Predicting Grammatical Cases in Slovenian Varieties in Italy: A Use Case from the LORIS 1.1 Language Assistant

David Bordon[1]

[1] University of Ljubljana, Slovenia – david.bordon@ff.uni-lj.si

ABSTRACT (ENGLISH)

This work showcases the LORIS 1.1[1] language assistant – a web-based application developed to support the Slovenian speaking minority living in Italy, with a specific focus on the linguistic challenges arising from contact with Italian. Alongside integration of different language technology tools for Slovenian, LORIS includes a database of language-contact phenomena specifically created to target the language minority. Our main focus is on the technical challenges of operationalizing this dataset—namely, predicting all possible grammatical case forms in Slovenian and ensuring that the system avoids false positives. We discuss our approach to handling non-standard forms and highlight how this technology can promote linguistic inclusivity and preservation for minority languages.

Keywords: Slovenian Language; Language Technology; Minority Languages; Morphology Prediction; Digital Humanities

ABSTRACT (ITALIANO)

Previsione dei casi grammaticali nelle varietà di sloveno in Italia: Esempio pratico dell'assistente linguistico LORIS 1.1

Questo contributo presenta l'assistente linguistico LORIS 1.1, un'applicazione web-based sviluppata per supportare la minoranza di lingua slovena che risiede sul territorio italiano, con un interesse specifico per i fenomeni linguistici derivanti dal contatto con l'italiano. Oltre all'integrazione di diversi strumenti di tecnologia linguistica per lo sloveno, LORIS include un database di fenomeni di contatto linguistico creato appositamente per la minoranza slovena. Il presente lavoro si concentra principalmente sulle sfide tecniche legate all'operatività di questo set di dati, ovvero la previsione di tutte le possibili forme di casi grammaticali in sloveno, garantendo che il sistema eviti i falsi positivi. Discutiamo il nostro approccio alla gestione delle forme non standard e sottolineiamo come questa tecnologia possa promuovere l'inclusività linguistica e la conservazione delle lingue minoritarie.

Parole chiave: Lingua Slovena; Tecnologia Linguistica; Lingue Minoritarie; Previsione Morfologica; Digital Humanities

1. INTRODUCTION

Slovenia and Italy share a 232-kilometer border that stretches across much of Slovenia's western region. Because of a complex historical background and socio-economic factors, each country hosts a minority of the other's population – Slovenians in Italy and Italians in Slovenia – leading to mutual linguistic influence over time.

The Slovenian community in Italy has thus developed specific linguistic traits, which have been the subject of extensive research over the past two decades (Grgič, 2017; Jagodic et al., 2017). Data indicate that the language code in areas where the Slovenian minority in Italy is present is gradually diverging from the central standard of Slovenian to the point that "local usage sometimes significantly distances itself from the Slovenian linguistic continuum" (Grgič, 2017).

A study (Grgič & Popič, 2023) employing modern corpus linguistics methods analyzed approximately 20 million words from two annual volumes (2020–2021) of *Primorski dnevnik*, a Slovenian-language newspaper published in Italy focused on COVID-19-related terminology – an area where terminological consistency is to be expected due to its technical nature and international relevance. However, the analysis revealed a significant degree of lexical variation between the standard Slovenian used in Slovenia and the variety found in cross-border public mediatic discourse which demonstrates that the divergence is not limited to informal registers or locally contextualized reference but extends to structured, formal texts and standardized terminology.

[1] https://www.jeziknaklik.it/loris/

Minority, marginalized, or indigenous languages face the threat of decline or even (digital) extinction, while at the same time, language technologies offer them new opportunities (Grenoble & Whaley, 2006). The LORIS 1.1 language assistant, a browser-based application, is designed to provide linguistic advice to the Slovene minority living in Italy, navigating between two formal languages and systems.

While in recent times there have been many clones of the Grammarly app, and after the surge of generative AI tools there has been a great influx of AI grammar checkers, tools targeted at low resource languages, especially minority ones, are not as frequent. For example, initiatives such as GaelSpell[2] or Divvun[3], which are applications focused on minority languages, mostly act as spellcheckers. In this respect, Loris with its focus on language contact linguistic phenomena and by providing definitions and explanations of these phenomena, is among the more unique and specialized ones.

