Language Resources and Evaluation
https://doi.org/10.1007/s10579-024-09758-4

ORIGINAL PAPER

Šolar, the developmental corpus of Slovene

Špela Arhar Holdt[1,2] · **Iztok Kosem**[1,2]

Accepted: 19 June 2024

Abstract

The paper presents the Šolar developmental corpus of Slovene, comprising the written language production of students in Slovene elementary and secondary schools, along with teacher feedback. The corpus consists of 5485 texts (1,635,407 words) and includes linguistically categorized teacher corrections, making the corpus unique in reflecting authentic classroom correction practices. The paper addresses the corpus compilation, content and format, annotation, availability, and its applicative value. While learner corpora are abundant, developmental corpora are less common. The paper bridges the gap by introducing the evolution from Šolar 1.0 to 3.0, emphasizing improvements in text collection, error and correction annotation, and categorization methodology. It also underlines the challenges and unresolved issues of compiling developmental corpora, most notably the lack of openly available tools and standards for different steps of the compilation process. Overall, the Šolar corpus offers valuable insights into language learning and teaching, contributing to teacher training, empirical studies in applied linguistics, and natural language processing tasks.

Keywords Šolar · Developmental corpus · Slovene language · Student writing · Teacher feedback

✉ Špela Arhar Holdt
Spela.ArharHoldt@ff.uni-lj.si

Iztok Kosem
iztok.kosem@ff.uni-lj.si

[1] University of Ljubljana (Faculty of Arts & Faculty of Computer and Information Science), Ljubljana, Slovenia

[2] Jožef Stefan Institute, Ljubljana, Slovenia

Published online: 18 July 2024

1 Introduction

Developmental corpora represent "the language as used by native speakers whose competence has not yet reached maturity" (Leech, 1997 p. 19).[1] They must be distinguished from *learner corpora*, which sample the language production of speakers who are learning a language as their second/foreign language (Granger, 2008). Both types of corpora are primarily built to meet the needs of language didactics: they facilitate bottom-up language learning (Osborne, 2002), the creation of teaching and testing materials, quantitative and qualitative studies of student writing problems, and the training of systems for automatic evaluation of student language production, among other things. However, while the field of learner corpora and related research is thriving, especially when aimed at teaching English as L2,[2] developmental corpora are still relatively scarce.

In this paper, we present a state-of-the-art developmental corpus: the Šolar developmental corpus of Slovene. The main purpose of the corpus is to enable empirical research into the written language production of students in Slovene elementary and secondary schools, as well as into teacher feedback. The corpus comprises 5485 texts (1,635,407 words) that were written by Slovene elementary and secondary school students as part of their coursework. The corpus also includes 36,570 linguistically categorised teacher corrections. What makes the Šolar corpus unique, not only in the Slovene context but also internationally, is the fact that error annotation is based on the corrections made by teachers, rather than corpus designers. Thus, in addition to being representative of the language production of Slovene students, the corpus also shows the practice of Slovene teachers, reflected in the actual correction in a typical classroom environment.

The Šolar corpus has a long history, dating back to 2010 when the first version was published. Subsequent versions, Šolar 2.0 and Šolar 3.0 were released in 2019 and 2022 respectively. Despite shorter presentations of Šolar 1.0 by Kosem et al. (2011) and Arhar Holdt et al. (2017), as well as some contributions on Šolar 1.0 and 2.0 in the Slovene language (Arhar Holdt et al., 2022a; Kosem et al., 2012, 2016), the corpus has not yet been fully presented. In this paper, we aim to provide a comprehensive overview of the corpus, including its content, text collection and legal issues, transcription, anonymisation, error/correction annotation and categorisation, linguistic tagging, format, availability, and applications of the corpus use.

Each new edition of the corpus has presented an opportunity for evaluation and improvement of the corpus compilation methodology. We believe that highlighting these changes is of particular importance, as it illustrates the dependency of the results on the availability of specialized corpus-building tools and digital workflows,

[1] In the field of corpus linguistics, Leech differentiates between L1 and L2 developmental corpora, depending on whether the sampled language production is from native speakers or from second language learners. The term "learner corpora" is well-established for the latter, which is why we advocate for the use of the term "developmental corpora" specifically for L1 language production.

[2] The field of learner corpus research has a significant body of literature, as demonstrated by the extensive bibliography of the Learner Corpus Association, which includes over 1000 references (http://www.learnercorpusassociation.org/).

which although increasing, are not yet optimal. Additionally, providing insight into the several tailor-made solutions that were necessary to create the corpus Šolar can aid in the advancement of methodology for creating corpora that include language corrections, which will have positive effects on the availability of developmental and learner corpora, and similar language resources.

2 Related work

The LUCY Corpus is an electronic sample of modern written English produced in the UK by a spectrum of writers ranging from skilled published authors to young children, equipped with detailed annotation identifying grammatical and other linguistic structures (Sampson, 2003). The corpus includes texts written by young adults (33,000 words) and children between 9 and 12 years old (30,000 words). Similarly, the LOCNESS corpus comprises native English essays written by British pupils (60,209 words) and university students (95,695 words), as well as American university students (168,400 words), making it a valuable resource for studying language variation and development (Granger, 1998). Also noteworthy is the corpus created in the Growth in Grammar project (Durrant, 2019), studying how English children's written language develops as they progress through their school careers (from ages 6 to 16). 2898 texts from 983 children in 24 schools were collected, corrected for spelling, linguistically annotated, and analysed to understand differences in the use of grammar and vocabulary across year groups and text types. Other relevant work includes the corpus described by Parr (2010), consisting of 20,947 essays written by New Zealand students in years 4 to 12 of schooling, manually evaluated to track writing progress across school years and types of schools.

For German, there are also several available resources. Berkling (2016, 2018) describes a longitudinal collection of corpora (H1, H2, E2, ERK1) comprising 181,385 tokens in total, representing weekly writing by German schoolchildren aged 6–11 years, elicited within the normal classroom setting. The corpus contains transcriptions of the texts with and without spelling errors, aligned on a word-by-word basis. Additionally, the Litkey Corpus is available, containing 212,505 tokens of written texts produced by primary school children in Germany from grades 2 to 4, which have been transcribed and enriched with manual corrections of orthographic errors as well as semi-automatic linguistic annotations (Laarmann-Quante et al., 2019). To investigate the writing skills of German-speaking secondary-school students at the end of their school career, the KoKo Corpus has been created (Abel et al., 2014). The v3 of the corpus comprises 1503 argumentative essays (950,000 tokens) and includes manually performed transcriptions and linguistic error annotations.

For Italian, Barbagli et al. (2016) introduce CItA, a corpus of 1352 essays (369,456 tokens) written by Italian L1 learners in the first and second year of secondary school. The corpus includes error corrections and tracks the development of L1 writing competence of the same group of students over two school years and several students' background information. Some previous resources for Italian are also the collection of 5000 essays written by students from the first 5 years of elementary

school (Marconi et al., 1993) a collection of 2500 essays written by students from the first year of different high schools in Rome (Borghi, 2013). LEONIDE, a longitudinal corpus documenting the writing development of lower secondary school students in Italian, German, and English, contains 2512 texts from the multilingual Italian province of South Tyrol (Glaznieks et al., 2022).

Newer language resources include The Icelandic Error Corpus (Arnardóttir et al., 2021), which consists of modern Icelandic texts annotated for spelling, grammar, and other errors. It comprises 176 essays written by high school students aged 16–20. The Icelandic Child Language Error Corpus (Ingason et al., 2021) includes 119 error-annotated texts written by native Icelandic speakers aged 10–15. DOESTE v0.5 (Martins et al., 2020) is a developmental corpus of texts written by school-age children and adolescents in Brazil and Portugal. It includes 244 Portuguese and 450 Brazilian Portuguese narrative and argumentative essays from authors aged around 10–18 years. Recent corpora for French (Doquet et al., 2017; Garcia-Debanc et al., 2017; Jacques & Rinck, 2017; Wolfarth et al., 2017) were made available as a unified resource, È:CALM (Ho-Dac et al., 2020), comprising over 6,700 texts covering primary school to university and a variety of genres.

The Czech corpus Chyby (Pala et al., 2003) is similar to the Šolar corpus in that it includes teacher corrections; however, it differs in terms of the type of texts included. This resource comprises approximately 410,000 words from texts written by university students, specifically for the subject Element of Style. Two teachers reviewed the texts and returned them to the students, who then annotated the corrections.

