Chapter 7
Corpus-Based Resources for L1 Teaching: The Case of Slovene

Špela Arhar Holdt, Iztok Kosem, and Polona Gantar

Abstract The chapter highlights the potential of corpus-based resources for language education in K-12, more specifically for L1 teaching in the higher grades of elementary school and in the secondary school. Presented are two freely available online resources that were recently developed for teaching and learning Slovene as L1. Firstly, the Šolar corpus (www.korpus-solar.net), containing approximately one million words, comprises of texts written by Slovene elementary and secondary school students. More than half of the corpus texts include teacher corrections of language errors; furthermore, the errors have been manually categorised according to the classification scheme developed specifically for the project. The primary purpose of the corpus is to enable empirical research into communication competence of Slovene students and, based on that research, improve the methods and materials for Slovene language teaching. Secondly, the corpus-based Pedagogical Grammar Portal (http://slovnica.slovenscina.si) is an online language resource offering interactive explanations of language problems most commonly experienced by Slovene students when writing. The portal is aimed at students aged between 12 and 18 years. The content of the portal is based on the analysis of three corpora: the Šolar corpus, the reference Gigafida corpus, and the GOS corpus of spoken Slovene. The chapter provides a description of the Šolar corpus and the Pedagogical Grammar Portal, focusing on the applicative value of the results. Furthermore, the acquired know-how regarding the design and the implementation of such resources is presented, e.g. by highlighting the biggest challenges of the projects, and the pros and cons of the applied solutions. Finally, the chapter offers a wider discussion on the usefulness of the results for L1 teaching in K-12 education.

Keywords Language corpora • E-learning • Student writing • Language problems • Corpus Šolar • Pedagogical grammar portal • First language (L1) teaching • Elementary school • Secondary school • Digital tools • Language resources

Š. Arhar Holdt (✉) • I. Kosem
Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia

Institute for Applied Slovene Studies, Trojina, Slovenia
e-mail: Spela.ArharHoldt@ff.uni-lj.si; iztok.kosem@trojina.si

P. Gantar
Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia
e-mail: apolonija.gantar@guest.arnes.si

A. Marcus-Quinn, T. Hourigan (eds.), *Handbook on Digital Learning for K-12 Schools*, DOI 10.1007/978-3-319-33808-8_7

Introduction

Language corpora[1] have had a considerable impact on language education for nearly three decades, both indirectly and directly (cf. Aijmer, 2009; Aston, Bernardini, & Stewart, 2004; Campoy, Gea-Valor, & Belles-Fortuno, 2010; Hunston, 2002; O'Keeffe, Mccarthy, & Carter, 2007; Römer, 2005, and Scott & Tribble, 2006). Indirectly, corpora influenced language learning through corpus-based or corpus-driven dictionaries (starting with the COBUILD dictionary in 1987), grammars (e.g. Biber, Johansson, Leech, Conrad, & Finegan, 1999; Hunston & Francis, 2000), syllabi (McCarthy, McCarten, & Sandiford, 2005–2006; Willis, Willis, & Davids, 1988–1989), various language teaching and learning materials, and were even used in language testing (e.g. Ball & Wilson, 2002; Coniam, 1997). Direct use of corpora, as in data-driven learning (DDL), also has a long tradition, going back to Johns (1991).

Another strand of corpus influence on language teaching has been via the creation and analysis of learner corpora, which according to Osborne (2002) provide a bottom-up approach to language teaching. Learner corpus research bibliography is considerable, evidenced by the bibliography of the Learner Corpus Association[2] which contains over 1100 references.[3] In addition, the learner corpus community has recently founded an association and started the conference series. Furthermore, the results of learner corpus analyses have been used in dictionaries like Macmillan English Dictionary for Advanced Learners. However, the use of corpora in L1 language teaching and learning is considerably smaller, and developmental corpora—as we signify L1 equivalents of learner corpora, following terminology in (Leech, 1997, p. 19)—remain rare.

In Slovenia, corpora have only recently been introduced to language education, after a period of establishing their value in the Slovenian lexicography and lexical studies, where they were first introduced to the Slovenian linguistic community. However, in contrast with the international experience, particularly in relation to English as a Second Language, the first introduction of corpora to language pedagogy was made in L1 teaching and learning.[4] Among the reasons that induced interest of the field for the corpus approach were: literacy scores of young native speakers who achieved below average results in the PISA evaluations; findings that identified existing approaches to language teaching as systemic and structure-oriented (Rozman, Krapš Vodopivec, Stritar, & Kosem, 2012); lack of teaching materials

[1] "A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research." (Sinclair, 2005, p. 16).

[2] http://www.learnercorpusassociation.org/.

[3] Retrieved from https://www.uclouvain.be/en-cecl-lcbiblio.html.

[4] As for the L2 education, particularly noteworthy is PiKust, the pilot corpus of Slovene as a foreign/second language (Stritar, 2009). Currently, the applicative value of this corpus is somewhat limited due to its small size and unavailability to the general public. On the other hand, there have not yet been any larger-scale attempts to create corpora of Slovene native speakers communicating in foreign languages.

based on empirical evidence; and limited use of ICT in L1 language teaching compared to the use of ICT for teaching other school subjects.

In this chapter, we first present the Šolar corpus, a developmental corpus of Slovene writing. The compilation of the corpus is presented, and the methodological challenges and lessons learned are pointed out. As part of the project we describe, the Šolar corpus was made available to the general public via a customised concordancer, providing teachers and students with the opportunity to examine authentic written production in a qualitative, as well as quantitative way. Secondly, the data from the Šolar corpus was analysed and organised by the project team, and a set of ready-to-use corpus-based teaching materials were prepared. The materials were made available in a form of interactive, user-oriented resource, called the Pedagogical Grammar Portal. In this chapter, we present the features of the portal and the innovations the corpus-based materials bring to the Slovene language teaching. Following that is a discussion on the value of the portal for language pedagogy and a report on the initial feedback from the teachers. We conclude by outlining the plans for the future, and discussing the implications of the project for the use of corpora in L1 teaching and learning in Slovenia as well as in other countries.

Corpus Šolar

Idea and Implementation

The Šolar corpus was created as a part of the "Communication in Slovene" project, a national endeavour aimed at establishing language resources for the Slovenian language (different types of corpora and corpus annotation tools, language databases etc.).[5] The main purpose of the Šolar corpus was to enable empirical research into communication competence of Slovene students and, based on that research, improve the methods and materials for Slovene language teaching. To compile the corpus, a large quantity of authentic 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. The collection was conducted in close cooperation with the teachers, who have provided photocopies of the students' texts. A significant quantity of the received photocopies also included feedback that the teachers had provided to the students, namely corrections of students' language errors (spelling, morphology, syntax, vocabulary, punctuation, etc.). When transcribing the texts into digital form, these corrections were converted into annotations, increasing the value of the corpus data for research and educational purposes.[6]

[5] The project (2008–2013) was financed by European Social Fund and the Slovene Ministry of Education, Science and Sports. Information about the project is available at http://eng.slovenscina.eu/.

[6] Corpora containing texts of young L1 speakers are very rare, especially the ones with annotated errors; from this perspective, the Šolar corpus represents a potential for innovative research not only in L1 education but also in various fields of theoretical and applied linguistics, natural language processing, and language technology development.

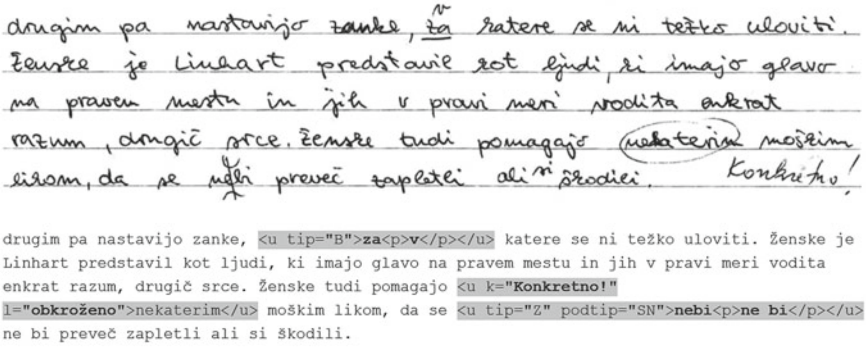
The design principles and the corpus building process were already presented to the international public in (Kosem, Rozman, & Stritar Kučuk, 2011) and in more detail to the Slovene public (Kosem et al., 2012). In this chapter, we therefore provide only a basic overview of the corpus content, followed by a presentation of the part of the corpus that includes teacher corrections, the corpus interface and the methodological challenges of the project.

