Slovenski jezik – Slovene Linguistic Studies 7 (2009): 135–152

Mojca Stritar
Filozofska fakulteta, Ljubljana

Slovene as a Foreign Language: The Pilot Learner Corpus Perspective

Korpus usvajanja tujega jezika je elektronska zbirka besedil nedomačih govorcev, ki usvajajo določen jezik. V prispevku je predstavljen pilotski korpus slovenščine kot tujega jezika PiKUST. Pri njegovem oblikovanju sta se kot osrednja izziva pokazala razvoj smiselnega nabora kriterijev za zbiranje in izbiranje materiala učečih se ter razvoj klasifikacije, ki olajša označevanje napak. Opisana sta klasifikacija napak in postopek označevanja, nazadnje pa je predstavljen primer praktične aplikacije korpusnih rezultatov.

A learner corpus is an electronic collection of texts produced by non-native speakers learning a certain language. The article presents the pilot Slovene learner corpus PiKUST. Its design has faced two major challenges: the development of a reasonable set of criteria for collection and selection of learner material, and the development of an error annotation scheme. The error taxonomy and tagging procedure are described and, finally, the application of corpus results to teaching material is demonstrated.

Introduction

After Slovene independence in 1991 and the accession to the European Union in 2004, the interest in Slovene as a foreign language (SFL) has been growing. Along with the increasing number of non-native speakers learning Slovene and the consecutive development of language learning materials, the notion that foreign speakers of Slovene should be involved in all stages of modern Slovene reference-book planning has also been gaining importance (Stabej 2004: 12). Current bilingual dictionaries are mostly outdated and do not cater to SFL learners' needs since information vital to non-native speakers, such as morphological or contextual information, is often omitted (Rozman 2004: 66). Textbooks should also take more of the most significant learner difficulties into account rather than follow a syllabus based on native speakers' intuition.

Modern approaches to linguistic research on SFL are thus increasingly desired. The productivity and relevance of learner corpora to language analysis have been proved by their expanding research scope and their use worldwide (Pravec 2002, Granger 2004, Stritar 2006b). If some years ago the majority of learner corpora dealt with English as a foreign language, today the number of other, lesser-used target languages, such as Dutch (Cucchiarini et al. 2008) or Finnish (Jantunen 2008), is increasing.

"Learner corpora, also called interlanguage (IL) or L2 corpora, are electronic collections of authentic foreign or second language data" (Granger 2003: 465). Texts are produced by non-native speakers of a certain language. They can be either written or spoken and are selected according to carefully balanced learner- and task-related criteria, for instance the learner's first language (L1) and language competence.

Learner corpora provide a "deviation from the standard, i.e. the language of the native speakers of a particular language" (Pravec 2002: 81). Error annotation in a huge amount of linguistic data enables the user to perform efficient and exact quantitative error analysis. Although this is not the only aspect of interlanguage research and is typically upgraded with qualitative methods such as contrastive interlanguage analysis, computer-aided error analysis still remains one of the most important issues (Granger 2004). Its results can be used for second language acquisition research or for application of corpus data to materials such as textbooks, dictionaries, CALL, syllabus design and classroom methodology.

While Slovene language technologies and resources have been developing rapidly, learner corpora have been lagging behind. The only important project is a collection of 138 essays written by Serbo-Croatian learners of Slovene. The error-tagged corpus was used for language-transfer research (Balažic Bulc 2004), but cannot be widely accessed.

The design of Slovene learner corpus faced two major scientific challenges: the development of a reasonable set of criteria for collection and selection of texts, and the development of an error-tagging system. To resolve these questions, pilot corpus PiKUST[1] was created which redefines the criteria for collection, selection and documentation of learner material. This corpus also develops and tests mark-up conventions and error-tagging principles. An important model for these was the Norwegian corpus, ASK (Norsk andrespråkskorpus).[2] With regard to some Slovene-specific problems, information was gathered by a questionnaire answered by 46 SFL teachers and translators.

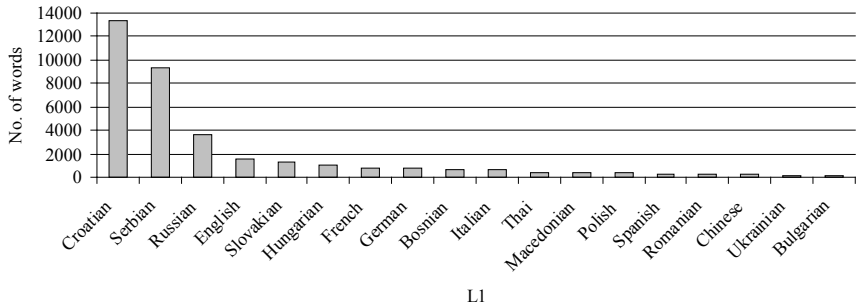
Design

PiKUST contains 35,000 words from 128 texts written by 119 learners of SFL with 18 different first languages: Bosnian, Bulgarian, Chinese, Croatian, English, French, German, Hungarian, Italian, Macedonian, Polish, Romanian, Russian, Serbian, Slovakian, Spanish, Thai and Ukrainian. The distribution of L1s is shown in Figure 1.

[1] PiKUST: Poskusni korpus usvajanja slovenščine kot tujega jezika. The corpus was designed and compiled during my Marie Curie host fellowship stay at Bergen Advanced Training Site in Multilingual Tools, Norway, in 2006/2007.

[2] <http://decentius.aksis.uib.no/corpus/askdemo-home.xml>.

