

# Leveraging Frequency List of Language Problems from Šolar 3.0

Špela ARHAR HOLDT, University of Ljubljana, Slovenia ([arharhs@ff.uni-lj.si](mailto:arharhs@ff.uni-lj.si))

Resource available at CLARIN.SI



Arhar Holdt, Špela; Rozman, Tadeja; Stritar Kučuk, Mojca; Krek, Simon; Krapš Vodopivec, Irena; Stabej, Marko; Pori, Eva; Goli, Teja; Lavrič, Polona; Laskowski, Cyprian; Kocjančič, Polonca; Klemenc, Bojan; Krsnik, Luka; Žagar, Aleš; Kosem, Iztok. 2022, Frequency list of language problems from Šolar 3.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, <http://hdl.handle.net/11356/1716>

- We created the *Frequency List of Language Problems from Šolar 3.0*.
- The dataset includes 36,570 sentences written by primary and secondary school students from various Slovene schools.
- Each sentence is accompanied by metadata detailing the text type, the educational stage of the author, and the school type and region where the text was produced.
- The sentences include teacher-provided corrections of language problems.
- Corrections were manually categorized into 180 types using a hierarchical labeling system (Arhar Holdt et al., 2023).
- We calculated and compared the relative frequencies of corrections, considering the author's educational stage and three levels of correction label robustness.

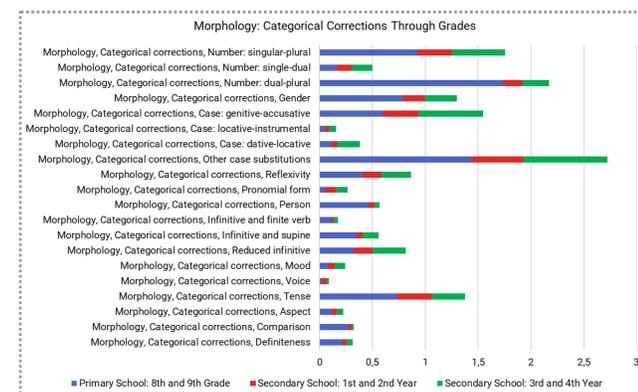
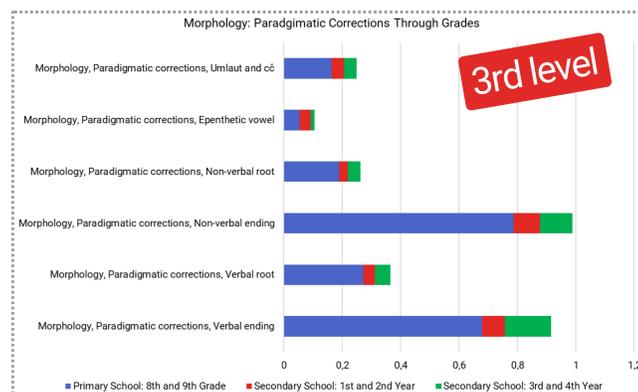
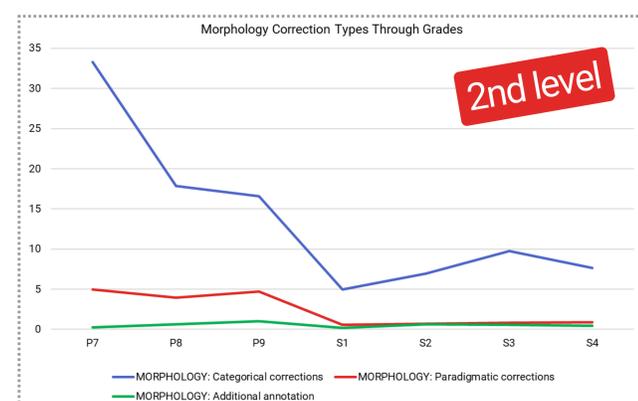
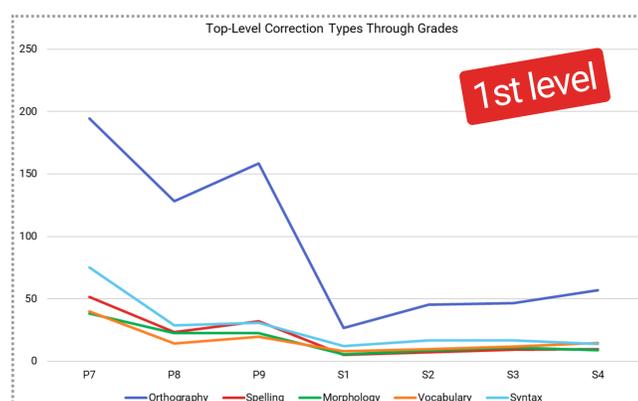
$$\text{Relative Correction Frequency (per Grade)} = \left( \frac{\text{Number of Specific Corrections}}{\text{Total Number of Corrections} \times \text{Total Number of Words}} \right) \times 10,000,000$$

O/KAT/sklon-dm

1st Level: Morphology, 2nd Level: Categorical corrections, 3rd Level: Case: dative-locative

- The results distinguish corrections prevalent at specific educational stages from those that persist throughout the years.
- Analyses pertaining to the region of the school proved less relevant; a more balanced corpus is needed.
- The findings provide valuable insights for Slovene language didactics and suggest improvements for structuring the Šolar corpus (on Šolar 3.0 see Arhar Holdt & Kosem, 2024).

