Validation of Proper Noun Lexemes in the *Digital Dictionary Database of Slovene* Using Wikidata and Large Language Models

Jaka Čibej
jaka.cibej@ijs.si
Jožef Stefan Institute
University of Ljubljana, Centre for Language Resources and Technologies
University of Ljubljana, Faculty of Arts
Ljubljana, Slovenia

Abstract

The *Digital Dictionary Database of Slovene* [7] (DDDS) is a modern digital lexicographic resource designed as a relational database. It is continuously expanded by extracting candidates from automatically tagged text corpora. The lemmatization and tagging errors inevitably present in these corpora were usually manually identified/corrected by human annotators before being included, but the advent of large language models has provided an additional tool that can help reduce the need for manual labor of lexicographers. We present an experiment on the identification and correction of inadequate proper noun lexemes in DDDS by using existing resources and knowledge bases (such as *Wikidata* and *Wikipedia*) and large language models. The results demonstrate that even zero-shot or contextless prompting can provide useful corrections and significantly reduce the amount of routine work.

Keywords

lexemes, proper nouns, Digital Dictionary Database of Slovene, Wikidata, large language models

1 Introduction

The *Digital Dictionary Database of Slovene* (DDDS) [7] is the largest open-access[1] database with machine-readable information on Slovene words, including information on their inflected forms[2] (e.g., *Slovenija* as the lemma and *Slovenija*, *Slovenije*, *Sloveniji*, ... as its inflected forms).[3] As a modern digital lexicographic resource, it is designed as a PostgreSQL[4] relational database. The data have multiple uses: aside from its obvious benefit as a language resource for human users (e.g., to obtain information on how to inflect/pronounce Slovene words and what senses they have) or as a source of examples for other language resources (such as the *Slovenian Normative Guide 8.0*)[5], the database has also been used to improve lemmatization and morphosyntactic tagging models for Slovene corpora [9][10] or to compile datasets for fine-tuning large language models for linguistic and lexicographic tasks [5][6]. In addition, DDDS can be used as a resource for RAG systems and agentic language models. As a fully machine-readable database, it can be linked to existing digital language resources and knowledge graphs (such as *Wikidata* [11]).

DDDS is continuously expanded with new candidates from frequency lists of lemmas extracted from corpora (such as the *Gigafida Corpus of Written Standard Slovene* [8] or the *Trendi Monitor Corpus of Slovene* [6]). However, such lists contain errors due to several factors. First, the lemmatization and morphosyntactic tagging models for standard Slovene perform well, but are still not entirely accurate (F1≈99.11 for lemmatization [10] and F1≈98.27 for morphosyntactic annotations [9]) and errors frequently occur with lexemes not yet included in DDDS [2], resulting in mislemmatized forms (e.g., *Polzelane* instead of *Polzelan* (demonym for a (male) inhabitant of *Polzela*, a settlement in Slovenia). Second, the texts in corpora contain human-made typographical errors (e.g., **Mancehster*, **Mancester*, **Manchaster*, **Manchetser*, **Manchster* → *Manchester*) and misspelled non-standard forms (such as (**Funlandija* instead of *Fundlandija*; from *Nova Fundlandija* 'Newfoundland').

In previous updates to the DDDS, a linguist (or group thereof) would manually validate the extracted candidates by observing their occurrences in corpora and consulting online resources (such as Wikipedia or other encyclopedic resources), especially when the lexeme is unfamiliar and/or could potentially be lemmatized in different ways (for instance, *Malenov* could be a possessive adjective derived from *Malen*; a legitimate masculine proper noun *Malenov*; a mislemmatized feminine proper noun *Malenova*; and so on).