Figure 1. Distribution of L1s in PiKUST



Most languages are represented with less than 2,000 words, so their parts are illustrative examples rather than a reliable basis for research. Larger and thus more representative are the Croatian, Serbian and Russian sub-corpora. The fact that around 67% of PiKUST texts were written by Croatian or Serbian speakers is a reflection of Slovene sociolinguistic reality – almost 89% of foreigners living in Slovenia are from one of the former Yugoslav republics (Antončič et al. 2006; 76, 96; Ilić et al. 2008: 98). The situation is similar at Slovene language exams; at the intermediate level 50% and at the advanced level 40% of participants speak Serbian or Croatian as their first language (Ferbežar 2006: 24).

92% of texts were written at the Slovene language exam for foreigners at the Centre for Slovene as a Second/Foreign Language in 2001.[3] 8% of texts were written at various Slovene language courses in 2005 and 2006. They provide material produced in an untimed and less restricted task setting.

Since the corpus was compiled opportunistically, learners' proficiency in Slovene was not carefully controlled. The majority of texts, 112, were written by advanced learners,[4] six by learners at the intermediate level[5] and three by beginners from South Slavic countries.[6] Texts by non-Slavic beginners were not included. It is a general principle among learner corpora compilers to exclude real beginners due to the instability of their interlanguage and the complexity that error tagging of their texts would require (Stritar 2006a: 137).

As in most other corpora (Stritar 2006a: 133), PiKUST texts are mostly argumentative essays with titles such as: My ideal job, Man and nature, Vernacular language in the media, Party without alcohol etc. Some other types were also included,

[3] <http://www.centerslo.net/l1.asp?L1_ID=3&LANG=slo>.
[4] Level C1 according to the Common European Framework of Language Reference.
[5] Level B2 according to the Common European Framework of Language Reference.
[6] At SFL courses they are usually considered to be a specific group, as the similarities between languages enable them to communicate in Slovene from the very start (Požgaj Hadži, Ferbežar 2001: 59).

namely formal letters which are a part of the exam, self-presentation which forms an important task at placement tests for language courses and diaries written as homework.

Compilation

The material was collected in the archives of the Centre for Slovene as a Second/Foreign Language, University of Ljubljana, and at various language courses organized by the same institution. Some texts were written as regular homework and some as a special home assignment.

The realization of corpus usually takes more time than anticipated (see Cucchiarini et al. 2008) and PiKUST was no exception. All texts were handwritten and needed to be digitalized. It took up to 20 minutes to type a 200-word text. Although the anticipated obstacle during this phase were typing errors made by the typist, it turned out that a greater problem was the native speaker's inclination to unintentionally type the correct form instead of the original error.

Mark-up and tagging

The header of each document contains meta-linguistic information about each text: learner and text ID, title, date of creation, source of text, learner's competence in Slovene, L1, age, gender, Slovene-speaking ancestors, education, profession, parents' L1, other languages spoken by the learner, years of learning Slovene and stays in Slovenia. Texts are tagged in XML following the TEI guidelines. But apart from the meta-linguistic data, the most important part of the PiKUST project was error annotation. By considering learner errors a part of the normal language acquisition process and by analyzing error-tagged learner corpora, researchers gain insight not only into errors, but also into learners' interlanguage as a whole (Granger 1998).

Usually different error taxonomies are used which make quantitative analysis more efficient and the research clearer and more detailed. Error annotation systems should be informative but manageable, reusable, flexible and consistent (Granger 2003: 467). Consistency can be achieved with a well-documented error-tagging manual, but despite that, inter-rater reliability is not always self-evident. Categorizing learner errors in general "is a laborious and oftentimes fruitless job, for there are various ways of classifying errors, depending on research interest and theories involved and it is often the case that the classification is only as valid as the theory it is based on" (Tono 2003: 801).

Error taxonomies based on surface strategy are often used in corpora. Its common categories are commission, ommission and insertion while misselection/misformation, misordering and contamination/cross-association (James 1998) are used less frequently. This classification in itself, however, is formal and does not offer useful information, which is why it is usually combined with classifications based on

linguistic categories, such as orthography, phonology, morphology, syntax, lexicon, discourse. Although distinctions between levels are not always clear and in some cases have to be agreed upon before tagging, these sorts of tags are relatively reliable and "descriptive rather than interpretative" (Granger 2003: 467).

Error classification and tagging in PiKUST

The process of error analysis includes five stages: collection of a sample of learner language, identification of errors, their description, explanation and evaluation (Ellis 1994: 48). Although the latter two are considered most important, they do not concern language corpora design, compilation and tagging. Researchers frequently confuse the explanatory and descriptive aspects of error analysis (Dulay, Burt, Krashen 1982: 141) but the task and objective of corpora is to offer relevant data for further linguistic research and not to perform it themselves. "[T]he accurate description of errors is a separate activity from the task of inferring the sources of those errors" (Dulay, Burt, Krashen 1982: 145). The description of errors involves "a comparison of the learner's idiosyncratic utterances with a reconstruction of those utterances in the target language. It requires, therefore, attention to the surface properties of learners' utterances (i.e. it does not attempt, at this stage, to identify the source of the errors)" (Ellis 1994: 54). Learner corpora do not explain errors, they simply offer material for further analysis. Therefore error classification and tagging have to be informative and descriptive, but non-biased or interpretive. They should also be simple, consistent, formal and general. Thus, even the popular theoretical distinction between mistakes and errors, based on learner's ability to self-correct his deviant forms,[7] has been largely avoided by learner corpora. "We have deliberately decided not to use distinctions such as 'errors' versus 'mistakes' or 'interlingual' versus 'intralingual' errors, which are difficult to assign and better left for a second stage in the analysis" (Granger 2003: 467). They are too biased because they "always [bear] with [them] the chance of a faulty assumption on the part of a teacher or researcher" (Brown 1980: 165). In short, error explanation and evaluation are not a part of the PiKUST design, compilation and tagging process and the corpus data can only be a basis for complete error analysis.

