How do Slovenian primary and secondary school students write and what their teachers correct: a corpus of student writing

Iztok Kosem[1], Tadeja Rozman[1], Mojca Stritar[2]

[1] Trojina, Institute for Applied Slovene Studies
[2] University of Ljubljana, Faculty of Arts

e-mails: iztok.kosem@trojina.si, tadeja.rozman@trojina.si, mojca.stritar@gmail.com

1. Introduction

In the past decades, corpora have established themselves as an important tool in language analysis. This is evident in the increase in the number of corpus-based research, and in the number of different disciplines, in which corpora are used (e.g. corpus linguistics, lexicography, literary studies, forensic linguistics, teaching). Corpora as large collections of actual language use represent a more reliable alternative to intuition; as Sinclair (1995) argues, "there are many facts about language that cannot be discovered by just thinking about it, or even reading and listening very intently."

In addition to being used for analysis and description of language of adult native speakers, corpora have been used since late 1980s for analyses of language used by language learners, e.g. young native speakers and L2 learners (Granger, 2004). Most interesting group of language users for researchers and authors of learning materials have been L2 learners of English, which is a direct consequence of the popularity and importance of the English language across the world. This is also supported by the fact that the largest learner corpora contain texts produced by L2 learners of English, e.g. the 35-million-word Cambridge Learner Corpus, the 25-million-word Hong Kong University of Science and Technology Learner Corpus, the 10-million-word Longman Learners' Corpus, and the 3,7-million-word International Corpus of Learner English (Granger et al., 2009). An important corpus database for analyzing the language of young native speakers is CHILDES (http://childes.psy.cmu.edu), which contains 130 spoken

corpora in more than 20 different languages. To date, the database has been used in over 3000 studies.

Corpora of Slovene have been built since the late 1990s, with the developers focusing on reference corpora such as the 100-million-word FIDA corpus, its upgrade, the 620-million-word FidaPLUS corpus (www.fidaplus.net), and the 162-million-word Nova Beseda corpus. More recent corpora, built as part of the "Communication in Slovene" project (www.slovenscina.eu), are the 1,13-billion-word Gigafida corpus (demo.gigafida.net) and the million-word GOS corpus of spoken Slovene (www.korpus-gos.net). Consequently, the majority of corpus-based research has focused on the language of adult native speakers. Corpora of texts produced by young native speakers or L2 learners of Slovene, i.e. the groups in most need of language support, are virtually non-existent. One of the rare corpora of this type is PiKUST, a corpus of texts produced by L2 learners of Slovene (Stritar, 2009).

Šolar, a corpus of student texts, is therefore a much needed and important addition to the corpora of Slovene. The corpus was built within the Communication in Slovene project, to serve as the basis for corpus-based pedagogic grammar. It contains texts produced by students in elementary and secondary schools. It is the first Slovene corpus of its kind and has been developed by taking learner corpora as a model. The texts in the corpus have two important features: they were produced as part of the curriculum rather than solely for project purposes, and they contain error tags and corrections that were made by teachers rather than researchers.

This paper first describes the design and development of the Šolar corpus, and some of the decisions made during the project. Corpus contents are also presented, including metatextual information such as text types, education level, etc. Annotation of teacher corrections is presented in more detail, with most attention being paid to error corrections. Then, the corpus analysis, conducted for the purposes of pedagogic grammar, is described, with initial findings provided. Finally, the conclusion discusses the value of the corpus for teachers, students and researchers, and outlines the plans for the future.

2. Design principles

The Šolar corpus was built to enable empirical research into written language production of students in elementary and secondary schools. It is expected that the analysis of the corpus will help in detecting language problems that students have when writing, and consequently provide the basis for the preparation of didactic solutions. Namely, the teaching of Slovene as native language is – despite being directed at the development of language ability – still too focussed on the language system and does not pay enough attention to the contents that pose problems for students. The Šolar corpus is also a key element in the development of a corpus-based pedagogic grammar (Arhar Holdt et al., 2011), one of the activities in the "Communication in Slovene" project, which will be developed considering the language needs of its target users, students aged 12 and above.

The collection of texts was conducted in several stages. Even before the collection, the decisions had to be made on which types of texts will be collected and how the collection will be done. These decisions were made by considering the project goals and the purposes for which the corpus would be used. In order to ensure that the corpus contents reflected actual school production, the texts had to be produced at different school subjects as part of the curriculum requirements, rather than 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, i.e. the texts produced with the help of adults (e.g. parents) or by copying the contents from other sources, such as the web. Another selection criterion was age of the author – we only collected texts produced by students aged between 12 and 18; which is also the target group of the corpus-based pedagogic grammar.

