-speech tags (@pos)[9] and MULTEXT-East Serbo-Croatian morphosyntactic specifications (@ana)[10] are provided (see Section 3.4). The attributes @start and @end point to the intervals in the audio-recordings defined in the <timeline> element (see Section 3.5).

7 *Inflectional lexicon srLex 1.3*. Available at: Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1233.

8 *TEI Guidelines Version 4.2.1 (Transcriptions of Speech)*. Available at: https://tei-c.org/release/doc/tei-p5-doc/en/html/TS.html.

9 *Universal POS tags*. Available at: https://universaldependencies.org/u/pos/.

10 *Serbo-Croatian MULTEXT-East Specifications*. Available at: http://nl.ijs.si/ME/V6/msd/html/msd-hbs.html. In the sixth and most recent MULTEXT-East release, Croatian, Serbian, and Bosnian specifications were replaced by Serbo-Croatian specifications, which cover the Croatian, Serbian, Bosnian and Montenegrin languages.

Example 2: *TEI version of the last turn shown in Example 1 (including the relevant lines in the element <timeline>)*

```
<TEI xmlns="http://www.tei-c.org/ns/1.0" version="4.1.0">
  <text>
    <body>
      <timeline unit="s" corresp="../audio/F8.wav" origin="#F8-u1-t0">
        […]
        <when xml:id="F8-u1-t0"/>
        <when xml:id="F8-u366-t1" interval="1024.2979996425654" since="#F8-u1-t0"/>
        <when xml:id="F8-u366-t2" interval="1029.1793571238047" since="#F8-u1-t0"/>
        […]
      </timeline>
      […]
      <u who="F8" xml:id="F8-u366" start="#F8-u366-t1" end="#F8-u366-t2">
        <w xml:id="F8-u366-w1" lemma="isključivo" pos="ADV" ana="mte:Rgp">isključivo</w>
        <w xml:id="F8-u366-w2" lemma="sa" pos="ADP" ana="mte:Si">sa</w>
        <w xml:id="F8-u366-w3" lemma="vi" pos="PRON" ana="mte:Pp2-pn">vi</w>
        <pause type="short" xml:id="F8-u366-p4"/>
        <vocal>
          <desc xml:id="F8-u366-v5">inhale (short)</desc>
        </vocal>
        <w xml:id="F8-u366-w6" lemma="ne" pos="PART" ana="mte:Qz">ne</w>
        <w xml:id="F8-u366-w7" lemma="oslovljavati" pos="VERB" ana="mte:Vmr3s">oslovljava</w>
        <del type="truncation" xml:id="F8-u366-w8">o</del>
        <w xml:id="F8-u366-w9" lemma="oslovljavati" pos="VERB" ana="mte:Vmr1s">oslovljavam</w>
        <w xml:id="F8-u366-w10" lemma="gospodin" pos="NOUN" ana="mte:Ncmsv">gospodine</w>
      </u>
      […]
    </body>
  </text>
</TEI>
```

3.4 Lemmatisation and morphosyntactic annotations

3.4.1 TAGGER

The normalised corpus was tagged with the tagger for Serbian and other South-Slavic languages CLASSLA-StanfordNLP (Ljubešić and Dobrovoljc, 2019), which is a fork of the StanfordNLP tagger.[11] The estimate of the accuracy on standard data for Serbian is 97.89 F1 for lemmatisation, and 95.23 F1 for morphosyntactic annotations. As in the first version (Lemmenmeier-Batinić et al., 2020), the corpus was tagged with a model trained on a set of all available training data for Serbian and Croatian: SETimes.SR 1.0 corpus of

11 *Classla 1.0.0* (CLASSLA Fork of Stanza for Processing Slovenian, Croatian, Serbian, Macedonian and Bulgarian). Available at: https://pypi.org/project/classla/.

newspaper texts (Batanović et al., 2018)[12], the hr500k Croatian reference training corpus (Ljubešić et al., 2016)[13], the ReLDI-NormTagNER, corpus of Serbian and Croatian tweets (Miličević and Ljubešić, 2016)[14,15], and the RAPUT corpus of Croatian non-professional writing (Štefanec et al., 2016). While in the first version of this corpus the tagger erroneously tagged several Ekavian words with Ijekavian lemmas (for instance, "hteo" 'wanted' was lemmatised as "htjeti" instead of "hteti" 'to want'), this feature was corrected in the second version, as the tagger was set to prefer Ekavian instead of Ijekavian variants.[16]

3.4.2 Evaluation of the TAGGER output

The accuracy of the tagger on our data was evaluated by checking the annotation of the first 500 tokens in one transcript.[17] The lemmatiser performed well with an accuracy of 98.2 F1. However, having both Serbian and Croatian corpora in the training set occasionally caused lemmatisation errors, since some word forms were annotated with lemmas characteristic of the Croatian, rather than the Serbian standard variety (such as the lemma "netko" [hr.] instead of "neko" [sr.] for the word form "neko" 'somebody').[18] The accuracy of morphosyntactic tags amounted to 92.2, which is, as expected, lower than the estimated accuracy for standard language data. Tagging errors are likely due to spoken

12 *Training corpus SETimes.SR 1.0*. Available at: Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1200.

13 *Training corpus hr500k 1.0*. Available at: Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1183.

14 *Serbian Twitter training corpus ReLDI-NormTagNER-sr 2.1*. Available at: Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1240.

15 *Croatian Twitter training corpus ReLDI-NormTagNER-hr 2.1*. Available at: Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1241.

16 The Proto-Slavic jat-vowel (ѣ in Cyrillic) has three different pronunciations in today's Shtokavian dialects: Ekavian (cf. first "e" in "vreme" 'time'), Ijekavian (cf. "ije" in "vrijeme") and Ikavian (cf. "i" in "vrime"). Standard Serbian has two variants: the Ekavian, which is spoken in most of Serbia, and the Ijekavian, which is spoken in south-west Serbia, but also in Croatia, Bosnia and Herzegovina and Montenegro. Since the corpus represents Serbian spoken by speakers using the Ekavian pronunciation (and living in Serbia), the tagger was set to prefer the Ekavian variants.

17 The evaluation of tagger's performance on this dataset was made in order to examine challenges related to tagging spoken language data. An elaborate evaluation of the tagger model would require a bigger and more diversified sample.

18 For this reason, in future versions we will test the tagger trained only on Serbian data.

character of the data, sometimes having a different word order, and extra-sentential elements that are rare in written (or standard) language data. One of the common tagging errors is the affirmative particle "da" ('yes', tag: "Qr"), which is frequently erroneously tagged as subordinating conjunction ("Cs"). Other erroneously tagged tokens are relative pronouns ("Pr") such as "koji" ('who') that are tagged as indefinite pronouns ("Pi"), as well as the interrogative particle "kako" ('how'), which is tagged as a subordinating conjunction ("Cs").[19]

Since Serbo-Croatian MULTEXT-East specifications do not propose tags for discourse particles, annotations that were compatible with the current MULTEXT-East specifications were regarded as correct in the evaluation process. For example, "znači" (literally: 'it means') was not counted as an error when it was tagged as verb, although it was used as discourse marker instead (see Halupka-Rešetar and Radić-Bojanić, 2014 on "znači" as discourse marker). Since specifications are missing for other discourse markers as well, they were also regarded as correct if their tags corresponded to the proposed MULTEXT-East specifications (for instance, "pa" 'well' was regarded as correct if it was tagged as coordinating conjunction "Cc"). However, in order to capture the peculiarities of spoken language, morphosyntactic specifications should ideally be extended to include discourse particles, hesitation signals, tag questions and other recurrent phenomena of the spoken register. Some examples of tagsets that were adapted to spoken language are STTS 2.0 for German (Westpfahl et al., 2017), and VOICE tagset (2014) for English. Extending Serbo-Croatian morphosyntactic specifications to suit spoken language phenomena would not only be of advantage for linguists interested in their use, but also for researchers developing other tools for processing spoken data.[20]

3.5 Aligning the corpus with audio segments

The transcripts were not originally aligned with the respective audio segments. This made searching for particular transcript segments in the audio

19 Specifications for tagging relative pronouns and interrogative particles are insufficiently documented in the MULTEXT-East specifications for Serbo-Croatian, which might have resulted in them being erroneously tagged not only in this, but also in other Serbian and Croatian corpora as well (see srWaC and hrWaC).

20 See Dobrovoljc and Martinc (2018) on the impact of discourse markers on spoken language dependency parsing for Slovene.

file an arduous task. In order to obtain alignments for each speaker turn, two forced alignment tools were tested: *aeneas*[21], and the model proposed by Plüss et al. (2020), using the Google Cloud STT Serbian ASR model. While *aeneas* offers support for aligning Serbian data, the model by Plüss et al. (2020) is not specifically tailored for Serbian, but requires an external ASR model.

For the first evaluation, we examined the difference in turn onset within the first minute in 9 different transcripts (88 turns). A comparison of turn beginnings produced by these two forced alignment tools against manual alignments showed that the model by Plüss et al. (2020) performs convincingly better than *aeneas* on our data (see Table 4). An assessment of the accuracy of alignment of 200 consecutive turns (17.5 minutes) is shown in Table 5.

Table 4: *Average absolute difference between turn beginnings calculated by forced alignment tools compared to manual alignment (measured in seconds)*

Absolute difference in turn onset $\lvert \textbf{turn start}_{\text{forced alignment}} - \textbf{turn start}_{\text{reference alignment}} \rvert$	**Plüss et al. (2020)**	***aeneas***
mean	1.17	10.32
median	0.58	2.75
standard deviation	1.88	15.14

Table 5: *Comparison of aeneas and the model by Plüss et al. (2020) regarding the accuracy of turn alignment in the transcript F1 (including non-lexical backchannels and affirmative particles)*

	Erroneously aligned turns	**Turns corresponding to the audio segments to a certain extent**			**Total**
		Partially correct	**Predominantly correct**	**Fully correct**	
Model by Plüss et al. (2020)	90 (45.0%)	21 (10.5%)	53 (26.5%)	36 (18.0%)	200 (100.0%)
aeneas	96 (48.0%)	27 (13.5%)	34 (17.0%)	43 (21.5%)	200 (100.0%)

At first glance in Table 5, both tools seem to produce unsatisfactory results: they both generate a high amount of erroneously aligned turns. *Aeneas* outputs more 'fully correct' alignments, but also more misalignments than the

21 *Aeneas*. Available at: https://www.readbeyond.it/aeneas/.

model by Plüss et al. (2020). The high amount of errors is due to a high rate of turns consisting only of affirmative particles ("da" 'yes') and non-lexical backchannels such as "mhm", or "aha", which are frequently misaligned (respectively, not-aligned) by both tools. [22] However, when turns consisting only of non-lexical backchannels and affirmative particles (n=66), are omitted, it becomes evident that the model by Plüss et al. (2020) outputs better alignments on our data than *aeneas* (see Table 6).

Table 6: *Comparison of aeneas and the model by Plüss et al. (2020) regarding the accuracy of turn alignment in the transcript F1 (excluding non-lexical backchannels and affirmative particles)*

	Erroneously aligned turns	Turns corresponding to the audio segments to a certain extent			Total
		Partially correct	Predominantly correct	Fully correct	
Plüss et al. (2020)	24 (17.9%)	21 (15.7%)	53 (39.5%)	36 (26.9%)	134 (100.0%)
aeneas	54 (40.3%)	25 (18.7%)	24 (17.9%)	31 (23.1%)	134 (100.0%)

Misalignments produced by the model by Plüss et al. (2020) are fewer (17.9% in comparison to 40.3% by *aeneas*), and they always consist of short speaker turns, whereas *aeneas* frequently misaligns longer turns as well. Therefore, the corpus has finally been aligned with the model proposed by Plüss et al. (2020).[23] With the help of turn alignments, users can navigate the transcripts while being able to hear the respective turns in the same time (or detect their approximate location in the audio segment in case they are not fully correct). The alignments are provided for each turn in the TEI version of the corpus (see attributes @start and @end in Example 2).

22 *Aeneas* has the advantage of *sometimes* producing correct alignments for these turns. However, the model by Plüss et al. (2020) has the advantage of pointing at empty alignments for these turns, so that they don't stand out as false positives during a manual inspection of alignments with transcription editors. The failed alignment of short and non-lexical backchannels is likely due to the fact that their transcription does not exactly correspond to their vocal realisation. A possible solution would be to add these alignments using transcription editors such as Partitur Editor (EXMARaLDA). However, this would require extensive manual adjustments, since non-lexical backchannels are frequent in our corpus (a search of all "aha", "hm", and "mhm" returns 5028 occurrences).

23 Only one transcript (id: F2) could not be aligned with the audio segments with either of the two tools, probably due to the low quality of the recording.

4 DATA SHARING

The corpus is available on CLARIN.SI.[24] In addition to the TEI-XML version of the corpus presented in this paper, we also provide raw transcripts including all annotations. The work in progress is documented at the GitLab repository of ZuCoSlaV corpora (Zurich Corpora of Slavic Varieties).[25] In accordance with the data privacy agreement, audio files are available on request. The corpus is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike (CC BY-NC-SA).[26]

5 POSSIBLE APPLICATIONS

The corpus presents a valuable resource for researchers interested in interactional linguistics, since it contains long fragments of natural language in interaction transcribed in great level of detail. The length of the transcripts, averaging to one hour of conversation, additionally allows one to study speaker-related peculiarities and different types of disfluencies produced in spontaneous conversation (pauses, truncations, self-repetitions, etc.). The almost equal number of male and female speakers allows for gender comparisons regarding content, as well as form-related phenomena. The corpus can be used for studying prosodic, lexical and morphosyntactic patterns of spoken Serbian. For instance, it is currently being used for investigating the use of simple past tenses and auxiliary omission in Serbian (Escher and Sonnenhauser, in preparation).

By providing semi-orthographic transcripts, this corpus may contribute to the development of tools for automatic speech recognition and forced alignment. Lastly, the XML encoding and annotation of the corpus also facilitates the study of forms of address, which are now normalised, lemmatised and tagged, and can be examined more easily by a quantitative approach.

24 *Corpus of Serbian Forms of Address 1.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1422.

25 *ZuCoSlaV: Zurich Corpora of Slavic Varieties*. Available at: https://gitlab.uzh.ch/uzh-slavic-corpora.

26 Licence details are available at: https://creativecommons.org/licenses/by-nc-sa/4.0/.

6 DISCUSSION

The Corpus of Serbian Forms of Address represents a significant step towards filling the gap of missing linguistic resources for spoken Serbian. While converting existing transcriptions requires substantial amount of manual work in the preprocessing step, in our case, the gain was worth the effort, since the interviews are long, the speaker metadata is provided, and the corpus has been meticulously and relatively consistently transcribed. Therefore, it cost less effort to clean and convert the corpus to a TEI-format and include all annotations, than it would to collect and transcribe new data of spoken Serbian from scratch.

The processing steps presented in this paper are useful for other researchers wanting to re-use existing material to create annotated corpora, and thereby enhance the study of spoken language. However, before starting the work on converting existing transcripts to a standardised format such as TEI-XML, it is important to carefully examine the quality of the transcripts, given that, depending on transcription consistency, the length of the corpus, or data formatting issues, it might take more time to preprocess the data than to transcribe it again with recent transcription tools. Transcription tools (such as for instance, FOLKER[27]) can control the syntax of transcribing conventions and align text with audio/video segments. Using these tools would not only assist the transcriber him/herself, but it would also significantly reduce the amount of work invested in enabling data re-use on the part of any third parties.

Another important issue that would facilitate data re-use is resolving possible data-privacy issues from start by ensuring that participants are willing to permit data re-use for general research purposes (and not only for one specific project they are originally taking part of). Making own transcripts available to a larger audience guarantees the transparency of research, and enables development of further work based upon it. Hopefully, considerations discussed in this paper will encourage data sharing of further collections of transcripts, and assist other researchers in converting existing transcript collections into annotated corpora of transcriptions of speech.

27 *FOLKER*. Available at: https://exmaralda.org/de/folker-de/.

7 CONCLUSION

Spoken language has long been overlooked not only when it comes to corpus resources, but also in regard to annotation conventions and development of models for automatic language processing. In addition to assessing the implications of data re-usability, and presenting a new resource for spoken Serbian, this paper addressed some unresolved issues regarding part-of-speech tags for spoken language phenomena, which are often left unspecified in the tagset specifications. An important step for further development of Serbian spoken language corpora would be to define the specifications for phenomena that are particular for the spoken register, such as discourse markers, non-lexical backchannels, hesitation markers, etc. The evaluation of forced alignment tools showed that there is also place for improvement regarding the implementation of Serbian models within current forced alignment tools. Using the approach of Plüss et al. (2020) via an open-domain ASR system for Serbian and resolving the issue of misaligned response tokens in future work would be a promising development for processing spoken Serbian data.

Acknowledgments

I would like to thank to Sonja Ulrich for sharing the transcripts and recordings she collected for her PhD thesis, and Tanja Samardžić (URPP Language and Space, Zurich) and Barbara Sonnenhauser (Department of Slavonic Languages and Literatures, Zurich) for enabling work on this corpus. I am also thankful to Nikola Ljubešić for tagging the corpus and for his insightful suggestions, Michel Plüss for aligning the corpus with his forced alignment model, and Miro Rodin, Petra Abramović and Luka Jovanović for their assistance in the evaluation of the automatic tools used for creating this corpus.

REFERENCES

Corpora, tools and tagsets

Aeneas. Retrieved from https://www.readbeyond.it/aeneas/

Classla 1.0.0 (CLASSLA Fork of Stanza for Processing Slovenian, Croatian, Serbian, Macedonian and Bulgarian). Retrieved from https://pypi.org/project/classla/

Croatian Twitter training corpus ReLDI-NormTagNER-hr 2.1. Retrieved from Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1241

FOLKER. Retrieved from https://exmaralda.org/de/folker-de/

Inflectional lexicon srLex 1.3. Retrieved from http://hdl.handle.net/11356/1233

Serbian Corpus of Early Child Language (SCECL). Retrieved from https://sla.talkbank.org/TBB/childes/Slavic/Serbian/SCECL

Serbian Twitter training corpus ReLDI-NormTagNER-sr 2.1. Retrieved from Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1240

Serbo-Croatian MULTEXT-East Specifications. Retrieved from http://nl.ijs.si/ME/V6/msd/html/msd-hbs.html

Spoken corpus of the Serbian minority in Hungary (SrMaCo). Retrieved from http://spokencorpus.eu/cms/bosco-2/

TEI Guidelines Version 4.2.1 (Transcriptions of Speech). Retrieved from https://tei-c.org/release/doc/tei-p5-doc/en/html/TS.html

Training corpus hr500k 1.0. Retrieved from Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1183

Training corpus SETimes.SR 1.0. Retrieved from Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1200

Universal POS tags. Retrieved from https://universaldependencies.org/u/pos/

ZuCoSlav: Zurich Corpora of Slavic Varieties. Retrieved from https://gitlab.uzh.ch/uzh-slavic-corpora

Other

Anđelković, D., Ševa, N., & Moskovljević, J. (2001). Serbian Corpus of Early Child Language. Laboratory for Experimental Psychology, Faculty of Philosophy, and Department of General Linguistics, Faculty of Philology, University of Belgrade.

Batanović V., Ljubešić, N., & Samardžić, T. (2018). SETimes.SR – A Reference Training Corpus of Serbian. *Proceedings of the Conference on Language Technologies & Digital Humanities 2018 (JT-DH 2018)* (pp. 11–17). Ljubljana, Slovenia.

Batinić, J., Frick, E., & Schmidt, T. (in press). Accessing spoken language corpora: An overview of current approaches. *Corpora. Edinburgh University Press*.

Delić V., Sečujski, M., Jakovljević, N., Pekar, D., Mišković, D., Popović, B., Ostrogonac, S., Bojanić, M., & Knežević, D. (2013). Speech and Language Resources within Speech Recognition and Synthesis Systems for Serbian and Kindred South Slavic Languages. In M. Železný, I. Habernal, A. Ronzhin (Eds.), *Speech and Computer. SPECOM 2013. Lecture Notes in Computer Science: Vol. 8113* (pp. 319–326). Springer, Cham. doi: 10.1007/978-3-319-01931-4_42

Dobrić N. (2012). Language Corpora in The West Balkans – History, Current State and Future Perspective. *Slavistična revija, 60*(4), 677–692.

Dobrovoljc, K., & Martinc, M. (2018). Er … well, it matters, right? On the role of data representations in spoken language dependency parsing. *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)* (pp. 37–46). Brussels, Belgium.

Escher, A., & Sonnenhauser, B. (in press). Simple Past Tenses in the Timok dialect.

Halupka-Rešetar, S., & Radić-Bojanić. B. (2014). The discourse marker *znači* in Serbian: An analysis of semi-formal academic discourse. *Pragmatics, 24*(4), 785–798.

Kostić, A. (2003). Đorđe Kostić electronic corpus of the Serbian language. In *Zbornik Matice srpske za slavistiku: Vol. 64* (pp. 260–264).

Krstev, C., & Vitas, D. (2005). Corpus and Lexicon – Mutual Incompleteness. In *Proceedings of the Corpus Linguistics Conference*, 14–17 July 2005, Birmingham. United Kingdom (hal-01108218).

Lemmenmeier-Batinić, D., Ljubešić, N., & Samardžić, T. (2020). XML-Encoding of a spoken Serbian corpus targeting forms of address. In D. Fišer in T. Erjavec (Eds.), *Proceedings of the Conference on Language Technologies & Digital Humanities* (pp. 127–130). Ljubljana: Institute of Contemporary History.

Ljubešić N., & Klubička. F. (2014). {bs,hr,sr}WaC – Web Corpora of Bosnian, Croatian and Serbian. *Proceedings of the 9th Web as Corpus Workshop (WaC-9)* (pp. 29–35). Gothenburg, Sweden.

Ljubešić, N., Klubička, F., Agić, Ž., & Jazbec. I. (2016). 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)* (pp. 4264–4270). Portorož, Slovenia.

Ljubešić, N., & Dobrovoljc, K. (2019). What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing* (pp. 29–34). Florence, Italy.

Miličević, M., & Ljubešić. N. (2016). Tviterasi, tviteraši or twitteraši? Producing and analysing a normalised dataset of Croatian and Serbian tweets. *Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research, 4*(2), 156–188.

Plüss, M., Neukom, L., & Vogel, M. (2020). Swiss Parliaments Corpus, an Automatically Aligned Swiss German Speech to Standard German Text Corpus. Retrieved from https://arxiv.org/abs/2010.02810

Popović, Z. (2010). Taggers Applied on Texts in Serbian. *INFOtheca, 11*(2), 21–38.

Schmidt, T. (2016). Construction and Dissemination of a Corpus of Spoken Interaction – Tools and Workflows in the FOLK project. *Corpus linguistic software tools, 31*(1), 127–154.

Selting, M., Auer, P., Barden, B., Bergmann, J., Couper-Kuhlen, E., Günthner, S., Quasthoff, U., Meier, C., Schlobinski, P., & Uhmann, S. (1998). Gesprächsanalytisches Transkriptionssystem (GAT). *Linguistische Berichte 173*, 91–122.

Selting, M., Auer, P., Barth-Weingarten, D., Bergmann, J., Bergmann, P., Birkner, K., Couper-Kuhlen, E., Deppermann, A., Gilles, P., Günthner, S., Hartung, M., Kern, F., Mertzlufft, C., Meyer, C., Morek, M., Oberzaucher, F., Peters, J., Quasthoff, U., Schütte, W., Stukenbrock, A., & Uhmann, S. (2009). Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion,* (10), 353–402.

Suzić, S., Ostrogonac, S., Pakoci, E., & Bojanić. M. (2014). Building a Speech Repository for a Serbian LVCSR System. *Telfor Journal*, *6*(2), 109–114.

Štefanec, V., Ljubešić, N., & Kuvač Kraljević. J. (2016). Croatian Error-Annotated Corpus of Non-Professional Written Language. *Proceedings of the*

Tenth International Conference on Language Resources and Evaluation (LREC 2016) (pp. 3220–3226). Portorož, Slovenia.

Ulrich, S. (2018). *Anredeformen im Serbischen*. Wiesbaden.

Utvić, M. (2011). Annotating the Corpus of Contemporary Serbian. *INFOtheca 12*(2), 36–47.

VOICE (2014). Part-of-Speech Tagging and Lemmatization Manual. With assistance of Barbara Seidlhofer, Stefan Majewski, Ruth Osimk-Teasdale, Marie-Luise Pitzl, Michael Radeka, Nora Dorn. The Vienna-Oxford International Corpus of English. Retrieved from http://www.univie.ac.at/voice/documents/VOICE_tagging_manual.pdf

Vuković, T. (2021). Representing variation in a spoken corpus of an endangered dialect: the case of Torlak. *Language Resources and Variation*. Springer Nature. doi: 10.1007/s10579-020-09522-4

Westpfahl, S., Schmidt, T., Jonietz, J., and Borlinghaus, A. (2017). STTS 2.0. Guidelines für die Annotation von POS-Tags für Transkripte gesprochener Sprache in Anlehnung an das Stuttgart Tübingen Tagset (STTS). Working paper. Mannheim: Institut für Deutsche Sprache.

PRETVORBA ZBIRKE SUROVIH ZAPISOV V ANOTIRAN IN SPREMENJEN TEI-XML KORPUS: PRIMER KORPUSA SRBSKIH OBLIK NASLAVLJANJA

V prispevku je opisan postopek gradnje TEI-XML korpusa govorjenega srbskega jezika, začenši s surovimi prepisi. Korpus sestavljajo polstrukturirani intervjuji, ki so bili zbrani z namenom raziskati oblike naslavljanja v srbščini. Intervjuji so bili temeljito prepisani v skladu s konvencijami o prepisovanju GAT. Prepis pa je bil izveden brez orodij, ki bi nadzorovala veljavnost sintakse GAT ali poravnala prepis z zvočnimi zapisi. Da bi ta vir ponudili širši publiki, smo odpravili nedoslednosti v izvirnih prepisih, normalizirali polortografske prepise in korpus pretvorili v format TEI za prepise govora. Nadalje smo korpus obogatili z označevanjem in lematizacijo podatkov. Nazadnje smo z orodjem za prisilno poravnavo v korpusu poravnali govore posameznih govorcev s pripadajočimi segmenti govornega signala. Ta članek poleg predstavitve glavnih korakov pri pretvorbi korpusa v format XML razpravlja tudi o trenutnih izzivih pri obdelavi govorjenih podatkov ter o implikacijah ponovne uporabe podatkov pri prepisih govora. Korpus srbskih oblik naslavljanja lahko uporabimo za preučevanje srbščine z vidika interakcijske lingvistike, za raziskovanje morfosintakse, leksike in fonetike govorjenega srbskega jezika, za preučevanje disfunkcij ter za preizkušanje modelov za samodejno prepoznavanje govora in prisilno poravnavo. Korpus je prosto dostopen za raziskovalne namene.

Ključne besede: govorjena srbščina, jezikovni biografski intervjuji, oblike naslavljanja, ponovna uporabnost podatkov

HEDGING MODAL ADVERBS IN SLOVENIAN ACADEMIC DISCOURSE

Jakob LENARDIČ, Darja FIŠER
Faculty of Arts, University of Ljubljana; Jožef Stefan Institute

Lenardič, J., Fišer, D. (2021): Hedging modal adverbs in Slovenian academic discourse. Slovenščina 2.0, 9(1): 145–180.

DOI: https://doi.org/10.4312/slo2.0.2021.1.145-180

This paper first presents a comparative analysis of modal adverbs in doctoral theses in the humanities and social sciences on the one hand, and in natural and technical sciences on the other from the 1.7-billion-token corpus of Slovenian academic texts KAS (Erjavec et al., 2019a). Using a randomized concordance analysis, we observe the epistemic and non-epistemic usage of the modal adverbs and show that epistemic adverbs are more characteristic of the humanities and social sciences theses. We also show that the non-epistemic dispositional meaning of possibility, which is most commonly used in natural and technical sciences theses, is not used as a hedging device. In the second part of the paper we compare the usage of a selected set of modals in bachelor's, master's and doctoral theses in order to chart how researchers' approach to stance-taking changes at different proficiency levels in academic writing, showing that the observed increase in hedging devices in doctoral theses seems to be less a function of an increased proficiency level in academic writing as such and more the result of conceptual differences between undergraduate and postgraduate theses, only the latter of which are original research contributions with extensive discussion of the results.

Keywords: epistemic modality, root modality, hedging, semantics, pragmatics, corpus linguistics

1 INTRODUCTION

Modal expressions offer an interesting insight into academic discourse because they can pragmatically function as *hedges* (Lakoff, 1972; Hyland, 1996, 1998), which are used by authors to present their claims with varying degrees of tentativeness. In academic writing, hedging is a particularly important pragmatic device, as it "enables writers to express a perspective on their statements, to present unproven claims with caution, and to enter into a dialogue with their audiences" and is therefore an "important means by which professional scientists confirm their membership in research communities" (Hyland, 1996, pp. 251–252).

In related work, which has primarily focused on English academic discourse, it is often shown that hedging is more characteristic of humanities and social sciences rather than natural and technical sciences (Hyland, 1998; Takimoto, 2015), which reflects the general idea that humanities and social sciences are more interpretative and less rooted in empirical research than natural and technical sciences (Takimoto, 2015). In this paper, we try to confirm whether this is also the case for Slovenian academic discourse on the basis of the doctoral theses in the *KAS corpus of Slovenian academic writing* (Erjavec et al., 2019a).[1] We present a quantitative analysis of the most frequent modal adverbs that display epistemic and possibly non-epistemic meanings and then conduct a randomized concordance analysis to determine whether the modals that pragmatically serve as hedging devices are also used more frequently in the humanities and social sciences.

Apart from cross-disciplinary comparisons, hedging in academic discourse has also been studied from the perspective of its developmental trajectory (Hyland, 2004; Lancaster, 2016) where it is compared between early forms of academic writing such as (under)graduate research papers on the one hand and published academic writing on the other in order to chart how researchers' approach to stance-taking changes as they gain experience in academic

1 This paper is an extended version of the conference paper Lenardič and Fišer (2020). We have employed a more fine-grained classification of epistemic modality, which has allowed us to take additional evidential/assumptive modals into consideration as well. Furthermore, we now also compare the prominence of hedging in PhD theses with hedging in bachelor's and master's theses on the basis of a relevant subset of the analysed modals.

writing (Aull and Lancaster, 2014). We contribute to this line of research by comparing a subset of the most frequent modal adverbs between the doctoral theses on the one hand and the bachelor's and master's theses in the *KAS corpus* (Erjavec et al., 2019a) on the other, namely, the subset of those modals that invariably play a hedging role in terms of discourse pragmatics and thus correspond to the authors' stance taking.

The paper is structured as follows. In Section 2, we lay out the relevant linguistic theory on modality and present the pragmatic notion of hedging. In Section 3, we discuss previous treatments of modality in Slovenian linguistics as well as related work on corpus-based treatment of hedging in academic discourse. In Section 4, we present the corpus we used for our analysis from the perspective of the extra-linguistic metadata relevant for our purposes as well as discuss the selection criteria of the modal adverbs that we have analysed. In Section 5, we present and discuss the results. In Section 6, we conclude the paper.

2 THEORETICAL FRAMEWORK

2.1 Epistemic and Non-Epistemic Modalities

Modality has been defined in many different ways in the literature, but it is perhaps von Fintel (2016, p. 21) who most succinctly summarizes the notion:

> Modality is a category of linguistic meaning having to do with the expression of possibility and necessity. A modalized sentence locates an underlying or prejacent proposition in the space of possibilities [...] *Sandy might be home* says that there is a possibility that Sandy is home. *Sandy must be home* says that in all possibilities, Sandy is home.

Modality thus evaluates a proposition from the perspective of the gradient from possibility to necessity. Notions such as *possibility*, *likelihood*, and *necessity*, which are logically related by entailment, are also referred to as the modal force (Kratzer, 2012). Aside from this, modality is polysemous and the usual linguistic distinction is made between epistemic modality on the one hand and non-epistemic modality on the other (Palmer, 2014), the latter of which is usually referred to as root modality (Coates, 1983) or circumstantial modality (Kratzer, 2012). In this paper, we use the term root modality.

Epistemic modality encompasses the speaker's judgement about the truth of the proposition (Palmer, 2014, p. 50). A modal like *mogoče* in sentence (1) is

epistemic, expressing that the speaker is not completely certain that the prejacent i.e. unmodalised proposition *Ana je doma* "Ana is home" is true.[2]

(1) Ana je *mogoče* doma.
"Ana is possibly home."

By contrast, root modality also evaluates the proposition in the domain of possibility (and necessity), but, unlike epistemic modality, does not tie the evaluation to the speaker's knowledge. An example of a non-epistemic modal is *lahko* in sentence (2).

(2) Ta program se *lahko* namesti na Windows.
"This program can be installed on Windows."

Here, *lahko* is not used to indicate the speaker's knowledge about the truth of the expressed proposition but rather to attribute possible qualities to the subject NP *ta program* "this program".

A single modal often allows for more than one reading that is contextually determined. For instance, *lahko* in sentence (3) has an epistemic reading that can be paraphrased as "It is possible that Ana is at home or at school" and a root meaning that denotes permission that Ana is granted by someone else ("Ana is allowed to stay at home or in school"), which is typically disambiguated by the context it appears in.[3] This motivates the manual concordance analysis of the Slovenian modal adverbs that will be presented in Section 5.2.

(3) Ana je *lahko* doma, *lahko* pa je v šoli.
"Ana may be at home or school."
"Ana can be at home or school."

Finally, many root modal expressions display prominent meta-discursive usage, as in the case of reader-oriented meta-commentary clauses like the one in example (4). Such use along with the purely epistemic meaning often corresponds to the pragmatic notion of hedging (Hyland, 1996, 1998; Grabe and Kaplan, 1997), which we introduce in Section 2.2.

2 For ease of exposition, we use simple constructed linguistic examples to showcase the relevant semantic characteristics of modality in this section.

3 The modal meaning involving obligation/permission is referred to as *deontic modality* by Palmer (2014).

(4) Kot *lahko* vidimo iz rezultatov …
"As can be seen from the results…"

2.2 Hedging – a Pragmatic Strategy

In linguistics, Lakoff (1972, p. 471) was the first to use the term *hedges* to refer to "words whose meaning implicitly involves fuzziness – words whose job is to make things fuzzier or less fuzzy". Lakoff (1972)'s basic concept is further explicated by Hyland (1996, p. 251), who claims that hedges are "any linguistic means used to indicate either (a) a lack of complete commitment to the truth of a proposition, or (b) a desire not to express that commitment categorically". Additionally, hedging not only involves markers of tentativeness but is typically extended to include rhetoric communicative strategies, e.g., politeness, by means of which the author implicitly includes the addressee in the discourse her or she is presenting (Grabe and Kaplan, 1997, p. 154).

Hyland (1996)'s definition of hedging overlaps quite significantly with that of epistemic modality defined in the previous section, but there is an important difference: a hedge is not a lexical property that holds of a specific category like modality, but rather a pragmatic device that can in principle hold for any lexical category given the suitable communicative context.

In terms of grammatical categories, hedging corresponds not only to modal verbs or adverbs, but also to other lexical categories such as the use of certain reporting verbs that indicate the author's tentativeness (e.g., *we believe that*) as well as syntactic strategies such as the use of the passive rather than the active voice to syntactically omit the otherwise entailed agent of the verbal event (Rizomilioti, 2006, p. 56) or the use of inclusive plural pronouns to help establish rapport between the reader and the writer (Hyland, 1996).

3 RELATED WORK

3.1 The Slovenian Modal System

Slovenian linguists generally discuss Slovenian modals either in relation to highly specialised topics in theoretical linguistics or in the context of applied and descriptive comparative linguistics. Theoretical linguists usually focus on discussing the formal properties of individual selected modal lexemes;

for instance, Marušič and Žaucer (2016) propose a syntactic explanation why the modal adverb *lahko* is a positive-polarity item (i.e., it cannot syntactically co-occur with negation), while Hladnik (2015, p. 86) discusses the fact that the lexeme *da*, which is syntactically a subordinator, triggers an epistemic meaning in relative clauses (e.g., *človek, ki* da *pride* "the person who *supposedly* is coming"). In applied/comparative linguistics, researchers usually use the modals as a springboard for studying broader pragmatic topics; for instance, Pisanski Peterlin (2015) discusses how Slovenian epistemic modals are used in English–Slovenian translation in comparison to original Slovenian texts in order to determine how epistemic modality is influenced by language transfer, while Pihler Ciglič (2017) compares the use of assumptive modals like *morda* with related lexemes in American Spanish in the context of literary translations.

However (and to our knowledge), no one has yet attempted a comprehensive typological study of the general syntactic and semantic properties of the Slovenian modal system in the context of descriptive Slovenian linguistics on par with Palmer (2014)'s work on English modal auxiliaries. What is especially noteworthy in relation to modal adverbs is that the Slovenian reference grammar *Slovenska slovnica* (Toporišič, 2004) only lists them as examples of the particle word class, but does not devote any attention to their syntactic characteristics nor to a more fine-grained semantic classification that would disentangle notions such as the modal force from the modal base for a given modal. As we will see in Section 4.2, such an uncomprehensive classification of modal adverbs in the reference grammar seems to have, at least from the perspective of syntactic consistency, also negatively affected the morphosyntactic tagging in Slovenian corpora, which is based on the reference grammar, as modal lexemes that are syntactically adverbs seem to be arbitrarily assigned to either the adverb or the particle classes.

In our paper, we take into account the fact that modals display a complex semantics. Although our primary aim is to investigate academic discourse, we nevertheless believe that certain aspects of our study, such as the rate at which a modal conveys a particular modal reading (Section 5.2), also positively contribute to the general understanding of the lexical-semantic characteristics the Slovenian modal system. However, a more comprehensive description of the

modal system, which should also compare the use of Slovenian modality in registers other than academic discourse, goes far beyond the scope of this paper.

3.2 Modal Adverbs and Hedging in Academic Discourse – Cross-Disciplinary Comparisons

In related work on hedging in academic discourse, researchers (Hyland, 1998; Rizomilioti, 2006; Pisanski Peterlin, 2010; Takimoto, 2015, a.o.) have generally taken into account all of the major categories that can in principle be used to hedge discourse, such as modal auxiliaries, modal and non-modal (e.g., approximators) adverbs and adjectives, and lexical verbs.

