UNIVERSITY OF BELGRADE
FACULTY OF PHILOLOGY

Saeed G. Safari

# CONSTRUCTING AND ANALYSING AN ERROR-TAGGED LEARNER CORPUS OF PERSIAN

Doctoral Dissertation

Belgrade, 2017

UNIVERZITET U BEOGRADU
FILOLOŠKI FAKULTET

Said G. Safari

# IZRADA I ANALIZA ANOTIRANOG KORPUSA PERSIJSKOG JEZIKA KAO STRANOG

Doktorska disertacija

Beograd, 2017.

УНИВЕРСИТЕТ БЕЛГРАДА
ФАКУЛЬТЕТ ФИЛОЛОГИИ

Саид Сафари

# Формирование и анализ аннотированного корпуса персидского языка

Докторская диссертация

Белград, 2017 г.

**Podaci o mentoru i članovima komisije**

**Mentor:**
**dr Maja Miličević Petrović**, vanredni profesor
Filološki fakultet, Beograd

**Članovi komisije:**

1. **dr Ljiljana Marković, redovni profesor**
   Filološki fakultet, Beograd

2. **dr Jelena Filipović, redovni profesor**
   Filološki fakultet, Beograd

3. **dr. Reza Morad Sahraei**
   Fakultet za persijsku književnost i strane jezike, Teheran
   (Faculty of Persian Literature and Foreign Languages, Allameh Tabataba'i University, Tehran)

Datum odbrane:
Beograd, _____

به نام خداوند جان آفرین

حکیم سخن در زبان آفرین

*I would like to express my sincere gratitude to my mentor, Dr Maja Miličević Petrović for the continuous support of my thesis research and her advice, comments, guidance and immense knowledge.*

*I would like to thank my esteemed professors, Dr Ljiljana Marković, Dr Julijana Vučo and Dr Jelena Filipović for their constant enthusiasm and encouragement during my doctoral studies.*

*I would also like to thank Reza Morad Sahraei, from Allameh Tabataba'i University in Tehran for reviewing my research and his valuable comments and feedback.*

*My deepest and endless gratitude goes to my amazing family, to whom this thesis is dedicated, especially to my loving and supportive wife, Solmaz Taghdimi.*

**CONSTRUCTING AND ANALYSING**
**AN ERROR-TAGGED LEARNER CORPUS OF PERSIAN**


**Summary**

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.

The present thesis, based on a methodological approach for building a learner corpus, is generally in line with the domain of error analysis and the field of Learner Corpus Research. The thesis describes the process of constructing and developing an 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. 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 are described. The SFLC software is equipped with four main tools in order to function as an error-tagged learner corpus and provide the statistical reports. These tools include a tool for submitting data and metadata into the corpus database, a computer-aided error editor to facilitate error tagging, filters and search, and data statistics tools which show various statistical data related to the corpus.

Based on the SFLC statistical reports, the frequency and error distribution in the whole corpus and the comparison of error distributions across different proficiency levels are discussed. The corpus statistics show that the most frequent errors made by the Serbian learners of the Persian language are initially to be found in the domain of orthography, while later on the most frequent errors lie in the domains of lexis and syntax. Word Order is marked as the most frequent error type in the corpus as a whole. As for the distribution of errors across specific proficiency levels, the results show that the total number of errors drops from level A2 to level C1, while errors in syntax increase, due to the use of more

complex syntactic structures at higher proficiency levels. The SFLC not only provides authentic data gathered from learners at different proficiency levels, but also statistics regarding error tags and metadata. Research into Persian as a second/foreign language thus can clearly benefit from the SFLC as a resource.

**Keywords**: Learner Corpus, Error Analysis, Second Language Acquisition, Teaching Persian as a Foreign Language.

Research area: Linguistics

Research subarea: Corpus linguistics, Second Language Acquisition

UDC number:

# IZRADA I ANALIZA ANOTIRANOG KORPUSA PERSIJSKOG JEZIKA KAO STRANOG


## Rezime

Lingvistički korpusi predstavljaju značajan izvor i sredstvo analize empirijskih jezičkih podataka. Njihova upotreba vrlo je raširena, između ostalog, u oblasti istraživanja usvajanja drugog/stranog jezika i nastavi jezika, gde posebno treba naglasiti značaj učeničkih korpusa. U ovoj disertaciji opisuje se izrada jednog takvog korpusa – učeničkog korpusa persijskog jezika, pod nazivom *Salam Farsi Learner Corpus* (SFLC). Ovaj korpus je izrađen na osnovu tekstova koje su tokom pohađanja kurseva persijskog jezika pisali učenici čiji maternji jezik je srpski. Pored toga što su tekstovi prebačeni u digitalni format, u korpusu su označene greške koje su učenici pravili prilikom pisanja.

Tri glavne faze u izradi korpusa bile su njegovo koncipiranje i digitalizovanje, predlaganje sistema anotacije grešaka i razvijanje alata za izradu i pretragu korpusa. Sve tri faze detaljno su opisane u disertaciji. Konkretno, pažnja je posvećena opisu praktičnih koraka poput prikupljanja podataka i metapodataka, kao i konceptualnih zadataka kakvi su definisanje kriterijuma za izradu korpusa, sastavljanje oznaka za greške i idejno osmišljavanje korpusnog interfejsa, softvera i alata. SFLC se softverski oslanja na četiri glavna alata koji omogućuju unos podataka i metapodataka u korpusnu bazu, označavanje grešaka, preuzimanje i pretragu dokumenata (prema površinskim oblicima reči ili prema greškama) i generisanje statističkih izveštaja o greškama.

Na osnovu statističkih izveštaja koje SFLC daje, u disertaciji se sprovodi i analiza grešaka – proučavaju se učestalost i raspodela grešaka u korpusu kao celini i na različitim pojedinačnim nivoima znanja persijskog jezika. Rezultati ove korpusno zasnovane analize pokazuju da učenici kojima je maternji jezik srpski na nižim nivoima znanja persijskog jezika najčešće prave greške u domenu ortografije, dok se kasnije greške češće nalaze u domenima leksike i sintakse. Greške vezane za red reči označene su kao ukupno gledano najčešći tip greške u čitavom korpusu. Ukupni broj grešaka smanjuje se kako se učenici kreću od nivoa A2 ka nivou C1. Međutim, kada je reč o sintaksi, broj grešaka raste, usled korišćenja složenijih sintaksičkih struktura na višim nivoima.

SFLC ne samo da obezbeđuje autentične podatke prikupljene od učenika na različitim nivoima znanja, već pruža i statističke podatke o označenim greškama i drugim korpusnim parametrima. Stoga se zaključuje da korpus može biti od velike koristi za istraživanje i nastavu persijskog jezika kao drugog/stranog.

**Ključne reči**: Učenički korpus, analiza grešaka, usvajanje drugog jezika, nastava persijskog kao stranog jezika.

Naučna oblast: Nauka o jeziku

Uža naučna oblast: Korpusna lingvistika, primenjena lingvistika

UDK broj:

**TABLE OF CONTENTS**

## LIST OF TABLES

**LISF OF FIGURES**

**APPENDIX**

# 1. Introduction

The Greek philosopher Heraclitus is famously quoted as saying: "The only thing that is constant is change". We live in an era of communication which is constantly dominated by the change, exchange and sharing of information. Although this era is named 'the information age', I believe it could also be called "the age of information classification", characterized by the use of computers in processing information and changing it into analyzable data.

Consequently, research in the field of language and linguistics has also been affected by the theme of "change" and "data classification". In Hunston & Francis' view (2001: 17), "language is not a system that is realised in actual instances, but a set of actual instances that may be regarded as construing an approximate and ever-changing system." Over the past few decades, Corpus Linguistics (CL) has changed the research models and methodologies in linguistics from theoretical to experimental investigations of the language. Therefore, applying corpus-based and corpus-driven approaches to the study of language is now inevitable and as Biber (2010: 159) indicates, leads to "results in research findings which have much greater generalizability and validity than would otherwise be feasible."

Today, corpus linguistics methodology is widely used in Second Language Acquisition (SLA) research and this field of study is equipped with corpora resources which are used for FL/SL (foreign language/second language) processing. Since the success of SLA research relies mainly on access to authentic data, applying 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 my research is to contribute to this effort.

## 1.1 Learner Corpora, Second Language Acquisition and Error Analysis

Linguistic corpora provide reliable sources and empirical means for analysing linguistic data. They are also widely used in the field of SLA and Foreign Language Teaching (FLT) research, where they are specifically referred to as learner corpora (LC). Granger pioneered the compilation of learner corpora and defined the term LC as "electronic collections of authentic FL/SL [foreign language/second language] textual data assembled according to explicit design criteria for a particular SLA/FLT purpose" (Granger, 2002:7). She also indicated that the main purpose in compiling a learner corpus is to gather objective data that can aid in describing learner language (ibid).

The field of Learner Corpora Research (LCR) is an emerging one and only dates back to the late 1980s (Granger: 2015), however, the number of learner corpora has noticeably increased in the past twenty years thus indicating a growing interest in corpus-based research in SLA. According to Granger (2008), using LC to analyse learner language can contribute to SLA research by providing a better description of interlanguage (i.e. transitional language produced by second or foreign language learners) and a deeper understanding of the factors that influence it. This in turn can be used to develop pedagogical tools and methods which more accurately target the needs of language learners.

In terms of the LC contribution to SLA, Gilquin *et al.* (2007) believe that LC can reveal learning problems related to various linguistic features such as, orthographic, lexical, grammatical, phraseological, stylistic, and pragmatic, as well as identify certain patterns of overuse, underuse and misuse through the application of a wide range of linguistic annotations (e.g. morphosyntactic tagging, discourse tagging, and error tagging). Similarly, Dash (2003) acknowledges that the systematic analysis of the data stored in LC provides authentic evidence of the linguistic efficiencies learners have acquired as well as the deficiencies they carry in the process of learning.

Granger (2002) identified two methodologies through which LC are studied: Contrastive Interlanguage Analysis (CIA) and Computer-aided Error Analysis (CEA). She termed such approaches "linguistic exploitations of learner corpus". In her view, CIA involves quantitative and qualitative comparisons between native language and learner language (L1 vs. L2) and also between different varieties of interlanguage (L2 vs. L2) (Granger, 2009). The International Corpus of Learner English (ICLE) (Granger *et al*. 2002, Granger, 2003) is an example of a corpus used by researchers for such comparisons.

On the other hand, CEA focuses on learners' errors and uses computer tools to tag, retrieve and analyse them (ibid). To this aim, one frequent type of LC, known as an **Error-tagged Learner Corpus** (ETLC), is used to identify and analyse learners' errors. ETLC can provide SLA/FLT researchers, educators as well as language materials developers with a valuable data resource in the field of Error Analysis (EA). In addition, such corpora also serve as a useful resource to determine the types and frequency of errors and to measure the extent to which learners can improve their performance in various aspects of the target language (Buttery & Caines, 2012; Nesselhauf, 2004). Analysing learners' errors via ETLC may function as the basis for both pedagogical purposes and the development of learning materials. For instance, the editions of the Longman Dictionary of Contemporary English (LDOCE) (2003) and the Cambridge Advanced Learner's Dictionary (CALD) (2003) both contain error notes based on their respective learner corpora, which are intended to help learners avoid making common mistakes (Granger, 2008). The Longman Dictionary of Common Errors (Turton & Heaton, 1996) is another example of a learner-corpus-informed dictionary.

## 1.2 Overarching Goals and Motivation

The process of learning a foreign language is to a certain degree one of making errors, correcting them and thus improving acquisition. In order to study language learning errors, a systematic procedure which attempts to collect, identify, describe, and evaluate the errors made is needed. LC in general and ETLC in particular can be used to follow such a procedure. Discovering and analysing errors through ETLC enables researchers and educators to gain a deeper understanding of the general trend of language learning, the

interlanguage, and the frequency and types of linguistic errors so as to discover the weaknesses and strengths of the language learning process.

The lack of a specific error-tagged learner corpus for learning the Persian language as a second/foreign language was the main inspiration of the present thesis. To this end, the written essays and texts from Serbian learners of the Persian language were compiled as the raw materials for the corpus. The two main aims of the thesis can be outlined as follows:

1. To construct and develop an error-tagged Persian learner corpus including a system for error annotation to be used as a new source for research in the field of learning the Persian language as a foreign language.
2. To investigate the frequency distribution and types of lexical, grammatical and orthographical errors made by Serbian learners of the Persian language. In other words, based on the error taxonomy of the corpus, (i) to find out what the frequent error categories are (ii) to determine the types of errors which are high/low in frequency.

In the present thesis, the main focus will be on designing and building an error-tagged learner corpus of Persian, the **Salam Farsi Learner Corpus (SFLC)**, and as is expected of such a corpus, to provide authentic, empirical data for subsequent analysis. It should be noted that since the thesis research is set to design and construct the very first ETLC of the Persian language and to detect and analyze learners' errors, the research is generally in line with the domains of SLA and EA.

## 1.3 Specific Objectives and Thesis Research Methodology

To achieve the goals in developing the error-tagged learner corpus of Persian, a number of specific objectives are defined as following:

1. To review some well-known learner corpora in order to develop the corpus design criteria. Such corpora will be reviewed based on 10 criteria (corpus purpose, size, target language, availability, learners' nativeness, learners' level of proficiency, learners' first language, materials genre, task type, and data annotation) so as to set the theoretical and structural basis of the research.

2. To develop an error annotation schema for creating the SFLC error tagset. This includes introducing a specific error taxonomy based on the theoretical model for error categorization by Dulay *et.al* (1982). The annotation schema will lead to the creation of a unique tagset for the SFLC.
3. To develop a corpus interface and tools. The corpus needs some specific tools for entering, saving, tagging, filtering and searching data.
4. To discover and analyze the error frequency distributions of Serbian learners of the Persian language based on the corpus data.

## 1.4 Thesis Research Methodology

The methodology for building an ETLC of Persian consists of three major stages: (i) constructing the corpus, (ii) proposing a system for error annotation and (iii) developing tools and software. In the first stage, for constructing the corpus and in line with the purpose of the SFLC, which is to identify learners' errors, the proposed design criteria by Tono (2003) are adopted. Based on these criteria, three major types of features are identified as follows: (a) language-related criteria (mode, medium, genre, topic), (b) task-related criteria (cross-sectional; prepared), and (c) learner-related criteria (age, gender, mother tongue).

Later, the data collection process is explained whereby the data is collected from Serbian learners, who studied the Persian language at the Iranian Cultural Center (ICC) in Belgrade in the academic years 2012 – 2015. The texts consist of excerpts from their creative writing (free writing) homework assignments at intermediate and advanced levels. Almost 700 authentic written texts were collected as well as information regarding gender, age, topic and level, which are used as metadata for the corpus. Before transcribing the texts into electronic format, some standards for the transcription process are defined due to the wide variety of writing styles in Persian, which allows writing words in different ways in terms of joining the letters. The next phase is the manual transcription of the raw texts into electronic content following the proposed transcription standards to ensure that they are subsequently able to handle error tags. Finally, the database set up for storing and managing the content is explained.

The second stage proposes a system for error annotation and consists of two phases: (i) developing an error annotation scheme for the SFLC; the main focus in this phase is on developing an error taxonomy to identify and detect categories and types of the most frequent errors to meet the needs of the SFLC and (ii) creating a tagset for SFLC error annotation; based on the corpus error annotation schema, a specific tagset is created for annotating the errors and is specifically called the Salam Farsi Error Tagset (SFET).

The third stage is devoted to developing the corpus interface and the software tools. It includes creating some tools for submitting data and metadata, a computer-aided error tagging tool, with a smart-selection function and tools for searching and the statistics. Table 1.1 summarises these three main stages in developing the SFLC and links each phase to the thesis chapters.

Table 1.1: The stages and phases of developing the SFLC with links to the thesis chapters

| Stage | Phases | Chapter |
|---|---|---|
| 1. Constructing the Corpus | Defining the corpus design criteria | 3 |
| | Data collection | |
| | Defining standards for the transcription process | |
| | Transcribing the texts | |
| | Setting up the SFLC database | |
| | | |
| 2. Proposing a System for Error Annotation | Developing an error annotation schema for the SFLC | 4 |
| | Creating a Tagset for Error Annotation of the SFLC | |
| | | |
| 3. Developing Corpus Tools | Setting up the corpus webpages and interface. | 5 |
| | Setting up corpus tools. | |

## 1.5 Outline of the Thesis

The thesis is organized into 7 chapters as follows:

**Chapter 1** provides an introduction to the thesis and defines the terms learner corpus, error analysis and error-tagged learner corpus which are the key terms in this thesis, as well as the connections between LC and SLA and EA. This is followed by a description of the goals and motivation behind the thesis.. The methodology adopted for the construction of the error-tagged leaner corpus of Persian is also explained in detail so as to provide a clear view of the project.

**Chapter 2** reviews the literature and consists of two sections. The first section provides the background to corpus linguistics and the application of corpora in SLA, followed by an explanation of the types of learner corpora. It also gives a review of 10 well-known leaner corpora under 10 categories to derive the design criteria for developing the learner corpus for the Persian language. The second section offers a brief description of Persian and its main characteristics, and the chapter ends with an overview of existing corpora and tools for Persian.

**Chapter 3** describes both the contents of the SFLC and how the data and metadata were collected. The transcription process is explained in detail, and the design criteria on which the corpus development is based are also discussed. This is followed by a description of the database design which is used to manage the SFLC data and automate the generation of corpus files in different formats.

**Chapter 4** provides an introduction to error analysis, error categories and types. It describes the annotation schema for the SFLC and the design of Error Tagset for Persian which is based on the annotation schema of the corpus. The chapter concludes with a discussion of the Error Tagging Manual.

**Chapter 5** describes the SFLC webpages and tools. The chapter introduces the SFLC webpages and the corpus interface and then explains the development of the four types of different tools and their specifications. and the information about the corpus applications.

**Chapter 6** highlights the possibilities of the use of the SFLC and provides data analysis. Based on the corpus data, the error frequency distribution and error types of Serbian learners are described in this chapter.

**Chapter 7** summarises the main contributions of the thesis and ends with a summary of the possible uses of the SFLC.

# 2. Review of the Literature

**Chapter Summary**

This chapter provides an overview of the learner corpora research domain. More specifically, it starts with the definitions of corpora and corpus linguistics and discusses types of corpora, followed by the background to learner corpora contribution in second language acquisition, and their types, development and applications are introduced. The chapter also takes a look at learner corpora around the world and reviews 10 projects related to the corpora design criteria. It also gives a brief description of the Persian language and its main morphological, syntactic and orthographic features.

## 2.1 Corpora and Corpus Linguistics

The term *corpus* derived from Latin, literally means "body" and traditionally refers to "a text collection" or "an archive". According to Casas-Pedrosa *et al*. (2013) and based on the *Oxford English Dictionary* (s.v. *corpus*), the first recorded written example of the word corpus, understood as "the body of written or spoken material upon which a linguistic analysis is based" appears in the following excerpt by Allen (1956: 128): "The analysis presented here is based on the speech of a single informant (…) and in particular upon a *corpus* of material, a large proportion of which was narrative, derived from approximately 100 hours of listening". Regarding a commonly agreed definition for the term, Flowerdew (2012:3) points to some leading researchers in the field of corpus linguistics (e.g. Sinclair 1991; Stubbs 1996; Biber *et al*. 1998; Hunston 2002), who all view a corpus as a collection of authentic language, either written or spoken, which has been compiled for a particular purpose. McEnery *et al.* (2006: 5) consider some notable specifications which limit the term corpus to the collection of (1) machine-readable (2) authentic texts (including transcripts of spoken data) which are (3) sampled to be (4) representative of a particular language or language variety.

For Sinclair (2005), "corpus design criteria" are a key issue; therefore in his definition a corpus is "a collection of pieces of a language text in an electronic form, selected according to external "criteria" to represent, as far as possible, a language or a language variety as a source of data for linguistic research". Finally, the definition offered by McEnery *et al.* (2006:4) seems to cover most of the aspects of the term *corpus*: "In modern linguistics, it [*corpus*] can best be defined as a collection of sampled texts, written or spoken, in machine-readable form which may be annotated with various forms of linguistic information."

The computer-aided analysis of large databases of text, first used in linguistic research in the late 1950s, led to the emergence of a new field of study that was later called "Corpus Linguistics" (henceforth abbreviated to CL). The term itself first appeared in the early 1980s (Leech, 1992: 105), however, according to Biber & Finegan (1991:207) "the early use of corpus methodologies in modern linguistic research dates back to the pre-

Chomskyan period when it was used by field linguists such as Boas and linguists of the structuralist tradition, including Sapir, Newman, Bloomfield and Pike."