The Šolar corpus comprises of texts written by Slovene elementary and secondary school students (aged between 12 and 18 years).[7] In its current state, the corpus consists of 2703 texts, more than half of which include teacher corrections of language errors. The total number of words in the corpus is 939,243 (excluding the teacher corrections). The texts, many of them graded, were produced as part of the coursework, mainly at Slovene (82.3 % of the texts). Much fewer texts (between 0.1 and 4.6 % of the texts) were obtained from other school subjects such as psychology, sociology, history, and geography. Majority of the texts are essays (79 %), the rest are tests (14 %) and other written school products such as letters, memos etc. (7 %). The texts were produced by students at high schools (43 %), students at technical schools (31 %), pupils at elementary schools (14 %) and students at vocational schools (5 %). Due to rich dialectal variation in Slovenia and the potential influence of specific dialectal features on the production in standard Slovene, it is also important to have a regionally balanced corpus. The Šolar corpus is only partly successful in achieving this aim: the texts are from all Slovenian regions; however, some regions have a very low share, especially Gorenjska (northwestern part of Slovenia) and certain smaller regions (e.g. Postojna region). However, the ratio between texts coming from South-West regions and North-East regions is approximately 3:2, which somewhat reflects the size of the area and the population.

The included texts were mainly produced in the school year 2009/2010. As already mentioned, the compilation of the corpus was conducted in cooperation with teachers of the selected schools who helped obtain permission from students and parents (for students and pupils under 18) for making the texts freely available, prepared photocopies of the material, and provided the necessary metatextual information. The second step of the corpus creation was the transcription of the written material to the digital form. In this process, the transcribers used XML tags to annotate language errors and teacher corrections (see Fig. 7.1). The transcription process proved to be more time-consuming than expected, and hence not all the collected texts could be included in the corpus.[8]

[7] 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, and secondly, including more sensitive data would make obtaining permission from the parents more difficult.

[8] Out of 8594 texts collected only 2703 were included in the corpus. It is also noteworthy that around 14 % of the gathered material was not suitable to be included due to lack of metalinguistic information, low quality photocopies and similar problems.



drugim pa nastavijo zanke, za katere se ni težko uloviti.
Ženske je Linhart predstavil kot ljudi, ki imajo glavo
na pravem mestu in jih v pravi meri vodita enkrat
razum, drugič srce. Ženske tudi pomagajo nekaterim moškim
likom, da se nebi preveč zapletli ali si škodili. Konkretno!

```
drugim pa nastavijo zanke, <u tip="B">za<p>v</p></u> katere se ni težko uloviti. Ženske je
Linhart predstavil kot ljudi, ki imajo glavo na pravem mestu in jih v pravi meri vodita
enkrat razum, drugič srce. Ženske tudi pomagajo <u k="Konkretno!"
l="obkroženo">nekaterim</u> moškim likom, da se <u tip="Z" podtip="SN">nebi<p>ne bi</p></u>
ne bi preveč zapletli ali si škodili.
```

Fig. 7.1 Student's text with teacher's corrections, and the annotation in the Šolar corpus

Language Errors and Corrections in the Corpus

To clarify the process of the annotation of language corrections, we present an example in Fig. 7.1. Above is an excerpt from an essay written by a student and corrected by a teacher. Below is the same text, as prepared for the Šolar corpus: the transcriber retyped the text and included the corrections in a form of XML tags.

As can be seen from the XML tags (attributes "tip" and "podtip", i.e. Category and Subcategory) in Fig. 7.1, the transcribers also assigned linguistic categories to errors and corrections. The annotation scheme was based on the classification designed for error tagging in a pilot corpus of Slovene as L2 (Stritar, 2009), with some minor adaptations (e.g. the category "abbreviation" was added for the purposes of the Šolar corpus). The categories of errors are presented in Table 7.1.

Annotated language errors and corrections can be found in 56 % of texts in the Šolar corpus. Nearly all of the corrected texts were produced at Slovene, which is understandable, as this subject is the most oriented towards the development of pupils' and students' writing skills. The predominant genres are argumentative and narrative essays. The text origin according to the type of school is shown in Table 7.2.

The teacher corrections in the texts are of different types, from underlined text and crossed out text to comments and suggestions for improvement of style. All the original corrections have been included in the digital form, and no additional corrections have been applied. The corpus data thus reflects not only the most typical language errors of students, but also highlights the correcting practices of the teachers.[9] Basing the error annotation and analysis solely on teacher corrections is an important methodological decision that needs to be taken into account when

[9] While the Šolar corpus is mainly intended for the investigation of the language production of students, the corpus can also be used as a consultation tool for language teachers. In teacher training, the corrections can be examined to better understand and to some extent standardise the strategies of interventions in student texts.

Table 7.1 Language errors in corpus Šolar

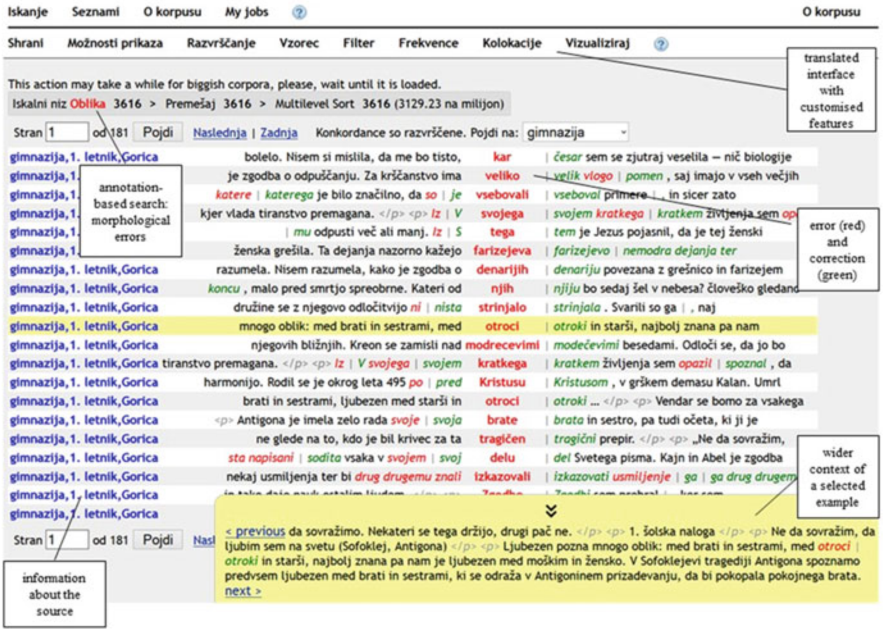
Error category	Error subcategory	Number of errors	Example from the Šolar corpus
Orthography	Spelling	2672	V času *razsvetljenjstva* ǀ *razsvetljenstva* so se ljudje začeli zavedati razuma.
	Together/Apart	1179	Zato se Simon *nemore* ǀ *ne more* osvoboditi ustanove.
	Capitalisation	2125	Znano je, da se ga je *Baron* ǀ *baron* Naletel skušal znebiti, saj je osvajal Nežko.
	Punctuation	15,371	To nam pove tudi verz" ǀ „ Ne maram ga, kdor le z besedo ljubi!"
	Abbreviation	23	V šoli sem 5 *h* ǀ *ur*.
	Numeral	50	Ko mi je bilo *12* ǀ *dvanajst* let, smo se s starši in prijatelji odpravili v Gardaland.
Vocabulary		3807	*Prizna* ǀ *Spozna*, da je Matiček njegov sin in Smrekarica je tako Matičkova mati.
Morphology		3618	*Vojno* ǀ *Vojne* ne jemlje tako resno kot Gregor.
Syntax	Word order	1265	Šele zdaj vidim, da je *za tisti čas bila* ǀ *bila za tisti čas* zelo napredna.
	Missing text	1607	Po navadi ǀ *tistega časa* naj bi bile žene poslušne možem.
	Redundant text	2665	Cankar je v drami prav tako razgalil politično prilagodljivost *tudi* ǀ učiteljev.
	Erroneous structure	653	*Tu je pomembno to* ǀ *Pomembno je*, kako so kmetje vztrajali pri maternem jeziku.