The presented resources are comparable to the corpus Šolar in terms of their structure and purpose, with a focus on languages that are geographically close to Slovene. In addition to the resources mentioned, there are several other types of corpora that serve diverse purposes. While our review primarily focused on written resources, spoken and transcribed corpora, such as those gathered in learning environments, can provide valuable insights into language development and learning disabilities. Moreover, written corpora can be designed to target typical errors associated with disabilities like dyslexia. While our review concentrated on texts from primary and secondary school students, many corpora concentrate on university students and academic language skills. Some corpora have been developed to analyse genre, discourse, and other linguistic features beyond language errors and corrections. Although our review focused on resources targeting L1, it is worth noting that L1 texts can also be used as a baseline in some L2 corpora. Lastly, error-annotated corpora are extensively utilized in NLP for various tasks, such as GEC, machine translation error classification, OCR correction, and more, highlighting their value in advancing NLP research and applications.

3 Šolar 1.0 to Šolar 3.0

The history of the Šolar corpus shows its continuous development over the past 15 years, either as a part of larger projects or smaller, focused endeavours. Šolar was conceptualised and its first version compiled as a part of the ‘Communication in

Slovene' project (2008–2013), financed by the European Social Fund and the Slovenian Ministry of Education, Science and Sports.[3] The purpose of the Šolar corpus, as defined in the project, was to enable empirical research into the communication competence of Slovene students in elementary and secondary schools, and, based on the findings, to improve the methods and materials for Slovene language teaching. Subsequent development of the Šolar corpus occurred within the 'Upgrading the Šolar corpus' project (2015–2018), resulting in the creation of Šolar 2.0 and its various versions. The aims of the project were two-fold: a) to improve the text representativeness by region and school level, ensuring better balance, and b) to improve the error categorisation system used in the Šolar corpus and adjust the existing error annotations accordingly.

Recently, two projects provided opportunities for further development of the corpus and corpus-building protocols. The 'Development of Slovene in a Digital Environment' project (2020–2023),[4] financed by the Slovenian Ministry of Culture and the European Regional Development Fund, focused on developing computational tools and services in the field of language technologies for Slovene. One project goal was to evaluate the methodology for the preparation of selected corpora, including Šolar, and establish procedures for their continuous creation. The 'Empirical foundations for digitally-supported development of writing skills' project (2021–2024),[5] financed by the Slovenian Research Agency, aims to create empirical data and digital tools to assist teachers in correcting and grading student writing. One of the outcomes of the project is Šolar 3.0, richly linguistically tagged with state-of-the-art tools, serving as an empirical basis for various research purposes.

3.1 Content

The Šolar corpus was built to enable empirical research into the written language production of students in Slovene elementary and secondary schools. To ensure that the corpus reflected authentic school production and authentic teacher feedback, we only collected texts that were produced as part of the curriculum requirements, rather than prompted for project purposes. In addition, we only collected texts that were produced in the classroom, as we did not want to include texts that the students might have not produced alone. As for the age of the author, we began by collecting texts produced by students aged between 12 and 18, with possible future extensions of this span in mind. Other information on the authors of the texts, such as gender and Special Educational Needs Status was not collected for two reasons: firstly, more information would increase the possibility of identifying the authors after the corpus was made available to the public, and secondly, including more sensitive data would make obtaining permissions from the parents more difficult.

Due to the rich dialectal variation in Slovenia and the potential influence of specific dialectal features on the production in standard Slovene, it was important to

[3] Information about the project is available at http://eng.slovenscina.eu/.
[4] https://slovenscina.eu/en
[5] https://www.cjvt.si/prop/en/

aim for a regionally balanced corpus. Considering the size of the regions, 60% of the corpus texts would optimally come from schools in the southwest of Slovenia, and 40% from schools in the northeast. Furthermore, we attempted to achieve text balance by school (the ratio between elementary and secondary schools, and the ratio between different types of secondary education), grade, and city. Because the vast majority of texts in Slovene schools are produced at the subject of Slovene, balancing the corpus according to the subject seemed less feasible. Nevertheless, we strived to collect texts produced at other subjects (history, philosophy, geography, etc.) and prioritized their inclusion to warrant at least some diversity across the subjects. From the beginning of the corpus creation, the internal balance of the corpus was seen as an ideal, not a prerequisite, as the process highly depended on the voluntary participation of school teachers and on the legal consent of the authors, while the text collection only covered one school year (see Chapter 3.2).

At the end of the first attempt, Šolar 1.0 comprised 2,703 texts (967,477 words) written by Slovene elementary and secondary school students, 56% of which included teacher corrections of language errors. The texts were produced as part of the coursework, mainly at Slovene (82% of the texts). Fewer texts were obtained from other school subjects. The majority of the texts were essays (64%), and the rest were tests (18%) and other written school products such as letters, memos, etc. (18%). The texts were produced by students at gymnasiums (43%), students at technical schools (31%), pupils at elementary schools (19%), and students at vocational schools (7%).

As part of the upgrade, 2782 new texts were added, resulting in 5,485 texts with metadata as presented in Tables 1, 2, 3, and 4. The content of the corpus is the same in versions 2.0 and 3.0, as no new texts were added for version 3.0. Tables 1, 2, 3, and 4 have been divided into two sections. Initially, the data encompassing the entire corpus is outlined in grey cells. Subsequently, the data pertaining to the corpus section containing teacher corrections is displayed in white cells. Across all tables, the count and percentage of texts within a specific category are presented, along with the count and percentage of words found within these texts.

Table 1 outlines the distribution of corpus texts/words across Slovene regions. The texts from the northeast regions (Celje, Maribor, Murska Sobota, Slovenj Gradec) amount to 23.9%, and the texts from southwest regions (Gorica, Koper, Kranj, Krško, Ljubljana, Novo mesto, Postojna) amount to 76.1%. Among the listed categories, Ljubljana (the capital region) exhibits the highest number of texts (1495) and words (453,030), constituting 27.3% and 27.7% respectively. The most underrepresented regions are Murska Sobota with 0.3% words and Postojna with 1.7% words.

Table 2 outlines the distribution of corpus texts/words per school type. The majority of texts originate from different types of secondary schools, while elementary schools account for 19.7% of texts and 16.3% of words. Technical schools and gymnasiums stand out with the highest shares, contributing 41.2% of texts and 37.5% of words and 28.2% of texts and 37.6% of words, respectively. Vocational schools constitute 9.8% of texts and 7.2% of words.

Table 3 provides a comprehensive breakdown of text and word distribution according to grade (primary school) and year (secondary school). The categories related to school type exhibit a relatively balanced representation. Notably, Year 4

Table 1 Number of texts and words per Slovene region in Šolar 3.0

Region	Num. of texts	Perc. of texts (%)	Num. of words	Perc. of words (%)	Num. of corrected texts	Perc. of corrected texts (%)	Num. of words in corrected texts	Perc. of words in corrected texts (%)
Celje	623	11.4	177644	10.9	32	0.6	11084	0.7
Maribor	271	4.9	71258	4.4	92	1.7	27097	1.6
Murska Sobota	43	0.8	4733	0.3	22	0.4	3223	0.2
Slovenj Gradec	372	6.8	97966	6.0	102	1.9	22313	1.4
Gorica	521	9.5	263852	16.1	321	5.8	205477	12.6
Koper	111	2.0	32898	2.0	74	1.3	21420	1.3
Kranj	380	6.9	75524	4.6	10	0.2	501	0.0
Krško	656	12.0	205366	12.6	147	2.7	40,637	2.5
Ljubljana	1495	27.3	453030	27.7	467	8.5	166221	10.2
Novo mesto	924	16.8	224862	13.7	249	4.5	83798	5.1
Postojna	89	1.6	28274	1.7	0	0.0	0	0.0
Together	5485	100.0	1635407	100.0	1516	27.6	581771	35.6

Table 2 Number of texts and words per school type in Šolar 3.0

School type	Num. of texts	Perc. of texts (%)	Num. of words	Perc. of words (%)	Num. of corrected texts	Perc. of corrected texts (%)	Num. of words in corrected texts	Perc. of words in corrected texts (%)
Elementary school	1081	19.7	267146	16.4	395	7.2	110932	6.8
Technical school	2262	41.2	613483	37.5	574	10.4	186809	11.4
Vocational school	540	9.9	117886	7.2	143	2.6	44878	2.8
Gymnasium	1549	28.2	615067	37.6	404	7.4	239152	14.6
Unknown	53	1.0	21825	1.3	0	0.0	0	0.0
Together	5485	100.0	1635407	100.0	1516	27.6	581771	35.6

Table 3 Number of texts and words per grade (primary school) and year (secondary school) in Šolar 3.0