The PiKUST error annotation scheme is based on linguistic categories, frequent in pre-corpora linguistic research (Dulay, Burt, Krashen 1982: 147, Ellis 1994: 54) as well as in existent learner corpora (Pravec 2002, Tono 2003), combined with surface strategy. "[T]here is a great benefit to combining them into a single bidimensional taxonomy" (Granger 2003: 467). Tags were added manually. This is common in learner corpora; until now, automatic error detection has been tested in only one

[7] A mistake is an error of performance, failure to utilize a known system correctly. It can be corrected by its author. An error is a deviation from the adult grammar of a native speaker, reflecting the interlanguage competence of the learner. It cannot be self-corrected until further relevant input has been provided and converted into intake by the learner (Corder 1974a: 25, Brown 1980: 165, James 1998: 84).

learner corpus but with limited success (Izumi et al. 2004). The error-tagging process in PiKUST proved extremely time-consuming; annotation of each text required from thirty minutes to one hour, depending on error frequency. Some months after initial tagging, additional texts were tagged to recheck the adequacy and reliability of the error taxonomy and the comprehensibility of the error-tagging manual.

PiKUST errors are classified in two levels: the error domain and, if necessary, a more detailed linguistic category or surface strategy. The basic classification is shown in Table 1; the categories are discussed in more detail below.

Table 1. Error annotation scheme in PiKUST[8]

Level 1	Level 2	Example
Orthographic errors	Spelling	***glasba*** *(= glasba)*
	Word fusion/division	***naprimer*** *(= na primer)*
	Capitalization	*Evropska* ***Komisija*** *(= Evropska komisija)*
	Punctuation	*konec* ***19*** *st. (= konec 19. st.)*
	Secondary error	
Lexical errors	Existent word	*Življenje je zabavno, če* ***veš*** *živeti! (= znaš)*
	Nonexistent word	*Dopoldne sem* ***lopatil*** *sneg (= kidal)*
	Secondary error	
Morphological errors	No sub-types	*nimam samo* ***nemška*** *državljanstva (= nemškega)*
	Secondary error	
Errors in structure	Structure	*Menim, da* ***časa vedno ima*** *(= je vedno čas)*
	Unclear meaning	*Jaz menim da je dobro biti lepo oblečen in* ***držati do sebe***
	Word order	*Mogoče* ***res je****, da (= je res)*
	Ommission	*da zapomnite samo eno (= da si zapomnite samo eno)*
	Insertion	*človek* ***si*** *zgubi kontrolo (= človek zgubi kontrolo)*
	Secondary error	

Orthographic errors affect the written form: spelling, word fusion/division, capitalization and punctuation. Spelling errors regard phoneme groups which are also problematic for native speakers since the pronunciation does not match the standard written form. More learner-specific problems are difficulties with letters <č>, <š>, <ž>, phonetic spelling (*kniževnost* vs. *književnost* 'literature')[9] or non-Slovene spell-

[8] All examples in the article are from PiKUST. Secondary errors of all types are discussed below.

[9] The letter <j> in the group <nj> is normally not pronounced.

ing (*concerti* vs. *koncerti* 'concerts', *večinoma* vs. *večinoma* 'mostly'). Word fusion/division is a problem which can also be found in native speakers' written production. Words that should be written separately are written together or the opposite (*naprimer* vs. *na primer* 'for example'; *prvi krat* vs. *prvikrat* 'for the first time'). Capitalization rarely influences understanding and is not given special attention either in SFL courses or at the language exam. Still, capitalization errors are tagged in PiKUST, either the erroneous use of minuscule (***zemlja*** *kot planet* vs. *Zemlja kot planet* 'the planet Earth'*)* or capital letters (***Novo*** *leto* vs. *novo leto* 'New Year'). Similarly, punctuation errors are frequent but do not obstruct communication seriously and are seldom given particular attention in SFL learning or testing (Ferbežar 2007: 32). Punctuation can be redundant (***i.t.d.*** vs. *itd.* 'etc.'), missing (*Ne strinjam se s trditvijo po kateri* vs. *Ne strinjam se s trditvijo, po kateri* 'I don't agree with the statement according to which') or incorrect (*Ali gre za izgovor ali za "modno muho".* vs. *Ali gre za izgovor ali za "modno muho"?* 'Is it an excuse or a "fashion trend"?').

Learners and native speakers tend to consider lexical errors more serious than grammatical errors (Ellis 1994: 63). The second-level distinction of lexical errors in PiKUST is between existent and nonexistent words. The former include the use of Slovene words in an inappropriate context and the latter the use of words that do not exist in Slovene (*sem* ***lopatil*** *sneg* vs. *kidal* 'I cleared away the snow'; ***življenjski*** *sopotnik* vs. *življenjski* 'fellow traveler through life'). Incorrect verbal aspect, verbal and other prefixes (***promišljevanje*** vs. *premišljevanje* 'thinking') or the use of vernacular expressions (*bi radi* ***probali*** vs. *poskusili* 'would like to try') were also tagged as lexical errors, as well as the use of inappropriate word class (*Bom* ***praznik*** *rojstni dan* vs. *praznoval* 'I will celebrate my birthday').

Slovene is a richly inflected language and morphological errors are common whenever word forms, mostly inflections, do not match the standard form. This includes incorrect case, number or gender of nouns, adjectives and pronouns, incorrect conjugations in present tense (***postajo*** vs. *postanejo* 'they become') or the form of past participle (***podčrtao*** *bi* vs. *podčrtal* 'I would underline') etc. When developing the annotation scheme, it seemed appealing to classify morphological errors into errors of case, number, tense etc., but it proved impossible since subjective decisions could not be avoided. In the example *evropski očeti in* ***mami*** 'European fathers and mothers' the plural ending of noun *mama* is incorrect and indeed could be the masculine ending. However, it would be interpretive and thus unjust to say that the learner chose it because he thought *mama* was a masculine noun. Since the decision would have been based on logic, inferences and guesswork, it has been avoided and morphological errors have no subtypes.