Table 7.2 Texts with language corrections regarding the type of school

Type of school	Texts	Percentage of texts	Words	Percentage of words
Elementary school (6th–9th grade)	395	26.1	110,828	19.1
High school	404	26.6	239,146	41.1
Technical school	574	37.9	186,784	32.1
Vocational school	143	9.4	44,719	7.7
Total	1516	100	581,477	100

interpreting the results of corpus analysis. Namely, when correcting texts, teachers consider student competence and other contextual specifics of the text. This means that the treatment of language errors is not completely comparable and consistent from student to student.[10] Nevertheless (or consequentially), the corpus is representative of school writing, and the valuable insight into the process of language correcting in schools extends the corpus value beyond merely statistics on language errors in student writing.

[10] For the time being, teacher corrections have not yet been thoroughly analysed; however, some preliminary findings reveal a certain level of inconsistency between the practices of different teachers, also in relation to the existing language norm.


Iskanje
Seznami
O korpusu
My jobs
?
O korpusu
Shrani
Možnosti prikaza
Razvrščanje
Vzorec
Filter
Frekvence
Kolokacije
Vizualiziraj
?
translated interface with customised features
annotation-based search: morphological errors
error (red) and correction (green)
wider context of a selected example
information about the source

This action may take a while for biggish corpora, please, wait until it is loaded.

Iskalni niz Oblika 3616 > Premešaj 3616 > Multilevel Sort 3616 (3129.23 na milijon)

Stran 1 od 181 Pojdi Naslednja | Zadnja Konkordance so razvrščene. Pojdi na: gimnazija

```
gimnazija,1. letnik,Gorica                       bolelo. Nisem si mislila, da me bo tisto,     kar          | česar sem se zjutraj veselila – nič biologije
gimnazija,1.                                   je zgodba o odpuščanju. Za krščanstvo ima     veliko       | velik vlogo | pomen , saj imajo v vseh večjih
gimnazija,1.                       katere | katerega je bilo značilno, da so | je    vsebovali    | vseboval primere | , in sicer zato
gimnazija,1.                   kjer vlada tiranstvo premagana. </p> <p> Iz | V     svojega      | svojem kratkega | kratkem življenja sem op
gimnazija,1.                                    | mu odpusti več ali manj. Iz | S     tega         | tem je Jezus pojasnil, da je tej ženski
gimnazija,1.                                ženska grešila. Ta dejanja nazorno kažejo     farizejeva   | farizejevo | nemodra dejanja ter
gimnazija,1. letnik,Gorica                  razumela. Nisem razumela, kako je zgodba o     denarijih    | denariju povezana z grešnico in farizejem
gimnazija,1. letnik,Gorica              koncu , malo pred smrtjo spreobrne. Kateri od     njih         | njiju bo sedaj šel v nebesa? človeško gledan
gimnazija,1. letnik,Gorica                družine se z njegovo odločitvijo ni | nista     strinjalo    | strinjala . Svarili so ga | , naj
gimnazija,1. letnik,Gorica                      mnogo oblik: med brati in sestrami, med     otroci       | otroki in starši, najbolj znana pa nam
gimnazija,1. letnik,Gorica                   njegovih bližnjih. Kreon se zamisli nad     modrecevimi  | modečevimi besedami. Odloči se, da jo bo
gimnazija,1. letnik,Gorica tiranstvo premagana. </p> <p> Iz | V svojega | svojem     kratkega     | kratkem življenja sem opazil | spoznal , da
gimnazija,1. letnik,Gorica           harmonijo. Rodil se je okrog leta 495 po | pred     Kristusu     | Kristusom , v grškem demasu Kalan. Umrl
gimnazija,1. letnik,Gorica                   brati in sestrami, ljubezen med starši in     otroci       | otroki ... </p> <p> Vendar se bomo za vsakega
gimnazija,1. letnik,Gorica          <p> Antigona je imela zelo rada svoje | svoja     brate        | brata in sestro, pa tudi očeta, ki ji je
gimnazija,1. letnik,Gorica                ne glede na to, kdo je bil krivec za ta     tragičen     | tragični prepir. </p> <p> „Ne da sovražim,
gimnazija,1. letnik,Gorica          sta napisani | sodita vsaka v svojem | svoj     delu         | del Svetega pisma. Kajn in Abel je zgodba
gimnazija,1. letnik,Gorica           nekaj usmiljenja ter bi drug drugemu znali     izkazovali   | izkazovati usmiljenje | ga | ga drug drugem
gimnazija,1. letnik,Gorica
gimnazija,1. letnik,Gorica
```

Stran 1 od 181 Pojdi Nasl

< previous da sovražimo. Nekateri se tega držijo, drugi pač ne. </p> <p> 1. šolska naloga </p> <p> Ne da sovražim, da ljubim sem na svetu (Sofoklej, Antigona) </p> <p> Ljubezen pozna mnogo oblik: med brati in sestrami, med otroci | otroki in starši, najbolj znana pa nam je ljubezen med moškim in žensko. V Sofoklejevi tragediji Antigona spoznamo predvsem ljubezen med brati in sestrami, ki se odraža v Antigoninem prizadevanju, da bi pokopala pokojnega brata. next >

Fig. 7.2 Morphological errors in the Šolar corpus

Corpus Concordancer

As the corpus contained annotated errors and corrections, available concordancing tools were not suitable for its use. Therefore, we developed a new interface, based on the widely used Sketch Engine corpus tool (Kilgarriff, Rychly, Smrz, & Tugwell, 2004), which was localised and customised for the purposes of displaying data from error-annotated corpora. A similar Sketch Engine-based concordancer is already used by Cambridge University Press for their learner corpora, but their concordancer is aimed at researchers who are more advanced users of corpus tools. Contrarily, our target users were primarily of wider audience (e.g. teachers), so a great deal of attention was paid to the means of clear and simple data presentation. The localisation included translating the interface into Slovene, developing help and tips on how to search the corpus, and simplifying the interface language (eliminating abbreviations and terminology). The customisation addressed mainly the development of innovative ways of presenting corpus data, especially demonstrating language errors and corrections in a user-friendly manner, and developing functionality for searching and manipulating corpus data (Fig. 7.2).

In addition to regular functions of the Sketch Engine tool, the Šolar concordancer also allows searching by language errors and corrections. The users can look up specific error (e.g. the occurrences of errors containing the personal pronoun *moj*); error-correction combination (e.g. all the errors where the pronoun *moj* was replaced with reflexive pronoun *svoj*); or all the errors of a particular error category (e.g. all the instances that were annotated as morphological errors). The results are shown in KWIC format, as in most concordancers, with the searched error or correction shown in the centre. For easier interpretation and overview, the language errors are shown in red and the corrections in green. The default view does not show tags (which can be activated if needed), as is the case with the Cambridge University Press interface, since testing has shown that tags make the concordances difficult to read. The concordances can be manipulated, e.g. the users can sort, sample or filter them, as well as save them for further use. In addition, the concordancer offers some possibilities of data summarisation, such as a collocation list, frequency distribution of errors by error type and metatextual information (type of school, region, year/grade etc.).

Methodological Challenges and Lessons Learned

One of the main challenges of the compilation of the Šolar corpus was the transcription of the student texts. In fact, initially, more problems were expected with obtaining the student texts, nevertheless the response from the Slovene teachers at schools was overwhelming. On the other hand, the transcription was problematic not only because the texts were handwritten (and as such sometimes difficult to decode), but also because the received photocopies were in black and white (not in colour), which often made it difficult to distinguish teacher corrections from student text.[11]

However, by far the most demanding and time-consuming part of the transcription proved to be annotation and categorisation of student 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 error using the attribute in the tag. 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 categorisation of errors) with technical ones (using XML tags and macros), which resulted in the mistakes on the linguistic side, e.g. incorrect category for the error used, and 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 annotated errors, 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 a fact that fewer texts were included in the corpus.

[11] While teachers might have been able to help us solve some of these problems, that would have been time-consuming both for them and the project, and it would considerably prolong the transcription process.

