Mapping Slovene Learner Vocabulary to CEFR Scales with AI-assisted Methods

Mojca Stritar Kučuk

University of Ljubljana, Faculty of Arts, Aškerčeva 2, 1000 Ljubljana
E-mail: mojca.stritarkucuk@ff.uni-lj.si

Abstract

This paper examines how a learner corpus can support lexicographic work by classifying learner vocabulary according to the CEFR scale. Using a corpus-driven methodology, I explore the potential of AI to complement traditional analysis. The study focuses on a selection of texts from the Slovene learner corpus KOST, balanced according to the pragmatically assigned levels of learners' language proficiency: non-Slavic beginners, South Slavic beginners, other Slavic beginners, intermediate and advanced learners. Lemma lists were generated using Sketch Engine and compared with the core vocabulary for Slovene as L2 (up to level B1) and other reference sources. Two advanced language models (ChatGPT and Copilot) were then used to automatically assign CEFR levels to the lemmas. The study compares traditional corpus-derived classifications with AI-generated classifications, evaluates their accuracy and bias, and aims to assess the feasibility of using LLMs in corpus-based CEFR annotation and vocabulary profiling in a lesser-resourced language such as Slovene.

Keywords: CEFR classification; Slovene learner corpus; large language models (LLMs); vocabulary profiling; second language acquisition

1. Introduction

Learner corpora are an established linguistic resource in the field of language acquisition research, but can also be valuable in lexicology (cf. Cobb & Horst, 2015) and lexicography, offering insights into the learners' linguistic production.

This paper examines how a learner corpus can serve as a lexicographical resource by investigating how vocabulary used by learners at different levels of language proficiency can be systematically classified. Several studies have already been conducted in this area, but they mostly focus on non-native English (cf. Ballier et al., 2020; Pitura, 2024) or they do not include AI-assisted techniques (cf. Volodina et al., 2016; Berešova, 2019). In this article, Slovene as a non-native language serves as a case study for less widely spoken and lesser-resourced languages. For Slovene, a generally useful dictionary aimed at non-native speakers has yet to be developed, and the existence of a dictionary indicating the difficulty level of headwords is also uncharted territory. Therefore, I propose a method for identifying the most frequently used lemmas in a Slovene learner corpus and assign them levels according to the well-established Common European

Framework of Reference for Languages (CEFR). It comprises six levels that are suitable for the organisation of language learning and the public recognition of language competence (Svet Evrope, 2011): A1 (Breakthrough Level), A2 (Waystage Level), B1 (Threshold Level), B2 (Vantage Level), C1 (Effective Operational Proficiency Level), C2 (Mastery Level).

In the second step of the research, the inclusion of large language models (ChatGPT/GPT-4 and Copilot) in the determination of CEFR levels is proposed, together with an evaluation of their effectiveness compared to existing methods. This could significantly improve the automation of the process. An important methodological goal was to carry out the whole process using widely available computer programmes, freely available LLMs and, more generally, a methodology that is accessible to researchers with limited facilities and skills in computational linguistics. A noteworthy limitation of the study, however, is that it focuses on lemma-level classification and not (yet) on sense-level classification. This would require dictionary material that does not yet exist, as mentioned previously.

2. Methodology

One of the starting points of the study was the desire to determine the CEFR level at which non-native speakers of Slovene use words at different stages of learning. To determine this, the material from the Slovene learner corpus KOST (Stritar Kučuk, 2024a)[1], a written corpus of Slovene with more than 1.5 million tokens, was used. Since analysing the entire corpus would be too time-consuming, a selection of texts from the corpus was prepared specifically for the purposes of this analysis.

2.1 Selection of texts and the initial lemma list

The main criterion considered when selecting the corpus texts for the study was that they were written by learners without external aids such as machine translation (cf. Stritar Kučuk, 2024b). Although all texts were essays, the second criterion was to select them in such a way that they were thematically diverse. In principle, only one text from one learner was included so that the selection would cover texts from as many different learners and on as many different topics as possible[2].

The final criterion was to balance the selected texts according to the level of language proficiency of the learners. The texts in KOST are categorised into four levels based on the learners self-reported language background, various placement tests, etc.: beginner, South Slavic beginner, intermediate and advanced (Stritar Kučuk, 2024b). Drawing upon my personal teaching experience, I have introduced an additional level in this study, that of Slavic beginner, including beginner texts by speakers of Slavic (but not

[1] https://viri.cjvt.si/kost/sl/ (accessed 19 June 2025).

[2] The list of topics and titles is too extensive to be included in this paper.

South Slavic) languages. The rationale for distinguishing South Slavic from other Slavic languages is that linguistic proximity to Slovene can strongly affect lexical choices and learning trajectories (Klinar et al., 2022). According to classroom experience, at the intermediate and advanced levels, differences between speakers of languages from different language groups become less relevant so distinctions between Slavic and non-Slavic speakers are not necessary anymore. The given sub-corpora classification is of course pragmatic, not necessarily fully reliable and depends on non-linguistic factors such as the number of learners in the particular language course. Taking all these factors into account, I created five sub-corpora for this study, as shown in Table 1. For each sub-corpus, I selected texts from KOST with about 10,000 tokens. The final analysis was thus carried out on a sample of 56,107 tokens, making it a pilot study for research that will be carried out on a larger sample in the future.

Sub-corpus	**No. of tokens**	**No. of learners**	**Learners' L1**
non-Slavic beginner	10,327	74	Dutch, English, French, German, Hungarian, Italian, Romanian, Spanish, Turkish
South Slavic beginner	10,370	57	Bosnian, Croatian, Macedonian, Montenegrin, Serbian
Slavic beginner	10,923	20	Bulgarian, Czech, Polish, Russian, Slovak, Ukrainian
intermediate	12,141	55	Bosnian, Bulgarian, Chinese, Croatian, Czech, Dutch, German, Italian, Macedonian, Polish, Romanian, Russian, Serbian, Slovak, Slovene[3], Spanish, Ukrainian
advanced	12,346	41	Albanian, Chinese, Croatian, Czech, Dutch, German, Hungarian, Italian, Macedonian, Polish, Russian, Serbian, Slovene, Ukrainian

Table 1: Sub-corpora included in the study.

