Building a CEFR-Labeled Core Vocabulary and Developing a Lexical Resource for Slovenian as a Second and Foreign Language

Matej Klemen[1], Špela Arhar Holdt[1,2], Senja Pollak[3], Iztok Kosem[1], Eva Pori[1], Polona Gantar[1], Mihaela Knez[1]

[1] Faculty of Arts, University of Ljubljana, Aškerčeva ulica 2, 1000 Ljubljana, Slovenia

[2] Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, 1000 Ljubljana, Slovenia

[3] Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia

E-mail: matej.klemen@ff.uni-lj.si, spela.arharholdt@ff.uni-lj.si, senja.pollak@ijs.si, iztok.kosem@ff.uni-lj.si, eva.pori@ff.uni-lj.si, apolonija.gantar@ff.uni-lj.si, mihaela.knez@ff.uni-lj.si

Abstract

This article introduces two newly available datasets: the KUUS 1.0 corpus and the list Core Vocabulary for Slovenian as L2 1.0. The KUUS 1.0 corpus consists of seventeen textbooks published by the Center for Slovene as a Second and Foreign Language at the University of Ljubljana, and it contains a total of 520,796 words accompanied by various linguistic tags and metadata. Using the KUUS 1.0 corpus, we compiled the list Core Vocabulary for Slovenian as L2 1.0. The list includes 350 words labeled as A1-core, 864 words as A1-larger, 1,451 words as A2, and 2,608 words as B1. The A1 vocabulary was used as pilot data for a project focused on developing a lexical description for learning Slovenian as a second and foreign language. Our methodology involved combining the data from the new datasets with existing, openly available lexical information on modern Slovenian, with the aim of achieving didactic adaptation and maximal reusability of the results.

Keywords: Lexicography and CEFR; Slovenian; second and foreign language; textbook corpus; core vocabulary

1. Introduction

Most existing CEFR-based language documents and curricula for Slovenian as a second and foreign language—for example, *Preživetvena raven v slovenščini* (Breakthrough Level in Slovenian, 2004, revised version 2016); *Sporazumevalni prag za slovenščino* (Threshold Level for Slovenian, 2004), and *Slovenščina kot drugi in tuji jezik: Izobraževalni program za odrasle* (Slovenian as a Second and Foreign Language: Adult Education Program, 2020)—are based on consensual expert group knowledge. The documents present a general description of language skills and contain a list of illustrative vocabulary. Over the past twenty years, these documents have been the basis for developing learning materials aimed at different target groups (e.g., adolescents, students, adult speakers, etc.) learning Slovenian as a second and foreign language in Slovenia and in other countries. The different communicative needs of these learners are reflected in the different choice of vocabulary in the learning materials.

Our aim was to create a corpus-based list of core vocabulary[1] covering different CEFR levels. This article presents KUUS 1.0, a corpus of textbooks for learning Slovenian as a second and foreign language, and how it was used to create the corpus-based list Core Vocabulary for Slovenian as L2 1.0,[2] which contains single-word vocabulary labeled as A1, A2, and B1. We then present how the newly available datasets have been included in developing the CEFR-labeled lexical resource for Slovenian as a second and foreign language.

2. The KUUS 1.0 corpus

The work presented in this article is based on the KUUS 1.0 corpus, which is a collection of seventeen textbooks specifically created for teaching Slovenian as a foreign and second language. These textbooks, published between 2002 and 2022 by the Center for Slovene as a Second and Foreign Language (Sln. *Center za slovenščino kot drugi in tuji jezik*, CSDTJ) at the University of Ljubljana, are widely used in both Slovenia and other countries to teach Slovenian to learners of different ages at various CEFR levels (Gril, 2022: 123; Knez et al., 2021: 261–262, 342–343). The KUUS corpus was developed as a companion project to the CSDTJ's publishing of graded readers series and aims to provide a standardized, linguistically annotated, and openly accessible dataset of this nature for Slovenian.

KUUS 1.0 includes metadata for each textbook, including the title, subtitle, authors, year of publication, publisher, CEFR level, target audience, and estimated number of lessons. The corpus was linguistically annotated with the CLASSLA v1.1.1 pipeline (Ljubešić & Dobrovoljc, 2019) at the levels of tokenization, sentence segmentation, lemmatization, MULTEXT-East v6 MSD-tags, JOS dependency syntax, and named entities. The current version of the corpus comprises 520,796 words and is available as a database at the CLARIN.SI repository (Klemen et al., 2022a).

The selection of textbooks was made to cover different CEFR levels, contain the bulk of the textbook production of the CSDTJ, and comprise a significant part of the current textbooks for learning Slovenian as a foreign and second language. The texts were converted from PDF or DOC format into TXT format. Parts of the textbooks that are not intended for the student or for direct use in teaching were manually removed. These typically included the introduction, table of contents, colophon, and sources of pictures and texts. In addition, any recurring text in the header or footer of pages was deleted, except for page numbers. Foreign-language instructions were marked with special codes so that they are easily separable from the Slovenian part of the text. We furthermore

[1] Similar to Volodina et al. (2022), we understand core vocabulary as vocabulary known to most learners at a certain level of language proficiency. In terms of building a vocabulary list, we understand the "core" as a consensually agreed-upon and stable but expandable starting point for learning.