The latest major update within the *Development of Slovene in a Digital Environment* (RSDO) project [1] expanded the database with approximately 220,000 proper noun lexemes extracted from *Gigafida 2.0* [8] (along with approximately 45,000 candidates from other part-of-speech categories). Due to time and resource constraints, the proper noun lexemes were not manually validated (only filtered to exclude the ones with a frequency below 5 to remove outliers). Their inflected forms were automatically generated with the assumption that the majority of candidates were still correctly lemmatized considering the performance of the lemmatization model. Although other parts-of-speech also

[1] The DDDS is available as an SQL dump (the latest version is published at https://wiki.cjvt.si/books/digital-dictionary-database-of-slovene/page/domain-and-data-model-links) or through a public API documented at *CJVT Wiki*: https://wiki.cjvt.si/books/digital-dictionary-database-of-slovene/chapter/rest-api

[2] Slovene is a morphologically rich Slavic language which features inflected word forms with affixes.

[3] The *Sloleks Morphological Lexicon of Slovene* [3], which until now has been presented as the largest open-access database on Slovene morphological data, has been integrated into DDDS as its morphological basis. It is no longer developed separately and exists only as an export of the morphological part of DDDS.

[4] PostgreSQL: https://www.postgresql.org/

[5] *Pravopis 8.0* (*Slovenian Normative Guide 8.0*): https://pravopis8.fran.si/

[6] For instance, the data from DDDS was used to fine-tune the *GaMS-9B-Instruct-Lex* and *GaMS3-12B-Instruct-Lex* models from *GaMS-9B-Instruct* and *GaMS3-12B-Instruct*, respectively. All are available at HuggingFace.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2026.sikdd.18

Table 1: Proper noun lexemes in DDDS by status and grammatical gender (masculine, feminine, neuter).

Gender → Status ↓	M	F	N
Automatic	183,862	38,725	691
Manual	5,525	4,668	159

contain errors, we focus on proper noun lexemes in this paper, as they are currently the largest group with the least previously done manual work (and thus the segment that most likely contains large amounts of errors).

Proper noun lexemes in DDDS contain personal names and surnames (*Deborra*, *Zeljković*), demonyms (*Filipinec*, *Ukrajinka*), names of companies (*Microsoft*, *Agrosoršak*), locations (*Gainesville*, *Catanzaro*), and so on. Any errors in lemmas and their morphosyntactic features result in further errors downstream when inflected forms are generated using *Pregibalnik*,[7] a custom-built tool for Slovene lexicon expansion. For instance, an incorrect lexeme masculine proper noun *Gruzijki* (which should be correctly lemmatized as *Gruzijka* 'female inhabitant of Georgia' and tagged as feminine) is incorrectly inflected as **Gruzijki*, **Gruzijkija*, **Gruzijkiju*, ... (as opposed to *Gruzijka*, *Gruzijke*, *Gruzijki*, ...).

The first step toward correcting these errors is to identify the mislemmatized candidates. We present an experiment on validating proper noun lexemes in DDDS using both existing resources as well as large language models. The main goal is to test whether the method can successfully complement the work of lexicographers and to what extent LLMs can be implemented in future updates of DDDS to ensure clean data.

2 Proper Noun Lexemes in the *Digital Dictionary Database of Slovene*

As of the time of writing this paper (June 2026), the *DDDS* contains a total of 233,630 proper noun lexemes.[8] 10,352 (4.4%) were validated manually (the validation included lemmas, morphosyntactic tags, and their inflected forms), while 223,278 were added entirely automatically from the *Gigafida 2.0 Corpus of Written Standard Slovene* [8]. Table 1 shows the distribution of lexemes by status and grammatical gender.[9]

There is a disproportionately high number of masculine lexemes among the automatically added units, which reflects a frequently occurring error of the lemmatization model: the tendency to interpret inflected forms of out-of-vocabulary units as basic forms, along with a bias to tag them as masculine nouns. Multiple candidates should in fact be converged into a single lexeme (e.g., **Ukrajinkama* (feminine), **Ukrajink* (masculine), **Ukrajinko* (masculine) → *Ukrajinka* (feminine); **Filipince* (masculine), **Filipincem* (masculine), **Filipincev* (masculine) → *Filipinec* (masculine)).

It should be noted that in DDDS, some proper noun lexemes (such as surnames) can be both masculine and feminine because they are inflected differently depending on grammatical gender. In such cases, duplicate lexemes are generated (e.g., *Novak* (masculine) and *Novak* (feminine)).