Pedagogical Grammar Portal

Idea and Implementation

Since it is freely available online, the Šolar corpus facilitates insight into authentic language practices of Slovene students to the interested general public. Teachers might, for example, use the corpus data to improve their teaching materials (e.g. find examples, create exercises and tests) or to prioritise the topics of their teaching to better address the language problems, reflected in the corpus. However, such use of corpora has its limitations, the most important one being the time one has to invest into the preparation of such materials. Thus, one aim of the "Communication in Slovene" project was to develop a set of ready-to-use corpus-based teaching materials and make them available to the teachers and students in a form of an interactive online resource. The result of this project activity was the Pedagogical Grammar Portal, a freely available multimodal resource, consisting of several units (chapters) that focus on the most typical language problems students encounter while writing in standard Slovene (see Table 7.3).

As mentioned previously, teacher corrections in the Šolar corpus were first categorised into 12 robust linguistic categories (Table 7.1). For the identification of the most typical student problems, a more fine-grained categorisation was needed. Our decision was to conduct a manual (sub)categorisation of the corrections using a bottom-up categorisation approach; however, the annotation procedure had to be improved. The overview of existing annotation tools showed that there are very few tools for error annotation available, especially tools that would have allowed us to import error-annotated corpus and categorise existing errors further. In the end, WordSmith Tools (Scott, 2008) proved the tool most suitable for our purposes, even though its main shortcoming was that it only allowed categorisation within the tool; making changes directly into the corpus file(s) was not possible.

As a result of this process, 692 different categories of language problems were identified (Kosem et al., 2012): by far the most frequent problem was the use of a comma, also often causing problems was the declension of certain nouns, spelling of certain words (e.g. *življenje*), use of possessive and reflexive-possessive pronouns, use of infinitive and supine, use of modal verbs *moči* and *morati*, use of comparative and superlative forms of adjectives etc. In the second step, priority lists with problems of different type/frequency/regional distribution were created (Arhar et al., 2011, pp. 50–57), and used as a basis for the conceptualisation of the portal, the development of its wireframe and design, and the customisation of the selected Content Management System (CMS). In the final stage of the project, 24 chapters[12] of interactive corpus-based teaching material were prepared by a team of linguists

[12] The Pedagogical Grammar Portal consists of *chapters*, i.e. self-standing ready-made teaching units that revolve around a specific topic (e.g. the use of supine in Slovene). The structure of the chapters on the portal is presented in more detail in section "Structure of the PGP Chapters" (see Figs. 7.3 and 7.4).

Table 7.3 Language problems presented on the Pedagogical Grammar portal

	Topic	Example from the Šolar corpus
1	Spelling of negated verbs	Tak odnos nima smisla in se *nemore* | *ne more* srečno končati.
2	Declension of the irregular noun *otrok*	Dandanes je vez med starši in *otroci* | *otroki* veliko večja.
3	Use of prepositional variants *s/z*	Na koncu je odšla *s* | *z* botrom Esadom v Nemčijo [.]
4	Use of prepositional variants *k/h*	Kasneje sta se Raymond in Mersault vrnila *h* | *k* Arabcem [.]
5	Reduction of the infinitive	[N]aučila jih je, da je v življenju treba *delat* | *delati* in se *borit* | *boriti*.
6	Use of supine	Pobriše mizo in se odpravi *pomivati* | *pomivat* posodo.
7	Umlaut in noun declensions	Oba se ranita z zastupljenim *mečom* | *mečem*.
8	Use of verbs with –te/-ta and –ste/-sta	Sklenila sta |, da se *bota* | *bosta* poročila.
9	Use of modal verbs *morati* and *moči*	Kajn je *mogel* | *moral* bloditi po svetu.
10	Spelling of past active participles with *-l*	*Delav* | *Delal* se je norca iz njega [.]
11	Spelling of preposition *v*	Janez je *u* | *v* mestu zelo veliko razmisljal o njej |, kako je lepa.
12	Negation of adjectives	[V]anj se *ne normalno* | *nenormalno* zaljubi [.]
13	Words with *izs-*	Opazovalec prispe na cilj in *iztopi* | *izstopi* iz avtobusa [.]
14	Words with *-lj-*	Povezanost pomeni sožitje, medsebojno spoštovanje in *prijatelstvo* | *prijateljstvo* narodov.
15	Words with *-nj-*	Ta mu pove, da ga vidi *zadnič* | *zadnjič*, ker se je dala krstiti.
16	Spelling of words with double letters	Vendar župnik na koncu Jermanu *vseno* | *vseeno* prizna, da je ravnal prav.
17	Use of *nobeden* and *noben*	Pripovedovalec je ugotovil tudi, da se *nobeden* | *noben* drug potnik ni zmenil za to gospo [.]
18	Use of pronouns *nobeden* and *nihče*	Bog pa mu je dal znak, da ga *noben* | *nihče* ne bi ubil.

Table 7.4 Online language resources, presented on the Pedagogical Grammar Portal

Topic			
1	Corpus Gigafida	4	SSKJ dictionary of standard Slovene
2	Corpus GOS	5	Slovene orthography 2001
3	Sloleks morphological lexicon	6	Orthography Guide

and teachers: 18 chapters focusing on specific language problems (Table 7.3), and 6 chapters describing available online language resources for Slovene (Table 7.4).

Each presentation of a language resource starts with a description of how the resource was created, then offers 3–5 video tutorials on how to use the resource to solve different types of language questions, points out any limitations of the resource, and finally challenges the user to answer a set of language questions using the presented resource. These chapters were added to the portal so they can be referred to when describing solutions for specific language problems.

The preparation of the chapters on language problems, on the other hand, posed a considerable challenge because no large-scale corpus-based description of Slovene grammar was available at the time. Therefore, the first step of preparing each chapter included a detailed analysis of corpus data for the selected problem. Three corpora were used: the Šolar corpus, a corpus of spoken Slovene GOS,[13] and a reference corpus of written Slovene Gigafida.[14] As the first two corpora are smaller and content specific, the analyses were conducted in their corresponding concordancers. Gigafida, however, facilitated a more synthetic, quantitative approach with automated data extraction and organisation. Some of the findings of these studies were presented to the (Slovene) linguistic audience, together with the methodology used (Arhar Holdt & Stritar Kučuk, 2012; Može, 2013).

The results of corpus analyses brought another challenge: the newly acquired insight into the chosen language phenomena differed significantly from the existing language description, and in some instances in opposition with the current linguistic norm. This gap was to some extent expected, since it is widely recognised that "corpora have provided evidence for our intuitions about language and very often they have shown that these can be faulty when it comes to issues such as semantics and grammar (O'Keeffe et al., 2007, p. 21)". The extent to which traditional dichotomies (written vs. spoken, formal vs. informal, correct vs. erroneous, lexis vs. grammar) were challenged by the new data was nevertheless surprising. Consequentially, an important task in the design of the PGP was to present these new findings without creating unnecessary conflicts with the existing reference books and teaching materials. A middle way, determined in cooperation with experts from the fields of linguistics and language teaching, was demonstrating actual language use with a democratic view of different language choices, while at the same time highlighting the specifics of the current (codified) standard the students are expected to master in the educational process. Our decision confirmed anticipations that corpus-based studies will impact language teaching by replacing monolithic grammar descriptions with register-specific ones, integrating teaching of grammar with teaching of vocabulary, and shifting emphasis from "accurate" to "appropriate" (Conrad, 2000, p. 549). Larger-scale evaluations of the Pedagogical Grammar Portal, planned for future work, will aid in establishing the success of this integration.

[13] GOS is the first corpus of spoken Slovene. It consists of approximately 120 h of recorded speech in various situations. The content, structure, and availability of the corpus, as well as the specifics of its custom-designed interface are described in (Verdonik, Kosem, Zwitter Vitez, Krek, & Stabej, 2013).

[14] Gigafida is the most recent, and with nearly 1.2 billion words also the biggest, of Slovene written corpora. Information about the corpus is available in (Logar & Krek, 2012).

Use of Corpus Data on the Portal

Corpora are not (yet) systematically incorporated in the Slovenian education process and are also not included in existing curricula. Consequently, teacher training on the use of corpora in the classroom is not provided.[15] Hence our decision was to design PGP as a resource that uses corpora indirectly (according to typology in Leech, 1997), meaning that corpus material is not presented to the user directly via a concordancer, but rather preselected and organised for particular educational purposes. Another reason for such use of corpus data was the already mentioned desire to produce ready-made materials that facilitate instant integration into the teaching progress and are easy for students to access and use both in classroom and at home.