As Le´on (2005: 35) notes, "what is called 'corpus linguistics' covers various heterogeneous fields ranging from lexicography, descriptive linguistics, and applied linguistics – language teaching or Natural Language Processing – to domains where corpora are needed because introspection cannot be used, such as studies of language variation, dialect, register and style, or diachronic studies". As an evolving and fast-growing field of study and due to its multidimensional nature, finding an explicit definition for CL is not easy, hence the wide spectrum of definitions. Taylor (2008) discusses this issue in more detail by making a corpus on corpus linguistics and finally concludes that the term "corpus linguistics" has been defined variously, as a tool, a method, a methodology, a methodological approach, a discipline, a theory, a theoretical approach, a paradigm (theoretical or methodological), or a combination of these. Reviewing some well-known linguists' standpoints regarding the term, we notice such varieties in their definitions. Leech (1992: 106), for instance, argued that "computer corpus linguistics defines not just a newly emerging methodology for studying language, but a new research enterprise, and in fact a new philosophical approach to the subject". Kennedy (1998: 1) defines the term as a new scholarly enterprise which emerged in the last three decades of the twentieth century and can serve as a basis for linguistic analysis and description. Tognini-Bonelli (2001:1) describes CL as "a pre-application methodology which possesses theoretical status". Granger (2002: 4) believes that CL is a linguistic methodology "which is neither a new branch of linguistics nor a new theory of language, but the very nature of the evidence it uses makes it a particularly powerful methodology, one which has the potential to change perspectives on language." Meyer (2002) also defines the term as a methodological principle in linguistic research, while Teubert (2005: 2) describes it as "a theoretical approach to the study of language" and Thompson & Hunston (2006: 8) state that "corpus linguistics is a methodology that can be aligned with any theoretical approach to language". For McEnery, Xiao and Tono (2006: 7), "[a]s corpus linguistics is a whole system of methods and principles of how to apply corpora in language studies and teaching/learning, it certainly has a theoretical status. Yet the theoretical status is not a theory in itself" and

finally they conclude that "corpus linguistics is a methodology" (ibid). McCarthy & O'Keeffe (2010) consider it an attempt to make sense of linguistic phenomena in large texts or collections of smaller texts.

Although linguists are unable to agree on one explicit definition of corpus linguistics, there are some descriptions for the term which define the general realm of the field. Biber *et al.* (1998: 4) describe corpus linguistics as having four main features: 1) it is an empirical approach in which patterns of language use that are observed in real language texts (spoken and written) are analysed, 2) it uses a representative sample of the target language stored as an electronic database (a corpus) as the basis for the analysis, 3) it relies on computer software to count linguistics patterns as part of the analysis, and 4) it depends on both quantitative and qualitative analytical techniques to interpret the findings. Dash (2010:1) points to the idea that "the uniqueness of corpus linguistics lies in its way of using modern computer technology in the collection of language data, methods used in processing language databases, techniques used in language data and information retrieval, and strategies used in the application of these in all kinds of language-related research and development activities".

In general and based on the definition for the term *corpus*, it can be concluded that corpus linguistics covers a large amount of machine-readable data of actual language use which includes the collections of literary and non-literary text samples to reflect on both the synchronic and diachronic aspects of a language. Corpus linguistics, which aims to investigate the language and all its properties by analysing large collections of text samples, has opened up new horizons in the scientific study of language by providing authentic and objective samples of language.

### 2.1.1 Types of Corpora

As corpus linguistics methodology has become more popular in linguistic research, many types of corpora have been developed to serve different types of research. During the last three decades thousands of corpora have been built, most of which are not publically available and are designed and created for specific research projects. Such corpora are of varying sizes with different contents and purposes.

Corpora can be classified according to different features, for instance, Atkins *et al.* (1991) focused mainly on the contrastive parameters of "corpora content" and introduced eight groups of corpora: (1) Full-Text, Sample, Monitor (2) Synchronic, Diachronic (3) General, Terminological (4) Monolingual, Bilingual, Plurilingual (5) Single, Parallel (6) General, Shell, (7) Core, Periphery and (8) Languages of Corpus.

According to Kennedy (1998:19), corpora can be simply classified into three main groups based on (1) use and application, (2) mode and (3) size. In terms of 'use and application', corpora can be classified into two major categories: general corpora and specialized corpora. General corpora refer to a collection of texts for unspecified linguistic research and are sometimes referred to as 'core corpora', which can be used as the basis for comparative studies (ibid). On the other hand, specialized corpora are designed for particular research projects, such as 'dialect corpora', 'regional corpora', 'test corpora' and 'learner corpora'. As for 'mode', corpora may be classified into 'written' or 'spoken' (ibid).

Some researchers tried to make distinct classifications for the available corpora in terms of their usage and features. Xiao (2008) introduced a classification for corpora of what he calls 'well-known and influential corpora'. Such corpora are grouped into eleven categories according to their primary uses and supported with examples of corpora for each group in the following way:

(1) National Corpora, such as the British National Corpus (BNC), the American National Corpus (ANC), the Czech National Corpus (CNC), the Russian National Corpus (RNC), etc.;
(2) Monitor Corpora, such as the Bank of English (BoE) and the Global English Monitor Corpus;
(3) The Brown Family of Corpora, such as the LOB (the Lancaster/Oslo-Bergen corpus of British English), the WWC (the Wellington Corpus of Written New Zealand English) and the Kolhapur Corpus (the Kolhapur Corpus of Indian English);
(4) Synchronic Corpora, for comparing language varieties, such as the International Corpus of English (ICE) and the Longman/Lancaster Corpus;
(5) Diachronic Corpora, containing texts from the same language gathered from different time periods, used to track changes in language evolution, such as the

Helsinki Corpus of English Texts and the Lampeter Corpus of Early Modern English Tracts;

(6) Spoken Corpora, such as the London-Lund Corpus (LLC), the Lancaster/IBM Spoken English Corpus (SEC), the Bergen Corpus of London Teenage Language (COLT) and the Switchboard Corpus (SWB);

(7) Academic and Professional English Corpora, such as the British Academic Spoken English (BASE) corpus, the Michigan Corpus of Academic Spoken English (MICASE) and The Corpus of Spoken Professional American English (CSPAE);

(8) Parsed Corpora, also called treebanks, such as the Lancaster-Leeds Treebank, the Lancaster Parsed Corpus (LPC) and the British component of the International Corpus of English (ICE-GB);

(9) Developmental and Leaner Corpora, such as the Child Language Data Exchange System (CHILDES) , the Louvain Corpus of Native English Essays (LOCNESS) and the International Corpus of Learner English ( ICLE);

(10) Multilingual or Parallel Corpora, such as the Canadian Hansard Corpus, the English-Norwegian Parallel Corpus (ENPC), the Oslo Multilingual Corpus (OMC) and the European Corpus Initiative Multilingual Corpus I (ECI/MCI);

(11) Monolingual Corpora, such as COSMAS (the Corpus Search, Management and Analysis System), the Institute for Dutch Lexicology Corpus (INL), and the Scottish Corpus of Texts and Speech (SCOTS).

Finally, Serena (2012) suggests a comprehensive classification of corpora and believes it should be based on fundamental features such as: Size (small, medium, large), the Number of Text Languages (monolingual, bilingual, multilingual), Mode (spoken, written, mixed), the Nature of Data (general, specialized), the Nature of Application (research, illustrative, learner, translation, aligned comparable, parallel, reference), Dynamism (monitor, static), Temporal Characteristic (diachronic, synchronic), Authorship (one author or more), Annotation (non-annotated, annotated, morphological, semantic, syntactic, prosodic, etc.) and Access (free, commercial, closed).

### 2.1.2 Types of Corpora in Language Learning and Teaching

In the field of language learning and teaching, numerous corpora have been developed and different types of corpora introduced. Bennett (2010: 13) believes that the most useful types of corpora in the field of language teaching and learning can be limited to (1) Generalised, (2) Specialised, (3) Leaner, and (4) Pedagogical corpora. According to Bennett (ibid), generalised corpora, which are usually very large and consist of different types of texts, can give the user a broader picture of a language, and can therefore be consulted by language learners. 'Specialised corpora', on the other hand, contain texts of a certain type and aim to be representative of the language of this type. Such corpora are often used in Language for Specific Purposes (LSP) settings. Coxhead (2002), for example, developed the Academic Word List (AWL), which is a corpus-derived wordlist used as an important tool in learning and teaching EAP (English for Academic Purposes).

'Learner corpora' are a type of specialised corpora containing written texts and/or spoken transcripts of the language used by students who are currently acquiring a given language. They are often tagged and can be examined, for example, to find the common errors learners make. The International Corpus of Leaner English (ICLE) (Granger, 2003) is a well-known learner corpus and contains 14 different native languages. And finally, 'pedagogical corpora' are those which contain the language used in classroom settings. Such corpora may include academic textbooks, transcripts of classroom interactions, or any other written text or spoken transcripts encountered by learners in an educational setting. (Bennett, 2010:14).

Xiao (2008: 426) indicates that "two types of corpora are particularly relevant to language learning: developmental corpora and learner corpora." Developmental corpora consist of data produced by children acquiring their first language (L1) and the most well-known corpus in this category is The Child Language Data Exchange System (CHILDES) which is a large corpus of child language and child-directed speech (MacWhinney, 2000, 2007). On the other hand, learner corpora deal with acquiring a second language (L2) and are in fact a collection of the writing or speech produced by learners. This will be thoroughly discussed in the following section.

## 2.2 Learner Corpora

Learner corpora are categorized as "specialized corpora" (Bennett, 2010: 14) and according to Granger (2008), have all the characteristics commonly attributed to corpora. Therefore, like any other corpora, a learner corpus can be defined generally as a "collection of machine-readable authentic texts (including transcripts of spoken data) which is sampled to be representative of a particular language or language variety" (McEnery *et al.*, 2006: 5). According to Granger *et al.* (2015: 9), what makes the learner corpus special is that "it represents the language as produced by foreign or second language (L2) learners and what makes it different from the data used in earlier SLA studies is that it seeks to be representative of this language variety."

Pravec (2002: 81) defines the learner corpus as "a computerized textual database of the language produced by foreign language learners". She also indicates that such a learner language database could be a very useful resource for discovering how languages are learned and improving the learning process (ibid).

Nesselhauf (2004: 125) believes that learner corpora are "systematic computerized collections of texts produced by language learners." She explains that by 'systematic', she means "the texts included in the corpus were selected on the basis of a number of criteria (e.g. learner level(s), learners' L1(s) [mother tongue(s)]) and that the selection is representative and balanced" (Nesselhauf 2004: 127). Callies & Paquot (2015:1) also indicate the systematic collection of data for the learner corpora and define LC as "systematic collections of authentic, continuous and contextualized language use (spoken or written) by L2 learners stored in electronic format which are a special type of empirical data used by scholars in a variety of disciplines."

On the other hand, in her definition of learner corpora, Granger (2008a: 338) places emphasis on "a degree of naturalness" and defines them as "electronic collections of natural or near-natural data produced by foreign or second language learners and assembled according to explicit design criteria." She indicates that such corpora may include texts that do not naturally occur because for learners (especially foreign language learners) the target language fulfils only a limited number of functions, most of which are restricted to the

classroom context. And finally, Pravec (2002: 81) indicates the importance of LC which provides "a deviation from the standard", so that through the investigation of such data researchers are able to focus on theoretical and/or pedagogical issues, while educators can concentrate on the needs of learners. Granger (1998:6) also asserts that the main purpose in compiling a learner corpus is to gather objective data that can aid in describing learner language or learners' total interlanguage.

### 2.2.1 Learner Corpus Research

Studies in the field of learner corpus research have increased over the past two decades, numerous learner corpora have been developed and researchers have started to show a greater interest in using corpus linguistics tools and methodology in their research, especially in the field of second and foreign language learning (L2). All these developments have resulted in the emergence of a new field of study called Learner Corpus Research (LCR). Granger, one of the pioneers of LCR and the initiator of the International Corpus of Learner English (ICLE), which was developed in 1990, predicted the emergence of such a new field of study:

"I have no doubt that the investigation of computerized learner corpora may well be able to achieve the kind of spectacular results which we have witnessed in lexicography, opening up new avenues of research and giving rise to a new generation of grammars, dictionaries, vocabulary books and language software programs developed with the difficulties of the learner in mind." (Granger 1994: 29)

According to Callies & Paquot (2015: 1), "Learner Corpus Research emerged as a field at the turn of the 1990s from the developing field of Corpus Linguistics when academics and publishing houses, simultaneously but independently, started to realize the considerable potential of large computerized datasets of learner production to describe learner language and/or develop new pedagogical tools and methods that target language learners' specific needs." Granger *et al.* (2015:3) also indicate that "LCR has become a truly interdisciplinary field at the crossroads between corpus linguistics, second language acquisition, language teaching and natural language processing." Granger (2009: 15) introduced the core components of LCR and represented them in a model. This proposed

model shows the direct interaction between LCR and the research domains in 'corpus linguistics', 'foreign language teaching', 'linguistic theory' and 'second language acquisition'. It could be concluded that those wishing to work and carry out research in this interdisciplinary field should be equipped with the knowledge of SLA principles and linguistic theories, and have expertise in corpus linguistics methodologies and a good understanding of teaching foreign languages.

Although LCR is still in its early stages, the increasing number of different types of learner corpora and research studies in this field, such as those by Granger (1998, 2002, 2003, 2004, 2007, 2008, 2012), Pravec (2002), Tono (2003), Nesselhauf (2004), Granger *et al.* (2013), Myles (2008), Díaz-Negrillo & Thompson (2013) and Granger *et al*. (2015), all highlight the important role of learner corpora in providing a valuable data resource, thus indicating a growing interest in this field.

It should be noted that although research on leaner corpora was initially restricted to the English language, it is now being undertaken in many different languages, creating a diverse and rapidly expanding international network of researchers. According to Callies & Paquot (2015), this is evidenced by the number and variety of learner corpus compilation projects listed on the 'Learner Corpora around the World' webpage maintained by the *Centre for English Corpus Linguistics* (CECL) at Louvain-la-Neuve, Belgium. This database[1] currently (in October 2017) contains 159 learner corpora, with 93 (58%) representing L2 English and the rest focusing on other languages (Arabic, French, German, Korean, Spanish, etc.). In a similar study, Alfaifi (2015) found and consequently reviewed 159 learner corpora around the world. The exact number of learner corpora, however, is not known.

The rapid developments and growing interest in LCR have also resulted in some academic products. Following the publishing of many academic papers and textbooks, now LCR has its very first handbook, *The Cambridge Handbook of Learner Corpus Research* (Granger *et al.,* 2015), an international peer-reviewed journal dedicated solely to LCR, the *International Journal of Learner Corpus Research* (IJLCR), and its own international

[1] https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

academic association, the Learner Corpus Association, officially founded in 2013, which coordinates the organization of a biennial international conference in the field of LCR.

## 2.3 Types of Learner Corpora

Different types of learner corpora are being created for different purposes and usages around the world. The varieties in developing learner corpora provide the possibility for drawing up different classifications. In this thesis, in order to establish the main categories of learner corpora, I only review the different LC typologies presented by Granger (2008), Hana *et al.* (2010) and Diaz-Negrillo & Thompson (2013). Granger introduced a typology based on comparative descriptions, while Hana *et al.* and Diaz-Negrillo & Thompson tried to classify the learner corpora in terms of the main features and the corpus design criteria.

### 2.3.1 Types of LC Based on Comparative Descriptions

In her comparative descriptions, Granger (2008) introduces 7 types of learner corpora:

1. **Commercial vs. Academic**

Commercial learner corpora which are compiled in educational settings are mainly initiated by major publishing companies and academic learner corpora compilations. The *Longman Learners' Corpus (LLC)* and the *Cambridge Learner Corpus (CLC)* are two major commercial learner corpora (Tono, 2009: 184). Academic corpora can come in all shapes and sizes and usually cover learners from only one mother tongue background. The *International Corpus of Learner English* (Granger 2003d) is an example of one.

2. **Big vs. Small**

The size of the corpora is a matter of consideration as it is a major asset in terms of the representativeness of the data and the generalizability of the results; however, small corpora are also of considerable value. According to Flowerdew (2004), there is no ideal size for a corpus; it all depends on what the corpus contains and what is being investigated. As pointed out by Ragan (2001: 211), "the size of the sample is less important than the

preparation and language product and its subsequent corpus application to draw attention to an individual or group profile of learner language use".

**3. English vs. Non-English**

This is the simplest way to classify learner corpora as English clearly dominates the learner corpus scene and numerous projects are developed for English. Some well-known English learner corpora are the International Corpus of Learner English (Granger, 2003d), the Longman Learners' Corpus (Gillard & Gadsby, 1998) and the Hong Kong University of Science and Technology Learner Corpus (Milton, 1998).The languages covered include Arabic (Aflifi & Atwell, 2013), Swedish (Hammarberg, 1999), Norwegian (Tenfjord *et al*., 2004), Dutch (Degand & Perrez, 2004), Spanish (Ife, 2004) and German (Lüdeling *et al.,* 2005).

**4. Writing vs. Speech**

Due to the difficulty of collecting and transcribing learners' speech, written learner corpora are more frequent than spoken learner corpora. Based on the CECL database, out of a total of 149 leaner corpora registered, 102 corpora are written (68%), 11 written and spoken (7%) and only 37 corpora spoken (25%). The difficulty is compounded in the case of multimedia learner corpora, which contain learners' texts linked to audio-video recordings (Reder *et al*., 2003).

**5. Longitudinal vs. Cross-sectional**

Longitudinal corpora, where data from the same learners are collected over time, are few and far between, therefore, the overwhelming majority of learner corpora covering more than one type of interlanguage data are cross-sectional, i.e. they contain data gathered from different categories of learners at a single point in time. For this reason, researchers interested in the trend of interlanguage development tend to collect quasi-longitudinal corpora, i.e. corpora gathered at a single point in time but from learners of different proficiency levels (Granger, 2008).

**6. Immediate vs. Delayed Pedagogical Use**

According to Granger (2008), learner corpora for immediate use refer to the idea that learners are at the same time both the producers and users of corpus data. On the other hand, delayed pedagogical corpora are not used directly as teaching/learning materials by

the learners who have produced the data. Such corpora are compiled with a view to providing a better description of one specific interlanguage and/or designing tailor-made pedagogical tools which will benefit similar-type learners, i.e. learners with the same profile as the students who have produced the corpus data (the same mother tongue background, the same level of proficiency, etc.).

7. **Raw vs. Annotated**

Another notable classification is based on LC annotation. Granger (2008) defines "raw learner corpora" as those which contain learner texts with no added linguistic annotation, while "annotated learner corpora" contain extra layers of information which can be counted, sorted and compared. She cites "grammar annotation" and "error annotation" as the two main types of annotation for learner corpora (ibid).

### 2.3.2 Types of LC based on Corpus Features and Design Criteria

Hana *et al.* (2010) classify learner corpora mainly in terms of corpus features. They introduced 5 main features which make up different types of learner corpora:

1. Target language (TL): learner corpora which cover the language of learners of a second or foreign language.
2. Medium: learner corpora can capture written or spoken texts, the latter being much harder to compile, and thus less common.
3. L1: the data can come from learners with the same L1 or with various L1s.
4. Proficiency in TL: some corpora gather texts from students at the same level, while others include texts of speakers at various levels. Most corpora focus on advanced students.
5. Annotation: many learner corpora contain only raw data, possibly with emendations, without linguistic annotation; some include part-of-speech (POS) tagging. Several include error tagging. Despite the time-consuming manual effort involved, the number of error-tagged learner corpora is growing.

Another classification of LC based on corpus features was introduced by Diaz-Negrillo & Thompson (2013). They introduced 6 main features, which could result in the construction of different types of learner corpora. The proposed features are as follows:

(1) Mode (spoken/written LC): (2) Annotation (unannoatated/annotated LC), (3) Language (multilingual/monolingual LC), (4) Data Collection Conditions (controlled/uncontrolled LC), (5) Time (longitudinal/cross-sectional LC), (6) Breadth (general/specialised LC).

Considering annotation as an important criterion in classifying linguistic corpora, an error-tagged learner corpus (ETLC) could be regarded as a type of learner corpora which is annotated by learners' errors and consequently designed and built to detect them. McEnery & Hardie (2012: 83) assert that error tagging was a development in learner corpus research strongly advocated by Granger. Granger (2003:10) emphasizes the potential benefits of error tagging, noting that 'once the corpus is error-tagged, the return on investment is huge'.

The error-tagged learner corpus can also be subjected to computer-aided error analysis (CEA), which is not restricted to errors seen as a deficiency, but which are understood as a means to explore the target language and to test hypotheses about the functioning of L2 grammar (Štindlová *et al*., 2010). The notion of error-tagging, error taxonomies, error types and error annotation will be discussed in depth in chapter 4.

## 2.4 Learner Corpora and SLA Research

Research on second language acquisition (SLA) is somewhat new and although this field of study is active and growing, according to Long (2012:35), it is still a "young science". SLA research dates back to the mid-1960s and the vast majority of work has been completed since the 1980s which implies an immaturity in SLA studies in many ways such as collecting, organizing and analysing reliable data about the language learning process and proposing theories based on them (ibid).