The selected corpus texts were imported into Sketch Engine[4], and the Wordlist function was used to create a comprehensive list of all lemmas used. The data was then exported to Microsoft Excel in tabular form and manually reviewed.

[3] L1 speakers of Slovene living in countries other than Slovenia. When listing the learners' L1s, our approach is to adhere to their self-definition, which they stated in their consent to participate in the corpus.

[4] https://app.sketchengine.eu/ (accessed 19 June 2025).

2.2 Defining the CEFR level manually

In order to determine the level of the individual lemmas on the CEFR scale, I compared the lemma list with the corpus-based list from core vocabulary for Slovene as L2[5], which was compiled on the basis of the KUUS textbook corpus (Klemen et al., 2022)[6] and the reference list of Slovene frequent common words (Arhar Holdt et al., 2020)[7]. The core vocabulary list comprises 5273 unique lemmas, 1214 at A1 level[8], 1451 at A2 level and 2608 at B1 level. Unfortunately, no lemma list has yet been compiled for levels B2 and above, so these lemmas are not the focus in this analysis.

Two additional sources, based on consensual expert group knowledge (Klemen et al., 2023), were used to classify the lemmas along the CEFR scale. The core vocabulary list, for instance, does not contain main numerals, as they do not occur systematically in the KUUS corpus (Klemen et al., 2023). But main numerals are set to level A1 in the document defining the Breakthrough Level (Pirih Svetina, 2016). I have tagged the ordinal numerals up to *šesti* 'sixth' as A1, as they are listed as such in the list of core vocabulary. *Deseti* 'tenth', however, is listed as A2 in the core vocabulary. Following this, I have tagged higher ordinal numerals as A2. At the B1 level, I have included some lemmas that are listed in the document defining the Threshold Level (Ferbežar et al. 2004) but are not included in the core vocabulary.

For lemmas that could not be found in any of the existing lists, I first checked whether there was an error in spelling, morphology or automatic lemmatisation. For example, the lemmatiser defined the possessive adjective *otrokov* 'child's' as the base form for the occurrence *otrokov*, although the context shows that this is an inappropriate inflexion of the noun *otrok* 'child': *Ni treba otroško sobo ker ne želim otrokov* ('I do not need a nursery because I do not wish to have children'). I have corrected errors of this type manually. The remaining lemmas that are not included in the list of core vocabulary have been categorised into various additional categories, which are explained in more detail in section 3.1.

It should be noted that at this stage of the research, the CEFR level is given for each lemma as a headword and not separately for each of its word senses, as is already the case in certain dictionaries for other languages[9]. This limitation is acknowledged and planned for later stages of the project.

[5] http://hdl.handle.net/11356/1697 (accessed 19 June 2025).

[6] http://hdl.handle.net/11356/1877 (accessed 19 June 2025).

[7] http://hdl.handle.net/11356/1346 (accessed 26 June 2025).

[8] In the list of Core Vocabulary, A1-core and A1-larger are differentiated according to the number of textbooks at A1 level in which the word occurs (Klemen et al. 2022). This distinction is not relevant for the present study and the two categories were combined into a single A1 level.

[9] Cf. Oxford Advanced Learner's Dictionary, https://www.oxfordlearnersdictionaries.com/ (accessed 26 June 2025).

For the purposes of this research and to facilitate comparisons with the CEFR grades assigned by AI, I assume that the CEFR grades based on the core vocabulary and other sources mentioned above are valid and are thus considered referential. However, certain words may have been placed differently.

2.3 Defining the CEFR level with AI

To determine how AI assigns CEFR level to lemmas from the initial KOST list, I used several LLMs. The initial hypothesis was that the process would be quick and easy; however, getting the results was more complicated than expected.

In the position of a semi-professional user, I have tried out some LLMs that are available for free: ChatGPT (GPT-4)[10], Copilot[11], Deepseek (Deep Think R1)[12] and Gemini (2.5 Flash)[13]. For each LLM, the same prompt was administered, using the zero-shot model but with further corrections of the output, requiring the assignment of a CEFR level to the words on the list provided. If the word was not recognised as Slovene, it should be tagged as non-Slovene. Useful results were only provided by ChatGPT, Copilot and Deepseek. Gemini explained that it could not assign a grade "because the official CEFR grade requires much more than just recognising individual words. It is about understanding the wider context, grammar, vocabulary as a whole and communication skills"[14]. Instead, only a descriptive definition of how common a particular word is was provided (Figure 1). Of the three remaining LLMs, none would directly tag lemmas in an uploaded Excel spreadsheet. This was particularly surprising in the case of ChatGPT, where a paid version based on the more advanced GPT-5 model was also tested. The file generated by the model was unusable, as the lemma list output contained different lemmas than the original list. The lemmas in the output list matched the lemmas from the core vocabulary list (Table 2), including additional data about them.