[2] In the article, we refer to the official names of institutions and published resources, which can result in some discrepancies, such as the contrast between "Slovene as a Second and Foreign Language" and "Slovenian as L2."

corrected any errors that occurred during the conversion process, including problems with characters such as *č*, *š*, *ž*, upper- and lower-case letters, punctuation, and hyphenated words. In some cases, we had to add text that was erroneously omitted during the conversion due to specific fonts or layouts. The preparation of the KUUS corpus is presented in greater detail by Klemen et al. (2022).

Some of the textbooks included in the KUUS corpus have a part of the book that is structurally similar to workbooks and includes grammar exercises. These parts of textbooks have been included in the corpus because they are part of a single publication. However, in the current version, the corpus only includes textbooks and not the corresponding workbooks.

3. Core Vocabulary for Slovenian as L2 1.0

Using the KUUS 1.0 corpus, we prepared the list Core Vocabulary for Slovenian as L2 1.0 (Klemen et al. 2022b). The list comprises 5,273 lemmas, classified into the first three CEFR levels: 350 lemmas with the assigned label A1-core, 864 words with the label A1-larger, 1,451 words with the label A2, and 2,608 words labeled B1.[3] The current version of the list is available at CLARIN.SI in a tab-separated format containing the lemma, part-of-speech (following the MULTEXT-East tagset for Slovenian), information on whether the lemma appears in the Reference List of Slovene Frequent Common Words (Pollak et al., 2020), and the relative average frequency. An example of the data is presented in Table 1.

CERF level	**Lemma**	**POS**	**Lemma in Reference List of Slovene Frequent Common Words**	**Sum of relative frequencies across textbooks**
A1-core	*biti* 'to be'	g	Yes	124.87740
A1-core	*v* 'in'	d	Yes	38.03003
A1-core	*se* 'oneself'	z	Yes	34.44841
A1-core	*in* 'and'	v	Yes	34.28150

[3] In the article, we intentionally make a distinction between "level" and "label." Here, "level" refers to the CEFR level, while "label" pertains to the corpus-based annotation of a specific lemma in the core vocabulary list. In our methodology, the current labels serve as a baseline and are subject to potential modifications in subsequent stages of our work.

When creating the list of core vocabulary, we did not distinguish between criterion levels and plus levels (e.g. A2 and A2+) as conceived in the CEFR Companion Volume, as the labels on the textbooks do not differentiate between them. Therefore, we have used the labels B1 and A2, and for A1 we have introduced two labels: A1-core and A1-larger. The former was assigned to words that appear in all five A1 textbooks included in the KUUS 1.0 corpus, the latter to words that appear in four or fewer A1 textbooks.

A1-core	*na* 'on, at'	d	Yes	26.39539
A1-larger	*ki* 'which'	v	Yes	6.74070
A1-larger	*svoj* 'one's one'	z	Yes	3.67359
A1-larger	*če* 'if'	v	Yes	3.17442
A1-larger	*človek* 'human'	s	Yes	3.16109
A1-larger	*res* 'really'	r	Yes	3.14526
A2	*treba* 'necessary'	r	Yes	0.98788
A2	*saj* 'because'	v	Yes	0.96636
A2	*pomemben* 'important'	p	Yes	0.95674
A2	*zaradi* 'because of'	d	Yes	0.92400
A2	*svet* 'world'	s	Yes	0.92355
B1	*nekdo* 'someone'	z	Yes	0.32863
B1	*glede* 'regarding'	r	Yes	0.28649
B1	*sodoben* 'contemporary'	p	Yes	0.27790
B1	*lastnost* 'characteristic'	s	Yes	0.26160
B1	*dejanje* 'action'	s	Yes	0.24973

Table 1: First five lemmas for each CEFR level in the Core Vocabulary for Slovenian as L2 1.0 with associated data.

In summary, our approach involved importing the corpus into the Sketch Engine tool (Kilgarriff et al., 2014), exporting the frequency lists for each separate textbook, and calculating the relative frequency of each word (lempos) across the seventeen textbooks. We compiled these data (23,068 words of different types) into a single table that included information on word frequency and occurrence across textbooks at each CEFR level. Next, we compared the data to the Reference List of Common Frequent Words (Pollak et al., 2020). This reference list consists of 4,768 common general lemmas compiled by comparing the most frequent 10,000 lemmas by word type from four Slovenian text corpora: Kres 1.0, GOS 1.0, Janes 1.0, and Šolar 2.0.[4] We found an

[4] The Kres corpus (Logar et al., 2012) is a balanced sub-corpus of the Gigafida reference corpus, with almost 100 million words from various written sources. The Janes corpus (Fišer et al., 2020) consists of online user-generated content, and the Šolar corpus (Kosem et al., 2016) consists of written texts created independently by primary- and secondary-school

overlap of 4,603 words between the two lists, with only 166 words appearing solely in the list of common general vocabulary but not in the KUUS corpus, and 18,465 words appearing only in the KUUS corpus (Klemen et al., 2022a: 170).