Error in structure is a general term for syntactical errors and erroneous multi-word units. They include structure errors, multi-word units with unclear meaning, incorrect word order and omission or insertion of a word or phrase. A phrase is tagged as a structure error when the meaning is clear but the combination of words is incor-

rect (*moram živeti v Švici **za poznati** tudi državo moje mame* vs. *moram živeti v Švici, da bi poznala tudi državo moje mame* 'I have to live in Switzerland to get to know my mother's country'). The opposite errors are units where grammar is acceptable but their meaning cannot be inferred (*Jaz menim da je dobro biti lepo oblečen in **držati do sebe***). Incorrect word order is a common problem in Slovene and can include difficulties with the position of enclitic strings, within them (*s tem **težava bi bila** manjša* vs. *s tem bi bila težava manjša* 'with this, the problem would be smaller') or within noun phrases (*imam **slovenščine tečaj*** vs. *imam tečaj slovenščine* 'I have my Slovene language course'). Negations, auxiliary verbs, pronouns and prepositions are frequently omitted (*Pomembno je da zapomnite samo eno* vs. *Pomembno je, da si zapomnite samo eno* 'It is important you remember only one thing'), while different types of pronouns are erroneously inserted (*Vsak glasbenik, posebno če **on** je koncertant* vs. *Vsak glasbenik, posebno če je koncertant* 'Every musician, especially if he performs at concerts').

Secondary errors are a specific category for forms that are correct but need to be changed once an erroneous form in their vicinity has been corrected. For example, in *hvala **na** razumevanju* 'thank you for your understanding', the preposition *na* is incorrect. After it has been tagged and corrected to the appropriate *za*, the noun *razumevanju* also needs to be added a tag so its locative case is changed to the accusative *razumevanje* which is required by the new, correct preposition. It would be unjust to tag the noun as a genuine error since it is correct in the original context, so secondary errors have been added to all error categories and the user has to be aware of their specific status during error explanation and evaluation. In PiKUST, 77% of secondary errors are morphological.

Although PiKUST was not POS-tagged, POS-tags were manually added to one-word errors. Thus, it is possible to sort them by grammatical categories and to draw up lists of relevant error categories for each one. To enhance the relevance of the results, the simplified version of POS-classification used in different Slovene corpus projects, such as the reference corpus FIDA (Erjavec et al. 2001), was applied to PiKUST. Its categories are noun, verb, adjective, adverb, pronoun, numeral, conjunction, preposition, particle, interjection and abbreviation.

All aspects of PiKUST error tagging cannot be addressed in this article, for instance tagging of forms with multiple errors. But it should be borne in mind "that error annotation will always contain an element of subjectivity as the very notion of error is far from clear cut" (Granger 2003: 474). Some errors can be placed into two or more categories. In the example ***avtobuska** postaja* vs. *avtobusna* 'bus station', the adjective *avtobuska* could be tagged as a spelling error as only one letter is erroneous. But the confusion of *-sna* with *-ska* exceeds common spelling problems, so it was tagged as a lexical error. The erroneous use of pronouns (*vsak otrok gre v šolo ker **ji** je všeč* vs. *mu* 'each child goes to school because he likes it') can be tagged either as a lexical (speaker has chosen the wrong word) or morphological error (speaker has cho-

sen the wrong word form, i.e. the feminine instead of masculine). Since confusion of pronouns is a lexical more than a morphological issue, the former category has been chosen. However, all these cases stress the importance of a detailed error-tagging manual and the user's awareness of the fundamental subjectivity of tagging.

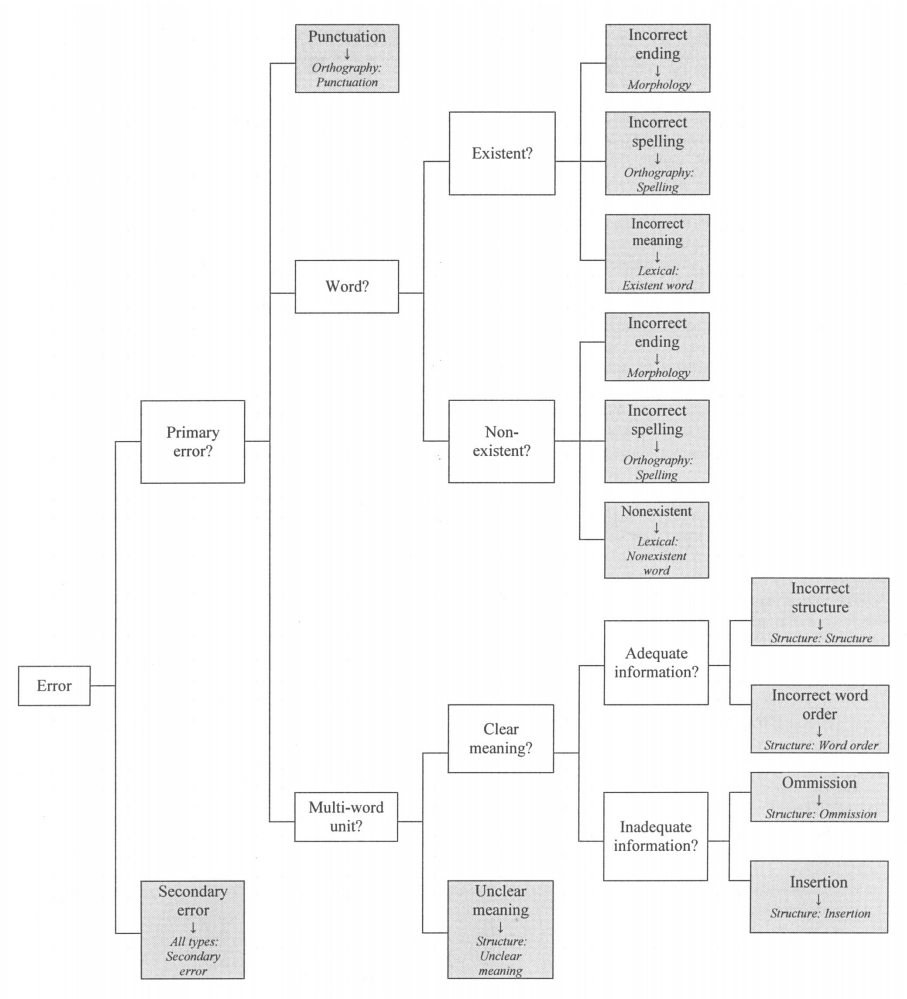
One controversial issue is the insertion of correct forms into error tags. Without the possibility to contact the learner, the reconstructions are based on form and context and are thus only probable (Corder 1982: 38). In the sentence *Ker je sijala polna luna, nisem rabil* ***prvih*** *luči* 'Because the full moon was shining, I didn't need my first lights' the adjective *prvih* 'first' is unsuitable. To correct it means to guess while trying to infer the meaning – was the learner referring to his car's headlights and the corrected form would be *sprednjih luči* 'front lights', *dolgih* 'high beam headlights' or *kratkih* 'low beam headlights', or was he trying to convey something completely different? Still, even if the reconstructions are probable, it is common to annotate correct forms in learner corpora. One always tries to find a meaning and thus to reconstruct the errors. Reading is pragmatic and error descriptions are based on "translations" of erroneous forms (Corder 1982: 37). Reconstructions are a chance to offer further explanation, increase intelligibility and facilitate subsequent interpretation of error annotations. From a technical perspective they are necessary for automatic lemmatization, POS-tagging, syntactic parsing or simply automatic sorting on the correct forms (Granger 2003: 469).