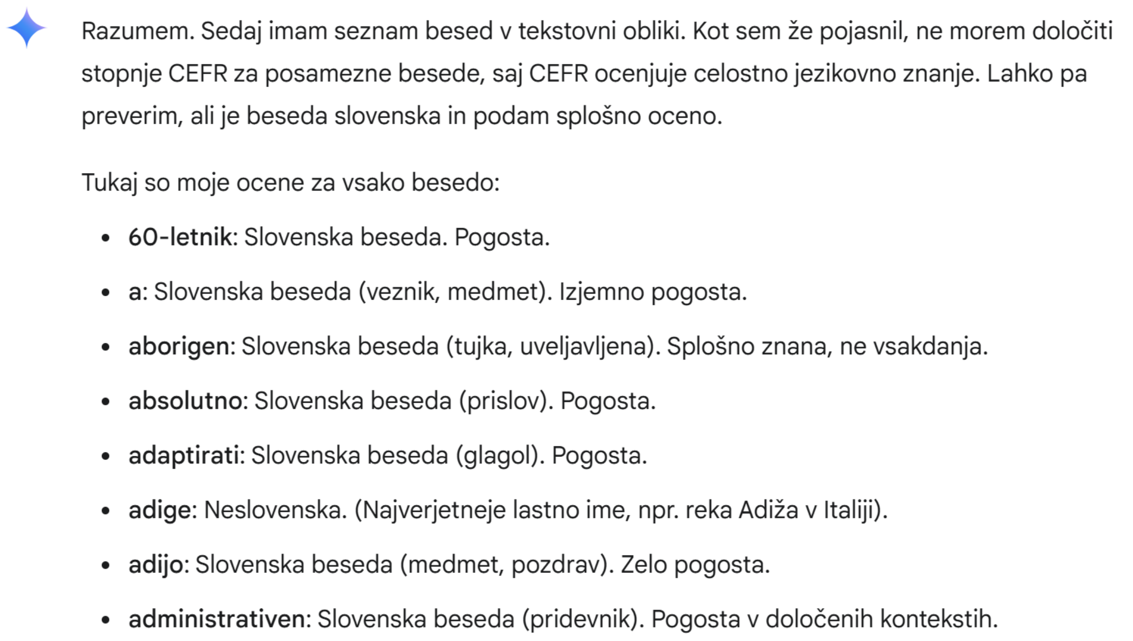
[10] https://chatgpt.com/ (accessed 20 June 2025).

[11] https://copilot.microsoft.com/ (accessed 20 June 2025).

[12] https://chat.deepseek.com/ (accessed 20 June 2025).

[13] https://gemini.google.com/app (accessed 20 June 2025).

[14] https://gemini.google.com/app/d4d9081f115a6d61 (20 June 2025).



Razumem. Sedaj imam seznam besed v tekstovni obliki. Kot sem že pojasnil, ne morem določiti stopnje CEFR za posamezne besede, saj CEFR ocenjuje celostno jezikovno znanje. Lahko pa preverim, ali je beseda slovenska in podam splošno oceno.

Tukaj so moje ocene za vsako besedo:

- **60-letnik**: Slovenska beseda. Pogosta.
- **a**: Slovenska beseda (veznik, medmet). Izjemno pogosta.
- **aborigen**: Slovenska beseda (tujka, uveljavljena). Splošno znana, ne vsakdanja.
- **absolutno**: Slovenska beseda (prislov). Pogosta.
- **adaptirati**: Slovenska beseda (glagol). Pogosta.
- **adige**: Neslovenska. (Najverjetneje lastno ime, npr. reka Adiža v Italiji).
- **adijo**: Slovenska beseda (medmet, pozdrav). Zelo pogosta.
- **administrativen**: Slovenska beseda (pridevnik). Pogosta v določenih kontekstih.

Figure 1: Beginning of the answer from Gemini.

KOST selection	ChatGPT	Core Vocabulary
60-letnik	*a*	*a*
a	*a*	*a*
aborigen	*abeceda*	*abeceda*
absolutno	*abonma*	*abonma*
adaptirati	*absolutno*	*absolutno*
adige	*abstinenčen*	*abstinenčen*
adijo	*abstrakten*	*abstrakten*
administrativen	*adijo*	*adijo*
afera	*adjektiv*	*adjektiv*
afrika	*adrenalinski*	*adrenalinski*

Table 2: The first lemmas from the KOST selection, from their initial ChatGPT processing and from the core vocabulary.

The lemmas thus had to be manually entered into the chatbot, just as the levels assigned to them by the LLMs had to be manually entered into the Excel spreadsheet. Although API-based processing would be more scalable and reproducible, this manual work was motivated by the goal of testing what is feasible for an average user without programming knowledge. Since the process was relatively time-consuming, I decided to compare the CEFR level for the first 1,000 lemmas on the KOST list, which was organised alphabetically from *60-letnik* to *itn.*, and to use only two LLMs, ChatGPT and Copilot.

The basis on which the LLMs evaluate words is not fully known. However, to answer this question, ChatGPT has provided the following decision criteria: semantic frequency and usefulness (part of the basic vocabulary or a more specific word), abstractness and semantic complexity, formal complexity, thematic domain (everyday topics or specialist areas), how frequently it occurs in teaching materials, and the form and origin of the word (foreign or adopted word or an incorrect form)[15]. Copilot's assessment is based on the frequency, abstractness, complexity of meaning or form and the occurrence of the word in teaching materials for Slovene as a second/foreign language[16].

3. Results and discussion

3.1 The lemma list

The final list from KOST comprises 4,423 lemmas, which are distributed fairly evenly across the levels: 951 at A1, 780 at A2, 780 at B1 and 830 beyond B1 (Figure 2). Among the words that are not included in the core vocabulary list and are thus most likely at a level beyond B1, I included 23 colloquial words that were manually identified as common in spoken but not standard Slovene (*cajt* 'time*, štempelj* 'brand, impress'). 557 lemmas cannot be found in any list of Slovene words and are thus referred to as "non-Slovene" in this paper. Among them there are unadapted loanwords, cases of code-switching or words in the learners' L1 or other languages that are not Slovene[17]. In the "other" category which covers 525 lemmas, I included abbreviations (*FRI, SOS;* 13 in the list) and proper names (512 in the list), whether personal (*Andrej, Bobby*), geographical (*Amerika*) or names of institutions, brands, etc. (*Snapchat*). Both categories are not included in the aforementioned data sources (Klemen et al., 2022). However, in the Breakthrough Level it is assumed that the speaker knows the name of their country in Slovene (Pirih Svetina, 2016). Due to the specifics of the corpus method, which does not allow to automatically recognize whether a geographical name refers to the learner's country or not at every occurrence, I excluded all proper names and abbreviations from further analysis and only kept those non-Slovene lemmas where the learners should have used a common Slovene word.

[15] https://chatgpt.com/share/685e336a-d54c-8010-b91f-da1ca41cb4e7 (accessed 27 June 2025).