We wanted to make the corpus freely accessible to researchers and teachers, so the matter of copyright was carefully considered from the very beginning of the project. With the help of legal advisors, contracts were prepared for the authors, in which the authors (or in the case of under aged students, parents/legal representatives) gave the 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.

3. Building the Šolar corpus

The collection of texts was conducted over one academic year (from fall 2009 to spring 2010) with the help of teachers at participating schools (altogether 59 teachers from 39 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; all this information is provided in the document header. At the end of the collection process, 8594 texts in total were collected.

Due to time and financial constraints of the project, not all the collected texts could be included in the corpus. Most texts were written by hand, which meant that they had to be transcribed. In addition, texts had to be anonymized to ensure personal data protection, and assigned unique identification codes to enable the validation of digitalization process. The examination of the collected texts revealed that 14% (more than 1000 texts) were not suitable for inclusion in the corpus, for example they did not contain the required information (e.g. the information on education level or grade was missing), were too short, contained only graphics (e.g. mind maps), or the photocopies were of poor quality.

Slovene is dialectally very diverse language, with dialects being specific to particular regions, so the main criterion for text selection was regional balance as this ensured that the corpus would be representative of the Slovene language. Thus, approximately 60% of the texts in the corpus come from schools in the southwest of Slovenia, and 40% of the texts come 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 were

produced at the subject of Slovene, all the texts produced at other subjects (history, philosophy, business and management, geography, etc.) were included in the corpus to ensure some diversity across the subjects.

The Šolar corpus contains 2703 texts, comprising approximately 1 million words. The texts were produced at 11 different subjects, with 82.3% at Slovene. The texts in the corpus are divided in essays, written products, and tests. Essays, which were written in class and graded, represent the majority of texts in the corpus, namely 64.2%. Written products, representing 18% of texts, were produced at Slovene as part of the lesson and were in most cases not marked; they include summaries, descriptions, formal letters, etc. Tests are divided into two groups: answers to questions are classic tests with questions and (longer) answers, produced at different subjects, while longer tests consist of different practical texts (e.g. letters of request) that students had to produce as part of a longer test at Slovene.

4. Teacher corrections

One of the features that makes the Šolar corpus unique not only in the Slovene context but also in the international context is the fact that error annotation of texts is not based on the error definitions of corpus designers but on the corrections made by teachers. Thus, in addition to being representative of the language production of Slovene students, the corpus also shows the language practice of Slovene teachers, reflected in the actual correction that takes place in a typical classroom environment on a daily basis.

Annotation was added to approximately 50% of texts in Šolar and 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 atributes within the tag mark the type of intervention. The error corrections also contain the <p> tag which contains the correction suggested by the teacher. Tag structures of different types of teacher interventions are presented in Table 1.

Type of teacher intervention	Tag structure
Textual comment	**<u k="*teacher comment*">***student text***</u>**
Symbol	**<u k="*symbol*">***student text***</u>**
Error correction	**<u type="*error type*" sub-type="*error sub-type*">***student text*** ***teacher correction*** </u>**
Illegible student text	**<u n=""/>**
Partially legible student text[1]	**<u n="***abc***"/>**
Illegible teacher correction/comment	**<u kn="">***student text***</u>**
Partially legible teacher correction/comment[1]	**<u kn="*abc*"/>**

Table 1: Teacher interventions and tags in Šolar

When possible and reasonably probable, all types of interventions were interpreted as error corrections to facilitate automatic error analysis; for instance, when the teacher crossed out a part of the text, this was tagged as an error of redundant text rather than a symbol comment. The information about certain other, linguistically irrelevant teacher marks (e.g. the number of words, legibility corrections, or content corrections regarding non-linguistic subjects) was not recorded in the corpus.

Textual comments and symbols

To inform students of allegedly unsuitable parts of their texts or to mark those elements for their own purposes, teachers add numerous comments when correcting student writing. These comments can be textual and/or in the form of various, usually commonly known symbols. Textual comments are notes written by teachers above, under or beside the student text, comprising single words, longer comments, or mere question marks or exclamation marks, such as: <u k="*too informal!*">.[2] Also relatively frequently used in the correction process are symbols, i.e. graphic signs commenting on the content aspect of student writing,

[1] The legible part, e.g. part of the word/phrase, is transcribed.
[2] All examples of tags were originally written in Slovene and were translated into English by the authors of this paper.

for example different types of underlining, arrows or parenthesis: <u l="*underlined"*>luksuz</u>.