As Granger (2002:5) argues, "much current SLA research favours experimental, metalinguistic and introspective data, and tends to be dismissive of natural language use data". Therefore, theoretical views, experimental data and the interpretations based on them could result in theory proliferation which is a big problem nowadays in the field of SLA. By some counts, there are as many as 40-60 "theories" of SLA, or at least, theories in SLA. Most SLA researchers, like scientists everywhere, are rationalists of one kind or another,

not relativists, and from that perspective, theory proliferation is one of the chief obstacles to progress (Long, 2012).

The central source of evidence for all of these SLA theories is what Myles (2005:374) describes as "the language produced by learners", whether spontaneously or through data elicitation procedures. Therefore, both the authenticity of L2 theories and the success of research rely on the validity and reliability of data elicitation and data collection procedures.

Ellis (1994: 38) cites two goals for SLA research: "description and explanation" i.e. to describe learner's linguistic or communicative competence and then try to explain how and by which means learners acquire and develop a second language. To achieve these goals, he asserts that the researcher must examine the learner's usage of the L2 in actual performance, through the collection and analysis of samples of learner language (ibid).

According to Granger (2008:337), "analysing learner language is the key component of second and foreign language research and serves two main purposes: it helps researchers better understand the process of SLA and the factors that influence it, and it is a useful source of data for practitioners who are keen to design teaching and learning tools that target learners' attested difficulties."

Housen (2002: 78) remarks that "computer-aided language learner corpus research provides a much needed quantificational basis" for current SLA hypotheses and makes it possible to "empirically validate previous research findings obtained from smaller transcripts, as well as to test explanatory hypotheses about pace-setting factors in second language acquisition" (ibid: 108). When it comes to the contribution of LC in describing the developmental stages of interlanguage, Granger (2008: 259) also believes "Learner Corpora can contribute to Second Language Acquisition theory by providing a better description of interlanguage (i.e. transitional language produced by second or foreign language learners) and a better understanding of the factors that influence it; and they can be used to develop pedagogical tools and methods that more accurately target the needs of language learners".

## 2.5 Stages in Learner Corpora Research

Conducting research into LC, like any other field of study, needs to follow certain stages. Granger (2012:13) defined the seven main stages in Learner Corpora research including Choice of methodological approach, Choice of methodological approach, Selection and/or compilation of learner corpus, data annotation, data extraction, data analysis, data interpretation and the pedagogical implementation of learner corpora. She believes that these stages, with the exception of "data annotation' and "pedagogical implementation" are mandatory features in setting up a learner corpus. These stages are reviewed as follows:

(1) The Choice of Methodological Approach

According to Tognini-Bonelli (2001:2), corpus analysis is "an empirical approach, because it is derived from observing and describing authentic data", or more precisely the analysis and description of language use as realised in text(s). Therefore, describing or analyzing linguistic instances will lead to the distinction between two approaches: corpus-driven and corpus-based language studies (ibid). Granger (2012) also confirms that any researcher embarking on a corpus project chooses one of two main methodological approaches – **corpus-based** or **corpus-driven**.

Tognini-Bonelli (2001: 84-5) believes that corpus-based studies typically use corpus data in order to explore a theory or hypothesis, aiming to validate, refute or refine it. The definition of corpus linguistics as a method underpins this approach and as Granger (2012) asserts, it is essentially a deductive one. On the other hand, a corpus-driven approach rejects the characterisation of corpus linguistics as a method and claims instead that the corpus itself should be the sole source of any hypothesis about language. Therefore, it is an inductive approach which progressively generalizes from the observation of data to build up the theory or rule (Granger, 2012). According to Granger (2012), in the field of SLA research, most studies so far have used corpus-driven mythology and few studies have used learner corpora to test an SLA hypothesis.

(2) Selection and/or Compilation of the Learner Corpus

Since compiling a learner corpus is a time-consuming and difficult undertaking, Granger (2012) suggests that it is advisable to first survey the field to find out whether any suitable and available corpora for the research have already been compiled. However, if no learner corpus has yet been assembled (or made available) that could meet the purposes and requirements for conducting the required research, a suitable corpus needs to be compiled. The key point in compiling a learner corpus is to set appropriate design criteria. Leech (1998: 17) stressed the importance of careful and practical design criteria for corpora and stated: "the creation of corpora demands a great deal of spadework to be done before any research results can be harvested." It is also notable that according to Tono (2003: 800), the corpora design will vary from project to project based on researchers' interest in different aspects of learner language.

(3) Data Annotation

Granger (2012) believes that data annotation should not necessarily be a mandatory stage in developing a learner corpus because it varies based on different projects and in many cases data can even be successfully analysed in the format of raw (unannotated) corpus. Besides part-of-speech (POS) tagging, which is a common annotation for most corpora, the type of annotation most naturally connected to learner corpora is error tagging, where the corpus needs to be preliminarily annotated with the help of comprehensive error classification (Granger, 2012: 13; also Dagneaux *et al*., 1998: 163). It should be noted that although corpora annotation facilitates linguistic analysis, there are some challenges involved in learner corpora annotations. Gries & Berez (2015) point to the fact that since non-native language use (collected in learner corpora) contains nonstandard spellings, lexical items, and grammatical constructions, the annotation for such corpora requires great care in choosing the right tagset and more manual checking than is customary for native language corpora.

(4) Data extraction

The "data extraction" stage refers to using different search tools and programs for searching the corpora and extracting the required information. Granger (2012) introduced the most common functionalities of such programs/software to retrieve information: 1)

Word Lists, 2) Keyword Lists, 3) Concordancing, 4) Distribution /Range 5) Collocates, and 7) Clusters. It should be mentioned that data extraction is directly connected to the research purpose and therefore specific tools could be developed for error detection.

(5) Data Analysis

As already mentioned in 1.1, the two main approaches in LC data analysis are Computer-Aided Error Analysis (CEA) (Dagneaux *et al*., 1998) and Contrastive Interlanguage Analysis (CIA) (Granger, 1996 and Gilquin, 2000/2001). The first method is contrastive, and consists of carrying out quantitative and qualitative comparisons between native (NS) and non-native (NNS) data or between different varieties of non-native data. The second focuses on errors in interlanguage and uses computer tools to tag, retrieve and analyse them (Ganger, 2002).

(6) Data Interpretation

The concept of data interpretation is in contrast to data description. Granger (2012: 20) states that since the majority of LCR is focused on varieties of interlanguage description, data interpretation is not given due consideration. Theoretical interpretations can be studied by using multilingual and leaner corpus data, however, learner corpora could provide a solid base from which to interpret L2.

(7) Pedagogical Implementation

LCR provides not only a theoretical perspective on language learning and interlanguage studies, but also establishes a practical framework for improving pedagogical tools and methods. The main application of LC could be categorized as creating learner dictionaries, language courseware and syllabus design. For example, Rundell & Granger (2007) describe how learner and native corpus data was used to devise materials for inclusion in the Macmillan English Dictionary for Advanced Learners. Using error-tagged leaner corpora, Chuang & Nesi (2006) designed a remedial online self-study package called *GrammarTalk* which targets high frequency errors such as article errors (Granger, 2012). Another implementation of LC could be in language assessment when language instructors can use learner corpora to select and rank testing material for learners in different proficiency levels.

## 2.6 Learner Corpora Applications

A practical categorization for the language-pedagogical applications of learner corpora could be based on the direct and indirect usage of corpora. Such categorization was introduced by Leech (1997) in relation to native corpus data, which can be applied to learner corpus data. While a direct approach involves hands-on use of learner data on the part of the teacher and/or student, the indirect approach restricts the direct manipulation of learner corpus data to the researcher or publisher who produces the pedagogical resources (Granger, 2015).

According to McEnery & Xiao (2011: 365), the use of corpora in language teaching and learning has been more indirect than direct. This is perhaps because the direct use of corpora in language pedagogy is restricted by a number of factors including the level and experience of learners, time constraints, curricular requirements, the knowledge and skills required of teachers for corpus analysis and pedagogical mediation, and access to resources such as computers and appropriate software tools and corpora, or a combination of these (ibid.).

### 2.6.1 Delayed Usage vs. Immediate Usage of LC

Granger (2009a) presents another classification for the way learner corpora are collected as well as their pedagogical usage: delayed pedagogical usage (DPU) and immediate pedagogical usage (IPU). In DPU leaner corpora can contribute to the presentation of learner-corpus-informed reference tools (monolingual learners' dictionaries and pedagogical grammars) and instructional materials (course books and computer-assisted language learning programs). She uses the term DPU for the "indirect application of LC" and believes:

"In a DPU situation, learner corpora are not used directly as teaching/learning materials by the learners who have produced the data [but] are compiled by academics or publishers with a view to providing a better description of one specific interlanguage and/or designing tailor-made pedagogical tools which will benefit similar-type learners". (Granger 2009: 24-25)

In the case of IPU, learners are both the providers and users of the data, while corpus compilers are usually teachers who collect data from their students as part of their daily activities, thereby generating 'local' learner corpora. DPU corpora are usually bigger and therefore have wider generalizability. A good example of this type of corpus is the Longman Learners' Corpus or the Cambridge Learners' Corpus, which contain several million words of data from learners with a wide range of L1s. They are ideal resources for designing generic pedagogical tools like EFL dictionaries or grammars (Gillard & Gadsby 1998). On the other hand, IPU corpora are usually much smaller and therefore, as pointed out by Ragan (2001: 210), not representative in the usual sense of the word as they only represent themselves "providing specific information and a basis for generalizations concerning the limited range of the variety of language".

Based on the distinction between DPU and IPU, Meunier (2010) examines two DPU issues - syllabus design and material design. She has a critical view and argues that recourse to learner corpora for syllabus and material design is still relatively rare and holds the belief that LC provide "incomprehensible input", which is detrimental to learning and the topics covered in most LC are often far from the everyday needs of the vast majority of language instructors (ibid).

### 2.6.2 Specific Applications of LC

The specific applications of LC have been discussed by some researchers. Diaz-Negrillo & Thompson (2013) presented a comprehensive categorization of the users and activities surrounding learner corpora. They believe that the two main research user groups of LC are those of foreign language teachers and SLA researchers, and therefore, the specific applications of LC could be limited to those activities associated with language teaching and learning acquisition. For instance, in language testing and assessment (LTA), Callies *et al*. (2014) propose a threefold distinction of how learner corpora can be used. They suggest they should be either corpus-informed (the way corpus data are actually put to use), corpus-based (the aims and outcomes for LTA) or corpus-driven (the degree of involvement of the researcher in data retrieval, analysis and interpretation). In a similar study, Callies & Gotz (2015) confirm that learner corpora have the potential to increase

transparency, consistency and comparability in the assessment of L2 proficiency, and in particular, to inform, validate, and advance the way L2 proficiency is assessed.

## 2.7 An Overview of Some Learner Corpora Projects

Learner corpora research is a growing and expanding field of study. As mentioned in 2.3.4, a total of 149 projects have been listed on the 'Learner Corpora around the World' webpage maintained by the *Centre for English Corpus Linguistics* (CECL) at Louvain-la-Neuve, Belgium. In addition, Alfaifi (2015) listed and consequently reviewed 159 learner corpora in his research in order to determine the general trend of research in this area. In this section and for the purpose of familiarity with the existing projects, I review 10 learner corpora based on 9 aspects of corpus design criteria, which are presented in Table 2.1 with their references

Table 2.1: Learner corpora reviewed with their references

| No. | Learner corpus | Reference |
|---|---|---|
| 1 | Arabic Leaner corpus | Alfaifi (2015) |
| 2 | The Czech as a Second/Foreign Language Corpus (CzeSL) | Hana *et al.* (2010) |
| 3 | The Cambridge Learner Corpus (CLC) | Cambridge University (2012) |
| 4 | The International Corpus of Learner English (ICLE) | Granger (1993)<br>Granger (2003b)<br>Granger *et al.* (2010) |
| 5 | The Gachon Learner Corpus (GLC) | Price (2013) |
| 6 | The Louvain International Database of Spoken English Interlanguage (LINDSEI) | Granger *et al.* (2012)<br>Kilimci (2014) |

| | | |
|---|---|---|
| 7 | The Japanese English as a Foreign Language Learner Corpus (JEFLL) | Tono (2011) |
| 8 | The PELCRA Learner English Corpus (PLEC) | Pęzik (2012) |
| 9 | The Longman Learner Corpus (LLC) | Longman Corpus Network (2012) |
| 10 | The Corpus of Learner German (CLEG13) | Maden-Weinberger (2013) |

This review is similar to "*the corpora design criteria*" introduced by Tono (2003). He identified the following three major types of features for learner corpora: (a) language-related criteria, (b) task-related criteria, and (c) learner-related criteria. To cover all the aspects for reviewing the selected learner corpora, I have modified the categories and added one more criterion called "corpus criteria". Based on these criteria, the corpora are reviewed by 9 quantitative features as shown in Table 2.2.

Table 2.2: The features of corpora design criteria

| **Learner-related Criteria** | **Language-related Criteria** | **Corpus-related Criteria** |
|---|---|---|
| Target language | Mode | Data Annotation |
| Learner proficiency level | Task type | Size |
| Learners first language | Genre | Availability |

### 2.7.1 Learner-related Criteria

Some learner- related criteria which we will discuss are target language, proficiency level and learners first language.

(1) Target language

The first feature to review is the corpus content or what is termed "the target language". Although most of the learner corpora contain data from a single language, there are some corpora that include more than one language as the target language. Therefore, in terms of target language, corpora can be categorised into three groups: monolingual, bilingual and multilingual corpora. According to Alfaifi (2015), bilingual learner corpora can be used to undertake interlanguage studies, provided that they include comparable materials. Corpora involving multiple languages are beneficial when researchers need to investigate the effect of learners' L1 on second or foreign language acquisition, particularly if the corpus contributors share the same L1. It is not surprising that as Granger (2008:262) mentions "English clearly dominates the learner corpus scene", however, many languages are now targeted as learner corpora construction is developing fast. Table 2.3 categorizes the reviewed corpora in terms of the target language feature.

Table 2.3: Reviewed learner corpora based on the target language feature

| **Learner Corpora of English Language** | **Learner Corpora of Other Languages** |
|---|---|
| International Corpus of Learner English (ICLE) | Arabic Learner Corpus (ALC) |
| The Cambridge Learner Corpus (CLC) | The Gachon Learner Corpus (GLC) |
| The Louvain International Database of Spoken English Interlanguage (LINDSEI) | The Corpus of Learner German (CLEG13) |
| Japanese leaner corpus of English(JLCE) | The Czech as a Second/Foreign Language Corpus (CzeSL) |
| The PELCRA Learner English Corpus(PLEC) | |

(2) Learner Proficiency Level

Identifying learner proficiency levels can be tricky as it mostly depends on the language learning evaluation system. The learner corpora reviewed (Table 2.4) use the classic three levels of proficiency known as 'Beginner', 'Intermediate', and 'Advanced' with the sublevels of upper and lower, and if they contain different levels, they are marked as 'various'. However, other categorizations for marking proficiency levels can be used such as the well-known Common European Framework of Reference for Languages (CEFR) three-tier ranking system of A, B and C, therefore, it completely depends on the developers and thus cannot be regulated.

(3) Learners' First Language

In terms of learners' first language, corpora can be classified into two main categories: learner corpora with a single L1, for example the Japanese English as a Foreign Language Learner Corpus (Tono, 2011), and those with various L1s such as the Corpus of Academic Learner English (Callies & Zaytseva, 2011a, 2011b; Callies *et al.*, 2012). Table 2.4 gives a summary of the features of the learner-related criteria for the selected corpora.

Table 2.4: Features of the learner-related criteria

| No. | Learner corpus | Learner-related Criteria | | |
|---|---|---|---|---|
| | | **Target language** | **Level of proficiency** | **Learners first language** |
| 1 | Arabic Leaner corpus | Arabic | Intermediate and advanced | Various languages |
| 2 | The Czech as a Second/Foreign Language Corpus | Czech | Various | Various languages |
| 3 | The Cambridge Learner Corpus | English | Various | Various languages |
| 4 | The International Corpus of Learner English (ICLE) | English | High-intermediate to advanced | Various languages |
| 5 | The Gachon Learner Corpus | English | Lower intermediate | Korean and Chinese |
| 6 | The Louvain International Database of Spoken English Interlanguage (LINDSEI) | English | High-intermediate to advanced | Various languages |
| 7 | The Japanese English as a Foreign Language Learner Corpus (JEFLL) | English | From beginning to intermediate | Japanese |
| 8 | The PELCRA Learner English Corpus (PLEC) | English | Beginners to post-advanced | Polish |
| 9 | The Longman Learner Corpus | English | Various | English |
| 10 | The Corpus of Learner German (CLEG13) | German | Intermediate to advanced | English |

### 2.7.2 Language–related Criteria

Three main language-related criteria which we will discuss in this section are mode, task type and genre.

(1) Mode

Traditionally and according to Sinclair (2005:1), the mode of the text in corpora refers to whether the language originates in speech or writing, therefore the two common types of corpora based on mode are written and spoken corpora. Generally, there are many more written corpora as such learner material is easier to collect. Alfaifi (2015), who reviewed 159 learner corpora, found that two-thirds (66%) of the learner corpora included written data, and for example, The Cambridge Learner Corpus (Cambridge University Press, 2012), which is the biggest learner corpora in terms of size with 50 million words, consists of solely written corpora. Those that include solely spoken data, such as the French Learner Language Oral Corpora (Myles and Mitchell, 2012) represent 26% of the entire corpora. We can also take two more modes into consideration: the combination of written and spoken data which could be determined as a "written/spoken" mode and "multimodal" corpora which include written, spoken, and video data. The Santiago University Learner of English Corpus (Diez-Bedmar, 2009) is an example of the first type of corpora, while the Multimedia Adult ESL Learner Corpus (Stephen *et al*., 2012) is an example of the second.

(2) Task Type

This feature represents the style of the task or the collected data. Granger (2011: 12) categorizes two favourite text types represented in LC as "*Argumentative Essays*" for writing and "*Informal Interviews*" for speech. Alfaifi (2015) listed the task types of the LC and found 51 types including the "essay", "interview" and "test" at the most frequent.

(3) Genre

Corpus materials can be classified according to their specific genre such as argumentative, narrative, descriptive etc. According to Alfaifi (2015), few corpora focus on collecting material from a single genre such as the Scientext English Learner Corpus (Osborne *et al.*, 2012) which includes argumentative materials. Most corpora include different genres. For instance, the Taiwanese Corpus of Learner English (Shih, 2000) contains four genres: argumentative, narrative, descriptive, and expositive. The results of the review of the selected corpora (Table 2.4) for the 'language-related criteria' features are summarised in Table 2.5.

Table 2.5: Features of the language-related criteria

| No. | **Learner corpus** | **Language-related criteria** | | |
|---|---|---|---|---|
| | | **Mode** | **Task type** | **Genre** |
| 1 | Arabic Leaner corpus | Written and Spoken | Essays, interview | Argumentative, Descriptive, Narrative |
| 2 | The Czech as a Second/Foreign Language Corpus | Written and Spoken | Essays, interview | Descriptive, Narrative |
| 3 | The Cambridge Learner Corpus | Written | Exam scripts | Descriptive |
| 4 | The International Corpus of Learner English (ICLE) | Written | literary essays | Argumentative |
| 5 | The Gachon Learner Corpus | Written and Spoken | Written Journal Assignments | Descriptive, Narrative |
| 6 | The Louvain International Database of Spoken English Interlanguage (LINDSEI) | Spoken | Interviews and picture descriptions | Descriptive |
| 7 | The Japanese English as a Foreign Language Learner Corpus (JEFLL) | Written | Student essays | Descriptive |
| 8 | The PELCRA Learner English Corpus (PLEC) | Written and Spoken | Essays; formal letters | Argumentative, Descriptive |
| 9 | The Longman Learner Corpus | Written | Essays and exam scripts | Descriptive, Narrative |
| 10 | The Corpus of Learner German (CLEG13 | Written and Spoken | Free compositions | Argumentative |

### 2.7.3 Corpus-related Criteria

There corpus-related criteria: data annotation, size and availability will be discussed in this section.

(1) Data Annotation

Like other corpora, learner corpora are mostly annotated with different types of annotations. According to Alfaifi (2015), error annotation, part-of-speech (PoS) annotation and structural features (e.g. titles, sections, headings, paragraphs, questions, examples, etc.) are the most frequent types of LC annotations.

(2) Size

The size of a corpus is a controversial issue. While Sinclair (2005) believes that size is not a significant factor, so there is no maximum corpus size, Granger (2004: 125) argues that "learner corpora tend to be rather large, which is a major asset in terms of representativeness of the data and generalizability of the results". She also believes that learner corpora cannot be simply assessed according to the number of words compared with large general corpora, but the number of contributing learners is an equally important factor (Granger, 2003b). The present review shows commercial corpora, such as The Longman Learner Corpus and The Cambridge Learner Corpus, which are the biggest in terms of size.

(3) Availability

Access to the corpus data varies from one project to other. Based on the information for the learner corpora listed in the website of the *Centre for English Corpus Linguistics* (CECL) at Louvain-la-Neuve, five types of access to LC can be determined: freely available online, restricted availability, restricted (commercial: paid access), under development and not publicly available. Table 2.6 shows the specifications of the reviewed learner corpora (Table 2.4) based on the corpus-related criteria features.