For instance, Takimoto (2015) analyses how hedges corresponding to 5 syntactic categories (adverbs, adjectives, auxiliaries, nouns, and verbs) are used across 4 different natural sciences disciplines and 4 humanities/social sciences disciplines, showing that "70% of all hedges and boosters were found in humanities and social sciences" (2015, p. 103) and that philosophy contains "almost 5.3 times as many hedges and boosters as electrical engineering" (*ibid.*).[4] Similarly, Rizomilioti (2006, p. 64) compares the use of hedging between a 200,000 token corpus of journal papers in literary criticism and a comparable corpus of papers in biology, showing that there are more adverbs of uncertainty in the literary criticism corpus than in the biology corpus.

Given the high degree of lexical polysemy and the consequent likelihood that not all of the observed lexemes in the studied corpus function as hedges, a prominent strategy to filter out irrelevant data relies on the close reading of all the concordances that potentially correspond to hedges in order to single out only the relevant occurrences. For this to be possible, the corpora used in the related literature are often quite small, generally consisting of 100,000–500,000 tokens and around 50–60 research articles (Thompson, 2000; Pisanski Peterlin, 2010; Hyland, 1998; Rizomilioti, 2006; Takimoto, 2015).

Nevertheless, despite such a strategy of close reading, the epistemic and non-epistemic notions of possibility seem conflated in some of the related

4 Some authors use the term *boosters* to describe those hedges that convey the author's certainty rather than tentativeness; since our analysis, presented in Section 5.1, does not show prominent differences between hedges and boosters, we use *hedges* as a general term for expressing both tentativeness and certainty.

work. For instance, Piqué-Angordans et al. (2002), who survey how English modal auxiliary verbs (e.g., *can*, *may*, *should*) vary between their epistemic and root/deontic senses across 3 corpora of research articles in medicine, biology, and literary criticism, provide the following 2 examples as expressing epistemic modality in their corpus of research articles in medicine (2002, p. 53):

(5) Tricyclic antidepressants, however, *can* also have significant adverse effects, such as arrhythmias, postural hypotension, sedation, dry mouth, constipation, confusion, and urinary retention.

(6) The quantities of the factors *could* limit the amount of renin mRNA that can be produced, even under conditions of normal salt loading and in the absence of pharmacological interventions.

While the use of *could* in sentence (6) undoubtedly expresses an epistemic judgement, i.e., that the authors are not certain whether the "quantities of the factors" do in fact "limit the amount of renin mRNA", the use of *can* in sentence (5) plays a different i.e. non-epistemic modal role, in contrast to Piqué-Angordans et al. (2002)'s claim.[5] That is, *can* in (5) simply expresses that "tricyclic antidepressants" have properties that can cause adverse effects under certain undefined conditions. As we will see in Section 5.2, the distinction between the two meanings is crucial from the perspective of hedging; we will claim that only expressions of possibility like that in (6) but not in (5) constitute this pragmatic strategy.

We therefore attempt to make our quantitative analysis of the modals more precise by making such a distinction between the modality types introduced in Section 2.1, arguing that only those instances of possibility expressed by the modals that correspond either to epistemic modality or to the meta-discursive usage function as hedges, whereas non-epistemic meanings of possibility that correspond to dispositional ascriptions do not.

5 This sentence is taken from the introduction of the paper by Rowbotham et al. (1998), where the co-text affirms that the use of *can* here is not meant to convey the authors' epistemic judgement. It is also worth noting that Portner (2009, p. 30) claims that *can* is never used epistemically (e.g., *It can be raining* does not seem to admit an epistemic reading unless it is negated).

Our corpus, which we introduce in Section 4.1, is also significantly larger than those in the related literature, consisting of approximately 1.7 billion tokens. Because close reading of such a large corpus was not a feasible approach for us and because we wanted to reduce the amount of irrelevant data that in part arises from the often unpredictable lexical polysemy,[6] we limit our analysis to a single word class, i.e., modal adverbs, which can be queried systematically via its morphosyntactic tag and at the same time arguably constitute the most prominent category for expressing sentential modality in Slovenian.

3.3 Modal Adverbs and Hedging in Academic Discourse – Between Academic Stages

In another major strand of related work (e.g., Aull and Lancaster, 2014; Aull et al., 2017; Crosthwaite et al., 2017), it is shown that there are prominent differences in the use of markers of stance between early and advanced academic writing. For instance, Aull and Lancaster (2014) survey the distribution of English approximative hedges (e.g., *generally*, *evidently, somewhat*) in the context of research papers written by students at US universities, comparing them between 3 corpora: first, a corpus of argumentative essays by first-year undergraduate students (abbr. *FY*); second, a corpus of upper-level essays by third-year students and graduate students (abbr. *UP*); and third, published scholarly writing from peer-reviewed journals in the academic subcorpus of

6 It is also often quite unclear whether research that observes hedging across multiple word classes (and broader syntactic patterns) takes into account the idiosyncratic grammatical features of a category that distinguish it from others and could serve as potential caveats for studying pragmatic effects. An example of this is modal adjectives. Modality in NP-modifying adjectives exhibits sub-sentential semantic scope (Portner, 2019), which means that it does not take scope over the asserted proposition in contrast to prototypical modals but rather over an implicit proposition that is presupposed in the semantics of the noun phrase (DeLazero, 2011).
Crucially, what is then hedged in such cases is a non-overt claim; for instance, *možno* in a sentence like *To so možne analize* "These are the possible analyses" takes scope over a non-overt presupposed proposition in the noun phrase *možne analize*, with the resulting modalised meaning being either something like *these analyses might be correct* (epistemic) or *these analyses can be correct under certain circumstances* (root), which however is not something that is asserted by the original sentence. Since the modalised proposition is thus non-overt, it is often quite unclear if and how the claim is being hedged in such cases. None of the reviewed related work on hedging that looks at modal adjectives takes this into account.

the *Corpus of Contemporary American English* (abbr. *COCAA*). It is shown that the frequency of such approximative hedges increases between all three corpora: from 109.5 per 100,000 words in the *FY sub*corpus to 173.5 in the *UP* subcorpus, that is a 58% increase from *FY*, and finally to 203.8 per 100,000 words in *COCAA,* that is an 86% increase from *FY* (Aull and Lancaster, 2014, p. 162).

Interpreting this increase observed in American English academic writing, Aull and Lancaser (*ibid.*) claim that students are "often encouraged to take a 'critical stance' with regard to others' arguments" and that a "highly attitudinal, forceful, and assertive stance is less valued in advanced student writing than stances that are implicitly attitudinal [...] or open to other views in the surrounding discourse" (*ibid.*, p. 155). Similarly, Aull et al. (2017, p. 32) claim that published academic writing more prominently displays "qualified and circumscribed arguments" than the writing of incoming college students. In sum, advanced writers use hedge to obviate a forceful, asserted stance by more frequently using hedging devices.

However, such an increase in hedging from less mature to more advanced writing is not necessarily a universal trend. Crosthwaite et al. (2017), who compare the use of stance expressions between learner and professional research reports in dentistry, observe that hedging in their dentistry professional corpus is *less* frequent than in the learner corpus. This is precisely the opposite of the results reported by Aull and Lancaster (2014). In the second part of the paper, we therefore attempt to determine this trend for Slovenian academic writing by comparing the frequency of hedging adverbs between Slovenian bachelor's, master's, and doctoral theses, which are the final works signalling the completion of each of the three major stages of tertiary education in Slovenia.

4 METHODOLOGY

4.1 The KAS Corpus of Academic Slovenian

The study presented in this paper has been carried out on the 1.7-billion-token *KAS corpus of Slovenian academic writing* (Erjavec et al., 2019a). The theses in the corpus were written between 2000 and 2018 at Slovenian universities

and other academic institutions.[7] The corpus is linguistically annotated and is also marked up for several extra-linguistic metadata categories that are tailored to the genre of academic theses, the most relevant for our purposes being the publisher and CERIF (Common European Research Information Format). The corpus is accessible online through the CLARIN.SI *noSketch Engine* concordancer,[8] which is an open-source version of *Sketch Engine* corpus query system.

The Publisher information corresponds to the institution or faculty where the thesis was defended. There are a total of 70 different publisher abbreviations, 55 of which are faculties of the Universities of Ljubljana, Maribor, Nova Gorica, and Primorska. The remaining 15 are research institutes with their own study programmes or private and semi-private colleges. The corpus represents a very diverse breadth of scientific (sub)disciplines, so each thesis has been assigned to (at least) one of the five top-level CERIF[9] categories: BIO(MEDICAL SCIENCES), HUM(ANITIES), PHYS(ICAL SCIENCES), SOC(IAL SCIENCES), and TECH(NOLOGICAL SCIENCES). Since the CERIF categories represent a generalised division of academic disciplines, they are particularly well-suited for comparative corpus analyses of academic genres, especially given the diverse disciplinary scope of the individual publishers included in the corpus.

The CERIF division of the theses in the *KAS* corpus is given in Table 1.

Table 1: *The five disciplinary subcorpora of KAS*

CERIF	Size (in tokens and %)	
BIO	100,514,116	7%
HUM	150,634,867	10%
PHYS	147,690,128	10%
SOC	1,018,235,132	66%
TECH	121,360,503	8%
∑	1,538,434,746	100%

7 The morphosyntactic annotation and lemmatisation of the corpus was performed with the ReLDI morphosyntactic tagger and lemmatizer (https://github.com/clarinsi/reldi-tagger), which gives an accuracy of 98.94% on the parts of speech and 94.27% on the complete morphosyntactic descriptions. For a comprehensive description of the corpus, see Erjavec et al. (2020).

8 https://www.clarin.si/noske/.

9 https://eurocris.org/services/main-features-cerif. Accessed on 16 June 2021.

As shown in Table 1, the five CERIF subsets of *KAS* are unequal in size, with the SOC(IAL SCIENCES) subset accounting for over half of the corpus. Consequently, we will provide frequency counts for our modal adverbs that are relativised to a million tokens. Furthermore, the total token size (1,538,434,746) listed in Table 1 is slightly smaller than that of the entire *KAS* corpus (1,699,097,710); this is because approximately 9% of the theses are assigned to multiple CERIF categories, while the texts that we take into account include all the theses with only one CERIF label.

In the first part of our analysis, we focus on the subcorpus of doctoral theses, *KAS-dr* (Erjavec et al., 2019c), which consists of 1569 doctoral theses, amounting to a total of 100 million tokens or roughly 7% of the entire *KAS* corpus. In the second half of our analysis, we compare the results obtained for the *KAS-dr* subcorpus with the subcorpora of master's (*KAS-mag*; Erjavec et al., 2019b) and bachelor's theses (*KAS-dipl*; Erjavec et al., 2019d), which contain 496,000,000 tokens (31% of the entire *KAS* corpus) and 1.1 billion tokens (72% of the entire *KAS* corpus), respectively. Because of this inequality in size, and because the theses are unequally distributed among the CERIF categories in all three subcorpora in roughly the same ratio as in Table 1 (i.e., SOC theses account for more than half of each subcorpus), we will again use normalized frequencies to compare the findings in the three subcorpora.

4.2 Modal Adverbs

The modal adverbs analysed in this paper are listed in Table 2. There are 6 adverbs that denote possibility (*lahko, mogoče, možno, morda, menda*, *morebiti*), 3 adverbs that denote likelihood (*najbrž, domnevno, verjetno*), and 3 adverbs that denote certainty (*nedvomno*, *zagotovo, gotovo*).

The modals were selected in the following way. We first extracted all the lemmas in the *KAS-dr* subcorpus that are morphosyntactically tagged as either adverbs or as particles. It is important to note that the Slovenian descriptive grammar *Slovenska slovnica* (Toporišič, 2004), which is the basis for the MULTEXT tagset[10] used by the *KAS* corpus (Erjavec, 2012), postulates that the particle is a separate word class. Toporišič (2004, pp.

10 https://www.sketchengine.eu/slovene-tagset-multext-east-v5.

Table 2: *The most frequent epistemic modal adverbs in the KAS-dr subcorpus*

MODAL	Meaning	AF	RF
lahko	possibly	296,311	2,920
verjetno	likely	12,958	128
morda	possibly	9,727	96
zagotovo	certainly	3,291	32
gotovo	certainly	3,152	31
nedvomno	certainly	2,534	25
mogoče	possibly	1,878	19
možno	possibly	1,346	13
najbrž	likely	1,082	11
domnevno	likely	969	10
morebiti	possibly	811	8
menda	possibly	315	3

Note. AF lists the absolute frequencies while RF lists the relative frequencies per 1 million tokens.

445–449) exceptionally defines the particle class solely in terms of its semantic rather than syntactic properties, claiming that the category is distinct from adverbs in that it consists of semantically abstract clausal modifiers (i.e., propositional operators) rather than event modifiers such as adverbials of manner or time. While most of the lexemes in Table 2 are tagged as adverbs in *KAS*, *morda*, *najbrž*, *morebiti*, and *menda* are tagged as particles, even though their syntactic distribution is prototypically adverbial. In other words, there are no categorical differences between *verjetno*, which is tagged as an adverb, and *najbrž*, which is tagged as a particle. For simplicity's sake, we thus refer to all the 12 lexemes in Table 2 as adverbs. From this extracted list of adverb and "particle" lexemes in the corpus, we selected all that semantically correspond to epistemic modals and are not stylistically marked; because of this latter criterion, we omitted the infrequent colloquial hearsay modals *bržda* "likely", *baje* "possibly", *nemara* "likely", and *bojda* "possibly".

The 12 lexemes in Table 2 largely correspond to the epistemic modal adverbs identified for Slovenian by Pisanski Peterlin (2015, p. 31). However, in contrast to her approach, our selection criteria were stricter in that we excluded

those adverbs that are frequently ambiguous between a modal and non-modal (e.g., manner) interpretation.[11]

Such an ambiguous modal is *očitno* "apparently", as shown by the two possible paraphrases of example (7), taken from *KAS-dr*, where the first corresponds to a modal interpretation denoting the speaker's attitude towards the proposition while the other to a non-modal interpretation in which the adverb specifies the manner of the verbal event.

(7) Z naraščajočim deležem titana se je *očitno* zmanjšala količina ter velikost evtektičnih karbidov M7C3.
"It appears that with the increasing amount of titanium, the quantity and size of eutectic carbides M7C3 has decreased."
"With the increasing amount of titanium, the quantity and size of eutectic carbides M7C3 has decreased in an obvious manner/to a great degree."

Discounting such ambiguous adverbs reduces the amount of irrelevant data; that is, it ensures that our comparative analysis is not hindered by the noise due to polysemy.

5 THE RESULTS

5.1 Quantitative Analysis of Modal Adverbs Across Disciplines in Doctoral Theses

Table 3 compares the distribution of the 12 modal adverbs in focus between the humanities (i.e., HUM) and social sciences (SOC) disciplines in *KAS-dr* on the one hand and the biotechnical (BIO), physical sciences (PHYS), and technological (TECH) disciplines on the other. The size of HUM and SOC is 68,207,965 tokens in total, while the size of BIO, PHYS, and TECH is 39,679,476 tokens in total. The AF columns reports the absolute frequency and RF the relative frequency, which is normalised to 1 million tokens.

11 The adverb *lahko* also has a manner interpretation, i.e., "easily". However, this use is very rare – in our analysis of a randomized set of 250 concordance examples (see Section 5.2) for this adverb, there was only 1 example, given in (i), where *lahko* is used in its comparative form *lažje* and corresponds to the non-modal manner usage:
(i) [...] zaradi česar *lažje* in pogosteje prihaja do sprememb v vrednostih indikatorjev.
"[...] because of which changes in the values of the indicators occur more frequently and more easily."

Based on a comparison of the relative frequencies, the modals in Table 3 are divided into two groups. The first group consists of the modals *lahko* ("possibly"), *verjetno* ("likely"), and *možno* ("possibly"). Each modal in this group is more frequent in the biotechnical, physical sciences, and technological sciences than in the humanities and social sciences, as indicated by the BPT:HS ratio reported in the fourth column. On the whole, this group is 1.1 times more frequent in BIO, PHYS, and TECH than it is in HUM and SOC.

The second group consists of 9 modals, that is *morda* ("possibly"), *zagotovo* ("certainly"), *gotovo* ("certainly"), *nedvomno* ("certainly"), *mogoče* ("possibly"), *najbrž* ("likely"), *domnevno* ("likely"), *morebiti* ("possibly"), and *menda* ("possibly"). Each modal in this group is more frequent in the humanities and social sciences than in the biotechnical, physical, and technological sciences; on the whole, this group is 2.2 times more frequent in the humanities and social sciences.

Table 3: *Modal adverbs in KAS-dr across academic disciplines*

	HUM, SOC		BIO, PHYS, TECH					
MODAL	AF	RF	AF	RF	BPT:HS	LLV	*p*	DIN
lahko	194,386	2,850	119,639	3,015	1.1	234.167	0.0000	−2.817
verjetno	8,635	127	5,089	128	1.0	0.539	0.4627	−0.649
možno	760	11	713	18	1.6	82.812	0.0000	−23.45
∑	203,781	2,988	125,441	3,161	1.1	247.631	0.0000	−2.825

	HUM, SOC		BIO, PHYS, TECH					
MODAL	AF	RF	AF	RF	HS:BPT	LLV	*p*	DIN
morda	8,028	118	2,123	54	2.2	1198.072	0.0000	37.497
zagotovo	2,655	39	844	21	1.9	257.012	0.0000	29.329
gotovo	2,695	39	568	14	2.8	590.887	0.0000	46.811
nedvomno	2,223	33	448	11	3.0	518.854	0.0000	48.542
mogoče	1,449	21	593	15	1.4	54.460	0.0000	17.406
najbrž	891	13	227	6	2.2	142.948	0.0000	39.088
domnevno	665	10	173	4	2.5	102.498	0.0000	38.199
morebiti	821	12	187	5	2.4	160.011	0.0000	43.726
menda	306	4	12	0	6.0	202.431	0.0000	87.369
∑	19,733	289	5,175	130	2.2	2994.528	0.0000	37.855

To check for statistical significance, we have tested the individual distributions using *Calc: Corpus Calculator* (Cvrček, 2021), an online statistical tool that offers a module for evaluating whether the difference between a pair of absolute frequencies is statistically significant. We report the log-likelihood values (LLV) for each pair of frequencies and the associated *p* values calculated by the module, where the cut-off point for significance is $p < 0.05$. The calculation of the log-likelihood score is based on Andrew Hardie's implementation of Ted Dunning's (1993) original formula (Václav Cvrček, p.c.) and is as follows:

$$2 \times (O_1 \times \ln(\frac{O_1}{E_1}) + O_2 \times \ln(\frac{O_2}{E_2}))$$

where O_1 and O_2 are the observed absolute frequencies and E_1 and E_2 the expected frequencies. In Table 3, all the differences in the absolute pairwise frequencies are significant except for *verjetno*; LLV = 0.539, $p = 0.4627 > 0.05$.

However, as noted by Fidler and Cvrček (2015, p. 226), a problem of large corpora is that the *p*-value of a test does not take into account the practical importance (effect size) of the difference – i.e., "the larger the amount of data, the higher the likelihood that the resulting difference is significant" (2015, p. 227). To take the effect size into account, Table 3 also reports the Difference Index (DIN; also calculated by *Calc*) in the last column. DIN is calculated with the following formula (2015, 230):

$$100 \times \frac{RelFq(AF_{\text{HUM,SOC}}) - \text{RelFq}(AF_{\text{BIO,PHYS,TECH}})}{RelFq(AF_{\text{HUM,SOC}}) + \text{RelFq}(AF_{\text{BIO,PHYS,TECH}})}$$

The values of DIN range from –100 to 100, where –100 would mean that the word is present only in BIO, PHYS, and TECH; 0 would mean that the word occurs equally often in HUM and SOC on the one hand and BIO, PHYS, and TECH on the other, and 100 would mean that the word occurs only HUM and SOC.

In Table 3, the DIN values for all the 3 modals in the first group are negative, which reflects the fact that they occur more frequently in PHYS, SOC, and TECH. The –2.825 score for the overall difference for this group reflects the small BPT:HS ratio. Conversely, the DIN scores for the second group are much higher, where the overall difference between HUM and SOC on the one hand

and BIO, PHYS, and TECH on the other has a DIN score of 37.855, reflecting the much higher HS:BPT ratio in this group.

5.2 Comparison of Epistemic and Non-Epistemic Usage Across Disciplines

In order to gain more insight into the pattern observed in the previous section, according to which 9 out of the 12 analysed modal adverbs occur most frequently in the humanities and social sciences in *KAS-dr* while the remaining adverbs are more prominent in the biotechnical, physical, and technological sciences, we have manually classified a randomized set of 250 concordance examples for each of the 12 adverbs into one of the three categories:

a) epistemic modality;

b) meta-discursive root modality; or

c) dispositional root modality.

The results of the concordance analysis are presented in Table 4.[12] It shows that the distribution of epistemic and non-epistemic meanings of the adverbs generally follows the distribution of the modals between the academic disciplines (Table 3). Eight modals, namely *morda, najbrž, zagotovo, nedvomno, domnevno, gotovo, morebiti*, and *menda*, are used almost exclusively to denote epistemic modality. The modal *mogoče* is also used mostly as an epistemic modal (60% of the concordance). Crucially, all these modal adverbs are precisely those which are more frequently used in the humanities and social sciences (cf. the second group in Table 3). By contrast, the modals *možno* and *lahko*, which are more prominent in natural and technical sciences, infrequently convey the epistemic meaning (11% of the concordances in the case of *lahko* and 2% of the concordances in the case of *možno*). An exception is the modal *verjetno*, which despite its purely epistemic meaning is

12 Note that, in Table 4, the number of included concordances for each modal is not always exactly 250, like 248 in the case of *možno*. The lower number in these cases is due to a few instances of incorrect part-of-speech tagging in the corpus (e.g., some syncretic premodifying adjectives, like *možno* in the accusative/instrumental NP *možno analizo* "possible analysis", are incorrectly tagged as adverbs); we have discarded such irrelevant occurrences from our analysis. Furthermore, *menda* had the largest number of irrelevant examples (i.e., 49), all of which were sentences in which the modal was used in a quoted context, so it did not reflect the author's perspective.

Table 4: *The epistemic/root distribution of the modal adverbs in KAS-dr*

MODAL	EPISTEMIC		META-DISCURSIVE		DISPOSITION	
	Freq.	**%**	**Freq.**	**%**	**Freq.**	**%**
lahko	25	11%	105	42%	117	47%
verjetno	250	100%	0	0%	0	0%
možno	6	2%	9	4%	233	94%
morda	240	96%	7	4%	0	0%
najbrž	250	100%	0	0%	0	0%
zagotovo	243	100%	0	0%	0	0%
nedvomno	250	100%	0	0%	0	0%
mogoče	150	60%	3	1%	97	39%
domnevno	250	100%	0	0%	0	0%
gotovo	245	98%	5	2%	0	0%
morebiti	250	100%	0	0%	0	0%
menda	201	99%	2	0%	0	0%

more prominent in the natural and technical sciences. In the remainder of this section, we take a closer look at the results of the annotation process for each of the three categories and relate the use of modality to the notion of hedging that was introduced in Section 2.2.

5.2.1 Epistemic Modality

Let us first take *morda*, which is used as an epistemic modal in 240 (96%) of the randomized concordances and only in 7 (4%) as a non-epistemic modal in the meta-discursive sense, as being representative of the group that is almost exclusively epistemic. Sentence (8), which is taken from a thesis defended at the Faculty of Social Sciences at the University of Ljubljana, exemplifies this epistemic usage.

(8) *Morda* je to eden od razlogov, da znanstvena skupnost ni bila uspešna pri svojem "programu" izboljšanja javnega razumevanja znanosti in znanstvene pismenosti.
"Perhaps this is one of the reasons that the scientific community wasn't successful in implementing their proposed program for improving the public understanding of science and scientific literacy."

Pragmatically, this corresponds to Hyland (1996, pp. 256–257)'s notion of an *accuracy-based hedge*, as it is used by the writer to denote their uncertainty about the validity of the proposition in the example; i.e., that whatever is denoted by the demonstrative *to* "this" in the main clause is indeed one of the reasons for the lack of success on part of the scientific community.

Similarly, *menda* and *domnevno* are also used mainly as epistemic modals in the sense that they convey the author's uncertain about what they are claiming. However, in contrast to *morda*, the adverbs *menda* and *domnevno* are additionally used to signal that the claim is an assumption, possibly one that is shared within the author's research community.[13] Sentence (9), which is taken from a thesis defended at the Faculty of Arts at the University of Maribor, exemplifies this usage:

(9) Klun je nato v svojem govoru zavrnil očitke, da je bil pobudnik interpelacij, kot je to *menda* trdil Schwegel.
"In his speech, Klun then denied the accusations that he was the instigator of the interpellations, as was supposedly claimed by Schwegel."

In this example, the writer uses *menda* to signal that it is not universally certain whether Schwegel indeed claimed that Klun had been the instigator of whatever the interpellations were, but that it is merely assumed that he made the claim; because *menda* thereby conveys the author's uncertainty (although with an additional assumptive meaning lacking with *morda*), its role in terms of hedging is also accuracy-based in Hyland (1996)'s terms.

All the epistemic examples with the remaining modals (which we do not exemplify here due to space constraints) also function as similar accuracy-based hedges, where the sole semantic and pragmatic difference is in the modal force of the lexeme in question; that is, a modal like *najbrž* "likely" denotes a greater degree of the speaker's commitment to the truth of the proposition than *morda* or *morebiti* "possibly".

13 As Pihler Ciglič (2017) notes, there is an on-going debate in the literature whether evidential/hearsay modals like *menda* and *domnevno* constitute a category that is distinct from other epistemic modals. We follow Palmer (2001) and von Fintel and Gillies (2007) in assuming that the evidential adverbs we analyse are an epistemic subtype since they invariably signal the speaker's uncertainty. In any case, this is a complex issue that hinges on quite a few technical and formal assumptions about modality; see Portner (2009, section 4.2.2) for a good overview of this issue.

5.2.2 Meta-Discursive Root Modality

Sentence (10), taken from a thesis defended at the Faculty of Pedagogy at the University of Ljubljana, exemplifies one of the few cases of the non-epistemic meta-discursive use of *morda*.

(10) Zato lahko *morda* na tem mestu poudarim strinjanje z Banduro (1997), da je samoučinkovitost precej povezana s samouravnavanjem […]
"This is why I can (perhaps) emphasise my agreement with Bandura (1997) that self-effectiveness is related to self-regulation."

In contrast to its epistemic use in (8), *morda* in this sentence clearly does not denote the writer's uncertainty and could be freely omitted from the sentence without a change in the propositional truth-commitment. It is rather used as part of a meta-discursive strategy with which the writer "acknowledge[s] the reader's role in ratifying knowledge" (Hyland, 1996, p. 258), in the sense that the lexical meaning of possibility, which is inherently entailed by the modal, "subtly hedges the universality of a writer's claim by implying that a position is an individual interpretation" (*ibid.*).

Such meta-discursive use is most prominent with the modal *lahko*, having been observed in 105 (42%) out of a total 250 of the randomized set of concordances. The sentence in (11), which is taken from a thesis from the Biotechnical Faculty at the University of Ljubljana, exemplifies this usage.

(11) Zaključimo *lahko*, da alkidni premazi na osnovi organskih topil izkazujejo nižje kontaktne kote na obeh substratih kot vodni akrilni premazi […]
"We can conclude that alkyd coatings on the basis of organic solvents show smaller contact angles on both substrates than aqueous acrylic coatings…"

In all the 105 examples with the meta-discursive use of *lahko*, the modal adverb is used with directive verbs that are inflected for the so-called inclusive plural, like *zaključimo* "we conclude" in example (11). According to Takimoto (2015, p. 99), the use of "inclusive pronouns (e.g., *we*) […] enables the writers to produce more interpersonal signals to the readers, which may allow the writers to share contexts with the readers and draw on their assumed belief

specific to a particular field of study". In other words, the inclusive inflection emphasises the meta-discursive use of *lahko* as a hedge that is reader-oriented rather than accuracy-oriented (Hyland, 1996). Note that the remaining modals which are also used in this meta-discursive role (*mogoče, možno, morda, zagotovo, morebiti*, *menda*) do not pattern with the inclusive plural inflection (cf. example (10), where the first person is used) as consistently, which may possibly correlate with the fact that their use in this role is much less frequent in comparison to *lahko*, this being the de-facto modal for expressing meta-discursive commentary.

5.2.3 Dispositional Root Modality

Finally, we turn to the dispositional root modality of *lahko*, *mogoče*, and *možno*. Sentence (12), which is taken from a thesis defended at the Faculty of Medicine at the University of Ljubljana, exemplifies this meaning with the modal *možno*, which is by far the most frequently used in this sense (233 or 94% examples), while sentence (13), which is from a thesis in the former Faculty of Electrical Engineering, Computer Science and Information Sciences at the University of Ljubljana, contains the modal *mogoče*, which is used in the dispositional sense in 97 (39%) of the concordance examples.[14]

(12) Upliniti je *možno* najrazličnejšo biomaso (les, oglje, kokosove olupke, riževe lupine).
"It is possible to gasify many kinds of biomass (wood, charcoal, coconut peels, rice husks)."

(13) Celoten grafični vmesnik je zasnovan tako, da ga je *mogoče* hitro prilagoditi potrebam metode […]
"The entire GUI is designed in such a way that it can be easily tailored to the needs of the method."

14 In standard descriptive Slovenian linguistics, the lexemes *možno* and *mogoče* are usually referred to as adverbs in sentences like (12) and (13); see, e.g., the *Dictionary of Standard Slovenian* entry for *možno* (Bajec et al., 2014). Note, however, that in both examples *možno* and *mogoče* require that the VP be infinitival. It would therefore be more precise to analyse the two lexemes as predicative adjectives, on par with those heading extrapositional *it*-constructions in English like *It is possible to*+VP_{inf} (Van linden and Davidse, 2009). Conversely, adverbs in clausal adjunct positions are unable to govern the syntactic properties of other sentential constituents in such a way.

In such cases, the modals are used to denote possibility in its root non-epistemic sense. This kind of modality is not concerned with the knowledge or attitude of the writer (as in the case of epistemic modals and those used in the meta-discursive sense), but is rather used to convey the characteristic properties (i.e., the disposition) on the basis of which the underlying subject NP can be used in some way; for instance, example (13) says that the GUI is such that it is possible to tailor it to the needs of whatever is the method in question.

Palmer (2014, p. 38) claims that such subject-oriented modality is actually "not strictly a kind of modality at all, modality being essentially subjective", and that such modals are used "to make purely objective statements about the subject of the sentence" (*ibid.*). From the perspective of pragmatics, it does not seem that such dispositional modals actually constitute hedging of any kind given that they are used to convey objective properties of what the authors are describing in a given example. It should be noted that Hyland (1998, p. 5) claims that "hedges are the means by which writers can present a proposition as an opinion rather than a fact: items are only hedges in their epistemic sense, and only when they mark uncertainty". Examples (12) and (13) do not involve the speaker's opinion one way or the other; hence, they are not hedges. Lastly, we note that *možno* is used the most frequently in the BIO, PHYS, and TECH disciplines out of all the observed modals (see Table 3). We speculate that because it is used almost exclusively as a non-attitudinal dispositional modal, it is also well suited for the natural sciences, which are generally objective in that they deal "with numerical data, which is more likely to generate a more precise picture of their findings" Takimoto (2015, p. 95) than, e.g., the presumably more subjective and less empirical humanities.[15]

5.2.4 Discussion

With the manual concordance analysis, we have shown that adverbs which mainly convey epistemic modality (and thus pragmatically function as

15 We do note, however, that the empirical vs. non-empirical divide partially transcends the distinction between humanities/social sciences on the one hand and natural/technical sciences on the other, but is rather influenced by the methodological framework adopted by the researcher. Thus, a thesis in a humanities discipline may be more concerned with empirical data than other theses in the same discipline.

accuracy-based hedges) are exactly those that are more frequent in the humanities and social sciences in our corpus. This result is generally consistent with related studies that compare the use of adverbial hedging between humanities disciplines on the one hand and natural sciences on the other. For instance, Takimoto (2015, p. 105) shows that, in his corpus, the English adverbs of epistemic possibility are used two times more frequently in the humanities than they are in the natural sciences. Similarly, Rizomilioti (2006, p. 64) shows that adverbs of uncertainty are used 1.2 times more frequently in her literary criticism corpus than in her comparable biology corpus, whereas the difference we have shown is even greater – on average, all the mainly epistemic modals (except for *verjetno*) in our corpus are 2.2 times more frequent in the humanities and social sciences.

Lastly, a note on *verjetno*: this modal is on average the most frequent in natural sciences discourse despite its purely epistemic meaning, as shown in Tables 3. We speculate that this is because *verjetno* does not seem to be completely synonymous with *najbrž*, which also entails likelihood. *Verjetno* seems to have a stronger evidential meaning, in the sense that it conveys that the speaker has some empirical evidence for judging the given proposition as likely, whereas *najbrž* seems more rooted in introspective speculation. A similar claim has been made for the distinction between the certainty modal auxiliaries in English, where the "difference between *will* and *must* is that *will* indicates what is a reasonable conclusion, while *must* indicates the only possible conclusion on the basis of the evidence available" (Palmer, 2014, p. 57).

To see whether *verjetno* truly has a stronger evidential meaning than *najbrž*, we have used the Collocations tool in the *noSketch Engine*, with which *KAS-dr* can be queried online. This tool allows us to observe how the two keywords differ in the collocates (i.e., co-occurring lexemes) that they pattern with, thus revealing larger co-textual differences between them. In the BIO subset of *KAS-dr*, the top-ranking collocates of *verjetno*, based on the MI Score,[16] are words directly related to empirical phenomena in biomedicine, such as *nevroinvazije* ("neuroinvasion"), *nepatogen* ("non-pathogenic"), and *polieter* ("polyether"), while the top-ranking collocates of *najbrž* are non-empirical,

16 The MI score "expresses the extent to which words co-occur compared to the number of times they appear separately" (https://www.sketchengine.eu/guide/glossary/).

meta-discursive expressions like *učinki* ("effects"), *posledica* ("consequence"), and *dejavnikov* ("factors"). If *verjetno* truly has a stronger evidential meaning than *najbrž*, as is hinted at by its collocational profile, then it comes as no surprise that it is the most frequent in biomedical sciences, where empirical evidence abounds.

5.3 Comparison of Epistemic Modal Adverbs Across Academic Stages

In this section, we compare the use of hedging in bachelor's, master's, and doctoral theses in *KAS-dipl*, *KAS-mag*, and *KAS-dr*, respectively. We do this for the following 9 modal adverbs: *verjetno, morda, zagotovo, gotovo, nedvomno, najbrž, domnenvo, morebiti*, and *menda*. These are the modals that almost exclusively (i.e., in more than 96% of the analysed concordances; see Table 4) convey epistemic modality, as was discussed in the previous section.[17] Because of their epistemic meaning, these modals invariably constitute *accuracy-based hedges* (Hyland, 1996) in terms of discourse pragmatics. Consequently, their distribution across the three *KAS* subcorpora offers a window into how authors' stance in relation to truth commitment changes from early (i.e., bachelor's and master's theses) to more proficient academic writing (i.e., doctoral theses).[18] Their distribution across the disciplines is also independent of thesis type, which is shown in Table 5, where each modal (save for *verjetno* in *KAS-dr*) is more frequent in the HUM and SOC disciplines than in BIO, PHYS and TECH in all the three subcorpora of *KAS*.

In Table 6, we now compare the frequencies of the 9 hedging adverbs between the bachelor's theses in *KAS-dipl* and master's theses in *KAS-mag*. The size of *KAS-dipl* is 1,101,796,659 tokens, while the size of *KAS-mag* is 495,827,656 tokens.

The frequencies of all the hedging adverbs are generally stable in both the bachelor's theses in *KAS-dipl* and the master's theses in *KAS-mag*. Overall, there is a negligible 0.6% decrease in the frequency of hedging from bachelor's

17 This is also independent of thesis type; for instance, *morda* in *KAS-dipl* is used as an epistemic modal in 97% cases in a random sample, which is similar to its modal-sense distribution in *KAS-dr* in Table 4.

18 For this reason, we omit the modals *lahko*, *možno*, and *mogoče* in this section. That is, they are not used exclusively in their epistemic sense and thus do not always relate to the authors' stance; see also the discussion of *možno* in the previous section.

Table 5: *The relative frequencies of the modals normalized to a million tokens in the 3 KAS subcorpora*

	KAS-dipl		*KAS-mag*		*KAS-dr*	
MODAL	**HS**	**BPT**	**HS**	**BPT**	**HS**	**BPT**
verjetno "likely"	110	89	105	94	127	128
morda "possibly"	95	57	91	57	118	54
zagotovo "certainly"	50	33	49	34	39	21
gotovo "certainly"	34	18	30	15	40	14
nedvomno "certainly"	29	12	28	13	33	11
najbrž "likely"	12	7	10	6	13	6
domnevno "likely"	6	3	5	4	10	4
morebiti "possibly"	9	6	11	7	12	5
menda "possibly"	2	1	2	0	4	0
∑	347	226	331	230	396	243

Table 6: *Hedging adverbs in bachelor's theses (KAS-dipl) and master's theses (KAS-mag)*

	KAS-dipl		*KAS-mag*				
MODAL	**AF**	**RF**	**AF**	**RF**	**LLV**	***p***	**DIN**
verjetno "likely"	115,248	105	51,487	104	1.892	0.1690	0.364
morda "possibly"	93,030	84	41,983	85	0.228	0.6325	–0.141
zagotovo "certainly"	49,783	45	22,932	46	8.520	0.0035	–1.166
gotovo "certainly"	32,710	29	13,425	27	81.751	0.0000	4.601
nedvomno "certainly"	27,058	25	12,519	25	6.561	0.0104	–1.387
najbrž "likely"	11,849	11	4,548	9	85.103	0.0000	7.938
domnevno "likely"	5,509	5	2,168	4	28.515	0.0000	6.695
morebiti "possibly"	9,028	8	4,853	10	97.841	0.0000	–8.863
menda "possibly"	2,019	2	639	1	63.710	0.0000	17.42
∑	346,234	314	154,554	312	7.024	0.008	0.405

theses (314 tokens per million) to master's theses (312 tokens per million). We have again used the *Calc: Corpus Calculator* (Cvrček, 2021) tool to compare the absolute pairwise frequencies statistically. The log-likelihood values (LLV), the related *p* scores, and the difference indices (DIN) calculated by the tool are given in the last three columns in Table 6 (see also Section 5.1 for how the LLV and DIN values are calculated). All the differences are statistically significant except for *verjetno* (LLV = 1.892; $p = 0.1690 > 0.05$) and *morda*

(LLV = 0.228; p = 0.6325 > 0.05). A negative DIN value indicates that the modal is more frequent in the second group (i.e., master's theses), while a positive value indicates that the modal is more frequent in the first group (i.e., bachelor's theses), though the closer the value is to 0, the less prominent is the difference. The DIN value for the overall difference (LLV = 7.024; p = 0.008 < 0.05) is 0.405, which reflects the fact that the epistemic modal adverbs are generally used at roughly the same frequency in bachelor's theses and in master's theses.