In the following sections, we perform several cross-comparisons with other available resources to first determine whether the lemmas of automatically added candidates are legitimate.

3 Validation with Existing Resources

First, since the automatically added proper noun lexemes were extracted from Slovene texts from *Gigafida 2.0* [8] and contain large amounts of Slovene names and surnames, we cross-compare them with the lists of names and surnames obtained from the *Statistical Office of the Republic of Slovenia* (SURS)[10]. The SURS lists contain the majority of names and surnames used in Slovenia (in their basic forms). Female and male names are listed separately. This comparison resulted in 2,948 lexemes validated by the list of male names, 2,828 confirmed by the list of female names, and 23,141 validated by the list of surnames.[11]

At this stage, we considered further validating lexemes with 196,798 Slovene Wikipedia pages from *Wikipedia Monthly* [4],[12] a repository that provides monthly, multilingual dumps of Wikipedia, processed and prepared for easy use in NLP projects. However, the resource is problematic because all titles of Wikipedia articles are title-cased regardless of whether they are proper nouns or not (**Jezikoslovje* → *jezikoslovje* 'linguistics', **Stegodon* → *stegodon* 'stegodon', **Slovenščina* → *slovenščina* 'Slovene (language)', **Rehabilitacija* → *rehabilitacija* 'rehabilitation'). This ambiguity also frequently occurs with the CLASSLA-Stanza lemmatization/tagging models: common nouns are lemmatized as title-case and tagged as proper nouns (either because of their sentence-initial position or use in multiword named entities), so direct validation using Wikipedia articles would result in false positives. Validation through Wikipedia pages was thus postponed to a later stage.

The capitalization problem does not occur in Wikidata [11],[13] a free open-access knowledge base with over 122 million data entities that is readable by both machines and humans. In Wikidata, each item is listed in its basic canonical form, which makes it more suitable for lexeme validation. The database can be searched through the *Wikidata Query Service*[14] using SPARQL Query Language[15]. We compiled a SPARQL query to test how many of the 223,278 automatically added lexemes from DDDS are present in Wikidata as an item (each query contained a batch of 50 lexemes; the total number of queries was 4,466). The query returned the Wikidata identifier of the queried item itself, as well as the identifier and label of the hierarchically superior item linked to the queried item through the *instanceOf* property (*P31*). An example for the lexeme *Petelinjek* is shown in Table 2. The lexeme is

[7] *Pregibalnik* is available as open-access code at https://github.com/clarinsi/SloInflector and as an API service at https://orodja.cjvt.si/pregibalnik/docs; the *Form Generator* component, which deals with generating inflected forms, is also available as an API at https://orodja.cjvt.si/pregibalnik/form-generator/docs; depending on part-of-speech, the current version achieves an accuracy between 75-98% (see [1])

[8] Within DDDS, there is an important distinction: *lexemes* are single-word units that form the morphological basis for *lexical units* (which can also consists of multiple words). For instance, the lexical unit *Šempeter pri Gorici* (a town in Slovenia) is linked to three lexemes: *Šempeter* (noun, proper, masculine), *pri* (preposition), and *Gorica* (noun, proper, feminine)); lexemes are carriers of purely morphological information (such as inflected forms), while lexical units are semantic in nature (i.e. they are assigned senses, definitions, or links to other knowledge bases).

[9] In line with the *Multext-East v6 Slovene Morphosyntactic Specifications* (https://nl.ijs.si/ME/V6/msd/html/msd-sl.html), the only lexeme-level morphosyntactic feature for proper nouns is grammatical gender.

[10] SURS - SiStat: https://pxweb.stat.si/SiStat/sl

[11] Note that there is some overlap between the categories and some lexemes were validated by multiple lists, potentially warranting multiple lexemes in DDDS. Likewise, the names present on the lists but not included in DDDS can be added during the next database expansion.

[12] The latest export of *Wikipedia Monthly* at the time of writing this paper was from March 2026.