2. LORIS 1.1 Overview

LORIS 1.1 was designed to suggest standardized forms and explanations to users based on a database of common linguistic errors and interferences (e.g., false friends ("Avtist" (SI) *person with authism* vs. "Autista" (IT) *chauffeur*), incorrect case endings), as well as normative sources. The program is highly modular in its design, combining manually compiled and exported lexical lists of cases of language-contact, alongside more computationally heavy services integrated as "plugins".

The current version includes the following key resources:

- Data related to language contact between Slovenian and Italian (examples of paronyms, false friends, calques e.g. *ne vidim ure* (I cannot see the clock)/*non vedo l'ora* (I cannot wait))
- An internal database of toponyms in both, Slovenian and Italian (Slovene names of towns in Italy, e.g. *Devin/Duino* or *Nabrežina/Aurisina*)
- Standardized vocabulary from the *Slovenian Orthographic Dictionary 2001* (including words labeled as *incorrect*, *prohibited*, or *discouraged* – mostly tied to declination of proper names)
- Standardized vocabulary from the Slovenian morphological lexicon *Sloleks*[4]
- Synonyms for some of the most common lexemes, sourced from the *Synonyms 1.0* dictionary[5]

Except for internal SLORI lists, all resources were obtained from open-access and freely available databases. The custom lists were created by domain experts, aggregating materials from dictionaries, editorial and advisory services, and user suggestions to name a few.

In addition, LORIS incorporates various applications based on machine-learned models, all of which are based on the neural model BERT or sloBERTa. The tools in question are the automatic comma insertion tool *Vejice 1.0* (Božič et al., 2020) with an accuracy of 94 %[6], a neural spellchecker (Klemen et al., 2024) with $F_{0,5}$ scores on evaluation tasks exceeding 0.9, and the *EssayHelper* (Petrič, 2021) system for automatic correction of case forms and verb numbers (dual vs. plural).

The application works only on written text – there is no speech component due to the fact that open-source solutions for a low resource language such as Slovene are not readily available, while dialects and non-standardized vocabulary are not supported in current applications.

From launch in February 2023 until summer 2024 (last analysis), the tool was used 4970 times.

The full program code is available upon request on a GitHub repository curated by the authors[7]. Due to contractual obligations and third-party API integration, it is not possible to make it fully open source under a Creative Commons license.

3. Use Case – Predicting Grammatical Cases

An existing challenge during the development of LORIS was that none of the existing systems could predict non-standard cases. This issue is amplified by Slovenian's inherent morphological complexity – six cases, three grammatical genders and three grammatical numbers. To address these gaps, we developed a custom solution capable of predicting *all* possible variants for a given lemma, including unusual or non-standard forms.

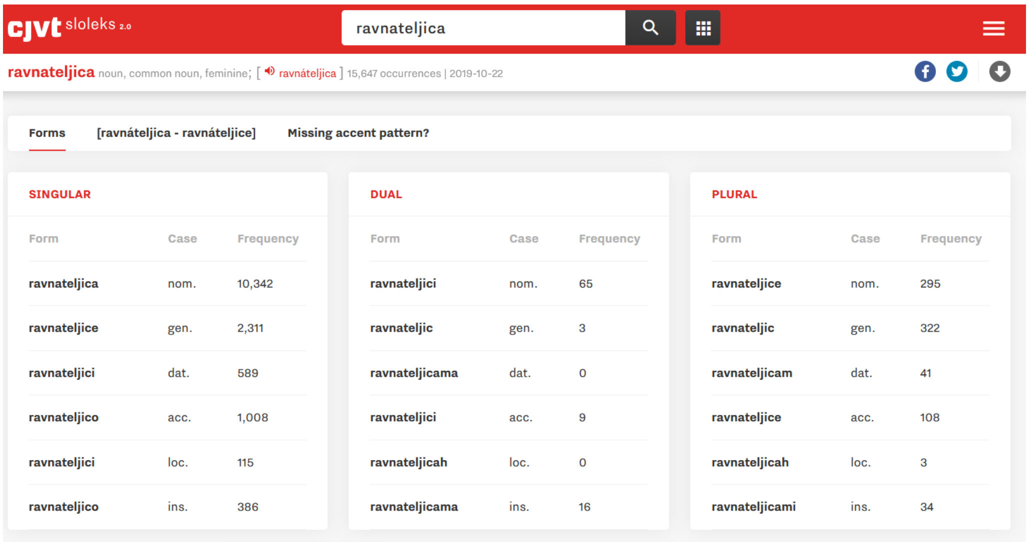
[2] https://cadhan.com/gaelspell/index-en.html
[3] https://divvun.no/en/korrektur/speller-demo.html
[4] https://viri.cjvt.si/sloleks/slv/
[5] https://viri.cjvt.si/sopomenke/slv/
[6] https://orodja.cjvt.si/vejice/about
[7] To obtain access please send an e-mail specifying how you'd wish to use the data.

