Saeed Safari*
Faculty of Philology
University of Belgrade

https://doi.org/10.18485/analiff.2018.30.2.13
811.222.1'33
Originalni naučni rad
Primljen: 12.10.2018.
Prihvaćen: 22.11.2018.

THE SALAM FARSI LEARNER CORPUS - INTRODUCING THE ERROR TAGGING SYSTEM

Linguistic corpora constitute reliable sources and empirical means for analyzing linguistic data. They are also widely used in the fields of Second/Foreign Language Acquisition and Foreign Language Teaching research, where the most commonly used type are Learner Corpora. This paper aims to introduce the the error annotation and tagging system of the very first error-tagged Persian learner corpus, called the *Salam Farsi Learner Corpus (SFLC),* as well as an analysis of linguistic errors based on a collection of written texts produced by Serbian learners of the Persian language. To set up the SFLC, three major stages, namely, constructing the corpus, proposing a system of error annotation and developing tools and software, were followed, and the practical phases such as the systematic collection of data and metadata, defining the corpus design criteria, creating the error tagsets and developing the corpus interface, software and specific tools were developed. The SFLC software is equipped with four main tools in order to function as an error-tagged learner corpus and provide the statistical reports.

Keywords: Learner corpus, Teaching Persian to Serbian, Corpus linguistics, Error analysis.

1. Introduction

Linguistic corpora provide reliable sources and empirical means for analysing linguistic data. They are also widely used in the field of Second Language Acquisition (SLA) and Foreign Language Teaching (FLT) research, where they are specifically referred to as learner corpora (LC). Today SLA research and this field of study is equipped with corpora resources which are used for FL/SL processing. Since the success of SLA research relies mainly on access to authentic data, applying Corpus Lin-

[*] Email: saeed.safari@fil.bg.ac.rs

guistics (CL) methods in collecting and analysing samples of what learners have produced during their learning could help researchers to define certain parameters on the way a second language is learned and investigate the second language acquisition process. Nowadays, many languages use CL tools and resources for annotating and analysing linguistic data in SLA research. In the case of the Persian language there is a great need to develop specialized corpora for research in Farsi as a Second/Foreign Language and to create the required tools and resources. The aim of developing an error-tagged learner corpus for the Persian language is to contribute to this effort. In this paper, the construction of the Salam Farsi Leaner Corpus (SFLC) (safari, 2017) is reviews shortly and later the error tagging system is introduced.

2. Developing the SFLC Design Criteria

The first step in constructing a corpus, including a learner corpus, is to identify the design criteria. The importance of adopting some criteria has been emphasized by many corpus developers and experts, such as Atkins *et al.* (1991), Biber (1993), Biber *et al.* (1998), Granger (1993a). When it comes to developing learner corpora, as indicated by Gilquin (2015: 12), "design criteria are even more crucial given the highly heterogeneous nature of interlanguage, which can be affected by many variables related to the environment, the task and the learner him-/herself." Therefore, exactly what will be included in the learner corpus should be clearly determined in advance. The issue of learner corpus design and its features were briefly discussed in 2.7 and it can be concluded that the corpus design criteria as well as the features and variables usually change based on the corpus research purposes. Tono (2003) emphasized such changes and concluded: "it is quite natural that the design of learner corpora will vary from project to project", as researchers are interested in different aspects of learner language. As for developing the SFLC, the proposed model consists of two types of features: (i) The Main Criteria for LC Design and (ii) The Specific Metadata for LC Design. Therefore, two types of data were collected: written texts and metadata variables.

The SFLC data were collected from two groups of learners: first, the students at the Faculty of Philology, University of Belgrade, and second, the learners who attended courses in the Persian language at the Iranian Cultural Center (ICC) in Belgrade, Serbia. The corpus data were collected from Serbian learners over three academic years between 2012 and 2015. The texts consist of excerpts from their homework in free writing and compositions (on specific subjects). The SFLC consists only of written productions (text type) and they are compositions and examples of free writing produced by Serbian learners (task type). Some other features are discussed in detail in the following sections. The SFLC, consists of 300 authentic written texts which in total contain 26,978 words. The corpus defines a target size of 100,000 words. The SFLC has been designed to identify the type and frequency of learners' errors; therefore it is an error-tagged learner corpus for academic purposes. The intended users of the corpus are researchers and scholars who wish to conduct research into the problems of learning Persian as a foreign language. A summary of the SFLC corpus criteria is given in Table 1.