On the PGP, corpus material is usually provided in a form of short sentences that exemplify specific features of language use. The only editing procedure in the preparation of corpus examples was the (optional) shortening of sentences, as we believed long and complex examples would shift the focus of the users away from the actual purpose of the chapter.[16] The second type of corpora use was to prepare supportive visual material for the chapters, such as frequency-based word clouds, word lists, and charts.

On the portal, different corpora are used for different purposes. The Gigafida corpus, as a reference corpus of written Slovene, is used to explain the discussed language phenomenon in general (e.g. how supine is used in written Slovene), and to visualise supportive language data (e.g. a word cloud with most common Slovene verbs to help the user identify and understand the part-of-speech category). The Šolar corpus, on the other hand, is used to exemplify the specifics of the discussed language problem, e.g. to represent what errors/corrections typically occur when students use supine in their writing. Thirdly, as the aim of the PGP is to improve the written production of the students, examples from the GOS corpus are included to highlight the differences between written and spoken Slovene. GOS is also used for demonstrating specific dialectal features in comparison to the standard language.

Innovation in Slovene Language Didactics

Using the corpus approach when developing the PGP introduced several novelties to Slovene language didactics. We discuss some of these decisions in section "The Value of the Pedagogical Grammar Portal for the Slovene Language Education":

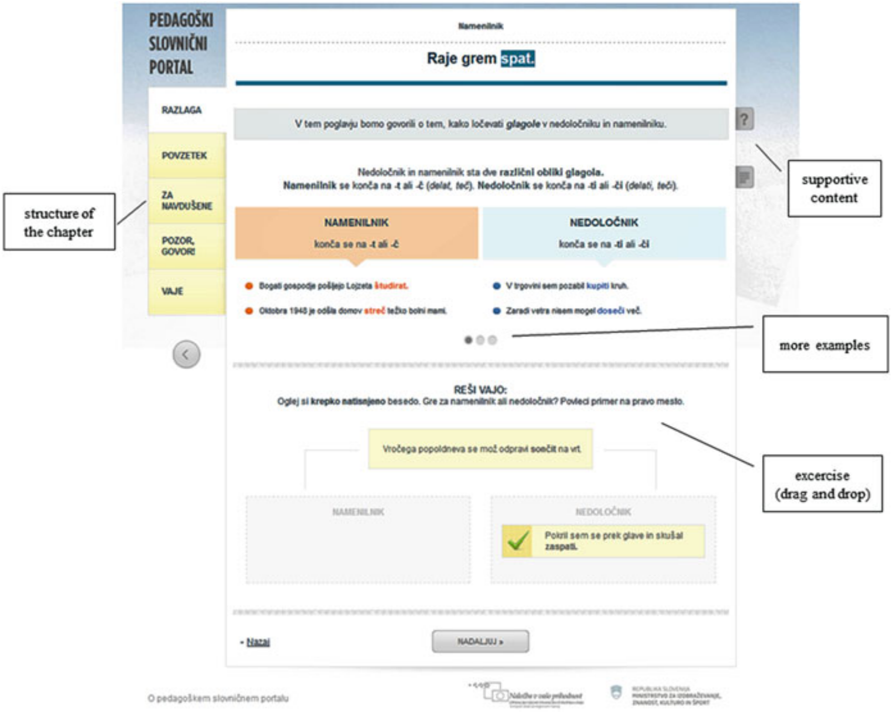
[15] Despite this fact, more and more teachers are becoming aware of the existence of corpora and their educational potential. In section "Preliminary Feedback from Teachers" we summarise the experience from the workshops for teachers on the use of language resources and technologies in language teaching.

[16] A debate in corpus linguistics on (non)authenticity of decontextualised corpus data is summarised in (O'Keeffe et al., 2007, pp. 25–27).

- prioritisation of the teaching content according to the frequency of language errors in student writing;
- conceptualisation of explanations from specific language problems rather than from the grammar system;
- use of authentic corpus examples to support explanations;
- representation of language use as it appears in various genres (written and spoken, standard and non-standard);
- inclusion of a high number of interactive corpus-based exercises;
- different treatment of language problems, common to all or nearly all Slovenian regions, and problems limited to individual region(s).

With these features as starting points, the online interface of the Pedagogical Grammar Portal was designed by a team of experts: linguists, language teachers, designers and programmers. One of the most important decisions in the preparation of the portal was connected to the organisation of the content. Unlike existing resources for teaching Slovene as L1, which are based on topics from the grammar system selected with top-down approach, the topics in the PGP were selected bottom-up from the student language production. Consequentially, we decided to present each language problem as an independent, self-standing chapter. At the beginning of each chapter, the user is equipped with the metalinguistic knowledge needed to understand the explanation of the language problem. From then on, the focus is on the problem: its characteristics; possible reasons for it; and the suggested solutions, i.e. the use of mnemonics to remember the grammar rules, the use of reference books or resources, or the use of techniques to improve writing strategies. The explanation of a language problem is divided into smaller units, and each of the units is supplemented with a short exercise to activate the users and promote continuous self-evaluation of the progress (Fig. 7.3). A high number of interactive corpus-based exercises at the end of each chapter facilitates automatisation of the use of language rules and consequently the transfer of the knowledge into practice (Fig. 7.4).

A great deal of attention has been paid to the language of the explanations. Unnecessary use of terminology is avoided, as is the use of complex syntactic structures. The language used is concrete and concise. Supportive information—definitions of terms, theoretical background, additional statistical information, tables with word forms, word lists etc.—is removed from the central explanation and made available to the user in the form of clickable side tags (Fig. 7.3). As the PGP is aimed at students of different ages (12–18), we made every effort to make the primary content as clear and straightforward as possible, hoping to ensure comprehensibility for the younger users while still providing the older users with additional useful information for better understanding of the problem in a wider context. Another reason for the simplification of the language was our aim to facilitate students' independent use of the PGP. Namely, the portal was conceived and developed to support individualised approach to improving students' writing skills: after the evaluation of the students' written production, the teacher can choose to assign each student only the chapters relevant for his/her language problems, instead of covering the same grammar topics with the entire class regardless of student-specific strengths and weaknesses.



PEDAGOŠKI SLOVNIČNI PORTAL

Namenilnik

Raje grem spat.

V tem poglavju bomo govorili o tem, kako ločevati glagole v nedoločniku in namenilniku.

Nedoločnik in namenilnik sta dve različni obliki glagola.
Namenilnik se konča na -t ali -č (delat, teč). Nedoločnik se konča na -ti ali -či (delati, teči).

NAMENILNIK konča se na -t ali -č	NEDOLOČNIK konča se na -ti ali -či
● Bogati gospodje pošljejo Lojzeta študirat.	● V trgovini sem pozabil kupiti kruh.
● Oktobra 1948 je odšla domov streč težko bolni mami.	● Zaradi vetra nisem mogel doseči več.

REŠI VAJO:
Oglej si krepko natisnjeno besedo. Gre za namenilnik ali nedoločnik? Povleci primer na pravo mesto.

Vročega popoldneva se mož odpravi sončit na vrt.