After conducting a comprehensive first review of the data, we established robust numerical criteria with the aim of obtaining core (i.e., relevant or typical) vocabulary for each level from the textbook material. The criteria were used to assign a baseline CEFR-level label to the words. The criteria were considered sequentially, starting with the A1-core criteria, followed by the A1-larger criteria check, and so on. When preparing the criteria, we considered that there are fewer textbooks available for B1 than for A1 and A2, and that a textbook covering two levels (A2–B1, see Klemen et al. 2022) also appears in the material.

– For the A1-core label, the word must appear in all five level-A textbooks (e.g., *nov* 'new', *dober* 'good', *slovenski* 'Slovenian', *star* 'old', *velik* 'big', *lep* 'beautiful', *majhen* 'small', *mlad* 'young', *sam* 'alone', *zanimiv* 'interesting').[5]

– For the A1-larger label, the word must appear in four, three, or two level-A textbooks (e.g., *ustrezen* 'relevant', *srednji* 'middle', *prijazen* 'friendly', *prost* 'free', *visok* 'high', *beseden* 'word', *ženski* 'feminine', *naslednji* 'next', *deloven* 'working', *oseben* 'personal').

– For the A2 label, the word appears in no more than one A1 textbook, but it appears in five, four, three, or two A2 textbooks (e.g., *pomemben* 'important', *znan* 'known', *različen* 'different', *svetoven* 'global', *evropski* 'European', *kulturen* 'cultural', *šolski* 'school', *osnoven* 'basic', *zadovoljen* 'satisfied', *posloven* 'business'). (If a word appears in two textbooks at level A2, and one of them is the A2–B1 textbook, then it is considered a B1 word.)

– For the B1 label, the word does not appear in A1 textbooks and can appear in at most one A2 textbook. It must appear in one or two B1 textbooks, and it must have a frequency of at least 2 in the entire corpus (e.g., *sodoben* 'contemporary', *državen* 'state-owned', *družben* 'social', *socialen* 'social', *skupen* 'common', *lasten* 'own', *današnji* 'today's', *prepričan* 'convinced', *vprašan* 'asked', *posamezen* 'individual').

We manually reviewed the labeled words and eliminated any irrelevant instances that we considered to be noise, such as erroneously lemmatized or POS-tagged data, proper nouns, and numerals that would require separate addition because they are not represented systematically in the corpus. However, we decided to retain linguistic terminology and metalanguage commonly found in textbooks, symbols, and

students. The GOS corpus (Verdonik & Zwitter Vitez, 2011) is a spoken Slovenian reference corpus with 120 hours of recordings, spanning a wide range of contexts.

[5] As an example, the first ten adjectives of each tag are given. The English glosses do not necessarily cover all the meanings and are for general information only. Because of the identical form of adjective and noun in English, certain adjectives may appear as nouns in translation, e.g. *šolski* 'school' as in *šolske počitnice* 'school holidays'.

abbreviations. During our examination of the words in a wider textual context, we encountered some cases that belonged to a higher level than B1 due to mislabeling, homonymy, or polysemy. Nonetheless, in the vast majority of cases, the automatically assigned CEFR labels were found adequate. It is worth noting that our methodology in the first step is purposefully permissive because we prefer to include a word too many rather than too few.

4. Developing a Lexical Resource for Slovenian as a Second and Foreign Language

This section presents how the corpus and the list described in sections 2 and 3 are being utilized to develop a new lexical resource for Slovenian as a second and foreign language. Because the resource is still a work in progress, we explain the methodological considerations and present the work on sample entries.

4.1 Project framework

An opportunity to utilise the newly prepared datasets was offered as part of the project *Nadgradnja učnega gradiva* Čas za slovenščino 1 *v digitalnem okolju in prilagoditev gradiva za pouk slepih in slabovidnih mladostnikov* (Expanding the Teaching Material *Čas za slovenščino 1* in the Digital Environment and Adapting the Material for Teaching Blind and Partially Sighted Adolescents), led by the CSDTJ. As part of the project, funded by the Slovenian Ministry of Culture,[6] we committed ourselves to enriching the vocabulary previously labeled as A1 (see section 3) with user-adapted grammatical, semantic, and multimedia information (e.g., pronunciation recordings) in Slovenian, and to including translations of the headwords into three foreign languages (i.e., Albanian, English, and Hungarian),[7] thus combining monolingual and multilingual dictionary approaches. For this purpose, the project envisages using all relevant information in the lexicographical and other resources produced by the Center for Language Resources and Technologies (Sln. *Center za jezikovne vire in tehnologije*, CJVT) at the University of Ljubljana, revisiting them through the approaches developed at the CSDTJ on the basis of experience in teaching Slovenian as a second and foreign language.

As part of the project, we aim to prepare a lexical resource that could be used by A1 users of Slovenian because no such dictionary for Slovenian has been developed yet. We

[6] The aims of the project are threefold: (a) preparation of a digital platform with interactive activities for learning, (b) development of a lexical resource (as described in this article), and (c) adaptation of the textbook for teaching blind and partially sighted learners (cf. https://centerslo.si/za-otroke/projekti/nadgradnja-ucnega-gradiva-cas-za-slovenscino-1-v-digitalnem-okolju-in-prilagoditev-gradiva-za-pouk-slepih-in-slabovidnih-mladostnikov/).