Considering all of the above, correct forms were assigned to each error in PiKUST, unless it was impossible to infer the meaning. For instance, it is easy to assume that the corrected form of misspelled *življenje* would be *življjenje* 'life' but it is difficult to guess the correction of the inadequate verb in *Ampak to je človeški last da se spreminja in* ***ugaja*** *s svojim časom* 'But it is typically human to change and ? with one's time'. The correction of phrase ***Prvič z internetom je****, da je povsod pri nas* would require a total paraphrase such as ***Prva lastnost interneta je****, da je povsod pri nas* 'The first characteristic of internet is that it is all around us'. Only 0.7% of PiKUST errors have reconstructions that are impossible or too interpretive. They are marked with "?" instead of a correction. But the PiKUST user has to bear in mind that some corrections are indisputable while other only have an indicative value.

The objectivity of error annotation increased with a manual parsing procedure that brought tagging closer to the reliability of automatic systems. The procedure is shown in Figure 2.

Statistics extracted from the corpus show the distribution of error domains is quite balanced: there are 35.1% of orthographic errors, 21.8% lexical, 25.7% morphological and 17.2% of errors in structure. Using the corpus statistics, the error types have also been ranked in decreasing order of frequency, shown in Table 2. These results confirm the mostly intuitive notion of SFL teachers that learners have most difficulties with morphology.

Figure 2. Error classification procedure in PiKUST; final categories are in italics



Practical application

As the pilot corpus PiKUST is not balanced and the sub-corpora for different first languages are small, it does not give results which could be generalized to all SFL learners or applied to language resources. However, the Croatian and Serbian sub-corpora have together 22,679 words which is enough to demonstrate a possible application of corpus results to teaching material.

Table 2. Distribution of error types in PiKUST

Error type	No. of occurrences	Percent
Morphology	1219	23.9%
Orthography: Punctuation	1102	21.7%
Lexical errors: Existent words	671	13.2%
Structure: Word order	406	7.9%
Orthography: Spelling	311	6.1%
Lexical errors: Nonexistent words	255	5%
Secondary errors	250	4.9%
Structure: Omission	130	2.5%
Structure: Insertion	126	2.4%
Orthography: Word fusion/division	75	1.4%
Structure: Structure	29	0.5%
Orthography: Capitalization	29	0.5%
Structure: Unclear meaning	18	0.3%
Total errors	**5085**	

The Centre for Slovene as a Second/Foreign Language has been developing *Pot do izpita iz slovenščine (PIS)*,[10] a textbook aimed at learners preparing for the Slovene language exam at intermediate or advanced level. As already mentioned before, the majority of participants at the exam and preparatory courses speak Serbian or Croatian as L1, so they are the anticipated main users of *PIS*. The textbook is composed of different exam-type exercises. To see if these exercises take into account specific problems and needs of the main target group, the Croatian and Serbian learners of Slovene, we can compare them with findings from PiKUST.

Most frequent errors by Croatian and Serbian learners in PiKUST are shown in Table 3.

In PIS, there are 8 exercises in which learners have to recognize and correct errors in a short text. An example is shown in Figure 3.

In these exercises, there are 55 errors to be identified and corrected. Their types and frequencies are shown in Table 4 along with statistics for the same error types in Croatian and Serbian sub-corpora of PiKUST.

If we compare data from Table 3 and Table 4, we see that the most common error types South Slavic learners make are adequately covered in PIS. Still, some adjustments could be made according to the corpus results. Although the number of

[10] Alič, Tjaša, Huber, Damjan, Jerman, Tanja, Kern, Damjana, Stritar, Mojca, forthcoming. *Pot do izpita iz znanja slovenščine*. Ljubljana: Center za slovenščino kot drugi/tuji jezik Filozofske fakultete Univerze v Ljubljani.