Table 1: The SFLC design criteria

The SFLC Design Criteria		
1	Mode	Written
2	Size	26,978
3	Purpose	Academic use
4	Availability	Limited access
5	Users	Researchers
6	Text type	Written
7	Task type	Compositions, Creative writing (Free Writing)
8	Genre	Descriptive, Narrative
9	First Language	Serbian
10	Target Language	Farsi (Persian)
11	Level of Proficiency	A2 – C1
12	Annotation	Errors

3. Developing the SFLC Error Tagging System

Developing a system for error tagging is a basic theoretical requirement for constructing an error-tagged learner corpus; however, since linguistic errors differ from one language to another and error detection is generally for the purposes of research, there is no comprehensive error-tagging system to refer to. Therefore, researchers try to develop their own system of error annotation. Diaz-Negrillo & Fernandez-Dominguez (2006:86) believe that "research groups often appear to design their own error-tagging systems and explore different tagging models and error typologies. Indeed, the diversity of error-tagging systems seems to be evidence of the constant questioning of emerging approaches to error annotation, and also of the need for a benchmark for the analysis of computerized learner errors." However, Granger (2003) suggests that some requirements need to be met for the development of an error tagging system. According to Granger (ibid), an error system should be 'informative', 'reusable', 'flexible' and 'consistent' based on "observable criteria and be well described, in order to keep the degree of subjectivity low and thus ensure reliability."

4. The SFLC Error Taxonomy

The SFLC is an error-tagged corpus aimed at 'detecting', 'tagging' and 'reporting' the linguistic errors made by Serbian learners of the Persian language. To achieve this aim and for the purpose of detecting and tagging errors, the model of descriptive error classification and error taxonomies introduced by Dulay, Burt, & Krashen (1982) has been employed and expanded in the SFLC.

Dulay et al. (1982: 145) tried to introduce a comprehensive model for error taxonomies which "classify errors according to some observable surface feature of the error itself, without reference to its underlying cause or source." The model which is called 'error descriptive taxonomies' contains four main error taxonomies: (1) Linguistic Category (2) Surface Strategy, (3) Comparative Analysis and (4) Communicative Effect.

Taxonomy based on 'Linguistic Errors', as explained by Dulay *et al.* (1982) refers mainly to errors in the language component such as phonology, syntax and morphology, semantics and lexicon, and discourse. 'Surface Strategy' taxonomy concentrates on how learners modify target forms and the ways surface structures are altered. Dulay *et al*. (1982: 150) suggested four main categories for this taxonomy: (1) omission, (2) additions, (3) misformation, and (4) misordering.

‘Comparative Errors’ taxonomy deals with the comparison between the structure of L2 errors and other types of constructions, most commonly the errors made by children during their L1 acquisition. Dulay *et al.* (1982: 163-164) proposed four error categories related to this taxonomy: (1) developmental errors, (2) interlingual errors, (3) ambiguous errors, and (4) the ‘grab bag category’ of other errors.

The last proposed error taxonomy by Dulay *et al.* is ‘Communicative Effect’ which refers to those errors which impact on the listener or reader and hinder successful communication. Some groups of errors, known as global errors, affect the overall organization of the sentence and subsequently impede successful communication, while others, termed local errors, affect a single element of the sentence and do not hinder communication.

The SFLC uses the descriptive error taxonomy system by Dulay *et al*. (ibid) as the basic model for error classification and applies the first two subtypes (a) the Surface Strategy taxonomy and (b) Linguistic Category for developing the SFLC error taxonomy as explained below.