Unlike regex-based approaches, which struggle to differentiate between forms like *ravnatelj* (male) and *ravnateljica* (female), LORIS employs a manually curated list of lemma-specific cases. This approach ensures that every lemma and its potential non-standard cases are accounted for individually. The same principle applies to non-standard forms, where incorrect case endings must be manually listed to guarantee reliable detection.

The pipeline for implementing case detection and correction in LORIS follows a structured approach and is carried out manually using open-source tools for the Slovenian language.

First, all cases for a specific lemma are extracted from Sloleks. For lemmas not available in the database, cases are added following the declension rules for the Slovene language. For multi-word expression, every word in an expression is processed individually. The extracted cases are compiled into a structured dataset for manual cleaning. For multi-word expressions, every possible combination is tested and the ones not grammatically adequate are removed. Next, all duplicate entries are removed, and non-standard cases are manually added. The final step is to structure the data into strings and to add syntax coding for implementation in the build.

By curating every potential variant, from conventional to non-standard, this pipeline reduces false positives and enhances detection accuracy.



ravnateljica noun, common noun, feminine; [ravnáteljica] 15,647 occurrences | 2019-10-22

SINGULAR

Form	Case	Frequency
ravnateljica	nom.	10,342
ravnateljice	gen.	2,311
ravnateljici	dat.	589
ravnateljico	acc.	1,008
ravnateljici	loc.	115
ravnateljico	ins.	386

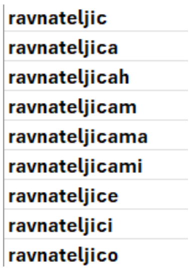
DUAL

Form	Case	Frequency
ravnateljici	nom.	65
ravnateljic	gen.	3
ravnateljicama	dat.	0
ravnateljici	acc.	9
ravnateljicah	loc.	0
ravnateljicama	ins.	16

PLURAL

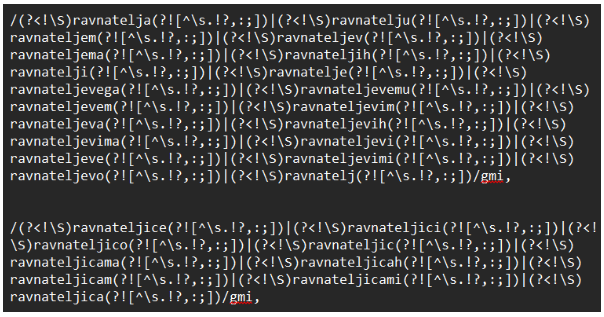
Form	Case	Frequency
ravnateljice	nom.	295
ravnateljic	gen.	322
ravnateljicam	dat.	41
ravnateljice	acc.	108
ravnateljicah	loc.	3
ravnateljicami	ins.	34

Figure 1. Lemmas for "ravnateljica" as viewed in Sloleks



ravnateljic
ravnateljica
ravnateljicah
ravnateljicam
ravnateljicama
ravnateljicami
ravnateljice
ravnateljici
ravnateljico