In Table 7, we compare the use of hedging adverbs between the bachelor's theses in *KAS-dipl* and the doctoral theses in *KAS-dr*. The size of *KAS-dr* is 101,473,395 tokens.

Table 7: *Hedging adverbs in bachelor's theses (KAS-dipl) and doctoral theses (KAS-dr)*

	KAS-dipl		***KAS-dr***				
MODAL	**AF**	**RF**	**AF**	**RF**	**LLV**	***p***	**DIN**
verjetno "likely"	115,248	105	12,958	128	439.879	0.0000	−9.943
morda "possibly"	93,030	84	9,727	96	137.020	0.0000	−6.336
zagotovo "certainly"	49,783	45	3,291	32	374.346	0.0000	16.429
gotovo "certainly"	32,710	30	3,152	31	5.816	0.0159	−2.262
nedvomno "certainly"	27,058	24	2,534	25	0.644	0.4221	−0.836
najbrž "likely"	11,849	11	1,082	11	0.072	0.7880	0.427
domnevno "likely"	5,509	5	969	10	296.129	0.0000	−31.268
morebiti "possibly"	9,028	8	811	8	0.465	0.4952	1.246
menda "possibly"	2,019	2	315	3	66.565	0.0000	−25.762
Σ	346,234	314	34,839	344	242.231	0.0000	−4.423

All the hedging adverbs (except for *zagotovo, najbrž*, and *morebiti*) are used more frequently in doctoral theses than in bachelor's theses. Overall, there is a 9.5% increase in the frequency of hedging from bachelor's theses (314 tokens per million) to doctoral theses (344 tokens per million). All the differences are statistically significant except for *nedvomno* (LLV = 0.644; p = 0.4221 > 0.05), *najbrž* (LLV = 0.072; p = 0.7880 > 0.05), and *morebiti* (LLV = 0.465; p = 0.4952 > 0.05). The DIN value for the overall difference (LLV = 242.231; p = 0.0000 < 0.05) between bachelor's and doctoral theses is −4.423, which reflects the fact that doctoral theses employ the adverbs more frequently. In

sum, while hedging adverbs are used almost equally frequently in bachelor's and master's theses, their use increases in doctoral theses.

In Section 3.3, we saw that related work done in the context of English academic writing reports significant differences in hedging between different stages of the writers' academic progress. Aull and Lancaster's (2017) report results similar to ours in Table 7 in that they also see an increase in the use of hedging devices from less mature forms of academic writing such as students' research papers to more mature forms such as published journal papers. They interpret this difference by claiming that advanced academic writers are more likely to avoid an assertive stance in presenting their research than less experienced writers, favouring an approach to writing that is "implicitly attitudinal" and "open to other views in the surrounding discourse" (*ibid.*).

We propose that this also explains why hedging adverbs are more frequent in Slovenian doctoral theses (Table 7) in comparison to bachelor's and master's theses (Table 6). Relatedly, we speculate that the lack of such an increase from bachelor's theses to master's theses is because bachelor's theses together with master's theses constitute a uniform group in relation to research content and academic maturity. That is, most of the master's theses in *KAS-mag* (roughly 80%) are post-Bologna-reform master's theses that are in terms of academic maturity similar to the pre-Bologna bachelor's theses, in the sense that they are not (post)graduate research dissertations in contrast to doctoral theses.

This difference is evidenced in the official guidelines for (post)graduate programmes that are based on Slovenia's Higher Education Act, in which the aims of post-Bologna master's theses are more broadly defined than those of doctoral theses. For instance, according to the guidelines of the Faculty of Economics at the University of Ljubljana,[19] a master's thesis must present results that are "either achieved by the candidate's independent research or his or her expert evaluation of previous work". By contrast, similar guidelines for doctoral studies specify the aims of a doctoral thesis in narrower terms, in that it must necessarily present an original scientific contribution.[20] It is fur-

19 See Article 4 in http://www.ef.uni-lj.si/media/document_files/katalog_info_jav_znacaja/PravilaOMagistrskihDelihBolonjskiMagistrskiPrograms.pdf. (Accessed on 4 January 2020.)

20 See Article 35 in https://www.pef.uni-lj.si/fileadmin/Datoteke/Pravni_akti/Pravilnik_o_podiplomskem_%C5%A1tudiju_3.stopnje.pdf. (Accessed on 4 January 2020.)

thermore noteworthy that, at the University of Ljubljana, doctoral students (but not bachelor's and master's students) are required to publish at least one scientific paper in a peer-reviewed scientific journal before they are allowed to defend their thesis.

Post-Bologna master's theses may thus include only a discussion and evaluation of related work and need not present original research, whereas doctoral students hedge their novel claims in order to "negotiate solidarity with a reader who [might] hold contrary points of view" (Aull and Lancaster, 2014, 154), a pragmatic goal that is especially important in the context of peer review. In other words, it is precisely because Slovenian doctoral students are expected to present novel research that they more frequently employ accuracy-based hedges like the surveyed modal adverbs than undergraduate students writing bachelor's or post-Bologna-reform master's theses.

We wanted to confirm this by comparing the pre-Bologna master's theses, which used to be scientific works, with the post-Bologna master's theses, which inherited the old university diploma status of the concluding requirement at the undergraduate level. Although the *KAS-mag* subcorpus is not marked up for metadata that would distinguish these two master's thesis types, it is possible to demarcate them by publication date. The Bologna reform started to be implemented in Slovenia in 2004, so all the theses prior to this date must necessarily correspond to the old pre-Bologna scientific master's thesis. The pre-Bologna master's programme was gradually phased out in the 2010s, and the master's students enrolled in this system had to defend their theses by the end of the academic year of 2015/2016; consequently, all the theses in the last two publication dates in the subcorpus – 2017 and 2018 – correspond to the post-Bologna master's theses. (Conversely, the master's theses published in the remaining period – especially after 2010 and before 2016 – may correspond to either variant and it is difficult to distinguish between the two given the lack of mark-up, although the post-Bologna theses seem to be in the majority.)

By limiting our query to these two periods (2001–2004 and 2017–2018) in *KAS-mag*, which has yielded 449 theses (17,819,133 tokens) in the pre-Bologna subset and 2647 theses (65,764,329 tokens) in the post-Bologna subset, we are able to determine whether the frequency of hedging adverbs

changes between post-Bologna master's theses published in 2017–2018 and the pre-Bologna theses published in 2001–2004. The comparison is shown in Table 8.

Table 8: *The relative frequencies of hedging adverbs (per one million tokens) in KAS-mag*

	post-Bologna (2017–2018)		pre-Bologna (2001–2004)				
MODAL	**AF**	**RF**	**AF**	**RF**	**LLV**	***p***	**DIN**
verjetno "likely"	6,890	105	2,261	127	60.428	0.0000	–9.548
morda "possibly"	5,395	82	1,426	80	0.696	0.4039	1.240
zagotovo "certainly"	2,956	45	618	35	36.333	0.0000	12.893
gotovo "certainly"	1,186	18	961	54	586.587	0.0000	–49.881
nedvomno "certainly"	943	14	713	40	392.534	0.0000	–47.236
najbrž "likely"	457	7	167	9	10.424	0.0012	–14.845
domnevno "likely"	308	5	25	1	47.438	0.0000	54.897
morebiti "possibly"	585	9	156	9	0.0314	0.8593	0.798
menda "possibly"	74	1	39	2	10.410	0.0013	–32.09
∑	18,794	286	6,366	357	228.236	0.0000	–11.116

Note. The 2017–2018 theses are all post-Bologna master's theses, while the 2001–2004 theses are all pre-Bologna master's theses.

The majority of the hedging adverbs (5 out of 9) are more frequent in pre-Bologna master's theses (the so-called scientific masters), especially *gotovo* (DIN = –49.881) and *nedvomno* (DIN = –47.236), which are three times more frequent in the pre-Bologna subset. The frequency of two of the hedging adverbs, *morda* (DIN = 1.24) and *morebiti* (DIN = 0.798), is stable in both subsets, and their differences are not statistically significant ($p > 0.05$). There are only two hedging adverbs, *zagotovo* (DIN = 12.893) and *domnevno* (DIN = 54.897), which are more frequent in the post-Bologna theses. In total, pre-Bologna master's theses published before 2004 employ the hedging adverbs 24% more frequently than the post-Bologna master's theses published after 2017, which is a an even greater difference (LLV = 228.236; $p = 0.0000 < 0.05$; DIN = –11.116) than the one observed from bachelor's theses to doctoral theses reported in Table 7. This confirms our hypothesis that hedging is more common in original scientific contributions as is the case with doctoral and the pre-Bologna master's theses, which are in Slovenia referred to as *znanstveni*

magisterij ("scientific master's degree"), in contrast to their post-Bologna counterparts, which are referred to as *strokovni magisterij* ("professional/expert master's degree").

8 CONCLUSION

In this paper, we have first analysed modal adverbs in the 100-million-token *KAS subcorpus of Slovenian doctoral theses*, comparing their frequency and use between humanities and social sciences on the one hand and natural sciences and technical sciences on the other. As one of our main contributions to research on hedging, we have taken into account the fact that modals are in actual usage often unpredictably ambiguous between epistemic and non-epistemic readings, and argued that only those modals that either convey epistemic judgements or meta-discursive commentary also function as hedges, whereas those that express dispositional possibilities do not. On the basis of this distinction, we have shown that the modals that are mainly used in the epistemic sense (and that thereby constitute accuracy-based hedges displaying varying degrees of the authors' tentativeness about the truth of the proposition) are used more frequently in Slovenian doctoral theses in the humanities and social sciences rather than the natural and technical sciences, which is generally in line with the related work (e.g., Takimoto, 2015; Hyland, 1998).[21]

Next, we have compared the use of the exclusively epistemic modal adverbs in theses at different stages of university education: bachelor's, master's and doctoral theses. We have shown that such modals are more frequent in doctoral theses than in bachelor's and master's theses, which is in line with the increase in hedging observed by Aull and Lancaster (2014) from first-year undergraduate writing to published research articles in the context of

21 It is difficult to say to what degree this trend can be generalised to hedging expressions other than modal adverbs. A problem here, as mentioned in Section 2.2, is that hedging is a pragmatic strategy and not a linguistic property (in the narrow sense), which means that a hedge can correspond not only to virtually any of the (open class) lexical categories (i.e., adverbs, adjectives, lexical verbs, nouns), but to many syntactic devices as well (the use of voice, mood, impersonalisation devices, etc.). To study this would require a manual analysis of the texts in the corpus, whereas for *KAS*, which is a very large corpus that is not syntactically parsed, we could only rely on the MSD-tags assigned to the tokens. We therefore leave such an analysis for future work.

American English academia. We have argued that such an increase in hedging observed in Slovenian doctoral reflects an important conceptual difference between bachelor's and post-Bologna master's theses on the one hand and doctoral theses on the other – that is, it is only doctoral theses that are research dissertations whose primary aim is presentation of novel research, the careful and responsible interpretation and discussion of which often needs to be properly hedged. We have confirmed this hypothesis by comparing the pre- and post-Bologna master's theses, the status of which has changed with the Bologna process from what was once a scientific degree to what is now a professional degree.

In our future work we would like to extend our analysis of the modals in the *KAS-dr* subcorpus to classes such as epistemic adjectives and verbs, while taking special care to properly account for the way their unique semantics interacts with the pragmatics. This will enable us to further ascertain whether expressions of epistemic modality are really more characteristic of humanities and/or social sciences disciplines across the board, as claimed by Takimoto (2015) and Hyland (1998), or whether they are a quirk of a specific word class, such as adverbs, as is claimed by Rizomilioti (2016). Furthermore, there might be prominent differences in the frequency of hedging between different parts of a thesis; for instance, the section dedicated to the discussion of results might contain many more hedging devices than the section dedicated to the research methodology (see also Thompson 2000 for precisely such findings for English). We also leave this for future work, as the *KAS* corpus is not annotated for thesis sections, nor is any other available Slovenian corpus.

Lastly, the extra-linguistic metadata in the *KAS* corpus also includes author-related information such as the name of the student and the advisor of the thesis. The second analysis presented in this paper could therefore be extended by taking into account how the use of hedging devices, such as epistemic adverbs, changes not only from undergraduate to (post)graduate theses in general, but also in the case of individual authors who first wrote a bachelor's or a master's thesis and then went on to pursue a doctoral degree. This would provide an even greater insight into the developmental trajectory of young Slovenian researchers as they advance through the higher educational system.

Acknowledgments

We would like to thank Maja Miličević Petrović for help with the statistics, and all the anonymous reviewers for their helpful comments. The work described in this paper was funded by the Slovenian Research Agency within the national research programme *Slovene Language – Basic, Contrastive, and Applied Studies* (P6-0215) and within the national basic research project *Slovene Scientific Texts: Resources and Description* (J6-7094, 2014-2017).

REFERENCES

Aull, L. L., & Lancaster, Z. (2014). Linguistic Markers of Stance in Early and Advanced Academic Writing: A Corpus-based Comparison. *Written communication, 31*(2), 151–183. doi: 10.1177/0741088314527055

Aull, L. L., Bandarage, D., & Miller, M. R. (2017). Generality in student and expert epistemic stance: A corpus analysis of first-year, upper-level, and published academic writing. *Journal of English for Academic Purposes, 26*, 29–41. doi: 10.1016/j.jeap.2017.01.005

Bajec, A., et al. (Eds.). (2014). Možno (lexicographic entry). In *Slovar slovenskega knjižnega jezika*.

Coates, J. (1983). *The Semantics of the Modal Auxiliaries*. London and Canberra: Croom Helm.

Crosthwaite, P., Cheung, L., & Jiang, F. K. (2017). Writing with Attitude: Stance expression in learner and professional dentistry research reports. *English for Specific Purposes*, *46*, 107–123. doi: 10.1016/j.esp.2017.02.001

Cvrček, V. (2021). *Calc v1.02: Corpus Calculator*. Czech National Corpus. Retrieved from https://www.korpus.cz/calc/

DeLazero, O. E. (2011). On the Semantics of Modal Adjectives. *University of Pennsylvania Working Papers in Linguistics*, *17*(1), 87–94. Retrieved from https://repository.upenn.edu/pwpl/vol17/iss1/11/

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, *19*(1), 61–74.

Erjavec, T., Fišer, D., & Ljubešić, N. (2019a). *Corpus of Academic Slovene KAS 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1244

Erjavec, T., Fišer, D., & Ljubešić, N. (2019b). *Corpus of Academic Slovene (MSc/MA theses) KAS-mag 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1266

Erjavec, T., Fišer, D., & Ljubešić, N. (2019c). *Corpus of Academic Slovene (doctoral theses) KAS-dr 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1265

Erjavec, T., Fišer, D., & Ljubešić, N. (2019d). *Corpus of Academic Slovene (BSc/BA theses) KAS-dipl 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1267

Erjavec, T., Fišer, D., & Ljubešić, N. (2020). The KAS corpus of Slovenian academic writing. *Language Resource and Evaluation*. doi: 10.1007/s10579-020-09506-4

Erjavec, T. (2012). Mutext-East: Morphosyntactic Resources for Central and Eastern European Languages. *Language Resources and Evaluation*, *46*, 131–143. doi: 10.1007/s10579-011-9174-8

Fidler, M., & Cvrček, V. (2015). A Data-Driven Analysis of Reader Viewpoints: Reconstructing the Historical Reader Using Keyword Analysis. *Journal of Slavic Linguistics, 23*(2), 197–239. Retrieved from https://www.jstor.org/stable/24602151

von Fintel, K. (2006). Modality and language. In D. M. Borchert (Ed.), *Encyclopedia of Philosophy – Second Edition* (pp. 20–27). Detroit: MacMillan Reference USA.

von Fintel, K., & Gillies, A. (2007): An opinionated guide to epistemic modality. *Oxford studies in epistemology*, *2*, 32–63.

Grabe, W., & Kaplan, R. B. (1997). On the writing of science and the science of writing: Hedging in science text and elsewhere. In J. S. Petöfi (Ed.), *Hedging and Discourse* (pp. 151–167). De Gruyter, Berlin and New York.

de Haan, F. (2001). The Relation Between Modality and Evidentiality. *Linguistic Reports*, *9*, 201–216.

Hladnik, M. (2015). *Mind the Gap: Resumption in Slavic Relative Clauses*. LOT Publications. Retrieved from https://www.lotpublications.nl/mind-the-gap-resumption-in-slavic-relative-clauses

Hyland, K. (1996). Talking to the Academy: Forms of Hedging in Science Research Articles. *Written Communication*, *13*(2), 251–281. doi: 10.1177/0741088396013002004

Hyland, K. (1998). *Hedging in Scientific Research Articles*. Amsterdam: John Benjamins.

Hyland, K. (2004). Patterns of engagement: Dialogic features and L2 undergraduate writing. In L. Ravelli & R. A. Ellis (Eds.), *Analysing academic writing: Contextualized frameworks* (pp. 5–23). London, UK: Continuum.

Kratzer, A. (2012). The notional category of modality. In *Modals and Conditionals: New and Revised Perspectives* (pp. 27–69). Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199234684.003.0002

Lancaster, Z. (2016). Expressing stance in undergraduate writing: Discipline-specific and general qualities. *Journal of English for Academic Purposes*, *23*, 16–30. doi: 10.1016/j.jeap.2016.05.006

Lakoff, G. (1972). Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of Philosophical Logic*, *2*(4), 458–508. Retrieved from https://www.jstor.org/stable/30226076

Lenardič, J., & Fišer, D. (2020). Epistemic modal adverbs in Slovenian academic discourse. *Proceedings of the Conference on Language Technologies and Digital Humanities* (pp. 34–41).

Van Linden, A., & Davidse, K. (2009). The clausal complementation of deontic-evaluative adjectives in extraposition constructions: a synchronic-diachronic approach. *Folia Linguistica*, *43*(1), 171–211. doi: 10.1515/FLIN.2009.005

Marušič, F., & Žaucer, R. (2016). The modal cycle vs. negation in slovenian. In F. Marušič & R. Žaucer (Eds.), *Formal Studies in Slovenian Syntax* (pp.167–192). Amsterdam: John Benjamins. doi: 10.1075/la.236.08mar

Palmer, F. R. (2001). *Mood and Modality* (2nd ed.). Cambridge: Cambridge University Press.

Palmer, F. R. (2014). *Modality and the English modals*. Abingdon-on-Thames: Routledge.

Pihler Ciglič, B. (2017). Evidencialna branja prislova dizque v nekaterih različicah ameriške španščine in njegove ustreznice v slovenščini. *Ars & Humanitas*, *11*(2), 85–103. doi: 10.4312/ars.11.2.85-103

Piqué-Angordans, J., Posteguillo, S., & Andreu-Besó, J. V. (2002). Epistemic and Deontic Modality: A Linguistic Indicator of Disciplinary Variation in Academic English. *LSP & Professional Communication*, *2*(2), 49–65.

Pisanski Peterlin, A. (2010). Hedging Devices in Slovene-English Translation: A Corpus-Based Study. *Nordic Journal of English Studies*, *9*(2), 171–193. doi: 10.35360/njes.222

Pisanski Peterlin, A. (2015). So prevedena poljudnoznanstvena besedila v slovenščini drugačna od izvirnih? Korpusna študija na primeru izražanja epistemske naklonskosti. *Slavistična revija*, *63*, 29–44. Retrieved from https://srl.si/ojs/srl/article/view/COBISS_ID-57701986

Portner, P. (2009). *Modality*. Oxford: Oxford University Press.

Rizomilioti, V. (2006). Exploring Epistemic Modality in Academic Discourse Using Corpora. In *Information Technology in Languages for Specific Purposes*, Educational Linguistics, *7*, 53–71. Boston, MA: Springer. doi: 10.1007/978-0-387-28624-2_4

Rowbotham, M., Harden, N., Stacey, B., Bernstein, P., & Magnus-Miller, L. (1998). Gabapentin for the Treatment of Postherpetic Neuralgia: A Randomized Controlled Trial. *JAMA*, *280*(21), 1837–1842. doi: 10.1001/jama.280.21.1837

Takimoto, M. (2015). A Corpus-Based Analysis of Hedges and Boosters in English Academic Articles. *Indonesian Journal of Applied Linguistics*, *5*(1), 95–105. doi: 10.17509/ijal.v5i1.836

Thompson, P. (2000). Modal Verbs in Academic Writing. In B. Kettemann & G. Marko (Eds.), *Teaching and Learning by Doing Corpus Analysis – Proceedings of the Fourth International Conference on Teaching and Language Corpora* (pp. 305–328).

Toporišič, J. (2004). Slovenska Slovnica. Maribor: Založba Obzorja.

NAKLONSKI PRISLOVI KOT PRAGMATIČNI OMEJEVALCI V SLOVENSKIH ZNANSTVENIH BESEDILIH

V članku najprej primerjamo rabo epistemskih naklonskih prislovov v doktorskih disertacijah v humanistiki in družboslovju po eni strani ter naravoslovnih in tehničnih znanosti po drugi v korpusu slovenskih znanstvenih besedil KAS (Erjavec idr., 2019a). Z naključnim vzorčenjem korpusnih zgledov pokažemo, da so tisti naklonski prislovi, ki skoraj izključno izkazujejo epistemski pomen in se posledično uporabljajo kot ti. pragmatični omejevalci (angl. *hedges*), najbolj značilni za doktorske disertacije v humanistiki in družboslovju. Pokažemo tudi, da neepistemski dispozicijski pomen naklonske možnosti, ki se najpogosteje pojavlja v naravoslovju in tehničnih vedah, ni rabljen kot pragmatični omejevalec. V drugem delu članka primerjamo rabo epistemskih naklonskih prislovov v diplomskih in magistrskih delih ter doktorskih disertacijah z namenom, da ugotovimo, ali se pristop do podajanja in prikazovanja izsledkov z vidika pragmatičnega omejevanja v znanstvenem diskurzu spreminja glede na izkušenost avtorjev z znanstvenim pisanjem. Pokažemo, da doktorski študentje pogosteje uporabljajo naklonske prislove v omejevalni funkciji, za kar trdimo, da je posledica vsebinskih in konceptualnih razlik med diplomskimi in magistrskimi nalogami po eni strani ter doktorskimi nalogami po drugi, saj v okviru Bolonjske reforme zgolj slednje morajo obvezno predstaviti izvirni znanstveni prispevek, katerega poglavitni cilj je poglobljena predstavitev novih rezultatov.

Ključne besede: epistemska naklonskost, jedrna naklonskost, pragmatično omejevanje, pomenoslovje, pragmatika, korpusno jezikoslovje

UČNO E-OKOLJE *SLOVENŠČINA NA DLANI*: IZZIVI IN REŠITVE

Darinka VERDONIK, Simona MAJHENIČ, Špela ANTLOGA, Sandi MAJNINGER, Marko FERME, Kaja DOBROVOLJC
Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru

Simona PULKO, Mira KRAJNC IVIČ, Natalija ULČNIK
Filozofska fakulteta, Univerza v Mariboru

Verdonik, D., Majhenič, S., Antloga, Š., Majninger, S., Ferme, M., Dobrovoljc, K., Pulko, S., Krajnc Ivič, M., Ulčnik, N. (2021): Učno e-okolje Slovenščina na dlani: izzivi in rešitve. Slovenščina 2.0, 9(1): 181–215.

DOI: https://doi.org/10.4312/slo2.0.2021.1.181-215

Prispevek izhaja iz treh izzivov, ki jih zaznavamo pri pouku slovenščine v višjih razredih osnovnih šol in v srednjih šolah: kako odpraviti napake knjižne norme, ki vztrajajo v pisnih izdelkih učencev; kako izboljšati frazeološko kompetenco; kako izboljšati sporazumevalno jezikovno zmožnost. Ti izzivi so osrednja točka razvoja sodobnega učnega e-okolja *Slovenščina na dlani*, ki temelji na jezikovnih in informacijsko-komunikacijskih tehnologijah ter prinaša podporo prožnim oblikam poučevanja, poučevanju na daljavo, lajša učiteljevo delo, omogoča pa tudi motiviranje učencev prek elementov igrifikacije. V prispevku predstavljamo zasnovo in izvedbo vsakega od štirih vsebinskih sklopov e-okolja: pravopis, slovnica, frazeologija in besedila.

Ključne besede: učenje slovenščine, računalniško podprto učenje jezika, e-učenje

1 UVOD

Pojem prožne oblike poučevanja (po definiciji Evropske komisije)[1] vsebuje učenje in poučevanje, ki je odzivno na potrebe učečega se ter njegove močne strani.[2] Fleksibilno učenje učečemu se ponuja izbiro načina, okolja in časa učenja z namenom spodbujanja motivacije in vztrajnosti, lahko tudi v primerih, ko je prisotnost na lokaciji izvajanja učnega procesa otežena. Prožne oblike učenja in poučevanja tako omogočajo fleksibilno obliko dela, saj gre za sistem poučevanja in učenja, v katerem imajo učeči se možnost, da del učenja opravijo tudi izven šolskega okolja.

Digitalno okolje ponuja obetavna izhodišča za uresničitev prožnih oblik poučevanja, saj omogoča določeno stopnjo avtomatizacije, ki je lahko učitelju v pomoč pri spremljanju učenčevega napredka, učenca pa lahko delno samodejno vodi skozi učni proces. Z izzivom, kako s pomočjo digitalnega okolja podpreti prožne oblike poučevanja slovenščine v osnovnih in srednjih šolah, smo se spopadli v projektu *Slovenščina na dlani*.[3]

V prispevku predstavljamo učne izzive, ki jih zaznavamo pri pouku slovenščine v osnovnih in srednjih šolah, ter načine, kako se nanje odzvati z uporabo sodobnih jezikovnotehnoloških in informacijsko-komunikacijskih pristopov. V drugem poglavju predstavljamo pregled obstoječih e-pripomočkov za pouk slovenščine ter kako se mednje uvršča e-okolje *Slovenščina na dlani*. V tretjem poglavju opišemo pristopanje k problemu pravopisnih in slovničnih napak, ki so pri mnogih prisotne tudi še po koncu osnovnega in srednjega šolanja. V četrtem poglavju navajamo, kako smo se lotili spoznavanja frazemov in pregovorov ter izboljšanja frazeološke kompetence med mladimi. V petem poglavju predstavimo pripravo sklopa nalog za boljšo sporazumevalno jezikovno

1 Definicija je dostopna na spletni strani: https://www.igi-global.com/dictionary/flexible-learning/11249.

2 Odpiranje izobraževanja: inovativno poučevanje in učenje za vse z novimi tehnologijami in prosto dostopnimi učnimi viri https://eur-lex.europa.eu/legal-content/SL/TXT/PDF/?uri=CELEX:52013DC0654&from=HU.
Strateški okvir – Izobraževanje in usposabljanje 2020: http://ec.europa.eu/education/policy/strategic-framework_sl.

3 Projekt sofinancirata Republika Slovenija in Evropska unija iz Evropskega socialnega sklada. Izvaja se na Filozofski fakulteti, Pedagoški fakulteti in Fakulteti za elektrotehniko, računalništvo in informatiko Univerze v Mariboru.

zmožnost, in sicer tvorjenje ter razumevanje večpredstavnostnih klasičnih ali elektronskih pisanih in govorjenih besedil. V šestem poglavju na enem mestu predstavimo programiranje in vrednotenje vaj ter izdelavo razlag k vajam.

2 RAČUNALNIŠKO PODPRTO UČENJE JEZIKOV IN UČNA E-GRADIVA ZA SLOVENŠČINO

Z računalniki podprto učenje jezikov (angl. *computer assisted language learning* – CALL) ima začetke že v šestdesetih letih 20. stoletja (Davies, 2016); z razširitvijo osebnih računalnikov v sedemdesetih letih je dobilo precejšen zagon, a je bilo sprva omejeno na vnaprej sprogramirana navodila ter je bilo videti kot vnaprej posnet in sprogramiran linearni jezikovni laboratorij. S prelomno objavo Higginsa in Johnsa (1984) je bil predstavljen nabor novih možnosti za alternativne načine uporabe računalnikov pri učenju jezika; razdeljen je bil v štiri skupine: (1) sledenje navodilom (angl. *do what I tell you*), pri katerem računalnik uporabniku pove, kaj mora narediti (vaje z izbiranjem, vpisovanjem, kvizi …); pri tem ni zanemarljivo, da učenca ne zanima samo, ali in zakaj je nekaj narobe, ampak tudi, zakaj je nekaj pravilno, kar pomeni, da je nujno tudi vključevanje razlag; (2) ugibanje (angl. *guess what was there*), pri katerem računalnik izbriše del besedja/besedila, uporabnik pa mora ugotoviti, kaj se je tam nahajalo; (3) računalnik kot pomagalo (angl. *can I help you*?), kjer lahko učitelj pri učenju jezika na inovativne načine uporablja obstoječa računalniška orodja, kot so tezavri, slovarji sopomenk, konkordančniki; (4) računalniške simulacije (angl. *how do I get out of this*?), kamor sodijo razne igre, sestavljanke (npr. sestavljanje besed iz ponujenih črk, delov besed) ipd.

Danes lahko računalniško podprto učenje jezika ločimo glede na načine uporabe računalnika (Davies, 2016). Prvi se nanaša na računalniško okolje, sprogramirano namensko za učenje in utrjevanje jezikovnih vzorcev ali za prepoznavanje in popravljanje uporabnikovih napak. Drugi se nanaša na raziskovanje jezika, pri čemer so pogosto uporabljeno orodje različni konkordančniki in drugi jezikovni viri. Tretji način se nanaša na multimedijsko podprto učenje jezika prek avdio ali video vsebin, ki so pogosto pripravljene tako, da uporabnika vodijo pri učenju jezika (npr. zgoščenke za učenje tujega jezika). Sem lahko sodijo tudi razpoznavalniki govora, ki uporabniku pomagajo odpravljati težave pri branju ali govorjenju v tujem jeziku. Zadnji, četrti način se nanaša

na možnosti učenja jezika, ki jih odpira internet, zlasti z različnimi virtualnimi učilnicami, uporabo posnetkov z Youtuba ipd.

Računalniško podprto učenje se je večinoma razvijalo v povezavi z usvajanjem tujega jezika. V povezavi z učenjem o jeziku (materinščini) pa je ob tem prav tako nastajalo veliko elektronskih učbenikov ali delovnih zvezkov. Učitelji slovenščine tako lahko pri pouku slovenščine uporabljajo kar nekaj različnih e-gradiv in okolij. Založba Rokus svoje učbenike, berila in delovne zvezke ponuja tudi v e-obliki. Za prvo triletje osnovne šole je na voljo izobraževalni portal Lilibi,[4] ki vključuje tudi slovenščino. Za četrti in peti razred ponuja učno serijo Radovednih pet, ki poudarja medpredmetno povezovanje. Vključuje interaktivne samostojne delovne zvezke, interaktivne učbenike in interaktivno berilo. Interaktivno gradivo v napredni obliki je obogateno z videoposnetki, animacijami, interaktivnimi vajami in drugimi dodatki. Celotna serija Rokusovih gradiv je zamišljena na principu t. i. kombiniranega učenja, tj. prepletanja tiskanih in interaktivnih komponent. Založba Mladinska knjiga trži portal UČIMse,[5] ki ponuja interaktivne, grafično oblikovane in igrificirane vaje za celotno osnovno šolo, v posebnem sklopu za razredno stopnjo od 1. do 5. in v posebnem sklopu za predmetno stopnjo od 6. do 9. razreda. Portal Devetka[6] predstavlja zbirko spletnih nalog, v kateri najdemo povezave na različne ponudnike raznovrstnih e-gradiv. Za slovenščino najdemo zelo različne vsebine, od nacionalnih preverjanj znanja do posameznih vaj ali pripomočkov, ki so jih objavili različni avtorji. Na javno financiranem portalu iUčbeniki[7] so pod prosto licenco Creative Commons na voljo interaktivni učbeniki za slovenščino za 8. in 9. razred osnovnih šol in prvi letnik gimnazij. Podobne vrste so tudi e-gradiva Projekta slovenščina,[8] ki so na voljo za 8. razred osnovnih šol in 2. letnik srednjih šol ter gimnazij. Portal Interaktivne vaje[9] vsebuje povezave na interaktivne vaje, tudi iz slovenščine, za celotno osnovno šolo. Nekatere vaje so narejene v okviru portala, pogosto pa nas portal samo preusmeri na drug

4 https://www.lilibi.si/

5 https://www.ucimse.com/

6 http://devetka.net

7 http://eucbeniki.sio.si

8 http://www.s-sers.mb.edus.si/gradiva/w3/slo8/000_mapa/index.html

9 https://interaktivne-vaje.si/index.html

spletni naslov, kjer so vaje na voljo. Pedagoški slovnični portal[10] obravnava teme, ki šolarjem povzročajo največ težav pri pisanju. Izbrana poglavja celostno obdela, od razlage prek primerov do vaj. Ta portal je prvi v slovenskem okolju, ki izkorišča korpuse in korpusne pristope za definiranje tem, oblikovanje razlag in pripravo raznovrstnih vaj za izbrane teme. Temelji na skrbno premišljenem in izvedenem metodološkem postopku (Rozman idr., 2020), pokriva pa manjši del slovničnih problemov.

Navedeni pregled kaže, da med pregledanimi gradivi prednosti digitalnega formata v največji meri izkoriščajo na portalu UČIMse, saj uporabljajo bogato animacijo, grafiko in zvočne učinke; izkoriščen je element igrifikacije, ki vključuje virtualno okolje, nagrajevanje, vodenje skozi vaje z animiranimi junaki ipd. Zelo dobro grafično animacijo in igrifikacijo izkorišča tudi večina vaj, ki so na voljo prek spletnega portala Interaktivne vaje. Pomemben korak naprej pri izrabi potencialov digitalnega medija pa predstavlja tudi Pedagoški slovnični portal. Nabor e-vsebin za učenje slovenščine je v pregledanih gradivih sicer dokaj širok, vendar večinoma osredotočen na osnovno šolo ali celo na razredno stopnjo, za srednje šole je gradiv veliko manj. Opazno je, da se pogosto uporabljajo animacija, grafika, video vsebine in igrifikacija, ni pa še omogočene avtomatizirane individualizacije v smislu, da bi se vsebina in zahtevnost vaj samodejno prilagajali znanju učečega se. Uporabnik se mora tako v veliki količini razpoložljivih vsebin znajti sam ali pa ga mora skoznje voditi učitelj, ki pa ima prav s prilagajanjem dela vsakemu učečemu se največ težav in v tem segmentu potrebuje največ podpore. Osnovni poudarki novega e-okolja *Slovenščina na dlani*, ki ga predstavljamo, so zato: (1) obravnavanje vsebin, ki se usvajajo v višjih razredih osnovne šole in v srednji šoli, (2) samodejno prilagajanje vaj potrebam učečega se in (3) olajšanje učiteljevega dela pri formativnem spremljanju napredka posameznikov.

3 POGOSTE PRAVOPISNO-SLOVNIČNE NAPAKE PRI PISANJU

Dva od štirih vsebinskih sklopov e-okolja *Slovenščina na dlani* se nanašata na napake, ki se pri mnogih tudi še po koncu osnovnega in srednjega šolanja pojavljajo pri pisanju besedil. Na vrsto tovrstnih napak so opozarjali tudi učitelji, sodelujoči v projektu, nekatere pa izpostavljajo tudi strokovnjaki v

10 http://slovnica.slovenscina.eu/

strokovnih objavah, razpravah in priročnikih (Križaj in Bešter Turk, 2018; Gomboc, 2019).

3.1 Vsebinska področja iz pravopisa in slovnice

Pri definiranju tem in vsebin s področja pravopisa in slovnice smo se v projektu oprli na analizo napak v korpusu Šolar (Rozman idr., 2020), ki so jo predstavili Kosem idr. (2012). Na podlagi te analize in na podlagi napak, na katere so opozarjali sodelujoči učitelji, smo definirali vsebinska področja iz pravopisa in slovnice, ki jih obravnavamo v učnem e-okolju. Pri tem smo upoštevali, da lahko napake oz. odstopi od norme nastajajo zaradi: (1) nepoznavanja pravopisnih in slovničnih pravil učečega se, zato ga v navodilih k nalogam usmerjamo k prepoznavanju *pravilnega* oz. *napačnega* zapisa, kar je smiselno ovrednoteno s točkami za pravilni odgovor; (2) neustrezne jezikovne izbire glede na zvrst, zato ga v navodilih k nalogam usmerjamo k prepoznavanju *najprimernejšega* oz. *najustreznejšega* zapisa, za kar bo učeči se s točkami nagrajen šele, ko prepozna najustreznejši zapis; (3) rahljanja jezikovne norme (npr. *bodo* vs. *bojo*), zato ga v navodilih k nalogam usmerjamo npr. k prepoznavanju pogovorne oblike zapisane besede.

Za pravopis te vsebine vključujejo:

- uporabo ločil: končna ločila (vprašaj, tri pike), nekončna ločila (problematiko postavljanja vejice, in sicer pri podredjih – predmetni, osebkov, časovni, krajevni, načinovni, vzročni, pogojni, dopustni in prilastkov odvisnik; pri priredjih – vezalno, stopnjevalno, protivno, ločno, posledično ter pojasnjevalno in sklepalno priredje; pri zahtevnejših primerih z vezniki ter pri pastavkih, pristavkih in vrivkih; pomišljaj; tri pike) in ločila pri premem govoru;
- uporabo velike in male začetnice pri pridevnikih na *-ski* in *-ški* ter *-ov* in *-ev*, pri naselbinskih in nenaselbinskih imenih, pri pisanju imen bitij ter pri stvarnih imenih;
- pisanje skupaj oz. narazen, in sicer pri veznikih, predlogih in členkih, glagolih, pridevnikih, prislovih in zvezah z njimi, pri samostalnikih, zaimkih, številnikih in pri okrajšavah; v tem sklopu je obravnavano tudi pisanje z vezajem;

- zahtevnejše primere zapisa, ki se nanašajo na zapis prevzetih besed, zapis sklopov z neobstojnim in vrinjenim samoglasnikom, zapise s podvojenimi črkami, zapise s sičniki, z zvočniki in *u* ter t. i. besede nagajivke (npr. *stremeti* vs. *strmeti*).