Formatting corrections

Formatting corrections address various formatting issues in student texts, e.g. lack of paragraphs or indentation. In the original texts, the corrections were textual or in the form of established symbols, however in the corpus they were recorded uniquely; for instance, the teacher either wrote that the student should start a new paragraph at a certain point or simply used a graphic sign, but in the corpus both forms of correction were recorded with the tag <u obl="*new paragraph*">. This type of correction is particularly important in practical texts (e.g. formal letters), which have more specific tags regarding their formal aspect, such as justification of date or subject line.

Error corrections

Linguistically most important part of Šolar annotation are the error tags. They are based on a classification designed for error tagging of Slovene foreign learners' production (cf. Stritar, 2009), which was adapted to suit the specifics of native-speaker writing. For instance, the subcategory of abbreviations had to be added to the category of orthography, since these errors did not appear in the writing of non-native speakers.

The error classification tends to be formal, general, objective, descriptive and non-interpretative, especially since its intention is only to record and systematize the error corrections applied by teachers. Before the beginning of the error tagging process, which was performed along with the transcription, a manual with detailed instructions, particularly for the cases that could be interpreted ambiguously, was prepared to ensure consistency among the transcribers. To avoid subjective interpretation, there is no differentiation between slips, mistakes and errors in Šolar (cf. James, 1998).

The classification has 4 main types: orthography, vocabulary, morphology, and syntax. Orthography and syntax have six and four sub-types respectively. Orthography errors include various more formal or spelling aspects of writing and have the following subtypes: spelling (e.g. *bljižnih* instead of *bližnjih*),

writing together or apart (e.g. *neglede* instead of *ne glede*), lower or capital case (e.g. *mesto tebe* instead of *mesto Tebe*), punctuation (e.g. missing or redundant commas), abbreviations (e.g. *oz.* instead of *oziroma*), and numerals (e.g. *30* instead of *trideset*). Vocabulary errors include different lexical problems, such as the use of grammatically, semantically or pragmatically unsuitable words or phrases (e.g. *vzrok* instead of *razlog*), erroneous word formation etc. Morphology errors are ascribed to erroneous forms of words with declensional endings, i.e. wherever the wrong grammatical number, tense, case etc. are used (e.g. *možnostima* instead of *možnostma*). Finally, syntax errors include erroneous syntagmatic structures and have subtypes of word order (e.g. *šel domov je* instead of *šel je domov*), missing text (e.g. *Abel pa pasel* instead of *Abel pa je pasel*), redundant text (e.g. *izpove ljubezen do njega* instead of *izpove ljubezen*) and erroneous structure (e.g. *rek Jezusa* instead of *Jezusov rek*). If necessary, occurrences in texts have two or more types of error tags, for example when a word is spelled incorrectly and is at the same time part of an incorrect word order.

Also important is the concept of related errors. Using the attribute *pov* in the error tag, secondary errors were tagged; in the case of these errors, (originally) correct forms have to be corrected after an erroneous form in their vicinity has been changed. In the following example, the teacher replaced the student's comma with a full stop, so the lower case word following the comma becomes the sentence-initial word and has to be capitalized:

> Student text: Vendar ta stališča teh oseb niso vedno dobra**, naj** jih potrdim s primerom iz mojega življenja.
>
> Corrected text: Vendar ta stališča teh oseb niso vedno dobra**. Naj** jih potrdim s primerom iz mojega življenja.

5. Preliminary corpus analysis, based on teacher corrections

The Šolar corpus can be used for various analyses of student writing. The very first error analysis on the corpus was conducted for the purposes of the corpus-based pedagogic grammar, which will address the most common language problems of Slovene primary and secondary school students. The

starting point of the analysis was the basic error classification in Šolar (shown in Table 2), according to which 35,029 errors were tagged.