[16] https://copilot.microsoft.com/shares/Mpt6xZux6UAwgDffrn9pr (accessed 27 June 2025).

[17] Similarly, Copilot tags words as non-Slovene lemmas when they are words from other languages, misspellings, proper names, unadapted foreign words, unusual forms, and dialectal or rare forms that are not recorded in official sources or that are not used in language corpora, cf. https://copilot.microsoft.com/shares/MjQeuudcjkYaz2jHwh7SG (accessed 27 June 2025).

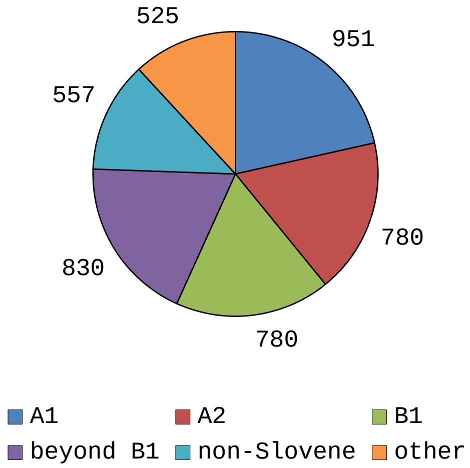


Figure 2: Number of lemmas from the KOST list according to CEFR levels.

Focusing on the lemma list according to the initial proficiency level of the learners, it can be seen that the largest number of different lemmas at A1 level was used by non-Slavic beginners and the smallest by advanced learners (Figure 3). Slavic and South Slavic beginners also used a considerable number of A2 and even B1 lemmas based on the close linguistic relationship between Slovene and their L1. At the higher levels, however, intermediate and advanced learners used a much greater variety of lemmas. Of course, the sub-corpora analysed are too small for the analysis not to reveal differences based on specific writing topics, but it is nevertheless evident that the vocabulary of higher-level speakers is at much higher CEFR levels and thus more abstract than that of lower-level speakers. Key differences between non-Slavic and Slavic beginners are also clearly recognizable, with (South) Slavic beginners using significantly more lemmas at higher levels than non-Slavic beginners. It is also interesting to note that the proportion of non-Slovene words used by learners at the lowest level is relatively low and comparable to that of advanced learners. Obviously, the use of non-Slovene words poses a greater problem for speakers of languages more closely related to Slovene or for advanced learners who are more confident in their language use but consequently exercise less control over their lexical choices.

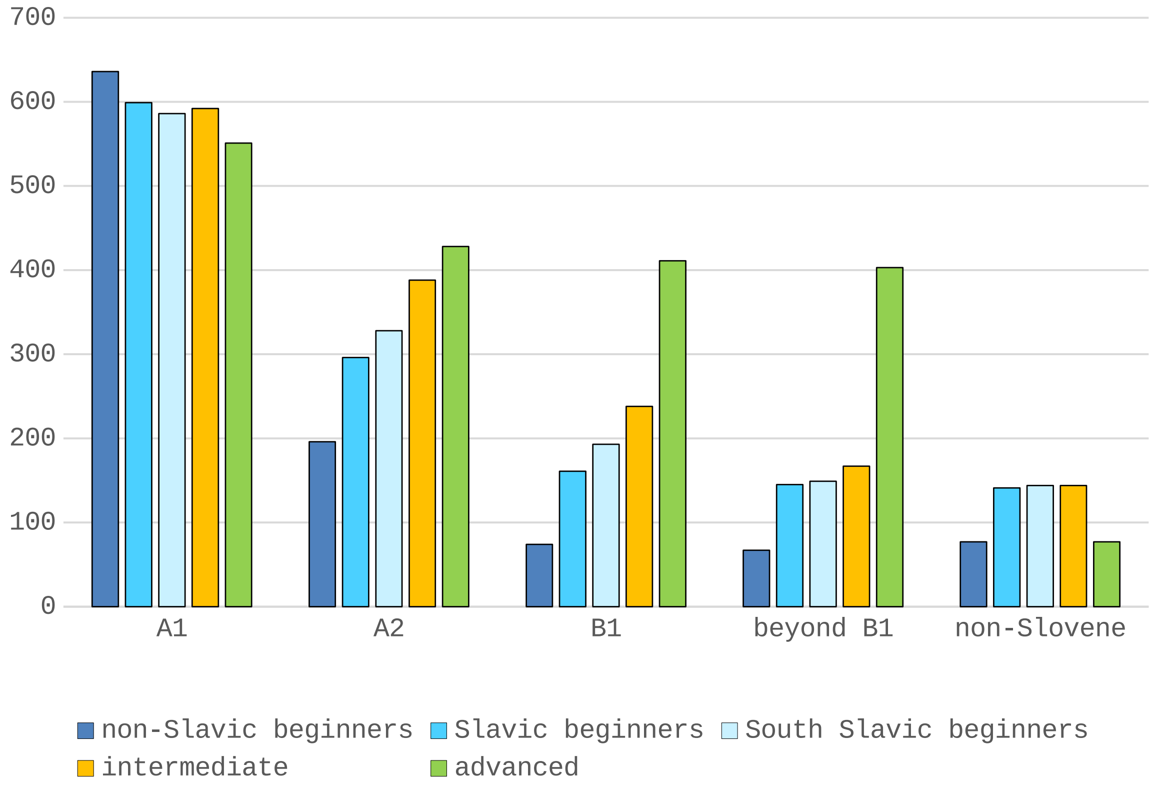


Figure 3: Number of lemmas depending on the CEFR level and the learners' language proficiency.

3.2 The AI determination of the CEFR level

The following section compares the results of the manual and AI determination of the CEFR level for the first 1000 lemmas in the list, excluding proper names and abbreviations. The results were compared for 795 lemmas. As can be seen in Figure 4, almost twice as many lemmas were manually determined at A1 level and the manual results were also higher for non-Slovene lemmas. LLMs identified more lemmas at other levels, i.e. A2, B1 and beyond B1, with Copilot particularly standing out at the levels beyond B1.