Za slovnico te vsebine vključujejo:

- težave pri pisanju, povezane s samostalniki, in sicer upoštevanje preglasa pri sklanjanju, sklanjanje lastnih imen, zahtevnejše primere sklanjatev (npr. *mati*, *hči*, *gospa*, *možje*, *človek*, *starši*), zanikani rodilnik;
- rabo pridevnikov: določna in nedoločna oblika, stopnjevanje ter knjižna raba svojilnih pridevnikov iz lastnih imen (npr. *Markov* vs. *Markotov*);
- rabo glagolov: dvojinske oblike, oblika sedanjika (npr. *bodo* vs. *bojo*), prihodnjika (npr. *boš* vs. *boš bil*) in preteklega deležnika (npr. *odločil* vs. *odloču*), ujemanje osebka s povedkom, uporaba namenilnika, nedoločnika in kratkega nedoločnika, uporaba glagolov *morati* in *moči* ter *vedeti* in *znati*, uporaba vikanja;
- rabo zaimkov: oziralni zaimek (npr. *ki* vs. *kateri*), svojilni in povratno svojilni zaimek, ujemanje zaimka (npr. *z njim* vs. *z njem*), zaimek v dvojini (problem opuščanja dvojine), zaimek v mestniku ali dajalniku (npr. *njem* vs. *njemu*) in zaimek v rodilniku (npr. *je* vs. *jo*);
- rabo predlogov: *čez*, *do*, *na*, *nad*, *poleg*, *pri*, *skozi* ter *h*, *iz*, *k*, *o*, *s*, *v*, *z* in *za* v kontekstih, kjer se pogosto uporabljajo manj ustrezni ali neustrezni predlogi (npr. *pregovori okrog ljubezni* vs. *pregovori o ljubezni*);
- rabo veznikov: enodelni in enobesedni vezniki (npr. *in*, *ter*, *pa*; *temveč*, *marveč*, *ampak*, *vendar*; *ne* in *brez*); dvodelni in večbesedni vezniki; vsebine so usmerjene v utrjevanje ustreznih vzorcev rabe v kontekstih, kjer učitelji pogosto opažajo neustrezno izbiro veznikov (npr. *Odklonil je tako kosilo in večerjo*);
- besedni red: zaporedje stavčnih členov (npr. veznik *ker* + pomožni glagol + glagol/samostalnik/prislov/zaimek: *Predvsem zato, ker smo prireditev prestavili s torka na soboto* vs. *Predvsem zato, ker*

prireditev smo prestavili s torka na soboto) in naslonski niz (npr. zaporedje *da naj se*).

3.2 Postopek izdelave nalog za vsebine iz pravopisa in slovnice

Pri obravnavi vsebin iz pravopisa in slovnice smo izhajali iz namere, v čim večji meri izkoristiti jezikovnotehnološke pristope in metode za pripravo učnega e-okolja. Strateško smo sledili principu »od prakse k teoriji«, kar pomeni, da učenci ob vajah prepoznavajo, kje jim slabo poznavanje pravopisnih in slovničnih vzorcev ter pravil knjižne norme povzroča težave pri pisanju, da z dodatnimi vajami utrjujejo predvsem ta področja in da ob tem hkrati spoznavajo tudi razlage in razloge, ki stojijo za posameznimi pravili knjižne norme. Pri tem smo v veliki meri izhajali iz rezultatov Pedagoškega slovničnega portala (Rozman idr., 2020; Kosem idr., 2012), a je med obema tudi nekaj ključnih razlik: (1) v e-okolju *Slovenščina na dlani* je poudarek na obravnavi velike količine različnih pravopisnih in slovničnih tem, posledično ni bila narejena podrobna dodatna jezikoslovna analiza posameznih tem, ampak smo se opirali na obstoječe jezikoslovne priročnike; (2) v ospredju so vaje: učeči se vstopa v e-okolje skozi vaje, razlagam je posvečene manj pozornosti; (3) učeči se je avtomatsko voden skozi e-okolje, ni se mu treba odločati, katere vaje bo delal; (4) e-okolje je prilagojeno učiteljem in jim omogoča vodenje in spremljanje učečih se ter komunikacijo z njimi.

Postopek izdelave velike količine vaj za vsebine, predstavljene v poglavju 3.1, je bil izveden po naslednjih korakih:

1. priprava korpusnega gradiva, ki bo osnova za priklic velike količine avtentičnih primerov za vsako vajo – korpus MAKS;
2. definiranje vaj z didaktičnega in jezikoslovnega vidika;
3. priklic primerov za vsako vajo iz zbranega korpusa, oblikovanje baze primerov in ročni pregled primerov;
4. programiranje vaj, omogočanje interaktivnega reševanja vaj in vzpostavitev hranjenja uporabnikovih odgovorov;
5. definiranje in izdelava algoritmov za vrednotenje uporabnikove uspešnosti reševanja vaj;
6. izdelava razlag k vsebinam vaj.

Posamezne korake predstavljamo v nadaljevanju. Korake programiranja in vrednotenja vaj ter izdelave razlag opisujemo v zadnjem delu članka za vse vsebinske sklope e-okolja skupaj.

3.3 Priprava korpusnega gradiva – korpus MAKS

Korpus MAKS (akronim za MlAdinski KorpuS) obsega pribl. 10 mio. besed oz. pribl. 12 mio. pojavnic. Dobro polovico tega sestavljajo besedila iz mladinskega in drugega leposlovja ali priročnikov. Dobrih 40 % besedil je zajetih iz publicistike, manjši delež, dobrih 300.000 besed, pa s spleta. Besedilodajalci[11] so bili večinoma založbe, posamezni leposlovni avtorji, kar nekaj besedil pa je bilo prevzetih tudi iz korpusa Gigafida (Logar Berginc idr., 2012).

Vsa zajeta besedila so bila ročno pregledana; preverjeno je bilo, ali po vsebini (vsebina ni oglasna, ideološko zaznamovana, nasilna, spolna, zelo strokovna in težko razumljiva ipd.) in jezikovno (rabljen je knjižni jezik) ustrezajo specifičnim potrebam e-okolja *Slovenščina na dlani*. Pokazalo se je, da so vsebine, ki za učeče se niso primeren vir povedi za vaje iz pravopisa in slovnice, zelo pogoste: v leposlovju, zlasti nemladinskem, so bile pogoste vsebine, povezane z nasiljem, občasno pa smo morali vsebine izločati tudi zaradi slenga, vulgarizmov, narečnega ali starinskega jezika. V publicistiki so se po drugi strani pojavljale propagandne ali ideološko zaznamovane vsebine, občasno pa tudi jezikovno nezadostno pregledana besedila. Zahtevnim strokovnim vsebinam smo se skušali izogniti že pri izboru virov.

Ob pregledu smo besedilom ročno pripisali vir, leto objave, naslov, avtorja, primernost za osnovno ali srednjo šolo ter teme, vsa besedila pa smo nato strojno označili še z vidika oblikoslovja, skladnje in imenskih entitet. Za oblikoslovno in skladenjsko označevanje smo za čim večjo natančnost pripisanih oznak vzpostavili delotok, ki združuje več različnih splošno uporabljenih orodij za slovenščino, in sicer orodje Obeliks4J (Grčar idr., 2012) za segmentiranje besedil na besede in povedi, orodje ReLDI (Ljubešić in Erjavec, 2016) za lematizacijo besed ter orodje Stanford Parser V3 (Qi idr., 2018) za pripisovanje oblikoslovnih oznak po sistemu JOS (Erjavec idr., 2010) ter skladenjskih oznak po sistemu Universal Dependencies (Nivre idr., 2016). Na koncu smo z orodjem Janes NER (Fišer idr., 2018) izvedli še označevanje

11 Navedeni so na spletni strani projekta http://projekt.slo-na-dlani.si/.

imenskih entitet v besedilu. Vsa orodja so bila naučena na učnem korpusu ssj500k (Krek idr., 2019).

3.4 Definiranje vaj

Za vsako temo iz pravopisa in slovnice smo definirali vaje. Pri tem smo bili pozorni, da so bile pri vsaki temi vaje različnih tipov in zahtevnostnih stopenj.

Vaje smo definirali v več korakih, in sicer smo za vsako vajo določili identifikacijsko številko, temo, h kateri sodi, zahtevnostno stopnjo, tip naloge, navodila za programiranje priklica primerov in načina reševanja ter navodilo za uporabnika.

Pri nekaterih tipih vaj je bilo treba poleg teh elementov pripraviti še nekatere dodatne. Za poseben tip vaje, ki od učečega se zahteva, da svojo izbiro odgovora utemelji, smo pripravili predloge pravilnih in napačnih utemeljitev, pri čemer smo upoštevali različne priklicane primere, ob katerih se bodo lahko prikazale. Primer take vaje je, ko mora uporabnik odgovoriti, ali je vejica ob večbesednem vezniku pravilno uporabljena ali ne. V drugem koraku mora uporabnik pravilno dopolniti vnaprej definirano utemeljitev odgovora: *Med deli večbesednega enodelnega veznika se vejica* ***, pri čemer lahko izbira med *piše* in *ne piše*.

Za posamezne vaje je bilo treba izdelati seznam ustreznih besednih kandidatov, tj. besed, besednih zvez ali daljših kolokacijskih nizov, na podlagi katerih so bili priklicani primeri. Kot vire za iskanje ustreznih besednih kandidatov smo uporabili: korpus MAKS, korpus Gigafida, druge relevantne jeziko(slov)ne vire (SSKJ, Slovenska slovnica, jezikoslovne razprave) ali nejeziko(slov)ne vire (Wikipedija, enciklopedije ipd.). Glavna vodila so bila razumljivost, aktualnost, frekventnost iskanih kandidatov in možnost čim bolj nazorne in realistične ponazoritve problematike konkretne vaje oziroma prikaz dejanske jezikovne rabe. Načeloma so vse enote na seznamih v lematizirani obliki, razen kadar je za priklic ustreznega primera potrebna točno določena skladenjska oblika (npr. *tekem*, *dekel*, *oken* pri preverjanju zapisa neobstojnega samoglasnika v množinski rodilniški obliki).

Oblikovanje seznamov besednih kandidatov za priklic primerov je potekalo na štiri načine, in sicer:

(1) na podlagi nekorpusnih virov, npr. za iskanje primerov zaključenega nabora besed ali besednih zvez, ki so relevantne za preverjanje določene vsebine pri nalogi in so že popisane v jezikovnih priročnikih (pri nalogi, ki preverja sklanjanje zaimkov *kaj*, *malokaj*, *marsikaj*, *mnogokaj* in *nekaj*, natančneje njihovo rodilniško in tožilniško obliko, so ustrezni primeri priklicani iz nabora vseh iskanih zaimkov v rodilniški obliki, kot napaka pa so v priklicane primere vstavljeni njihovi pari s seznama v tožilniški obliki) ali drugih pisnih in spletnih virih (za nabor besednih kandidatov za priklic primerov pri preverjanju začetnice imen praznikov smo uporabili relevantne spletne strani z informacijami javnega značaja,[12] Wikipedijo ipd.);

(2) na podlagi korpusa MAKS z iskanjem preko posebnega spletnega vmesnika (konkordančnika) NoSketch Engine;[13]

(3) z združevanjem nekorpusnih virov in korpusa MAKS, tako da smo najprej pripravili model v obliki popisa ustreznih jezikovnih zakonitosti iskanih besednih kandidatov, ki je nato v drugi fazi služil za luščenje relevantnih zadetkov v korpusu MAKS, npr. za iskanje samostalniških (po podobnem principu tudi glagolskih, pridevniških) sestavljenk (in tudi nekaterih drugih besedotvornih vrst) smo oblikovali seznam potencialnih predponskih obrazil (*nad-*, *pod-*, *anti-*, *pra-*, *raz-*, *super-*, *eks-*, *ultra-*, *ne-*, *a-*, *proti-* itd.), v drugi fazi pa smo v korpusu MAKS poiskali tiste leme, ki so sestavljene iz take predpone in nekega drugega znanega samostalnika (npr. po seznamu iz Sloleksa): *pod* + *odbor*, *anti* + *oksidant* itd.;

(4) s paberkovanjem, npr. za priklic primerov, ki preverjajo bodisi poznavanje razlikovanja med zapisom in pomenom določenih besednih parov bodisi stopnjo podomačitve določene prevzete besede (glagoli *ustaviti – vstaviti*, *uročiti – vročiti*; **coca-cola – *koka kola – kokakola* itd.).

Skupaj sta bila za naloge pri vsebinskih sklopih pravopis in slovnica pripravljena 102 seznama besednih kandidatov.

12 https://www.gov.si/teme/drzavni-prazniki-in-dela-prosti-dnevi/

13 https://www.clarin.si/noske/

Vaje smo razdelili v tri zahtevnostne stopnje: osnovna, srednja in zahtevna. Na osnovni stopnji uporabnik prepoznava napake (npr. z izbiranjem, iskanjem, označevanjem (ne)pravilnega odgovora). Pri srednje zahtevnih nalogah mora uporabnik napake ne samo prepoznati, ampak tudi odpraviti (npr. s popravljanjem, premikanjem, vstavljanjem). Najzahtevnejše naloge pa od uporabnika zahtevajo tudi, da ve, zakaj je neka rešitev pravilna ali napačna.

Skupno smo definirali več kot 500 različnih vaj in za vsako smo v naslednjem koraku priklicali primere iz korpusov.

3.5 Priklic primerov za vaje

Po jezikoslovno-didaktičnem definiranju vaj, opisanem v razdelku 3.4, so bila za vsako vajo oblikovana še podrobnejša jezikovnotehnološka navodila za samodejni priklic vseh konkretnih primerov rabe obravnavanih jezikovnih pojavov v korpusu, najpogosteje v obliki zaključenih povedi. Pri nekaterih vajah so ta navodila zelo preprosta, saj se opirajo zgolj na obliko besed ali besednih zvez (npr. iskanje pojavnic z nizom števk in/ali črk, vezajem in nizom malih črk za priklic vseh povedi z zvezo podstave in končaja, kot so *70-letni, LDS-ov, a-jevski*) ali na vnaprej pripravljene sezname, omenjene v razdelku 3.4 (npr. priklic vseh povedi, v katerih se kot neprva pojavnica pojavi lema s seznama zemljepisnih lastnih imen na *-sko/-ško;-ska/-ška*, npr. *Dogodki na Koroškem so ga pretresli.*). Za veliko večino vaj pa se je bilo treba za priklic ustreznih primerov opreti tudi na višje ravni označenosti korpusa, kot so oblikoslovne oznake (npr. iskanje neprvih pojavnic z veliko začetnico in oznako za osebni ali svojilni zaimek v drugi osebi za priklic povedi s spoštljivimi ogovori, npr. *Vabimo Vas, da se nam pridružite*), skladenjske oznake (npr. za priklic povedi s posameznimi tipi stavkov) ali njihove kombinacije (npr. iskanje nedoločnih oblik pridevnikov moškega spola ednine v vlogi povedkovega določila za priklic vseh relevantnih povedi za vaje o rabi nedoločnih pridevniških oblik, npr. *Šopek je lep*). Zaradi bogate označenosti korpusa MAKS, zlasti na ravni skladnje in imenskih entitet, kakršne drugi referenčni korpusi za slovenščino še nimajo, je bilo tako mogoče samodejno priklicati tudi korpusne primere za skladenjsko kompleksnejše slovnične in pravopisne pojave.

Ker vaje znotraj posameznih tematskih sklopov pogosto vsebujejo enake jezikovne pojave (npr. glavni stavek, odvisnik, veznik za načinovno podredje,

priredna zloženka, stični pomišljaj), temelj jezikovnotehnoloških navodil za priklic primerov k vajam predstavlja seznam tovrstnih jedrnih gradnikov, ki jih mora računalnik razpoznati, da lahko sestavi programsko predstavitev problema. Znotraj učnega e-okolja smo formalno definirali več kot 300 gradnikov, navodila za priklic primerov k posamičnim vajam pa temeljijo na njihovih različnih kombinacijah.

Po opredelitvi gradnikov je bilo treba definirati način njihovega združevanja, pri čemer je bil uporabljen enostaven domensko specifični jezik z osnovnimi logičnimi pravili in parametri za zapis pravil izbire primerov. Zapis pravila tako omogoča osnovne logične operacije, kot so *in*, *ali* in *ne* za združevanje gradnikov v kompleksno pravilo, vsakemu gradniku pa je mogoče pripisati še specifične parametre, pri čemer je najpogosteje uporabljen parameter število ponovitev gradnika z operatorji *večje*, *manjše* in *je enako* (npr. 'poved z vsaj enim odvisnikom' ali 'poved z največ enim stičnim pomišljajem'). Izdelan je bil uporabniški vmesnik za zapis takšnih pravil in interpreter, ki na podlagi pravil prikliče ustrezne primere povedi iz korpusa MAKS. Za vsako posamezno vajo so bila ustvarjena pravila povezovanja gradnikov, na podlagi katerih je interpreter priklical vse ustrezne povedi. Za njihovo pregledovanje je bil izdelan poseben uporabniški vmesnik, v katerem smo kandidate ročno pregledali in z izločanjem neustreznih primerov oblikovali končno množico povedi, iz katerih se v učnem e-okolju tvorijo vaje.

Iz izbranih povedi je bilo treba v naslednjem koraku izdelati ustrezno podatkovno strukturo, ki poleg pravilno izbranega primera pripravi vsebino vaje. Struktura je odvisna od tipa vaje, pri čemer ločimo vaje, (1) kjer je uporabniku treba ponuditi poved, v kateri določen del manjka in ga mora dopolniti, (2) kjer je določen del povedi spremenjen in se mora do nje opredeliti, (3) kjer mu je potrebno ponuditi več povedi, on pa mora izbrati pravilne, in (4) kjer mu je ponujeno več delov različnih povedi, on pa jih mora ustrezno povezati. Prvi korak v takšnih spremembah je vedno določanje dela povedi, ki bo spremenjen. Za navedeno je bil izdelan domensko specifičen jezik, ki uporablja izdelane gradnike in njihove parametre. Slednji omogoča, da z njim označimo del povedi, ki bo spremenjena, in tudi zapišemo vrsto spremembe. Te so lahko preproste, kadar samo izbrišemo oziroma zakrijemo besedo ali del besede, ali pa kompleksne, kadar je treba za uporabnika pripraviti nepravilne primere, ki

pa morajo delovati verodostojno. V slednjih primerih gre največkrat za zamenjavo dela originalne besede, celotne besede ali skupine besed. Spremembe so lahko preproste, kot prestavitev mesta vejice, zamenjava velike začetnice ali menjava končnice v nedoločniku, ali kompleksne, kjer je določeno besedo treba zamenjati s pomensko sorodno besedo, v istem sklonu, spolu in številu, kot je to v primeru menjave besed *vedeti* in *znati*. Za slednje smo uporabili Sloleks 2.0 (Dobrovoljc idr., 2015), kjer smo iz osnovne oblike, ki je podana v pravilih za spremembe, na podlagi oblikoslovnih oznak besede, ki jo želimo zamenjati, dobili ustrezno izpeljanko. Izdelan je bil interpreter, ki iz pravil sprememb in potrjenih primerov pri posamezni vaji izdela zapis primerov v obliki JSON, ta pa se nato uporabi za prikaz vaj v uporabniškem vmesniku, njihovo reševanje in tudi vrednotenje. Z navedenimi orodji je v učno e-okolje mogoče dodajati tako nove primere kot tudi nove vaje.

4 POZNAVANJE FRAZEMOV IN PREGOVOROV

Eden izmed izzivov pri pripravi učnega e-okolja je bil, kako prispevati k spoznavanju frazemov in pregovorov ter s tem k izboljšanju frazeološke kompetence med mladimi. Znano je sicer, da nabor frazeoloških enot, ki jih poznamo in razumemo, narašča s starostjo (prim. Meterc, 2019), vendar pa osnovno- in srednješolski učitelji slovenščine menijo, da bi bilo koristno frazeologiji nameniti nekoliko več pozornosti, saj si, kot opažajo, učenci oz. dijaki »pogosto napačno interpretirajo frazeme« (Voršič, 2018, str. 91). Težave jim torej povzroča razumevanje enot, kar nadalje vodi v negotovost pri rabi. Poseben problem predstavlja tudi dejstvo, da v slovenskem prostoru še ni na voljo priročnika, ki bi se posebej posvečal slovarski obravnavi frazeoloških enot in bi bil prilagojen šolajoči se populaciji.[14] Naš namen je bil v e-okolju ponuditi raznovrstne vaje, ki bodo pomagale pri spoznavanju teh aktualnih in učinkovitih jezikovnih sredstev, s katerimi se začnemo srečevati že v zgodnjem otroštvu in ki pomembno sooblikujejo naše sporazumevanje na različnih področjih (Jesenšek, 2018).

14 Novost na tej ravni je slovar pregovorov in sorodnih enot, ki je od konca leta 2020 na voljo na portalu Fran (prim. Meterc, 2020).

4.1 Vsebinska področja

S sklopom, vezanim na frazeme in pregovore, smo želeli izpostaviti vsebine, ki so nekoliko manj prisotne v učnih načrtih,[15] kljub temu pa so za osnovno- in srednješolce zelo zanimive in motivirajoče. Sledili smo trem ciljem: (1) pripraviti slovarske opise izbranih frazemov in pregovorov, (2) zasnovati raznovrstne vaje ter (3) vaje podpreti z nazornimi teoretičnimi razlagami, ki bodo služile kot pomoč pri reševanju. Osredinili smo se na frazeme kot osnovne frazeološke enote in na pregovore, ki jih umeščamo k frazeologiji v širšem smislu. V prvi fazi smo pripravili izbor sto frazemov in sto pregovorov (prim. Ulčnik, 2019; Ulčnik in Meterc, 2019), pri čemer smo izhajali iz treh osnovnih kriterijev: (1) aktualnost enot, (2) didaktična relevantnost enot, (3) pokrivanje različnih tematskih skupin. Aktualnost enot smo vezali na prisotnost v učbeniških gradivih (pomeni, da obstaja velika verjetnost, da se s temi enotami srečajo pri pouku slovenščine),[16] pri čemer so bili frazemi in pregovori izpisani iz izbranih gradiv založb Mladinska knjiga in DZS (npr. iz delovnih zvezkov in beril od 6. do 9. razreda in od 1. do 4. letnika), in na zadostno pogostost v korpusu sodobnih pisnih besedil Gigafida v. 2.0 Dedup (Krek idr., 2019); enote, ki so v korpusu imele le posamične zglede rabe, niso bile upoštevane.[17] Pri frazemih je poseben problem predstavljalo dejstvo, da v slovenskem prostoru nimamo veliko raziskav o poznavanju in pogostosti rabe frazemov oz. nimamo seznamov (naj)pogostejših enot, medtem ko smo pri pregovorih lahko izhajali iz t. i. paremiološkega optimuma oz. seznama tristo najbolj poznanih in uporabljanih enot (prim. Meterc, 2017). Kot didaktično relevantne smo opredelili tiste enote, pri katerih je (praviloma zaradi višje stopnje idiomatičnosti) izkazano oteženo razumevanje. Pri tem smo upoštevali konkretne predloge učiteljev; ti so izpostavili frazeme in pregovore, za katere so opazili, da jih učenci in dijaki slabše poznajo oz. jih ne razumejo, npr. *gordijski*

15 Kržišnik (2015, str. 132) pri tem ugotavlja, da so zlasti na srednješolski ravni učni načrti vsebinsko že bolj natančni, kot so bili npr. pred tridesetimi leti, in omogočajo boljši vpogled v seznanjanje s frazeologijo.

16 Več o gradivu in poteku analize v Ulčnik, 2019; Ulčnik in Meterc, 2019.

17 Iz gradiva je bil na primer izpisan frazem *luna trka koga*, ki na koncu ni bil izbran, saj ima v korpusu manj kot deset pojavitev. Preverjanje v korpusu je sicer potekalo na podlagi t. i. fraznih jeder, da bi lahko v čim večji meri zajeli variantnost enot, npr. *bojen sekira* za iskanje frazema *zakopati bojno sekiro*; za pridobivanje variant smo uporabljali tudi iskanje v okolici (*bojna sekira – zakopati/zakopan/izkopati/izkopan*).

vozel, labodji spev, posuti se s pepelom, priti z dežja pod kap, *vleči dreto*. Tretji kriterij se je nanašal na idejo o vključevanju frazemov in pregovorov iz različnih tematskih skupin, s čimer smo želeli izpostaviti, da se frazeološke enote pojavljajo na različnih semantičnih poljih; z njimi lahko spregovorimo o človeku (npr. *stisniti zobe, videz vara*), medosebnih odnosih (npr. *pogledati skozi prste, obljuba dela dolg*), dejavnostih in bivanju (npr. *dobiti zeleno luč, vaja dela mojstra*), predmetnosti in pojavnosti (npr. *kamen spotike, lakota je najboljši kuhar*), času in prostoru (npr. *na vrat na nos, hiti počasi*), pa tudi o količini, meri in stopnji (npr. *levji delež, v tretje gre rado*). Z upoštevanjem osnovnih kriterijev smo prišli do izbora sto frazemov in sto pregovorov, ki izkazujejo zadostno aktualnost, didaktično relevantnost ter jih lahko umestimo v različne tematske skupine.

4.2 Priprava slovarskih opisov frazemov in pregovorov – FRIDA

Za izbranih dvesto enot smo v nadaljevanju pripravili celostne slovarske opise in jih zbrali v zbirki, ki smo jo poimenovali FRIDA (Frazemi in pRegovorI na DlAni). Najprej smo v programskem vmesniku izdelali večnivojsko podatkovno shemo in vanjo začeli vnašati podatke. Na prvem nivoju smo opazovali oblikovne in tipološke lastnosti izbranih enot, na drugem nivoju pomensko-pragmatične lastnosti, na tretjem nivoju pa njihove slovnične lastnosti. Opisne formulacije smo vseskozi prilagajali primarnim uporabnikom e-okolja, torej učencem in dijakom, ter pri tem sledili nazornosti, jasnosti in razumljivosti. Četrti nivo je bil namenjen navajanju dodatnih zgledov rabe, ki bi bili uporabni pri pripravi vaj. Pri tem smo bili posebej pozorni na zglede z več frazeološkimi enotami, na pojasnjevanje izvora posameznih enot (npr. na mitološka pojasnila in etimološke razlage), na opažene dobesedne pomene pri enotah, ki omogočajo t. i. dvojno branje (Kržišnik, 2006, str. 260), na prisotnost uvajalnih sredstev ipd. Vseskozi smo pazili na zadostno navajanje besedilnega okolja, iz katerega je možno razpoznati pomen predstavljenih frazeoloških enot. Uporabniku se opisi prikazujejo v prilagojeni zaslonski sliki, pri čemer sledimo njegovemu interesu in omogočamo večnivojski prikaz z možnostjo selektivnega izbiranja podatkov v razponu med osnovnim in razširjenim oz. celostnim prikazom (prim. tudi Jesenšek in Ulčnik, 2014, str. 286). Za frazeme so tako sprva prikazani le izhodiščna enota, parafraza in zgled, šele v naslednji fazi pa pomenski opis, pragmatična pojasnila, morebitne so- in

protipomenske enote ipd. Pri pregovorih izhodiščni enoti sledita pomenski opis in zgled, dodane pa so tudi opažene pragmatične značilnosti, sopomenske enote in variante. Izdelani slovarski opisi imajo dvojno funkcijo – služijo podrobnejšemu seznanjanju s konkretnimi frazemi in pregovori, preverjanju njihovega pomena in rabe, obenem pa smo zbrane podatke skušali v čim večji meri izkoristiti tudi pri zasnovi in pripravi vaj.

4.3 Priprava vaj za frazeološke vsebine

Pripravljeni slovarski opisi so bili torej z množico podatkov izhodišče za zasnovo vaj s področja frazeologije. V osnovi smo želeli ponuditi vaje različnih zahtevnostnih ravni (osnovna, srednja, zahtevna) in obenem zajeti različne tipe vaj (izbiranje, razvrščanje, dopolnjevanje …). Na podlagi spoznanj teoretične frazeologije in skladno z izsledki frazeodidaktike (Kržišnik, 2015; Kacjan in Jesenšek, 2010) smo vaje kategorizirali v tri vsebinske podsklope: (1) podoba, (2) pomen in (3) raba frazemov ter pregovorov. V prvem podsklopu se pojavljajo vaje osnovne in srednje zahtevnostne ravni. Njihov namen je doseči, da uporabniki e-okolja frazeme in pregovore prepoznajo v besedilu, da se zavedajo njihove oblikovne oz. sestavinske podobe (večbesednost, ustaljenost) in da znajo ločevati med frazemi ter pregovori. Te vaje so konkretno vezane na prepoznavanje frazemov in pregovorov v besedilu, označevanje njihovih sestavin, dopolnjevanje manjkajočih sestavin ipd. Pri posameznih vajah so dodani tudi elementi igrifikacije, npr. spomin – povezovanje frazema s sliko, povezovanje prvega in drugega dela pregovora; sestavljanka – sestavljanje frazemov oz. pregovorov iz danih sestavin. V drugem podsklopu so vaje vseh treh zahtevnostnih ravni, pri čemer preverjamo razumevanje pomena frazeoloških enot, (pre)poznavanje medfrazemskih razmerij (so- in protipomenskost), zmožnost pomenskega pojasnjevanja enot ter sposobnost njihovega nadomeščanja s slogovno nezaznamovanimi besedami. Uporabnik mora npr. ustrezno povezati enoto s pomenom, prepoznati zgled brez frazema (v njem je besedna zveza uporabljena v dobesednem pomenu, npr. *pustiti na cedilu*: *Jajčevca narežemo na debelejša kolesca, nasolimo in pustimo na cedilu 15 minut*), povezati so- in protipomenske enote, na podlagi pomena enote razvrstiti v ustrezne tematske skupine, namesto frazemov uporabiti nevtralne besede, razložiti pomen frazemov in pregovorov, uporabljenih v zgledu. V tretjem podsklopu so vaje srednje in zahtevne ravni. Osredinjamo

se na ustrezno frazeološko rabo ter sposobnost nadomeščanja nezaznamovanega izražanja s frazeološkim. Pri pregovorih smo posebej pozorni tudi na uveljavljeno rabo s t. i. uvajalnimi sredstvi. Uporabnik mora npr. izbrati ustrezni frazem in z njim dopolniti zgled, dokončati poved z ustreznim frazemom, za označeni del povedi uporabiti ustrezni frazem, izbrati zgled, ki vsebuje uvajalno sredstvo (npr. *znana modrost*, *ljudski pregovor, kot pravijo*). Vajam smo dodali še nekaj ustvarjalnih nalog oz. izzivov. Pri teh se uporabnik npr. preizkusi kot slovaropisec (s seznama izbere enoto in jo po navodilih in skladno z danim zgledom slovarsko opiše), izpolni atlas pregovorov, tvori kratko besedilo (npr. horoskop, šalo) in v njem uporabi frazeme in/ali pregovore, izraža svoje mnenje o resničnosti in aktualnosti izbranega pregovora (npr. *obleka naredi človeka*).

5 KOMPETENCE RAZUMEVANJA IN TVORJENJA BESEDIL

Jezik kot pomen, izražen z zvokom, ni le sporazumevalno sredstvo, ampak je neposredno povezan z razmišljanjem. Kot tak je v različnem obsegu sestavina skoraj vsake človekove aktivnosti v skupnosti ali različnih skupinah. Zato naj bi posameznik za delovanje in sodelovanje v skupnosti ali tudi za samorealizacijo razvijal sporazumevalno jezikovno zmožnost tvorjenja in razumevanja večpredstavnostnih klasičnih ali elektronskih pisanih in govorjenih besedil. To je tudi temeljni cilj področja o besedilih znotraj projekta *Slovenščina na dlani*. S tem ciljem so povezani cilji razvijanja kritičnega branja, pridobivanja znanja o jezikovni rabi in načrtno ter usmerjeno učenje o prvinah določenega besedila kot predstavnika določene besedilne skupine. Navedeni cilji so nujni, da lahko govorimo o razvijanju sporazumevalne jezikovne zmožnosti, saj le primerno visoka žanrska pismenost »uporabnikom jezika zagotavlja prepoznavanje in učinkovito rabo žanrov« (Nidorfer Šiškovič, 2013, str. 273). Razumevanje besedila pa ni le proces prepoznavanja prvin nejezikovnega in jezikovnega konteksta s posameznikovega stališča, zato je pri nalogah, zlasti vezanih na vsebino in smisel besedilnega sporočila, bila potrebna precejšnja mera natančnega branja ter posledično oblikovanja enoumno zastavljenih in smiselnih nalog. V ta namen smo za področje besedil oblikovali zbirko BERTA in vmesnik Berta.

5.1 Priprava gradiva za vaje in naloge s področja besedil[18]

Za pripravo gradiva, tj. oblikovanje zbirke BERTA, smo najprej pregledali učne načrte in učbeniško gradivo od šestega razreda osnovne šole dalje, da bi lahko oblikovali nabor besedilnih skupin, učnih ciljev in uporabljenih strokovnih terminov, ki naj bi jih učeči se poznali oz. dosegli na določeni stopnji.

Pri oblikovanju nabora besedilnih skupin in tudi pri drugih teoretičnih odločitvah se je bilo treba opredeliti do različnih terminov za poimenovanje in pojmovanje, npr. množice besedil. V pregledanem osnovno- in srednješolskem gradivu se za poimenovanje množice besedil uporablja termin besedilna vrsta. Besedilne vrste, kot jih razumemo tukaj, se z večrazsežnostnega vidika besedila oblikujejo glede na značilne skupne kontekstualne in strukturne prvine, ki so medsebojno povezane in soodvisne. Besedilne vrste oblikujejo okvir za prototipične prvine konkretnega besedila. Te prvine so osnovane na dogovorih jezikovnih uporabnikov o jezikovnih/govornih vzorcih. Hkrati besedilne vrste izkazujejo prototipične funkcijske, medijsko-položajne in tematske prvine ter skladno s temi prvinami značilno strukturo besedila (Gansel in Jürgens, 2007). Besedilne vrste so produkti konvencionalnih jezikovnih dejanj znotraj posameznega komunikacijskega področja (Heinemann, 2000). Tovrstna opredelitev besedilne vrste je pokazala, da v učnih procesih učeči se pravzaprav ne spoznavajo prvin besedilnih vrst ali le redko, temveč v veliki večini primerov spoznavajo značilnosti besedilnih tipov, npr. prošnja, zahvala, voščilo. Besedilni tip je namreč skupina besedil s skupnimi ali podobnimi jezikovnimi značilnostmi in je vir za oblikovanje besedilnih vrst. Da bi se izognili nejasnostim in zmedi, smo vpeljati termin, ki vključuje tako besedilne tipe kot vrste, to je besedilna skupina.

Pri izbiranju besedilnih skupin smo dajali prednost tistim, ki nastajajo v za večino govorcev slovenščine pomembnih ali priljubljenih, v vsakem primeru pa živih in aktualnih komunikacijskih položajih, zato smo se odločili za nabor 29 besedilnih skupin, med drugimi za pritožbo, novico, vremensko napoved,

18 Dogovor znotraj projektne skupine je, da termin *vaja* razumemo kot vadenje, urjenje znane učne snovi, torej ponavljanje in utrjevanje že znanega, medtem ko termin *naloga* razumemo kot aktivnost, ki zajema reševanje novih jezikovnih problemov in vključuje ustvarjalnost, več miselnega angažmaja učečih se, zato je razumljivo, da se na področju besedil uporabljata oba termina.

telefonski pogovor, govorni nastop, prijavnico, življenjepis, opis postopka, poljudnoznanstveni prispevek. Poleg prototipskih predstavnikov besedilne skupine smo želeli zbrati tudi nekaj neprototipskih, inovativnih, saj je tudi za vsakodnevne komunikacijske položaje značilna ustvarjalnost.

Sledilo je izbiranje tematskih sklopov in znotraj njih tematik, ki naj bi jih vsebovale izbrane besedilne skupine. Pri izbiranju tematskih sklopov in tematik smo izhajali iz sporazumevalnih tem za poučevanje slovenščine ali katerega koli drugega jezika kot drugega jezika (SEJO, 2011), npr. narava, zdravje, odnosi, poklic; SEJO je namreč »dokument, ki je pomemben za jezikovno poučevanje *nasploh*, ne samo poučevanje tujih jezikov« (SEJO, 2011, str. 5).

Na osnovi izbranih tematskih sklopov, besedilnih skupin in tematik smo oblikovali scenarij njihovega povezovanja in primernosti obravnave po razredih oz. letnikih. Del tega scenarija je slogan kot orientacijska točka, komu so besedila določene besedilne skupine, tematskega sklopa in tematike namenjena. Preglednica 1 prikazuje slogan *Lačen kot Lakotnik*.

Preglednica 1: *Izsek preglednice P-BERTA*

Tematski sklop	**Slogan**	**Tematika**	**Besedilna skupina**	**Naslov besedila**
Zdravje	*Lačen kot Lakotnik*	čokolada	novica	Temna čokolada je lahko tudi zdrava
			kuharska oddaja	Prigrizek v somraku
		češnje	novica	Japonska v pričakovanju čarobnega cvetenja češenj
			kuharski recept	Češnjev sladoled
		agrumi	novica	Veste, zakaj se pomaranče vedno prodajajo v rdečih mrežastih vrečkah?
			kuharski recept	Solata z agrumi
		paradižnik	novica	Z razkritim genskim zapisom do okusnejšega paradižnika
			kuharski recept	Posušeni paradižniki
		sivka	novica	Sivka iz Brij se je preselila na slikarska platna
			kuharski recept	Sivkina limonada

Sledila sta zbiranje konkretnih posameznih besedil in pridobivanje soglasij za njihovo uporabo. Tako je nastala zbirka BERTA (BEsedil pRakTičnega sporazumevanjA), ki ima dva podkorpusa, in sicer govornega (G-BERTA, 59 besedil) in pisnega (P-BERTA, 216 besedil), skupno 275 besedil. Zajema besedila, ki so vsebinsko in po načinu obravnave vsebine blizu šolajoči se mladini od 11. do 19. leta starosti. Za lažjo orientacijo uporabnikov (učiteljev in učečih se)

Preglednica 2: *Izsek preglednice Metakazalo vaj*

Besedilo	*Šolske besedilne skupine 1*	Opis življenja osebe
		Pripoved o življenju osebe
		Življenjepis
		Prijava
		Prošnja
		Pritožba
	Šolske besedilne skupine 2	Opis postopka
		Esej
		Prijavnica
		Vabilo
		Govorni nastop
		Poljudnoznanstveni prispevek
	Besedilne skupine od tu in tam	Telefonski pogovor
		Razgovor
		Anekdota
		Kuharska oddaja
		Filmski napovednik
		Ocena
		Prepričevalni pogovor
	Besedilne skupine v medijih	Intervju
		Mali oglas
		Oglas
		Vremenska napoved
		Horoskop
		Novica
		Šala

smo scenarij s slogani nadomestili z razdelitvijo vseh besedilnih skupin na štiri množice (Preglednica 2).