A. The SFLC Surface Structure Error Taxonomy

The first taxonomy introduced by Dulay *et al*. (1982), termed ‘Surface Strategy’, as they indicated (1982:150), “highlights the ways surface structure are altered”. Adopted for the SFLC, the taxonomy is termed Errors in the Surface Structure, which is the first level for error description in the corpus. The taxonomy retains the same four categories as introduced by Dulay *et al*. (4.2), however, the terms Substitution and Permutation are used instead of Misselection and Misordering. Table 2 shows the SFLC surface structure error taxonomy.

Table 2: The SFLC surface structure error taxonomy

Error Category	**Description**
Omission	The absence of a required element
Addition	The presence of an unnecessary or incorrect element
Substitution	The use of an incorrect element
Permutation	The misordering or incorrect placement of elements

B. The SFLC Linguistic Error Taxonomy

The SFLC employs two levels of error classification in the linguistic error taxonomy:

1. The Error Domains, which consists of 5 domains, namely, Orthography, Morphology, Syntax, Lexis and Style.
2. The Error Types, which specify errors related to the error domains. This category involves 22 error types, namely, Consonant Character(s), Long Vowel Character(s), Short Vowel Character(s), Connections, the Ezâfe Particle, Dots, Adjective, Noun-Plural, Noun (other), Pronoun, Preposition, Postposition (râ), Conjunction, Verb Tense, Verb Agreement, Verb (other), Adverb, Word Order, Word Selection, Phrase Selection, Cohesion and Unclear Style.

The SFLC error taxonomy model is based on the combination of the surface structure error taxonomy and the linguistic error taxonomy. In this model, errors will be identified, and subsequently selected and marked for the error annotation process in three categories as illustrated in Table 3.

Table 3: The SFLC error taxonomy

Errors in Surface Structure	Addition, Omission, Substitution, Permutation
Error Domains	Orthography, Morphology, Syntax, Lexis, Style
Error Types	Consonant Character(s), Long Vowel Character(s), Short Vowel character(s), Connections, the Ezâfe Particle, Dots, Adjective, Noun-Plural, Noun (other), Pronoun, Preposition, Postposition (râ), Conjunction, Verb Tense, Verb Agreement, Verb (other), Adverb, Word Order, Word Selection, Phrase Selection, Cohesion and Unclear Style.

5. The SFLC Error Tagset

The SFLC error tagset is developed based on the SFLC Error Taxonomy and includes a total of 31 errors. The errors are marked in three levels of annotation and on the basis of the tagset model. Each error is marked by a four-letter error tag. The first letter symbolises the error in surface structure, the second letter indicates the error domain, and the two last letters represent error type. The taxonomy is flexible, and therefore errors can be freely selected and combined on three levels of annotation. For example, in the error tag <O_M_VT>, the letter *O* indicates 'Omission' in the surface structure modification, the letter *M* represents the error domain which is 'Morphology', while the two last letters, *VT*, identify the specific error type which in this case is 'Verb Tense'. Table 4 shows the SFLC error tagset.

Table 4: The SFLC error tagset

First Level		**Second Level**		**Third Level**	
Surface Structure	**Abbr**	**Error Domain**	**Abbr**	**Error Type**	**Abbr**
Addition	A	Orthography	O	Consonant character(s)	CC
Omission	O	Morphology	M	Long Vowel character(s)	VL
Substitution	S	Syntax	S	Short Vowel character(s)	VS
Permutation	P	Lexis	L	Connections	CO
		Style	T	Ezâfe Particle	EP
				Dots	DT
				Adjective	AJ
				Noun-Plural	NP
				Noun Other	NO
				Pronoun	PR
				Preposition	PP
				Postposition (râ)	PO
				Conjunction	CN
				Verb Agreement	VA
				Verb Tense	VT
				Verb Other	VO
				Adverb	AD
				Word Order	WO
				Word Selection	WS
				Phrase Selection	PS
				Cohesion	CS
				Unclear style	US

The following examples explain how the annotation can be employed using the SFLC error tagset The first bracket is the incorrect form and the second one identifies the error in the surface structure.