Grade / Year	Num. of texts	Perc. of texts (%)	Num. of words	Perc. of words (%)	Num. of corrected texts	Perc. of corrected texts (%)	Num. of words in corrected texts	Perc. of words in corrected texts (%)
Grade 6	208	3.8	45305	2.8	23	0.4	7685	0.5
Grade 7	229	4.2	54433	3.3	92	1.7	22949	1.4
Grade 8	325	5.9	93628	5.7	132	2.4	43505	2.7
Grade 9	319	5.8	73780	4.5	148	2.7	36793	2.2
Year 1	1024	18.7	317130	19.4	427	7.8	163610	10.0
Year 2	1018	18.5	252775	15.5	236	4.3	108411	6.6
Year 3	870	15.9	308496	18.9	252	4.6	99299	6.1
Year 4	1373	25.0	456196	27.9	181	3.3	92522	5.7
Year 5	86	1.6	21510	1.3	25	0.4	6997	0.4
Matura course	33	0.6	12154	0.7	0	0.0	0	0.0
Together	5485	100.0	1635407	100.0	1516	27.6	581771	35.6

Table 4 Number of texts and words per text type in Šolar 3.0

Text type	Num. of texts	Perc. of texts (%)	Num. of words	Perc. of words (%)	Num. of corrected texts	Perc. of corrected texts (%)	Num. of words in corrected texts	Perc. of words in corrected texts (%)
Classroom work	823	15.0	112107	6.9	201	3.7	31988	1.9
Essay	3218	58.7	1269793	77.6	1280	23.3	547169	33.5
Practical text	691	12.6	71455	4.4	0	0.0	0	0.0
Test	753	13.7	182052	11.1	35	0.6	2614	0.2
Together	5485	100.0	1635407	100.0	1516	27.6	581771	35.6

stands out as the most represented, contributing 25.0% of texts and 27.9% of words. This aligns with the practicality of text production, as Year 4 entails significant writing activity, resulting also in longer texts. Conversely, Year 5 and the Matura course show lower representation due to their limited attendance among secondary school students, thus explaining their lower presence in the corpus.

Table 4 outlines the distribution of corpus texts/words per text type. Essays dominate the corpus with 58.7% of texts and 77.6% of words, followed by classroom work (15.0% texts, 6.9% words), tests (13.7% texts, 11.1% words), and practical texts (12.6% texts, 4.4% words).

3.2 Text collection and legal issues

To compile the corpus, a large number of texts that students have written as a part of their coursework (essays, school tests, etc.) had to be collected from a number of Slovene schools. We wanted to make the corpus openly accessible to researchers and teachers, so the matter of copyright was carefully considered from the very beginning of corpus creation. With the help of legal advisors, contracts were prepared for the authors, in which the authors (or in the case of underaged students, parents/legal representatives) gave permission to the public consortium to use their texts to build a corpus for public use. At the same time, by signing the contract, the consortium partners declared that all personal data in students' texts would be anonymized and protected in accordance with the Slovenian Personal Data Protection Act. For later versions of the corpus, we have provided the necessary changes to make the contracts compliant with the General Data Protection Regulation. We also made changes that facilitated long-term text collection: we broadened the consent from a specific school year to all the texts produced by a certain student in the course of their education at a specific school.

The collection of texts was conducted in close cooperation with teachers from different Slovene schools. They had the task to find contributing authors (most often their students), obtain their written consent, photocopy the texts, and provide metatextual information, i.e., the information about the circumstances in which the texts were produced. The metatextual information included education level (elementary, secondary), school subject, grade (7th year, 8th year, etc.), region of the author's school, and text type. The teachers were encouraged to provide versions of the texts that included corrections of language errors and other potential feedback that they had provided to the students. The teachers participated on a voluntary basis, however, a documented participation in national projects granted them a reference for promotion to a higher title at work. Many teachers were also motivated to participate because they recognized the value of an openly available corpus of student texts for their own needs, e.g., to create teaching materials, exercises, and tests.

For Šolar 1.0, the texts were collected in the school year 2009/2010 with the help of teachers from 39 participating schools. For Šolar 2.0, teachers from 20 schools participated in the school years 2016/2017 and 2017/2018. Based on the lessons learned from the first project (see Chapter 3.3), we decided to substitute photocopying with the scanning of texts. We asked the teachers to deposit the scans together

with all the relevant metadata to the project repository. In rare cases where a teacher needed help with this process, one of the project team members, equipped with a portable scanner, visited them at the school. For Šolar 3.0, no new texts were collected. Instead, we conducted an expert evaluation of the protocols with the help of 20 teachers to further simplify the text collection. As a result, we have designed a portal for an easy, time-efficient upload of texts together with all the necessary metadata.[6] In the portal, the teachers can manage their own contributions and monitor the collective contribution of their school against the goals for a specific corpus upgrade. The uploaded texts are automatically renamed using information from the metadata, making them directly usable for the next steps of corpus compilation.

3.3 Transcription, anonymisation, error and correction annotation

One of the main challenges in the compilation of the Šolar corpus has always been the conversion of the student texts into a corpus-ready format. This process includes transcription, anonymisation, and error/correction annotation. Different approaches have been used in the compilation of different versions; on the basis of lessons learned, the process has been continuously improved. One thing that remained a problem was that the majority of collected texts were handwritten, and as such sometimes difficult to decode, either in parts or in full.

In the compilation of Šolar 1.0, the digitization process often posed a challenge because the received photocopies of student texts were in black and white (not in colour), thus making it difficult to distinguish teacher corrections from student text. Moreover, some photocopies were of such bad quality that the text was almost illegible. However, by far the most demanding and time-consuming part of the transcription proved to be the annotation and categorization of corrected language errors. The transcriber's task was to correctly transcribe a handwritten student text, annotate the errors in the text using XML tags, annotate the teacher corrections of the errors, and categorise each instance using the attribute in the tag. In the collected texts, the teacher interventions were of different types, from underlined text, crossed out or corrected text to comments and suggestions for improvement of style. For Šolar 1.0, annotation recorded various types of teacher interventions in student texts, from textual comments, symbols and formatting corrections to error corrections. The interventions were recorded with the <u> tag, and various attributes within the tag mark the type of intervention. The error corrections also contained the <p> tag with the correction suggested by the teacher.

The transcription was conducted in Microsoft Word, which was selected because transcribers were most familiar with it, and a number of macros were prepared to save transcribers' time. The subsequent evaluation revealed that the transcribers had many difficulties with combining linguistic skills (annotation and categorization of errors/corrections) with technical ones (using XML tags and macros), which resulted in mistakes on the linguistic side, e.g. incorrect category used, and

[6] https://zbiranje.cjvt.si/solar/login/

on the technical side, e.g. incorrect XML format. The former problem was mainly addressed by introducing a thorough check of all the texts with annotations, while the latter was addressed semi-automatically with validation tools after all the texts had been transcribed and checked. This prolonged the compilation process and resulted in the fact that fewer texts were included in the corpus.

Anonymization was also conducted during transcription. We anonymized not only names and surnames of text authors, if mentioned in the text,[7] but also any other names and surnames, names of places, addresses, in short, any information that could be used to identify the author of the text. In order to provide some contextual information in the text, we replaced the anonymized contents with standard strings using the patterns XImeX ('XNameX' in English), XPriimekX ('XSurnameX'), XKrajX ('XPlaceX'), etc.

For Šolar 2.0, many unused texts from the first text collection were included, which meant that the problems from the Šolar 1.0 remained. We decided to digitize all newly included texts, i.e., prepare PDF versions, as that meant the transcribers could easily access the files online as opposed to having to collect (and return) the photocopies. For newly collected texts, the improvements introduced to the text collection process, in particular obtaining colour scans in PDF format from teachers, also facilitated transcription and anonymization. Due to financial constraints, error/correction annotation was not conducted for newly collected texts; only the information whether the text includes teacher corrections or not was recorded. There was also a change in software used for transcription: Microsoft Word was replaced by an XML editor (Notepad++, Oxygen, or a similar tool selected by the transcriber), doing away with the need for a separate XML validation.