Zbirka BERTA je bila dokončno oblikovana, ko smo običajnemu korpusnemu označevanju dodali še oznake, pomembne za priklic besedila kot primernega za določeno stopnjo in doseganje želenega učno-vzgojnega cilja. Zajeta besedila so zato opremljena s podatki o tvorcu, naslovniku, mestu in času objave/nastanka besedila, javni dostopnosti besedila, jezikovni zvrsti, številu udeležencev, sporočevalnem namenu, funkciji besedila, slogovnem postopku, tematskem sklopu, tematiki, temi, prenosniku in zahtevnosti besedila (osnovna, srednja, zahtevna). Uporabnikom bodo prikazana večpredstavnostno in v celoti (npr. pdf-posnetek besedila, video/avdio posnetek), tako da bodo vidni tudi neverbalni elementi. Za doseganje nekaterih učnih ciljev, npr. za določitev ubeseditvenega načina, smo daljša besedila, npr. poljudnoznanstvene prispevke, členili na odlomke.

5.2 Priprava vaj in nalog za področje besedil – vmesnik BERTA

Besedila ni mogoče obravnavati na enak način kot slovnična ali pravopisna pravila, saj so besedila aktualizacija in uresničitev teh pravil. Zgodi se, da ni mogoče le eno in zato pravilno interpretiranje besedila. Zdi se, da je težnja po le eni pravilni interpretaciji prisotna v izobraževalnem procesu, ko sicer zanimivemu besedilu sledijo tri ali štiri naloge, ki od učečega se zahtevajo poznavanje kraja in časa nastanka besedila, udeležencev, njihovega družbenega razmerja. Tem nalogam sledijo še vprašanja o temi in sporočevalnem namenu ter določitev ubeseditvenega načina, ki temelji bolj na občutku kot na dejanskih značilnostih besedila. Ostale naloge se nanašajo na slovnično in/ali leksikalno raven. Temu načinu dela smo se želeli izogniti. Oblikovali smo vmesnik BERTA, v katerega k posameznemu besedilu vnašamo trditve, vprašanja in druge podatke, potrebne za programiranje in dokončno oblikovanje vaj in nalog, ki se nanašajo izključno na besedilne značilnosti posameznega besedila, njegove morebitne druge jezikovne značilnosti pa so izpostavljene le, če so povezane z značilnostmi besedilne skupine.

Tako v vmesnik BERTA k posameznemu besedilu v štiri sklope vnašamo ločene trditve, vprašanja in odgovore ter napačne alternative. Prvi sklop nalog je vezan na nejezikovne prvine konteksta, ki so bile skupaj s tematskim sklopom,

besedilno skupino, temo in jezikovno zvrstjo že popisane pri vnašanju besedila v zbirko BERTA. Drugi sklop vaj in nalog je vezan zlasti na jezikovne prvine besedila kot predstavnika določene besedilne skupine. Zanemarjen ni niti večpredstavnostni vidik, saj ta lahko bistveno prispeva k interpretaciji besedilnega sporočila, na kar opozarjata npr. S. Starc (2011, str. 434) in B. Vičar (2015, str. 802). Z nalogami je opozorjeno na morebitna odstopanja od pričakovanega, npr. vabilo na klekljarski krožek, ki je v obliki opisa postopka. Uporabnik bo prek nalog in vaj opazoval ter spoznaval značilno leksiko in oblikoslovne značilnosti posamezne besedilne skupine; za novico oz. vest je npr. značilna sintagma *po poročanju X*, za horoskop pa strokovni termini s področja astronomije in astrologije, npr. *Sonce v ozvezdju Raka* (Krajnc Ivič, 2019). Za opis postopka je značilna raba sedanjiških prvoosebnih množinskih glagolskih oblik, pojavi pa se celo tretjeosebna edninska. Uporabnik bo tako spoznaval, da je uresničitev jezikovnih prvin posameznega besedila odvisna od komunikacijskega področja, znotraj katerega je besedilo nastalo, od teme in namena besedila ter dalje od ubeseditvenega načina oz. slogovnega postopka. V tretji sklop sodijo vaje in naloge, vezane na vsebino posameznega besedila, ki lahko dodatno utemeljijo določitev sporočevalnega namena ali slogovnega postopka konkretnega besedila. Četrti sklop predstavljajo naloge tvorjenja besedil, pri katerih želimo, da uporabnik povzame, obnovi izhodiščno besedilo ali tvori besedilo iste ali druge besedilne skupine. To pomeni, da praktično prikaže razumevanje pridobljenega znanja. Tovrstne naloge od uporabnika zahtevajo še dvoje, in sicer kritičen razmislek o prebranem/slišanem/ogledanem in sodelovanje z vrstniki. Preglednica 3 prikazuje primere trditev, alternativ in vprašanj za vsak sklop nalog in vaj.

Preglednica 3: *Primeri trditev in vprašanj v vmesniku BERTA k besedilu o terapiji s konji*

TIP	TRDITEV/VPRAŠANJE IN S KREPKIM TISKOM OZNAČENA PRAVILNA REŠITEV/ODGOVOR	ALTERNATIVE[19]
Prvi sklop vaj in nalog		
PRAVILNE TRDITVE		
Tvorec	Tvorec izhodiščnega besedila je **Kristijan Skok**.	• Jen Mundy • Lis Hartel • M. Demšar
Drugi sklop vaj in nalog		
PRAVILNE TRDITVE		
Sporočevalni namen	Namen izhodiščnega besedila je **naslovnike informirati** o konjih in razlogih za izbiro konj v terapevtske namene.	• bralce prepričati • izraziti brezbrižnost • najti ugotovitve
VPRAŠANJA		
Leksika	Katero izrazje je značilno za izhodiščno besedilo kot poljudnoznanstveni prispevek? **Poljudnoznanstveno besedje, splošno znani strokovni termini in iz latinščine prevzete besede**.	• Znanstveno besedje, splošno znani strokovni termini in iz angleščine prevzete besede. • Poljudnostrokovno izrazje, znanstveni strokovni termini in iz latinščine prevzeto besedje. • Vsakdanje besedje, splošno znani strokovni termini in iz angleščine prevzete besede.
Tretji sklop vaj in nalog		
PRAVILNE TRDITVE		
Vsebina	Starost konja je mogoče določiti po **njegovih zobeh**.	• njegovi čeljusti • njegovi grivi • njegovih očeh
NAPAČNE TRDITVE		
Vsebina	Konji se hranijo predvsem s koncentrirano hrano, kot sta npr. trava in seno.	
VPRAŠANJA		
	Kako konji glede na izhodiščno besedilo izražajo svoje razpoloženje? **Z obračanjem uhljev, mahanjem z repom in s šobljenjem**.	• z obračanjem glave, mahanjem z repom in s poskakovanjem • z obračanjem uhljev, mahanjem z repom in s poskakovanjem • z obračanjem glave, mahanjem z repom in z rezgetanjem
KLJUČNE BESEDE		
terapija | konj | proces | jahanje | učinki | človek | odnos		
Četrti sklop vaj in nalog		
BESEDILNE NALOGE		
TB+T	Opiši, kako poteka terapija s konjem.	

19 Navedene so napačne možnosti.

Vsako pregledano besedilo ima povprečno od 15 do 25 trditev in okoli 4 vprašanja za vse štiri sklope. Število trditev in vprašanj je odvisno od dolžine besedila.

6 PROGRAMIRANJE IN VREDNOTENJE VAJ TER IZDELAVA RAZLAG

V fazi programiranja e-okolja smo morali razviti postopek, ki bo uporabnikom omogočal reševanje vseh zamišljenih vaj. Ta postopek sestoji iz več korakov. Najprej se morajo generirati primeri, ki so na voljo za reševanje. Osrednji del celotnega postopka predstavlja prikaz vaje in njeno reševanje. Pomembno vlogo v e-okolju ima tudi algoritem, ki skrbi za določanje zaporedja reševanja vaj. Na koncu je treba še ovrednotiti uporabnikove poskuse reševanja. Sočasno s programiranjem smo izdelali še bazo razlag, ki je uporabnikom v pomoč pri reševanju.

6.1 Generiranje primerov vaj

Pravila za generiranje primerov smo določili že v fazi definiranja vaj v posameznih vsebinskih področjih. Glavni izziv je predstavljalo generiranje primerov za vaje, ki uporabniku ponujajo napačne odgovore. To so na primer naloge tipa 'Izberi pravilen odgovor' ali 'Popravi odgovor'. Pomembno je namreč, da napačni primeri niso preveč nesmiselni, kar bi uporabniku olajšalo reševanje. S tem bi se zmanjšal miselni napor uporabnika pri reševanju teh nalog, kar bi privedlo do upočasnjenega napredka v njegovem znanju.

Na področjih pravopisa in slovnice smo s pomočjo gradnikov in vnaprej določenih pravil uspeli vzpostaviti popolnoma avtomatiziran proces, ki iz korpusa besedil tvori primere za posamezno vajo. S pravili smo uspeli avtomatsko tvoriti tudi pravopisno oz. slovnično napačne povedi za vaje, kjer je bilo to potrebno. Proces je podrobneje opisan v poglavju 3.5.

Na področju frazemov in pregovorov smo primere generirali na podlagi informacij v slovarskih opisih. Pri tem smo spoznali, da računalnik ni zmožen tvoriti ustreznih alternativnih primerov za vaje, kjer so bili ti potrebni.[20] Na

20 Npr. pri vaji, kjer je izločena ena sestavina frazema in mora uporabnik med štirimi ponujenimi odgovori izbrati ustreznega (primer: *narediti iz muhe …* – ob ustrezni sestavini *slona* z avtomatskim iskanjem težko dobimo relevantne alternativne odgovore, ki zagotavljajo tudi ustrezno zahtevnost naloge).

tem mestu je namreč treba tvoriti smiselne alternative, kar je prezahtevno za računalniške algoritme. Da smo rešili težavo, smo zapise v slovarskih opisih dopolnili še z dodatnimi informacijami, ki jih računalnik uporabi pri generiranju takih primerov vaj.

Na področju besedil se primeri nalog pogosto nanašajo na pomen obravnavanega besedila. Teh primerov nismo mogli tvoriti povsem avtomatsko, zato smo pri vsakem besedilu zapisali nekaj trditev, vprašanj in pravilnih odgovorov, na podlagi katerih lahko računalnik tvori naloge tipa 'Odgovori na vprašanje', 'Izberi pravilen odgovor' ali celo uporabi več vprašanj in tvori nalogo 'Poveži vprašanja s pripadajočimi odgovori'. S takim polavtomatskim pristopom smo prihranili čas, ki bi ga sicer potrebovali za ročno tvorjenje nalog, in hkrati povečali nabor vaj, ki se nanašajo na pomen zapisanega ali govorjenega besedila. Pri določenih trditvah in vprašanjih smo zapisali tudi alternativne napačne odgovore, ki se uporabljajo za tvorbo nalog, pri katerih mora uporabnik prepoznati oz. popraviti napačen odgovor.

6.2 Prikaz in reševanje vaj

Osrednji del programiranja vaj predstavljata prikaz vaje in njeno reševanje. V tej fazi smo poskrbeli, da e-okolje ponuja interaktiven način reševanja, ki uporabnika spodbuja k uporabi. Načrtovani uporabniški vmesniki za reševanje so prilagojeni različnim napravam in potrebam uporabnikov.

Programska predstavitev vaje je uporabniku prikazana v obliki spletnega uporabniškega vmesnika, katerega videz močno variira glede na tip naloge. Za ustrezen prikaz posameznega primera se uporablja podatkovni zapis primera v obliki JSON, v katerem je primer razdelan tako, da določa, kateri so tisti elementi, kjer bo mogoča interakcija. Iz slednjega se nato generira uporabniški vmesnik, kjer so v besedilo dodani vnosni elementi, določene elemente, kot so vejice ali besede, pa je mogoče izbrati ali jih premikati po besedilu. Prav tako ima uporabnik pri določenih tipih nalog na izbiro različne možnosti, sam pa se mora odločati o njihovi pravilnosti. Pri posebnih uporabniških vmesnikih, kot so za rešeta ali sestavljanke, so v zapisih primera vključeni seznami alternativnih besed ali delov povedi, tako da omogočajo generiranje vmesnika. V vseh vmesnikih je treba ustrezno hraniti uporabnikovo interakcijo, saj se poleg vnosa, kot je besedilo ali izbira, beležita tudi uporabnikovo napačno

ravnanje ter čas, ki ga potrebuje za posamezno akcijo. Akcije se sproti beležijo in posredujejo v zaledje učnega sistema, kjer so ovrednotene, saj se na njihovi podlagi po logični shemi prožijo različne druge akcije. Logična shema se prav tako obnaša dinamično, saj so podatki zapisani v podatkih primera in so odvisni tudi od vrste opravila, ki ga uporabnik opravlja, prav tako pa tudi od uporabniškega konteksta. Vse navedeno vpliva na odziv uporabniškega vmesnika, ki lahko nato ponudi pomoč pri neuspešnem reševanju, ovrednoti posamezno reševanje in na koncu vpliva tudi na izbor naslednjega primera za reševanje. Omenjeni proces je zasnovan tako, da ga lahko uporabimo za prikazovanje in reševanje vaj na vseh področjih v e-okolju. Manjša posebnost so le vaje s področja besedil, kjer je uporabniku ob reševanju vedno na voljo še vpogled v izhodiščno besedilo.

Zaporedje, po katerem se prikazujejo primeri za reševanje, je odvisno od mnogih dejavnikov. Vzpostavili smo sistem, ki na podlagi statističnih modelov in vnaprej določenih pravil adaptivno izbira vaje in primere, ki jih rešuje uporabnik. Ločimo več načinov delovanja, ki so odvisni od uporabniškega vnosa. Vaje in naloge lahko uporabnik rešuje samostojno; v tem primeru si sam izbere nabor poglavij, ki jih želi vaditi. Druga možnost je, da mu nabor za reševanje dodeli učitelj; pri tem učitelj izbere poglavja in tudi posamezne vaje ter naloge, ki jih bo uporabnik reševal. Tretji način pa predstavljajo predlogi e-okolja, ki na podlagi zgodovine reševanja uporabniku predlaga primerne vaje in naloge za nadaljnje utrjevanje znanja; znotraj izbranega načina delovanja zaporedje prikaza vaj, nalog in primerov določa e-okolje. Pri tem upoštevamo težavnostno stopnjo, uporabnikovo starost oz. razred, ki ga obiskuje, in tudi njegovo uspešnost pri preteklih poskusih reševanja. Določena težavnostna stopnja vaje, naloge oz. primera se med uporabo e-okolja prilagaja tudi glede na uspešnost, ki so jo imeli uporabniki pri reševanju. Na ta način lahko avtomatično odpravimo morebitne človeške napake pri določanju težavnostne stopnje.

6.3 Vrednotenje

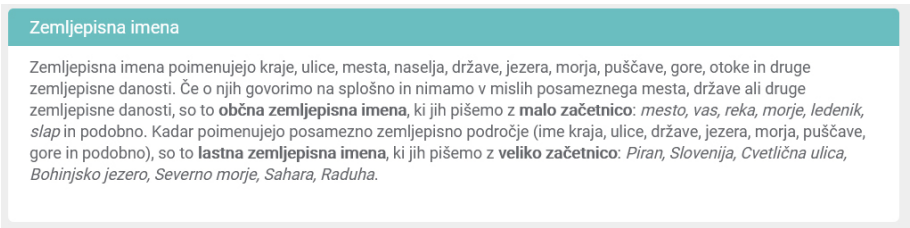
Ko uporabnik rešuje vaje, se izvaja vrednotenje. Najprej se ovrednoti pravilnost rešitve posameznega primera. Nadalje se vrednoti uporabnikova uspešnost pri reševanju posamezne zadolžitve, npr. domače naloge. Na koncu pa e-okolje vrednoti še uporabnikovo skupno znanje znotraj določenega poglavja.

Vrednotenje uporabnikovih rešitev poleg končnega odgovora upošteva še vmesne korake, porabljen čas in napačne poskuse. Vrednotenje lahko poteka na tri načine. Prvi način je samodejno vrednotenje, kjer vnaprej poznamo pravilno rešitev in jo lahko v času reševanja ovrednotimo. Tak način uporabniku ob reševanju daje neposredno povratno informacijo. Drugi način je samovrednotenje, ki izhaja iz principa navajanja možnih pravilnih odgovorov, ki se po reševanju prikažejo uporabniku, nato pa ta na podlagi ponujenih možnosti ovrednoti svoj odgovor in s tem presodi lastno uspešnost (Holcar Brunauer idr., 2019, str. 3). Tega smo uporabili pri nalogah, ki imajo različne možne rešitve (npr. razlaga pomena frazema ali pregovora) oz. so vezane na uporabnikovo ustvarjalnost (raba frazemov in pregovorov). Gre za naloge, ki izhajajo iz problemskega pristopa in pri katerih je vključeno doseganje višjih taksonomskih ravni (analiza, sinteza, vrednotenje). Pri tretjem načinu rešitev ovrednoti drug uporabnik, običajno v vlogi učitelja ali tutorja. Po tem načinu lahko poseže uporabnik, ki ni prepričan v svoje sposobnosti, da bi izvedel samovrednotenje, lahko pa ga določi tudi učitelj ob dodelitvi nalog, ki jih ni mogoče samodejno vrednotiti. Inovativne možnosti vrednotenja od uporabnikov zahtevajo kritično presojo in spodbujajo prevzemanje odgovornosti za lastno učenje oz. ustvarjajo možnosti za nudenje medvrstniške povratne informacije, s tem pa si učeči se pridobivajo pomembne učne izkušnje.

Vrednotenje uporabnikove uspešnosti pri reševanju posamezne zadolžitve temelji na seštevanju ovrednotenih posameznih primerov, ki jih je uporabnik rešil v sklopu zadolžitve. Pri vrednotenju uporabnikovega skupnega znanja znotraj določenega poglavja v e-okolju se upoštevajo vsi primeri iz določene tematike, ki jih je uporabnik rešil. Rezultat vrednotenja se uporabniku prikaže v obliki različno obarvanih medalj in dosežkov, ki ga dodatno motivirajo k reševanju. Rezultat se upošteva pri določitvi težavnostne stopnje primerov vaj in nalog, ki se uporabniku prikazujejo med reševanjem. Vrednotenje ni izvedeno enostavno linearno, ampak je prožno in upošteva različno stopnjo predznanja uporabnikov. Cilj je namreč, da lahko uporabnik z več znanja pri določenem poglavju hitreje napreduje in ga zaključi z manj vajami oz. začne hitreje dobivati zahtevnejše vaje iz tega poglavja.

6.4 Razlage k vsebinam vaj

Že Higgins in Johns (1984) izpostavljata, da učeči se želijo razlage k vajam, ki jim povedo, zakaj je nekaj pravilno ali napačno, in to ne samo takrat, ko ne poznajo prave rešitve, ampak tudi v primerih, ko vajo rešijo pravilno. Vsebinska področja vaj v učnem e-okolju *Slovenščina na dlani* so zato opremljena z razlagami o jezikovnih pravilih in vzorcih, na podlagi katerih so utemeljene pravilne rešitve. Učeči se lahko s preprostim klikom na ikono z vprašajem, ki je ves čas prisotna v zgornjem desnem kotu e-okolja, izve točno tisto razlago ob vaji, ki jo v nekem trenutku rešuje. Razlaga je najprej na voljo v krajši obliki in se prikaže v desnem delu okna ob primeru, ki ga uporabnik rešuje, dodana pa je povezava na daljšo razlago s primeri in ponazoritvami. Od slednje vodijo povezave tudi k sorodnim temam ali k izbranim relevantnim razlagam drugod na spletu. Slika 1 prikazuje primer kratke razlage za preverjanje začetnice pri zapisu zemljepisnih imen.



Slika 1: Kratka razlaga za preverjanje zapisa zemljepisnih imen.

Tehnično so razlage izvedene v obliki baze s približno 60 razlagami za pravopisne teme, približno 60 razlagami za slovnične teme, 39 razlagami za frazeme in pregovore ter 66 razlagami za besedila. Vsaka razlaga ima identifikatorje, ki določajo, pri katerih vajah naj se prikazuje. Razlage so sistematično urejene in dostopne tudi pod posebnim zavihkom Znanje, do katerega ima uporabnik dostop z osnovne strani.

7 SKLEP

V projektu *Slovenščina na dlani* s pomočjo jezikovnih virov in avtomatiziranih postopkov izdelujemo interaktivno učno e-okolje za podporo učenju slovenščine v osnovnih šolah od 6. razreda naprej in v srednjih šolah. Za potrebe tega e-okolja smo razvili manjši korpus besedil, primernih za mladino

(MAKS), večmodalno zbirko besedil praktičnega sporazumevanja (BERTA), slovarske opise frazemov in pregovorov (FRIDA) in bazo razlag (Znanje). Izdelani viri bodo na voljo tudi kot jezikovni vir pod licencami CC BY, razen v primerih, kjer to ni mogoče zaradi omejitev pri izvornih avtorskih pravicah besedil oz. vsebin. Dostopni so oz. bodo prek repozitorija CLARIN.SI oz. prek Clarinovih konkordančnikov (https://www.clarin.si/noske/).

Skupno je pripravljenih več kot 1000 različnih vaj in nalog s področja pravopisa, slovnice, frazeologije in besedil ter razlage za skupno več kot 200 različnih z vajami povezanih tematik. Za vsako vajo je na voljo večje število različnih primerov, pri pravopisu in slovnici tudi do 500 za eno vajo. S tem odgovarjamo na izzive, kako odpraviti napake knjižne norme, ki vztrajajo v pisnih izdelkih učencev, kako izboljšati njihovo frazeološko kompetenco in sporazumevalno jezikovno zmožnost. Ključna prednost novega e-okolja pred ostalimi e-priročniki za pouk slovenščine z vidika učitelja (in učenca) je, da se e-okolje *Slovenščina na dlani* samodejno prilagaja potrebam učečega se in tako olajša delo učitelja pri formativnem spremljanju napredka posameznikov.

V šolskem letu 2020/21 v projektu *Slovenščina na dlani* nadaljujemo z zadnjo fazo izgradnje vaj, tj. s priklicem vaj iz zalednih baz v spletni vmesnik e-okolja, dodajanjem besedilnih nalog in pisanjem razlag, izdelavo različnih funkcionalnosti, kot so prikazovanje doseženih rezultatov, prikazovanje razlag ob vajah ipd., ter začenjamo testiranje e-okolja in analizo vedenja uporabnikov. S šolskim letom 2021/22 bo e-okolje dostopno zainteresirani javnosti na spletni povezavi https://slo-na-dlani.si/.

Avtomatizacija s podporo jezikovnih tehnologij in digitalno okolje imata nekaj prednosti, ki jih papirni medij ne omogoča: veliko količino primerov in vaj, prilagajanje zahtevnosti vaj znanju uporabnika, avtomatsko vrednotenje in usmerjanje med vajami, enostavno prikliclјivo pomoč v obliki razlag, prilagojenih vsaki posamezni nalogi, podporo pri reševanju z namigi ali sprotno komunikacijo z drugimi uporabniki sistema ali sodelovanjem v skupinah. Ob raziskovanju možnosti za motivacijo in spodbujanje ustvarjalnosti pa smo ugotavljali tudi omejitve v primerjavi z osebnim stikom pri učenju. Ob tem da je učno e-okolje visoko avtomatizirano in v določeni meri individualizirano, namreč omogoča veliko manj ustvarjalnih in interaktivnih nalog. Avtorji

e-okolja *Slovenščina na dlani* zato že od zasnove naprej sledimo načelu, da je digitalno učno e-okolje koristno dopolnilo in popestritev pouka, nikakor pa ne nadomestilo za tradicionalne oblike poučevanja.

LITERATURA

Davies, G. (2016). *CALL (Computer assisted language learning)*. Centre for Languages, Linguistics & Area Studies. Pridobljeno s https://www.llas.ac.uk/resources/gpg/61#ref6

Dobrovoljc, K., Krek, S., & Erjavec, T. (2015). Leksikon besednih oblik Sloleks in smernice njegovega razvoja. V V. Gorjanc, P. Gantar, I. Kosem in S. Krek (ur.), *Slovar sodobne slovenščine: problemi in rešitve* (str. 80–105). Ljubljana: Znanstvena založba Filozofske fakultete.

Gansel, C., & Jürgens, F. (2007). *Textlinguistik und Textgrammatik. Eine Einführung*. 2. Auflage. Göttingen: Vandenhoeck & Ruprecht.

Gomboc, M. (2019). *Slovenščina. Po korakih do odličnega znanja*. Ljubljana: Mladinska knjiga.

Grčar, M., Krek, S., & Dobrovoljc, K. (2012). Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. V *Zbornik Osme konference Jezikovne tehnologij*e (str. 82–87). Pridobljeno s http://nl.ijs.si/isjt12/JezikovneTehnologije2012.pdf

Heinemann, W. (2000). Textsorte – Textmuster – Texttyp. V K. Brinker, G. Antos, W. Heinemann in S. F. Sager (ur.), *Text- und Gesprächslinguistik: ein internationales Handbuch zeitgenössischer Forschung*. Handbücher zur Sprach- und Kommunikationswissenschaft, (zv. 16) (str. 507–523). Berlin, New York: Walter de Gruyter.

Higgins, J., & Johns, T. (1984). Computers in Language Learning. London: Collins.

Holcar Brunauer, A., Bizjak, C., Cotič Pajntar, J. idr. (2019). *Formativno spremljanje. Samovrednotenje, vrstniško vrednotenje*. Ljubljana: Zavod Republike Slovenije za šolstvo.

Erjavec idr. (2010). The JOS Linguistically Tagged Corpus of Slovene. *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)* (str. 1806–1809). Pridobljeno s http://www.lrec-conf.org/proceedings/lrec2010/index.html

Jesenšek, V., & Ulčnik, N. (2014). Spletni frazeološko-paremiološki portal: redakcijska vprašanja ob slovenskem jezikovnem gradivu. V V. Jesenšek in S. Babič (ur.), *Več glav več ve: Frazeologija in paremiologija v slovarju in vsakdanji rabi* (str. 276–292). Maribor: Oddelek za germanistiko, Filozofska Fakulteta Univerze v Mariboru, ZRC SAZU Ljubljana, Inštitut za slovensko narodopisje.

Jesenšek, V. (2018). Zakaj in čemu frazeologija pri pouku materinščine. V N. Ulčnik (ur.), *Slovenščina na dlani 1* (str. 21–24). Maribor: Univerzitetna založba Univerze. Pridobljeno s http://press.um.si/index.php/ump/catalog/book/341

Kacjan, B., & Jesenšek, V. (2010). Pregovori pri učenju in poučevanju (tujega) jezika. V N. Holc (ur.), *Posodobitve pouka v gimnazijski praksi* (str. 59–67). Ljubljana: Zavod RS za šolstvo.

Kosem, I., Stritar, M., Može, S., Zwitter Vitez, A., Arhar Holdt, Š., & Rozman, T. (2012). *Analiza jezikovnih težav učencev: korpusni pristop*. Ljubljana: Trojina, zavod za uporabno slovenistiko.

Krajnc Ivič, M. (2019). Frazeološke enote v horoskopih in malih oglasih – besedilnovrstni vidik. V Ž. Macan (ur.) *Frazeologija, učenje i poučavanje* (str. 171–184). Reka: Sveučilište u Rijeci Filozofski fakultet.

Krek, S., Dobrovoljc, K., Erjavec, T., Može, S., Ledinek, N., Holz, N., ..., Zajc, A. (2019). Training corpus ssj500k 2.2, Slovenian language resource repository CLARIN.SI.

Križaj, M., & Bešter Turk, M. (2018). *Jezikovni pouk: Čemu, kaj in kako?* Priročnik za učitelje in učiteljice slovenščine v osnovni šoli. Ljubljana: Rokus Klett.

Kržišnik, E. (2006). Izraba semantične potence frazemov. *Slavistična revija*, *56*(1), 259–279.

Kržišnik, E. (2015). Frazeologija v šoli – drugič. *Jezik in slovstvo*, *60*(3–4), 131–142.

Ljubešić, N., & Erjavec, T. (2016). Corpus vs. Lexicon Supervision in Morphosyntactic Tagging: the Case of Slovene. *Language Resources and Evaluation Conference 2016*.

Logar Berginc, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š., & Krek, S. (2012). *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida*

in ccKRES: gradnja, vsebina, uporaba. Ljubljana: Trojina, zavod za uporabno slovenistiko, Fakulteta za družbene vede. Pridobljeno s https://www.fdv.uni-lj.si/docs/default-source/zalozba/pages-from-logar-et-al---korpusi.pdf?sfvrsn=2

Meterc, M. (2017). *Paremiološki optimum: Najbolj poznani in pogosti pregovori ter sorodne paremije v slovenščini*. Ljubljana: Založba ZRC, ZRC SAZU.

Meterc, M. (2019). Vpliv starosti na poznanost pregovorov, rekov in sorodnih paremij ter na paremiološko kompetenco slovenskih govorcev. V Ž. Macan (ur.), *Frazeologija, učenje i poučavanje* (str. 209–221). Rijeka.

Meterc, M. (2020). *Slovar pregovorov in sorodnih paremioloških izrazov. Rastoči slovar*. Pridobljeno s https://fran.si/

Nidorfer Šiškovič, M. (2013). Žanrskost funkcijskih besedilnih vrst. V A. Žele (ur.), *Družbena funkcijskost jezika: vidiki, merila, opredelitve*, *Obdobja 32* (str. 269–275). Ljubljana: Znanstvena založba Filozofske fakultete.

Nivre, J., Marneffe, M., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., ..., Zeman, D. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož: European Language Resources Association.

Odpiranje izobraževanja: inovativno poučevanje in učenje za vse z novimi tehnologijami in prosto dostopnimi učnimi viri. Pridobljeno s https://eur-lex.europa.eu/legal-content/SL/TXT/PDF/?uri=CELEX:52013DC0654&from=HU

Rozman, T., Krapš Vodopivec, I., Stritar, M., & Kosem, I. (2020). *Empirični pogled na pouk slovenskega jezika*. Ljubljana: Znanstvena založba FF UL. Pridobljeno s https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/227/327/5303-1

SEJO = Skupni evropski jezikovni okvir: učenje, poučevanje, ocenjevanje (2011). Irena Kovačič (pr.). El. knjiga. Ljubljana: Ministrstvo RS za šolstvo in šport, Urad za razvoj šolstva. Pridobljeno s https://centerslo.si/wp-content/uploads/2015/10/SEJO-komplet-za-splet.pdf

Starc, S. (2011). Stik disciplin v besedilu iz besednih in slikovnih semiotskih virov. V S. Kranjc (ur.), *Meddisciplinarnost v slovenistiki, Obdobja 30* (str. 433–440). Ljubljana: Znanstvena založba Filozofske fakultete.

Strateški okvir – Izobraževanje in usposabljanje 2020. Pridobljeno s http://ec.europa.eu/education/policy/strategic-framework_sl

Ulčnik, N. (2019). Izbor frazemov za bazo FRIDA. V N. Ulčnik (ur.), *Slovenščina na dlani 2* (str. 37–45). Ulčnik. Maribor: Univerzitetna založba Univerze. Pridobljeno s https://press.um.si/index.php/ump/catalog/book/447

Ulčnik, N., & Meterc, M. (2019). Izbor pregovorov za bazo FRIDA. V N. Ulčnik (ur.), *Slovenščina na dlani 2* (str. 47–55). Maribor: Univerzitetna založba Univerze. Pridobljeno s https://press.um.si/index.php/ump/catalog/book/447

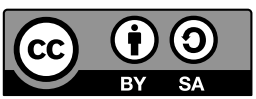
Vičar, B. (2015). Slovnični pristop k vizualni komunikaciji: vizualna analiza vojnih fotografij. V M. Smolej (ur.), *Slovnica in slovar – aktualni jezikovni opis, Obdobja 34* (str. 801–810). Ljubljana: Znanstvena založba Filozofske fakultete.

Voršič, I. (2018). Prvi odzivi učiteljic in učiteljev. V N. Ulčnik (ur.), *Slovenščina na dlani 1* (str. 89–91). Maribor: Univerzitetna založba Univerze. Pridobljeno s http://press.um.si/index.php/ump/catalog/book/341

E-LEARNING ENVIRONMENT »SLOVENŠČINA NA DLANI«: CHALLENGES AND SOLUTIONS

The paper describes three types of challenges that were detected in teaching Slovene as a mother tongue at schools. First, a number of orthographic and grammatic mistakes can be detected in pupils' writings (see Kosem et al., 2012; Križaj in Bešter Turk, 2018; Gomboc, 2019). Second, low phraseological literacy was noticed and the pupils often have problems understanding phrasemes (Voršič, 2018). Third, the challenges of communicative competence were addressed, referring to production and interpretation of different written, spoken as well as multimedia genres, as only appropriate genre literacy enables efficient use of different genres (Nidorfer Šiškovič, 2013). To address these challenges, we have developed a complex e-learning environment for improving writing and communication skills of Slovene pupils – "Slovenščina na dlani". The developed environment is divided into four general topics – orthography, grammar, phrasemes and texts. Each topic covers a number of subtopics, and for each sub-topic a number of exercises is available, along with explanations. We have used the most up-to-date language technologies and programming solutions in order to automatise the e-environment. The user's knowledge is automatically evaluated, and based on this s/he is automatically guided through the environment in a way to improve her/his writing and communication skills. The e-environment has also a special user interface for teachers which enables easy way to assign tasks as well as to track the performance of each pupil individually or a group of pupils as a whole. The gamification and professional graphic design fulfil the user experience. The "Slovenščina na dlani" will be freely available at https://slo-na-dlani.si from September 2021 on.

Keywords: learning Slovene, computer assisted language learning, e-learning



Slovenščina 2.0, 2021 (1)

NADGRADNJA ZGODOVINARSKEGA INDEKSA CITIRANOSTI

Katja MEDEN
Jožef Stefan Institut; Inštitut za novejšo zgodovino

Ana CVEK
Inštitut za novejšo zgodovino; Filozofska fakulteta, Univerza v Ljubljani

Meden, K., Cvek, A. (2021): Nadgradnja Zgodovinarskega indeksa citiranosti. Slovenščina 2.0, 9(1): 216–235.

DOI: https://doi.org/10.4312/slo2.0.2021.1.216-235

Začetki Zgodovinarskega indeksa citiranja segajo v leto 2003, ko so raziskovalci Inštituta za novejšo zgodovino začeli spremljati in sistematično popisovati citate za prijave projektov in programov na ARRS. Citatni indeks je doživel nekaj nadgradenj, poskusov harmonizacije podatkov in prečiščevanja relacijskih baz, vendar je bilo v zadnjih letih ugotovljeno, da sistem ne zadostuje potrebam indeksatorjev in uporabnikov. Pred nadgradnjo smo izvedli analizo podatkov, kjer so se identificirale največje težave. Nadgradnja je potekala v dveh delih; v prvem delu smo nadgradili administrativni del, v drugem delu pa spletno aplikacijo. Zgodovinarski indeks citiranja je bil med nadgradnjo tehnično posodobljen in s tem oblikovan tako, da je intuitiven za indeksatorje in uporabnike.

Ključne besede: Zgodovinarski indeks citiranosti, ZIC, nadgradnja, citatni indeksi

1 UVOD

Ocenjevanje uspešnosti raziskovalcev v humanistiki je v primerjavi z drugimi raziskovalnimi področji, predvsem naravoslovnimi, že od samih začetkov precej prikrajšano. Med drugim ocenjevanje temelji na frekvenci citiranosti, te podatke pa pridobimo iz različnih citatnih indeksov, kot sta na primer Web of Science (v nadaljevanju WOS) in Scopus. Monografije so primarni produkt raziskovalnega dela v humanistiki in družboslovju (Glänzel in Schoepflin, 1999; Hicks, 2004; Huang in Chang, 2008; Nederhof, 2006). V nasprotju z vrednotenjem raziskovalne uspešnosti v naravoslovju se ta področja teže vrednotijo, predvsem zaradi dejstva, da so monografije po večini bolj obsežne kot znanstveni članki (Kousha idr., 2011), in visokih kriterijev vključevanja publikacij v obstoječe indekse citiranja, na primer WOS in Scopus. Med pomembnejše kriterije spadajo redno izhajanje serijske publikacije, jezik publikacije, recenziranost, spoštovanje mednarodnih standardov (kot so informativni naslov, povzetek, popolna bibliografska informacija za vse citirane reference), poleg pogojev pa težavo predstavlja tudi indeksiranje monografij. Obstoječi citatni indeksi se namreč bolj osredotočajo na serijske publikacije. Neenakosti pri vključevanju publikacij v citatne indekse so na Inštitutu za novejšo zgodovino skušali zamejiti že v letu 2003. Raziskovalci so začutili potrebo po spremljanju in sistematičnem popisovanju citatov za prijave projektov in programov, kar predstavlja zametek Zgodovinarskega indeksa citiranja (v nadaljevanju ZIC). Osnovni namen je bil ustvariti bazo citatov iz slovenskih zgodovinskih monografij, osrednjih znanstvenih časopisov in revij (Lazarević in Zemljič, 2003). Začetna shema baze, ki je bila precej enostavna, je ob nastanku dobro zadovoljevala potrebe raziskovalcev, vendar so se sčasoma pokazale pomanjkljivosti (Pančur idr., 2014), ki so vodile v nadaljnje nadgradnje, poskuse harmonizacije podatkov in prečiščevanja relacijskih baz. ZIC trenutno vsebuje 4.837 vseh vnosov, od tega 2.901 vnos serijskih publikacij in 1.936 vnosov monografij in poglavij iz monografij, kar predstavlja razmerje 59,9 % serijskih publikacij ter 39,1 % monografij in poglavij iz monografij.

Zadnja nadgradnja je potekala leta 2012 in predstavlja osnovo in temelj nadgradnje, ki je predstavljena v nadaljnjem besedilu članka.