Table 2.6: Features of the corpus- related criteria

| No. | Learner corpus | Corpus- related Criteria | | |
|---|---|---|---|---|
| | | **Size** | **Types of Annotation** | **Availability** |
| 1 | The Arabic Leaner corpus | Written: 283,000 words Audio: 3h30 | Errors | Available |
| 2 | The Czech as a Second/Foreign Language Corpus | 2 million words | Errors and structural features | Available |
| 3 | The Cambridge Learner Corpus | 50 million words | Errors | Commercial |
| 4 | The International Corpus of Learner English (ICLE) | 2 million words | Part-of-speech | CD-Rom and handbook |
| 5 | The Gachon Learner Corpus | 2.5 million words | Part-of-speech, lemma | Available |
| 6 | The Louvain International Database of Spoken English Interlanguage (LINDSEI) | 800,000 words | Spoken phenomena | CD-Rom and handbook |
| 7 | The Japanese English as a Foreign Language Learner Corpus (JEFLL) | 700,000 words | N/A | Under license |
| 8 | The PELCRA Learner English Corpus (PLEC) | 3 million words | Part-of-speech, lemma | Available |
| 9 | The Longman Learner Corpus | 10 million words | N/A | Commercial |
| 10 | The Corpus of Learner German (CLEG13) | 320.000 words | N/A | Available |

## 2.8 The Persian Language

Since the present thesis aims at constructing an error-tagged learner corpus of the Persian language for detecting and analysing linguistic errors made by Serbian students, a

short overview of the Persian language is presented in this part and the main phonological, morphological and syntactic characteristics are analysed.

Persian is an Indo-European language and it is used as the official language in Iran and Tajikistan and as one of the two official languages (along with Pashto) in Afghanistan. This language is officially called Farsi in Iran, Dari in Afghanistan, and Tajik in Tajikistan. In this thesis by Persian I refer to the contemporary Persian as spoken in Iran.

### 2.8.1 The Phonological and Orthographic Characteristics of the Persian Language

Some specifications of the phonology and orthography of the Persian language are introduced as follows:

(1) The Persian Phonemes

The Persian language has a total of 29 phonemes, 6 vowels and 23 consonants. Modern Persian has six vowels whose manner and place of articulation is shown in Table 2.7. The vowels, /i, u, â /, are considered long vowels, and /e, o, a/ short vowels. However, this traditional categorization is disputed because in some phonological environments the length of the short vowels may be longer than that of the long ones. For example, in the word <dard> 'pain' the vowel /a/ is longer than the vowel / â/ in the word <gâz> 'biting' (Samareh, 1985: 102). To be more specific, some linguists, such as Lazard (1992), consider the term “unstable” for short vs. “stable” for long vowels. In the present thesis and with a view to simplifying the detection of orthographic errors, I will adhere to the traditional categorization. According to this categorization, long vowels (â, i, u) are usually considered to be conveyed by alphabet letters whereas short vowels (a, e, o) are represented by so-called “diacritics”. Diacritics are normally left unwritten in texts and are mostly used for beginners, since adult native speakers are expected to have already developed cognitive strategies for efficient linguistic performance. Absence of diacritics causes errors, especially for the Farsi learners, since words can be read in different ways. As an example, the word ‘ کرم ‘ [K-R-M] , without diacritics, i.e. vowels, and if not be in a context, can be read differently with different meaning such as: [‘KeRM: *worm*] , [KeReM: *cream*], [KaRaM: *generosity*].

Table 2.7: Persian vowels

| Tongue position | Front | Back |
|---|---|---|
| High | i | u |
| Middle | e | o |
| Low | a | â |

Persian consonants consist of 27 phonemes. Table 2.8 illustrates the Persian consonants based on the place and manner of articulation.

Table 2.8: Persian consonants

| Place/ manner | Bilabial | Labio-dental | Dental | Alveolar | Post-alveolar | Palatal | Velar | Uvular | Glottal |
|---|---|---|---|---|---|---|---|---|---|
| Plosive | m | | t<br>d | | | k<br>g | | q | ʔ |
| Nasal | b p | | | n | | | | | |
| Fricative | | | | s z | ʃ ʒ | | | | h |
| Affricative | | f<br>v | | | t͡ʃ d͡ʒ | | x | | |
| Central Approx. | | | | r | | j | | | |
| Lateral Approx. | | | | l | | | | | |

Regarding the Persian phonological difficulties for the Serbian learners, the absence of long vowel /â/ and three consonants: /q/, /h/ and /x/ in the Serbian phonological system, could be causes of some errors in pronunciation as well as dictation.

(2) The Syllable System

During the evolution of the Persian language, the syllabic structure has become simplified. The Persian syllable system is a controversial issue, however, the majority of

linguists such as Bateni (1975) and Meshkotod Dini (1995) have accepted that it has the structure of (C)V(C)(C).

(3) The Persian Script

The Persian script is written and read from right to left. The writing system is based on the Arabic script which has been modified to represent the Persian phonemes. Therefore, four additional letters (consonants) which are absent in the Arabic alphabet (namely, گ/g/, چ/ /t͡ʃ/, پ/p/, ژ/ /ʒ/ /) have been added to the alphabet. According to the official instruction for Persian transcription by the Academy of Persian Language and Literature (2010)[2], the Persian writing system consists of 33 letters. The way each letter is connected to the previous or following letter in a word depends on whether that letter is at the beginning, in the middle or at the end of a word, although not all letters connect to the following one. The existence of punctuation and the cursive nature of the script, the omission of short vowels in writing and multiple consonant forms can be said to be the most important Persian script characteristics.

(4) Multiple Consonant Forms

In Persian, some consonant phonemes may be represented by different letters and that is due to the multiple forms of consonants representing one phoneme. The reason for such multiple forms lies in the number of loanwords from Arabic, which have been kept unchanged in Persian writing. For example, although in Persian there is only one phoneme for /z/, loanwords like 'لذیذ' [/laziz/ (adj). delicious], 'ظلم' [/zolm/ (n). oppression], and 'مریض' [/mariz/ (n). sick] maintain the original Arabic phonemes which represent three different phonemes. Table 2.9 shows the multiple forms of consonants in Persian phonemes. Regarding the Persian orthographic difficulties for the Serbian learners, it is expected that such characteristic may cause spelling errors, especially for learners in the elementary levels of proficiency.

[2] http://www.persianacademy.ir/UserFiles/Image/Dastoor-e%20khat/d02.pdf

Table 2.9: Persian consonant multiple forms

| /z/ | /s/ | /t/ | /h/ | /q/ |
|---|---|---|---|---|
| ز | س | ت | ه | ق |
| ذ | ث | ط | ح | غ |
| ظ | ص | | | |
| ض | | | | |

(5) Spoken Persian, Literary Persian

One of the most important themes of this section on the phonetic characteristics of Persian is the difference between its two registers – spoken and literary. Persian is counted among those languages which have two registers. One register is used by people in everyday conversation, while the other is used in correspondence, documents, books and written media, as well as on official radio and television stations. Spoken Persian can be divided into two registers – formal and informal. Formal spoken Persian is similar to literary Persian. The main differences between the spoken and literary registers of Persian are based on phonetic changes; therefore there are not many changes in word morphology or sentence syntax. For example in literary standard Persian, the plural suffix "hâ" is added to the noun to make it plural and in spoken Persian it changes into "â", therefore the word 'books' can be written in two ways: [ketabhâ (literary Persian)] vs. [ketabâ (spoken Persian)]. The distinction between spoken and literary Persian is important while developing the learner corpus, especially in error tagging, as there is a possibility that some words or sentences were written in the spoken form. A solution could be to determine the specific register used in the texts via the corpus metadata. As an example, in the corpus metadata, two types of annotation can be added regarding the register, i.e. 'standard form' and 'spoken form', then if in a text there were some different registers, then could be marked. As an example, the word [xejabân: street] can be written as [xejabun], therefore it can be marked by the metadata as 'spoken form'.

### 2.8.2 The Morphological Characteristics of the Persian Language

Persian is described as a predominantly agglutinative language (Jeremiás, 2003; Seraji, 2015), therefore word formation is dominated by the affixal system which appears in the form of both prefixes and suffixes.

Two major word formation processes in the Persian language are derivation and compounding. In derivation, new words are formed by adding prefixes and suffixes to the root. For example, the plural of the noun دانشجو /dânešju/ 'student' and the adjective دانشمندان /dânešmandân/ 'wise' was formed in the following way:

دان /dân/ (root) + ش /eš/ (nominal suffix) + جو /ju/ (nominal suffix).

دان /dân/ (root) + ش /eš / (nominal suffix) + مند /mand/ (adjectival suffix) + ان /ân/ (plural suffix).

As Seraji (2015) states, using derivational agglutination and combining affixes, verb stems, nouns, and adjectives to derive new words is highly productive in word formation. She provides the most common and frequent examples of derivational word formation, using the root /dân / (to know) to show the process in Persian morphology. I have kept the same common examples (ibid) as presented in Table 2.10.

Table 2.10: Examples of word derivation forms in Persian morphology

| Components | Transcription | PoS | Translation |
|---|---|---|---|
| دان | /dân/ | Verbal stem | to know |
| دان + ش | /dân-eš/ | Noun | knowledge |
| دان + ش + جو | /dân-eš-ju/ | Noun | student |
| دان + ش + مند | /dân-eš-mand/ | Noun | scientist |
| دان + ش + گاه | /dân-eš-gâh/ | Noun | university |
| دان + ش + گاه + ی | /dân-eš-gâh-i/ | Adjective | academic |
| دان + ش + گاه + ی + ان | /dân-eš-gâh-i-ân/ | Noun | academics |
| هم + دان + ش+ گاه + ی | /ham-dân-eš-gâh-i/ | Noun | university-mate |
| دان + ش + کده | /dân-eš-kadeh/ | Noun | faculty |
| دان + ا | /dân-â/ | Adjective | wise |
| دان + نده | /dân-andeh/ | Noun | knower |
| نا + دان | /nâ-dân/ | Adjective | ignorant |
| نا + دان + ی | /nâ-dân-i/ | Noun | ignorance |

Another morphological word formation process in the Persian language is called compounding which is done by connecting two simple or derived words. The following examples show word formation of this type:

اسباب بازی /asbâbbâzi/ ‘toy’: اسباب /asbâb/ ‘thing, medium’ + بازی /bâzi/ ‘game’

کتابخانه /ketâbxâne/ ‘library’: کتاب /ketâb/ ‘book’ + خانه /xâne/ ‘house’

As for other notable morphological characteristics of Persian, the following should be mentioned:

1) Modern Persian has no case system; therefore, nouns, adjectives, pronouns and adverbs do not have a declension, while the function of words in a sentence is primarily expressed by prepositions.
2) Verb conjugation follows the suffixal system: personal suffixes which carry information on tense, aspect and mood are added to the stems (a present or a past stem) and make regular conjugations. Verbs usually agree in person and number with the subject.
3) Simple adjectives and adverbs take suffixes (tar/tarin) to make comparative and superlative forms respectively.
4) Pronouns are often found in the form of pronominal clitics (م-/-am/ 1sg, ت-/-at/ 2sg, ش-/-as/˜ 3sg, مان-/-emân/ 1pl, تان-/-etân/ 2pl, شان-/-ešân/ 3pl) which are the bound forms of personal pronouns.
5) Gender is not marked in Persian: nouns, pronouns and adjectives do not have any gender markers; instead, various words are used to denote gender.
6) Possessiveness is expressed in two ways: the first and the most frequent is by the genitive clitic /–e/ which is called ‘Ezâfe’. It is an unstressed enclitic particle that connects all the constituents of a noun phrase, adjective phrase or prepositional phrase demonstrating the semantic relation between it parts. It is represented by a short vowel /e/ when used after consonants and /ye/ if used after vowels. The second way is by pronominal genitive clitics as mentioned in number 4.
7) There are several plural markers /-hâ/, /-ân/, (with variants -gân and -yân), and some Arabic plural markers /-ât/, -in, un, attaching only to Arabic loanwords.
8) There is no definite article in Persian.

### 2.8.3 The Syntactic Characteristics of the Persian Language

Persian belongs to the group of languages with SOV word order. However, the scrambling characteristic of the Persian language (Karimi, 2003), especially in spoken Persian, is something that makes Persian word order highly flexible. Therefore, the syntactic pattern has a mixed typology. The language represents a hybridization of two opposite syntactic patterns belonging to a group of typically VO languages (as in Arabic) and a group of typically OV languages (as in Turkish) (Stilo, 2004). Some of the main syntactic characteristics of SOV order are:

1. The subject may appear only as personal clitics on the verb, person and number are also inflected on the clitics.
2. Verbs normally agree in person and number with their subject, however there are some exceptions.
3. Direct objects are characterized by the postposition را (râ) which is the only case marker in the language. In Persian, a complement other than the direct object is introduced by a prepositional phrase (Lazard, 1992).

In the Persian language a sentence consists of a subject and object, which are optional, and a verb, which is compulsory, that is, (S), (O), V. It is possible to place the subject anywhere in the sentence. Alternatively, it may be completely omitted given that Persian is a "pro-drop" language, whose verb system is inflectional, i.e. person and number are inflected on the verb. In short, Persian word order is highly flexible as the use and order of optional constituents is fairly arbitrary. Regarding the Persian syntactic difficulties for the Serbian learners, the SOV word order could causes errors, since the word order in the Serbian language, as in other Slavic languages, is Subject – Verb- Object (SOV). Such errors will be discussed in chapter 6 based on the corpus reports and the statistics.

# 3. The Salam Farsi Learner Corpus

**Chapter Summary**

This chapter describes both the contents of the SFLC and how the data and metadata were collected; later the transcription process is explained in detail. The design criteria on which the corpus development is based are also discussed. This is followed by a description of the database design which is used to manage the SFLC data and automate the generating of the corpus files in different formats.

## 3.1 Developing a Model for Learner Corpora 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.

In the present thesis and for the purpose of providing a comprehensive guideline for designing learner corpora, I have tried to propose a comprehensive model to cover all the features and criteria involved in designing learner corpora. The proposed model consists of two types of features: (i) The Main Criteria for LC Design and (ii) The Specific Metadata for LC Design.

### 3.1.1 The Main Design Criteria for LC

In the proposed model, the main design criteria for LC fall into the following four categories:

1. **Corpus Criteria**, which are the features related to the corpus construction; such features include the mode of the LC (written, spoken, mixed), size (in terms of the total number of words), purpose (for academic/research work or commercial), availability (available or restricted access, free or commercial, etc.), users (researchers, instructors, lexicographers, etc.), and representativeness (representative or non-representative).

2. **Data Criteria**, which are the unique features regarding the content of the LC; such features include text type (written, spoken, media, mixed), task type (creative writing (free writing), composition, exam, essay, etc.) and genre (narrative, descriptive, argumentative, etc.).
3. **Learner Criteria**, which introduce the criteria which identify the learner features; such as the first language (which should be given special consideration when the data are compiled from learners with different L1s), the target language and level of proficiency (based on the learning program such as beginner, intermediate, advanced or A, B, C, etc.).
4. **Types of Annotation**, which determines the specific markup used in the LC such as (errors, part of speech, morphosyntactic, lemma, etc.)

The proposed model contains metadata markup restricted to 'data' and 'learner'. The 'specific metadata' is introduced and discussed in 3.1.2. Table 3.1 shows the proposed comprehensive design criteria for learner corpora.

Table 3.1: The proposed design criteria for learner corpora

<table>
<tr><th>Corpus Criteria</th><th>Data Criteria</th><th>Learner Criteria</th><th>Types of Annotation</th></tr>
<tr><td>Mode</td><td>Text type</td><td>First Language</td><td>Errors</td></tr>
<tr><td>Size</td><td>Task type</td><td>Target Language</td><td>Part of Speech</td></tr>
<tr><td>Purpose</td><td>Genre</td><td>Level of Proficiency</td><td>Morphosyntactic</td></tr>
<tr><td>Availability</td><td rowspan="4">Specific Metadata for the Data</td><td rowspan="4">Specific Metadata for the Learner</td><td>Semantic</td></tr>
<tr><td>Users</td><td>Phonological features</td></tr>
<tr><td rowspan="2">Representativeness</td><td>Lemma</td></tr>
<tr><td>Orthography</td></tr>
</table>

### 3.1.2. The Specific Metadata for LC Design

In the proposed model for LC design (Table 3.1) two types of metadata are suggested. Before introducing the proposed metadata, we need to review the importance of the role of 'metadata' in LC design. Burnard (2005:3) defines the term simply as 'data about data". Collecting metadata records enriches and equips the LC with additional information to describe the corpus data and the learners specifically. Granger (2012) indicates the necessity of recording metadata, naming them "leaner" and "task" variables:

"Full details about these variables must be recorded for each text.... This documentation will enable researchers to compile subcorpora which match a set of predefined attributes and effect interesting comparisons, for example between spoken and written productions from the same learner population or between similar-type learners from different mother tongue backgrounds" (Granger, 2002: 10).

The main purpose of including metadata in the learner corpus is to enrich it with relevant variables and to generate different studies based on connecting the main content of the corpus to such variables. Therefore, to develop the learner corpus metadata, each learner text in the corpus needs to be accompanied by such information. It can be obtained directly via a questionnaire (directly from the learners), or by collecting information from the institutes where the learners underwent instruction (provided permission for such a procedure is obtained). Table 3.2 introduces the proposed metadata variables in designing the learner corpora.

Table 3.2: The proposed metadata variables in designing learner corpora

| Metadata for LC Data | Metadata for the Learner |
|---|---|
| Title of the text | Gender |
| Year/month of the production | Age |
| Country of the production | Nationality |
| City of the production | Number of languages spoken |
| Task setting (home, class, exam session) | Number of years learning L2 |
| Timing (limited, free) | Profession (job) |

| | |
|---|---|
| Length of the text (min./max.) | General level of education |
| Type of written data (typed/ handwritten) | Educational institution |
| Software correction use | Major (field of Study) |
| Dictionary use | Year / semester |
| Grammar book use | Language learning Motivation (job, faculty/research, personal interests) |

## 3.2 The SFLC Design Criteria

Based on the criteria proposed in section 3.1, the criteria and features will be reviewed for developing the Salam Farsi Learner Corpus (SFLC) design criteria; and finally, the 12 criteria used in the SFLC design will be introduced.

### 3.2.1 The SFLC Corpus Criteria

The SFLC is a written learner corpus. The corpus has a minimum size of 26,978 words, which is the current status of the corpus for the thesis research, but the target size is 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. The corpus has no sub-corpora. A summary of the SFLC corpus criteria is given in Table 3.3.

Table 3.3: The SFLC corpus design criteria

| **Corpus Criteria** | **The SFLC** |
|---|---|
| Mode | Written |
| Size | 26,978 |
| Purpose | Academic use |
| Availability | Limited access |
| Users | Researchers |
| Representativeness | Representative |

### 3.2.2 The SFLC Data Criteria

In terms of mode the SFLC is a written corpus (Table 3.3), which contains the texts of Serbian learners of the Persian language, thus limiting the text type criteria to 'written'. The task types consist of 'free writing and composition' and since the majority of the texts are either descriptions or narrative essays on certain topics, such as the students' life, family, hometown, country, likes and dislikes, daily activities, etc., the data genre can be identified as "descriptive/narrative". Table 3.4 provides a summary of the data criteria for the SFLC.

Table 3.4: The SFLC data criteria

| Data Criteria | The SFLC |
|---|---|
| Text type | Written |
| Task type | Compositions, Free writing |
| Genre | Descriptive, Narrative |

### 3.2.3 The SFLC Learner Criteria

In the SFLC, the learners are only Serbian students, making the first/native language Serbian. This feature makes the SFLC a unique leaner corpus, as it is the very first Persian learner corpus which collects data from learners with only one L1 background. The second criterion in the proposed criteria is the "learners' target language", which in this corpus is "Farsi". The third criterion determines the learner's proficiency. In the SFLC, the proficiency levels range from pre-intermediate to advanced. For developing the SFLC, the data were collected from two groups of learners. The first group consists of learners who study the Persian language at the Faculty of Philology, University of Belgrade. This group of students follow a two-year course in the Persian language, which is called "Savremeni persijski jezik", as an elective language. Their proficiency is graded from A1 to A2, based on the language proficiency levels introduced by *The Common European Framework of Reference for Languages* (CEFR) (2001). The data collected from the students are restricted to those who study Persian in the second year, and according to the curriculum,

their proficiency should be equal to the level of A2. The second and the biggest group of learners whose productions have been collected for the SFLC are those who have studied Persian and attended the courses held by the Center for the Persian Language at the Iranian Cultural Center (ICC) in Belgrade. The Persian courses at this center are designed in three levels: beginner, intermediate and advanced, and they have already adopted a modified proficiency level system based on the CEFR consisting of Sath-e jek (equal to A), Sath-e do (equal to B) and Sath-e se (equal to C). The data were collected from the learners at intermediate to advanced levels, which are equal to A2 to C1 as based on the CEFR. It should be noted that in the present thesis, the CEFR grading system for learners' proficiency was chosen in order to maintain consistency in the data analysis. Table 3.5 shows the learners' criteria for the SFLC.