[13] Wikidata: https://www.wikidata.org/

[14] Wikidata Query Service: https://query.wikidata.org/

[15] SPARQL Query Language: https://www.w3.org/TR/sparql11-query/

Table 2: Results of the SPARQL query used to search Wikidata to validate the lexeme *Petelinjek*.

Item	Identifier	InstanceOf	InstanceOfLabel
Petelinjek	Q2400950	Q486972	human settlement
Petelinjek	Q31528946	Q8502	mountain

Table 3: The 10 most frequent labels of queried Wikidata items.

Label	Frequency
family name	81,826
human	40,074
painting	38,264
human settlement	27,497
album	26,340
film	20,688
unincorporated community in the United States	17,413
male given name	16,239
ship	12,839
single	12,027

present in Wikidata under two items; the first is categorized as a human settlement and the second as a mountain.

A total of 134,316 lexemes were validated through Wikidata (approx. 60% of all automatically added lexemes). However, it should be mentioned that the presence of the lexeme in Wikidata does not necessarily indicate that this is a basis for a relevant lexical unit in DDDS. For instance, the query for the lexeme *Arcobaleno* returns 71 identifiers in Wikidata; all of them refer to specific kindergarten and primary school institutions in Italy. None of these are mentioned in the *Gigafida 2.0* corpus, where the occurrences of *Arcobaleno* refer to different associations in Gorica/Gorizia, Italy. The validated lexemes will be further analyzed based on their *instanceOf* labels to determine which ones are present in more relevant categories that are frequently inflected in Slovene (names, surnames, locations), while the overly specific ones can potentially be removed as the DDDS is not meant to be a comprehensive encyclopedic resource, but primarily a linguistic database. The 10 most frequently occurring labels from the queries are shown in Table 3.

At this point, we checked whether any of the queried Wikidata items had links to Slovene Wikipedia pages: for instance, the Wikidata item for *Benelux* contained the link to the Slovene Wikipedia page for *Beneluks*, confirming *Beneluks* as a legitimate lexeme. A total of 15,901 additional lexemes were validated this way.

4 Validation and Correction with Large Language Models

The remaining 77,083 lexemes not yet found in any of the previous resources were analyzed using large language models. In the prompt design phase, different models were tested on a small sample of 200 lexemes, including *GPT 5.4*, *Gemini 3.5 Flash*, *Claude 4.6*, and *GaMS-27B-Instruct-Nemotron*.[16] A qualitative manual analysis of the results showed that the most useful output was provided by *Gemini 3.5 Flash*, as it contained the most informative context for its decisions on lexeme validation, distinguishing between different problems (misspellings, encoding errors, mislemmatizations, etc.) as shown in Table 4. The following prompt was used (with a list of 500 lexemes per prompt):

Table 4: Example of the output of Gemini 3.5 Flash for the proper noun lexeme correction task (comments truncated for brevity).

ID	Original	Corrected	Comment
442307	Surinamec	Surinamec	Legitimate entity.
465061	Koktalj	koktajl	Misspelled form...
465111	Zlatoličja	Zlatoličje	Mislemmatized...

> Please take on the role of a lexicographer with native knowledge of Slovene. You are helping me validate a list of lexemes for the Digital Dictionary Database of Slovene. I will provide you with a list of lexemes and their IDs. Your task is to determine whether each lexeme refers to a legitimate entity or if it is a mislemmatized form. If the entry is not correctly lemmatized, please provide the correct lemma. Make sure you analyze the entire list. It is crucial that you keep the IDs as they are.
> Reply only with a Markdown table structured in the following manner: |lexeme_id|original_lemma|correct_lemma|additional_comment|
> You can add very short comments (in English!) in the table to explain your corrections. No other explanations outside the table are required.
> This is the list of lexemes: <list>