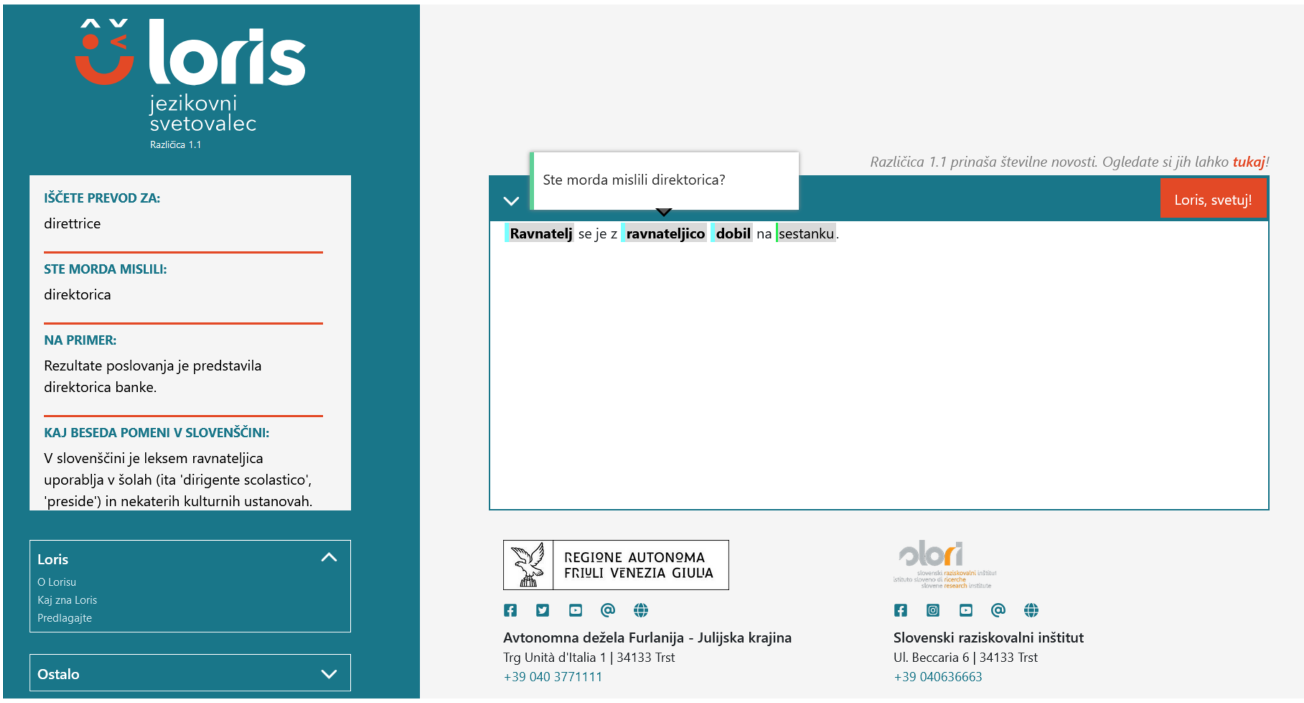
Figure 2. All cases for the lemma "ravnateljica"



```
/(?<!\S)ravnatelja(?![^\s.!?,:;])|(?<!\S)ravnatelju(?![^\s.!?,:;])|(?<!\S)
ravnateljem(?![^\s.!?,:;])|(?<!\S)ravnateljev(?![^\s.!?,:;])|(?<!\S)
ravnateljema(?![^\s.!?,:;])|(?<!\S)ravnateljih(?![^\s.!?,:;])|(?<!\S)
ravnatelji(?![^\s.!?,:;])|(?<!\S)ravnatelje(?![^\s.!?,:;])|(?<!\S)
ravnateljevega(?![^\s.!?,:;])|(?<!\S)ravnateljevemu(?![^\s.!?,:;])|(?<!\S)
ravnateljevem(?![^\s.!?,:;])|(?<!\S)ravnateljevim(?![^\s.!?,:;])|(?<!\S)
ravnateljeva(?![^\s.!?,:;])|(?<!\S)ravnateljevih(?![^\s.!?,:;])|(?<!\S)
ravnateljevima(?![^\s.!?,:;])|(?<!\S)ravnateljevi(?![^\s.!?,:;])|(?<!\S)
ravnateljeve(?![^\s.!?,:;])|(?<!\S)ravnateljevimi(?![^\s.!?,:;])|(?<!\S)
ravnateljevo(?![^\s.!?,:;])|(?<!\S)ravnatelj(?![^\s.!?,:;])/gmi,


/(?<!\S)ravnateljice(?![^\s.!?,:;])|(?<!\S)ravnateljici(?![^\s.!?,:;])|(?<!
\S)ravnateljico(?![^\s.!?,:;])|(?<!\S)ravnateljic(?![^\s.!?,:;])|(?<!\S)
ravnateljicama(?![^\s.!?,:;])|(?<!\S)ravnateljicah(?![^\s.!?,:;])|(?<!\S)
ravnateljicam(?![^\s.!?,:;])|(?<!\S)ravnateljicami(?![^\s.!?,:;])|(?<!\S)
ravnateljica(?![^\s.!?,:;])/gmi,
```

Figure 3. Final code output for "ravnatelj" and "ravnateljica"


loris
jezikovni
svetovalec
Različica 1.1
Loris, svetuj!
Loris
O Lorisu
Kaj zna Loris
Predlagajte
Ostalo
REGIONE AUTONOMA FRIULI VENEZIA GIULIA
slori
slovenski raziskovalni inštitut
istituto sloveno di ricerche
slovene research institute

Različica 1.1 prinaša številne novosti. Ogledate si jih lahko **tukaj**!