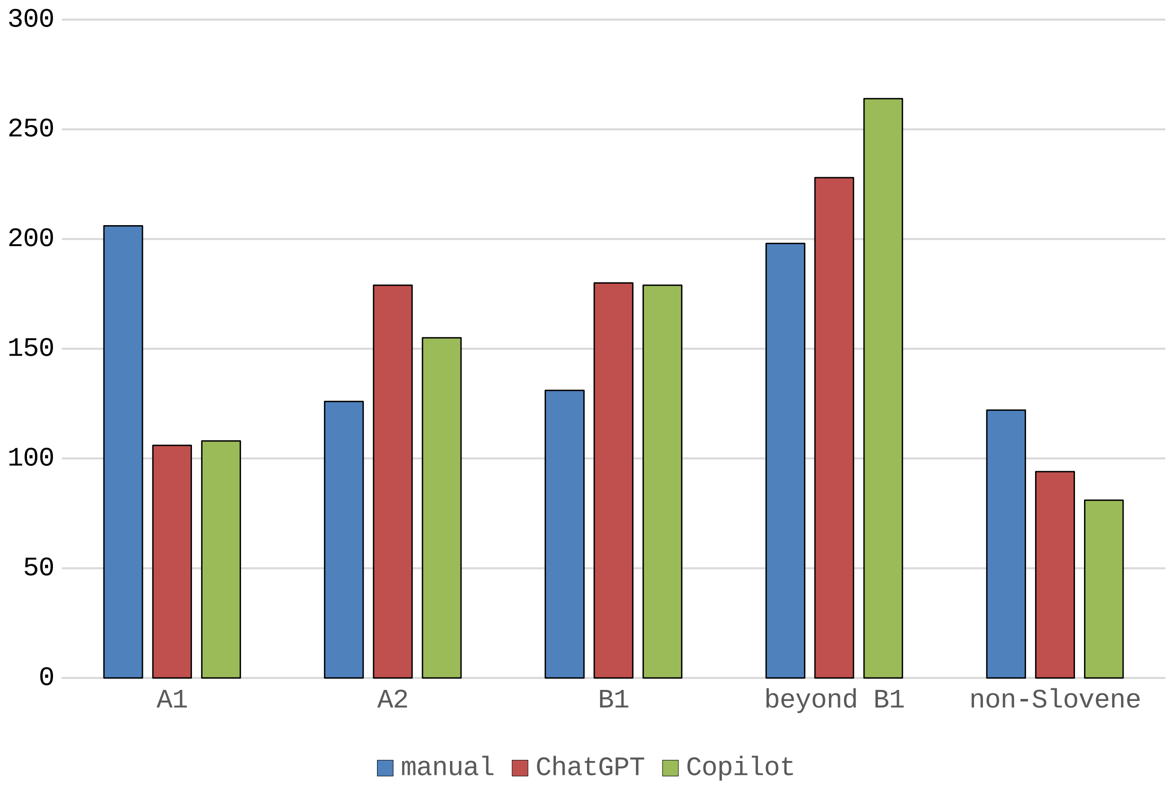


Figure 4: Number of lemmas at different CEFR levels according to different methods of determination.

Although the two LLMs tested did not deliver identical results, they nevertheless showed comparable performance. Compared to the manual method, ChatGPT identified 389 out of 795 lemmas equally, i.e. 48.9% successfully, while Copilot identified 381 lemmas, i.e. 47.9% successfully. If we exclude the categorizations for levels beyond B1, which are not accurate in the manual determination, ChatGPT even had a success rate of 66.6% and Copilot of 59.8%. Figures 5 and 6 show that both LLMs work best for levels beyond B1, but it should be noted that manual categorization puts all words beyond B1 into one category, while both LLMs separate them. Therefore, it is difficult to assess their success rate at these levels. Ignoring these levels, ChatGPT showed the greatest proficiency in identifying words at A1 and A2 levels, while Copilot was the most successful at A1 level.

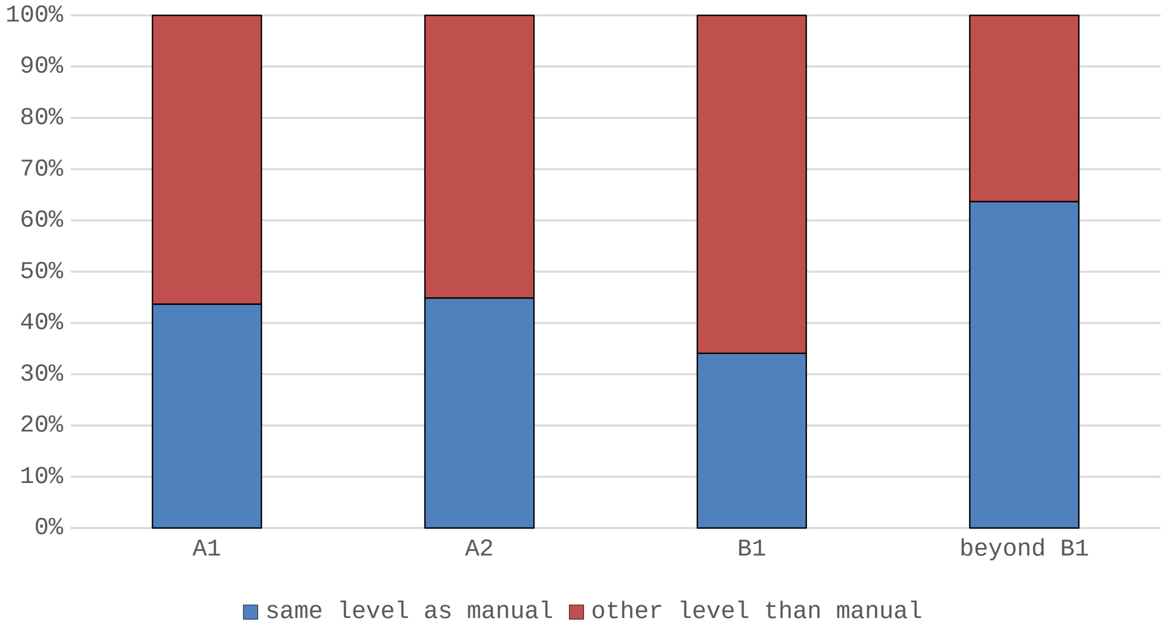


Figure 5: Performance of ChatGPT in determining words at different CEFR levels.