For Šolar 3.0, there have not been any new text transcriptions. However, the methodology for upcoming work was significantly enhanced by localizing and adapting the Svala tool, which simplifies several crucial steps in constructing corpora with language corrections (Wirén et al., 2019; Volodina et al., 2019). For the Slovene version, CJVT Svala, we've incorporated transcription, basic anonymisation, and error/correction annotation functions, while deferring automated anonymisation and annotation workflow management to a later phase. The tool was also translated into Slovene and upgraded to accommodate multiple languages and intricate annotation taxonomies, as described in Arhar Holdt et al. (2024).[8]

3.4 Categorisation of teacher corrections

Especially rich and detailed in the Šolar corpus are teacher corrections of language errors at different levels. Similar to other corpus features, there has been significant progress in methodology between the versions. For Šolar 1.0, the annotation system was based on a classification designed for error annotation of Slovene foreign learners' production (Stritar, 2009). We selected this annotation

[7] Names and surnames of text authors were not included as metadata in the document headers.

[8] CJVT Svala is available at https://orodja.cjvt.si/svala/ and as a code on https://github.com/clarinsi/swell-editor.

system as our foundational framework because it was developed in coordination with established practices in the domain of learner corpora and was subsequently piloted for its suitability in annotating the Slovene language (Stritar, 2012 pp. 129–195). This aligned our approach with the traditional annotation practices. However, we decided to enhance the system by establishing annotation subcategories from the ground up, directly from the corpus texts. Our aim was to devise a detailed and granular classification system that accurately captures the authentic occurrences of language errors and corrections, thereby providing immediate value for the creation of educational materials, language exercises, and tests (Arhar Holdt et al., 2017). The outcome of this approach was the categorization of 35,029 language corrections into broad categories, such as orthography, morphology, vocabulary, and syntax, and further into 692 specific bottom-up categories, such as Infinitive and Supine; Agreement in Noun Phrases; Agreement between Verb and Subject; Verbs in Dual, etc. (Kosem et al., 2012: 69). Nonetheless, for Šolar 1.0, a wholesome revision of the bottom-up categories was not conducted, and they remained somewhat scattered, heterogeneous, and poorly documented.

For Šolar 2.0, the annotation categories were revised, upgraded, and arranged into a three-level hierarchical system. While the focus on the specific problems was kept at the third level of the hierarchy, the first two provided a top-down linguistic organization that facilitated the summative use of annotated data and a comparison to more robust annotation systems. At the top level, the categories include Spelling (corrections involving the selection and sequence of vowels and consonants in words), Orthography (corrections related to capitalization, word separation, and punctuation), Morphology (corrections of word forms), Vocabulary (corrections involving the choice of words or phrases), Syntax (corrections of syntactical structures, including word order), and Related Corrections (changes to the text that result directly from the initial correction, e.g., when splitting one sentence into two, the adjustment of capitalization is considered a related correction). A challenging part of the re-categorisation process was technical as there was no tool available at the time which supported multiple-line concordance analysis and correction, by several people simultaneously. Our solution at the time was to use the annotation feature in the Sketch Engine used for Corpus Pattern Analysis (Baisa et al., 2015; Kilgarriff et al., 2004), which allowed the annotation of a keyword (or a related sentence, or in our case a correction category) with certain information. The entire procedure consisted of importing the Šolar corpus (in VERT format) into the Sketch Engine with the annotation option activated, selecting the assigned categories and annotating the corrections with new categories, exporting the annotations, implementing the annotations into the next version of the corpus (XML format), converting the new XML file into VERT, and re-importing the next version into the Sketch Engine for further annotation.

For Šolar 3.0, no new texts were incorporated. However, specific concerns regarding error segmentation from 2.0 were identified and rectified, utilizing the CJVT Svala tool (outlined in Chapter 3.3). At the end of the categorisation process, the Šolar corpus comprises 36,570 annotated teacher corrections.

The annotation guidelines employed for both Šolar 2.0 and 3.0 are presented in Arhar Holdt et al. (2022b). These guidelines provide a detailed description of the categories and offer a number of annotated corpus examples. Table 5 in the Online Appendix summarises the categories structured within a three-level hierarchy, and provides examples of errors and related corrections. The number next to the first- and second-level categories indicates the quantity of their corresponding subcategories.

3.5 Linguistic tags

The Šolar corpus in its various versions has been annotated with the most recent version of the annotation tools available for Slovene. Šolar 1.0 was annotated with the tools developed by the Amebis company, which rely on handwritten rules and data from the Ases database for their operation (Arhar & Holozan, 2009). One of the main benefits of the Amebis' rule-based tagger for the annotation of Šolar was that its rich lexicon included frequent misspellings and non-standard forms, which meant that such forms were tagged as their correct or standard versions, e.g. "živlenje" as a common misspelling of "življenje" was tagged as a noun. The disadvantage of Amebis' tools is that they are not open source,[9] which was one of the reasons other tools were used for the next versions of the Šolar corpus.

Šolar 2.0 was annotated with a tool called Obeliks (Grčar et al., 2012), which was developed in the 'Communication in Slovene' project. The Obeliks tool consists of three modules: a rule-based sentence splitter and tokenizer, a machine-learning morphosyntactic tagger, and a version of the machine-learning LemmaGen lemmatizer (Juršič et al., 2010) which works in combination with the tagger. As Erjavec et al., (2017 p. 146) explain, the tagger, rather than relying only on a model automatically generated from a training corpus, uses "handwritten expert rules, which filter hypotheses generated by the model, and [combine] the results of the lemmatiser and the tagger, assuring that they are not contradictory." Morphosyntactic tags used are based on the specifications developed in the project JOS 'Linguistic Annotation of Slovene' (Erjavec & Krek, 2008). The JOS specifications were based on the MULTEXT specifications (Ide & Véronis, 1994), specifically MULTEXT-East specifications for Slovene, Version 4.

Šolar 3.0 was linguistically annotated with the current state-of-the-art tools for modern Slovene, namely the CLASSLA v1.1.1 pipeline (Ljubešić & Dobrovoljc, 2019) at the levels of tokenization, sentence segmentation, lemmatization, MULTEXT-East V6 morphosyntactic tags, JOS dependency syntax, and named entities.[10] The corpus has additional levels of annotation other than morphosyntactic tagging, increasing the usability of the corpus. In terms of morphosyntactic tags, the difference from the JOS specifications (and MULTEXT-east V4) are some new values for Residual Type attribute and a new tag for Punctuation (Erjavec & Krek, 2018).

[9] The tools were used without any limitations for Šolar 1.0 due to Amebis being a partner in the 'Communication in Slovene' project in which the corpus was created.

[10] https://github.com/clarinsi/classla/

3.6 Formats

One of the aspects of the Šolar corpus that has undergone significant changes over various versions of the corpus has been the corpus format.

Šolar 1.0 was prepared in a customised XML format whose main objective was to provide a merged version of a student text and its "teacher" (corrected) version. This was reflected in separate indexing of student and teacher sentences (<st>) and paragraphs (<pa>). The tags with the corrected student text ended with the number 1, and the tags consisting of the teacher text (corrections) with the number 2, while the "joint" text (the student texts not changed by the teacher) was in tags ending with the number 3. An example is shown in a three-sentence extract in Example 1 in the Appendix. While the number of paragraphs in the student and teacher versions is the same in the corpus, the number of sentences is different; this is a direct result of the fact that teachers sometimes decided to split a student sentence into two sentences, or, less often, merged two sentences into one. The <w3> and <c3> tags contain tokens and punctuation, respectively, from the "joint" text, while error annotations are found in <u1>, <u2> etc. tags, consisting of <w1> tags, containing original student text, and <p1>, <p2> etc. tags consisting of <w2> tags containing teacher corrections. The <u> and <p> tags were numbered to allow the use of multiple types of errors for a certain token or a part of a sentence.

Šolar 1.0 was also converted to the VERT format for the project 'Finalisation of the Šolar corpus' in which the corpus was put in a localised and customised version of the Sketch Engine concordancer, which was made freely available online.[11] We were able to use the conversion tools used with other corpora of Slovene such as the reference corpus of written Slovene Gigafida (Logar et al., 2012), however special adjustments needed to be made to convert the annotations of errors and corrections. In this customised VERT format, each error and correction needed to be recorded separately—in the case of multiple types of errors attributed to a particular part of the text, it meant that the text needed to be repeated, sometimes resulting in a not very user-friendly way of showing the results in the concordancer.

The VERT format was also used for Šolar 2.0, but only for its versions without annotated errors and corrections. In this form, the corpus was ready for inclusion without any customisations into standard corpus concordancers such as noSketchEngine and KonText.

An important milestone for Šolar as far as format is concerned was the move to a standardised format, namely TEI,[12] when preparing Šolar 2.0. The main reason behind this decision was that we wanted to make the corpus more accessible and useful to other interested parties such as researchers, developers and so forth. The conversion process was not without its problems, as error/correction annotation, especially sublevels, again presented a challenge. For this reason, we ended up preparing different versions of Šolar 2.0 (see Sect. 3.7), with some of them fully TEI-compatible (but without error annotations), and some of them in a slightly adapted

[11] http://korpus-solar.net/
[12] https://tei-c.org/

TEI format to enable the recording of annotated errors (see Example 2 in the Online Appendix for a comparison of the formats).