Table 3.5: The SFLC learner criteria

| **Learner Criteria** | **The SFLC** |
|---|---|
| First Language | Serbian |
| Target Language | Farsi |
| Level of Proficiency | A2 – C1 |

### 3.2.4 The SFLC Annotation Type

The SFLC is designed to be an error-tagged learner corpus, therefore only one type of annotation is implemented in the corpus. According to Nagata *et al.* (2011: 1210), as compared to other annotations, especially POS, "there are very few error-tagged learner corpora among existing learner corpora". That is also true when it comes to the Persian language corpora, since, to the best of my knowledge, only one project has been launched for developing the Persian error-tagged learner corpus, the Persian Learner Corpus (PLC) (Safari, 2012). Therefore the SFLC is the second attempt at developing an error-tagged learner corpus of Persian. To this end, the corpus is annotated systematically, according to a specific annotation schema which is introduced in chapter 4.

### 3.2.5 The SFLC Metadata

On the basis of the proposed design criteria for learner corpora (3.1.2), two types of metadata variables were collected to enrich the SFLC: (i) metadata for the texts and (ii) metadata for the learners. Such information was obtained from the questionnaires the learners completed on receipt of the specific notebook for their writing, called 'daftar-e negâresh' (see section 3.3.1). The statistics and details for the variables will be explained in detail in 3.3.1. Table 3.6 shows the collected SFLC- metadata.

Table 3.6: The SFLC metadata

| **The SFLC Metadata** | |
|---|---|
| **Learner Metadata** | **Text Metadata** |
| Age | Text title |
| Gender (M/F) | Year of production |
| Nationality | Country of production |
| Number of languages spoken | City of production |
| Number of years learning Farsi | Where produced (Home, Class, Exam session) |
| General level of education | Timing (Free, Restricted) |
| Major | References use (Yes, No, N/A) |
| Educational Institution | Grammar book use (Yes, No, N/A) |
| | Dictionary use (Yes, No, N/A) |

### 3.2.6 Summary of the Design Criteria Used in the SFLC

In 3.2, all the proposed design criteria were reviewed for the detection and selection of those useful for designing the SFLC. Some of the criteria were not applicable for this corpus, therefore only 12 criteria were used in the design and development of the SLFC. Table 3.7 summarises the SFLC design criteria.

Table 3.7: 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 |
| 11 | Level of Proficiency | A2 – C1 |
| 12 | Annotation | Errors |

## 3.3 The SFLC Content

Based on the corpus criteria, two types of data were collected: written texts and metadata variables. These data make up the content of the SFLC. As mentioned in 3.2.3, 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.

### 3.3.1 The SFLC Data Specifications

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). Permission for using the learners' work had been

received in advance when the students were registering for the course, and it was also mentioned in the special writing notebook designed for collecting the data, called the 'daftar-e negâresh' (writing notebook). Permission was also granted to the thesis researcher by the Director of the Center for Persian Language at the ICC Belgrade. As for the students at the Faculty of Philology, whose texts were collected occasionally, i.e. not organized in the 'daftar-e negâresh' liker the learners at the ICC, their permission was also gained at the beginning of their studies by singing a consent application form (see Appendix).

Some features of the corpus data have already been defined as based on the design criteria in 3.2.1; 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.

#### 3.3.1.1 The Corpus Size

Collecting corpus data is a time-consuming and challenging task. This becomes more problematic when the corpus data is restricted to a compilation from a specific group of learners. In this case, an insufficient amount of collected data, i.e. the size of the corpus, could affect the results of the analysis, especially when the corpus identifies the type and frequency of learners' grammatical errors. For this reason, Pravec (2002: 90) argues that "consideration for the size of a learner corpus is important. Otherwise, the sample size may cause the investigation into learner language to be insufficient, or at the very least, to be more difficult." However, reviewing some learner corpora for academic use, Granger (2004: 129) indicates that "the academic corpora, far more numerous, are extremely variable in size (the *Hong Kong University of Science and Technology Learner Corpus* contains 25 million words while the *Montclair Electronic Language Database* only contains 100,000 words)." She confirms the claim made by Ragan (2001:211) that even small corpora compiled by teachers of their own pupils' work are of considerable value and that "the size of the sample is less important than the preparation and tailoring of the language product and its subsequent corpus application to draw attention to an individual or group profile of learner language use." The SFLC, as constructed for the present thesis,

consists of 300 authentic written texts which in total contain 26,978 words. The corpus defines a target size of 100,000 words.

#### 3.3.1.2 The level of the Learners' Proficiency

As already mentioned in 3.2.3, to observe the consistency of the proficiency level, in the present thesis the grading system of the *Common European Framework of Reference for Languages* (CEFR) (2001) has been adopted, and each text has been tagged on the basis of proficiency levels from A2 to C2. The proficiency level for all the students who studied the Persian language at the Faculty of Philology, University of Belgrade, was considered to be level A2. They were placed at A2 level on the basis of the textbook, *Salam Farsi* (2015), which covers levels A1 and A2. The other group, the Persian learners at the Iranian Cultural Center, were classified at 3 main levels and 6 sub-levels, as (A) for Beginners (A1 – A2), (B) for Intermediate (B1- B2) and (C) for Advanced (C1- C2). They completed each main proficiency level in one academic year, i.e. two semesters (e.g. A1 for the first and A2 for the second semester); therefore the collected data belongs to different levels and sub-levels. In the SFLC, the proficiency level is marked in 4 sub-levels (A2, B1, B2, and C1). The data collected from the two groups of learners can be categorized in the following proficiency levels as shown in Table 3.8.

Table 3.8: The SFLC proficiency levels

| Groups | Proficiency levels |
|---|---|
| Students at the Faculty of Philology | A2 |
| Learners at the Iranian Cultural Center | A2 – B1– B2 – C1 |

#### 3.3.1.3 Text Type, Task Type and Genre

The SFLC contains only written texts. The task types are restricted to two groups: firstly, compositions (on specific subjects), and secondly, creative writing (free writing) assignments. For the compositions, the students were given specific subjects to write about

for example, my city and my country, the best trip I ever had, my family, sports, the seasons, etc. In their creative writing assignments, they were free to write about anything which interested them. Based on these two types of tasks, the data genres are limited to description (i.e. describing a person, event, scene, etc.) and narration (i.e. the personal retelling of events, experiences, feelings, etc.). Table 3.9 shows the details of the task types and genres of the corpus data.

Table 3.9: The task types and genres in the SFLC

| Corpus Criteria | Data Description | Total Number of Documents |
|---|---|---|
| Task type | Compositions | 115 |
| | Creative Writing | 185 |
| Genre | Descriptive | 197 |
| | Narrative | 103 |

#### 3.3.1.4 Summary of the SFLC Data

Table 3.10 gives a summary of the SFLC data.

Table 3.10: A summary of the SFLC data

| Corpus Criteria | Data Description | Documents | Total Words | Percentage |
|---|---|---|---|---|
| Size | Texts collected from students at the Faculty of Philology (FPH) | 62 | 5,575 | 22.70% |
| | Texts collected from learners at the Iranian Cultural Center in Belgrade (ICC) | 238 | 21,403 | 79.30% |
| | | | | |

| | | | | |
|---|---|---|---|---|
| Levels of Proficiency | A2 | 62 | 5,575 | 21% |
| | B1 | 81 | 7,284 | 27% |
| | B2 | 101 | 9,082 | 33% |
| | C1 | 56 | 5,035 | 19% |
| | | | | |
| Task Type | Compositions | 115 | 10,342 | 38% |
| | Creative Writing | 185 | 16,636 | 62% |
| Genre | Descriptive | 197 | 17,716 | 66% |
| | Narrative | 103 | 9262 | 34% |

### 3.3.2 The SFLC Metadata Specifications

Based on the corpus design criteria, the SFLC metadata were introduced in section 3.2.5. Two groups of variables were collected as metadata, providing information about the learners (i.e. the producers of the data) and the data (i.e. written texts). .

#### 3.3.2.1 Learner Metadata

Eight variables were selected as the 'learner metadata' in the corpus design and the data was subsequently collected. Table 3.11 shows the learner metadata in the SFLC. With the exception of 'nationality', which is 'Serbian' for all the learners in the present corpus, and 'the general level of education' and the 'major' which provide a clear overview of the learners' educational backgrounds, the other variables can be used to compare different learners in order to investigate the effect of such variables when analyzing the type and frequency of learning errors.

Table 3.11: The SFLC learner metadata

| **The SFLC Learner Metadata** | | **Variables** |
|---|---|---|
| 1 | Age | Various 19 to 67 |
| 2 | Gender | Male, Female, N/A |
| 3 | Nationality | Serbian |

| | | |
|---|---|---|
| 4 | General level of education | BA, MA, PhD, N/A |
| 5 | Educational Institution | ICC, FPH |
| 6 | Major | Various |
| 7 | Number of years learning Farsi | Various (1 to 5 years) |
| 8 | Number of languages spoken | Various (2 to 6 languages) |

#### 3.3.2.2 Text Metadata

Nine variables were selected as the 'text metadata' in the SFLC as shown in Table 3.12. The 'text title' could be used to distinguish text types (i.e. compositions vs. creative writing). The texts with common titles indicate the composition text type, while those without titles or with various titles can be considered as creative writing. The corpus data were gathered in Belgrade, Serbia, in the academic years 2012 to 2015. The texts were produced at 'home' or 'in the classroom' with 'free or restricted timing'; the information about the use of references, dictionaries and grammar books remained unavailable. Table 3.10 shows the text metadata details.

Table 3.12: The SFLC text metadata

| NO. | **Text Metadata** | **Variables** |
|---|---|---|
| 1 | Text title | Various |
| 2 | Year of production | 2012-15 |
| 3 | Country of production | Serbia |
| 4 | City of production | Belgrade |
| 5 | Where produced | Home/Classroom |
| 6 | Timing (Free- Restricted) | Free / Restricted |

| | | |
|---|---|---|
| 7 | References use | Non-Applicable |
| 8 | Grammar book use | Non-Applicable |
| 9 | Dictionary use | Non-Applicable |

## 3.4 Digitizing the SFLC

The SFLC raw data consisted of hand-written texts; therefore the process of digitizing the data was implemented to convert the texts into electronic form to make them readable by the corpus software tools. The process consisted of the following 4 phases:

1. Scanning the hand-written texts and saving them in pdf format,
2. Defining the instructions for the transcription,
3. Manually transcribing the texts,
4. Creating the corpus database.

It should be noted that scanning the hand-written texts was done in order to obtain digital files in Portable Document Format (PDF) so as to save the original texts.

### 3.4.1 Defining an Instruction Format for the Transcription

Considering the specific characteristics of Persian orthography, especially the cursive nature of the script and the possibility of writing some words in different forms (i.e. the plural suffix –'ha' - can be written either in segmented or unsegmented form, for example the plural for the word 'book' in Persian 'کتاب /ketâb' can be written segmented like کتابـــها or unsegmented as کتاب ها) and in order to achieve consistency in the manual transcription, instructions for transcribing the raw texts have been defined. To the best of the researcher's knowledge, this is the first proposed set of instructions for transcribing the raw materials for a Persian learner corpus. Table 3.13 lists the proposed instructions for the transcription of the raw texts in the SFLC.

Table 3.13: The instructions for the data transcription in the SFLC

| No. | The Instructions |
|---|---|
| 1 | The texts should be transcribed without any corrections and should remain authentic |
| 2 | Except for the title of the text, all the metadata variables should be excluded from the text body |
| 3 | If the text has no title, it should be marked by *** |
| 4 | The teacher's corrections and comments should be excluded |
| 5 | Any struck-out texts should be excluded. |
| 6 | The diacritics should be excluded, except when indicating a proper name (person, city, etc.) |
| 7 | The segmented plural suffix (-ha) should be unsegmented |
| 8 | The semi-space should be applied in transcription based on the Persian Academy manual |
| 9 | In cases of illegibility, the form closest to the correct form should be transcribed |
| 10 | If a dot character is omitted, it should be transcribed as written by the learner |
| 11 | In cases of changing the place of a word by arrows, the correct form should be transcribed |
| 12 | Any shapes, ornamentations or underlined words or sentences should be excluded |

### 3.4.2 Transcribing the Texts

As already mentioned in 3.4, all the collected papers were scanned; however, in the process of transcription only the raw texts were transcribed, and some pages, such as the cover page for the 'daftar-e negâresh' containing the learners' information, were deleted; blank pages were also excluded. Later, after setting the transcription instructions, the texts were transcribed manually by the researcher himself and subsequently entered into the corpus by means of the Data Submitting and Metadata Tagging Tool (DSMT) (see 5.2.1). Table 3.14 shows the detailed information for the raw texts .

Table 3.14: The SFLC raw text data

| Documents | Total number of documents |
|---|---|
| Total of scanned pages | 610 |
| Total of deleted sheets | 25 |
| Total of transcribed texts | 300 |

Figure 3.2 provides an example of a scanned raw text.

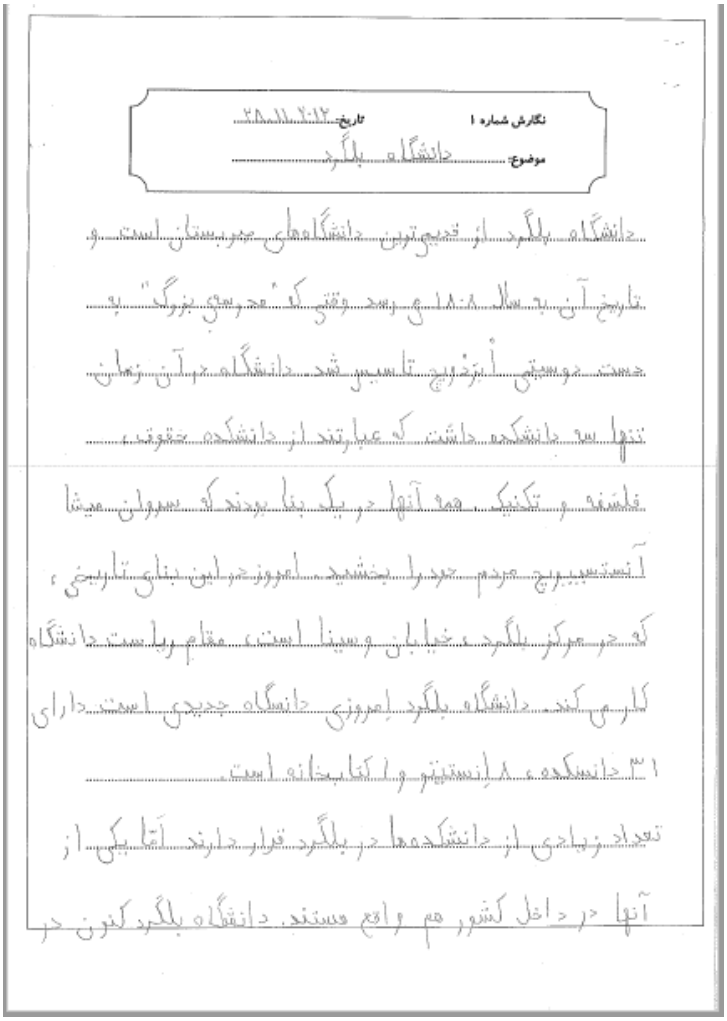



Figure 3.1: A scanned raw text in PDF format

### 3.4.3 Document Storing

The transcribed texts were saved in the SFLC database, which is a web-based, online database. To set the corpus database, and with the aim of providing corpus tools, an internet domain was registered (www.salamfarsi.com) and subsequently hosted to the Linux virtual private server (VPS) supporting the Python programming language. The SFLC uses a type of SQL database, PostgresSPL, so the data are submitted in the database via the DSMT application tool. The technical information about the process of data storing in database will be discussed in 5.3. Each text is considered as a new document and entered in the DSMT 'text box' while the metadata rerated to that text is need to be entered before submitting to database. Figure 3.3 illustrates the text submission in the DSMT panel.

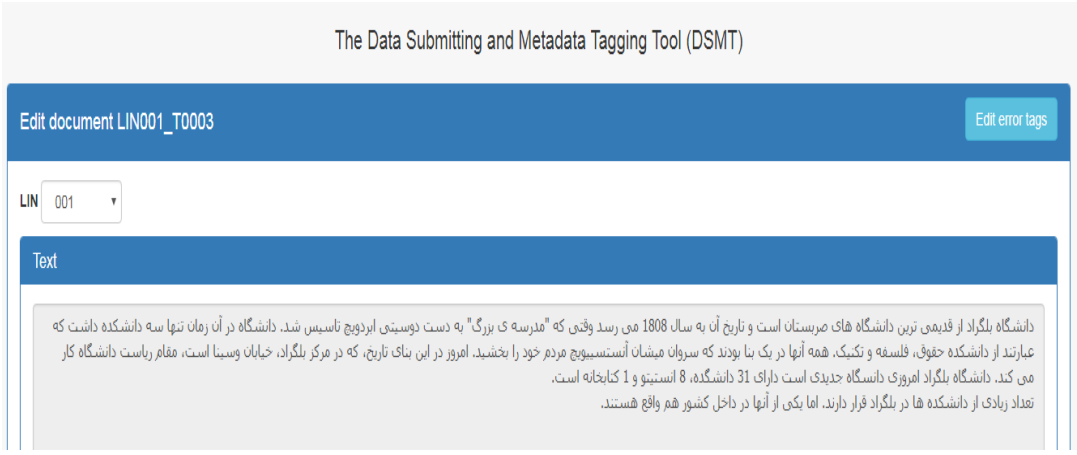



Figure 3.2: The raw text submission in the SFLC database

As already mentioned, in addition to the raw texts, metadata is entered, therefore, four types of variable data were manually entered into the database for each document:

1. Data criteria (text type, task type, genre)
2. Leaner criteria (first language, target language, proficiency level)
3. Text metadata (text title, year of production, country of production, city of production, place of production, timing, reference use, dictionary use and grammar book use)

4. Learner metadata (age, gender, nationality, number of languages spoken, number of years learning Persian, general level of education, major, educational institutes)

### 3.4.4 File Generation and Naming

The SFLC database will provide different files from the corpus raw text (i.e. without annotation) along with the variables in TXT and PDF formats. All the files were named based on the "Learner's Identification Number (LIN)", and the specific code for the "Text" (T). As an example, a generated file under the name of (LIN038_T0115) indicates the learner's code (038) and the text code (0115). In this way, if a learner has produced more than one text, the LIN code will remain the same and only the text code will change. The SFLC database enables users to study the corpus using concordances and frequency word lists, in addition to allowing a large amount of annotation to be added and utilised in the corpus which will be discussed in detail in chapter 4 (The SFLC tools) and chapter 5 (The SFLC uses and data analysis).

# 4. The SFLC Error Annotation System

**Chapter Summary**

This chapter provides an introduction to Error Analysis in Second Language Acquisition and the concept of EA in learner corpora. The idea of computer-aided error analysis in learner corpora is also discussed. After reviewing the theory, and with the aim of developing an annotation system for the SFLC, the error classification process is explained and the model of error taxonomy in the corpus introduced. The chapter concludes with the design of a specific tagset for annotating the errors based on the specific error taxonomy of the SFLC.

## 4.1 Error Analysis: An Overview

According to Richards & Schmidt (2002:184), Error Analysis (EA) is "The study and analysis of the errors made by second language learners." Corder (1967:19-27), as a pioneer in the study of errors in students' writing, points to the objectives of such a study and claims, "EA has two objectives: one theoretical and the other applied." The theoretical objective serves to "elucidate what and how a learner learns when he studies a second language. The applied objective is to enable the learner to learn more efficiently by using the knowledge of his dialect for pedagogical purposes. (ibid)" Brown (1980, cited by Hasyim, 2002:43) adopts another point of view regarding EA, and considers it to be a process of "observing, analysing, and classifying" errors, which he refers to as "deviations from the rules of the second language" for the purpose of revealing the systems operated by a learner. In this part, the concepts of Error Analysis in SLA research and learner corpora are discussed.

### 4.1.1 SLA Research and Error Analysis

Gass & Selinker (2008:1) define SLA as "the study of how learners create a new language system." As a research field, they add that SLA is "the study of what is learned of a second language and of what is not learned" (ibid). The goals of SLA research are to describe how second language acquisition proceeds, and to identify the factors that account for the reasons why learners acquire a second language in the way they do. As Davies (2013: 45) asserts, "many SLA researchers would argue that the formal study of SLA was launched in 1976 with Corder's publication, "*The Significance of Learners' Errors"*. Its construct of "transitional competence", together with research on "interlanguage" (Selinker, 1972) and data description through "Error Analysis" (Richards, 1974), laid the groundwork for most of the early studies in the field of SLA. In other words, the issue of "learner errors" was somehow the basic framework for the study and research of SLA.