NAMENILNIK

NEDOLOČNIK

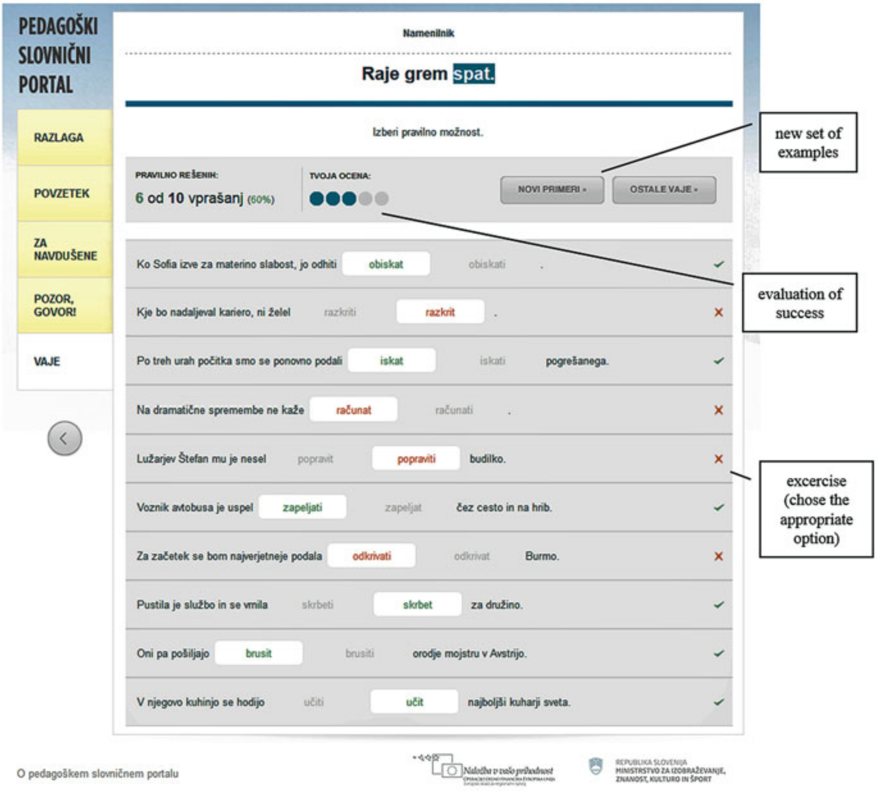
✓ Pokril sem se prek glave in skušal zaspati.

Fig. 7.3 Explanation of the use of supine (*namenilnik*)

Structure of the PGP Chapters

The chapters on the portal follow a standard structure, containing obligatory (1–5) and optional (6) elements:

1. Introduction, where the users can determine whether the topic of the chapter is relevant to them or not. To help them with deciding three examples of student errors (and teacher corrections) from the Šolar corpus are provided.
2. Explanation of the language problem with suggestions/techniques for its solution. The content is separated into 3–5 shorter units, each of them supplemented with short exercises and tasks to activate the users.
3. A one-page summary of the presented content.
4. A sub-chapter with additional information on the language problem (e.g. interesting facts from the history of the Slovenian language or corpus statistics) with links to external resources. This sub-chapter is aimed at motivating the users to utilise existing resources for Slovene for further investigation of the discussed topic.
5. A large number of exercises, presented to the users in chunks of 10. Exercises are automatically rated and the users can monitor their results over time.


PEDAGOŠKI SLOVNIČNI PORTAL
RAZLAGA
POVZETEK
ZA NAVDUŠENE
POZOR, GOVOR!
VAJE
NOVI PRIMERI »
OSTALE VAJE »
O pedagoškem slovničnem portalu
Naložba v vašo prihodnost
REPUBLIKA SLOVENIJA MINISTRSTVO ZA IZOBRAŽEVANJE, ZNANOST, KULTURO IN ŠPORT

Namenilnik

Raje grem spat.

Izberi pravilno možnost.

PRAVILNO REŠENIH: 6 od 10 vprašanj (60%)

TVOJA OCENA: ●●●○○

Ko Sofia izve za materino slabost, jo odhiti obiskat obiskati . ✓

Kje bo nadaljeval kariero, ni želel razkriti razkrit . ✗

Po treh urah počitka smo se ponovno podali iskat iskati pogrešanega. ✓

Na dramatične spremembe ne kaže računat računati . ✗

Lužarjev Štefan mu je nesel popravit popraviti budilko. ✗

Voznik avtobusa je uspel zapeljati zapeljat čez cesto in na hrib. ✓

Za začetek se bom najverjetneje podala odkrivati odkrivat Burmo. ✗

Pustila je službo in se vrnila skrbeti skrbet za družino. ✓

Oni pa pošiljajo brusit brusiti orodje mojstru v Avstrijo. ✓

V njegovo kuhinjo se hodijo učiti učit najboljši kuharji sveta. ✓

new set of examples

evaluation of success

excercise (chose the appropriate option)

Fig. 7.4 Exercises on the use of supine (*namenilnik*)

6. Additional warnings, typically about the exceptions to the presented language rules, regionally relevant aspects of the language problem, or noteworthy specifics of the relation between written and spoken Slovene.

Figures 7.3 and 7.4 present two pages from the chapter on the use of supine, the first page of the explanation of the problem and a set of (solved) grammatical exercises. There is not enough space to describe the structure of the portal in greater detail, but the readers are welcome to examine it on http://slovnica.slovenscina.eu/.

The Value of the Pedagogical Grammar Portal for the Slovene Language Education

When we think about the potential value that the PGP has for the Slovene language education, we need to consider it in the context of the existing situation in Slovene primary and secondary education where online language resources and ICT have only recently, and in small steps, been introduced to teaching Slovene as L1 (Rozman, Krapš Vodopivec, et al., 2012, pp. 91–100). The available online teaching resources are fairly basic in content and form (Arhar et al., 2011). What is more, teaching materials and methods are not corpus-based (see Aston, 2001 for arguments in favour of such practice); the same can be said for existing reference works such as dictionaries and orthographies. The main reasons for such a situation are uneven distribution and unbalanced prioritisation of content and the belief that ICT cannot play a significant role in language teaching as it does in other subjects such as natural sciences. The development of online language resources is language specific, i.e. materials cannot be taken from other languages and simply translated. Therefore, online language resources need to be developed from scratch, which requires time and money. In Slovenia, lack of resources and funding[17] has presented the biggest obstacle to development and implementation of ICT in L1 teaching in Slovenia. Thus, the PGP is faced with two main challenges. The first one is technical and has to do with taking advantages of the online medium such as multimedia, hyperlinks and customizability for presenting language content more effectively. The second challenge is about introducing corpus methodology and corpus-based materials into Slovene schools, which raises questions about whether the Slovene educational system is ready for such a pedagogical language resource[18] and whether the pedagogical purpose of the PGP and the corpus methodology used in its compilation are compatible.[19] The discussion on the linguistic potentials of the PGP for the Slovene education will focus mainly on the latter challenge.

Implementing Corpus Methods into L1 Teaching: Challenges, Advantages and Dangers

Introducing corpus methodology and corpora as language resources to L1 teaching in schools is one of the essential steps towards bridging the gap between real language use and student language problems, and language content in existing textbooks. The PGP achieves that by offering a large number of authentic corpus examples, both as part of the explanations of the problems and in the exercises. Because many examples are taken from a corpus of student writing, it is likely that their content will be familiar to the (student) users which will make the identification with the

[17] This has been exacerbated by the economic crisis, during which the funding allocated to education in Slovenia has been constantly decreasing.

[18] This would also mean providing regular training of teachers on how to work with corpora in the classroom (see section "Preliminary Feedback from Teachers").

[19] The developers of the PGP were unable to inform their decisions on how to implement corpus data into education process using online language resources (particularly in L1 teaching) using examples of good practice from abroad. Namely, the PGP is in many ways a unique pilot project where a new form of teaching and learning materials and strategies are being developed and tested.

language problem easier—this is one of the important advantages that PGP has over existing textbooks.[20]

This new approach to teaching Slovene is based on explaining linguistic phenomena with real language situations and shifts the focus from the accumulation of factual knowledge, found in existing textbooks, to language competence. Another advantage of the PGP lies in the fact that it attempts to utilise intertextuality and interactivity of the digital media.

In addition to more adequate representation of authentic Slovene, the corpus-based approach introduces new terminology (e.g. concordancer, tagger, language technologies) into language materials, as well as links to language resources and tools for Slovene—corpora, grammar checker, morphological lexicon etc.—which help raise awareness among students of the increasingly important role of information science in the field of linguistics. The absence of interdisciplinary links between linguistics and computer science in existing textbooks does not reflect the recent developments in linguistics and the reality of the digital age in which (soon to be) competent users of language live. In this aspect, the PGP presents a good model that should be used when revising and updating the Slovene syllabi and in planning online language resources.

A Shift in the Conceptualisation of Language Phenomena

The planned shift in the conceptualisation of language phenomena in comparison with the existing linguistic theory is evident in the presentations and explanations of the problems which are not tied to the structure-focused view of the language system, but are problem-driven. Moreover, the presentation of the solution to the problem links morphology and syntax, word-formation and semantics, grammar and norm, which highlights internal co-dependency of language phenomena. This is a novel approach in Slovene language teaching,[21] and benefits from the multidimensionality of the online medium.