Despite many advantages, (adapted) TEI format still posed problems when the Šolar corpus (especially versions with corrections) was imported into corpus tools. Storing both the student version and the teacher version in one file was far from ideal, and a new solution was sought. Once the decision was made to transition to the Svala tool, we also adopted its approach to data storage for Šolar 3.0. This means that Šolar 3.0 is available in a fully-compatible TEI format, where the original and corrected versions of the texts are encoded separately, while intertextual links with error labels give the relations between the two (see Example 3 in the Online Appendix for an example).

3.7 Availability of Šolar corpora

Different versions of Šolar are available as datasets in the CLARIN.SI repository, under different licenses:

- Šolar 1.0 (Rozman et al., 2013), containing 2703 texts, is available under Creative Commons—Attribution-NonCommercial-ShareAlike 4.0 International license (CC BY-NC-SA 4–0). It contains annotated teacher corrections. http://hdl.handle.net/11356/1036
- ccŠolar 1.0 (Kosem et al., 2019a). This corpus contains 1693 texts collected during 2016–2018, which are a part of Šolar 2.0, for which a more open license was obtained. Therefore, the corpus is available under Creative Commons—Attribution 4.0 International license (CC BY 4.0). http://hdl.handle.net/11356/1224
- Šolar 2.0 (Kosem et al., 2019b), containing 5485 texts, is available under CC BY-NC-SA 4.0 license. This corpus also contains annotated teacher corrections. A part of the corpus documentation are the guidelines for the annotation of teacher corrections. http://hdl.handle.net/11356/1214
- Šolar 2.0 Clear (Kosem et al., 2019c), containing 5485 texts, is available under CC BY-NC-SA 4.0 license. In this corpus, annotated teachers' corrections are not included, but, as mentioned above, it is fully TEI-compatible. http://hdl.handle.net/11356/1219
- Šolar 2.0 Error (Arhar Holdt et al., 2019). This corpus contains 2094 texts from the Šolar 2.0 corpus that include teacher corrections of student errors. It also includes the guidelines for the annotation of teacher corrections. It is available under CC BY-NC-SA 4.0 license. http://hdl.handle.net/11356/1231
- Šolar 3.0 (Arhar Holdt et al., 2022c), containing all 5485 texts with improved format, error segmentation, annotation guidelines, and state-of-the-art linguistic tags, available under CC BY-NC-SA 4.0 license. http://hdl.handle.net/11356/1589

An analysis of the statistics of views and downloads of the above-listed versions shows that four versions of the corpus have attracted interest: Šolar 2.0 Error, and all the versions of the entire corpus (Šolar 1.0, Šolar 2.0, and Šolar 3.0). Šolar 2.0

Error has been downloaded 1006 times from 2020 to June 2024; 277 times the users downloaded the corpus, and 729 times the annotation guidelines. Šolar 1.0 was downloaded 330 times from 2015 to 2019, and Šolar 2.0 was downloaded 161 times in 2021.[13] Šolar 3.0 was downloaded 1072 times since September 2022 when it was first uploaded to the CLARIN.SI repository.

The Šolar corpus has also always been freely available to the research community and other interested users through different concordancers. Since 2014, Šolar 1.0 has been available in a customised version of the Sketch Engine, with the focus being on a user-friendly display of annotated errors and corrections (Kosem et al., 2013). Despite the shortcomings of displaying parts of text annotated with more than one different type of error (Chapter 3.6), the tool proved to be very useful—it featured in nationwide teacher training workshops (see Chapter 4) and was used as the interface of the Lektor corpus,[14] a corpus containing texts with 30,258 corrections made by proofreaders (Popič, 2014).

Šolar 2.0 without annotated errors/corrections has been included in the concordancers available through CLARIN.SI: KonText[15] and two versions of noSketchEngine.[16] Šolar 3.0 is also available in these concordancers in two separate versions, one containing the original student texts and the other the corrected texts. Other than offering access to the most recent version of the Šolar corpus, these tools provide additional options which are not available in the standalone Sketch Engine installation, e.g. keyword extraction using one of the other available corpora as a reference corpus. Corrections are available for viewing as token annotations, similarly as the information on a lemma, part of speech etc. Therefore, despite the improved comparability and connectivity with other corpora, the search functionality and visualisation of data with annotated errors/corrections are still far from optimal (see Chapter 5 for more).

4 Applications of the Šolar corpus

An important attempt to popularise the use of the Šolar corpus in language teaching and learning was its inclusion in the curriculum of a workshop series called 'Language Technology training for teachers'. Workshops, funded by the Ministry of Culture, were conducted from 2012 to 2014 at elementary and secondary schools all over Slovenia and their aim was to inform teachers of Slovene as L1 (but also other teachers) about online reference works, corpora, concordancers and other language tools for Slovene, and teach them how to use them. Post-workshop analysis revealed that many teachers were particularly enthusiastic about using online reference works and the reference corpus of written Slovene Gigafida in class.

[13] This data is provided via the Pickwick Statistics tool, avaliable as part of the CLARIN.SI repository. The tool is currently in the beta version. There is a gap in statistics, with no data available for 2020.

[14] http://korpus-lektor.net/

[15] https://www.clarin.si/kontext/

[16] https://www.clarin.si/noske/ and https://www.clarin.si/ske/

As also confirmed by the workshops, for educators the corpus data can be a valuable resource to enhance teaching materials—finding examples, designing exercises and tests, and prioritizing teaching topics to effectively address language challenges evident in the corpus. However, this practical use of corpora comes with its limitations, notably the considerable time required for material preparation. Hence, the Pedagogical Grammar Portal,[17] a freely available multimodal resource, was developed. It is composed of multiple chapters, each centred on the most common language challenges students face while writing in standard Slovene (Arhar Holdt et al., 2017). Data from three distinct corpora inform its content. The Gigafida corpus (Logar et al., 2012) offers a general explanation of the discussed language phenomena and visualizes supportive language data. Examples from the GOS corpus (Verdonik & Zwitter Vitez, 2011), which captures spoken Slovene, are included to highlight differences between written and spoken Slovene and to showcase specific dialectal features in comparison to the standard language. Simultaneously, the Šolar corpus is employed to provide concrete illustrations of the particularities of the discussed language issues.

In recent years, language teachers and material developers have started to include corpus information in teaching materials more regularly, which has benefitted pupils and students. For example, the value of the Šolar corpus as a source of authentic (and corrected) language use has been evidenced by its inclusion in Slovene grammar textbooks for elementary and secondary schools called 'Kratkoslovnica: slovenska slovnica za osnovno šolo' (Ahačič, 2017a) and 'Slovnica na kvadrat: slovenska slovnica za srednjo šolo' (Ahačič, 2017b), respectively. In these textbooks, each chapter is concluded with interesting findings and example errors in language production from the Šolar corpus. To facilitate further comparable applications and empirical analyses, Šolar 3.0 was utilized to create a Frequency list of language problems (Arhar Holdt et al., 2022d). This list encompasses error codes alongside corresponding student and corrected sentences, along with pertinent metadata. This metadata includes details like text type, author's educational stage, and the school's type and region where the text originated.

Slightly more indirectly, the research using the Šolar corpus has brought benefits to the users of dictionaries and tools for Slovene. For example, Šolar 2.0 was employed to compile the Reference list of Slovene most frequent common words (Pollak et al., 2020). This compilation involved selecting vocabulary at the intersection of the top 10,000 most frequent lemmas from four corpora of Slovene. Alongside Šolar, this included the balanced reference corpus of written Slovene Kres (Logar et al., 2012), the reference corpus of spoken Slovene GOS (Verdonik & Zwitter Vitez, 2011), and the corpus of computer-mediated communication Janes (Fišer et al., 2020). The process of creating the list, which encompasses 4768 common general lemmas, along with their associated part-of-speech labels, relative average reduced frequency across each corpus, and the final average score derived from these values, is described in Arhar Holdt et al. (2020). This list played a pivotal role in constructing the Core vocabulary for Slovenian as a second language, as

[17] http://slovnica.slovenscina.eu/

presented by Klemen et al. (2023). Additionally, the list is used for the frequency filter for collocations in the second version of the Collocations Dictionary of Modern Slovene (Kosem et al., 2023),[18] and in one of the analysis features of the SENTA (Sentence simplification and analysis) tool.[19]

The Šolar corpus has been used in the testing and development of language tools for Slovene. Two examples of such tools are Besana,[20] a commercial grammar checker for Slovene, currently in version 4.29, and Vejice,[21] an automated comma placement tool. Both tools have used the Šolar corpus to test their precision in detecting errors in language production, especially in terms of comma placement (Holozan, 2013, 2015). For other types of language errors, especially at the level of morphology, new methodology was tested (Mokotar, 2023; Petrič, 2022), however, the integration of these methodologies into a comprehensive grammar checker is an ongoing endeavor.