(1)

*هر ماه به {کتاب فروش} <O_M_NO> می‌روم.

* har mâh be [ketâbforuš] <O_M_NO> miravam
The error tag: <O_M_NO> Omission_Morphology_Noun Other
Description: The noun suffix (i) has been omitted.
Correct Form: [ketâbforuši] هر ماه به کتابفروشی می‌روم.

Gloss[1]:
*har mâh be [ketâb-foruš]<O_M_NO> [ketâb-foruš=i] mi=rav=am
Every month to [book-sell] <O_M_NO> [book-sell.indef] cont-go.pres.1sg
“Every month I go to the bookstore”

(2)

*{خیلی} {اضافه} بارها این سوال پرسیده می شود.

The error tag: < A_L_AD > Addition_Lexis_Adverb
[xejli] [A_L_AD] bârhâ in so’âl porside mišavad
Description: An intensifier (xejli) has been added before another intensifier (an formed construction in Persian).
Correct Form: bârhâ in so’âl porside mišavad بارها این سوال پرسیده می‌شود

Gloss:
[xejli] < A_L_AD > bârhâ in so’âl pors=ide mi=šav=ad
[Many] < A_L_AD > times this question ask-PAST-pp cont-be-3sg
“This question is asked many times”

(3)

*دوستانم {بودند} در خانه.

The error tag: < S_S_VO > Substition_Syntax_Verb Other
*dustânam [budand] < S_S_VO > dar xâne.
Correct Form: dustânam dar xâne [budand].   دوستانم در خانه بودند
Description: The verb (budand) has been substituted with the adverb. Persian follows SOV, so verbs normally appear at the end.

Gloss:
dust-ân-am [budand] < S_S_VO > dar xâne [budand].

[1] Based on the Leipzig glossing rules,.segmentable morphemes are separated by hyphens, and clitic boundaries are marked by an equals sign.

friend-PL-POS [be-PAST.2sd] [S] at home [be-PAST.2sd]
"My friends were at home"

(4)

*{قفط} {جابجایی} به من بگو.

The error tag: < P_O_CC > Permutation_Orthography_Consonant Character
*[qafat] < P_O_CC > be man begu.
Correct Form: faqat be man begu. فقط به من بگو.

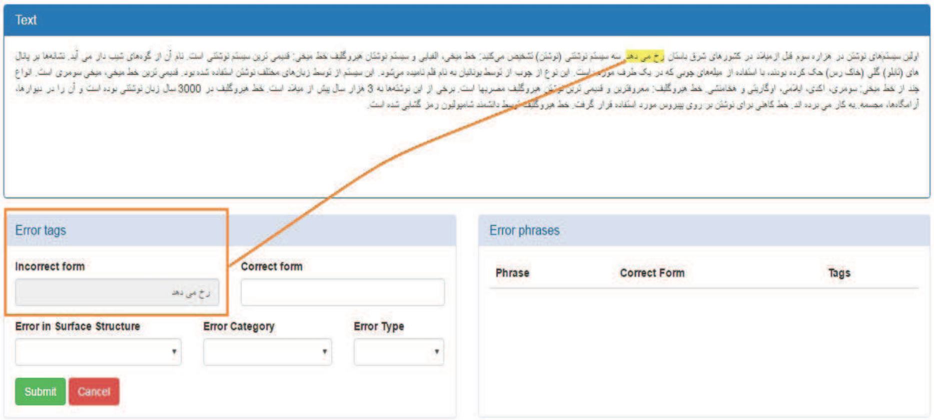
Description: In this word, the letter <f> has been misplaced with <q> due to the spelling similarity. They differ by one dot as <f فـ / <has one dot while <q / قـ > has two, which results in frequent mistakes in recognizing and spelling these letters.