4.1 Results

For 33,343 lexemes (43% of analyzed lexemes), the model confirmed their lemma and made no changes, whereas the lemmas for the remaining 43,740 lexemes (57%) were modified in different ways: 155 lexemes were only converted from lower-case to title-case (**vodafone* → *Vodafone*, **yamaha* → *Yamaha*); 7,859 lexemes were converted from title-case to lower-case (**Kamnolomstvo* → *kamnolomstvo*, **Spopad* → *spopad*); 35,726 lexemes were modified in more complex ways, usually by correcting lemmatization errors in morphology (**Nabucc* → *Nabucco*, **Pensilvani* → *Pensilvanija*, **Lenoro* → *Lenora*), suggesting preferred variants or alternatives (*Moldavijec* → *Moldavec* '(male) inhabitant of Moldova', *Singapurec* → *Singapurčan* '(male) inhabitant of Singapore') or correcting spelling errors (**Podvolovjek* → *Podvolovljek*).

We compared the corrected lexemes to the set of lexemes that were either manually confirmed or validated by any of the previous resources (SURS, Wikidata, Wikipedia). 21,861 of the corrected lexemes were essentially duplicates of existing validated lexemes: for instance, both **Jurkloštr* and **Jurkloštro* were correctly modified as *Jurklošter*, which is already included in DDDS and was validated by Wikidata and Wikipedia. Redundant candidates will be removed from DDDS. The high number of such candidates indicates that the use of large language models can successfully mitigate lemmatization errors when expanding the database with new lexeme candidates from corpora.

4.2 Evaluation

We performed a preliminary evaluation of the results by randomly sampling 250 lemmas that have been validated by the

[16] https://huggingface.co/cjvt/GaMS-27B-Instruct-Nemotron

Table 5: Confusion matrix for the evaluation sample of Gemini 3.5 Flash.

True → ↓ Predicted	Correct Lemma	Incorrect Lemma	∑
Correct Lemma	**218**	32	250
Incorrect Lemma	31	**219**	250
∑	249	251	-

model and 250 lemmas that have been modified (stratified by the categories mentioned in Section 4.1; so 83 lower-to-titlecase lemmas, 83 title-to-lowercase lemmas, and 84 lemmas with more complex modifications). We manually evaluated the sample to assess model performance. The results are shown in Table 5 (A=BA=P=R=F1≈87.40; a baseline majority classifier achieves an accuracy of 50.20%).[17]

Despite the very robust prompt and lack of any context for individual lemmas, in the majority of examples, the model correctly predicted whether the lemma is correct or whether it requires modification. It should be noted that this task is also difficult for human annotators without context or prior knowledge of the lemma in question. In some cases, lemmas can also be ambiguous: for instance, the lexeme candidate *Delgada* can be legitimate as part of the location *Ponta Delgada* (Portugal), but not legitimate if it is in fact an inflected form of the surname *Delgado*. Such examples are not frequent, however: in the evaluation sample, only 22 (4%) lexemes were ambiguous.

We also evaluated the quality of the model's lemma corrections, and the lemmas were correct in a total of 448 examples (89.60%), and even in the ambiguous cases, the correction usually corresponded to one of the valid options. Any ambiguities were also explained in the model's additional comment: for instance, the lexeme *lakota* was corrected as *lakota* 'hunger', but the comment ("*Legitimate common noun (hunger) or tribe (Lakota).*") also predicted it might refer to the Lakota people. The model seemed to perform well regardless of whether the correction was due to a lemmatization error (**Bjerringbra → Bjerringbro*), diacritic or grapheme simplifications (**Szekesfehervar → Székesfehérvár*), capitalization errors (*Mccormack → McCormack*), truncated forms or encoding artifacts (*drĹžava → država*), or misspelling (**Dormund → Dortmund*).

The manual evaluation of 500 lemmas (by a lexicographer) took approximately 5 hours, and the results show that annotations of a similar quality can be achieved with a single prompt, leaving more time for complex lexicographic tasks.

5 Conclusion

We have demonstrated an approach to validating proper noun lexemes in DDDS using both existing resources as well as large language models. The results of the experiment[18] show that the proposed method complements existing workflows used to update language resources with corpus-based data and can be used to (pre-)validate proper noun candidates, which provides reliable results and reduces the workload of lexicographers. The identified corrections will be implemented in the DDDS, and the validated data can also be used to expand Wikidata with missing items (and, by extension, Slovene Wikipedia).