5 Discussion

Acquisition of students' texts has undergone a significant change over various versions of the Šolar corpus, taking into account our experiences and teacher feedback. The current solution using the online portal for text submission facilitates text collection, digitisation, and storage. However, the main issue remains the lack of awareness of the need for constant monitoring of student language production among the top-level decision-makers and funding bodies. As long as text collection efforts will remain limited to occasional projects, it will be very difficult to ensure systematic and continuous analyses of student language production.

An important feature of the Šolar corpus are authentic teacher corrections and other types of teacher feedback. All the original corrections have been included in the digital form, and no additional corrections have been applied. Basing the error annotation solely on teacher corrections is an important methodological decision with considerable consequences for corpus use. When correcting texts, teachers consider student competence and other contextual specifics of the text. The treatment of language errors is thus not entirely consistent and comparable from student to student. However, with realistic corrections, the corpus provides valuable insight into the process of language education in schools, extending its value beyond mere statistics on language errors in student writing.

While basing error annotation on teacher corrections alone maintains a link with the language education process, it does bring certain limitations to the corpus use. For example, in certain texts, corrections are consistently provided only in the initial sections, often likely reflecting the decision of the teacher to leave the remaining errors for the student to discover and correct (also probably due to their repetitive

[18] https://viri.cjvt.si/kolokacije/eng/
[19] https://senta.cjvt.si/en
[20] https://besana.amebis.si/
[21] https://orodja.cjvt.si/vejice/

nature). In some texts that were added in Šolar 2.0 the corrections still have not been annotated due to the time constraints in the project. Therefore, anyone using the corpus, especially for tasks such as testing grammar and spell checkers where precision and recall of error detection is checked, needs to consider that not all the errors in the corpus texts are addressed.

On the other hand, a research avenue that becomes feasible with a corpus containing teacher corrections is the exploration of similarities and discrepancies in how teachers handle related writing problems. Error/correction annotation has revealed occurrences of diverse feedback and instructions, particularly concerning language variation. For instance, if a student selects a morphological variant that reference dictionary resources mark as colloquial, most teachers tend to correct it. However, some teachers also rectify variants that are standard but indicated as less frequently used. This situation raises concerns as it introduces a certain level of inconsistency in the guidelines for developing writing skills and could potentially impact the grading of written production across different schools. The Šolar corpus offers an opportunity for teachers, as well as implementers of faculty programs that educate for the teaching profession, to highlight challenging points in providing feedback, and to pinpoint good practices and solutions.

One problem that has not been addressed since the beginnings of the Šolar corpus was digitization, i.e. converting texts from a hand-written format to a computer-readable format (TXT, XML). In fact, the transcription process for Šolar 1.0 was so time-consuming that only one third of all collected texts made it into the corpus. This problem is shared by many institutions across Europe conducting similar projects, showing the need for an international project that would develop a tool for automatic transcription of (student) texts containing errors.

Considering the size of the research community working with error-annotated corpora,[22] it is surprising there are no standard corpus formats and tools available. Therefore, different researchers and research groups use different combinations of different solutions, which likely results in encountering similar or the same issues others have already experienced (and solved). The efforts by the Swedish research group at Språkbanken, the University of Gothenburg, who have developed the Svala annotation tool, are a step in the right direction as the tool and the format it uses have been developed specifically with the needs of the users of error-annotated texts in mind. As our experience has shown, the annotation of texts in Svala is very user-friendly, maintaining the original and the corrected text separated, and changes made clearly shown and categorised. This division is maintained in the format, with original texts, corrected texts, and changes made all stored separately, the text in XML and the changes in JSON. With Svala being open source and easily localisable (Arhar Holdt et al., 2024), we would strongly recommend this tool and/format to be considered as a possible standard for error-annotated corpora.

In addition, there is a need to develop a specialised corpus tool which would facilitate the analyses of data annotated with errors, corrections, and other types of

[22] We intentionally avoided using "developmental corpora" or "learner corpora" here, as from the technical perspective they have very similar, if not the same, requirements.

related information. Such a corpus tool would need to keep some resemblance to the existing corpus tools to enable knowledge transfer, but should at the same time offer additional functionalities, especially those related to efficient searching and visualisation of errors and corrections. Considering the differences in skills and needs between potential users, e.g., teachers/material developers on the one hand and linguists/researchers on the other, it may also be necessary for the tool to have simple and complex versions of the interface, the former for quick searches and overviews and the latter for complex analyses.

6 Conclusion and future work

In this paper, we presented the history and current status of the Šolar developmental corpus of Slovene. The most recent version, Šolar 3.0, contains over 1,6 million words. Slightly over a third of the texts in the corpus contain annotated teacher corrections. Over the years, improvements have been made in terms of corpus contents, text collection and digitization, corpus format, and annotation. The Šolar corpus has already been used for various purposes, including as a resource in the development of language didactic textbooks and materials, as well as in the testing and development of grammar and spell checkers.

The main focus of the corpus' future development is to expand its contents. This involves increasing the corpus size and improving its representativeness in terms of region, school type, and grade/year of the authors. We plan to expand the corpus contents towards the written production in lower grades on the one hand, and towards university student writing on the other. Our ultimate aim is to establish a regular update schedule, potentially every third school year. To facilitate this, the compilation methodology would need to be further enhanced. Leveraging the predictable patterns of teacher corrections, we could develop machine-assisted methods for identification, annotation, and categorization. Another approach is to integrate corpus building into teacher training programs at the faculty level, benefiting both students and researchers. Moreover, crowdsourcing tasks for a wider public could be considered, provided they are executed with proper quality control and incentives to ensure participation.

There is significant potential in comparing data from the Šolar corpus (student production) with data representing student reception (such as textbooks, youth literature, and user-generated web content). Conversely, there is value in comparing school writing in situations where Slovene is taught as a first language with situations when it is taught as a second/foreign language. These findings and comparisons across different educational levels can aid in constructing empirical descriptors and indicators for tracking the progression of writing skills. Analysing the unique bottom-up approach of the error annotation in the Šolar corpus versus traditional top-down systems, focusing on their respective benefits and limitations, will be a key future step towards comparative corpus studies. Future efforts also encompass adapting the Šolar corpus for machine learning applications, which requires consistent annotation of language errors. Regarding linguistic tagging, we intend to assess how errors impact the precision of state-of-the-art tools at specific tagging levels

and then implement necessary methodological enhancements or cautions. Lastly, to ensure the broadest possible exploitation, disseminating the corpus and offering training in its use are essential steps.

In terms of methodological goals, in particular, international endeavours, solutions, and effective approaches hold significant importance. This applies not exclusively to the field of corpus linguistics but also extends to the broader realm of digital humanities, encompassing tasks such as error-sensitive scanning of handwritten texts and transcription. The overarching objective is the establishment of global benchmarks and openly accessible tools for compiling and visualising corpora that incorporate language errors and corrections. In doing so, we enhance their capacity for comparison and streamline their practicality, which, furthermore, acts as a conduit for the smooth exchange of knowledge and solutions.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10579-024-09758-4.

Acknowledgements The work on this publication was financially supported by the Slovenian Research and Innovation Agency (ARIS) via the core programme Language Resources and Technologies for Slovene (P6-0411) and the project Empirical Foundations for Digitally Supported Development of Writing Skills (J7-3159).

Author contributions This paper was prepared in agreed cooperation between Š.A.H. (first author) and I.K. (lead author). Š.A.H. prepared Chapters 1, 2, 3.1, 3.2, and 3.4. I.K. prepared Chapters 3.3, 3.5, 3.6, and 3.7. Chapters 4, 5, and 6 were equally co-authored. Both authors reviewed the manuscript.

Data availability The resources described in this paper are accessible through the CLARIN.SI repository, each governed by distinct open licenses as outlined in Chapter 3.7 of the manuscript. These references are appropriately documented in the Reference list, following the guidelines set by the journal.