Gloss:
[qafat] < P_O_CC > [faqat] be man be=gu
Just to me tell- IMP
"Just tell me"

6. The Error Tagging Tool

The main purpose of developing the ETT, which can be called a computer-aided error annotation tool, was to facilitate the error annotation process in the corpus. The tool is created based on the development of the SFLC Error Tagset. By using this tool, the user is able to (1) select word(s), phrase(s) or sentence(s) for error annotation, (2) suggest a corrected form for the selected error, (3) annotate each error by selecting error tags from three levels, and (4) edit or delete the selected error tags. The ETT contains 3 levels of error tagging for the surface structure, error domain and error types, consisting of a total of 31 error tags. The tool was designed in 3 sections, namely, 'the text box', 'the error tags box' and 'the error phrase box', although it functions as an integrated unit.

'The Text Box' shows the raw text which has already been submitted into the corpus database. Each character, word, phrase, sentence or even paragraph can be selected for error annotation simply by clicking on it or selecting a group of characters. The selected segment is highlighted in yellow and consequently is shown in the 'incorrect form' in the error tags box, where the selected segment should be annotated and subsequently submitted. Figure 1 shows the ETT text box and its function.



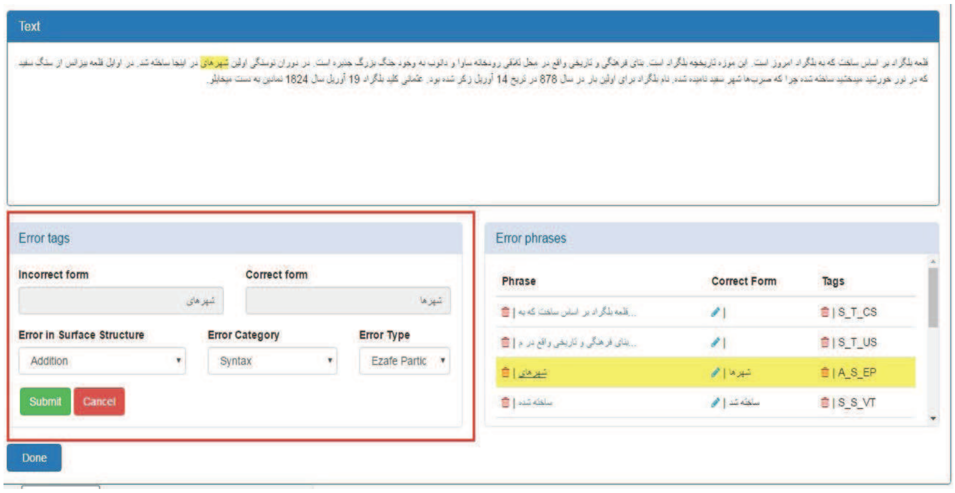
اولین سیستم‌های نوشتن در هزاره سوم قبل ازمیلاد در کشورهای شرق باستان رخ می دهد. سه سیستم نوشتنی (نوشتن) تشخیص می‌کنید: خط میخی، الفبایی و سیستم نوشتاری هیروگلیف خط میخی: قدیمی ترین سیستم نوشتنی است. نام آن از گوه‌های شیب دار می آید. نشانه‌ها بر پانل
های (تابلو) گلی (خاک رس) حک کرده بودند، با استفاده از میله‌های چوبی که در یک طرف مورب است. این نوع از چوب از توسط یونانیان به نام قلم نامیده میشود. این سیستم از توسط زبان‌های مختلف نوشتن استفاده شده بود. قدیمی ترین خط میخی، میخی سومری است. انواع
چند از خط میخی: سومری، اکدی، ایلامی، اوگاریتی و هخامنشی. خط هیروگلیف: معروفترین و قدیمی ترین نوشتن هیروگلیف مصریها است. برخی از این نوشته‌ها به 3 هزار سال پیش از میلاد است. خط هیروگلیف در 3000 سال زبان نوشتنی بوده است و آن را در دیوارها،
آرامگاه‌ها، مجسمه. به کار می برده اند. خط گاهی برای نوشتن بر روی پیپروس مورد استفاده قرار گرفت. خط هیروگلیف توسط دانشمند شامپولیون رمز گشایی شده است.