Although we have so far focused on *Gemini 3.5 Flash* because of its performance, additional experiments with other open-access models are warranted, particularly with the *GaMS* family of Slovene models. Additional prompts with more context (such as examples of use from corpora) will be tested, and the validated data from DDDS can be used to compile fine-tuning datasets for lemmatization correction.

In this experiment, we presented only the first step of correcting lexemes (lemma validation). We will test LLMs (as well as simpler machine-learning approaches such as logistic regression and support-vector classifiers) to complement the other steps in the workflow of expanding DDDS, such as determining basic morphosyntactic features of lexemes (such as gender for noun lexemes and aspect for verbs) and inflected forms – for instance, to determine whether a masculine noun is animate or inanimate; a semantic category which determines the accusative singular inflected form: *Kerssenberg* (surname, masculine, animate) → accusative singular *Kerssenberga*; *Kienhofen* (toponym, masculine, inanimate) → accusative singular *Kienhofen*. We leave these steps for our future work.

Acknowledgements

The research presented in this paper was carried out within the research programme *Language Resources and Technologies for Slovene* (P6-0411), the research project *Large Language Models for Digital Humanities* (LLM4DH; GC-0002), and the *CLARIN.SI Research Infrastructure* (I0-E004), all funded by the Slovenian Research and Innovation Agency (ARIS). The author also thanks Kaja Gantar for helping with the evaluation and the anonymous reviewers for their constructive comments.

References

[1] Jaka Čibej. 2023. Leksikon besednih oblik Sloleks. Poročilo projekta Razvoj slovenščine v digitalnem okolju Aktivnost DS1.3. Development of Slovene in a Digital Environment. (2023). https://www.cjvt.si/rsdo/wp-content/uploads/sites/18/2023/06/RSDO_Kazalnik_Sloleks_v2.pdf.

[2] Jaka Čibej and Tina Munda. 2024. Metoda polavtomatskega popravljanja lem in oblikoskladenjskih oznak na primeru učnega korpusa govorjene slovenščine ROG. (2024). doi:10.5281/zenodo.13936390.

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

[4] Omar Kamali and Omneity Labs. 2025. Wikipedia Monthly. (2025). doi:10.57967/hf/6575.

[5] Timotej Knez and Slavko Žitnik. 2026. Slovene Lexicographic QA Fine-Tuning Corpus SloLexQA 1.0. Slovenian language resource repository CLARIN.SI. (2026). http://hdl.handle.net/11356/2116.

[6] Iztok Kosem, Jaka Čibej, Kaja Dobrovoljc, Tomaž Erjavec, Nikola Ljubešić, Primož Ponikvar, Mihael Šinkec, and Simon Krek. 2026. Monitor corpus of Slovene Trendi 2026-05. Slovenian language resource repository CLARIN.SI. (2026). http://hdl.handle.net/11356/2219.

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

[8] Simon Krek et al. 2019. Corpus of Written Standard Slovene Gigafida 2.0. Slovenian language resource repository CLARIN.SI. (2019). http://hdl.handle.net/11356/1320.

[9] Nikola Ljubešić, Luka Terčon, and Jaka Čibej. 2023. The CLASSLA-Stanza model for morphosyntactic annotation of standard Slovenian 2.0. Slovenian language resource repository CLARIN.SI. (2023). http://hdl.handle.net/11356/1767.

[10] Luka Terčon, Jaka Čibej, and Nikola Ljubešić. 2023. The CLASSLA-Stanza model for lemmatisation of standard Slovenian 2.0. Slovenian language resource repository CLARIN.SI. (2023). http://hdl.handle.net/11356/1768.

[11] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57, 10, (Sept. 2014), 78–85. doi:10.1145/2629489.

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

[18] The prompts, SPARQL queries and evaluation data are available on Github: https://github.com/jakacibej/sikdd2026_proper_noun_validation