[7] The three languages were chosen for the following reasons: Albanian is a non-Slavic language spoken by migrants that have moved to Slovenia from Kosovo (cf. Knez et al., 2021); English is a lingua franca; and Hungarian is used in the Slovenian cross-border area and, as a non–Indo-European language, is the least similar to Slovenian among the four neighboring languages (German, Italian, Croatian, and Hungarian).

are thus targeting users able to understand the explanation (which we perceive as both the basic description and the illustrative material) of the headword, provided that it uses common everyday expressions and very basic phrases in simple grammatical and sentence structures referring to particular concrete situations (e.g., the most basic personal and family information, everyday routine activities and tasks, schooling, or employment) in which the users communicate in Slovenian in their everyday life and which they need to meet their concrete needs and perform linguistic tasks relevant to them (cf. Companion Volume, 2020: 54, 56, 60, 131–132, 175).

Furthermore, the idea is that the resource could be systematically expanded for users of Slovenian as a second and foreign language at higher levels of language proficiency (A2–C1) in the future.

4.2 Methodological background

The list Core Vocabulary for Slovenian as L2 1.0 provided the candidates for the lexical description, consisting of 350 words labeled as A1-core and 864 words as A1-larger. To ensure connectivity between the new lexical resource and the digital platform with interactive exercises (also being developed as part of this project) we supplemented the list with 247 additional words found in the textbook *Čas za slovenščino 1*, which did not meet the criteria for inclusion in the core vocabulary list.[8] The final wordlist included 1,461 words of various types (e.g., content and function words) requiring distinct lexicographic treatment.

In the first step, we selected fifteen headwords with different part-of-speech categories (common nouns, adjectives, verbs, and adverbs) and created diverse sample entries with the grammatical and semantically structured features (semantic indicators, collocations, and usage examples) to develop a model for a lexical description of Slovenian as a second and foreign language that is suitable for users at level A1 and can be expanded in the future with more complex semantic information relevant for higher-level users.

[8] The textbook *Čas za slovenščino 1* is aimed at teenagers, especially migrant children who are joining the Slovenian school system. Thus, it also includes specific vocabulary that is relevant for them at the beginner level (e.g., *radirka* 'eraser', *ravnilo* 'ruler'), but is less relevant for other users of Slovenian as a second and foreign language at this level and is therefore not included in other textbooks and consequently not part of the core vocabulary.

The additional 247 words have already been included in the baseline data but have remained unlabeled and will thus remain without a label in the new lexical resource for the time being. We plan to review them and assign them appropriate level in the subsequent stages of our work (see section 5).

The new lexical resource will be linked to the interactive exercises accompanying the textbook *Čas za slovenščino 1* (see footnote 6). This will allow learners to use it simultaneously while solving the exercises.

The test entries were created using a localized and customized version of the dictionary tool Lexonomy.[9] The grammatical and semantic information for the enrichment of the vocabulary list was taken from the following sources: the Slovene Morphological Lexicon Sloleks 2.0 (Čibej et al. 2022),[10] which contains essential information on Slovenian words (e.g., their part-of-speech category and their grammatical features) as well as recordings of word pronunciations and automatically generated recordings; the Collocation Dictionary of Modern Slovene 1.0 (Kosem et al. 2019)[11] with information on the most common and statistically typical collocations and collocations for the selected vocabulary; the Thesaurus of Modern Slovene 1.0 (Krek et al. 2018),[12] which offers synonyms as well as certain antonyms; and the Comprehensive Slovenian–Hungarian Dictionary 1.0 (Kosem et al. 2021)[13] with information on semantic indicators, dictionary labels, manually reviewed corpus examples, and Hungarian translations of words. For words not covered by these sources, the data were updated in accordance with the methodology used. Currently automatically prepared data were also manually reviewed and corrected.

The concept of the lexical resource for Slovenian as a second and foreign language includes the following elements: (a) a **semantic indicator**; (b) a **set of collocations**; (c) **usage examples**; (d) **translations of the headword** into Albanian, English, and Hungarian, and, where possible, (e) **multimedia elements** (images and recordings) that effectively illustrate the sense of the headword (Figure 1).