Developmental corpus Šolar: 5,485 texts (primarily school essays) written by students in Slovene secondary schools (age 15-19) and pupils in the 7th-9th grade of primary school (13-15), small percentage also from the 6th grade. Part of the corpus (2,094 texts) is annotated with teachers' corrections. Available at CLARIN.SI: <http://hdl.handle.net/11356/1589>.

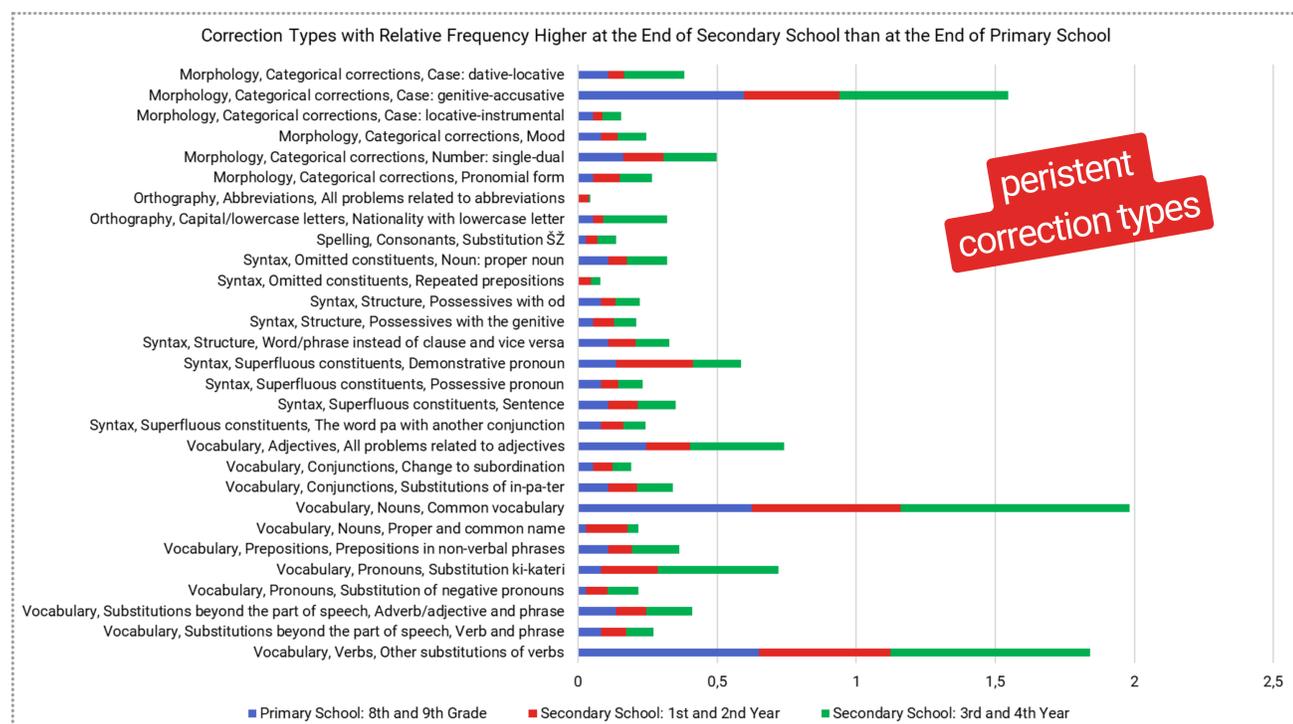


Paradigmatic = the corrected word form does not exist in the standard language:

- Ko družini **izveta**, da sta se Romeo in Julija poročila se pobotata. (*Verbal ending*)
- Ko družini **izvesta**, da sta se Romeo in Julija poročila, se pobotata.
- When the families **find out** that Romeo and Juliet have married, they make up.

Categorical = the corrected word form exists but is not grammatically appropriate for the context:

- Antigona pa se ni bila pripravljena prilagoditi, zato **jo** ni čakal dober konec. (*Case: genitive-accusative*)
- Antigona pa se ni bila pripravljena prilagoditi, zato **je** ni čakal dober konec.
- However, Antigone was not willing to adapt, so a good end did not await **her**.



Qualitative analyses with the help of a specialised concordancer (<https://solar.cjvt.si/>).

Affiliation



Project



Financer



Acknowledgements

The authors acknowledge the financial support of the Slovenian Research and Innovation Agency through the project Empirical foundations for digitally-supported development of writing skills (J7-3159), the programme Language Resources and Technologies for Slovene (P6-0411) and the research infrastructure CLARIN.SI.

ARHAR HOLDT, Špela, LAVRIČ, Polona, ROBLEK, Rebeka, GOLI, Teja, BON, Mija, 2023: *Categorizing Teachers' Corrections: Guidelines for Annotating the Šolar Corpus*. Version 1.2. Available from: <https://wiki.cjvt.si/books/11-developmental-corpus-solar/page/annotation-guidelines>.  
ARHAR HOLDT, Špela, KOSEM, Iztok, 2024: Šolar, the developmental corpus of Slovene. *Language Resources and Evaluation*, in print.