2 CITATNI INDEKSI IN HUMANISTIKA

Kot omenjeno, sta humanistika in družboslovje pri vrednotenju znanstvene uspešnosti v nasprotju z naravoslovnimi vedami nekoliko prikrajšana pri vključevanju raziskovalne produkcije v mednarodne citatne indekse, kot sta Web of Science (WOS) in Scopus. V Sloveniji vrednotenje raziskovalne uspešnosti poteka prek Informacijskega sistema o raziskovalni dejavnosti (SICRIS), v katerem je popisana celotna slovenska raziskovalna produkcija in je povezan s prej omenjenima mednarodnima citatnima indeksoma WOS in Scopus (Curk idr., 2006). Pomembno je poudariti, da so točke, pridobljene prek SICRIS, osnovno merilo za točkovanje raziskovalne uspešnosti in so neposredno povezane s procesom financiranja raziskovalnih projektov in programov prek Agencije za raziskovalno dejavnost Republike Slovenije (ARRS).

Z vprašanjem vključenosti humanistike in družboslovja v WOS in Scopus se je ukvarjalo več raziskav (Ball in Tunger, 2006; Bartol idr., 2014), kjer obstaja konsenz o tem, da je za vključevanje humanistike in družboslovja Scopus občutno bolj primeren kot pa WOS. Vendar kot omenjeno, je monografija primarna oblika znanstvene produkcije v humanistiki, ki pa ji citatni indeksi niso najbolj naklonjeni. Podatki kažejo, da WOS zajema okoli 12.000 znanstvenih revij in samo okoli 50.000 monografij, medtem ko Scopus zajema več kot 21.500 znanstvenih revij in 113.000 znanstvenih monografij. Število monografij v indeksu Scopus odraža večji obseg monografij v primerjavi z WOS, pa vendar monografije v primerjavi s številom znanstvenih člankov v revijah predstavljajo zgolj zanemarljiv del citatnega indeksa (Južnič, 2017).

Podobno stanje je tudi pri vključevanju slovenske raziskovalne produkcije v humanistiki. Južnič in Čadej (2016) v svoji raziskavi ugotavljata, da baza Scopus bistveno bolje zajema slovensko humanistično in družboslovno znanstveno publikacijo v primerjavi z WOS. Razlogi za to so različni: od dejstva, da je Scopus neprimerno bolj naklonjen vključevanju neangleških revij slabše razvitih in manjših držav vzhodne Evrope, do milejših meril vključevanja publikacij (Pajić, 2015).

Ne glede na dejstvo, da je Scopus bolj primeren za vključevanje slovenskih znanstvenih revij in monografij v humanistiki, pa še vedno obstaja vrzel pri vključevanju teh publikacij v Scopus. To pa poskušamo zamejiti s citatnimi indeksi, kot je npr. ZIC, ki so prilagojeni specifičnim lastnostim področja, ki ga

pokrivajo (v primeru humanistike je torej največje odstopanje v vključevanju monografskih publikacij).

3 CILJI IN POTEK NADGRADNJE

Pri postopku nadgradnje smo z uporabo sodobnih tehnologij in estetsko privlačne grafične podobe želeli preoblikovati administratorski spletni vmesnik in indeksatorju omogočiti prijazno in pregledno izkušnjo pri urejanju podatkov. Najpomembnejši cilj nadgradnje je bila postavitev ZIC kot ločene aplikacije. Ker je baza MySQL trenutno integralni del portala SIstory in se upravlja s pomočjo skupne administracije, je treba podatkovno bazo ZIC postaviti kot ločeno aplikacijo na poddomeni portala SIstory. Razlog za to je načrtovana postavitev nove digitalne knjižnice portala SIstory kot samostojnega repozitorija z ločeno administracijo. Poleg ločene baze in administracije smo pri nadgradnji upoštevali naslednje sklope problemov. V prejšnji nadgradnji uvoz in izvoz podatkov nista bila mogoča, zato smo želeli to omogočiti. Prav tako smo želeli, da je spletna aplikacija narejena modularno, kar bo omogočalo dodajanje novih funkcionalnih rešitev. Pri uporabniškem vmesniku smo želeli, da je stran prijazna za mobilne obiskovalce, pri iskalniku pa smo želeli doseči hitro in pregledno iskanje po podatkih. Nadgrajeni administracijski modul naj bi omogočal enostavnejši dostop in upravljanje vseh podatkov ter z geslom zaščiten dostop do administracije. Izbrani osnovni podatki morajo biti z ustreznim vmesnikom prosto dostopni strojnemu zajemu podatkov (Pančur, 2019b).

Pri postavljanju ciljev in procesu nadgradnje smo izhajali iz temeljnih načel Raziskovalne infrastrukture slovenskega zgodovinopisja (v nadaljevanju RI INZ), ki vključujejo uporabo uveljavljenih in razširjenih tehnologij, ki jih člani infrastrukture dobro poznajo in obvladajo (načeli enostavnosti in poznavanja), modularno nadgrajevanje obstoječih tehnologij (načelo fleksibilnosti) in uporabo odprtih ali lastniških standardov (načelo odprtosti) (Pančur in Šorn, 2019). V procesu nadgradnje smo tako uporabljali tehnologije, ki jih priporoča RI INZ (Pančur, 2019a) in upoštevajo načeli enostavnosti in poznavanja HTML5 in CSS3, najnovejše verzije PHP, MySQL, ElasticSearch engine, JavaScript in JavaScript knjižnice. Pomemben vidik nadgradnje je tudi vidik interoperabilnosti, ki se v svojem pomenu prepleta z načelom fleksibilnosti. Fleksibilnost in interoperabilnost sistema želimo doseči z implementacijo

aplikacijskega profila MODS za uvoz in izvoz metapodatkov v različnih formatih, ki podpirajo nadaljnjo diseminacijo in izmenjavo podatkov z drugimi informacijskimi sistemi. Nadgradnja je potekala v posameznih sklopih, ki so opisani v nadaljevanju besedila.

4 REZULTATI NADGRADNJE

Nadgradnja je potekala v dveh delih: prvi del se nanaša na administrativni sistem SIstory. Nadgradnja v tem delu zajema preoblikovanje mask in njihovih polj, postavitev nove sheme XML po standardu MODS za uvoz in izvoz podatkov, iskalnik, ki temelji na tehnologiji ElasticSearch, ter migracije vrednosti ločenih polj Avtor(ji). Drugi del se osredotoča na nadgradnjo spletne aplikacije in uporabniškega vmesnika. Pri programski nadgradnji smo sodelovali z zunanjimi sodelavci Infrastrukture.

4.1 Administrativni sistem Sistory

4.1.1 Maske za vnos podatkov

Glavna sprememba v administracijskem sistemu (admin) je prehod s prej enotne maske na dve ločeni. Enotna maska je vsebovala tri razdelke: *Splošni podatki, Podatki o viru* in *Vsebinska obdelava*. Vnos podatkov v maske poteka ročno, podatkovna polja v enotni maski pa so bila nejasna (npr. ponavljanje polja za vnos id številke COBISS, imena avtorja idr.), nekatera tudi brez pomena za potrebe citatnega indeksa. Tako je bil na primer razdelek Vsebinska obdelava za citatni indeks povsem neuporaben, saj vsak zapis vsebuje identifikatorje s povezavami na zapise publikacij (COBISS, SIstory) s polnim metapodatkovnim opisom.

Iz enotne maske sta nastali dve neodvisni maski za vnos podatkov v ZIC V2. Iz maske za vnos publikacije sta nastali dve: *maska za vnos monografij* in *maska za vnos serijskih publikacij*, ki dovoljujeta natančnejši opis glede na publikacijo, ki jo indeksiramo. Vsaka izmed mask, tako kot v prejšnji verziji, vsebuje tudi masko za vnos citatov. Maske so bile oblikovane na podlagi zaznanih težav v prejšnjem administracijskem sistemu, o katerih so poročali indeksatorji, ter na podlagi potreb za opis določene publikacije in citatnega indeksa. Spodnja preglednica (Preglednica 1) prikazuje polja oziroma metapodatke za opis posameznih del in citatov.

Preglednica 1: *Metapodatki mask za vnos podatkov*

Metapodatek	min/max. št	Podatkovni tip	Maska (Mono, Serijska, Citat)	Primer
Cobiss ID	0,1	ID	M, S, C	3278924
Sistory ID	0,1	ID	M, S, C	handle. net/11686/4320
ISBN	0,1	ID	M	987-961-3421-43
ISSN	0,1	ID	S	0353-0329
Jezik	1,1	ISO639-2b	M, S	slv - slovenski
Tipologija	1,1	COBISS tipologija	M, S	1.16 – Samostojni znan. sestavek
Tip	0,1	interni seznam	M	Poglavje v monografiji
Avtorji	1,neomejeno	niz	M, S, C	Marko Zajc
Naslov	1,1	niz	M, S, C	Slovenski intelektualci in ...
Vzporedni naslov	0,1	niz	M, S	Slovenian Intellectuals ...
Naslov zbornika	0,1	niz	M	Slovenija v Jugoslaviji
Naslov vira	0,1	niz	S	Prispevki za novejšo zgodovino
Uredniki	0,neomejeno	niz	M	Zdenko Čepič (ur.)
Kraj	0,1	niz	M, S, C	Ljubljana
Založba	0,1	niz	M, S, C	Založba INZ
Leto	0,1	številčna vrednost	M, S, C	2015
Letnik	0,1	številčna vrednost	S, C	57
Številka	0,1	številčna vrednost	S, C	1
Zbirka	0,1	niz; št. vrednost	M	Vpogledi; 10
Stran	0,1	št. vrednost	M, S, C	241–256
DOI	0,1	ID	S, C	10.1090/019339135
Baza citatov INZ	0,1	gumb	M, S	DA
Citat na strani	1,1	št. vrednost	C	34
Vir	0,1	niz	C	Prispevki za novejšo zgodovino

Večina elementov, potrebnih za opis publikacij, je ostala nespremenjena. Po opravljeni analizi elementov mask smo izpostavili ključna polja za potrebe

opisa publikacij in njihovih citatov. Večina polj je splošne narave (npr. avtor, naslov, leto, kraj itd.), publikacije, ki jih vnašamo (monografije in serijske publikacije), pa se med seboj razlikujejo v določenih vidikih. Ločeni maski s prilagojenimi polji omogočata (z indeksatorskega vidika) kakovostnejšo indeksacijo publikacije. Elementi so bili spremenjeni ali prilagojeni, saj določeni niso bili ažurirani (na primer element Tipologija) ali niso omogočali dovolj natančnega opisa (element Avtor). Pri poljih Avtor in Urednik smo metapodatkovno polje ločili na dve polji: Ime in Priimek. S tem smo zagotovili natančnejši, bolj strukturiran opis in posledično boljše prikazovanje podatkov. Zaradi nove strukture polja je bilo za povezovanje vrednosti polj treba opraviti migracijo vrednosti iz starih, neločenih polj v nova, strukturno ločena polja v obliki Priimek, Ime (za namen prikaza). Nekaterih elementov iz stare maske v novih maskah nismo vključili, npr. Ključne besede ali Država, saj so bili za opis publikacij v citatnem indeksu nepotrebni. Dodani so bili tudi novi elementi, ki jih starejša maska za vnos podatkov ni vsebovala, ker ti podatki še niso bili potrebni. Tu govorimo predvsem o maski za vnos serijskih publikacij in citatov, kjer smo dodali polji DOI in URL, ki omogočata enoznačno, trajno identifikacijo, prav tako pa poleg polja Sistory ID uporabniku omogočata hiter dostop do publikacije.

Pri analizi obstoječih zapisov se je izkazalo, da so pomanjkljivi in neenotni. Do takšnih napak je prihajalo predvsem zato, ker indeksatorji niso imeli nobenih konkretnih navodil in so publikacije v maski (glavni vnos in citat) vpisovali po lastni presoji. Zato smo se pri nadgradnji odločili, da indeksatorjem ponudimo pomoč, ki jim bo olajšala vnos podatkov, še bolj pomembno pa je, da bi s temi navodili oz. pomočjo radi zagotovili čim bolj enotno indeksacijo ter pravilnejše in natančnejše zapise v indeksu. Ob vsakem polju je pri vseh treh maskah opis polja z navodili za vnos in primeri, ki naj bi bili indeksatorju v pomoč oz. oporo pri vpisovanju podatkov. Tu velja poudariti, da se zavedamo, da se bodo napake kljub pomoči še vedno pojavljale, saj se podatki vpisujejo ročno. S tem, da dajemo navodila za vnos, poskušamo zmanjšati število pogostih napak.

4.1.2 ElasticSearch iskalnik in filtriranje

Iskalnik ElasticSearch je distribucijsko, odprtokodno in analitično orodje za vse vrste podatkov, skupaj z besedilnimi, številčnimi, geoprostorskimi, strukturiranimi in nestrukturiranimi podatki (What is ElasticSearch, b.d.). ElasticSearch temelji na knjižnici Lucene Apache, ki je odprtokodna Java knjižnica za besedilno iskanje. ElasticSearch ponuja najrazličnejše možnosti, kot so prilagodljiva mapiranja podatkovnih polj, shranjevanje vrednosti ključev (ang. Key Value Store) itd., sam delovni tok pa je sestavljen iz petih korakov (What is ElasticSearch, b.d.; Divya in Goyal, 2013):

- **Zajem podatkov** (ang. *Data ingestion*): Postopek zajema vrednosti se začne s tako imenovanim *data ingestion*, v katerem so surovi podatki zajeti v iskalnik iz različnih virov. Podatki, ki jih zajamemo, so lahko v kateremkoli formatu in kakršnekoli velikosti.
- **Pretvorba v format JSON**: Zajete podatke pretvorimo v format JSON *JavaScript Object Notation*), ki omogoča interoperabilnost podatkov med različnimi sistemi.
- **Tokenizacija**: Zajete podatke je potrebno ločiti na posamezne besede, kar dosežemo z uporabo funkcije *Tokenizer*.
- **Indeksacija**: V naslednjem delu se oblikuje ElasticSearch index, ki je zbirka med seboj povezanih dokumentov. Vsak izmed dokumentov je povezan s ključi (imena, podatkovna polja ali lastnosti) in njihovimi vrednostmi (niz, številke, Boolovi operatorji, nabor vrednosti ...).
- **Parsiranje podatkov** (*Data parsing*): Parser bo procesiral iskalno poizvedbo (ang. search query), preiskal indeksirani dokument in poiskal morebitne ustrezne zadetke.

Za implementacijo iskalnika ElasticSearch za ZIC v administrativnem sistemu podatke zajamemo iz relacijske baze, ki temelji na tehnologiji MySQL (*What is ElasticSearch, b.d.*). Indeksirani ključi so v tem primeru podatkovna polja, ki bodo namenjena iskalnim poizvedbam, in njihove vrednosti (ki so večinoma besedilni nizi ali številčne vrednosti). Iskalnik ponuja izvajanje kompleksnih iskalnih poizvedb, ZIC uporablja funkcijo *simple string query*:

```
GET /_search
{
   »query«: {
      »simple_query_string« : {
      »«query«: »Mojca + Šorn +
      \«Življenje Ljubljančanov
      med drugo svetovno vojno\««
      »fields«: [»title^5«, »body«],
      »default_operator«: »and«
      }
   }
}
```

Funkcija uporablja preprosto sintakso za besedilne iskalne poizvedbe, na podlagi katere vrača iskalne rezultate z uporabo parserja.

Za iskalnik v spletni aplikaciji indeksiramo zgolj polji *Avtor* in *Naslov*, filtri v spletni aplikaciji pa imajo indeksirana polja (in njihove vrednosti) *Identifikator*, *Avtor*, *Naslov*, *Tipologija*, *Leto*, *Kraj* in *Št. citatov*. V administrativnem sistemu je bil filter nadgrajen. Prej je omogočal filtriranje po naslednjih parametrih: *Avtor*, *Leto*, *Naslov*, *Vir* in *Kraj*. Ti po mnenju indeksatorjev niso omogočali učinkovitega in natančnega iskanja zapisov znotraj baze. Novi filtri vsebujejo večje število parametrov: *Tip (monografija/serijska publikacija)*, *ID*, *Avtor*, *Naslov*, *Leto* in *Vir*. Iskalnik ElasticSearch podpira tudi funkcijo samodokončanja iskalne poizvedbe, poznano tudi pod imenom *Autocomplete* ali *Completion suggester*. Funkcija je optimizirana za hitrost tipkanja, saj se prilagaja hitrosti tipkanja iskalne poizvedbe, ki jo uporabnik vnese. Podpira izključno funkcijo *type as you go* in ni mišljena za samodejno korekcijo iskalne poizvedbe ali funkcije *Ali ste mislili* (What is ElasticSearch, b.d.). V našem primeru se na funkcijo samodokončanja, enako kot pri osnovnem iskalniku, vežeta zgolj polji Avtor in Naslov.

4.1.3 Uvoz in izvoz metapodatkov – MODS aplikacijski profil

XML ali eXtensible Markup Format prihaja iz družine označevalnih jezikov, kot sta SGML in HTML. Vendar pa se od omenjenih formatov razlikuje predvsem po fleksibilnosti – v primerjavi s HTML omogoča oblikovanje lastnih označevalcev oz. elementov (angl. tag) in s tem predstavlja enega izmed najpogosteje uporabljenih standardov za izmenjavo podatkov v digitalni humanistiki (Extensible markup language (XML) 1.0 (fifth edition), b. d.). Že v

prejšnjih verzijah baze je izvoz podatkov bil mogoč v formatu XML. Shema je predpostavljala lastne elemente (npr. *OpTipBiblEnote* za označevanje tipologije vpisanega vnosa ali *OpSistoryUrnId* za vnos SIstory identifikatorja) in ni upoštevala kateregakoli metapodatkovnega standarda, kot je na primer Dublin Core. Kot je bilo že omenjeno, to pomeni zmanjšano stopnjo interoperabilnosti podatkov, saj gre za unikatne elemente oz. označevalce, ki jih drugi (informacijski) sistemi ne uporabljajo. Pri prenosu podatkov lahko zaradi neujemajočih shem (oziroma elementov) prihaja do izgube določenega dela podatkov ali celo do izgube konteksta, v katerem so podatki. Čeprav je med metapodatkovnimi standardi najbolj razširjen in uporabljen standard Dublin Core ali njegova razširjena različica, DCTERMS, pa imata oba standarda precej omejen nabor elementov, ki ne zadostuje našim potrebam. Čeprav bi z implementacijo enega izmed omenjenih standardov dosegli višjo stopnjo interoperabilnosti, pa smo se zaradi omejitev nabora elementov odločili za metapodatkovni standard MODS.

Metadata Object Description Schema (MODS) je shema XML z bibliografskimi elementi (oziroma naborom elementov), ki jo lahko uporabljamo za najrazličnejše potrebe. Shema izhaja iz standarda za bibliografske zapise MARC21, vendar za svoje elemente namesto številčnega zapisa (na primer polje 222 za glavni naslov (ang. *Key Title*) in 210 za skrajšan naslov (ang. *Abbreviated Title*) uporablja besedilne označevalce oziroma elemente (ang. *language-based tags*) (MODS User Guidelines, Version 3 (Metadata Object Description Schema), b.d.).

MODS namreč vsebuje dovolj obsežen nabor elementov, ki ustreza našim potrebam, hkrati pa je še vedno dovolj razširjen in zato omogoča zaželeno stopnjo interoperabilnosti naših podatkov z minimalno izgubo konteksta.

Postopek prenosa podatkov iz interne sheme v metapodatkovno shemo MODS je vključeval tri faze:

- Pregled elementov stare sheme, ki je za svoje elemente upoštevala imena, kot so *OpTipBiblEnote* ali *OpSistoryUrnId;* del elementa 'Op' se nanaša na publikacijo, ki jo opisujemo (Op = original publication), 'Pv' pa označuje podatke za vir publikacije, sledi interno poimenovanje polja (ki ustreza imenu polja, iz katerega vzamemo podatke).

- Preslikava internih polj (poimenovanje po meri) v metapodatkovni standard MODS in komentiranje kode (navodila za programerja, iz katerih polj v stari metapodatkovni shemi se vežejo vrednosti v nove elemente). Iz ene sheme sta nastali dve novi, upoštevali smo novo strukturo mask za vnos podatkov, tako kot smo predhodno enotno masko razdelili na masko za monografije in serijske publikacije. V aplikacijskem profilu v skupnem metapodatkovnem zapisu v formatu XML sta ločena zapisa mask definirana z elementom mods in identifikatorjem ID=pub za oznako zapisa za monografijo ali serijsko publikacijo (na primer *mods ID=pub.224*) ali elementom relatedItem in identifikatorjem za oznako navedenih del, na primer *relatedItem type=referencesID=ref.1*.

- Prenos vrednosti iz starih internih polj v polja MODS ima svoje prednosti; poleg dejstva, da tako povečamo interoperabilnost svojih podatkov z drugimi sistemi, s tem pridobimo večjo strukturiranost in pogosto

```
<?xml version="1.0"?>
<root>
  <delo>
      <field name="ID" >4916</field>
      <field name="OpTipBiblEnote" >1</field>
      <field name="OpTipologija" >1</field>
      <field name="OpZvrst" >1</field>
      <field name="OpJezik" >21</field>
      <field name="OpDrzava" >17</field>
      <field name="OpStAvtorjev" >1</field>
      <field name="OpCobId" >3955316</field>
      <field name="OpSistoryUrnId" >0</field>
      <field name="OpAvtor0" >Hadalin Jurij</field>
      <field name="OpAvtor1" ></field>
      <field name="OpAvtor2" ></field>
      <field name="OpUrednik" ></field>
      <field name="OpNaslov" >Jugoslovanska malica</field>
      <field name="OpVzpNaslov" ></field>
      <field name="OpPodnaslov" ></field>
      <field name="PvCobId" >302353920</field>
      <field name="PvISSN" >978-961-6386-99-9</field>
      <field name="PvTip" >2</field>
      <field name="PvAvtor" ></field>
      <field name="PvNaslov" >Mimohod blaga : materialna kultura potrošniške družbe na Slovenskem</field>
      <field name="PvNaslovKratki" ></field>
      <field name="PvPodnaslov" ></field>
      <field name="PvVzporedniNaslov" ></field>
      <field name="PvZbirka" >Vpogledi ; 22</field>
      <field name="PvKraj" >Ljubljana</field>
      <field name="PvZalozba" >Inštitut za novejšo zgodovino</field>
      <field name="PvLetnik" ></field>
      <field name="PvLeto" >2019</field>
      <field name="PvSt" ></field>
      <field name="PvStran" >151-165</field>
      <field name="KwDeskriptor1" ></field>
      <field name="KwDeskriptor2" ></field>
```

Slika 1: Metapodatkovna polja maske za vnos podatkov pred nadgradnjo.

tudi dodatne podatke, ki jih v stari shemi ne bi mogli implementirati. Element OpJezik ima za svojo vrednost na primer le številčno vrednost »21«, kar se navezuje na interni nekontroliran seznam jezikovnih vrednosti, novi element pa v svoji strukturi dovoljuje navedbo avtoritete in tipa poimenovanja. Tako poleg jezikovne kode pridobimo tudi podatek o standardu oziroma kontroliranem seznamu, ki je bil uporabljen, s tem pa tudi standardiziramo vrednost zapisa. Slika 1 prikazuje strukturo in del elementov stare, interne metapodatkovne sheme.

Spodaj so prikazani stari in novi način poimenovanja ter primerjava strukture posameznega zapisa:

Interna shema ZIC (element Avtor):

```
<field name="OpAvtor0">Hadalin Jurij</field>
```

Aplikacijski profil v XML:

```
<name type="personal">
      <namePart>Priimek Ime avtorja</namePart>
   <role>
      <roleTerm type="code">cre</roleTerm>
      <roleTerm>Avtor</roleTerm>
   </role>
      <namePart type="family">Priimek</namePart>
      <namePart type="given">Ime</namePart>
</name>
```

Interna shema ZIC (element Jezik)

```
<field name="OpJezik" >21</field>
```

Aplikacijski profil v XML:

```
<language>
   <languageTerm type="code">slv</languageTerm>
   <scriptTerm type="code">Latin</scriptTerm>
</language>
```

Interna shema ZIC (element Tipologija):

```
<field name="OpTipologija" >1</field>
```

Aplikacijski profil:

```
<classification authorityURI="https://www.izum.si/">101</classification>
```

Z novim aplikacijskim profilom, ki izhaja iz metapodatkovnega standarda MODS, smo namesto internih metapodatkovnih elementov v shemi uporabili obstoječi in razširjeni metapodatkovni standard MODS. S tem smo naslovili dve izmed temeljnih načel: poznavanje oziroma uporabo poznanih in razširjenih tehnologij ter načelo interoperabilnosti. Format XML nam namreč zagotavlja lažje izmenjevanje in diseminacijo podatkov z drugimi sistemi.

4.1.4 Migracija vrednosti polj avtorji

Enega izmed večjih problemov, ki nam ga je delno uspelo rešiti med nadgradnjo, predstavlja migracija vrednosti polja Avtor(ji) iz skupnega polja v dve ločeni. Problem je nastal zaradi neenotnega zapisa oziroma različnih oblik vrednosti Priimek in Ime (oblike: *Priimek, Ime; Ime in Priimek, Ime, Priimek* …) ter naštevanja več avtorjev v enem polju (*Avtor1; Avtor2* …), ki so bili med seboj ločeni z različnimi ločili. Ta problem nam je uspelo rešiti zgolj delno: migracija, ki je potekala strojno, je bila uspešna na poljih, ki so se med seboj ujemala, pri določenih zapisih pa to ni bilo mogoče (primer *Ime Ime, Priimek*), zato zahteva ročne popravke. Te napake bomo lahko odpravili po začetku procesa prečiščevanja baze, ki pa za zdaj še ni predviden.

4.2 Spletna aplikacija in uporabniški vmesnik

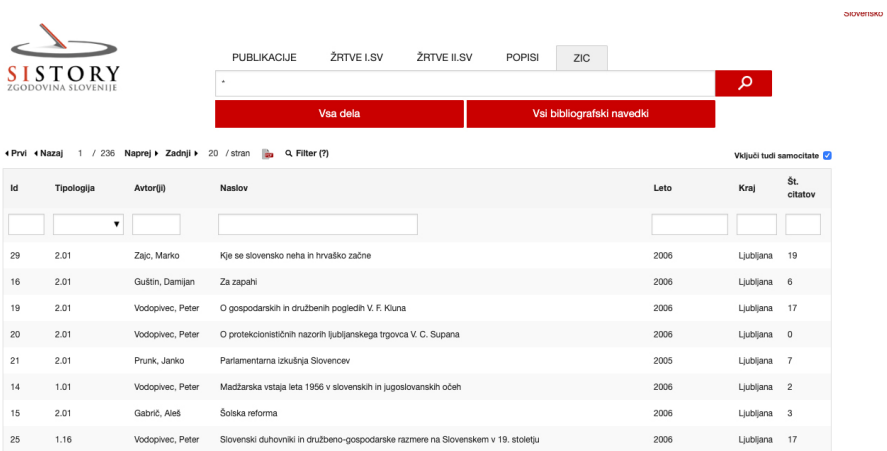
4.2.1 Podatkovna baza Vseh del in podatkovna baza Vseh bibliografskih navedb

Spletna aplikacija vsebuje dve podatkovni bazi: bazo *Vsa dela* in podatkovno bazo *Vse bibliografske navedbe*. Razlog za dve medsebojno ločeni bazi je v prikazu rezultatov, še natančneje v prikazu števila prejetih citatov pri določenem zapisu. Pri izpisu rezultatov je na voljo število citatov, ki jih je določeno delo prejelo, vendar ti podatki morda niso pravilni, ker se število prejetih citatov določenega dela veže na ujemanje naslova pri glavnem vnosu (maska za vnos glavnega zapisa) in pri citatu (maska za vnos citata). Kot pa smo omenili že zgoraj, nemalokrat pride do napak. Zaradi tega je potrebna druga baza *Vse bibliografske navedbe*, po kateri je omogočeno brskanje z uporabo filtrov. Ta baza dovoljuje uporabniku dodaten in bolj natančen vpogled v citate, saj tu dejansko vidimo vse vnesene citate, indeksatorjem pa predstavlja dodatno orodje za lažje popravke že obstoječih zapisov (preglednejše iskanje zapisov slabše kakovosti).

4.2.2 Prikaz iskalnih rezultatov

Iskalni rezultati so prikazani v obliki tabel, ki uporabnikom ponujajo tudi filtriranje rezultatov oziroma omogočajo ožanje iskalne poizvedbe znotraj tabele. Rezultate je mogoče tudi razvrščati. Poleg filtriranja je uporabniku omogočen izvoz zadetkov na seznamu rezultatov in posameznega zadetka v formatu PDF. Za uporabnike sta prav tako pripravljeni tudi dve vrsti pomoči: osnovna razlaga uporabe citatnega indeksa na prvi strani ZIC (iskanje/brskanje) in manjši namig pri uporabi filtrov s primeri uporabe ločil. Prikaz posameznega zapisa uporabniku dovoljuje vpogled v osnovne podatke (metapodatke dela), osnovne podatke vseh del, v katerih je bil citiran, in avtorjev seznam literature. Podatki so prikazani v dveh ločenih tabelah, *Citirano v* in *Seznam literature*, zapisi so med seboj povezani.

Med oblikovanjem vmesnika so v vmesnih fazah sodelovali raziskovalci/uporabniki, s katerimi smo testirali odzive na novi vmesnik, novo podatkovno strukturo in nove funkcionalnosti. Največ težav je predstavljala terminologija, predvsem na podlagi dejstva, da se zgodovinarsko dojemanje terminov literature in virov precej razlikuje od pojmovanja na področju tehnologije. Nerodna poimenovanja iz prejšnje verzije vmesnika (*Avtor citira, Citiranost Avtorja*) je bilo treba nadomestiti s terminom, ki bo uporabnikom razumljiv. Kot že omenjeno, smo se na podlagi tega odločili za osnovno iskanje in dve ločeni bazi, ki sta po številnih preimenovanjih pridobili ime *Vsa dela* in *Vsi bibliografski*



Slika 2: Trenutni uporabniški vmesnik ZIC-a.

navedki. Čeprav sta imeni daljši, smo prednost namenili razlagi terminov, saj so uporabniki menili, da sta ti poimenovanji najbolj jasni in logični.

Poleg terminologije je problem predstavljala tudi postavitev elementov na spletni strani (predvsem gumbi). Tu se je izkazalo, da je uporabnike precej zmedla postavitev gumbov za obe bazi, saj so mislili, da s klikom na npr. *Vsa dela* dobijo vsa dela iskanega avtorja. Težavo smo odpravili tako, da smo ustvarili različne statične verzije uporabniškega vmesnika in s pomočjo uporabnikov določili tisto, ki je najbolj jasna in intuitivna.

4.2.3 Uporaba indeksa citiranosti

Primarni uporabniki citatnega indeksa so raziskovalci, ki lahko v sistemu enostavno preverijo št. prejetih citatov za posamezno avtorsko delo; če je to indeksirano v sistem. Poleg izpisa iz sistema SICRIS (Slovenian Current Research Information System), ki je osnova za vrednotenje znanstvene uspešnosti na posameznem raziskovalnem področju, lahko izpis iz ZIC predstavlja dodano vrednost pri prijavljanju projektov ali programov na področju humanistike in pri obnavljanju ali napredovanju v višje znanstvene nazive. Poleg raziskovalcev si z ZIC lahko pomagajo tudi uredniki revij, ki želijo preveriti, kolikokrat so bili posamezni članki citirani, in s tem upravičijo obstoj revije. Poleg primarne naloge, ki je zagotavljanje vpogleda v število prejetih citatov, pa indeks ponuja tudi druge možnosti, ki jih stari ZIC ni ponujal. Te naj bi uporabniku omogočile prijetnejšo interakcijo s sistemom. Ena izmed takšnih funkcionalnosti je npr. možnost *prijaznega kopiranja*, ki uporabniku omogoča lažje navajanje virov v svojih delih, saj ZIC ponuja skoraj popolne bibliografske podatke, ali npr. izpis števila citatov v formatu PDF ipd. Indeks ponuja tudi možnost dostopa do polnega besedila, če je le-to na voljo na sestrskem spletnem portalu Zgodovina Slovenije – SIstory.

5 SKLEP

Sistem je bil že v začetni zasnovi izjemno ambiciozen in zaradi načina objavljanja v zgodovinopisju izjemno potreben. Vendar je Zgodovinarski indeks citiranja zadnja leta nekoliko stagniral. Po pregledu in analizi podatkov smo ugotovili, da je nadgradnja potrebna, saj sistem ne zadostuje potrebam indeksatorjev in uporabnikov. Začeli smo nadgradnjo administrativnega dela,

kjer smo preoblikovali oz. nadgradili nove maske, nadgradili metapodatkovno shemo oziroma ustvarili nov aplikacijski profil na podlagi metapodatkovnega standarda MODS, filtre in dodali pomoč indeksatorjem, ki naj bi pripomogla k poenotenim zapisom. Poleg administrativnega dela smo nadgradili tudi uporabniški vmesnik z občasnim testiranjem baze in njenih komponent z raziskovalci. Z omenjeno nadgradnjo smo rešili večino zaznanih problemov, od nejasnih in nepotrebnih polj vnosa podatkov in razčlenitve mask, ki indeksatorju omogočajo lažje in natančnejše oblikovanje zapisov, oblikovanja aplikacijskega profila MODS, ki omogoča lažji uvoz in izvoz podatkov, do uporabniku prijaznejšega vmesnika itd. Vseh težav pa zaradi omejitev, povezanih z ročnim vnosom podatkov, ni bilo mogoče v celoti rešiti. To velja predvsem za postopek migracije polja Avtorji, kjer bo problem v celoti rešen šele po prečiščenju celotne baze podatkov. Postopek prečiščenja bo pripomogel tudi k poenotenju zapisov, kar bo omogočalo, da uporabniki v sistemu pridobijo zanesljive in kakovostne informacije. Pri nadgradnji Zgodovinarskega citatnega indeksa smo dosegli zastavljene cilje. Sistem smo tehnično posodobili in ZIC postavili kot ločeno spletno aplikacijo na poddomeni portala SIstory. Spletna aplikacija je narejena modularno, zato je mogoče dodajati nove funkcionalne rešitve, iskalnik s tehnologijo ElasticSearch pa omogoča natančnejše in preglednejše iskanje po podatkih.

V prihodnosti želimo poleg že obstoječih funkcionalnosti dodati še druge možnosti, ki bi olajšale delo indeksatorjem, uporabnikom pa omogočile prijetnejšo uporabniško izkušnjo. Te možnosti so npr. avtomatizirano vnašanje osnovnih podatkov iz vnosov, ki so povezani in dostopni na portalu SIstory, ter možnost samodejnega generiranja citatov po različnih citatnih stilih (npr. APA, Chicago idr.). Z nadgradnjo Zgodovinarskega indeksa citiranosti smo tako oblikovali sistem, ki je intuitiven za indeksatorje in uporabnike, s tem pa zagotovili, da ZIC izpolni svoj namen.

Zahvala

Raziskavo je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije v okviru programa Raziskovalne infrastrukture slovenskega zgodovinopisja (I0-0013) in slovenske raziskovalne infrastrukture DARIAH SI.

LITERATURA

Ball, R., & Tunger, D. (2006). Science indicators revisited-Science Citation Index versus SCOPUS: A bibliometric comparison of both citation databases. *Information Services and Use*, *26*(4), 293–301.

Bartol, T., Budimir, G., Dekleva-Smrekar, D., Pušnik, M., & Južnič, P. (2014). Assessment of research fields in Scopus and Web of Science in the view of national research evaluation in Slovenia. *Scientometrics*, *98*(2), 1491–1504.

Curk, L., Budimir, G., Seljak, T., & Gerkes, M. (2006). Linking the SICRIS-COBISS.SI-Web of Science systems. *Organizacija znanja*, *11*(4), 230–235.

Divya, M. S., & Goyal, S. K. (2013). ElasticSearch: An advanced and quick search technique to handle voluminous data. *Compusoft*, *2*(6), 171.

Extensible markup language (XML) 1.0 (fifth edition). Pridobljeno https://www.w3.org/TR/xml/

Glänzel, W., & Schoepflin, U. (1999). A bibliometric study of reference literature in the sciences and social sciences. V *Information Processing & Management* (str. 31–44).

Hicks, D. (2004). The four literatures of social science. V *Handbook of quantitative science and technology research* (str. 473–496).

Huang, M. H., & Chang, Y. W. (2008). Characteristics of research output in social sciences and humanities: From a research evaluation perspective. *Journal of the American Society for Information Science and Technology*, *59*(11), 1819–1828.

Južnič, P. (2017). *Bibliometrijski indikatorji*. Pridobljeno s https://www.youtube.com/watch?v=l9W5glZl97I&feature=youtu.be

Kousha, K., Thelwall, M., & Rezaie, S. (2011). Assessing the citation impact of books: The role of Google Books, Google Scholar, and Scopus. *Journal of the American Society for information science and technology*, *62*(11), 2147–2164.

Lazarević, Ž., & Zemljič, I. (2003). *Slovenski zgodovinarski indeks citiranosti – izhodišča in pomisleki*. [Neobjavljena dokumentacija.]. Ljubljana: Inštitut za novejšo zgodovino.

MODS User Guidelines, Version 3 (Metadata Object Description Schema). Pridobljeno s https://www.loc.gov/standards/mods/userguide/introduction.html

Nederhof, A. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: A review. *Scientometrics*, *66*(1), 81–100.

Pajić, D. (2015). Globalization of the social sciences in Eastern Europe: genuine breakthrough or a slippery slope of the research evaluation practice? *Scientometrics*, *102*(3), 2131–2150.

Pančur, A. (2019a). *Preprosta raziskovalna infrastruktura za kompleksne raziskovalne podatke v humanistiki – si4 (Simple research Infrastructure FOR complex research data in digital humanities)*. [Neobjavljena dokumentacija.]

Pančur, A. (2019b). *Specifikacije za izvedbo naročila izdelave Zgodovinarskega indeksa citiranosti (ZIC)*. [Neobjavljena dokumentacija.]

Pančur, A., & Šorn, M. (2019). Na začetku je bil SIstory: raziskovalna infrastruktura slovenskega zgodovinopisja. V J. Hadalin in Ž. Lazarević (ur.), *Inštitut za novejšo zgodovino: 60 let mislimo preteklost* (str. 47–58). Ljubljana: Inštitut za novejšo zgodovino.