[9] https://lexonomy.cjvt.si/

[10] https://viri.cjvt.si/sloleks/slv/

[11] https://viri.cjvt.si/kolokacije/slv/headword/69883#

[12] https://viri.cjvt.si/sopomenke/slv/

[13] https://viri.cjvt.si/slovensko-madzarski/slv/

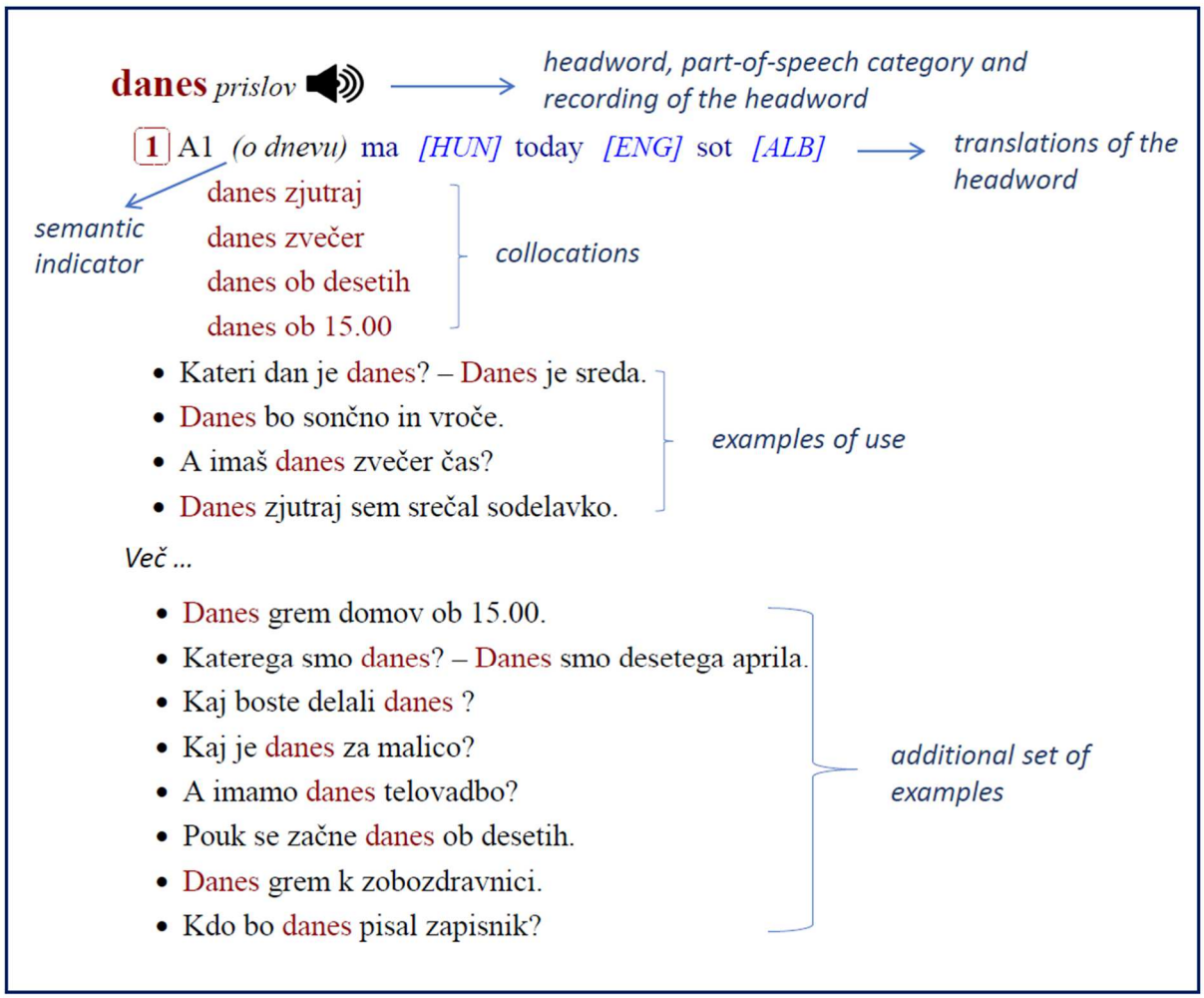

Figure 1: Sample entry for the adverb *danes* 'today'.

4.3 Elements of the semantic description in the model presented

Considering the target users and the tool being user friendly, certain guidelines have been found relevant when creating entries and semantic descriptions for level A1. These guidelines are explained below.

The **semantic indicator** is one of the three segments of semantic information in the CJVT dictionary resources.[14] It defines the meaning of a word concisely and distinctively in relation to its other meanings. For this purpose, semantic indicators are primarily used to create a "sense menu," which is familiar from foreign language dictionaries and introduced in CJVT dictionary resources (cf. Collocation Dictionary of Modern Slovene 1.0, Comprehensive Slovenian–Hungarian Dictionary 1.0).

The semantic indicator should be informative for the target users. It should be short and clear. The semantic indicator is either a thematic category (as in similar language learning resources—e.g., English Profile—these are in line with language documents

[14] In addition to indicators, the semantic description includes a label and an explanation.

such as *Threshold Level for Slovenian*)[15] or hypernym (e.g., for the headword *bel* 'white', the indicator is *lastnost – barva* 'characteristic – color'). Where this proves to be relevant and helpful, the semantic indicator should be supplemented by a synonym (e.g., in describing the meaning of some nouns: *punca:dekle* 'girl') and/or antonym (e.g., in describing the meaning of qualitative adjectives: *lep:grd* 'beautiful:ugly').

The **set of collocations** provides information about the most typical textual environment of the headword. When creating the set of collocations, we took into account collocations from the KUUS corpus that also showed semantic and statistical relevance in the Gigafida 2.0 reference corpus of written Slovenian (Krek et al. 2020). For example, in the case of the entry *bel* 'white', the collocation *bela barva* 'white color' was accepted because it has been verified as a collocation in both corpora. In addition to the aforementioned criterion of typicality, the criterion of a variety of syntactic relations or structures was taken into account; for example, the use of a noun in different cases, with different prepositions, verb valency, and so on (e.g., for the entry *babica* 'grandmother': *draga babica* 'dear grandmother', *obiskati babico* 'to visit one's grandmother', *počitnice pri babici* 'vacation at grandmother's', *dedek in babica* 'grandfather and grandmother'). However, the linguistic competence of the target user has been taken into account.