Incorrect form: رخ می دهد

Figure 1: The ETT text box

‘The Error Tags Box’ enables users to assign error tags in 3 layers to the selected error after suggesting a correct form for it. The error annotation is based on the SFLC Error Tagset. The first layer selects the error in the surface structure in four categories (addition, omission, substitution, permutation), the second layer selects the error domain in five groups (orthography, morphology, syntax, lexis, style), and the third layer the specific error, categorised as the error type, which is the biggest group, with 22 types of errors, namely: consonant character(s), long vowel character(s), short vowel character(s), connections, the Ezâfe article, dots, adjective, noun-plural, noun other, pronoun, preposition, postposition (râ), conjunction, verb agreement, verb tense, verb other, adverb, word order, word selection, phrase selection, cohesion, and unclear style. The tool provides the possibility of assigning more than one tag to the selected error.

When the errors have been selected, the ‘submit’ button will enter the tags into the ‘Error Phrase Box’ where all the errors are listed, and subsequently they will be saved in the corpus database. Figure 2 shows the ETT error tags box.



Text

قلعه بلگراد بر اساس ساخت که به بلگراد امروز است. این موزه تاریخچه بلگراد است. بنای فرهنگی و تاریخی واقع در محل تلاقی رودخانه ساوا و دانوب به وجود جنگ بزرگ جنیره است. در دوران نوسنگی اولین شهرهای در اینجا ساخته شد. در اوایل قلعه بیزانس از سنگ سفید
که در نور خورشید میدرخشید ساخته شده چرا که صربها شهر سفید نامیده شده. نام بلگراد برای اولین بار در سال 878 در تریخ 14 آوریل زکر شده بود. عثمانی کلید بلگراد 19 آوریل سال 1824 تماین به تست میخابلو.

Error tags

Incorrect form: شهرهای

Correct form: شهرها

Error in Surface Structure: Addition

Error Category: Syntax

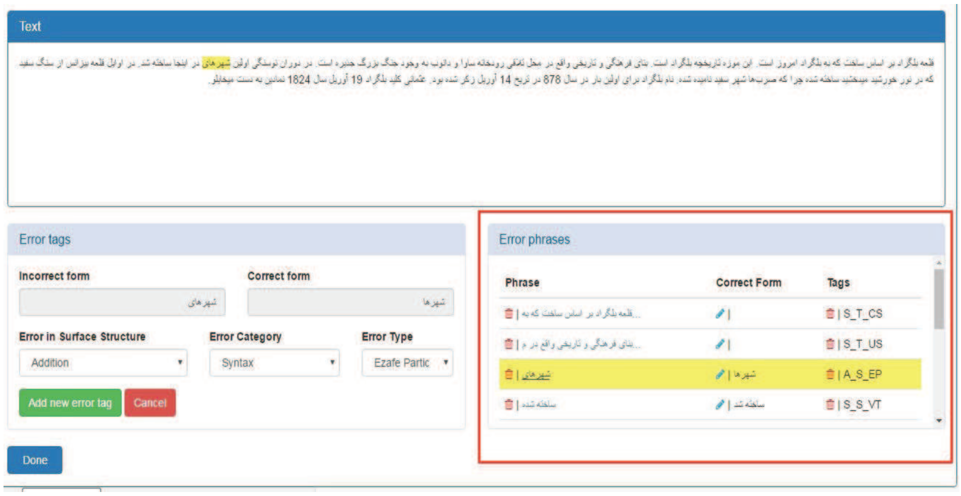
Error Type: Ezafe Partic

Error phrases

Phrase	Correct Form	Tags
| ...قلعه بلگراد بر اساس ساخت که به	|	| S_T_CS
| ...بنای فرهنگی و تاریخی واقع در م	|	| S_T_US
| شهرهای	| شهرها	| A_S_EP
| ساخته شده	| ساخته شد	| S_S_VT

Figure 2: The ETT error tags box

The annotated errors are listed in ‘The Error Phrase Box’. This box consists of three parts: (1) the ‘Phrase’ which copies the selected error segment (character(s), word(s), phrase(s), sentence(s) or text); (2) the ‘Correct Form’ which will be shown only if the correct form has been inserted into the Error Tags box - if not, it remains blank; and (3) the ‘Tags’, where the selected error tag codes are shown. It is possible to delete the error phrase or error tags or to edit the correct form in this box. Figure 3 shows the ETT error phrase box. The annotation process will be completed by the annotator pressing ‘Done’ at the bottom.