As for the term 'errors', Coder (1973: 260) defines them as "those features of the learner's utterances which differ from those of any native speaker". He also believes that linguistic errors are systematic and reflect a defect in knowledge; i.e. linguistic competence

(ibid). According to Dulay *et al.* (1982: 150-160), the term generally refers to a systematic deviation from a selected norm or set of norms. Lennon (1991:182) defines an error as "a linguistic form or combination of forms which in the same context and under similar conditions of production would, in all likelihood, not be produced by the speakers'native speaker counterparts".

Regarding SLA research and its connection to Error Analysis, Saville-Troike (2006: 37) believes "Error Analysis is the first approach to the study of SLA which includes an internal focus on learners' creative ability to construct language." It is based on the description and analysis of actual learner errors in L2, rather than on idealized linguistic structures attributed to native speakers of L1 and L2. In the late 1960s and early 1970s, several studies pointed out that the language of second language learners is systematic and that learner errors are not random mistakes but evidence of rule-governed behaviour. Corder (1967:19-27) was the pioneer in developing "Linguistic Error Analysis" who highlighted the importance of studying errors in learners' writing. He suggested that by classifying the errors that learners make, L2 researchers can learn a great deal about the processes and strategies used by language learners (ibid).

The EA process has been the focus of research for some time. Corder (1976), cited in Ellis (1994), suggests the following steps for EA research:

(1) Collection of a sample of learner language,
(2) Identification of errors,
(3) Description of errors,
(4) Explanation of errors,
(5) Evaluation of errors.

Corder also emphasizes the importance of the data selected for analysis, and how this data has been collected in particular (Castillejos Lopez, 2009). Gass & Selinker (2008:103) identified six steps to be followed in conducting Error Analysis, namely, collecting data, identifying errors, classifying errors, quantifying errors, analysing the sources of errors and remediating errors. Finally, it should be mentioned that despite some criticism of Error Analysis in terms of its weaknesses in methodological procedures and its limited scope (Maicusi *et al*.,1999), EA still preserves its merits as an effective approach

for dealing with L2 learner errors, often used alongside other analytical techniques (Ellis, 1994).

### 4.1.2 Learner Corpora and Error Analysis

Learner corpora could play a supportive role in meeting existing criticism of Error Analysis for methodological reasons. Castillejos Lopez (2009) explains such criticism in brief, and enumerates some of the criticisms as follows: weaknesses in error evaluation judgments, lack of precision in defining the point of view under which an utterance is considered erroneous, difficulty in finding the interlingual or intralingual source of error, and difficulty in the classification and interpretation of errors. She concludes that as "the authentic data" constitute the convergence point in Corpus Linguistics and Error Analysis, while neither of them are theories of language acquisition but methodologies, learner corpora could provide the object of study and EA could determine the techniques. Both could keep their own rules but make mutual contributions in order to offer results that would enrich SLA theories and language teaching. Another important issue in using learner corpora for the purposes of EA is the theoretical aspect of error classification. Learner errors may be classified according to different aspects and various criteria. During the last few decades some different/new categorizations have been introduced, based on the aims and purposes of the required analysis. As a pioneer in this field, Richards (1971) believed that the forms of learner errors could be grouped into four categories: (1) Overgeneralization, (2) Ignorance of Rule Restriction, (3) Incomplete Application of Rules, and 4) False Concepts Hypothesized (Ellis, 1994: 59). Lee (1990) suggested a four-level error classification based on learner performance, and introduces them as (1) Grammatical Errors; (2) Discourse Errors; (3) Phonologically-induced Errors and (4) Lexical Errors.

Saville-Troike (2006) proposed three main error categories based on (1) Language Level: whether an error is phonological, morphological, syntactic, etc.; (2) General Linguistic Category: e.g. auxiliary system, passive sentences, negative constructions; and (3) Specific Linguistic Elements: e.g. articles, prepositions, verb forms.

Other researchers like Dulay, Burt & Krashen (1982) argue the need for descriptive taxonomies of errors that focus only on the observable, surface features of errors (Ellis, 1994: 54). Dulay *et al.* (1982: 146) discuss four major types of descriptive error taxonomies in depth. They propose such taxonomies as constituting the (1) Linguistic Category, (2) Surface Strategy, (3) Comparative Analysis, and (4) Communicative Effect. They believe the two major descriptive error taxonomies to be (1) Linguistic Categories, such as morphology, lexis, and grammar (more specifically, auxiliaries, passives, and prepositions), and (2) Surface Structures Alternation or Modification (ibid). Such views and categorizations may also be considered the theoretical background to developing an error-tagged learner corpus.

#### 4.1.2.1 Computer-aided Error Analysis

As already mentioned in 1.1, Granger (2002) indicates that the two complementary approaches to learner corpus analysis are Contrastive Interlanguage Analysis (CIA) and Computer-aided Error Analysis (CEA) which together make up a powerful 'methodology' for the quantitative and qualitative study of learner language. CEA shares the same aims as traditional Error Analysis (i.e., the study and analysis of the errors made by L2 learners). However; according to Dagneaux *et al.* (1998), there is a difference between the two approaches because CEA methodology uses a wide range of linguistic software tools to store and process the learner language thus providing automatic linguistic analysis. They also propose 5 steps for the entire CEA process (ibid) in which the two main software-oriented stages are (1) the insertion of error tags and corrections in the text files, and (2) the retrieval of lists of specific error types and error statistics.

The tagging procedure and correction insertion is usually accompanied by an 'error editor' which allows researchers to mark errors in a text (Granger, 2002). The 'error editor', such as the UCLEE (Université Catholique de Louvain Error Editor), is a menu-driven editor which enables the annotator to insert an error tag at the relevant point in the text by clicking on the appropriate tag from the error tag menu (ibid).

Once the error-tagging process has been completed it is possible to perform Error Analysis with the help of text retrieval software tools. This is possible by searching the

corpus data by means of the error tags, sorting the concordance lines in various ways to obtain relevant error patterns and examining them in the context of the other interlanguage phenomena which exist in the linguistic context of the co-text, as well in as the wider context. Researchers are thus able to obtain and present reliable quantitative and qualitative descriptions of learners' difficulties in the context of the relevant subsystem of their interlanguage (Dagneaux *et al*., 1998; Granger, 2003).

It can be concluded that CEA represents a major improvement in the development of error analysis methodology for at least two reasons: firstly, it helps to overcome the limitations of traditional error analysis; and secondly it examines errors in the full context of the surrounding text while simultaneously exploiting a wide range of linguistic errors with the help of software tools. In the present thesis, the error analysis process is based on CEA methodology and the error annotation and data retrieval procedures are carried out by specific software tools which will be discussed in chapter 5.

## 4.2 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."

The development of the SFLC error-tagging system includes (1) a 'model for error taxonomy (the SFLC error taxonomy)' and (2) a 'tagset designed for annotating errors (the SFLC error tagset) which is described in the following section.

### 4.2.1 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 4.1 introduces the SFLC surface structure error taxonomy.

Table 4.1: 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 4.2.

Table 4.2: 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. |

### 4.2.2 The SFLC Error Tagset

The SFLC error tagset is developed based on the SFLC Error Taxonomy introduced in 4.2.1 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 4.3 shows the SFLC error tagset.

Table 4.3: 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[3]:

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

(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].

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"

# 5. The SFLC Software Interface and Tools

**Chapter Summary**

This chapter introduces the four main tools with which the SFLC is equipped in order to function as a learner corpus. To this end, first of all the software interface is introduced, from where the tools are accessed, and then each tool is discussed in detail. These four tools are: the Data Submitting and Metadata Tagging Tool (DSMT),which deals with storing data in the corpus database and marking with metadata tags; The Error Tagging Tool (ETT), which functions as a computer-aided error editor and facilitates the error tagging; the Filter and Search Tool (FST), which includes different filters and enables searches for specific errors or words in the corpus; and finally the Data Statistics Tool (DST), which shows various statistical data related to the corpus.

## 5.1 The SFLC Webpages/Interface

The SFLC uses a web-based interface for submitting, tagging, filtering, searching and downloading the corpus data, as well as providing statistics by means of four technical corpus tools. The SFLC website (http://www.corpus.salamfarsi.com), including all the tools, were designed by the author of this thesis based on the corpus design criteria (outlined in Section 3.2), then subsequently created by a group of software development technicians (3.4.3) paid by the researcher (see 5.3.4). The corpus is hosted on the web-hosting service of the Faculty of Philology, University of Belgrade, with official permission granted to the author by the Dean of the Faculty of Philology.

The SFLC can also be accessed through a link on the website www.salamfarsi.com, which was developed by the researcher independently to provide details about learner corpora research, related publications and other information regarding this field of research. The corpus website consists of 5 main pages: (1) the 'Login' page, where the user can log in or sign up to access the corpus, (2) the 'Data and Tagging' page, where the data-submitting and tagging tools are located, (3) the 'Filter and Search' page, where specific tools enable filtering and searching the corpus data, (4) the 'Data Statistics' page, where statistical data are shown, and (5) the 'About' page, which provides general information about the corpus, as well as the links to access the tools directly. An SFLC User Guide, located at the top of each page in PDF format, was also created with instructions on how to use the corpus (see Appendix).

### 5.1.1 The SFLC Login Page

On the login page, users can enter their username or password to enter the corpus if they have obtained it in advance. Otherwise, since the corpus is not available for open access, new users need to sign up to obtain access from the corpus administrator. On the login page, the logo of the Faculty of Philology is provided to indicate the corpus affiliation, and a dedicated logo has been designed to represent the SFLC. A short description of the corpus is also provided on this page. It should be noted that since the

corpus will be developed further, the current version is called the SFLC Version 0.1. Figure 5.1 shows the SFLC corpus login page.


Figure 5.1: The SFLC login page

### 5.1.2 The SFLC Data and Tagging Page

Two access mode options, for the administrator/annotator and for the user, are defined for the corpus. The admin mode provides access to the data, i.e. to enter texts, insert tags and edit, while the user mode provides limited access only and does not allow data and tag insertion. This section focuses on the full access mode, the administrator/annotator access.

The data and tagging page contains a list of submitted documents referred to as the 'Document List', and four groups of filters. The Document List shows the data submitted, with five specifications: LIN Code, Date Creation, Last Modification, Error Tag Check and Operations. The filters contain four groups of tags, associated with 'Data Criteria', 'Learner Criteria', 'Text Metadata' and 'Learner Metadata'. The data can be sorted and shown in the Document List by changing the values of the variables included in the filters.

The page provides access to the two main corpus tools, namely, the Data Submitting and Metadata Tagging Tool (DSMT) and the Error Tagging Tool (ETT). To access the DSMT, the administrator needs to click on the 'Submit New Document' option, which leads to the tool where text and metadata can be submitted. By submitting the tagged data in the DMST, the next tool, the ETT, which is an error editor designed to facilitate error tagging, will become available for inserting error tags. Both tools are only accessible to the corpus administrator(s)/annotator(s) and are not shown to other corpus users (i.e. they do not see the 'Submit New Document' option). The function of these tools will be explained in Sections 5.2.1 and 5.2.2. Figure 5.2 shows the 'Data and Tagging Page' in the corpus.

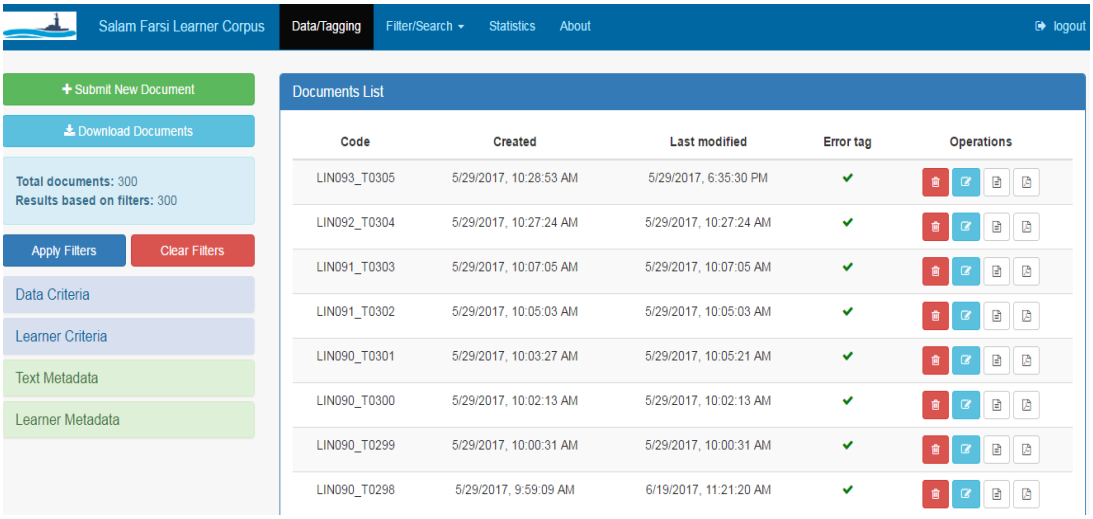



Figure 5.2: The SFLC data and tagging page

### 5.1.3 The SFLC Filter and Search Pages

These pages are developed based on the Filter and Search Tool (FST). Two separate pages were created for the FST to allow for filtering and searching the annotated errors and corpus data. These pages are named the 'Errors' page and the 'Words' page. The FST on the 'Errors' page provides the error occurrence in the corpus based on the filter variables, while the 'Words' page is for searching any words or phrases in the whole corpus or

applying filters to obtain specific results. Both pages show the search results in context. The FST function in these pages is described in detail in Section 5.2.3.

### 5.1.4 The SFLC Data Statistics Page

The Data Statistical Tool (DST) consists of two main parts, the filters and the diagram window, where the tool provides the statistics based on the different filters. The DST function will be discussed in detail in 5.2.4. Figure 5.3 shows the SFLC Statistics Page.

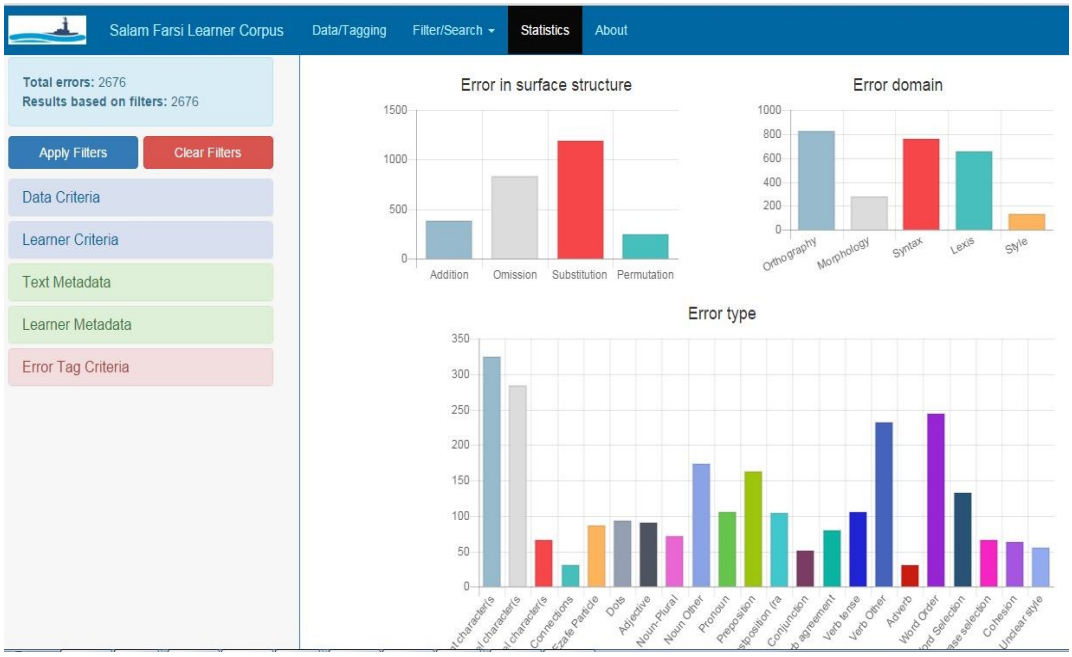



Figure 5.3: The SFLC statistics page

### 5.1.5 The About Page

The ‘About’ page presents general information about the corpus and briefly describes the corpus data, metadata, error tags and tools. Some basic information which gives an overview of the corpus is also provided in a table called ‘SFLC Quick Info’. The

lists of corpus tools with direct links to the pages where they are located are also provided. Figure 5.4 shows the ‘About’ page in the corpus.

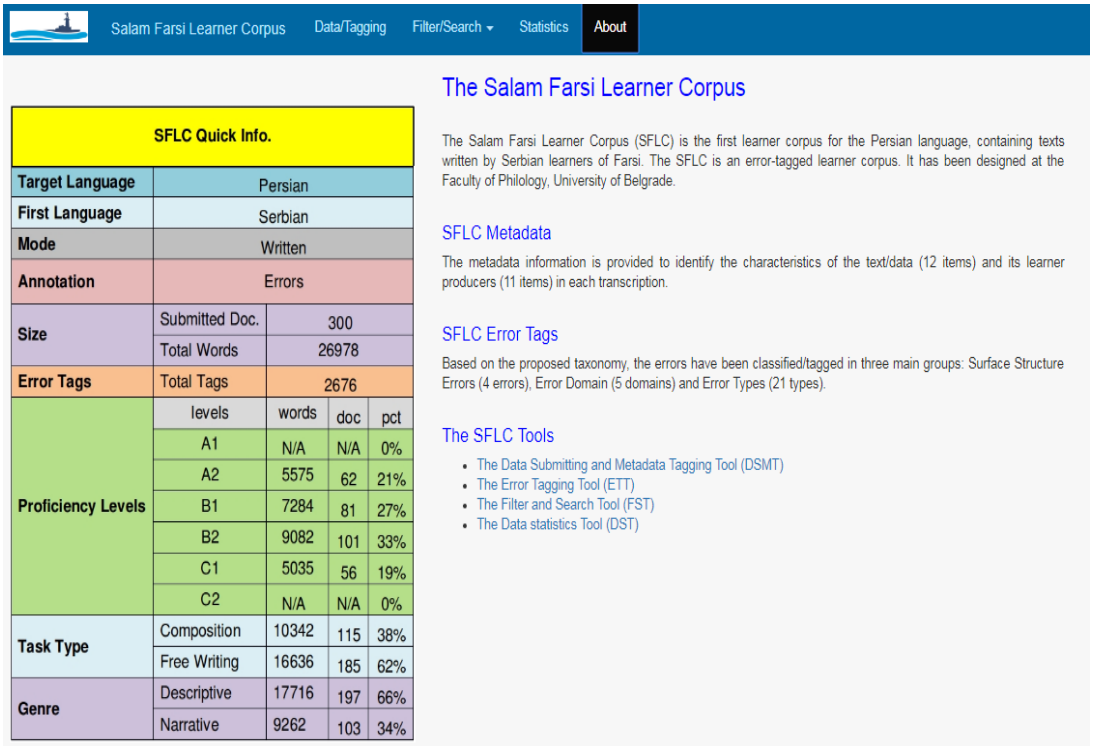



Figure 5.4: The SFLC about page

## 5.2 The SFLC Tools

The SFLC uses four main tools which were created and developed for submitting the raw data to the corpus database, annotating metadata and errors, filtering and searching for specific data in the whole corpus, as well as providing the corpus statistics. These tools are accessed from the pages described in Section 5.1, which in this section are introduced separately in detail.

The first tool, ‘The Data Submitting and Metadata Tagging Tool’, was developed to store and save the raw data in the corpus database, and to assign metadata tags based on the proposed SFLC metadata specifications (3.3.2). The second tool, ‘The Error Tagging Tool’,

was designed according to the SFLC Error Tagset (4.2.2) to function as a computer-aided error-tagging tool for annotating the errors. The third tool, 'The Filter and Search Tool', enables users to filter the metadata or/and error tags, then to search through the corpus and obtain the results in concordance format. The fourth tool, 'The Data Statistics Tool', includes filters to sort specific data and a diagram box which shows the distribution of errors in the corpus based on the SFLC Error Tagset (4.2.2). This section explores and discusses these four corpus tools and their functions.

### 5.2.1 The Data Submitting and Metadata Tagging Tool

The DSMT was developed on the basis of the SFLC design criteria (3.2). The tagging tool consists of 5 boxes, namely 'the Document Submission Box', 'the Data Criteria Tagging Box', 'the Learner Criteria Tagging Box', 'the Metadata Tagging Box' and 'the Learner Metadata Tagging Box'.

'The Document Submission Box' is for submitting each item of raw data which has already been transcribed. The text box is associated with a specific code assigned for each learner. This code is called the 'Learner Identifier Number' or 'LIN' and it is used as a learner ID. If a learner has more than one text, the LIN will remain the same for all his/her texts; therefore all the texts provided by one learner are labeled identically, and it is also possible to filter the data based on the LIN. Figure 5.5 shows the document submission box in the DSMT.