Table 3. Distribution of error types in Croatian and Serbian sub-corpora

	Croatian and Serbian sub-corpus		Croatian sub-corpus		Serbian sub-corpus	
Error type	**No. of occurrences**	**Percent**	**No. of occurrences**	**Percent**	**No. of occurrences**	**Percent**
Orthography: Punctuation	802	32.4%	448	26.7%	354	33.9%
Morphology	551	22.2%	383	22.8%	168	16.1%
Lexical: Existent words	351	14.2%	212	12.6%	139	13.3%
Structure: Word order	187	7.5%	102	6.1%	85	8.1%
Lexical: Nonexistent words	159	6.4%	103	6.1%	56	5.4%
Orthography: Spelling	136	5.5%	92	5.5%	44	4.2%
Secondary errors	124	5%	81	4.8%	43	4.1%
Structure: Insertion	44	1.8%	30	1.8%	14	1.3%
Structure: Omission	40	1.6%	22	1.3%	18	1.7%
Orthography: Word fusion/ division	34	1.4%	17	1%	17	1.6%
Structure: Structure	32	1.3%	20	1.2%	12	1.1%
Orthography: Capitalization	11	0.4%	10	0.6%	1	0.1%
Structure: Unclear meaning	5	0.2%	5	0.3%	0	0%
Total	**2476**		**1679**		**1043**	

errors that need to be corrected in PIS is relatively low, it appears that too much attention is given to declination errors, comparison of adjectives and the use of verbal aspect which have a significantly lower frequency rate in PiKUST. On the other hand, spelling, insertion of pronouns (19 occurrences in PiKUST, 0.7%) and omission of prepositions (12 occurrences, 0.5%) should be more emphasized in the exercises. Also, no particular attention in PIS has been paid to the confusion of Slovene verbs *vedeti, znati, poznati* 'to know'. Such errors appeared 9 times in PiKUST since Serbo-Croatian learners do not have the Slovene semantic distinction. Punctuation errors

Figure 3. Exercise from textbook *PIS*

V besedilu »Razstava otroških knjig« je v vsaki vrstici največ ena napaka. Popravite besedilo: če je v vrstici napaka, jo označite, v desni stolpec pa vpišite pravilno obliko, če napake ni, pustite prazno. V besedilu je še 6 napak.

Razstava otroških knjig

0. Primer:	*V knjigarni Konzorcij bo med 18. in 22. ~~maja~~*	*majem*
1.	potekala 12. razstava tujih otroških knjig z	0-2
2.	naslovom Bologna po Bologni. Na razstavo	0-2
3.	bo predstavljeno najboljše, kar	0-2
4.	bilo je mogoče videti na letošnjem 41.	0-2
5.	mednarodnem sejmu knjig za otroci v	0-2
6.	italijanski Bologni, to je skoraj 500 knjig, ki	0-2
7.	so bile nagrajene ali pa so zanimive za slovenske	0-2
8.	bralce. Večinoma so napisale v angleškem jeziku.	0-2
9.	Razstavo, ki je jo pripravila Bedita Mlinar,	0-2
10.	bodo odpirali v torek, 18. maja.	0-2

Table 4. Distribution of error types in PIS exercises and PiKUST (Croatian and Serbian sub-corpora)

	PIS		**PiKUST**	
Error type	**No. of occurrences**	**Percent**	**No. of occurrences**	**Percent**
Morphology: Declination	21	38.18%	268	10.8%
Structure: Word order	7	12.73%	187	7.5%
Lexical errors: Pronouns	6	10.91%	153	6.2%
Lexical errors: Prepositions	5	9%	41	1.6%
Lexical errors: Verbal aspect	5	9%	11	0.4%
Morphology: Adjective comparison	4	7.27%	4	0.2%
Lexical errors: Conjunctions	3	5.45%	44	1.8%
Morphology: Conjugation	3	5.45%	146	5.9%
Structure: Omission	1	1.81%	12	4.8%
Total	**55**		**2467**	

present almost a third of Serbo-Croatian errors in the corpus but have been excluded completely from PIS. As it has been already pointed out, teachers rarely focus on them since they are less crucial for successful communication and are usually not graded at the Slovene language exams (Ferbežar 2007: 32). Nevertheless, further

research should be conducted on how native speakers of Slovene evaluate such errors. A decision on whether more pedagogical attention should be paid to punctuation errors could be reached on this basis. So, it can be concluded that the error-correction exercises in PIS prepare the learners for this sort of exercise at the Slovene language exam, but more attention should be paid to specific needs of its users.

The example described in this section has shown how PiKUST results can be applied to specific exercises for a narrow target group. They can be implemented to more widely oriented language resources as well, but basic issues should be resolved first: who is the "average" learner of Slovene as a foreign language and how specific should the "typical" errors be? For example, language-learning resources that focus only on errors made by speakers from one L1 group will have limited utility for other groups (e.g., English speakers do not need to learn not to write *ć* which is necessary for South Slavic learners), and vice-versa (e.g., Croatians do not need instruction in pro-drop necessary for the English). To answer these questions, a full-size learner corpus of Slovene with balanced sub-corpora for different L1s is needed. It should also be borne in mind that "error tagging, in spite of its numerous advantages, is only concerned with learner misuse. It fails to uncover other aspects of interlanguage such as the under- and overuse of words and phrases, which together with downright errors contribute to the nonnativeness of learner productions" (Granger 2003: 475). Learner corpora should not be seen as "a panacea, but rather as one highly versatile resource which SLA/FLT researchers can usefully add to their battery of data types" (Granger 2004: 129).

Conclusion

The pilot corpus PiKUST is a corpus kindergarten where basic principles have been tested before setting off for a Slovene learner corpus, larger in respect of size and meta-linguistic information. While PiKUST has been compiled opportunistically, more attention will be paid to the balance of different design criteria, especially learners' L1 and the task setting, during the compilation of the bigger corpus.