The **usage examples** are taken from the level-A1 textbooks included in the KUUS corpus. They are typically one-sentence utterances. Regarding the form, declarative sentences (*Rad bi naročil pico.* 'I would like to order a pizza.'), negative sentences (*Naša učilnica ni velika.* 'Our classroom is not big.'), and interrogative sentences (*A greš zvečer na pijačo?* 'Are you going for a drink tonight?') are included as usage examples for each headword, if relevant. In principle, one-clause sentences have been included. In some cases, simple coordinated and subordinated sentence structures (e.g., *Kaj delaš, ko prideš domov?* 'What do you do when you get home?') with the conjunctions *in* 'and', *ali* 'or', *ampak* 'but', *ker* 'because', *ko* 'when' (limited to expressing time), *če* 'if' (limited to use with present and future forms), or *ki* 'which' (limited to the nominative case) have been used if the usage has been documented in the corpus. Especially when expressing time or location, dialogue forms were also included in the usage examples so the question word could provide contextual clues and/or illustrate the grammatical limitations of use (e.g., *Kje si? – Doma.* 'Where are you? – At home.'; *Kdaj greš na dopust? – Avgusta.* 'When are you going on vacation? – In August.').

Vocabulary and morphosyntactic patterns in the usage examples correspond to the expected lexical and morphosyntactic ability of the target user. Inflectional word types are shown in their various forms (e.g., *Danes je sreda.* 'Today is Wednesday.'; *V sredo imamo angleščino.* 'We have English on Wednesday.'; *Tečaj imamo ob sredah.* 'We have classes on Wednesdays.'). We have taken into account the fact that users usually have

[15] Cf. the documents *Preživetvena raven za slovenščino* (Pirih Svetina, 2004, 2016) and *Sporazumevalni prag za slovenščino* (Ferbežar et al., 2004).

a slightly higher receptive ability than productive ability, and that they are able to use some reception strategies, especially if the examples are supported by pictures, if they can use their general knowledge or first language to help them understand the meaning, if the examples show a predictable communicative situation, or if the circumstances are familiar to the user (cf. CEFR Companion Volume, 2020: 54, 59–60, 175).

Due to transparency and the pedagogical maxim of progressivity, the usage examples are presented in two categories. The first set of examples, the three to five "core examples", appear on the screen automatically, whereas an additional set of examples appears only on demand. Within the core examples, the headword is presented in various general domains or contexts (e.g., for the headword *danes* 'today': *Danes je sreda.* 'Today is Wednesday.'; *Kateri dan je danes?* 'What day is today?'; *Danes bo sončno in vroče.* 'Today will be sunny and hot.'; *A imaš danes zvečer čas?* 'Do you have time this evening?'). In the additional set of examples, the use of the headword in specific domains is illustrated; for example, in the context of school or work (e.g., *Kaj je danes za malico?* 'What's for (school) lunch today?'; *Danes imamo geografijo.* 'Today we have geography.'; *Danes ne morem priti na sestanek.* 'Today I can't come to the meeting.'), and some usage examples beyond level A1 are shown (e.g., *Danes ponoči sem sanjala o tebi.* 'Last night I dreamed about you.').

The newly created lexical resource includes usage examples that function as self-sufficient even in isolation from a wider textual context, and that are comprehensible, accessible, and useful for the user (e.g., *Komaj čakam počitnice.* 'I can't wait for vacation to begin.'). Examples that were not semantically coherent without a context were not included (e.g., *Lepo, komaj čakam.* 'Nice, I can't wait'). The usage examples are selected from the level-A1 textbooks included in the KUUS corpus. In some entries the examples from the workbooks that complement the textbooks have also been manually included in the resource because the plan is to expand the corpus with workbooks (see section 5).

Where possible, **multimedia elements** (i.e., photographs and/or recordings of the headword) are included. As mentioned in the previous paragraph, these have an important explanatory function for users with limited linguistic ability.

4.4 Import of data into the Digital Dictionary Database for Slovene (DDDS)

Because it is essential for languages with fewer speakers to facilitate optimal connectivity and reusability of language resources and data, special care has been taken to ensure that all newly produced data are available for further use. The presented lexical information will be included in the Digital Dictionary Database for Slovene (DDDS) (Kosem et al., 2021a), which is being developed at the CJVT at the University of Ljubljana. The main aim of the DDDS is to offer a uniform set of concepts (i.e., senses) for various monolingual and bilingual dictionaries (with Slovenian as the source

language) and similar resources. Naturally, the integration of a resource targeted at nonnative speakers requires a few special features in the database; features that have been predicted since the beginning. For example, each sense in the DDDS can have more than one definition, depending on the type of resource. Similarly, examples can be attributed to one or many (or all) resources drawing on the data in the DDDS.