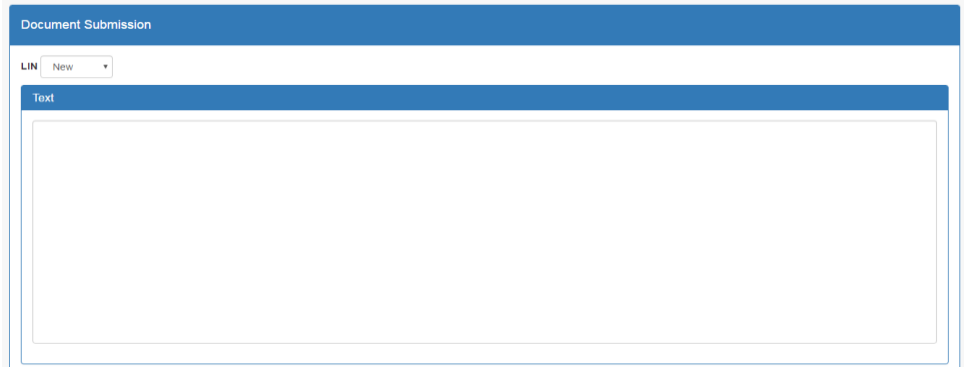



Figure 5.5: The DSMT document submission box

Based on the SFLC Data Criteria discussed in 3.2.2, 'the Data Criteria Tagging Box' provides tagging items in text type (written, spoken, mixed, media, N/A), genre (descriptive, narrative, argumentative, discursive, N/A), task type (composition, interview, creative writing, exam, essay, N/A). The tags are added to the text following the insertion of the raw text into the text submission box. Figure 5.6 shows the DSMT data criteria tagging box.


Figure 5.6: The DSMT data criteria tagging box

'The **Learner Criteria Tagging Box'** provides 3 types of tagging items related to the learner criteria, as introduced in 3.3.2.1. Each raw text is tagged with the learner's first language (which is Serbian by default), target language (which is Persian by default) and proficiency level (A1, A2, B1, B2, C1, C2, N/A). Figure 5.7 shows the tagging boxes for the learner criteria.


Figure 5.7: The DSMT learner criteria tagging box

‘The Text Metadata Tagging Box’ provides 9 metadata tagging items based on the SFLC text metadata introduced in 3.3.2.2. The text metadata includes tagging items for text title, year of production, country of production (Serbia by default), city of production (Belgrade by default), where produced (home, class, exam session, N/A), timing (free, restricted, N/A), reference use (yes, no, N/A), grammar book use (yes, no, N/A), and dictionary use (yes, no, N/A). Figure 5.8 shows the tagging box for the text metadata.

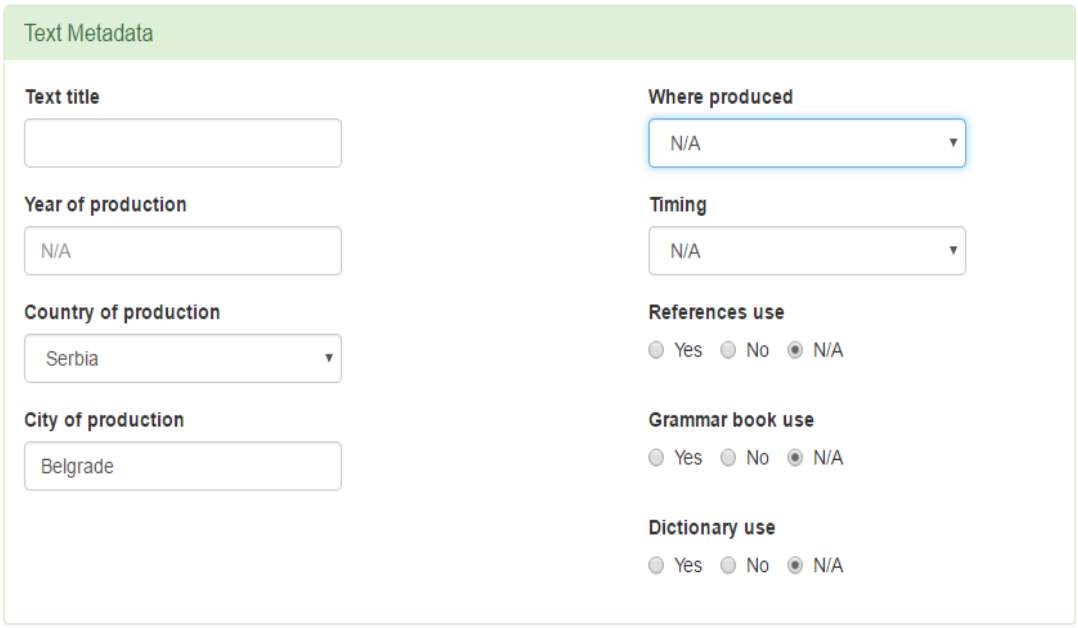



Figure 5.8: The DSMT text metadata box

‘The Learner Metadata Tagging Box’ submits the metadata related to the learners to the corpus. In this part, each raw text is tagged with 8 tagging items, namely, age, gender (male, female, N/A), nationality (Serbian by default), number of languages spoken, number of years learning Farsi, general level of education (ST, BA, MA, PhD, N/A), major and educational institute (where learning Persian, which in this research is limited to the Iranian Cultural Center or Belgrade University Faculty of Philology). Figure 5.9 shows the tagging box for the learner metadata.

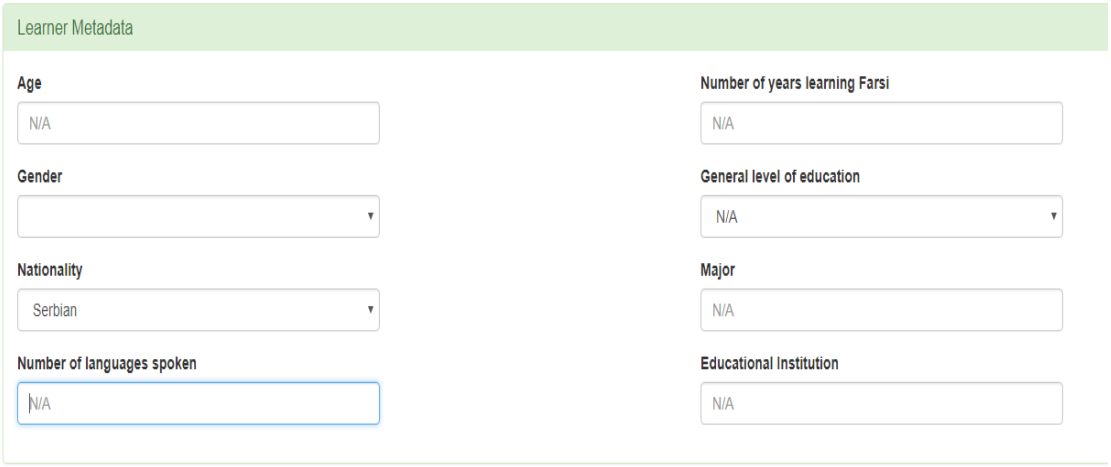



Figure 5.9: The DSMT learner metadata tagging box

### 5.2.2 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 (4.2.2). 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 5.10 shows the ETT text box and its function.

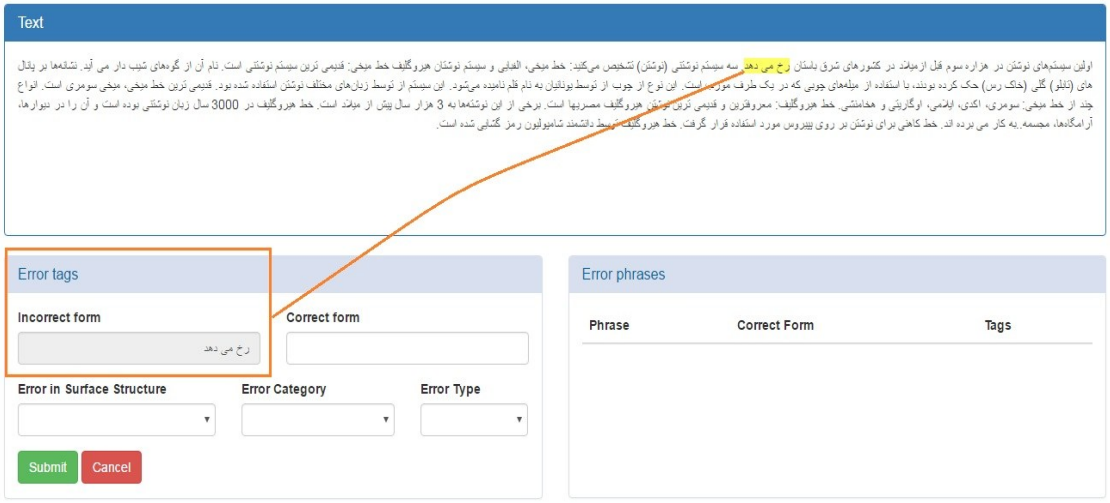



Figure 5.10: The ETT text box

'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 (4.2.2). 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 5.11 shows the ETT error tags box.

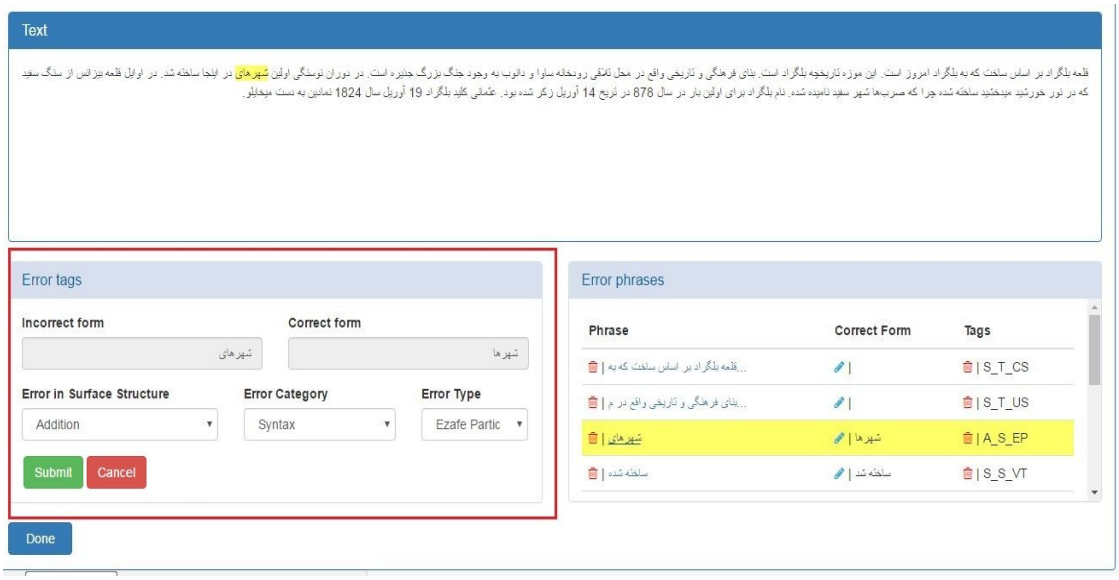



Figure 5.11: The ETT error tags box

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 5.12 shows the ETT error phrase box. The annotation process will be completed by the annotator pressing 'Done' at the bottom.

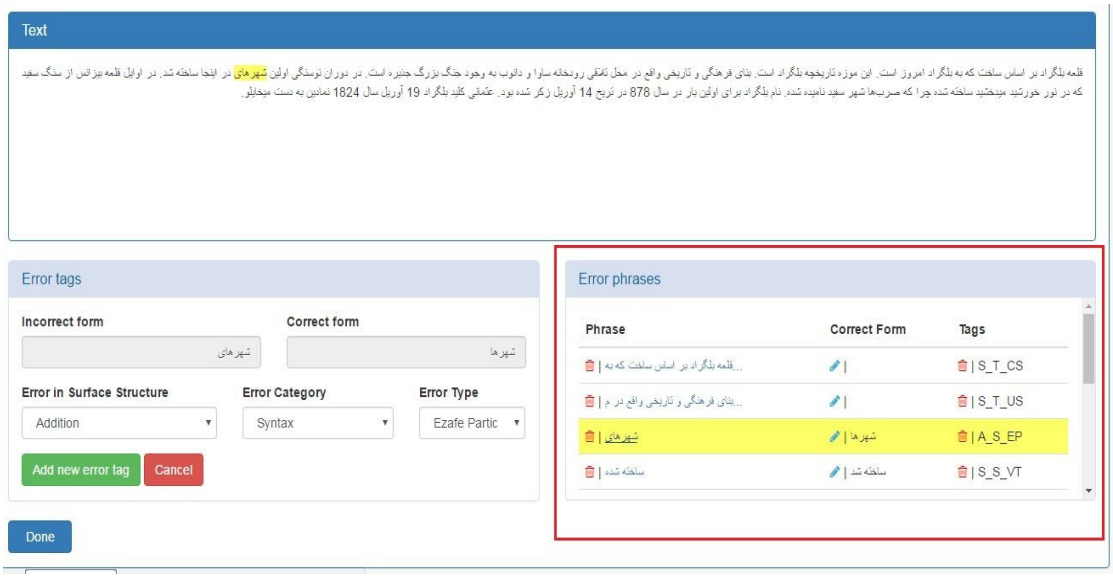



Figure 5.12: The ETT error phrase box

### 5.2.3 The Filter and Search Tool

This tool enables filtering the metadata and/or errors for the purpose of searching specific errors or words/phrases in the whole corpus. The tool is developed in two separate sections: 'The Error Filter and Search' and 'The Word Filter and Search'.

The FST consists of 'the filters' and the 'error box', which show error occurrence in the corpus. There are six groups of filters, which enable the user to obtain specific data by selecting from among the filter variables. On applying the filters, the results are shown in the 'error box', which consists of a document code, the context of the specific errors, and the error codes (i.e. the error tags). The 'data box' shows the total errors and results based on the filters. Figure 5.13 shows the error page in the filter and search tool.


Figure 5.13: The error page in the FST

The word filter and search component is designed to assist with the free search for a word or phrase throughout the whole corpus, as well as with searches using the filters to find the occurrence of a specific word/phrase in the corpus. This tool is made up of four parts: the search box, the data box, the filters' and the results box. Four groups of filters, namely, data criteria, learner criteria, text metadata and learner metadata, can be used selectively to find a specific word/phrase occurrence. The results box shows the word/phrase in context, with the associated document codes. The data box shows the number of total searched queries and the number of results based on the filters. Figure 5.14 shows the word page in the filter and search tool.

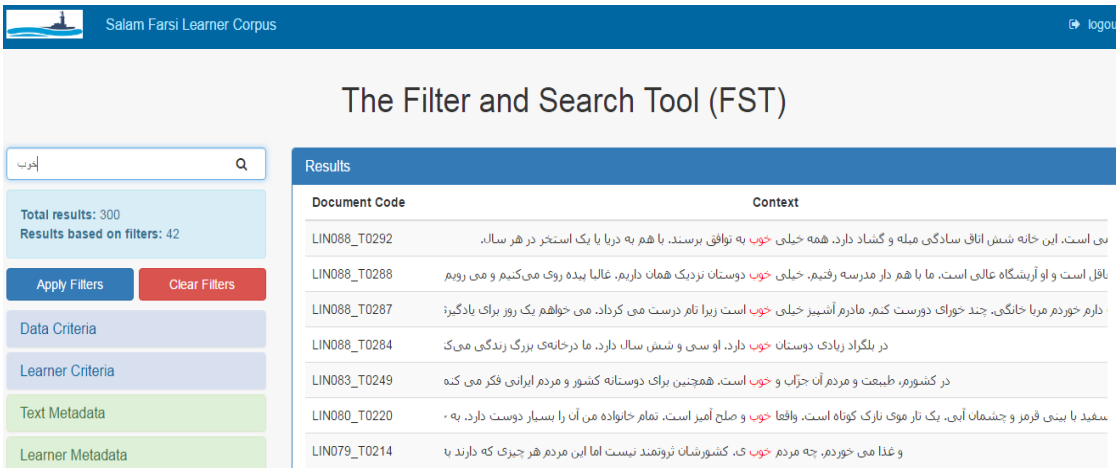



Figure 5.14: The word page in the FST

### 5.2.4 The Data Statistics Tool

This tool is designed for the purpose of collecting and sorting the corpus numerical data for various types of data analysis. The DST mainly provides the numerical statistics regarding the distribution and frequency of error tags in the whole corpus, presenting the results in diagrams. The statistics can be used in analysis of the corpus data for measuring learner performance at different levels and for obtaining an overview of the learners' linguistic strengths and weaknesses by identifying the most common errors they make. The tool is enhanced by four groups of metadata filters (data criteria, learner criteria, text metadata, and learner metadata) and one error filter (error tag criteria). The filters allow a flexible selection of the variables to obtain the specific data statistics required. By default, the diagrams show the latest frequency of error tag distribution in the surface structure, error domain and error type. A data box is also provided to show the total errors and the number of results based on the selected filters. In addition to the diagrams, the tool also enables the downloading of the statistics in XLS (Microsoft Excel) format through the download results link. Figure 5.15 shows the Data Statistics Tool.


Figure 5.15: The data statistics tool

## 5.3 The SFLC Software Application

The SFLC software is a web-based application, i.e. the users can access the corpus via a web browser. Therefore, the corpus is searchable online and at the disposal of researchers. The software application uses the Client-Server model, in which the data processing is done by connecting the browser to the server. The SFLC uses the PostgreSQL database which is set on its server. The PostgreSQL[4] is a type of 'relational database', which is specifically used for storing, querying and maintaining the data. The database is set on an Ubuntu Server 14.04 bit (1024 MB Ram memory and 20 GB of hard drive space).

### 5.3.1 The SFLC Database Structure

The database structure consists of four main tables which function as follows:

1. **The Learner Table**: the learner metadata and a specific code are assigned for each leaner which is called the Learner Identifier Number (LIN).
2. **Document Table**: the table contains the texts and metadata for each document, the raw 'learner _id' shows the connections between this table and the leaner table.
3. **Error Phrase Table**: in this table, the specific code for the error (ID), error text, starting place of the error (based on the number of characters), the document ID where the error is registered and the correct form of the error are entered and saved.
4. **Error Tag Table**: the error tags are specified in three levels (surface structure, domain and type).

Figure 5.16 shows the structure of the SFLC database

[4] PostgreSQL is an object-relational database management system (ORDBMS), developed at the University of California at Berkeley Computer Science Department. POSTGRES pioneered many concepts that only became available in some commercial database systems much later.

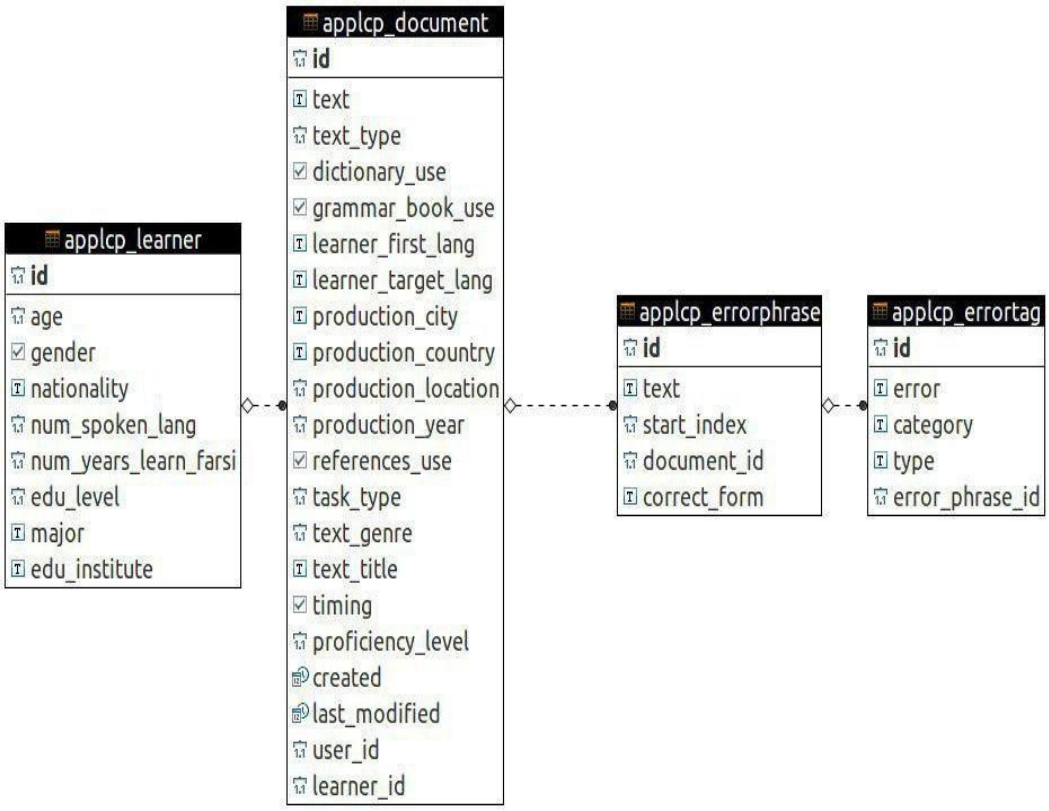



Figure 5.16: The structure of SFLC database

### 5.3.2 The SFLC Data Retrieval

As described in 5.3.1, the data is stored on the database and the frequency of errors is counted by counting the tags registered on the database. The following figures (5.18 and 5.19) show the error tag registration on the database. Figure 5.17 shows that the error phrase (ID 243) is registered on document number 243.