Pančur, A., Šorn, M., & Hadalin, J. (2014). Slovenski indeks citiranosti (SICI): Načrt izgradnje in delovanja. Tehnično poročilo. Pridobljeno s https://www.sistory.si/11686/36153

What is ElasticSearch. Pridobljeno s https://www.elastic.co/what-is/elasticsearch

THE HISTORIOGRAPHY CITATION INDEX UPGRADE

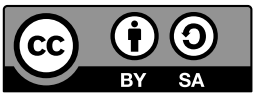
The fields of humanities and social sciences are often deprived of inclusion within the international citation indexes such as Scopus and Web of Science (WOS). The reason for this offshift in the indexes are commonly associated with the format of published works, e.g. the most common type of published works in humanities are monographs (though the scientific journals are on the rise), which are not typically included in WOS and Scopus. Even though Scopus is far more inclusive of such types and fields in comparison to WOS, there is still a gap to be filled. As a response to this predicament the Institute of Contemporary History developed its own citation index – the Historiography Citation Index (HCI), which was first meant to only track the research production within the institution, but has since been expanded to cover the production of the whole field of Slovene historiography. Over the years HCI was a subject of several upgrades and data harmonization attempts. Even with the upgrades, several shortcomings of the systems were apparent, and therefore, another upgrade was taken into consideration, and after the extensive analysis was performed, we identified the most problematic aspects of the index and began working on another upgrade.

The upgrade was performed in two parts – in the first one, we took upon ourselves to improve the administrative system in which we implemented the ElasticSearch technology to improve our search engine and filtration system, as well as improving the data masks to increase the precision and accuracy of the data input into the index. As a part of the administrative system upgrade we also modeled the MODS application profile to increase the interoperability of our data and therefore, enabling the exchange of our data between different information systems without losing data and its context. In the second part, we upgraded the user interface of the citation index to be more user friendly. In order to increase the coherence of the data display, we implemented a table-like design of the search result, equipped with filters in each column. To increase the visibility of the most important factor of the citation index, number of citations the work has received, we included additional column just for that information. The index aims to enable researchers access to the information on the number of citations, cited works ect. It is also recognised by the Slovenian Research Agency (ARRS) as a valid source of citations and could be used to provide proof

of the researchers achievements and scientific excellency, though it is still not recognised as equal to the SICRIS information system.

With the upgrade we increased the efficiency of the citation index, as well as its usability, and with it ensured a more intuitive system to its indexators and users.

Keywords: the Historiography Citation Index, HCI, upgrade, citation indexes



TRI SPLETNE APLIKACIJE O SLOVENSKIH NAREČJIH

Rok MRVIČ
Inštitut za slovensko narodopisje, ZRC SAZU

Špela ZUPANČIČ
Filozofska fakulteta, Univerza v Ljubljani

Mrvič, R., Zupančič, Š. (2021): Tri spletne aplikacije o slovenskih narečjih. Slovenščina 2.0, 9(1): 236–261.

DOI: https://doi.org/10.4312/slo2.0.2021.1.236-261

Potreba po večji prisotnosti narečnih vsebin na spletu in njihovi interaktivni multimedijski predstavitvi, predvsem strokovno zasnovanih dialektoloških virov in orodij, je spodbudila interdisciplinarno sodelovanje različnih fakultet Univerze v Ljubljani, zlasti Filozofske fakultete (FF) in Fakultete za računalništvo in informatiko (FRI), ki je v letih 2017 in 2018 obrodilo sadove v obliki treh prostodostopnih in odprtokodnih spletnih aplikacij o slovenskih narečjih – to so *Slovenski narečni atlas* (SNA, 2017), *Interaktivna karta slovenskih narečnih besedil* (IKNB, 2018) in *Slovar starega orodja v govoru Loškega Potoka* (SSOLP, 2018). Članek v prvem delu prinaša splošen pregled slovenskih spletnih dialektoloških virov in orodij, v drugem delu pa podrobnejšo predstavitev funkcionalnosti navedenih treh aplikacij, ki so uporabnikom trenutno na voljo. V diskusijskem delu pregleda je izpostavljen del okoliščin nastanka obravnavanih aplikacij in z nastankom povezanih omejitev, nakazane pa so tudi možne rešitve, ki bi jih veljalo preudariti za zagotovitev njihovega dolgoročnega razvoja.

Ključne besede: slovenska narečja, spletna aplikacija, narečni atlas, narečni slovar, interaktivna karta

1 UVOD

Z digitalizacijo in hitrim tehnološkim razvojem se je zlasti v zadnjem desetletju v slovenski dialektologiji pojavila potreba po prenosu jezikovnih orodij in priročnikov na splet. V sodobnih slovenskih narečnih govorih prihaja do velikih sprememb – tako v zemljepisnem prostoru kot v novih funkcijah oz. položajih rabe (Smole, 2019, str. 21) –, zanimanje zanje v sodobni slovenski družbi pa vse bolj narašča.[1] Spremembe so v zadnjih petih letih spodbudile razvoj več spletnih orodij, ki omogočajo strokovno in ciljno objavo narečnega gradiva, namenjenega zlasti jezikoslovcem in študentom, vendar poskušajo ob tem k uporabi pritegniti tudi širšo javnost. Med taka orodja uvrščava aplikacije *Slovenski narečni atlas* (SNA), *Interaktivna karta slovenskih narečnih besedil* (IKNB) in *Slovar starega orodja v govoru Loškega Potoka* (SSOLP), ki so nastale v interdisciplinarnem sodelovanju različnih fakultet Univerze v Ljubljani.[2] Vse tri spletne aplikacije so prostodostopne, odprtokodne,[3] interaktivne in rastoče. V drugem poglavju strneva splošen pregled slovenskih dialektoloških virov in orodij, v tretjem poglavju pa po kronološkem zaporedju od najstarejše (SNA, 2017) do najmlajše (SSOLP, 2018) nadaljujeva s predstavitvijo bistvenih informacij o aplikacijah SNA, IKNB in SSOLP,[4] in sicer z vidika funkcionalnosti, ki so uporabnikom trenutno na voljo.

1 Na večje zanimanje za narečja vpliva preplet več družbenih dejavnikov, povezanih z jezikovno identiteto narečnih govorcev, ki se je z digitalizacijo družbe začela jasno odražati v obliki diskusijskih skupin, forumov in predstavitvenih strani krajev, pokrajin in njihovih narečnih govorov na sodobnih družbenih omrežjih, kot sta Facebook in Instagram. O nezanemarljivem vplivu spletnih mest manifestacije narečne zavesti pričajo podatki o številu sledilcev oz. članov tovrstnih skupin in podatki o njihovi dejavnosti, na podlagi samoiniciativnih objav narečnega gradiva (domnevno narečno specifičnih frazemov, pregovorov, kletvic, pozdravov, vzklikov ipd.) pa so se začeli vzpostavljati tudi turistični projekti ter samostojne publikacije, ki so izšle z namenom predstavitve narečnih prvin širši javnosti.

2 Pri izdelavi IKNB in SNA sta sodelovali Filozofska fakulteta Univerze v Ljubljani (FF) in Fakulteta za računalništvo in informatiko Univerze v Ljubljani (FRI), pri izdelavi SSOLP pa še Naravoslovnotehniška fakulteta Univerze v Ljubljani (NTF).

3 Izvorne kode predstavljenih aplikacij so objavljene v repozitoriju Bitbucket.

4 Pri pripravi vsebine aplikacij so v okviru seminarjev in projektov ter zaključnih študijskih del sodelovali študenti Oddelka za slovenistiko Filozofske fakultete Univerze v Ljubljani. Zapis narečnega gradiva v navedenih aplikacijah je bil pripravljen z vnašalnim sistemom ZRCola (http://zrcola.zrc-sazu.si), ki ga je na Znanstvenoraziskovalnem centru SAZU v Ljubljani (http://www.zrc-sazu.si) razvil Peter Weiss.

2 SPLETNI DIALEKTOLOŠKI VIRI IN ORODJA

Dialektološki viri[5] so lahko 1) prvotno objavljeni v tiskani obliki in kasneje digitalizirani ter prilagojeni za objavo na spletu ali 2) izhodiščno digitalni, torej namensko razviti za spletno objavo. Med slednje spadajo tudi spletne aplikacije, o katerih v kontekstu tega besedila govoriva kot o specializiranih jezikoslovnih oz. dialektoloških orodjih, ki izkoriščajo različne možnosti digitalnega medija, s čimer uporabniku omogočajo interaktivno spoznavanje narečnega gradiva na več ravneh. V Sloveniji so dialektološke vsebine spletnih jezikovnih virov v veliki večini primerov rezultat strokovnega in znanstvenega preučevanja, izjemoma pa tudi ljubiteljskega zbiranja narečnega gradiva.[6] V nadaljevanju je predstavljen kratek pregled nekaterih slovenskih spletnih narečnih virov, ki so prosto dostopni in pri izdelavi katerih so sodelovali dialektologi, torej virov, ki naj bi uporabnikom nudili relevantne, strokovno pregledane vsebine.

2.1 Spletni dialektološki viri

Med temeljne slovenske digitalizirane dialektološke vire uvrščava pet narečnih slovarjev in *Slovenski lingvistični atlas* (SLA) – do vseh lahko dostopamo na spletnem portalu Fran.[7] Digitalizirane različice slovarjev so uporabnikom portala Fran na voljo predvsem v obliki faksimilov,[8] kar jih v primerjavi s

5 Pojem spletni dialektološki vir uporabljava kot krovni pojem za vse oblike virov dialektološko obdelanih jezikovnih podatkov (slovarjev, atlasov, korpusov, interaktivnih kart), ki so dostopni na spletu.

6 Pregled trenutno dostopnih spletnih virov pokaže, da na izbiro vrste končnega prikaza zbranega narečnega gradiva vpliva strokovno znanje zbirateljev takega gradiva. Z izdelavo narečnih slovarjev se npr. ukvarjajo tudi nejezikoslovci (rezultate njihovega dela, zlasti na spletu, zaradi manjkajočih leksikografskih podatkov pogosteje obravnavamo kot zbirke narečnih besed in jih zato v najin pregled ne uvrščava, prim. Benko, 2016, str. 127), medtem ko prikaz narečnega gradiva v atlasu ali na zemljevidu ostaja v domeni dialektologov. Zbiranje narečnega gradiva je v vsakem primeru (tudi če pri tem ne sodelujejo dialektologi) pomembno, saj lahko zbrano gradivo, še posebej posneto, predstavlja osnovo za nadaljnje dialektološke raziskave. Kot vir jezikoslovnih raziskav npr. lahko služi tudi gradivo, zbrano v okviru etnološkega in folklorističnega dela (gl. Ivančič Kutin, 2017, str. 65–69).

7 Na spletu najdemo tudi monografijo *Besedotvorni atlas slovenskih narečij: Kulturne rastline* (Kumin Horvat, 2018), ki je izšla tako v tiskani kot v digitalni obliki.

8 Prvotno tiskani slovarji pred objavo na spletu niso šli skozi proces optičnega prepoznavanja znakov (ang. *optical character recognition*, OCR), ki bi uporabnikom omogočalo lažje in bolj ciljno usmerjeno iskanje po gradivu.

spletnimi slovarji dela precej neprijazne za uporabo, saj uporabniku znotraj enega okna v brskalniku ni omogočen takojšnji vpogled v vsebino slovarskega sestavka – slednjo uporabnik najde na ločeno objavljenih slovarskih straneh v formatu PDF, ki so v Franovi spletni bazi povezane z vsemi slovarskimi gesli, ki jih vsebujejo. V tej obliki povezav gesel s slovarskimi stranmi so uporabnikom dostopni *Črnovrški dialekt* (Tominec, 2015),[9] obsežen, abecedno urejen slovar, ki prinaša leksiko črnovrškega narečja, *Slovar govorov Zadrečke doline med Gornjim Gradom in Nazarjami (A–H)* (Weiss, 2015)[10] ter *Slovar bovškega govora* (Ivančič Kutin, 2015).[11] Gesla *Slovarja govorov Zadrečke doline med Gornjim Gradom in Nazarjami (A–H)* ter *Slovarja bovškega govora* so tako kot v Tominčevem slovarju glasovno poknjižena, slovarski sestavki pa so v primerjavi s sestavki *Črnovrškega dialekta* podrobneje in bolj enotno strukturirani. Weiss je pri tem natančnejši kot Ivančič Kutin, vendar je lahko njegov slovar za splošnega uporabnika prav zato zahtevnejši.

Poleg narečnih slovarjev sta na spletišču Fran vključena tudi oba zvezka *Slovenskega lingvističnega atlasa* (2014, 2016)[12] – »atlas[a], ki obsega celoten slovenski jezikovni prostor in predstavlja temeljno delo slovenske dialektologije in geolingvistike« (Bon, 2018, str. 42). Uporabniška izkušnja iskanja po atlasu je delno podobna izkušnji iskanja po digitaliziranih slovarjih, saj lahko uporabnik do gradiva dostopa prek posameznih datotek PDF.[13] Razlika je v tem, da je gradivo za atlas šlo skozi proces optičnega prepoznavanja znakov, kar uporabniku olajša pregledovanje komentarjev, kart in gradiv h kartam za posamezne lekseme. Atlas je bolj kot splošnim uporabnikom namenjen

9 *Črnovrški dialekt: Kratka monografija in slovar* je ob izdaji leta 1964 predstavljal »napredek v slovenski dialektologiji, saj smo imeli pred njim le rokopisne zbirke« (Benko, 2016, str. 126).

10 *Slovar govorov Zadrečke doline med Gornjim Gradom in Nazarjami: Poskusni zvezek (A–H)* je bil v tiskani obliki objavljen leta 1998 in je predstavljal »prvi slovenski model za izdelavo znanstvenega sinhronega enonarečnega razlagalnega slovarja« (Benko, 2016, str. 126).

11 Slovar je v tiskani obliki izšel leta 2007.

12 *SLA 1: Človek (telo, bolezni, družina)* je v tiskani izdaji izšel leta 2011 in v digitalni leta 2014, *SLA 2: Kmetija* pa je v tiskani in digitalni obliki izšel leta 2016.

13 Uporabnik lahko dostopa do komentarjev o posameznih knjižnih leksemih z ustreznicami v različnih narečjih, do kart z grafičnim prikazom narečnih leksemov in do gradiv h kartam, ki vsebujejo podatke o tem, kakšni so narečni izrazi za določen knjižni leksem v posameznih raziskovalnih točkah.

strokovnjakom in študentom jezikoslovnih smeri; za ustrezno uporabo atlasa je namreč potrebno osnovno dialektološko predznanje, da lahko uporabnik s kart in iz priloženih komentarjev pridobi iskane podatke.

Na spletišču Fran se nahajata tudi *Kostelski slovar* (Gregorič, 2015)[14] in *Slovar oblačilnega izrazja ziljskega govora v Kanalski dolini* (Kenda-Jež, 2019).[15] *Kostelski slovar* je bil prvotno izdan v tiskani obliki, njegova postavitev na splet pa se bistveno razlikuje od spletnega prikaza že omenjenih digitaliziranih slovarjev. Uporabnik namreč do slovarskih sestavkov v brskalniku dostopa neposredno na spletnem mestu slovarja – tj. s klikom na izbrano geslo –, ne pa več prek dokumenta digitalizirane strani iz knjige, ki vsebuje določeno geslo. Spletni slovarski sestavek v *Kostelskem slovarju* je oblikovan pregledno in je po leksikografski zasnovi slovarske mikrostrukture podoben slovarskim sestavkom, ki jih prinaša slovar Barbare Ivančič Kutin. Korak naprej v izrabi možnosti, ki jih ponuja spletni medij, predstavlja spletni slovar Karmen Kenda-Jež. Glavna odlika slovarja je, da ob narečnih zapisih vsebuje zvočne posnetke narečnega gradiva – na ta način slovar po besedah avtorice funkcionira kot »govoreči slovar« (Kenda-Jež, 2019, str. 2), kar je bil del načrtovane slovarske strukture že od samega začetka. Slovarski sestavki, zasnovani natančneje kot v *Kostelskem slovarju*, imajo pregledno podobo, ki vključuje slovnične podatke in pregibalne vzorce uslovarjenih leksemov (glede na zbrano narečno gradivo) ter pojasnila v pojavnih okencih ob kazalcu miške,[16] ki ne zahtevajo tako podrobnih legend znakov in uvodnih pojasnil krajšav in ikon. To splošnemu uporabniku omogoča enostavnejšo uporabo slovarja v primerjavi s sestavki slovarjev Weissa in Ivančič Kutin. Za razliko od ostalih omenjenih slovarjev, ki so zasnovani kot splošni narečni slovarji, je slovar Karmen Kenda-Jež tematski narečni slovar.

Zvočni posnetki so vključeni tudi v slovar *Narečna bera*, ki se nahaja na

14 Avtor slovarja je Jože Gregorič, njegovo gradivo pa so urejali Sonja Horvat, Ivanka Šircelj-Žnidaršič in Peter Weiss. Slovar je bil v tiskani obliki objavljen leta 2014.

15 Slovar, ki je bil razvit za objavo v spletni obliki, temelji na knjižnih izdajah monografije *Shranli smo jih v bančah: Slovarski prispevek k poznavanju oblačilne kulture v Kanalski dolini – Contributo lessicale alla conoscenza dell'abbigliamento in Val Canale* (Kenda-Jež, ¹2007, ²2015).

16 Npr. metapodatki o terenskem delu, kot so začetnice imena in priimka informatorjev ter letnice njihovega rojstva.

samostojnem spletnem mestu, (Benko, 2013),[17] vendar za razliko od slovarja oblačilnega izrazja ne pri vseh slovarskih sestavkih, imajo pa posamezni sestavki dodane celo videoposnetke. Iskanje po slovarju je od zgoraj navedenih slovarjev najenostavnejše in najpreglednejše.[18] *Narečna bera* prinaša veliko narečnega ponazarjalnega gradiva, pri nekaterih geslih tudi zvočne posnetke, videoposnetke in slikovno gradivo, vendar zgradba slovarskih sestavkov v marsičem sledi zgradbi slovarskih sestavkov tiskanih slovarjev.[19] V nekaterih slovarskih sestavkih so sicer izkoriščene določene medleksemske povezave, poleg tega pa pojasnila v pojavnih okencih ob kazalcu miške,[20] pripomorejo k temu, da je slovar dovolj razumljiv in informativen tudi za splošnega uporabnika.

Prvi slovenski dialektološki korpus, *Govorni korpus Koprive na Krasu – GOKO* (Šumenjak, 2013),[21] je bil najverjetneje tudi prvi spletni dialektološki vir, ki je vključeval zvočne posnetke narečnega govora. Korpus vključuje okoli 60 minut posnetega gradiva (Šumenjak, 2013, str. 35), razdeljenega na krajše posnetke; govorjeno besedilo je prikazano v fonetični in poenostavljeni transkripciji ter poknjiženem zapisu. Ob objavi je korpus predstavljal sodoben in svež pristop k predstavitvi narečnega gradiva širši javnosti, vendar se z vidika

17 Slovar vključuje leksiko s področja kmetijstva, ki je bila zbrana v štirih krajevnih govorih koroškega podjunskega narečja, in je ob objavi predstavljal »[p]rvi model za izdelavo strokovnega jezikovnega (slikovnega) narečnega slovarja« (Benko, 2016, str. 135).

18 Uporabniku se ob izbiri določene črke prikažejo vsa gesla, ki se z njo začnejo, s čimer pridobi boljši vpogled v nabor uslovarjene leksike. Na portalu Fran lahko lekseme v izbranem slovarju iščemo le s klikanjem skozi slovarske strani ali pa s pomočjo iskalne vrstice (to je v primeru iskanja po narečnem slovarju velikokrat nepraktično, saj kljub predlogom v spustnem meniju iskalne vrstice ne moremo vnaprej vedeti, kateri leksemi so vključeni v slovar).

19 Za glasovno poknjiženim geslom in narečno ustreznico so linearno nanizani slovnični razdelek, razdelek s krajevnimi označevalniki, pomen in morebitne sopomenke, spodaj pa še narečno ponazarjalno gradivo in etimološki razdelek, kar pri krajših sestavkih pušča precej neizkoriščenega prostora v oknu brskalnika. Številna grafična znamenja, prvotno namenjena racionalizaciji prostora v tiskanih slovarjih, so s sodobnimi spletnimi oblikovalskimi rešitvami v veliki meri postala odveč. V tem oziru slovar oblačilnega izrazja bolje izkorišča potencial medija, v katerem je objavljen.

20 Npr. poimenovanja posameznih razdelkov slovarja, poimenovanja raziskovalnih točk.

21 Benko (2016, str. 127) je korpus GOKO zabeležila kot enega od trinajstih slovenskih spletnih narečnih slovarjev (dialektoloških in ljubiteljskih). Med naštetimi so trije delo dialektologov in še vedno delujejo: korpus GOKO, *Narečna bera* in *Mali bisidnik za tö jošt rozajanskë pïsanjë* (zdaj *Resianica*; Steenwijk).

sodobnih jezikovnih tehnologij že kažejo številne možnosti za izboljšave, npr. izvedba iskanja in določanja iskalnih pogojev.[22] Podobno velja tudi za *Govorni korpus Ospa – GOSP* (Šumenjak, 2013), ki je bil pripravljen po zgledu korpusa GOKO.

2.2 Spletne aplikacije

Spletne narečne aplikacije med vsemi spletnimi dialektološkimi jezikovnimi viri najbolje izkoriščajo možnosti, ki jih nudi digitalno okolje (npr. združevanje jezikovnih podatkov s kartografskimi, vnos povezav na druge jezikovne vire, vzpostavljanje medleksemskih povezav med narečnim gradivom, dodajanje slikovnega, zvočnega in video ponazarjalnega gradiva, urejanje uporabniških vlog in odnosov med njimi), hkrati pa se od ostalih spletnih virov razlikujejo po tem, da lahko uporabnik kot skrbnik sam ustvarja in oblikuje nove vsebine, torej za razliko od virov v podpoglavju 2.1 govorimo ne le o virih, temveč tudi o orodjih.[23] Zaenkrat v slovenskem prostoru obstajajo tri tovrstne aplikacije – narečni atlas, interaktivna karta narečnih besedil in narečni slovar –, ki so predstavljene v nadaljevanju.[24]

22 Glavna pomanjkljivost je ta, da splošni uporabnik vnaprej ne ve, katere besede so vključene v korpus, zato lahko do gradiva pride le z naključnim vpisovanjem besed v iskalno vrstico, če pogoji in načini iskanja niso jasno opredeljeni. Ker je korpus nastal v pilotni raziskavi (Šumenjak, 2013, str. 35) in gre za prvi tovrstni prikaz narečnega gradiva na Slovenskem, so njegove tehnične omejitve razumljive.

23 S podrobnim razmejevanjem med vrstami aplikacij se nisva ukvarjala. *Narečna bera* denimo temelji na sistemu za upravljanje vsebin Joomla, kar jo uvršča med spletne aplikacije, vendar je med *Narečno bero* in v nadaljevanju predstavljenim SSOLP mogoče opaziti veliko razliko v zasnovi, ki se odraža zlasti v funkcionalnostih aplikacije. SSOLP ima namreč tudi skrbniški del vmesnika, ki uporabniku omogoča enostaven vnos novih vsebin.

24 V nastajanju so še tri spletne aplikacije – interaktivni *Slovenski lingvistični atlas* (Škofic in Vičič, 2013), *Frazeograf* (Mrvič in Žnidaršič, 2020) in *Narečni frazem* (Mezgec idr., b. l.). Interaktivni *Slovenski lingvistični atlas* (*e-SLA*) je spletna različica *Slovenskega lingvističnega atlasa*. Temeljil bo »na medsebojni povezanosti različnih podatkovnih zbirk« (Škofic, 2013, str. 98) – uporabnik bo namreč lahko prek »jezikovne karte dostopal do digitaliziranega arhivskega gradiva, zvočnih in video posnetkov v podatkovni zbirki ter do drugih spletnih povezav na bibliografske podatke o raziskavah krajevnega govora […] ter na podatke o krajih – točkah iz raziskovalne mreže jezikovnega atlasa« (prav tam, str. 96). Zaenkrat sta oblikovani interaktivni karti za besedi *kmetija* in *hiša* (Bon, 2018, str. 49). *Frazeograf*, ki bo na voljo od konca letošnjega leta dalje, je prostodostopna, odprtokodna, interaktivna in rastoča aplikacija za ustvarjanje in urejanje frazeološkega gradiva (Mrvič, 2020). V njej je bil leta 2020 v okviru magistrskega

3 APLIKACIJE SNA, IKNB IN SSOLP

3.1 Slovenski narečni atlas (SNA)

SNA je v sklopu interaktivnih aplikacij, ki omogočajo vnos in organizacijo podatkov na podlagi jezikovnih kart, v svojem magistrskem delu podrobneje predstavila Mija Bon (2018),[25] in sicer skupaj z interaktivnim *Slovenskim lingvističnim atlasom*, ki je spletna različica *Slovenskega lingvističnega atlasa* (SLA), in IKNB (gl. podpoglavje 3.2). SNA je nastajal in se razvijal pod mentorstvom Alenke Kavčič (FRI); aplikacijo je leta 2017 v okviru diplomskega dela izdelal Gregor Šajn, leto pozneje pa jo je nadgradil Nermin Jukan.[26] V osnovi SNA izpolnjuje temeljne pogoje, ki jih za predstavitev prostorske razširjenosti (na zemljevidih) posameznih jezikovnih pojavov potrebuje geolingvistika. Splet je jezikovnim virom omogočil dodatne funkcije, ki pripomorejo k natančnejšemu raziskovanju in večji informativnosti, zaradi česar se je tudi Bon odločila za vnos narečnega frazeološkega gradiva (primerjalnih frazemov s pomenom človeške lastnosti) v SNA, skupaj s komentarji in povezavami na druge vire (Bon, 2018, str. 28, 48).

3.1.1 Kaj aplikacija omogoča uporabnikom

sNA je spletno orodje, ustvarjeno za kartiranje narečne leksike iz različnih tematskih polj (trenutno zapolnjeno polje so primerjalni frazemi, deloma poimenovanja delov stare kmečke hiše in sadja, temi nekonvencionalnih replik in posode pa sta le nakazani). Osrednji element aplikacije je enaka narečna karta kot pri IKNB z jasno prikazanimi narečnimi skupinami, narečji in podnarečji. SNA je torej namenjen spoznavanju slovenske narečne leksike, vendar je njegov vmesnik za razliko od IKNB (podpoglavje 3.2) in SSOLP (podpoglavje 3.3) za uporabnika nejezikoslovca precej zahtevnejši in manj intuitiven. Od

dela ustvarjen poskusni narečni frazeološki slovar, ki ga je pod mentorstvom Vere Smole (FF) izdelal Rok Mrvič. *Narečni frazem* je pilotna spletna aplikacija (Vičič in Marc Bratina, 2015, str. 814); funkcionalnosti aplikacije so torej zaenkrat v celoti na voljo le sodelavcem projekta, ki so hkrati registrirani uporabniki. Izdelana aplikacija bo prostodostopna in bo »skupnosti pomagala pri zbiranju narečnih frazemov« (prav tam, str. 817), na ta način pa bi se sčasoma lahko oblikoval vsesloveski narečni frazeološki e-slovar (prav tam, str. 812).

25 Avtorica je magistrsko delo pripravila pod mentorstvom Vere Smole (FF).

26 Podrobni podatki o Šajnovem diplomskem delu se nahajajo na seznamu literature. Jukan je aplikacijo nadgradil pri predmetu Računalništvo v praksi II.

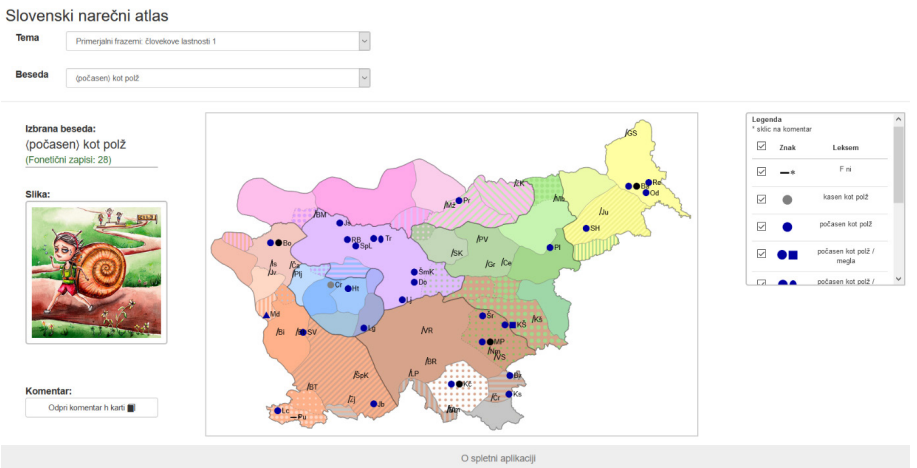
njega namreč pričakuje temeljno geolingvistično znanje o uporabi jezikovnih kart, ki omogoča branje podatkov s karte in priložene legende, ter poznavanje fonetične transkripcije, v kateri je zapisano vse narečno gradivo. Za večjo jasnost izhodiščne narečne karte bi morala aplikacija ponujati možnost dodatne legende ali preglednega seznama na karti obarvanih območij, ki predstavljajo narečne skupine, narečja in podnarečja. Do teh lahko uporabnik trenutno dostopa le s klikom na izpisane raziskovalne točke. Zaradi naštetega je aplikacija najbolj zanimiva za študente in jezikoslovce.

Uporabniku se ob izbiri tematskega polja v prvem spustnem meniju in leksema, navedenega v drugem spustnem meniju (Slika 1), prikažejo podatki o prostorski razširjenosti izbranega leksema, saj se izpišeta število in geografski položaj raziskovalnih točk, kjer je bil leksem zabeležen (zvočni zapis in fonetična transkripcija), desno od narečne karte pa je za interpretacijo rezultatov dodana tudi legenda znakov, uporabljenih za diferenciacijo gradiva. Legenda vsebuje glasovno poknjiženi[27] leksem s pripadajočim simbolom, ki se na karti pojavlja skupaj s kratico kraja. Simbolom je za jasno diferenciacijo narečnega gradiva na karti mogoče spreminjati obliko in barvo. Bon kot eno izmed pomembnih prednosti SNA navaja zlasti možnost oblikovanja lastnega nabora na karti prikazanih leksemov, ki ga ustvarimo z obkljukanjem želenih leksemov v legendi (2018, str. 54). To uporabniku omogoča ciljno brskanje in organizacijo narečnega gradiva z uporabo besedotvornih in/ali morfoloških kriterijev.

Ob kliku na izpisano raziskovalno točko, označeno s simbolom in kratico kraja, se odpre novo okno, ki vsebuje podatke o kraju, narečno umestitev govora ter glasovno poknjiženi leksem v slovarski obliki, ki mu je dodana fonetična transkripcija, na priloženem vtičniku pa lahko uporabnik posluša zvočni zapis leksema – slednji je lahko naveden samostojno ali znotraj daljšega besedilnega zgleda.

Poleg navedenih možnosti je h karti mogoče priložiti PDF datoteko komentarja in morebitno slikovno gradivo (fotografijo ali ilustracijo). Obe možnosti znatno povečata informativnost izpisa in uporabniku ponudita celovitejšo informacijo o iskanem narečnem gradivu.

27 Pod pojmom glasovna poknjižitev »je mišljen prenos glasovnega sistema narečnega govora v knjižnega, na vseh drugih jezikovnih ravninah pa je ohranjen narečni sistem« (Smole, 2019, str. 25).



Slika 1: Geolingvistični prikaz primerjalnega frazema (*počasen*) *kot polž*.

3.2 Interaktivna karta slovenskih narečnih besedil (IKNB)

Aplikacijo je leta 2018 v okviru diplomskega dela izdelal Ivan Lovrić, študent FRI,[28] še istega leta pa jo je nadgradil Nermin Jukan,[29] prav tako študent FRI. Vsebina aplikacije je nastala na podlagi brošure *Stara kmečka hiša: Narečna besedila z analizo I* (Smole in Horvat, 2016). IKNB tako vsebuje narečna besedila (posnetke, fonetične in poknjižene prepise ter analize govorov) na temo stare kmečke hiše[30] (prostori in oprema v njej). Večino gradiva so zbrali in pripravili študenti Oddelka za slovenistiko FF,[31] posnetke krajevnih govorov, ki so jih prispevali nekdanji študenti UL, pa je

28 Lovrić je aplikacijo izdelal pod mentorstvom Alenke Kavčič (FRI) in somentorstvom Vere Smole (FF). Podrobni podatki o diplomskem delu se nahajajo na seznamu literature.

29 Jukan je izdelano aplikacijo nadgradil z dodatnimi funkcionalnostmi pri predmetu Računalništvo v praksi I, in sicer pod mentorstvom Alenke Kavčič.

30 Izhodiščno besedilo v knjižnem jeziku *Stare kmečke hiše* je po delu vprašalnice za SLA, ki jo je sestavil Fran Ramovš, pripravila Vera Smole. Dostopno je na spletni strani IKNB, pod zavihkom O aplikaciji.

31 Zvočne posnetke, fonetične transkripcije in poknjižitve besedil so pripravili študenti, ki so do leta 2018 obiskovali izbirni predmet Slovenska narečja pod vodstvom Vere Smole in Mojce Kumin Horvat (ZRC SAZU, ISJFR), analize pa študenti seminarja pri predmetu Slovenska dialektologija in izbirnega predmeta Poglavja iz zgodovine slovenskega glasoslovja pod vodstvom Vere Smole.

transkribirala in poknjižila Vera Smole. V rastočo[32] aplikacijo je trenutno vključenih sto krajevnih govorov.

Osnovo aplikacije predstavlja *Karta slovenskih narečij*,[33] na kateri so z različnimi barvami in vzorci predstavljene vse narečne skupine, narečja in podnarečja, ter pripadajoča legenda. Karta je zgrajena na odprtokodni Javascriptovi knjižnici Leaflet in prostodostopnih zemljevidih OpenStreetMap (Kavčič idr., 2018, str. 122), njena uporaba pa je enostavna – povečuje in pomanjšuje se skupaj z zemljevidom. Na karti so z ikonami in kraticami označeni kraji,[34] katerih govori so vključeni v aplikacijo.

3.2.1 Kaj aplikacija omogoča uporabnikom

IKNB je spletno orodje za spoznavanje slovenskih narečij, namenjeno tako dialektologom kot širši javnosti. Zasnovano je tako, da uporabnikom omogoča jasen pregled nad celotnim sistemom razdelitve slovenskih narečij in natančnejšo predstavitev posameznih krajevnih govorov na več ravneh; ker so govori predstavljeni na enak način, jih ni težko primerjati med sabo.

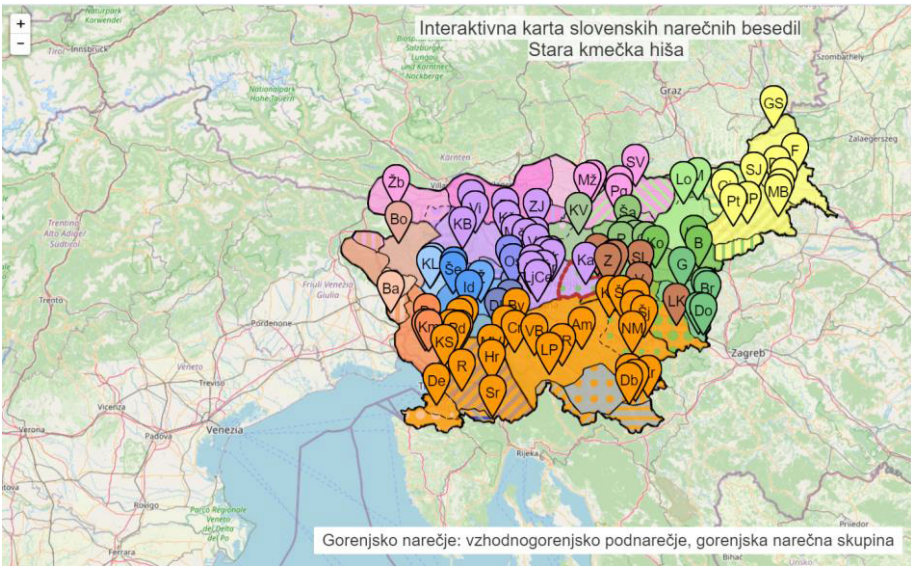
Vsaka narečna skupina je na zemljevidu označena s svojo barvo, vsako narečje in podnarečje pa vsebuje dodatne grafične simbole (pike ali poševne črte), kar uporabnikom omogoča, da spoznavajo, kako se na določenem območju narečja in podnarečja prepletajo med sabo in vplivajo drug na drugega (Kavčič idr., 2018, str. 122). Uporabniki lahko razdelitev slovenskih govorov na narečne skupine, narečja in podnarečja spoznajo in usvojijo na dva različna načina.

32 Ustvarjalci aplikacije poleg dodajanja novih krajevnih govorov pod temo Stara kmečka hiša načrtujejo razširitve z dodajanjem novih besedil in novimi temami. Vključiti želijo dve basni (*Čebela in Čmrlj* ter *Mravlji*), ki sta krajši in manj zahtevni besedili kot besedilo o stari kmečki hiši, s čimer bi se mladim najverjetneje bolj približali in jih pritegnili k uporabi aplikacije. Za razliko od besedil o kmečki hiši besedila basni ne bi vključevala diahrone analize, ampak sinhrono primerjavo s knjižnim jezikom (Smole, 2019, str. 25–27).

33 Karta je nastala na podlagi Dialektološke karte slovenskega jezika Frana Ramovša (1931), novejših raziskav in gradiva Inštituta za slovenski jezik ZRC SAZU. Priredili so jo Tine Logar in Jakob Rigler (1983), Vera Smole in Jožica Škofic (2011) ter sodelavci Dialektološke sekcije ISJFR ZRC SAZU (2016).

34 Postavitev ikone, ki predstavlja posamezni kraj, je »določena z geografskimi koordinatami (geografsko dolžino in širino) kraja« (Kavčič idr., 2018, str. 123).