The introduction of real language use into language teaching, a consequence of the PGP using corpus methodology, is accompanied by at least two more dilemmas directly connected with conceptualisation of language phenomena in existing textbooks and materials for Slovene. Firstly, examples of real language use inevitably show language variation, as opposed to language description in existing textbooks where only one variant of many is often presented as the correct one (Rozman, Krapš

[20] In the survey conducted by Rozman et al. (2012: 106–108) the students listed mind maps and charts as the preferred explanatory methods for language content, whereas examples and images received lower scores. This could be linked to the finding that the students find examples in existing textbooks uninteresting, unreal and atypical (ibid. 108).

[21] The survey conducted among teachers of Slovene as L1 (Rozman, Krapš Vodopivec, et al., 2012, p. 75) showed that most of them agree with structure-oriented grammar teaching (separated in subfields, such as phonology, morphology, lexicology, syntax), and believe that this leads to the development of students' communication skills.

Vodopivec, et al., 2012, p. 60). Secondly, language learners, as future competent language speakers are expected to learn to linguistically and pragmatically argument their language decisions. If the traditional approach is based mainly on knowing rules and norms of the language, the PGP builds the learner's/speaker's language competence by teaching functional variation which depends on the communication situation, or in other words, by teaching that language choices depend on the context (formal and informal), and regional, genre and stylistic characteristics. With other words, the main difference between both approaches is, that the PGP aims to raise the awareness of the speakers to make informed language choices, while the traditional approach focuses almost exclusively on the standard language.

To sum up, the value of the PGP for the Slovenian education lies in the combination of a synthetic, problem-oriented approach to language with attempts for user-friendliness, that introduce mnemonics to facilitate the acquisition of language rules and combine simple and easy-to-understand language descriptions with fun and interesting exercises; this combination results in showing the learners a specific language phenomenon in its variety, with more than one possible grammatical solution, yet exposes the solutions in accordance with the existing standard as the natural choice for the written production students are to master in the process of education. The question to what extent and in what manner the students, especially in elementary schools, are ready to learn about language variation, remains to be more thoroughly researched, and the creation of the PGP can facilitate an insight into this area. The same applies to teachers who need to be familiar with the characteristics of corpus approach in language teaching. In addition, they need to be able to adapt the teaching strategies to successfully teach their students how to choose the suitable variant in a given communication situation from the variety of data offered by corpora. In order to give teachers the necessary basis to achieve these goals, a series of workshops on language resources and technologies was organised. The feedback of the teachers on the workshops is presented in the next section.

Preliminary Feedback from Teachers

From 2012 to 2014 a series of workshops on language technologies for primary and secondary school teachers, mainly teachers of Slovene, have been held at various locations around Slovenia. The 8 h workshops, funded by the Ministry of Culture of the Republic of Slovenia, focussed on presenting existing freely available language resources for Slovene, such as corpora, language tools, dictionaries and other reference works.[22] The Šolar corpus and the PGP were included in the workshop material for the 2013 and 2014 workshops. The workshop presenters reported of very positive feedback from the teachers; few teachers were initially sceptical, but mainly due to their lack of experience in using electronic resources and computers in general. However, after getting hands-on experience with the resources, scepticism was often

[22] The project website: http://ucitelji.sdjt.si/.

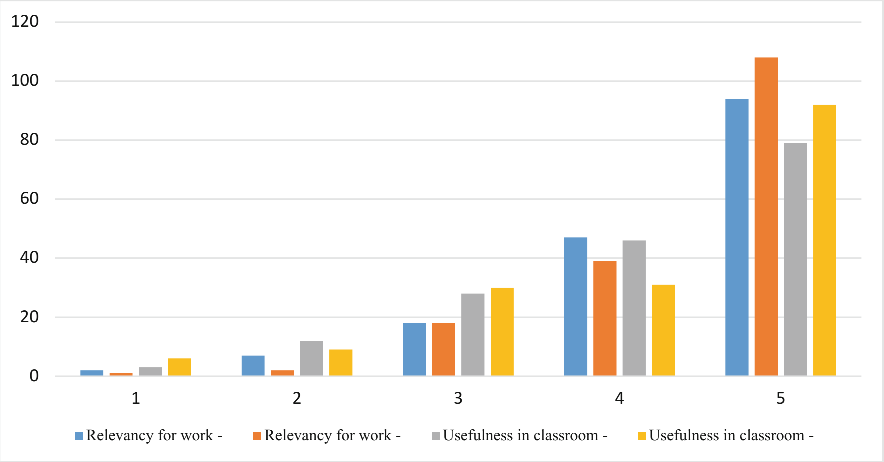


Fig. 7.5 Teacher evaluation of the Šolar corpus and the PGP

replaced with enthusiasm. The positive attitude of the teachers towards the Šolar corpus and the PGP was also confirmed by the results of the survey completed by the teachers after the workshops. The survey in 2013 was less detailed in so far that it did not ask teachers for their opinion on each of the resources presented; however, the teachers were given the option to name the resource that they thought was most useful for classroom work. The survey was completed by 208 primary school teachers, including 74 participants who teach Slovene in the last triad (12–14 years). Out of these, nearly a quarter (24 %) named the Šolar corpus as the most useful resource, while even more of them (36 %) thought the PGP to be the most useful.

A more detailed insight was provided by the survey in 2014, which was conducted among 168 teachers from different primary and secondary schools (Fig. 7.5). The participants had to evaluate the relevancy of the presented language resources for their professional (teaching) purposes. The grades ranged from 1 to 5, with 1 meaning that the resource is completely irrelevant and 5 meaning that the resource is highly relevant. The Šolar corpus was rated as relevant (with 4 or 5) by 84 % of the teachers, and the PGP by 88 %. Furthermore, the participants had to evaluate the usefulness of the resources for classroom use, again on a scale from 1 to 5. The Šolar corpus was evaluated as useful (with 4 or 5) by 74 % of the teachers, and the PGP by 73 %.[23] The results seem encouraging, considering the fact that the content on the PGP is still in its pilot version and thus rather limited in quantity. Nevertheless, this first feedback is very general, and hence a more detailed evaluation is planned.

[23] While interpreting these results, one has to keep in mind that teachers of different subjects were participating, as well as teachers of different grades. As a comparison: dictionaries of Slovene language SSKJ and SP 2001, which were among the highest rated resources, were recognised as relevant for work by 89 %, and useful in classroom by 84 % of the participants.

Conclusion and Future Work

In this chapter, we highlighted the potential of corpus-based language resources for language education in K-12, more specifically for L1 teaching. The basis of our study were two freely available online resources that were recently developed for teaching and learning Slovene as L1: the Šolar corpus and (utilising the corpus data) the Pedagogical Grammar Portal.

Despite the challenges presented in this chapter, the results seem promising. The Šolar corpus proved to be a highly suitable basis for the research of students' language competence and their language problems. The design of a specialised concordancer made the corpus data available not only to the research community, but also to the wider audience, such as teachers. The first feedback from the teachers has been very positive: not only did the teachers rate the corpus as useful for their work, they also recognised its value for the use in the classroom. Nevertheless, the initial enthusiasm was accompanied by remarks about the time-consuming use of the corpus for the preparation of teaching materials. These needs have been partially addressed by the creation of the Pedagogical Grammar Portal; however, further investigation into the attitude of the teachers towards the implementation of corpus data in the teaching process is needed. The second important task for the future is to increase the size of the Šolar corpus, ideally to around five million words, especially with the texts from the currently under-represented Slovenian regions. To accomplish this goal, the methodology would have to be improved as suggested in this chapter, especially to make the compilation process faster and more systematic.[24] Last but not least, we would like to supplement the corpus annotation with the fine-grained categories of language problems, as identified for the creation of the PGP. With the inclusion of these categories in the corpus XML file, the data will be available in the corpus concordancer, and consequentially generally accessible for the creation of materials for teaching Slovene.