Declarations

Conflict of interest The paper's preparation received financial support from the Slovenian Research and Innovation Agency (ARIS) through the core program Language Resources and Technologies for Slovene (P6-0411) and the project Empirical Foundations for Digitally Supported Development of Writing Skills (J7-3159). The resource being described was developed through several research projects, which are duly referenced in the manuscript. The dissemination of our work via this publication could have a positive impact on securing future research grants. Apart from that, the authors have no financial or proprietary, or any other competing interests in any material discussed in this article.

Ethical approval The authors affirm that the presented research has been conducted in accordance with ethical standards.

References

Abel, A., Glaznieks, A., Nicolas, L., & Stemle, E. (2014). KoKo: An L1 learner corpus for German. In N. Calzolari et al. (Eds.), *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland (pp. 2414–2421). European Language Resource Association (ELRA).

Ahačič, K. (2017a). *Kratkoslovnica: slovenska slovnica za osnovno šolo* (1st ed.). Rokus Klett.

Ahačič, K. (2017b). *Slovnica na kvadrat: slovenska slovnica za srednjo šolo* (1st ed.). Rokus Klett.

Arhar, Š, & Holozan, P. (2009). Leksikalna podatkovna zbirka ASES (Amebisov skupni elektronski slovar). In V. Mikolič (Ed.), *Jezikovni korpusi v medkulturni komunikaciji* (pp. 30–51). Založba Annales in Zgodovinsko društvo za južno Primorsko.

Arhar Holdt, Š., Erjavec, T., Kosem, I., & Volodina, E. (2024). Towards an ideal tool for learner error annotation. In N. Calzolari et al. (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italy (pp. 16392–16398). ELRA and ICCL.

Arhar Holdt, Š., Goli, T., Lavrič, P., Laskowski, C., Klemenc, B., Rozman, T., Stritar Kučuk, M., Krek, S., Krapš Vodopivec, I., Stabej, M., & Kosem, I. (2019). *Error-annotated developmental corpus Šolar 2.0 Error*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1231

Arhar Holdt, Š, Kosem, I., & Stritar Kučuk, M. (2022a). Metode in orodja za lažjo pripravo korpusov usvajanja jezika. In N. Pirih Svetina & I. Ferbežar (Eds.), *Na stičišču svetov: slovenščina kot drugi in tuji jezik. Zbirka Obdobja.* Ljubljana University Press.

Arhar Holdt, Š, Kosem, I., & Gantar, P. (2017). Corpus-based resources for L1 teaching: the case of slovene. In A. Marcus-Quinn & T. Hourigan (Eds.), *Handbook on digital learning for K-12 schools* (pp. 91–113). Springer.

Arhar Holdt, Š., Lavrič, P., Roblek, R., & Goli, T. (2022b). *Categorizing teachers' corrections: Guidelines for annotating the šolar corpus. Version 1.1.* Prepared in the project Development of Slovene in a Digital Environment. Retrieved August 15, 2023, from https://wiki.cjvt.si/books/2-razvojni-korpus-solar/page/oznacevalne-smernice

Arhar Holdt, Š., Pollak, S., Robnik Šikonja, M., & Krek, S. (2020). Referenčni seznam pogostih splošnih besed za slovenščino. In D. Fišer, & T. Erjavec (Eds.), *Proceedings of the Conference on Language Technologies and Digital Humanities*, Ljubljana, Slovenia (pp. 10–15). Institute of Contemporary History.

Arhar Holdt, Š., Rozman, T., Stritar Kučuk, M., Krek, S., Krapš Vodopivec, I., Stabej, M., Pori, E., Goli, T., Lavrič, P., Laskowski, C., Kocjančič, P., Klemenc, B., Krsnik, L., & Kosem, I. (2022c). *Developmental corpus Šolar 3.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1589

Arhar Holdt, Š., Rozman, T., Stritar Kučuk, M., Krek, S., Krapš Vodopivec, I., Stabej, M., Pori, E., Goli, T., Lavrič, P., Laskowski, C., Kocjančič, P., Klemenc, B., Krsnik, L., Žagar, A., & Kosem, I. (2022d). *Frequency list of language problems from Šolar 3.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1716

Arnardóttir, Þ., Xu, X., Guðmundsdóttir, D., Stefánsdóttir, L. B., & Ingason, A. K. (2021). Creating an error corpus: Annotation and applicability. In M. Monachini, & M. Eskevich (Eds.), *CLARIN Annual Conference Proceedings* (pp. 59–63). Virtual Edition.

Baisa, V, El Maarouf, I., Rychlý, P., & Rambousek, A. (2015). Software and data for corpus pattern analysis. In A. Horák, P. Rychlý, & A. Rambousek (Eds.), *Proceedings of recent advances in Slavonic natural language processing (RASLAN 2015)* (pp.75–86). Tribun EU.

Barbagli, A., Lucisano, P., Dell'Orletta, F., Montemagni, S., & Venturi, G. (2016). CItA: An L1 Italian learners corpus to study the development of writing competence. In N. Calzolari et al. (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016),* Portorož, Slovenia (pp. 88–95). European Language Resources Association (ELRA).

Berkling, K. (2016). Corpus for children's writing with enhanced output for specific spelling patterns (2nd and 3rd grade). In N. Calzolari et al. (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016),* Portorož, Slovenia (pp. 3200–3206). European Language Resources Association (ELRA).

Berkling, K. (2018). A 2nd longitudinal corpus for children's writing with enhanced output for specific spelling patterns and evaluation. In N. Calzolari et al. (Eds.), *Proceedings of the Eleventh*

International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan (pp. 2262–2268). European Language Resources Association (ELRA).

Borghi, C. C. (2013). *Analisi di produzioni scritte. Valutazioni e misure automatizzate di elaborati scolastici.* Tesi di dottorato in pedagogia sperimentale, Università di Roma.

Doquet, C., Enoiu, V., Fleury, S., & Maziotti, S. (2017). Problèmes posés par la transcription et l'annotation d'écrits d'élèves. *Corpus, 16*, 133–156. https://doi.org/10.4000/corpus.2776

Durrant, P. (2019). *Growth in grammar corpus 2015-2019.* [Data Collection]. UK Data Service. 10.5255/UKDA-SN-853809

Erjavec, T., Holozan, P., & Ljubešić, N. (2017). Language technologies and corpus encoding. In V. Gorjanc, P. Gantar, I. Kosem, & S. Krek (Eds.), *Dictionary of modern Slovene: Problems and solutions* (pp. 140–154). Ljubljana University Press, Faculty of Arts.

Erjavec, T., & Krek, S. (2008). Oblikoskladenjske specifikacije in označeni korpusi JOS. In T. Erjavec & J. Žganec Gros (Eds.), *Zbornik Šeste konference Jezikovne tehnologije.* Institut Jožef Stefan.

Erjavec, T. & Krek, S. (2018). *MULTEXT-East Morphosyntactic specifications, 3.6. Slovene specifications.* Retrieved August 14, 2023, from http://nl.ijs.si/ME/V6/msd/html/msd-sl.html

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

Garcia-Debanc, C., Ho-Dac, L.-M., Bras, M., & Rebeyrolle, J. (2017). Vers l'annotation discursive de textes d'élèves. *Corpus*. https://doi.org/10.4000/corpus.2783

Glaznieks, A., Frey, J. C., Stopfner, M., Zanasi, L., & Nicolas, L. (2022). LEONIDE: A longitudinal trilingual corpus of young learners of Italian, German and English. *International Journal of Learner Corpus Research, 8*(1), 97–120.

Granger, S. (1998). The computer learner corpus: A versatile new source of data for SLA research. In S. Granger (Ed.), *Learner english on computer* (pp. 3–18). Addison Wesley Longman.

Granger, S. (2008). Learner corpora. In A. Ludeling & M. Kyto (Eds.), *Corpus linguistics. An international handbook*. Mouton de Gruyter.

Grčar, M., Krek, S., & Dobrovoljc, K. (2012). Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. In T. Erjavec & J. Žganec Gros (Eds.), *Zbornik Osme konference Jezikovne tehnologije.* Institut Jožef Stefan.

Ho-Dac, L.-M., Fleury, S., & Ponton, C. (2020) É:calm resource: a resource for studying texts produced by French pupils and students. In N. Calzolari et al. (Eds.), *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, France (pp. 4327–4332). European Language Resources Association (ELRA).

Holozan, P. (2013). Uporaba strojnega učenja za postavljanje vejic v slovenščini. *Uporabna Informatika, 21*(4), 196–209.