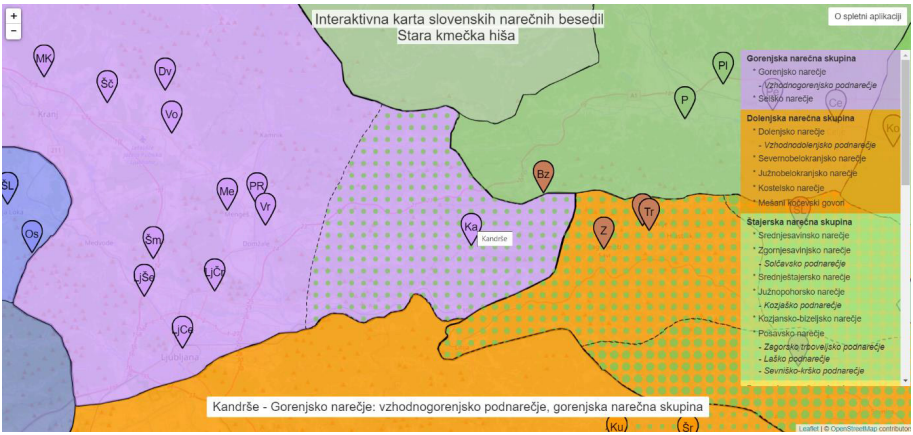
1. Ko se s kazalcem miške premaknejo na neko narečno območje, se to obrobi z odebeljeno rdečo črto, v spodnjem delu zaslona pa se prikaže bel okvirček s poimenovanjem narečja in narečne skupine, pa tudi podnarečja, če se nahajajo na območju podnarečja (Slika 2). Če se s kazalcem postavijo na ikono s kratico nekega kraja, se zgodi podobno; v spodnjem delu zaslona se prikaže bel okvirček s podatki o narečju, podnarečju in narečni skupini, spredaj pa je dodano še ime kraja.



Slika 2: Naslovna stran IKNB s kazalcem miške na območju vzhodnogorenjskega podnarečja.

2. Uporabniki si pri spoznavanju delitve slovenskih narečij lahko pomagajo tudi z legendo. Ko se s kazalcem miške postavijo na določen zapis (narečna skupina, narečje, podnarečje) v legendi, se ta zapis obarva rdeče, pripadajoče območje na karti pa se obrobi z odebeljeno rdečo črto.

Aplikacija omogoča poljubno premikanje po zemljevidu in približevanje, kar je še posebej praktično, kadar je na manjšem območju označenih več krajev (Slika 3); s približevanjem se tako lahko posamezne posnetke lažje loči med seboj (Kavčič idr., 2018, str. 122).



Slika 3: Povečava naslovne strani IKNB s kazalcem miške na ikoni kraja Kandrše.

Slika 4: Primer pojavnega okna ob kandrškem krajevnem govoru.

Obiskovalci spletne strani lahko krajevne govore, vključene v aplikacijo, spoznajo na več ravneh. Ko kliknejo na določeno ikono s kratico kraja, se odpre pojavno okno (Slika 4). To v zgornjem delu vsebuje poimenovanje kraja in kratico, uporabljeno na karti, ter podatke o narečju, podnarečju in narečni skupini. Sledijo podrobnejši podatki o kraju (pod katero pošto in v katero občino spada) ter podatki o avtorju posnetka, zapisovalcu in letu zapisa. Osrednji del v vsakem pojavnem oknu predstavljajo zvočni posnetek[35] narečnega bese-

35 Zaradi boljše uporabniške izkušnje je valovanje zvoka prikazano grafično, omogočeno pa je tudi premikanje nazaj in naprej po posnetku (Kavčič idr., 2018, str. 124).

dila ter njegova fonetična transkripcija in glasovna poknjižitev, ki se nahajata pod posnetkom. Pri nekaterih krajih je dodana tudi diahrona analiza govora (Smole, 2019, str. 25) z vidika značilnosti na sedmih jezikovnih ravninah (naglas, dolgi samoglasniki, kratki naglašeni in kratki nenaglašeni samoglasniki, soglasniki, oblikoslovni pojavi, leksika) (Kavčič idr., 2018, str. 123–124).

3.3 Slovar starega orodja v govoru Loškega Potoka (SSOLP)

Slovar je nastal leta 2018 v okviru štirimesečnega Študentskega inovativnega projekta za družbeno korist (ŠIPK), pri katerem so kot osrednje partnerske organizacije sodelovale FF, FRI, NTF in OŠ dr. Antona Debeljaka Loški Potok.[36] Študenta FRI pod mentorstvom Alenke Kavčič sta izdelala aplikacijo, študentki NTF pod mentorstvom Helene Gabrijelčič Tomc sta posneli in uredili video in foto gradivo ter ustvarili celostno grafično podobo aplikacije,[37] študenti FF pod mentorstvom Vere Smole pa so poskrbeli za vsebino slovarja (na terenu so posneli gradivo in ga uredili ter zasnovali in izdelali slovarska gesla).

SSOLP je narečni[38] tematski slovar. V okviru širše teme *staro orodje* vključuje podtemi *orodje za sekača in tesača* ter *orodje za spravilo sena*, na ta način pa so v slovarju zbrani izrazi za orodja in pripomočke tistih opravil, ki so v Loškem Potoku najbolj prisotna.[39] Poleg izrazov za orodja so v slovar vključena tudi poimenovanja za sestavne dele orodja in sopojavnice, to so »besede, ki se najpogosteje pojavljajo v sobesedilu« (Kenda-Jež v Smole idr., 2020, str. 1043); v konkretnem primeru so bili to glagoli z istim korenom, kot jih imajo

36 Podrobnosti o projektu so predstavljene na spletni strani SSOLP, pod zavihkom O projektu.

37 Oblikovalski vidik slovarja in izvedbeni vidik slovarja vključno z zasnovo in zgradbo slovarja ter skrbniškim in nadskrbniškim delom aplikacije sta predstavljena v Smole idr. (2020). Skrbniški del aplikacije omogoča dodajanje novih ali urejanje že naloženih vsebin, nadskrbniški del pa dodajanje novih ali urejanje že registriranih skrbnikov aplikacije – nadskrbnik torej ne more posegati v vsebine aplikacije.

38 Slovar vsebuje besedje krajevnega govora Loškega Potoka. V občini Loški Potok se govorita dve narečji: v severnem delu občine z osrednjim Hribom in okoliškimi vasmi Mali Log, Retje, Šegova vas in Travnik se govori krajevni govor tonemskega dolenjskega narečja, v južnem delu pa netonemsko kostelsko narečje. V raziskavo je bil vključen le tonemski govor Loškega Potoka, ki v SLA še ni zajet, njegove osnovne značilnosti pa so že predstavljene (gl. Smole idr. 2020, str. 1041–1042).

39 V okviru projekta je bilo zbranega ogromno narečnega gradiva, vendar sta podtemi zaenkrat zapolnjeni le delno, toliko, kot je bilo možno v omejenem času trajanja projekta.

orodja (npr. *kosa – kositi*), in samostalniki za izvajalce (npr. *kosec*). Slovenisti so narečno gradivo zbrali[40] s prostimi pogovori, pomagali pa so si tudi z usmerjevalnimi vprašalnicami Francke Benedik in Vere Smole ter z orodji in pripomočki informatorjev. Rastoči slovar omogoča dopolnjevanje obeh obstoječih tem in dodajaje novih.[41]

3.3.1 Kaj aplikacija omogoča uporabnikom

SSOLP je uporaben za vse, ki si želijo izvedeti več o lokalni snovni (starejša orodja in vsakdanji pripomočki) in nesnovni (narečni govor) kulturni dediščini. Uporabnikom se za iskanje po slovarju vanj ni treba prijaviti, po slovarju pa lahko brskajo na več načinov. 1) Želeni leksem lahko vpišejo v iskalno vrstico (leksem je mogoče tudi izbrati iz spustnega menija) ali pa 2) najprej v zavihku Stara orodja izberejo podtemo, nato pa iz nabora izpisanih gesel s klikom na določeno geslo odprejo slovarski sestavek. 3) Ko uporabniki kliknejo na izbrano geslo in se jim odpre slovarski sestavek, lahko prosto prehajajo med ostalimi gesli in njihovimi slovarskimi sestavki, saj so med njimi vzpostavljene medleksemske povezave.

Uporabniki lahko uslovarjene lekseme spoznajo z več vidikov. Naslovne strani slovarskih sestavkov so grafično razdeljene na dva dela (gl. primer gesla *plenkača* na Sliki 5):

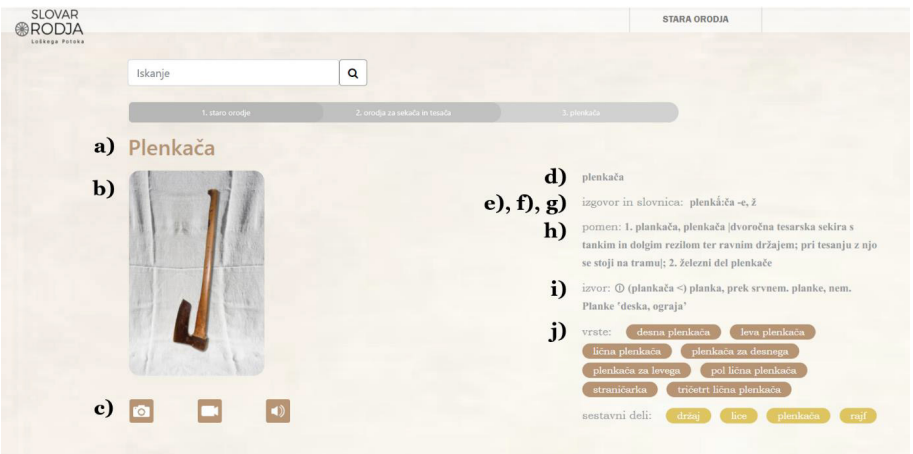
1. Na levi je najprej naslov – geslo v glasovno poknjiženem zapisu (a) s fotografijo (b), spodaj pa so ikone, ki ob kliku odprejo prikaz fotografij, videoposnetkov in zvočnih posnetkov (c). Z izjemo gesla je leva stran slovarskega sestavka v celoti namenjena ponazarjalnemu gradivu.

2. Na desni strani se nahaja besedilni, tj. jezikoslovni del slovarskega sestavka, ki pri vseh geslih vsebuje naslednje podatke: geslo v glasovno poknjiženem zapisu (d), slovarsko obliko narečnega leksema (e) z rodilniško končnico (f) (oboje je zapisano v fonetični transkripciji) in

40 Raziskovalci so obiskali 24 informatorjev. V slovar je zaenkrat vključeno le gradivo, pridobljeno pri Jožetu Anzeljcu (p. d. Štalarjevem stricu) iz Malega Loga (Smole idr., 2020, str. 1040).

41 Iz tematskega slovarja bo postopoma mogoče razviti splošni narečni slovar s prikazom slovarskih sestavkov po abecednem zaporedju. Nadgradnja aplikacije je v teku, namenjena pa bo Loškopotoškemu slovarju oz. Slovarju govora Loškega Potoka).

besednovrstno oznako (slovnični spol samostalnika)[42] (g) ter pomen leksema (h). Gre za t. i. obvezne razdelke, ki so ob vseh geslih zapolnjeni. Poleg obstojskega (d), izgovarjalnega (e), slovničnega (f), besednovrstnega (g) in pomenskega razdelka (h), ki se nahajajo v desnem, besedilnem delu slovarskega sestavka, je za vzpostavitev slovarskega sestavka obvezen tudi ponazarjalni razdelek (b, c) (fotografija ter zvočni posnetek in/ali videoposnetek). Nekaterim geslom so dodani podatki o izvoru leksema (i) ter medleksemske povezave (j), ki uporabniku ponudijo podatke o morebitni sopomenki, o tem, ali ima izhodiščno orodje več sestavnih delov, ali obstaja več vrst tega orodja, s katerimi orodji ga lahko vzdržujemo, ali ima geslo nadpomenko in kateri leksemi z istim korenom so vključeni v slovar. Vsi ti podatki so del t. i. neobveznih razdelkov, ki so ob nekaterih geslih zapolnjeni, ob nekaterih pa ne. Neobvezni so torej etimološki (i) in sopomenski razdelek ter pet povezovalnih razdelkov (j): vrste (orodja), nadpomenke, sestavni deli (orodja), orodja (za vzdrževanje) in besedje z istim korenom (Smole idr., 2020, str. 1044–1046).

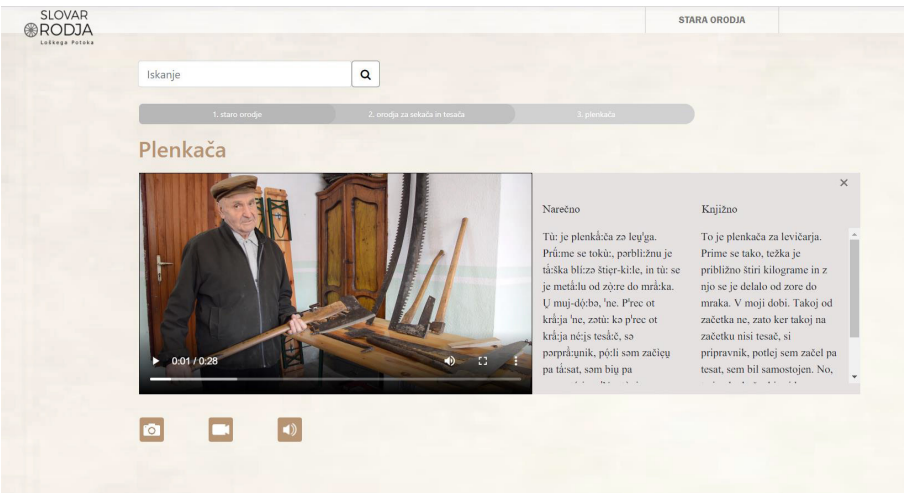


Slika 5: Prikaz naslovne strani slovarskega sestavka za geslo *plenkača*.

42 V SSOLP je tako kot v *Narečni beri* pri samostalniških geslih izpisana rodilniška končnica in dodana oznaka za besedno vrsto oz. spol samostalnika; v obeh slovarjih je izvedba slovnično-besednovrstnih razdelkov podobna tudi pri sestavkih pridevniških in glagolskih gesel. Poleg tega etimološki razdelek v SSOLP, ki je v izpisu poenostavljeno poimenovan izvor, uvaja simbol ⓘ, ki ga uporablja tudi Benko. Oboje je odraz zgledovanja po oblikovanju slovarske mikrostrukture v prvotno tiskanih slovarjih.

Leksemi, ki so vključeni v sopomenski razdelek (v slovarju: sopomenke) in v povezovalne razdelke (v slovarju: vrste, nadpomenke, sestavni deli, orodja in besedje z istim korenom), so zapisani v barvnih okvirčkih. Gre za medleksemske povezave, s katerimi je uporabnikom slovarja omogočeno raziskovanje besedja loškopotoškega govora v več smereh; ob kliku na določen okvirček (npr. *desna plenkača*) uporabnik pride do novega slovarskega sestavka, ki pripada geslu, na katerega je kliknil (v tem primeru geslu *desna plenkača*). Z medleksemskimi povezavami so ustvarjalci slovarja želeli kar najbolje izkoristiti potencial elektronskega medija in na ta način uporabnikom omogočiti čim bolj dinamično in nelinearno raziskovanje po slovarju (Smole idr., 2020, str. 1045).

Ko uporabnik klikne na ikono kamere ali zvočnika, se mu odpre druga stran slovarskega sestavka, ki vsebuje videoposnetek ali zvočni posnetek. Ob njem se v stolpcu *Narečno* prikaže zapis govorjenega besedila v fonetični transkripciji, v stolpcu *Knjižno* pa prevod v knjižni jezik.[43] Na Sliki 6 je prikazan pogled uporabnika, ko ob geslu *plenkača* klikne na ikono kamere.



Slika 6: Prikaz videoposnetka in zapisa govorjenega besedila ob slovarskem geslu *plenkača*.

43 Narečni sistem je v knjižnega prenesen na vseh jezikovnih ravninah (Smole idr., 2020, str. 1045).

4 IZHODIŠČA ZA NADALJNJI RAZVOJ

Aplikacije, ki jih predstaviva, omogočajo spoznavanje narečij na več ravneh: ne le pisno, kot so to doslej omogočala tiskana dialektološka dela, temveč interaktivno s ponazarjalnim slikovnim gradivom, prilagodljivimi kartografskimi podatki ter narečnimi avdio- in videoposnetki, ki uporabnikom omogočajo dostop do podatkov o slušnem vtisu. Kljub številnim prednostim, ki jih prinašata uporaba in razvoj predstavljenih aplikacij, želiva izpostaviti, da bi bilo vsako izmed njih mogoče še izboljšati, ob čemer se zavedava, da so sredstva za to omejena. Tako bi bilo v SNA dobro dopolniti obstoječa polja z novimi leksemi, posodobiti temeljno karto, dodati nove raziskovalne točke, narečno posneto gradivo vključiti ob več leksemih in poleg fonetične transkripcije dodati še poenostavljeni zapis za uporabnike brez dialektološkega predznanja, za uporabniku prijaznejšo izkušnjo pa bi bila smiselna tudi nadgradnja uporabniškega vmesnika, kot je bila letos izvedena za SSOLP.[44] V IKNB bi bilo smiselno vključiti več tem, primernih za vnose krajših in strukturno preprostejših besedil, kar so avtorji začeli z zbiranjem dveh basni, ki utegnejo aplikaciji prinesti dodano pedagoško vrednost v osnovnošolskem učnem procesu in seznanjanja učencev s slovenskimi narečnimi govori. IKNB izmed obravnavanih treh aplikacij pokriva največje število raziskovalnih točk oz. krajev, kjer so bili zbrani narečni posnetki, vendar kljub temu ostajajo slabše zapolnjena ali nezapolnjena območja, kjer bodo dobrodošli novi vnosi zvočnih posnetkov s pripadajočimi fonetičnimi prepisi, poknjižitvami prepisov in analizami govorov. SSOLP bi bilo treba dopolniti z novimi slovarskimi sestavki in fotografskim gradivom ob že obstoječih. Na ta način bi aplikacije, podobno pa tudi ostale spletne dialektološke vsebine, zares rasle in se nadgrajevale. Ob tem želiva v širšem kontekstu slovenskih spletnih dialektoloških virov dodati, da bi bilo treba metapodatke slovarjev

44 Najnovejšo nadgradnjo spletne aplikacije SSOLP je v času nastajanja tega besedila opravil Dimitrije Mitić v okviru diplomskega dela. Skrbniški del aplikacije je zdaj bolj intuitiven in uporabniku prijazen: slikovno gradivo v tabelah je opremljeno z možnostjo predogleda slik, ki jih je mogoče tudi povečati; dodano je modalno okno za ustvarjanje povezave med geslom in medijskimi vsebinami, ki bodo služile kot ilustrativno gradivo; ilustrativno gradivo je mogoče prostorsko razmeščati in urejati vrstni red prikaza; medleksemske povezave je mogoče poljubno ustvarjati ter jih oblikovno in vsebinsko prilagajati slovarskim podatkom; zelo pomembna je tudi prilagoditev vmesnika za nemoteno delovanje aplikacije na namiznih in mobilnih napravah (gl. Mitić, 2021).

in orodij na spletnih mestih jasno izpostaviti, s čimer bi omogočili ustrezno citiranje jezikovnih virov in s tem sledljivost podatkov ter ponovljivost rezultatov raziskav. Poleg objave celovitih metapodatkov bi bilo jezikovnim virom koristno dodati trajne enkratne identifikatorje za nedvoumno identifikacijo, kot priporočajo sodobne smernice na področju citiranja jezikovnih virov (Lenardič idr., 2020, str. 22).

Osrednje težave za razvoj aplikacij ne vidiva v pomanjkanju idej ali v kakovosti njihove izvedbe, temveč predvsem v kratkoročnosti projektov. Objave (npr. Škofic, 2013; Vičič in Marc Bratina, 2015; Benko, 2016; Smole, 2019) kažejo, da so slovenski dialektologi dobro seznanjeni z novostmi doma, v določeni meri tudi z novostmi v tujini, vendar je razvoj jezikovnih tehnologij, namensko razvitih za dialektološko rabo, počasen in v veliki meri odvisen od pilotnih projektov posameznikov in njihovega dela, ki v večini primerov, predstavljenih v podpoglavjih 2.1 in 2.2, poteka ali v okviru študentskih zaključnih del ali priložnostnih fakultetnih projektov. Takemu delu za dolgoročne uspehe in dosego ciljev manjka ustrezna in trajna institucionalna podpora, ki edina utegne vzdrževati izhodiščno vizijo in v končni fazi ponuditi jezikovni oz. dialektološki vir, kot je bil načrtovan,[45] ter ga nadalje razvijati glede na uporabniške potrebe.[46] Pri tem so lahko v veliko pomoč mehanizmi množičnega zunanjega izvajanja (ang. *crowdsourcing*), tj. pridobivanja podatkov s pomočjo množice izvajalcev (internetne javnosti), in množičnega financiranja (ang. *crowdfunding*), tj. pridobivanja denarne podpore nepovezanih posameznikov.[47] Prvi od mehanizmov je bil doslej predviden v več pilotnih spletnih aplikacijah (prim. Vičič in Marc Bratina, 2015; Mrvič, 2020), drugi še ne. Kot predpogoj uspešne implementacije enega in drugega mehanizma vidiva predvsem razvejano promocijsko strategijo trenutno dostopnih aplikacij, kajti izkušnje študentov v njihovem lokalnem okolju in okolju osnovnih in srednjih šol, kjer opravljajo

45 Ne le dosledna izvedba z zagotavljanjem potrebnih sredstev, institucionalna podpora je ključna tudi za zagotovitev (dolgo)trajnosti metapodatkov jezikovnih virov, kar je mogoče le s pomočjo za ta namen razvite računalniške infrastrukture (prim. Lenardič idr., 2020, str. 23).

46 S pomočjo podatkov, pridobljenih v okviru empiričnih raziskav (prim. Arhar Holdt, 2017), bi lahko strokovnjaki in študenti z različnih področij ponudili boljše programske rešitve in posledično omogočili kakovostnejše vnose narečnih jezikovnih podatkov.

47 Slovenski terminološki ustreznici uporabljava po terminološkem slovarju informatike (Islovar), ki je dostopen na http://www.islovar.org.

pedagoško prakso, kažejo, da je zelo malo ljudi seznanjenih z naštetimi aplikacijami in vsebinami, ki jih ponujajo.[48]

5 SKLEP

Z digitalizacijo in tehnološkim razvojem so številni jezikoslovni viri začeli izkoriščati prednosti, ki jih omogoča spletno okolje, mednje v slovenskem jezikoslovnem prostoru zlasti v zadnjem desetletju sodijo tudi dialektološki viri. Slednji uporabnikom – tako jezikoslovcem in študentom jezikoslovnih smeri kot širši javnosti – omogočajo hitrejši in lažji dostop do informacij o narečnih pojavih. Kljub pomanjkanju raziskav, ki bi to empirično potrdile, je na številnih spletnih mestih, predvsem na družbenih omrežjih, mogoče opaziti povečan interes javnosti za slovenska narečja (gl. op. 1).

Digitalizirani dialektološki viri, kakršnih je večina slovarjev na portalu Fran, so bili prvotno izdani v tiskani obliki, ob objavi na spletu pa možnosti medija niso izkoristili, kar uporabnikom otežuje iskanje po njih. Digitalni viri, prvotno zasnovani za splet, kakršen je npr. slovar *Narečna bera*, predstavljajo uporabniku prijaznejši pristop, ki bolje izkorišča možnosti spletnega okolja; v spletni slovar je npr. vključeno slikovno gradivo, zvočni posnetki, videoposnetki in v omejenem obsegu določene medleksemske povezave, še vedno pa slovarska mikrostruktura temelji na tiskanih slovarjih. V tem oziru

48 Zbiranje hišnih imen na Gorenjskem je primer ene izmed najuspešnejših praks, ki združuje uspešno promocijo in med drugim tudi dialektološko delo, potekala pa je v okviru več projektov od 2009 do 2016. V izvajanje projektov je bila vključena širša javnost: redno so bila organizirana srečanja z domačini, ki so prispevali narečno gradivo o hišnih imenih, vzpostavljeno pa je bilo tudi sodelovanje s 16 občinami in skoraj vsemi osnovnimi šolami v mreži 270 krajev, ki so bili vključeni v projekte. Rezultati več kot 12.000 zbranih hišnih imen so dostopni na spletni strani https://www.hisnaimena.si, raba hišnih imen pa je bila leta 2020 vpisana tudi v register nesnovne kulturne dediščine. Slednja se kaže kot pomembna motivacija številnih projektov – 2021 je bila namreč vzpostavljena spletna aplikacija *Zapisi spomina* (https://zapisi-spomina.dobra-pot.si), ki je namenjena deljenju informacij o nesnovni kulturni dediščini po načelu medgeneracijskega prenosa znanja. Aplikacija je uporabniku prijazna in temelji na participativni uredniški politiki, ki vključuje tako ljubiteljske uporabnike kot raziskovalce, med slednjimi zlasti etnologe, folkloriste in dialektologe. Aplikacija je zanimiva zlasti kot primer participativne platforme, ki je med drugim namenjena digitalnemu opismenjevanju najstarejših, čemur se pri razvoju dialektoloških aplikacij doslej ni namenjalo pozornosti, četudi so med informatorji na terenskem dialektološkem delu običajno najstarejši prebivalci raziskovanega območja.

je naprednejši *Slovar oblačilnega izrazja ziljskega govora v Kanalski dolini*, ki vsebuje povezave na druge slovarske vire, prevode v tuje jezike in se ponaša z uporabniku prijaznejšim vmesnikom. Možnosti spletnega prikaza dialektoloških vsebin najbolje izkoriščajo spletne aplikacije, o katerih v kontekstu tega besedila govoriva kot o specializiranih jezikoslovnih oz. dialektoloških orodjih, ki zaradi prednosti digitalnega medija raziskovalcem ponujajo nove možnosti raziskovanja. Tovrstne aplikacije uporabnikom omogočajo spoznavanje narečnih govorov na več ravneh in hkrati omogočajo prilagajanje prikaza narečnih vsebin v brskalniku. V slovenskem prostoru zaenkrat obstajajo tri – *Slovenski narečni atlas* (SNA), *Interaktivna karta slovenskih narečnih besedil* (IKNB) in *Slovar starega orodja v govoru Loškega Potoka* (SSOLP) –, ki so nastale v interdisciplinarnem sodelovanju različnih fakultet Univerze v Ljubljani ter so prostodostopne, odprtokodne, interaktivne in rastoče.

Glavni namen vseh treh aplikacij in hkrati njihova najpomembnejša skupna točka je interaktivno približati raznolikost slovenskih narečij različnim uporabnikom. Ker se v načinu predstavljanja gradiva razlikujejo, se razlikujejo tudi v tem, kateri skupini naslovnikov so namenjene, tj. v kolikšnem obsegu bodo lahko informativne. Najširšemu krogu uporabnikov je zaenkrat namenjen SSOLP – tematski narečni slovar, ki vključuje podtemi orodje za sekača in tesača ter orodje za spravilo sena. Iskanje po slovarju je preprosto, medleksemske povezave pa uporabnikom omogočajo razgibane možnosti raziskovanja narečne vsebine. V slovar je vključeno multimedijsko ponazarjalno gradivo – fotografije, zvočni posnetki in videoposnetki z dodanimi fonetičnimi prepisi govorjenih besedil in prevodi v knjižni jezik. Aplikacija je med vsemi tremi najbolj oblikovalsko dodelana, njen uporabniški vmesnik pa je bil v letu 2021 uspešno posodobljen. Jezikoslovcem in širši javnosti je namenjena tudi aplikacija IKNB, ki omogoča spoznavanje slovenskih narečij in posameznih krajevnih govorov na celotnem slovenskem jezikovnem prostoru. Vsak vključeni govor je predstavljen z zvočnim posnetkom narečne pripovedi, ki se sklada s krovno temo, v večini primerov je posnetkom dodana tudi fonetična transkripcija in glasovna poknjižitev, ponekod pa tudi diahrona analiza krajevnega govora. Trenutno so v aplikacijo vključene narečne pripovedi na krovno temo stare kmečke hiše in imajo sorodno vsebinsko strukturo, ki se opira na enotni model, kar olajša primerljivost in naredi zbrano gradivo bolj

informativno. SNA je za uporabnika nejezikoslovca precej zahtevnejši in manj intuitiven, kot sta SSOLP in IKNB, saj od njega pričakuje temeljno geolingvistično znanje o uporabi jezikovnih kart in poznavanje fonetične transkripcije, v kateri je zapisano narečno gradivo. Aplikacija je ustvarjena za kartiranje narečne leksike iz različnih tematskih polj; trenutno prinaša predvsem primerjalne frazeme s pomenom človeške lastnosti. Kartiranim narečnim leksemom so dodani fonetični zapisi, mestoma pa tudi zvočni posnetki.

Z vsako novo spletno narečno aplikacijo spremljamo vzpostavitev novih ali vsaj nadgradnjo obstoječih pristopov k digitalni predstavitvi slovenskega narečnega gradiva. Tako tudi aplikacije, obravnavane v prispevku, predstavljajo napredek v slovenskem jezikoslovnem prostoru in doprinos k postopnemu razvoju spletnih odprtokodnih orodij na področju dialektologije. Meniva, da so za pospešitev tega procesa nujni: 1) sodelovanje znotraj in zunaj jezikoslovne stroke, da se zagotovi kakovostno in učinkovito institucionalno podporo novim dialektološkim virom, 2) upoštevanje potreb in želja jezikovnega uporabnika, ki bi moralo izhajati iz izsledkov empiričnih raziskav, ter zlasti 3) promocija spletnih dialektoloških virov in spodbujanje strokovnega dialoga s širšo javnostjo na tem področju, ki bo z vnosom kritične presoje obstoječih virov lahko odprl nove, še nepreizkušene možnosti sodobnih mehanizmov za dolgoročni razvoj.

LITERATURA

Slovarski in drugi dialektološki viri

Lovrić, I., Jukan, N. idr. (2018). *Interaktivna karta slovenskih narečnih besedil* (IKNB). Pridobljeno s https://narecja.si

Nusheski, A., Mitić, D. idr. (2018). *Slovar starega orodja v govoru Loškega Potoka* (SSOLP). Pridobljeno s https://slovar-orodja.si

Šajn, G. idr. (2017). *Slovenski narečni atlas* (SNA). Pridobljeno s https://sna.si

Benko, A. (2013). *Narečna bera*. Pridobljeno s http://www.narecna-bera.si

Gostenčnik, J. idr. (2014). *Slovenski lingvistični atlas 1*. Pridobljeno s https://fran.si/204

Gregorič, J. (2015). *Kostelski slovar*. Pridobljeno s https://fran.si/197

Ivančič Kutin, B. (2015). *Slovar bovškega govora*. Pridobljeno s https://fran.si/196

Kenda-Jež, K. ([1]2007, [2]2015/spletna različica 2019). *Slovar oblačilnega izrazja ziljskega govora v Kanalski dolini*. Pridobljeno s https://fran.si/210

Kumin Horvat, M. (2018). *Besedotvorni atlas slovenskih narečij: Kulturne rastline*. Pridobljeno s https://doi.org/10.3986/9789610504214

Mezgec, T., Šukljan, T., & Vičič, J. *Narečni frazem*. Različica 0.9.1. Pridobljeno s http://frazem.famnit.upr.si

Mrvič, R., & Žnidaršič, T. (2020). *Frazeograf*. Pridobljeno s https://www.frazeograf.si

Razvojna agencija Zgornje Gorenjske (RAGOR) (2013). *Slovenska hišna imena*. Pridobljeno s https://www.hisnaimena.si

Slovensko društvo Informatika (2001). *Islovar*. Pridobljeno s https://www.islovar.org

Steenwijk, H. (2004). *Resianica*. Pridobljeno s http://147.162.119.1:8081/resianica/dictionaryForm.do

Škofic, J., & Vičič, J. (2013). *Interaktivni Slovenski lingvistični atlas*. Pridobljeno s https://sla.zrc-sazu.si/#v

Škofic, J. idr. (2016). *Slovenski lingvistični atlas 2*. Pridobljeno s https://fran.si/204

Šumenjak, K., & Vičič, J. (2013). GOKO. Pridobljeno s https://jt.upr.si/GOKO/index.html

Šumenjak, K., & Vičič, J. (2013). GOSP. Pridobljeno s https://gosp.upr.si/GOSP/index.html

Tominec, I. (2015). *Črnovrški dialekt*. Pridobljeno s https://fran.si/194

Zavod Dobra pot (2021). *Zapisi spomina*. Pridobljeno s https://zapisi-spomina.dobra-pot.si

Weiss, P. (2015). *Slovar govorov Zadrečke doline med Gornjim Gradom in Nazarjami (A–H)*. Pridobljeno s https://fran.si/195

Drugo

Arhar Holdt, Š. (2017). Uporabniške raziskave za potrebe slovenskega slovaropisja: prvi koraki. V V. Gorjanc, P. Gantar, I. Kosem in S. Krek (ur.), *Slovar sodobne slovenščine: problemi in rešitve* (str. 136–148). Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani. Pridobljeno s http://www.dlib.si/?URN=URN:NBN:SI:DOC-21CL5BT0

Benko, A. (2016). Slovensko narečno slovaropisje: Razvoj, stanje, prihodnost. V K. Šter, M. Žagar Karer (ur.), *Historični seminar 12* (str. 123–143).

Ljubljana: Založba ZRC, ZRC SAZU. Pridobljeno s http://hs.zrc-sazu.si/Portals/0/sp/hs12/Benko.pdf

Bon, M. (2018). *Geolingvistična interpretacija primerjalnih frazemov v slovenskih narečjih na interaktivni jezikovni karti: Primerjalni frazemi s pomenom človeške lastnosti*. Magistrsko delo. Ljubljana: Filozofska fakulteta Univerze v Ljubljani.

Ivančič Kutin, B. (2017). Gradivo za etnološko kontekstualizacijo muzejskih predmetov kot vir za jezikoslovne raziskave: študija primera. *Jezik in slovstvo*, *62*(4), 65–79. Pridobljeno s http://www.jezikinslovstvo.com/pdf.php?part=2017|4|

Kavčič, A., Lovrić, I., & Smole, V. (2018). Interaktivna karta slovenskih narečnih besedil. V D. Fišer in A. Pančur (ur.), *Zbornik konference Jezikovne tehnologije in digitalna humanistika* (str. 121–125). Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani. Pridobljeno s http://www.dlib.si/stream/URN:NBN:SI:doc-YWTL37V1/35baba0d-2fb8-4125-828a-d84827405afb/PDF

Lenardič, J., Erjavec, T., & Fišer, D. (2020). Citiranje jezikovnih podatkov v slovenskih znanstvenih objavah v obdobju 2013–2019. *Slovenščina 2.0, 8*(1), 1–34. Pridobljeno s https://doi.org/10.4312/slo2.0.2020.1.1-34

Lovrić, I. (2018). *Interaktivna spletna aplikacija za slovenska narečna besedila*. Diplomsko delo. Ljubljana: Fakulteta za računalništvo in informatiko Univerze v Ljubljani. Pridobljeno s https://repozitorij.uni-lj.si/Dokument.php?id=110326&lang=slv

Mitić, D. (2021). *Interaktivni tematski narečni slovar*. Diplomsko delo. Ljubljana: Fakulteta za računalništvo in informatiko Univerze v Ljubljani. Pridobljeno s https://repozitorij.uni-lj.si/Dokument.php?id=141440&lang=slv

Mrvič, R. (2020). *Koncept narečnega frazeološkega slovarja: tiskana in elektronska oblika*. Magistrsko delo. Ljubljana: Filozofska fakulteta Univerze v Ljubljani. Pridobljeno s https://repozitorij.uni-lj.si/Dokument.php?id=135102&lang=slv

Smole, V. (2019). Slovenska narečja v spletnih aplikacijah. V M. Smolej (ur.), *1919 v slovenskem jeziku, literaturi in kulturi. 55. seminar slovenskega jezika, literature in kulture* (str. 20–30). Ljubljana: Znanstvena založba Filozofske fakultete. Pridobljeno s https://centerslo.si/wp-content/uploads/2019/06/55-SSJLK_Smole.pdf

Smole, V., Gabrijelčič Tomc, H., & Kavčič, A. (2020). Uporaba novih medijev v narečnem slovaropisju na primeru *Slovarja starega orodja v govoru Loškega Potoka*. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje*, *46*(2), 1039–1057. Pridobljeno s https://hrcak.srce.hr/245482

Šajn, G. (2017). *Interaktivni atlas slovenskih narečnih besed*. Diplomsko delo. Ljubljana: Fakulteta za računalništvo in informatiko Univerze v Ljubljani. Pridobljeno s https://repozitorij.uni-lj.si/Dokument.php?id=102633&lang=slv

Škofic, J. (2013). Priprava interaktivnega Slovenskega lingvističnega atlasa. *Jezikoslovni zapiski*, *19*(2), 95–111. Pridobljeno s https://ojs.zrc-sazu.si/jz/article/view/2300

Šumenjak, K. (2013). *Opis govora Koprive na Krasu na osnovi dialektološkega korpusa*. Doktorska disertacija. Koper: Fakulteta za humanistične študije Univerze na Primorskem. Pridobljeno s https://repozitorij.upr.si/Dokument.php?id=12070&lang=slv

Vičič, J., & Marc Bratina, K. (2015). Narečni frazeološki slovar – prvi koraki. V M. Smolej (ur.), *Slovnica in slovar – aktualni jezikovni opis*. *Obdobja 34* (str. 811–818). Ljubljana: Znanstvena založba Filozofske fakultete. Pridobljeno s https://centerslo.si/wp-content/uploads/2015/11/34_2-Vicic-Bra.pdf

THREE ONLINE APPLICATIONS ON SLOVENIAN DIALECTS

The need for a greater presence of dialectal content on the internet and its interactive multimedia presentation, especially professionally designed dialectological sources and tools, has encouraged an interdisciplinary cooperation between various faculties of the University of Ljubljana, chiefly the Faculty of Arts and the Faculty of Computer and Information Science. This union bore fruit in 2017 and 2018 in the form of three free and open-source web applications on Slovene dialects – these are *Slovenski narečni atlas* (SNA, 2017), *Interaktivna karta slovenskih narečnih besedil* (IKNB, 2018) and *Slovar starega orodja v govoru Loškega Potoka* (SSOLP, 2018), which are a Slovene dialect atlas, an interactive map of Slovene dialect texts and a dictionary of old tools in the local speech of Loški Potok, respectively. The article begins with a general overview of Slovenian online dialectological resources and tools, while the second part provides a more detailed presentation of these three applications currently available to users in terms of functionality. In the discussion, the circumstances of said applications' development and the related limitations are considered, with suggestions on some possible solutions that ought to be regarded to ensure long-term development.

Keywords: Slovenian dialects, online application, dialect atlas, dialect dictionary, interactive map