In the case of dictionary entries for nonnative speakers with CEFR-labeled senses, we will use the sense indicators already found in the DDDS. We expect to find most of the collocations from our entries in the DDDS already; as we already observed during the entry compilation, the collocations that are "new" are often those that are less typical in the reference (written) corpus and more typical of spoken language. Examples selected for the entries will initially be linked to this particular resource only. The information on which CEFR label should be attributed to which sense(s) is based on the KUUS corpus. Currently, the focus is on A1, and the senses already present in A1 textbooks are thus labeled as such. Sometimes a sense that is suitable for A2 or higher levels can potentially occur in A1, however we found such cases to be rare. Overall, the majority of headwords have only one A1 sense, and few two A1 senses. Expectedly, A1 senses are almost always the first senses of the headword. It is worth noting that the lexical resource includes both single-word and multi-word headwords with CEFR level labels. At the moment, multi-word headwords consist of compounds only (e.g. *bela kava*, literally 'white coffee' meaning 'caffè latte'), but there are plans to add phraseology in the future.

5. Future work and Conclusion

This article presented the KUUS 1.0 corpus of textbooks for learning Slovenian as a second and foreign language, the core vocabulary list that was created on the basis of this corpus, and the construction of a lexical resource for Slovenian as a second and foreign language, currently with the vocabulary labeled as A1.

For the next version of the corpus, we aim to also include the workbooks because they contain examples of language use that are very valuable for the work we describe in section 4. The inclusion of workbooks will take place under the umbrella of a project called *Nadgradnja korpusov za slovenščino kot drugi in tuji jezik KOST in KUUS* (Expanding the KOST and KUUS Corpora of Slovenian as a Second and Foreign Language), with the improved version of the corpus available at the end of 2023.

For our next version of the list, we plan to manually review and label the words that were not included in the current list. These mainly consist of candidates for the levels B2 and C1, as well as some for lower levels that did not meet the inclusion criteria. In addition, we aim to confirm the CEFR labels assigned to each word by obtaining a wider consensus from experts that teach Slovenian as a second and foreign language.

The KUUS corpus has proven to be an invaluable resource in the development of a lexical resource for Slovenian as a second and foreign language. It is the first of its kind

for Slovenian. Moving forward, our aim is to expand this corpus-based lexical resource by incorporating vocabulary entries (both single-word and multi-word) for higher proficiency levels. Additionally, we plan to enhance the existing level-A1 explanations by including senses that are relevant to higher proficiency levels. The process of constructing the lexical descriptions, as described in this article, involves manual review and editing of the automatically extracted data. However, the subsequent major step, which involves creating pedagogically tailored sense definitions, will require more input from the authors. A specialized interface for the DDDS (Digital Dictionary of Slovene) is currently under development, which will streamline and enhance the efficiency of all stages of the lexicographic work. Once the lexical resource is published, we intend to evaluate its usability and gather feedback on user experience. This assessment will help identify priorities for further development, ensuring that future enhancements align with user needs and expectations.

6. Acknowledgements

This work was supported by the Slovenian Research Agency (ARRS) via the core programs Language Resources and Technologies for Slovene (P6-0411), Knowledge Technologies (P2-0103), and Slovene Language – Basic, Contrastive, and Applied Studies (P6-0215), and via the projects Empirical Foundations for Digitally Supported Development of Writing Skills (J7-3159), Quantitative and Qualitative Analysis of Unregulated Corporate Financial Reporting (J5-2554), and Computer-Assisted Multilingual News Discourse Analysis with Contextual Embeddings (J6-2581).

The KUUS corpus of textbooks for learning Slovenian as a second and foreign language and the vocabulary lists for levels A1, A2, and B1 were supported by CLARIN.SI. The project Expanding the Teaching Material *Čas za slovenščino 1* in the Digital Environment and Adapting the Material for Teaching Blind and Partially Sighted Adolescents is financially supported by the Slovenian Ministry of Culture.

7. References

Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume (2020). Strasbourg: Council of Europe Publishing. Available at: https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4

Čibej, J., Gantar, K., Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T., Romih, M., Arhar Holdt, Š., Krsnik, L. & Robnik-Šikonja, M. (2022). *Morphological lexicon Sloleks 3.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1745.

Gril, P. (2022). Na tečaj: učenje in poučevanje slovenščine kot drugega in tujega jezika v Sloveniji. In N. Pirih Svetina & I. Ferbežar (eds.) *Na stičišču svetov: slovenščina kot drugi in tuji jezik, Obdobja 41*. Ljubljana: Založba Univerze v Ljubljani, pp. 117–127.

Ferbežar, I., Knez, M., Markovič, A., Pirih Svetina, N., Schlamberger Brezar, M.,

Stabej, M., Tivadar, H. & Zemljarič Miklavčič, J. (2004). *Sporazumevalni prag za slovenščino*. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete, Ministrstvo RS za šolstvo, znanost in šport.

Fišer, D., Ljubešić, N. & Erjavec, T. (2020). The Janes project: language resources and tools for Slovene user generated content. *Lang Resources and Evaluation* 54, pp. 223–246. Available at: https://doi.org/10.1007/s10579-018-9425-z

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1), pp. 7–36.