Holozan, P. (2015). Možnosti uporabe jezikovnih tehnologij za določanje težav pri rabi vejice. In H. Dobrovoljc & T. Lengar Verovnik (Eds.), *Pravopisna razpotja: razprave o pravopisnih vprašanjih.* Založba ZRC.

Ide, N., & Véronis, J. (1994). MULTEXT: Multilingual Text Tools and Corpora. *Proceedings of the 15th International Conference on Computational Linguistics, COLING '94*. Kyoto, Japan (pp. 588–592).

Ingason, A. K., Arnardóttir, Þ., Stefánsdóttir, L. B., & Xu, X. (2021). *The icelandic child language error corpus (IceCLEC) Version 1.1*, CLARIN-IS, http://hdl.handle.net/20.500.12537/133

Jacques, M.-P., & Rinck, F. (2017). Un corpus de littéracie avancée: résultat et point de départ. *Corpus*. https://doi.org/10.4000/corpus.2806

Juršič, M., Mozetič, I., Erjavec, T., & Lavrač, N. (2010). LemmaGen: Multi-lingual lemmatisation with induced Ripple-Down rules. *Journal of Universal Computer Science, 16*(9), 1190–1214.

Kilgarriff, A., Rychlý, P., Smrz, P., & Tugwell, D. (2004). The sketch engine. In G. Williams, & S. Vessier (Eds.), *Proceedings of the Eleventh EURALEX International Congress*, Lorient, France (pp. 105–116). Universite de Bretagne-sud.

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. In M. Medveď, M. Měchura, C. Tiberius, I. Kosem, J. Kallas, M. Jakubíček, & S. Krek (Eds.), *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference*, Brno (pp. 654–668). Lexical Computing CZ. https://elex.link/elex2023/wp-content/uploads/118.pdf.

Kosem, I., Arhar Holdt, Š., Gantar, P., & Krek, S. (2023). Collocations Dictionary of Modern Slovene 2.0. In M. Medveď, M. Měchura, C. Tiberius, I. Kosem, J. Kallas, M. Jakubíček, & S. Krek (Eds.), *Electronic lexicography in the 21st century (eLex 2023): Invisible lexicography. Proceedings of the eLex 2023 conference*, Brno (pp. 491–507). Lexical Computing CZ. https://elex.link/elex2023/wp-content/uploads/100.pdf

Kosem, I., Arhar Holdt, Š., Stritar Kučuk, M., Krek, S., Krapš Vodopivec, I., Stabej, M., Pori, E., Goli, T., Lavrič, P., Laskowski, C., Kocjančič, P., Klemenc, B., & Rozman, T. (2019b). *Developmental corpus Šolar 2.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1214

Kosem, I., Arhar Holdt, Š., Stritar Kučuk, M., Krek, S., Krapš Vodopivec, I., Stabej, M., Kocjančič, P., Laskowski, C., Klemenc, B., Pori, E., & Rozman, T. (2019c). *Developmental corpus (without language corrections) Šolar 2.0 Clear*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1219

Kosem, I., Baisa, V., Kovář V. & Kilgarriff, A. (2013). The sketch engine interface for a learner corpus annotated with errors and corrections. In K. Tenfjord, A. Golden, F. Meunier, & K. De Smedt (Eds.), *Book of abstracts learner corpus research 2013*, Norway, September 2013 (pp. 82–84).

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.), *Proceedings of the conference on language technologies & digital humanities, September 29th–October 1st, 2016 Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia* (pp. 95–100). Ljubljana University Press, Faculty of Arts. http://www.sdjt.si/wp/wp-content/uploads/2016/09/JTDH-2016_Kosem-et-al_Solar-2-0-nadgradnja-korpusa-solskih-pisnih-izdelkov.pdf.

Kosem, I., Rozman, T., Pori, E., Arhar Holdt, Š., Kocjančič, P., Laskowski, C., & Klemenc, B. (2019a). *Developmental corpus ccŠolar 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1224

Kosem, I., Rozman, T., & Stritar Kučuk, M. (2011). How do Slovenian primary and secondary school students write and what their teachers correct: a corpus of student writing. In *Proceedings of the corpus linguistics conference 2011*, ICC Birmingham, 20–22 July 2011. University of Birmingham.

Kosem, I., Stritar Kučuk, M., Može, S., Zwitter Vitez, A., Arhar Holdt, Š, & Rozman, T. (2012). *Analiza jezikovnih težav učencev: korpusni pristop*. Trojina, zavod za uporabno slovenistiko.

Laarmann-Quante, R., Dipper, S., & Belke, E. (2019). The making of the Litkey Corpus, a richly annotated longitudinal corpus of German texts written by primary school children. In A. Friedrich, D. Zeyrek, & J. Hoek (Eds.), *Proceedings of the 13th Linguistic Annotation Workshop*, Florence, Italy (pp. 43–55). Association for Computational Linguistics.

Leech, G. (1997). Teaching and language corpora: A convergence. In A. Wichmann, S. Fliegelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 1–23). Routledge.

Ljubešić, N., & Dobrovoljc, K. (2019). What does neural bring? Analysing improvements in morphosyntactic annotation and lemmatisation of Slovenian, Croatian and Serbian. In T. Erjavec et al. (Eds.), *Proceedings of the 7th workshop on Balto-Slavic natural language processing*, Florence, Italy (pp. 29–34). Association for Computational Linguistics.

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

Marconi, L., Ott, M., Pesenti, E., Ratti, D., & Tavella, M. (1993). *Lessico elementare: dati statistici sull'italiano scritto e letto dai bambini delle elementari*. Zanichelli.

Martins, M., Janssen, M., Santos, T., Lopes, R., & Souza, T. (2020). *DOESTE v0.5*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-3262.

Mokotar, R. (2023). *Obvladovanje slovničnih napak v šolskih pisnih izdelkih z metodami za obdelavo naravnega jezika*. BA thesis. https://repozitorij.uni-lj.si/IzpisGradiva.php?id=144932

Osborne, J. (2002). Top-down and bottom-up approaches to corpora in language teaching. In U. Connor & T. A. Upton (Eds.), *Applied corpus linguistics: A multidimensional perspective* (pp. 251–265). Rodopi.

Pala, K., Rychlý, P., & Smrž, P. (2003). Text Corpus with Errors. In V. Matoušek & P. Mautner (Eds.), *Text, speech and dialogue (TSD 2003) lecture notes in computer science, 2807* (pp. 90–97). Springer.

Parr, J. M. (2010). A dual purpose data base for research and diagnostic assessment of student writing. *Journal of Writing Research, 2*(2), 129–150.

Petrič, T. (2022). *Predlogi jezikovnih popravkov v slovenščini z modelom SloBERTa*. BA thesis. https://repozitorij.uni-lj.si/IzpisGradiva.php?id=139688

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

Popič, D. (2014). Revising translation revision in Slovenia. In T. Mikolič Južnič, K. Koskinen, & N. Kocijančič Pokorn (Eds.), *New horizons in translation research and education 2*. University of Eastern Finland.

Rozman, T., Stritar Kučuk, M., Kosem, I., Krek, S., Krapš Vodopivec, I., Arhar Holdt, Š., & Stabej, M. (2013). *Learners' corpus Šolar 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1036

Sampson, G. (2003). *The LUCY Corpus: Documentation.* University of Sussex. Retrieved August 15, 2023, from https://www.grsampson.net/LucyDoc.html

Stritar Kučuk, M. (2009). Slovene as a foreign language: The pilot learner corpus perspective. *Slovenski Jezik - Slovene Linguistic Studies, 7*, 135–152.

Stritar Kučuk, M. (2012). *Korpusi usvajanja tujega jezika.* Zveza društev Slavistično društvo Slovenije: Znanstvena založba Filozofske Fakultete.

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

Volodina, E., Granstedt, L., Matsson, A., Megyesi, B., Pilán, I., Prentice, J., Rosén, D., Rudebeck, L., Schenström, C.-J., Sundberg, G., & Wirén, M. (2019). The SweLL language learner corpus: From design to annotation. *Northern European Journal of Language Technology, 6*, 67–104.

Wirén, M., Matsson, A., Rosén, D., & Volodina, E. (2019). SVALA: Annotation of second-language learner text based on mostly automatic alignment of parallel corpora. In I. Skadina & M. Eskevich (Eds.), *Selected papers from the CLARIN Annual Conference 2018* (pp. 227–239). Linköping University Electronic Press.

Wolfarth, C., Ponton, C., & Totereau, C. (2017). Apports du tal à la constitution et à l'exploitation d'un corpus scolaire. *Corpus*. https://doi.org/10.4000/corpus.2796