قلعه بلگراد بر اساس ساخت که به بلگراد امروز است. این موزه تاریخچه بلگراد است. بنای فرهنگی و تاریخی واقع در محل تلاقی رودخانه ساوا و دانوب به وجود جنگ بزرگ جنیره است. در دوران نوسنگی اولین شهرهای در اینجا ساخته شد. در اوایل قلعه بیزانس از سنگ سفید
که در نور خورشید میدخشید ساخته شده چرا که صرب‌ها شهر سفید نامیده شده. نام بلگراد برای اولین بار در سال 878 در تریخ 14 آوریل زکر شده بود. عثمانی کلید بلگراد 19 آوریل سال 1824 تماین به دست میخایلو.

Incorrect form: شهرهای

Correct form: شهرها

Error in Surface Structure: Addition

Error Category: Syntax

Error Type: Ezafe Partic

Phrase	Correct Form	Tags
قلعه بلگراد بر اساس ساخت که به...		S_T_CS
بنای فرهنگی و تاریخی واقع در م...		S_T_US
شهرهای	شهرها	A_S_EP
ساخته شده	ساخته شد	S_S_VT

Figure 3: The ETT error phrase box

7. The Frequency Distribution of Error Tags in the SFLC

The SFLC is designed and developed as an error-tagged learner corpus to investigate the frequency and types of linguistic errors made by Serbian learners of the Persian language. To achieve this aim, after developing the SFLC Error Tagset and setting up the corpus software and tools, the researcher carried out error annotation on 300 submitted documents. Using the Data Statistics Tool, the frequency distributions of errors are listed in accordance with the SFLC error taxonomy and the tagset. Based on the statistics, the 10 major error types of the Serbian learners of the Persian language in the SFLC are listed in Table 5. The table provides a clear view on the distribution of errors in the whole corpus since it is organized based on the error types, and then the error domain and the errors in the surface structure are listed in accordance to error types.

Table 5: The major error types in the SFLC

	Error Type	Error Domain
1	Consonant character(s)	Orthography
2	Long Vowel character(s)	
3	Word Order	Lexis
4	Verb Other	Syntax
5	Noun Other	Morphology
6	Preposition	Syntax
7	Word Selection	Lexis
8	Verb tense	Syntax
9	Pronoun	Morphology
10	Postposition (ra)	Syntax

The table statistic shows that the first 5 error types are the most frequent errors which account for 48%, or about half of the total error types in the SFLC. To review the statistics in detail and in order to gain a clear view of the learners' errors, they have been categorized into two major groups: (i) linguistics errors and (ii) orthographic errors based on the error domains. The SFLC linguistic error tags consist of 4 domains and 17 error tags as introduced in table 4. Based on the error tag distributions, the 5 major error types made by the Serbian learners of the Persian language are word order, verb other, noun other, preposition and word selection as illustrated in detail in Table 6.

Table 6: The 5 major error types made by the Serbian learners in the SFLC

	Error Type	Error Domain(s)	Surface structure (s)
1	Word Order	Lexis , style	Permutation, Substitution
2	Verb (other)	Syntax, Lexis	Substitution, Omission, Addition
3	Noun (other)	Lexis, Morphology	Omission, Addition, Substitution
4	Preposition	Syntax, Lexis	Substitution, Omission, addition
5	Word Selection	Lexis	Substitution, Addition

Table 7 illustrates the orthographic errors in the SFLC. The first two frequent error types, i.e. consonant character and vowel character, were marked as high frequent errors in the whole corpus. These errors are mainly tagged for substitution and omission at surface structure annotation level.