Ste morda mislili direktorica?

Ravnatelj se je z **ravnateljico** **dobil** na sestanku.

IŠČETE PREVOD ZA:
direttrice

STE MORDA MISLILI:
direktorica

NA PRIMER:
Rezultate poslovanja je predstavila direktorica banke.

KAJ BESEDA POMENI V SLOVENŠČINI:
V slovenščini je leksem ravnateljica uporablja v šolah (ita 'dirigente scolastico', 'preside') in nekaterih kulturnih ustanovah.

Avtonomna dežela Furlanija - Julijska krajina
Trg Unità d'Italia 1 | 34133 Trst
+39 040 3771111

Slovenski raziskovalni inštitut
Ul. Beccaria 6 | 34133 Trst
+39 040636663

Figure 4. The case detected in the program

4. Conclusion

LORIS 1.1 represents a useful tool for supporting Slovenian language use for the minority living in Italy and showcases how different open-source tools (if available) can be integrated into more focused applications to support a specific minority language community. Digital support for minority languages is crucial for their preservation and development – LORIS exemplifies how modern technologies can contribute to linguistic inclusion and awareness.

The authors hope this approach inspires comparable initiatives and strongly encourage the Italian speaking community in Slovenia to develop digital solutions to support their needs. We also envision cross-border collaboration that would allow other linguistic minorities to adopt similar strategies, fostering inclusivity and linguistic equity. In this way, LORIS 1.1 stands as both a practical resource for users and a model for developing digital tools that support minority language vitality and cultural identity.

ACKNOWLEDGEMENTS

The project was made possible with the support of the Slovene Research Institute SLORI and Central Office for the Slovenian Language of the Autonomous Region of Friuli–Venezia Giulia.

REFERENCES

Bird, S. (2020). *Decolonising speech and language technology*. In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 3504–3519). International Committee on Computational Linguistics.

Fabjan Bajc, D. (1994). *Lažni prijatelji: Slovensko-italijanski slovar paronimov / I falsi amici: Vocabolario italiano-sloveni dei paronimi*. Mladika.

Fabjan Bajc, D. (1995). *Dve muhi na en mah: Slovensko-italijanski frazeološki slovar / Due piccioni con una fava: Vocabolario fraseologico sloveno-italiano*. Zadruga Goriška Mohorjeva.

Grenoble, L. A., & Whaley, L. J. (2006). *Saving languages: An introduction to language revitalization*. Cambridge University Press.

Grgič, M. (2017). *Italijansko-slovenski jezikovni stik med ideologijo in pragmatiko*. *Jezik in slovstvo, 62*(1), 89–98.

Grgič, M., & Popič, D. (2023). *Procesi jezikovnega separatizma pri čezmejnih jezikovnih manjšinah: Series Historia et Sociologia, 33*(1), 151–166. http://www.dlib.si/details/URN:NBN:SI:doc-QIYGXKVD

Jagodic, D., Kaučič Baša, M., & Dapit, R. (2017). *Jezikovni položaj Slovencev v Italiji*. In N. Bogatec & Z. Vidau (Eds.), *Skupnost v središču Evrope: Slovenci v Italiji od padca Berlinskega zidu do izzivov tretjega tisočletja* (pp. 66–88). ZTT-SLORI.

Klemen, M., Božič, M., Holdt, Š. A., & Robnik-Šikonja, M. (2024). *Neural spell-checker: Beyond words with synthetic data generation*. In E. Nöth, A. Horák, & P. Sojka (Eds.), *Text, speech, and dialogue. 27th International Conference, TSD 2024* (Lecture Notes in Computer Science, Vol. 15048). Springer.

Petrič, T. (2022). *Predlogi jezikovnih popravkov v slovenščini z modelom SloBERTa* (Bachelor's thesis, University of Ljubljana). University of Ljubljana Repository.