PiKUST error classification has been successfully tested as well as already implemented into the Slovene part of the TOOL2 European project: Tools for Online and Offline language learning.[11] Some questions remain open, for instance how to deal with erroneous forms during automatic lemmatization and POS-tagging. The biggest challenge, however, that has so far been left aside is compilation and tagging of a spoken learner corpus.

References

Antončič, Ana (ed.). 2006. *Statistični letopis Republike Slovenije*. Ljubljana: Zavod Republike Slovenije za statistiko.

[11] Project website <http://toolproject.eu>.

Balažic Bulc, Tatjana. 2004. Jezikovni prenos pri učenju sorodnih jezikov (na primeru slovenščine in srbohrvaščine). *Jezik in slovstvo*, vol. 3–4/49: 77–89.

Brown, H. Douglas. 1980. *Principles of Language Learning and Teaching.* Englewood Cliffs: Prentice-Hall.

Corder, S. P. 1974a. The Significance of Learners' Errors. *Error Analysis: Perspectives on Second Language Acquisition*. London: Longman: 19–27.

Corder, S. P. 1974b. Idiosyncratic Dialects and Error Analysis. *Error Analysis: Perspectives on Second Language Acquisition*. London: Longman: 158–171.

Corder, S. Pit. 1982: *Error Analysis and Interlanguage*. London: Oxford University Press.

Cucchiarini, Catia, Driesen, Joris, Van hamme, Hugo, Sanders, E. 2008. Recording Speech of Children, Non-Natives and Elderly People for HLT. *LREC 2008 Proceedings*. <http://www.lrec-conf.org/proceedings/lrec2008/summaries/366.html> (December 2008).

Dulay, Heidi, Burt, Marina, Krashen, Stephen. 1982. *Language Two*. New York, Oxford: Oxford University Press.

Ellis, Rod. 1994. *The Study of Second Language Acquisition*. Oxford: Oxford University Press.

Erjavec, Tomaž, Gorjanc, Vojko, Holozan, Peter, Stabej, Marko. 2001. Application to Slovene. *Specifications and Notation for MULTEXT-East Lexicon Encoding*. MULTEXT-East / Concede.

Ferbežar, Ina. 2006. Izpitni center. *Letno poročilo Centra za slovenščino kot drugi/tuji jezik 2005*. Ljubljana: Center za slovenščino kot drugi/tuji jezik: 24–27.

Ferbežar, Ina. 2007. *Navodila za izvajalce izpitov iz znanja slovenščine z načeli dobre prakse.* Internal material. Ljubljana: Izpitni center Centra za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete Univerze v Ljubljani.

Ferbežar, Ina. 2008. Izpitni center. *Letno poročilo Centra za slovenščino kot drugi/tuji jezik 2007*. Ljubljana: Center za slovenščino kot drugi/tuji jezik: 38–41.

Granger, Sylviane. 1998. The computer learner corpus: a versatile new source of data for SLA research. *Learner English on Computer*. London, New York: Longman: 3–18.

Granger, Sylviane. 2003. Error-tagged Learner Corpora and CALL: A Promising Synergy. *CALICO Journal* 20/3: 465–480.

Granger, Sylviane. 2004. Computer Learner Corpus Research: Current Status and Future Prospects. *Applied Corpus Linguistics: A Multidimensional Perspective*. Amsterdam: Rodopi: 123–145.

Ilić, Milena, Kalin, Katja, Povhe, Janja, Razpotnik, Barica, Šter, Darja, Žnidaršič, Tina. 2008. Prebivalstvo. *Statistični letopis Republike Slovenije*. Ljubljana: Zavod Republike Slovenije za statistiko: 72–101.

Izumi, Emi, Uchimoto, Kiyotaka, Isahara, Hitoshi. 2004. SST speech corpus of Japanese Learners' English and automatic detection of learners' errors. *ICAME Journal*, vol. 28: 31–48.

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

Jantunen, Jarmo Harri. 2008. Corpus driven analysis of cotextual units of meaning in learner language. Lecture at the conference *New trends in corpus linguistics for language teaching and translation studies*, Granada (September 2008).

Leech, Geoffrey. 1998. Preface. *Learner English on Computer*. London, New York: Longman: xiv–xx.

Lennon, Paul. 1991. Error: Some problems of definition, identification, and distinction. *Applied Linguistics*, vol. 12: 180–195.

Požgaj Hadži, Vesna, Ferbežar, Ina. 2001. Tudi to je slovenščina. *37. seminar slovenskega jezika, literature in kulture*. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete: 57–68.

Pravec, Norma. 2002. Survey of learner corpora. *ICAME Journal*, vol. 26: 81–114.

Rozman, Tadeja. 2004. Upoštevanje ciljnih uporabnikov pri izdelavi enojezičnega slovarja za tujce. *Jezik in slovstvo*, vol. 3–4/49: 63–75.

Stabej, Marko. 2004. Slovenščina kot drugi/tuji jezik in slovensko jezikovno načrtovanje. *Jezik in slovstvo*, vol. 3–4/49: 5–16.

Stritar, Mojca. 2006a. Oblikovanje korpusa usvajanja slovenščine kot tujega jezika. *Informacijska družba – IS 2006: Proceedings of the 9th International Multiconference / Zbornik 9. mednarodne conference*. Ljubljana: Institut "Jožef Stefan": 134–139.

Stritar, Mojca. 2006b. Merila za oblikovanje korpusov usvajanja tujega jezika. *Jezik in slovstvo*, vol. 5/51: 59–74.

Tono, Yukio. 2003. Learner corpora: design, development and applications. *UCREL Technical Paper: Proceedings of the Corpus Linguistics 2003 Conference*. Lancaster: University: 800–809.