Table 7: Errors in Orthography

	Error Domain	Error Type	Surface structures
1	Orthography	Consonant character(s)	Substitution, Omission, Addition, Permutation
2		Long Vowel character(s)	Omission, Addition, Substitution
3		Dots	Omission, Substitution, Addition,
4		Short Vowel Character(s)	Omission, Substitution, addition

The Persian script and writing system has certain specific characteristics which are completely new for the Serbian learners of the Persian language, therefore, the major errors could be expected to belong to orthography. Although such errors decreased within the proficiency levels, they are the most frequent in the whole corpus as well as at each level of proficiency.

8. Conclusion

Learner corpora can be considered as 'language learning data resources' which generally provide empirical data and useful information about the language learning process and language skills development. The SFLC, as an error-tagged learner corpus, attempts to show us a clear view of difficulties which Serbian learners face during learning the process of learning the Persian language. The SFLC, also provides data resources which are expected be useful and provide helpful data sources for the Persian teachers, textbook and language material writers, lexicographers and even learners themselves.

References

Atkins, S., Clear, J., Ostler, N. (1991). Corpus Design Criteria. *Literary & Linguistic Computing* 7 (1): 1-16.

Biber, D. (1993). Representativeness in Corpus design. *Literary and linguistics computing*. 8 (4): 243-45.

Biber, D., Corad, S. & Reppen, R. (1998). *Corpus Linguistics. Investigating Language Structure and Language Use*. (pp. 246-250). Cambridge: Cambridge University Press.

Díaz-Negrillo, A., & Fernández-Domínguez, J. (2006). Error tagging systems for learner corpora. *Spanish Journal of Applied Linguistics (RESLA)*, 19, 83-102.

Dulay, H. C., Burt, M. K. & Kreshen, S. (1982). *Language Two (p 150-160)*. New York: Oxford University Press.

Gilquin, G. (2015). From design to collection of learner corpora. In: S. Granger, G. Gilquin & F. Meunier, *The Cambridge Handbook of Learner Corpus Research*, (pp. 9-34). Cambridge University Press: Cambridge.

Granger, S. (1993). The International Corpus of Learner English. In J. Aarts, P. de Haan, & N. Oostdijk (Eds.), *English language corpora: Design, analysis and exploitation* (pp. 57–69). Amsterdam, the Netherlands: Rodopi.

Granger, S. (2003). Error-tagged Learner Corpora and CALL: A promising Synergy. *CALICO Journal*. 20 (3).

Safari, S. (2917). Constructing and Analysisng an Error-tagged Learner Corpus of Persian. *Doctoral Dissertation*: Faulty of Philology, University of Belgrade.

Tono, Y. (2003). Learner corpora: design, development and applications. In D. Archer, P. Rayson, A. Wilson, & T. McEnery (Eds.), *Presented at the Corpus Linguistics 2003 Conference (CL 2003)* (Vol. 16, pp. 800–809). Lancaster (UK): Lancaster University: University Centre for Computer Corpus Research on Language.

Saeed Safari

SALAM FARSI KORPUS ZA UČENJE – UVOD U SISTEM TAGOVANJA GREŠAKA

Rezime

Jezički korpusi predstavljaju pouzdan izvor i empirijsko sredstvo za analizu jezičkih podataka. Uveliko se koriste na polju usvajanja drugog/stranog jezika, kao i u istraživanjima vezanim za predavanje stranih jezika. Najzastupljeniji u toj sferi su takozvani korpusi za učenje. Ovaj rad ima za cilj da predstavi sistem za anotaciju i tagiranje prvog za greške tagovanog korpusa za učenje persijskog jezika, koji se zove *Salam farsi korpus za učenje*, kao i da analizira jezičke pogreške napravljene u kolekciji pisanih tekstova koji su napisani od strane srpskih učenika persijskog jezika

Ključne reči: korpus za učenje, persijski kao strani jezik za Srbe, korpusna lingvistika, analiza grešaka