Error type	Error sub-type	No. of related errors	No. of all errors
Orthography	Spelling	18	2672
	Together/apart	/	1179
	Lower/capital case	872	2125
	Punctuation	775	15,371
	Abbreviation	/	23
	Numeral	/	50
Vocabulary	/	434	3807
Morphology	/	1241	3618
Syntax	Word order	212	1265
	Missing text	240	1607
	Redundant text	588	2665
	Erroneous structure	61	653

Table 2: Error distribution in Šolar

Since the error tags in Šolar are based on teachers' subjective definitions of error, not all are necessarily relevant for the corpus-based pedagogic grammar. Furthermore, the main four error categories are relatively heterogeneous, open and extensive in terms of number of corpus occurrences, so a subsequent manual error analysis was performed. Errors within the main categories were regrouped into more specific subcategories based on common morphological, semantic etc. characteristics, so a new, more in-depth but at the same time less universal error classification was designed, in which for instance morphological errors have 58 and syntactical errors have 328 subcategories. Again, not all of these subcategories are relevant for the new grammar, for example some are stylistic, limited to individual student's language use. Eventually, approximately 300 language problems were identified in the Šolar corpus, and among others include:

- use of modal verbs *moči* (could) and *morati* (must),
- use of possessive and reflexive-possessive pronouns *moj, tvoj, njegov* etc. and *svoj* (similar to *my* and *mine*),
- use of prepositions *z*, *iz*, *v*, *na* (from, in, on),
- use of pronouns *nobeden* (no one) and *nihče* (nobody),
- declension of nouns ending in –o,
- use of comparative and superlative forms of adjectives.

After correlating this data with two additional sources of information (online forums on language issues and a written survey among Slovene language teachers), most salient language problems were selected, which will be analysed and explained in the corpus-based grammar (cf. Arhar Holdt et al., 2011).

6. Conclusion

Language use of Slovene elementary and secondary school students has so far been relatively neglected by researchers, especially corpus linguists. Consequently, there is currently a lack of corpus-based language resources (e.g. dictionaries, grammars) and teaching material for Slovene at elementary and secondary level. Thus, the Šolar corpus of student writing, first of its kind in Slovenia, is an important resource that should help in filling this gap by enabling researchers to gain insight into the student writing and identifying common language problems of students. Šolar is also valuable due to annotated teacher corrections that represent partial analysis of the corpus and provide quick overview of student errors, albeit based on rather general categories.

The value of the Šolar corpus, and the teacher corrections it contains, has been confirmed by the results of the error analysis, conducted for the purposes of a corpus-based pedagogic grammar (Arhar Holdt et al., 2011); the analysis of teacher corrections has identified 300 language problems and most salient ones will constitute the basis for the pedagogic grammar. However, these initial findings should still be reviewed by a more focused error analysis. A further analysis of teacher corrections is also needed, e.g. which language issues do they

focus on during the correction process or how deeply do they interfere with their students' writing style.

Other potential uses of the corpus include the development of tools for detection and annotation of errors, which would speed up the creation and analysis of learner corpora of Slovene. The Šolar corpus could also be used for comparative analyses – using the existing corpora of adult native speakers of Slovene (e.g. FidaPLUS, Gigafida), the comparison can be made between the student and adult language production, while the PiKUST L2 learner corpus (Stritar, 2009) can be used for the comparison between the language production of native learners and non-native speakers of Slovene.

Future plans include enlarging the corpus by transcribing the remaining texts, which were initially excluded from the corpus. Furthermore, Šolar is expected to act as a model in the development of similar L1 or L2 learner corpora of Slovene, and its method of utilizing teacher corrections for initial error analysis could be followed by learner corpora of other languages.

6. References

Arhar Holdt, Š., Kosem, I., Krapš Vodopivec, I., Ledinek, N., Može, S., Stritar, M., Svenšek, T., Zwitter Vitez, A. 2011. *Pedagoška slovnica pri projektu Sporazumevanje v slovenskem jeziku: K16 – Standard za korpusno analizo slovničnih pojavov.* Ljubljana.

James, C. 1998. *Errors in Language Learning and Use: Exploring Error Analysis*. London, New York: Longman.

Granger, S. 2004. Computer learner corpus research: current status and future prospects. V Connor, Ulla M. in Thomas Upton (eds.): *Applied Corpus Linguistics: A Multidimensional Perspective (Language and Computers)*. Amsterdam/New York: Rodopi, 123-145.

Granger, S., Dagneaux, E., Meunier, F., Paquot, M. 2009. *The International Corpus of Learner English. Version 2. Handbook and CD-ROM*, Louvain-la-Neuve: Presses Universitaires de Louvain.

Sinclair, J. (ed.) (1995) Collins COBUILD English Language Dictionary, second edition. Glasgow: HarperCollins.

Stritar, M. 2009. Slovene as a foreign language: The pilot learner corpus perspective. *Slovenski jezik – Slovene linguistic studies* 7. 135-152.