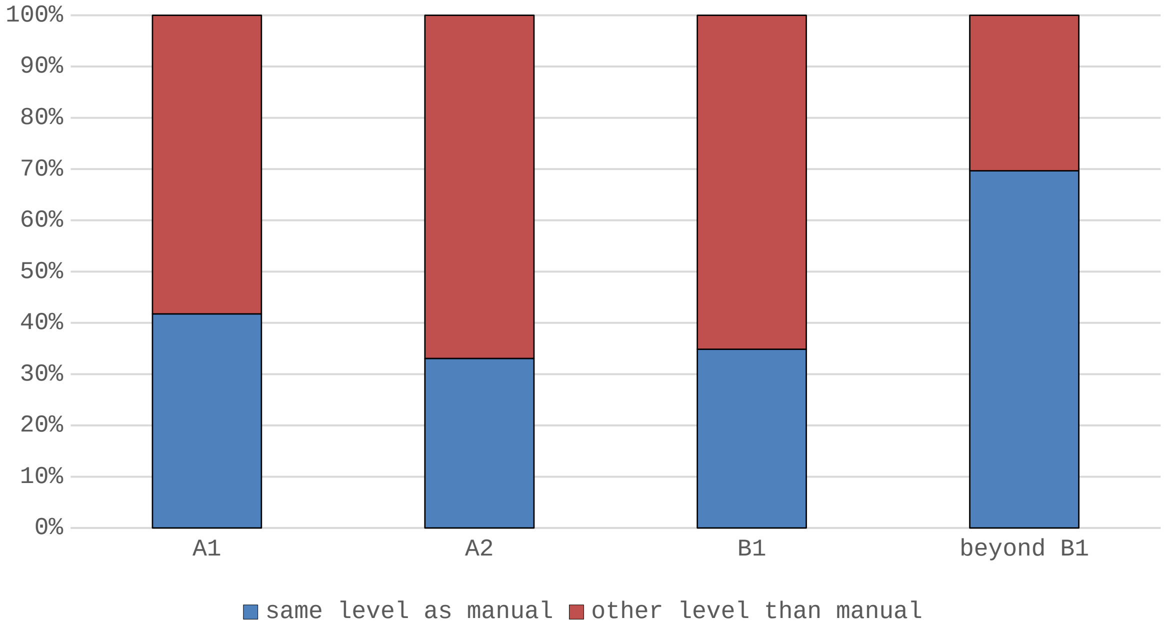


Figure 6: Performance of Copilot in determining words at different CEFR levels.

Let us now take a closer look at the manual lemma ratings and those determined by AI. To make disagreement patterns more transparent, confusion matrices were added (Figures 7 and 8). Almost 37% of all lemmas were determined at an identical CEFR level by all three methods. 23% of all lemmas were given an identical level by both LLMs, which differed from the manual level, and for 28% of the lemmas, the classifications were different for the three methods.

		Manual grade (predicted)				
		A1	**A2**	**B1**	**beyond B1**	**non-Slovene**
ChatGPT (actual)	**A1**	90	6	7	1	2
	A2	74	56	30	12	7
	B1	33	48	44	46	8
	beyond B1	5	14	50	128	31
	non-Slovene	5	2	2	14	73

Figure 7: Confusion matrix for the grades by ChatGPT vs. manual grades.

		Manual grade (predicted)				
		A1	**A2**	**B1**	**beyond B1**	**non-Slovene**
Copilot (actual)	**A1**	86	10	8	3	1
	A2	70	41	23	13	6
	B1	36	53	46	36	7
	beyond B1	11	21	55	125	37
	non-Slovene	1	1	0	9	70

Figure 8: Confusion matrix for the grades by Copilot vs. manual grades.

Let us concentrate on the lemmas that were identified at an identical level by all three methods. Most of them are at levels beyond B1 (e.g. *afera* 'affair', *adaptirati* 'to adapt', *bodisi* 'either, whether'). This is followed by A1 lemmas (e.g. *biti* 'to be', *danes* 'today, *adijo* 'bye'). A2 lemmas include words such as *avtomobil* 'car', *čistiti* 'to clean', *dovolj* 'enough', and B1 lemmas include *barven* 'colored', *dejansko* 'actually', *dokler* 'until, while'. The lemmas that were recognized as non-Slovene by all three methods include *angliško, bacal, bažič, bilala.*

In 140 cases, both LLMs consistently rated the lemmas at a higher level than manually. The difference is usually only one level: A2 instead of A1 (e.g. *do* 'to, until', *dobiti* 'to get', *dopoldne* 'morning, before noon'), B1 instead of A2 (e.g. *avtor* 'author', *bližati* 'to approach', *človeški* 'human') and B2 instead of B1 (e.g. *biološki* 'biological', *brezposelnost* 'unemployment', *drveti* 'to race, rush'). Less frequently, there is a difference of two or more levels: B1 instead of A1 (e.g. *dvigniti* 'to raise', *fakulteta* 'faculty', *hoteti* 'to want'), B2 instead of A1 (e.g. *elektronski* 'electronic', *enoposteljen* 'single-bed') and B2 instead of A2 (e.g. *arhitektura* 'architecture', *čim* 'as soon as, the').