Knez, M., Ferbežar I., Kern Andoljšek, D. & Stabej M. (2021). *Evalvacija modelov učenja in poučevanja slovenščine kot drugega jezika za učence in dijake, ki jim slovenščina ni materni jezik. Zaključno poročilo*. Ljubljana: Center za slovenščino kot drugi in tuji jezik.

Klemen, M., Arhar Holdt, Š., Pollak, S., Kosem, I., Huber, D. & Lutar, M. (2022). Korpus učbenikov za učenje slovenščine kot drugega in tujega jezika. In N. Pirih Svetina & I. Ferbežar (eds.) *Na stičišču svetov: slovenščina kot drugi in tuji jezik*, *Obdobja 41*. Ljubljana: Založba Univerze v Ljubljani, pp. 165–174.

Klemen, M., Kosem, I., Arhar Holdt, Š., Pollak, S., Huber, D. & Lutar, M. (2022a). *Corpus of textbooks for learning Slovenian as L2 KUUS 1.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1696.

Klemen, M., Arhar Holdt, Š. & Pollak, S. (2022b). *Core vocabulary for Slovenian as L2 1.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1697.

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I. & Dobrovoljc, K. (2020). In N. Calzolari (ed.) *Gigafida 2.0: the reference corpus of written standard Slovene. LREC 2020: Twelfth International Conference on Language Resources and Evaluation*, *Palais du Pharo, Marseille, France.* Paris: ELRA – European Language Resources Association, pp. 3340–3345. Available at: http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf

Krek, S., Laskowski, C., Robnik Šikonja, M., Kosem, I., Arhar Holdt, Š., Gantar, P., Čibej, J., Gorjanc, V., Klemenc, B. & Dobrovoljc, K. (2018). *Thesaurus of Modern Slovene 1.0.*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1166.

Kosem, I., Čibej, J., Dobrovoljc, K., Erjavec, T., Ljubešić, N., Ponikvar, P., Šinkec, M. & Krek, S. (2022). *Monitor corpus of Slovene Trendi 2022-10*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1681.

Kosem, I., Bálint Čeh, J., Ponikvar, P., Zaranšek, P., Kamenšek, U., Koša, P., Gróf, A., Böröcz, N., Harmat Császár, J., Szíjártó, I., Šantak, B., Gantar, P., Krek, S., Roblek, R., Zgaga, K., Logar, U., Pori, E., Arhar Holdt, Š. & Gorjanc, V. (2021). *Comprehensive Slovenian-Hungarian Dictionary 1.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1453.

Kosem, I., Krek, S. & Gantar, P. (2021a). Semantic data should no longer exist in isolation: the Digital Dictionary Database of Slovenian. In Z. Gavriilidou, L.Mitits

& S. Kiosses (eds.), EURALEX XIX, Congress of the European Association for Lexicography, Lexicography for inclusion. Komotini: Democritus University of Thrace, pp. 81–83. Available at: https://euralex2020.gr/wp-content/uploads/2021/09/EURALEX2020_BookOfAbstracts-Preview-1.pdf

Kosem, I., Gantar, P., Krek, S., Arhar Holdt, Š., Čibej, J., Laskowski, C., Pori, E., Klemenc, B., Dobrovoljc, K., Gorjanc, V. & Ljubešić, N. (2019). *Collocations Dictionary of Modern Slovene KSSS 1.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1250.

Kosem, I., Rozman, T., Arhar Holdt, Š., Kocjančič, P., Laskowski, C. A. (2016). Šolar 2.0: nadgradnja korpusa šolskih pisnih izdelkov. In T. Erjavec & D. Fišer (eds.) *Zbornik konference Jezikovne tehnologije in digitalna humanistika*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 95–100. Available at: http://www.sdjt.si/wp/wp-content/uploads/2016/09/JTDH2016_Kosem-et-al_Solar-2-0-nadgradnja-korpusa-solskih-pisnih-izdelkov.pdf

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

Logar, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š. & Krek, S. (2012). *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida inccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina; Fakulteta za družbene vede.

Pirih Svetina, N., Rigler Šilc, K., Lavrič, M., Ferbežar, I. & Jerman, T. (2004). *Preživetvena raven v slovenščini*. Krakov: TAiWPN Universitas.

Pirih Svetina, N. (2016). *Preživetvena raven za slovenščino: za potrebe programa Opismenjevanje v slovenščini za odrasle govorce drugih jezikov*. Ljubljana: Center za slovenščino kot drugi in tuji jezik. Available at: https://centerslo.si/wp-content/uploads/2016/07/IC_Prezivetvena_2016.pdf

Pollak, S., Arhar Holdt, Š., Krek, S., Robnik-Šikonja, M. (2020). *Reference List of Slovene Frequent Common Words*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1346.

Verdonik, D. & Zwitter Vitez, A. (2011). *Slovenski govorni korpus Gos*. Ljubljana: Trojina, zavod za uporabno slovenistiko.

Volodina, E., Alfter, D. & Lindström Tiedemann, T. (2022). Crowdsourcing ratings for single lexical items: a core vocabulary perspective. *Slovenščina 2.0*, 10(2), pp. 5–61.