The idea of the PGP was welcomed by the teachers as well, and its implementation into the teaching practice is at the time hindered primarily by the low number of chapters. For the future, an evaluation of the portal in an actual teaching environment is planned to determine how well the goals of the portal preparation have been achieved. After that, the pilot version of the portal will be upgraded and new chapters prepared. Our estimation is that for an applicative value in the classroom, at least 100 language problems need to be described (as mentioned before, almost 700 problems were identified in the corpus Šolar, though some of them will have to be described in more than one chapter). For the creation of the new material, the process of chapter preparation will have to be optimised, as the previously applied procedures have proven to be very time consuming. For example, the selection of

[24] At the end of 2015, a project aiming to at least double the size of the Šolar corpus was approved by the Ministry of Culture of the Republic of Slovenia. One of the main methodological changes is the collection of scans rather than photocopies of student texts—this will enable the development of digital repository of texts and thus facilitate the transcription process, as well as improve the archiving of the texts.

corpus examples for grammar exercises could be partially automatised, and in the second step conducted with the help of crowdsourcing. Secondly, further development of the portal's CMS could expedite the preparation of the online contents, as currently a lot of time is spent on manual formatting of specific elements. And finally, the creation of new content would have to be connected and synchronised with related corpus and lexicographic projects, e.g. the new dictionary of contemporary Slovene (Gorjanc, Gantar, Kosem, & Krek, 2015).

As it seems from the gathered experience, the ideal scenario would be to combine the compilation of corpora and development of didactic resources and materials into one seamless process. Such a process would consist of students writing essays on the computer, language technology tools would automatically identify errors for teachers, teachers would confirm the errors and their categorisation, the results (per text, student, class, region etc.) would be summarised and shared with other teachers and researchers, identified errors would be linked with resources with explanations and exercises (such as PGP), and ultimately, the teaching content would adapt to each student's individual needs and progress. There are early indications that we are not far from such process—the use of the digital media in schools is on the increase (e.g. in Slovenia, the 2015 PISA survey was for the first time conducted solely on computers), and there are websites such as Vocabulary.com and apps such as the Oxford Vocabulary Trainer (in testing at the time of writing) proving how different language technologies (automatic error detection, taggers, parsers etc.) can be combined to provide a more interactive and individual-oriented language learning experience. The only thing that the researchers and material developers need to ensure is that L1 language teaching and learning catches up with L2 and benefits from—as well as contributes to further progress of—these exciting new developments.

Acknowledgement The work described in this chapter was funded within the national project "Communication in Slovene" (2008–2013), financed by European Social Fund and the Slovene Ministry of Education, Science and Sports. The development of the concordance for the Šolar corpus was funded by the Ministry of Culture of the Republic of Slovenia.

References

Aijmer, K. (2009). *Corpora and language teaching*. Amsterdam: John Benjamins.

Arhar Holdt, Š., & Stritar Kučuk, M. (2012). Korpusna analiza podaljševanja samostalnikov moškega spola na -o. In N. Jakop & H. Dobrovoljc (Eds.), *Pravopisna stikanja: Razprave o pravopisnih vprašanjih* (pp. 179–190). Založba ZRC: Ljubljana.

Arhar Holdt, Š., Kosem, I., Krapš Vodopivec, I., Ledinek, N., Može, S., Stritar Kučuk, M., … Zwitter Vitez, A. (2011). *Pedagoška slovnica pri projektu Sporazumevanje v slovenskem jeziku: K16 - standard za korpusno analizo slovničnih pojavov*. Ljubljana: Ministrstvo za šolstvo in šport: Amebis, 2011. Retrieved from http://www.slovenscina.eu/Media/Kazalniki/Kazalnik16/Kazalnik_16_Pedagoska_slovnica_SSJ.pdf.

Aston, G. (2001). Learning with corpora: An overview. In G. Aston (Ed.), *Learning with corpora* (pp. 4–45). Houston, TX: Athelstan.

Aston, G., Bernardini, S., & Stewart, D. (Eds.). (2004). *Corpora and language learners*. Amsterdam: John Benjamins.

Ball, F., & Wilson, J. (2002). Research projects relating to YLE speaking tests. *Research Notes, 7*, 8–10.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written english*. London: Longman.

Campoy, M., Gea-Valor, M., & Belles-Fortuno, B. (2010). *Corpus-based approaches to english language teaching*. London: Continuum.

Coniam, D. (1997). A preliminary inquiry into using corpus word frequency data in the automatic generation of English language cloze tests. *CALICO Journal, 16*(2-4), 15–33.

Conrad, S. (2000). Will corpus linguistics revolutionize grammar teaching in the 21st century? *TESOL Quarterly, 34*, 548–560.

Gorjanc, V., Gantar, P., Kosem, I., & Krek, S. (Eds.). (2015). *Slovar sodobne slovenščine: Problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete UL.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Hunston, S., & Francis, G. (2000). *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.

Johns, T. (1991). Should you be persuaded: Two samples of data-driven learning materials. In T. Johns & P. King (Eds.), *Classroom concordancing ELR journal 4*. Birmingham: University of Birmingham.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). The sketch engine. In G. Williams & S. Vessier (Eds.), *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004 Lorient, France July 6–10, 2004* (pp. 105–116). Lorient: Universite de Bretagne-sud.

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. *Proceedings of 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 (Zbirka Sporazumevanje)* (1st ed.). Ljubljana: Trojina, zavod za uporabno slovenistiko.

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). London: Longmann.

Logar, N., & Krek, S. (2012). New Slovene corpora within the communication in Slovene project. *Prace Filologiczne, 63*, 197–207.

McCarthy, M., McCarten, J., & Sandiford, H. (2005–2006). *Touchstone* (Books 1–4). Cambridge: Cambridge University Press.

Može, S. (2013). Raba kratkega nedoločnika: Korpusni pristop. *Slovenščina 2.0, 1*(1), 155–175. Retrieved from http://www.trojina.org/slovenscina2.0/arhiv/2013/1/Slo2.0_2013_1_08.pdf.

O'Keeffe, A., Mccarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge: Cambridge University Press.

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). Amsterdam: Rodopi.

Römer, U. (2005). *Progressives, patterns, pedagogy: A corpus-driven Approach to English progressive forms, functions, contexts and didactics*. Amsterdam: John Benjamins.

Rozman, T., Krapš Vodopivec, I., Stritar, M., & Kosem, I. (2012). *Empirični pogled na pouk slovenskega jezika. Zbirka Sporazumevanje*. Ljubljana: Trojina, zavod za uporabno slovenistiko.

Scott, M. (2008). *WordSmith Tools, version 5*. Liverpool: Lexical Analysis Software.

Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education*. Amsterdam: John Benjamins.

Sinclair, J. (2005). Corpus and text - Basic principles. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice. Books: 1–16*. Oxford: Oxbow. Retrieved November 18, 2015, from http://ahds.ac.uk/linguistic-corpora/.

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

Verdonik, D., Kosem, I., Zwitter Vitez, A., Krek, S., & Stabej, M. (2013). Compilation, transcription and usage of a reference speech corpus: The case of the Slovene corpus GOS. *Language Resources and Evaluation, 47*(4), 1031–1048.

Willis, J., Willis, D., & Davids, J. (1988–1989). Collins COBUILD English course (Parts 1–3). London: HarperCollins.

Resources

Arhar Holdt, Š., Červ, G., Gantar, P., Kosem, I., Kosem, K., Krapš Vodopivec, I., … Zwitter Vitez, A. (2013). *Pedagoški slovnični portal*. Ljubljana: Ministrstvo za izobraževanje, znanost, kulturo in šport. Retrieved from http://slovnica.slovenscina.eu/.

Logar, N., Krek, S., Erjavec, T., Grčar, M., Holozan, P., & Šuster, S. (2012). *GIGAFIDA*. Ljubljana: Ministrstvo za izobraževanje, znanost, kulturo in šport. Retrieved from http://www.gigafida.net.

Rozman, T., Stritar Kučuk, M., Kosem, I., Krek, S., Krapš Vodopivec, I., Arhar Holdt, Š., & Stabej, M. (2012). *Šolar*. Ljubljana: Ministrstvo za izobraževanje, znanost, kulturo in šport. Retrieved from http://www.korpus-solar.net/.

Sinclair, J. (1987). *Collins COBUILD English language dictionary*. London: HarperCollins.

Zwitter Vitez, A., Zemljarič Miklavčič, J., Krek, S., Stabej, M., & Erjavec, T. (2012). *Gos*. Ljubljana: Ministrstvo za izobraževanje, znanost, kulturo in šport. Retrieved from http://www.korpus-gos.net/.