Less frequently, in 53 cases, both LLMs uniformly rated the lemmas at a lower level than manually. Here too, the difference is usually only one level: A1 instead of A2 (e.g. *ananas* 'pineapple', *božič* 'Christmas', *gor* 'up'), A2 instead of B1 (*cimer* ‘roommate’*,* *cvetje* ‘flowers’*,* *dijakinja* ‘high school student*,* *disko* ‘disco’) and B1 instead of beyond B1 (*60-letnik* ’60-year old’*,* *atmosfera* ‘atmosphere’*,* *biljard* ‘snooker’*,* *borov* ‘pine’, adj.). Less frequently, there is a difference of two or more levels: A2 instead of beyond B1 (*dedkov* ‘grandfather’s’*,* *gramatika* ‘grammar’*,* *hecati* ‘to joke’) and A1 instead of B1 (*babi* ‘granny’*,* *bombon* ‘candy’*,* *deklica* ‘girl’). Here, it is worth highlighting the nouns *baba* and *hči.* The first is an informal, pejorative word for ‘woman’ and is avoided in standard Slovene, while the noun *hči* ‘daughter’ is unmarked in standard Slovene, but is avoided in teaching non-native Slovene due to its irregular paradigm; instead, the regular noun *hčerka* is taught. Thus, both lemmas are obviously not considered A1 words in Slovene linguistics.

Both LLMs have also uniformly defined 25 words that were identified as non-Slovene during the manual identification as Slovene. One was placed at A1 level (the non-Slovene, South Slavic infinitive form *imati* ‘to have’), 4 at A2 (*amerikanec, babicin, dovoljno, gresti*), 3 at B2 (*amerikanski, bežigradski, greti*) and 17 at B1 (e.g. *birokratija, bližni, bratanec, brezdvomno, donesti*), which shows the tendency of LLMs to automatically place less frequent words at a higher level.

In 264 cases, the two LLM assessments were not compatible. No clear pattern can be found here, but almost all A2 and B1-level lemmas were tagged higher by the LLMs, rather than lower. The lemma *gps* is interesting; while it was manually categorized as beyond B1, ChatGPT placed it at B1 and Copilot even at A1.

The two LLMs were quite inconsistent with the words that were manually rated as non-Slovene. ChatGPT identified 14 such words as Slovene and placed them at levels A1 (*bulgarija),* A2 (*blizo),* B1 (*blizina, brež*), B2 (*dolgočasan, evropejski, ispravni*) and C1 (*aborigen, brezposlednji*). Copilot placed 11 such words at levels A2 (*enindvadeset*), B1 (*afrikanski*), B2 (*den, dogotek*), C1 (*anglicistika, besednik*). On the other hand, ChatGPT tagged several Slovene lemmas as non-Slovene: among A1-level lemmas, such are *avtobusen* ‘bus’, adj., and *dneven* ‘daily’, among A2-level lemmas *fin* ‘fine, delicate’, and among beyond-B1-level lemmas such are *artist*, *češ* ‘as if to say’, *daljina* ‘distance’, *fantazijski* ‘fantasy’, adj.

Among the words that were manually categorized as beyond B1, a large group should be highlighted. These are international words, mostly of Latin or Greek origin, whose form and meaning are similar in most European languages and therefore pose only minimal comprehension difficulties for learners (Uni, 2019), such as *agresiven* ‘aggressive’, *alfa*, *aplikacija* ‘application’. Both LLMs placed them mostly at B2 level, and to a lesser extent at C1 and B1. This high ranking of these words is probably due to the fact that they are mainly used to name abstract notions and concepts that also require certain cognitive skills on the part of the language user. However, it seems that once the user has acquired these concepts in their L1, their transfer to other languages does not pose major difficulties, and the user generally understands them long before their general language proficiency in the non-native language reaches a higher level.

Finally, let us examine the words that have been manually annotated as colloquial expressions and are not included in the list of core vocabulary. The colloquial word *gužva* ‘crowd’ was rated as non-Slovene by both LLMs, *bus* was rated as non-Slovene by ChatGPT and assigned an A1 rating by Copilot, while the phonetically spelled *influenser* was categorized as non-Slovene by Copilot and B2 by ChatGPT. The noun *ata* ‘dad’ was graded A1 and A2, the adjective *blond* and the common colloquial adverb *ful* ‘very, a lot’ were graded B1 and A2, respectively, and the common colloquial noun *cajt* ‘time’ was graded B2.

4. Conclusion

This paper proposes a corpus-driven approach to analysing the vocabulary used by non-native speakers in their texts, combined with the use of LLMs in the assessment of CEFR level. To this end, I tested two free LLMs, both of which performed relatively well, achieving a satisfactory score on approximately two thirds of the lemmas. In those cases where their results do not match the grading based on referential, professionally authored linguist sources, they tend to score higher than the classical assessment.

Some of the ratings where the LLMs deviate from the manual ratings are undoubtedly wrong (e.g. the irregular noun *hči* placed at A1), but on the other hand, for some lemmas it is questionable whether they are adequately rated in the referential sources (e.g. the lemma *gps* at a level beyond B1). Both LLMs proved to have difficulties in identifying non-Slovene lemmas, and often interpreted them as lemmas at higher levels of linguistic ability (mostly B2). However, it is difficult to interpret all the results comprehensively without a full understanding of the methods used by the LLMs to determine the lemma rating in the first place. The categorization of words is not in itself language-independent; for example, the Slovene word *hrbtenica* ‘spine’ is rated as B1 (A2 or B1 by LLMs), while the English word *spine* is set at C1 in the OALD dictionary[18].

The results of the test presented here for Slovene as a non-native language are therefore only partially satisfactory. The procedure itself was relatively time-consuming and required a great deal of manual effort and revision. However, I deliberately used an elementary methodology that is accessible to any linguist without special computer skills to see how modern AI methods are accessible to the average user. It is likely that a researcher with greater expertise in the field of computational linguistics and AI, using API-processing or similar, would have achieved better results in a comparable evaluation.

Learner corpora such as KOST can be an important source of data on the actual language production of the learners. The sample size in this analysis is of course too limited for generating a comprehensive educational vocabulary list for Slovene. Vocabulary usage in learner essays is likely influenced by specific task types and topics

[18] https://www.oxfordlearnersdictionaries.com/definition/english/spine?q=spine (accessed 26 June 2025).

so a different selection of corpus texts would result in a different set of lemmas. But when including these lemmas in the material, attention should also be paid to the usage data and the focus should be on lemmas with a sufficiently high frequency of use. The extraction of lemmas from texts written by non-native speakers of Slovene can be a valuable addition to existing language learning materials, such as dictionaries for non-native speakers like SLOGOST[19], which focuses exclusively on A1 level words. Of course, caution should be exercised with this approach, as the CEFR level cannot be determined solely on the basis of learners' written work or the lemmas used in it. Nevertheless, this is an aspect of language use that can be easily quantified.