Prispelo decembra 2008, sprejeto marca 2009
Received December 2008, accepted March 2009

Slovenščina kot tuji jezik: Pogled skozi pilotski korpus usvajanja tujega jezika

Slovenščina kot tuji jezik postaja vse pomembnejše področje jezikoslovnega raziskovanja, hkrati s tem pa raste potreba po korpusu usvajanja slovenščine kot tujega jezika. Korpus usvajanja tujega jezika je elektronska zbirka besedil nedomačih govorcev, ki usvajajo določen ciljni jezik. Besedila so pisna ali govorjena, izbrana pa so na podlagi natančno uravnoteženih dejavnikov, povezanih s tvorci besedil ali z okoliščinami njihovega nastanka. Označene napake v veliki količini jezikovnih podatkov omogočajo uspešno, natančno in relevantno kvantitativno analizo napak, ki je ponavadi dopolnjena s kvalitativnimi pristopi. Rezultati so uporabni tako za raziskave usvajanja tujega jezika kot za različne pedagoške namene.

Pri oblikovanju korpusa usvajanja slovenščine kot tujega jezika sta se kot osrednja izziva pokazala razvoj smiselnega nabora dejavnikov za zbiranje in izbiranje materiala učečih se ter razvoj klasifikacije, ki olajša označevanje napak. Namen pilotskega korpusa PiKUST je bil razrešiti oba izziva.

PiKUST vsebuje 128 besedil oziroma 35.000 besed 119 tvorcev z 18 prvimi jeziki (angleški, bolgarski, bošnjaški, francoski, hrvaški, italijanski, kitajski, madžarski, makedonski, nemški, poljski, romunski, ruski, slovaški, srbski, španski, tajski, ukrajinski). Besedila so bila napisana in zbrana na izpitih iz znanja slovenščine za tujce ter na različnih tečajih slovenščine. Večina učečih se je bila na izpopolnjevalni stopnji. Kot v drugih tovrstnih korpusih so besedila v glavnem utemeljevalni eseji.

Besedila so bila označena v jeziku XML. Poleg metajezikovnih oznak različnih podatkov o tvorcu in okoliščinah nastanka v glavi vsakega dokumenta so najpomembnejše ročno dodane oznake napak. Klasifikacija napak v PiKUST-u ima naslednje kategorije: napake zapisa (s podkategorijami črkovanje, pisanje skupaj ali narazen, velika oziroma mala začetnica, ločila), napake besedišča (s podkategorijami obstoječih oziroma neobstoječih besed), oblikoslovne napake in napake strukture (s podkategorijami strukture, neustreznega pomena ustrezno strukturiranih fraz, besednega reda, izpusta in vstavitve besede ali več besed). Vsi osnovni tipi imajo podkategorijo sekundarne napake za oblike, ki so v navedenem kontekstu ustrezne, a jih je treba spremeniti, ko je popravljena napaka v njihovi bližini.

Enobesedne napake imajo ročno dodane oblikoslovne oznake. Vse napake imajo pripisane pravilne oblike, razen če jih je bilo nemogoče predvideti. Objektivnost označevanja napak je bila povečana s formalnim postopkom, ki je povečal konsistentnost procesa in ga s tem približal avtomatiziranim sistemom.

Kot pilotski korpus PiKUST ni uravnotežen, zato so podkorpusi posameznih prvih jezikov premajhni za posploševanje rezultatov. Nekaj predlogov za izboljšavo učbeniškega gradiva na podlagi rezultatov hrvaškega in srbskega podkorpusa pa vendarle kaže možnosti za praktično uporabo korpusnih podatkov.

Slovene as a Foreign Language: The Pilot Learner Corpus Perspective

Slovene as a foreign language is gaining importance in language research and along with it the need for a Slovene learner corpus is growing. A learner corpus is an electronic collection of texts produced by non-native speakers learning a certain target language. Texts are either written or spoken, selected using carefully balanced learner- and task-related criteria. Error tagging in a huge amount of linguistic data enables the user to perform efficient, exact and relevant quantitative error analysis, usually upgraded with qualitative approaches. The results can be used for language acquisition research or pedagogical application.

The design of a Slovene learner corpus faced two major scientific challenges: the development of a reasonable set of criteria for collection and selection of learner material, and the development of an error-tagging system. To resolve these challenges, a pilot corpus named PiKUST was created.

PiKUST contains 35,000 words from 128 texts written by 119 learners with 18 different first languages (Bosnian, Bulgarian, Chinese, Croatian, English, French, German, Hungarian, Italian, Macedonian, Polish, Romanian, Russian, Serbian, Slovakian, Spanish, Thai and Ukrainian). Texts were written and collected at the Slovene language exam for foreigners and at various language courses. The majority of learners were at the advanced level of competence. As in other corpora, texts are mostly argumentative essays.

Texts were tagged in XML. Apart from meta-linguistic data in the header of each document regarding different learner- and task-related issues, manually added error tags are most important. PiKUST error classification has the following categories: orthographical errors (with sub-categories spelling, word division/fusion, capitalization, punctuation), lexical errors (with sub-categories existent or nonexistent word), morphological errors and errors in structure (with sub-categories erroneous structure, grammatically adequate phrase with unclear meaning, word order, word/phrase omission or insertion). All error categories have the sub-category secondary error for forms that are correct in the given context but need to be changed once an erroneous form in their vicinity has been corrected.

POS-tags were manually added to one-word errors. Corrected forms were assigned to each error unless it was impossible to infer the meaning. The objectivity of error tagging was increased with a manual parsing procedure, bringing the process closer to the reliability of automatic systems.

As the pilot corpus PiKUST is not well-balanced and the sub-corpora for different first languages are small, the results cannot be generalized. The article suggests some textbook improvements based on the Croatian and Serbian sub-corpora, thus demonstrating possible practical application of the corpus.