Furthermore, mere lemma data without context can quickly be misleading. For example, the lemma *greti* 'to heat' in the lemma set presented in this paper was used by two speakers of South Slavic languages. Only from the context (*Kak je čas grel naprej sam se upoznal s več ljudmi sa fakultete* 'Only as the time went on I met more people from the faculty') and knowing the typical errors of these speakers, it can be deduced that it is not actually the lemma *greti*, but that the learner generalised the present stem of the verb *iti* 'to go', *grem*, to its past participial form, which was then recognised by the LLMs as the infinitive of the wrong verb.

5. References

Arhar Holdt, Š., Pollak, S., Robnik-Šikonja, M. & Krek, S. (2020). Referenčni seznam pogostih splošnih besed za slovenščino. *Konferenca Jezikovne tehnologije in digitalna humanistika, Ljubljana*, pp. 10–15. Available at: http://nl.ijs.si/jtdh20/pdf/JT-DH_2020_Arhar-Holdt-et-al_Referencni-seznam-pogostih-splosnih-besed-za-slovenscino.pdf.

Ballier, N., Canu, S., Petitjean, C., Gasso, G., Balhana, C., Alexopolou, T. & Gaillat, T. (2020). Machine learning for learner English. *International Journal of Learner Corpus Research*, 6(1), pp. 72–103.

Berešova, J. (2019). Assigning reference levels to the meanings of words. *EDULEARN19 Proceedings*, pp. 1780–1785. Available at: https://doi.org/10.21125/edulearn.2019.0512.

Cobb, T. & Horst, M. (2015). Learner corpora and lexis. *The Cambridge Handbook of Learner Corpus Research.* Cambridge University Press, pp. 185–206.

Ferbežar, I., Knez, M., Markovič, A., Pirih Svetina, N., Schlamberger Brezar, M., Stabej, M., Tivadar, H. & Zemljarič Miklavčič, J. (2004). *Sporazumevalni prag za slovenščino 2004.* Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete Univerze v Ljubljani, Ministrstvo RS za šolstvo, znanost in šport.

Klemen, M., Arhar Holdt, Š., Pollak, S., Kosem, I., Huber, D. & Lutar, M. (2022). Korpus učbenikov za učenje slovenščine kot drugega in tujega jezika. *Na stičišču svetov: slovenščina kot drugi in tuji jezik. Obdobja 41.* Ljubljana: Založba Univerze

[19] https://lexonomy.cjvt.si/slovar-za-govorce-slovenscine-kot-drugega-in-tujega-jezika/ (accessed 24 June 2025).

v Ljubljani, pp. 165–174. Available at: https://doi.org/DOI:10.4312/Obdobja.41.165-174.

Klemen, M., Arhar Holdt, Š., Pollak, S., Kosem, I., Pori, E., Gantar, P. & Knez, M. (2023). Building a CEFR-Labeled Core Vocabulary and Developing a Lexical Resource for Slovenian as a Second and Foreign Language. *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference.* Brno: Lexical Computing CZ s.r.o., pp. 664–678.

Klinar, M., Pisek, S., Stritar Kučuk, M. & Šter, H. (2022). Poučevanje slovenščine za redno vpisane tuje študente na Univerzi v Ljubljani. *Na stičišču svetov: slovenščina kot drugi in tuji jezik: Obdobja 41.* Ljubljana: Založba Univerze v Ljubljani, pp. 185–194.

Pirih Svetina, N. (2016). *Preživetvena raven za slovenščino: Za potrebe programa opismenjevanje v slovenščini za odrasle govorce drugih jezikov.* Univerza v Mariboru, Filozofska fakulteta, Center za slovenščino kot drugi in tuji jezik. Available at: https://centerslo.si/knjige/ucbeniki-in-prirocniki/prirocniki-in-ucno-gradivo/prezivetvena-raven-za-slovenscino/.

Pitura, J. (2024). Enhancing advanced vocabulary in EFL writing: an AI-assisted intervention for English studies students in Poland. *Journal of China Computer-Assisted Language Learning.* Available at: https://doi.org/10.1515/jccall-2024-0014.

Stritar Kučuk, M. (2024a). KOST 2.0: Predstavitev korpusa in potek označevanja jezikovnih napak. *Zbornik konference Jezikovne tehnologije in digitalna humanistika.* Ljubljana, pp. 589–603. Available at: https://doi.org/10.5281/zenodo.13912515.

Stritar Kučuk, M. (2024b). Investigating the Usage of Machine Translation in L2 Learning and Its Impact on Writing Proficiency. *Lidil* 70). Available at: http://dx.doi.org/10.4000/12lmd.

Stritar Kučuk, M. (2024c). Prvi korpus slovenščine kot tujega jezika KOST 1.0. *Razvoj slovenščine v digitalnem okolju.* Ljubljana: Založba Univerze v Ljubljani, pp. 93–117. Available at: https://doi.org/10.4312/9789612972561.

Svet Evrope (2011). *Skupni evropski jezikovni okvir: Učenje, poučevanje, ocenjevanje.* Ljubljana: Ministrstvo RS za šolstvo in šport, Urad za razvoj šolstva.

Uni, K. (2019). Benefits of Vocabulary of Latin Origin for the Learners of Swedish and Danish. *The Journal of Social Sciences Research.* Available at: https://doi.org/10.32861/JSSR.52.431.435.

Volodina, E., Pilán, I., Llozhi, L., Degryse, B. & François, T. (2016). SweLLex: Second language learners' productive vocabulary. *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC 2016*, *Linköping Electronic Conference Proceedings*, (130), pp. 76–84.