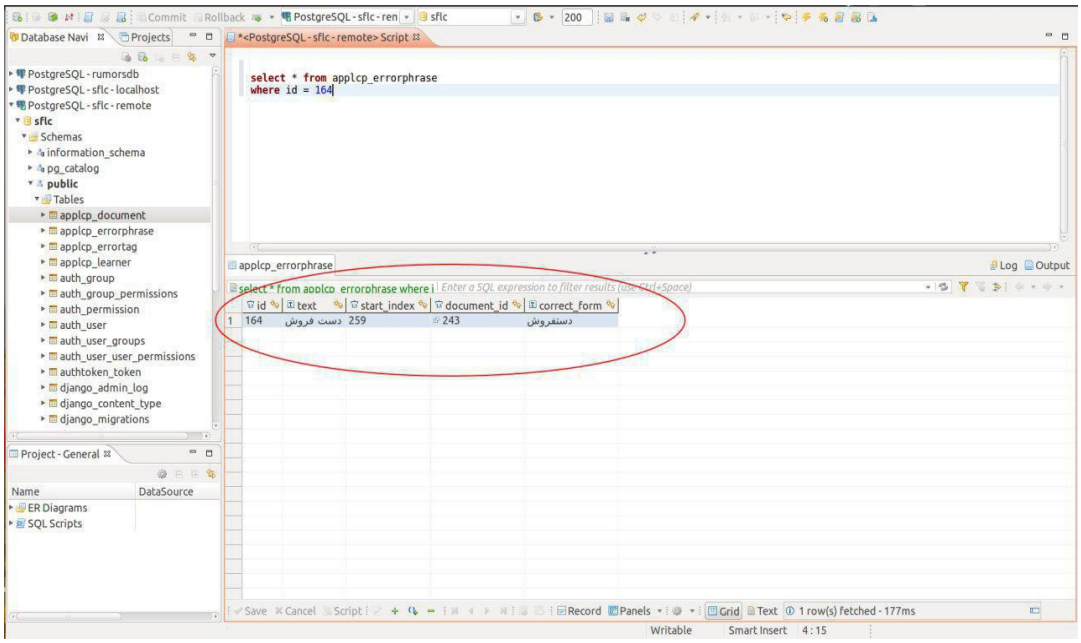



Figure 5.17: Error phrase registration on the database

Figure 5.18 shows that two errors have been registered for the error phrase with ID number 243.

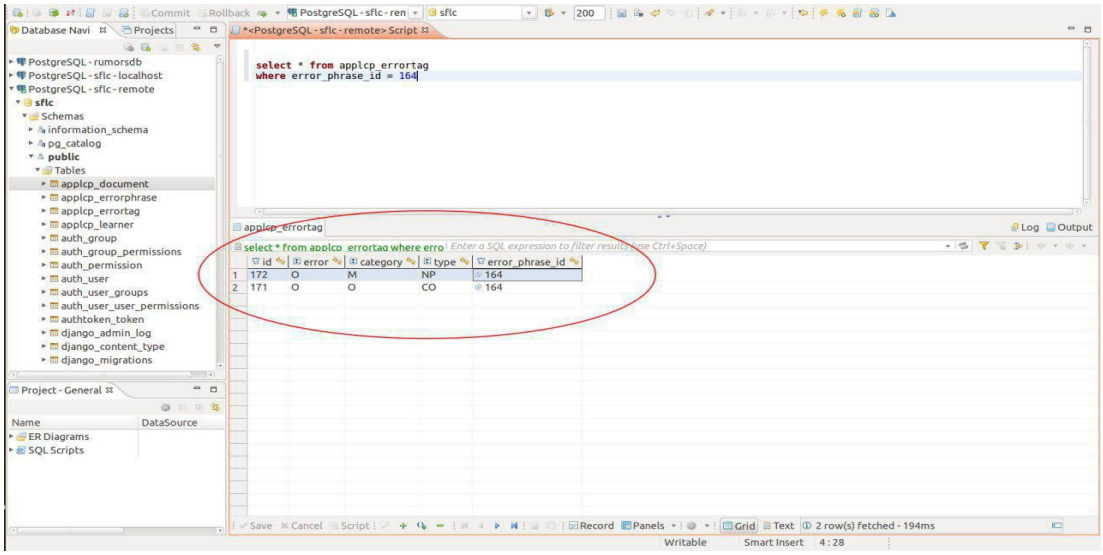



Figure 5.18: Error tags registration on the database

This example indicates that the FST (see 5.2.3) which functions as a query engine allows the retrieval of the data based on the ID number.

### 5.3.3 The File Export Operation

The file export operation enables the users to generate and download files from the database. The process starts with retrieving all the fields and records, i.e. a text with its metadata and error annotations, from the database. The operation constructs two formats: a text format (.txt) and a PDF format (.pdf). The files can be generated and exported either for one specific LIN or for the entire corpus. Figure 5.19 show the file export operations on the website and Figure 5.20 shows a PDF file exported from the corpus database.

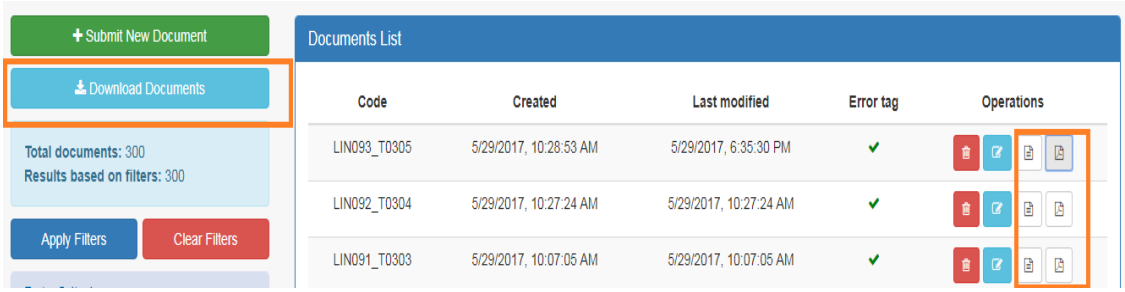
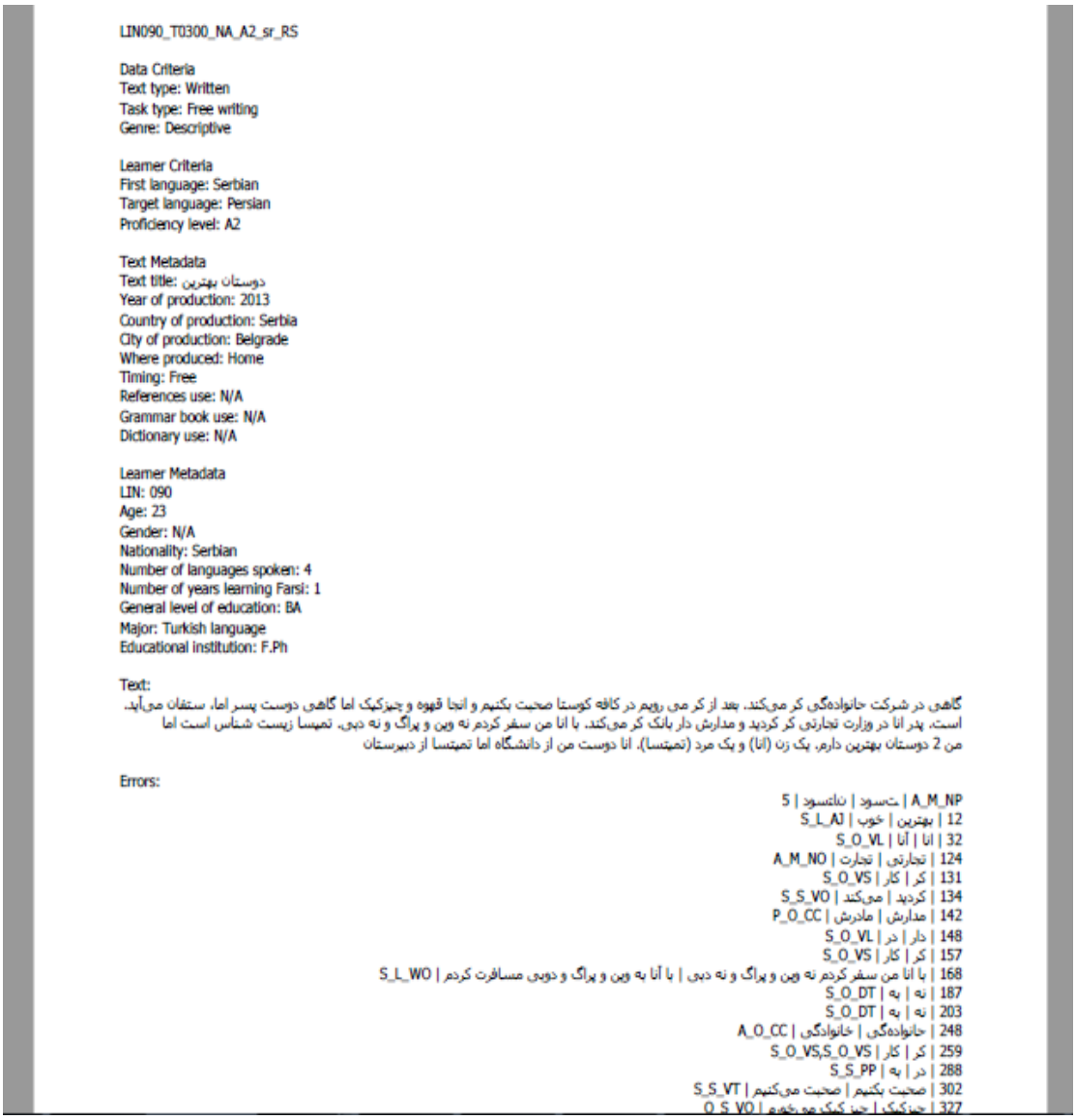



Figure 5.19: File export operations

The following (Figure 5.20) is an example of data retrieval by exporting the file in PDF format.



LIN090_T0300_NA_A2_sr_RS

Data Criteria
Text type: Written
Task type: Free writing
Genre: Descriptive

Learner Criteria
First language: Serbian
Target language: Persian
Proficiency level: A2

Text Metadata
Text title: دوستان بهترین
Year of production: 2013
Country of production: Serbia
City of production: Belgrade
Where produced: Home
Timing: Free
References use: N/A
Grammar book use: N/A
Dictionary use: N/A

Learner Metadata
LIN: 090
Age: 23
Gender: N/A
Nationality: Serbian
Number of languages spoken: 4
Number of years learning Farsi: 1
General level of education: BA
Major: Turkish language
Educational institution: F.Ph

Text:
گاهی در شرکت خانواده‌گی کر می‌کند. بعد از کر می رویم در کافه کوستا صحبت بکنیم و انجا قهوه و چیزکیک اما گاهی دوست پسر اما، ستفان می‌آید.
است. پدر انا در وزارت تجارتی کر کردید و مدارش دار بانک کر می‌کند. با انا من سفر کردم نه وین و پراگ و نه دبی. تمیسا زیست شناس است اما
من 2 دوستان بهترین دارم. یک زن (انا) و یک مرد (تمیتسا). انا دوست من از دانشگاه اما تمیتسا از دبیرستان

Errors:

A_M_NP | ـتسود | نداتسود | 5
S_L_AJ | خوب | بهترین | 12
S_O_VL | آنا | انا | 32
A_M_NO | تجارت | تجارتی | 124
S_O_VS | کار | کر | 131
S_S_VO | می‌کند | کردید | 134
P_O_CC | مادرش | مدارش | 142
S_O_VL | در | دار | 148
S_O_VS | کار | کر | 157
S_L_WO | با آنا به وین و پراگ و دوبی مسافرت کردم | با انا من سفر کردم نه وین و پراگ و نه دبی | 168
S_O_DT | به | نه | 187
S_O_DT | به | نه | 203
A_O_CC | خانوادگی | خانواده‌گی | 248
S_O_VS,S_O_VS | کار | کر | 259
S_S_PP | به | در | 288
S_S_VT | صحبت می‌کنیم | صحبت بکنیم | 302
O_S_VO | چیزکیک می‌خوریم | چیزکیک | 327

Figure 5.20: The generated file from the database

The generated file consists of 7 parts as follows: the LIN, data criteria, learner criteria, text metadata, learner metadata, the text and errors (see Appendix)

### 5.3.4 The SFLC Software Developers

As shortly mentioned in 5.1, the technical parts of the corpus, i.e. setting up the data-base, programming and producing the software tools and interface, were

constructed by a group of freelance software developers in Iran. Since the technical team, as general application developers, had no experiences in developing learner corpora, the corpus structure, webpages, software and tool functionalities as well as the graphical design were planned and suggested by the researcher and subsequently ordered to the developers. It should be noticed and confirmed that the contribution of the software developers in the SFLC is only limited to develop the technical parts as planned and ordered by the researcher. The SFLC software tools and the interface are authentic and original productions and all rights for the intellectual property are reserved for the researcher as well as the Faculty of Philology, University of Belgrade.

# 6. The SFLC Error Distribution and Analysis

**Chapter Summary**

This chapter describes the statistics and results obtained by means of the SFLC Data Statistics Tool in terms of the distribution and frequency of errors in the corpus as well as a comparison of the proficiency levels. The results of the error frequency distribution are introduced and discussed in three error categories according to the SFLC Error Tagset. The error frequency distribution is also presented separately for each level of proficiency and the overall distribution of high-frequency errors is introduced.

## 6.1 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 (see section 1.3). To achieve this aim, after developing the SFLC Error Tagset (described in section 4.2.2) and setting up the corpus software and tools (introduced in section 5.2), 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. This section presents the frequency of error distributions for each level of annotation, namely, errors in the surface structure, error domain and error type.

### 6.1.1 Error Frequency Distribution in the Surface Structures

Based on the SFLC error taxonomy, the first level of error annotation relates to the surface structure which consists of errors in four groups: addition, omission, substitution and permutation. Based on the total of 2,767 error tags counted for the 300 documents (texts) in the corpus, the frequency distribution of errors at surface structure level is listed in Table 6.1.

Table 6.1:Error tag distribution in the surface structure

| Errors in the Surface Structure | Abbr. | Absolute Frequency | Relative Frequency |
|---|---|---|---|
| Addition | A | 388 | 14% |
| Omission | O | 834 | 31% |
| Substitution | S | 1,198 | 45% |
| Permutation | P | 256 | 10% |
| Total | | 2,767 error tags | 100% |

The majority of the errors produced by the Serbian learners of Farsi at surface structure level involve substitution errors, i.e. the mis-selection of linguistic elements, which make up almost half of the total number of error tags (45%), while the smallest number were counted for the permutation or misordering of elements. The following are some examples of the error tags at surface structure level.

**(1) Addition**

Document ID: LIN022_T0088

*بلگراد از <A_S_PP> دو میلیون جمعیت دارد

*belgrâd az < A_S_PP> do miljun djamijjat darad

*The error tag:* < A_S_PP> Addition_Syntax_Preposition

Correct Form: [Ø]

Gloss:

* belgrâd az < A_S_PP> do miljun djamijjat dar=ad

Belgrade [from] < A_S_PP> [Ø] two million population has.PRS.3rd

"The population of Belgrade is 2 million people"

*Error Description*: The learner has added a preposition which is redundant.

**(2) Omission**

Document ID: LIN075_T0196

* هوا سرد و برف <O_M_AJ> است

*hâvâ sard va barf < O_M_AJ > ast.

*The error tag:* < A_S_PP> Omission_Morphology_Adjective

Correct Form: [barfi]

Gloss:

* hâvâ sard va [barf] < O_M_AJ > [barf=i] ast

weather cold and snow < O_M_AJ > [snow.adj.suffix] is.

"The weather is cold and snowy"

*Error Description*: The learner has omitted the adjective suffix.

**(3) Substation**

Document ID: LIN054_T0151

به آرایشگاح {آرایشگاه} <S_O_CC> رفتم

be ârâješgâh raftim

*The error tag:* < S_O_CC> Substitution_Orthography_Consonant character(s)

Correct Form: [آرایشگاه]

Gloss:

* be [ârâješgâh] < S_O_CC > raft=am

to [barber shop] < S_O_CC > [ârâješgâh] go.PST.1sg

"I went to the barber shop"

*Error Description*: The learner mis-selected the consonant character in the word, due to the multiple forms of consonants in Persian (see 2.8.1.4). The consonant /h/ has two forms in writing, [ح] and [ه].

**(4) Permutation**

Document ID: LIN011_T0040

دوست بهترین <P_L_WO> کی است؟

*dust behtarin < P_L_W > ki ast?

*The error tag:* < P_L_WO > Permutation_Lexis_Word Order

Correct Form: [behtarin dust]

Gloss:

* dust behtarin < P_L_WO > ki ast?

Best friend [dust behtarin] < P_L_WO > [behtarin dust] who is.COP?

"Who is (your) best friend?"

*Error Description*: The learner misordered the superlative adjective which comes before the noun in Persian.

Figure 6.1 shows the same data (absolute frequencies) graphically.

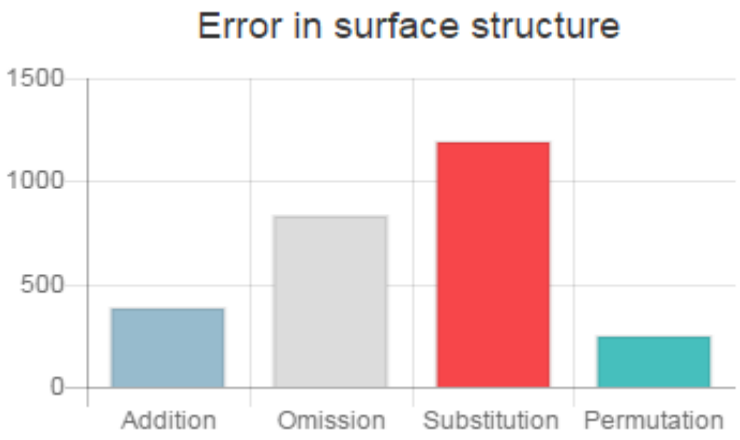



Figure 6.1: The distribution of errors in the surface structure

### 6.1.2 Error Frequency Distribution in the Error Domains

The second level of error annotation in the SFLC relates to error domain, which consists of 5 categories: orthography, syntax, lexis, morphology and style. The error frequency distributions for the error domains are listed in Table 6.2.

Table 6.2: The frequency distribution of errors in the error domains

| Error Domains | Abbr. | Absolute frequency | Relative frequency |
|---|---|---|---|
| Orthography | O | 833 | 31% |
| Morphology | M | 287 | 11% |
| Syntax | S | 763 | 28% |
| Lexis | L | 658 | 25% |
| Style | T | 135 | 5% |
| Total | | 2,767 error tags | 100% |

The majority of the error tags belong to the domain of orthography and the fewest to style. It should be noted that errors in syntax were in second place after orthography, which means the main errors in linguistic forms were tagged in syntax. Errors in lexis are also high in frequency, comprising a quarter of the errors at domain level.

Figure 6.2 illustrates the distribution of error domain tags (absolute frequencies) sorted according to their occurrence.

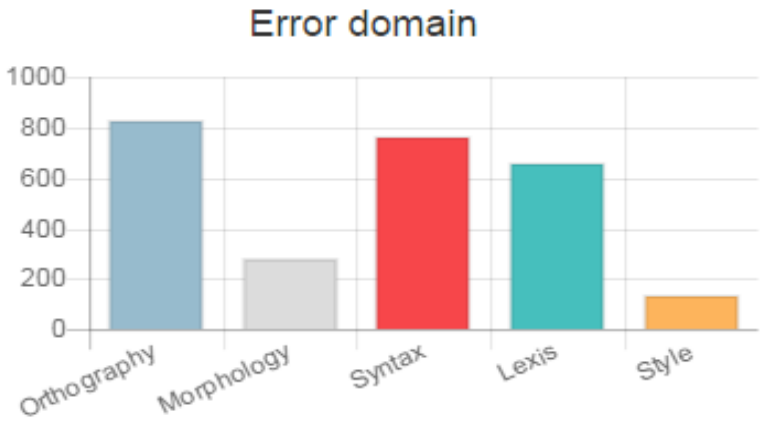



Figure 6.2: The distribution of error domains

### 6.1.3 Error Frequency Distribution in the Error Types

On the basis of the SFLC error taxonomy (4.2.1), error Type includes 22 types of error which determine the learners' errors in detail. This is the third level of error annotation and the frequency distribution of 2,767 error tags for 300 documents (texts) detected is shown in Table 6.3.

Table 6.3: The distribution of error tags sorted according to their occurrence

| No | Error Types | Abbr. | Absolute frequency | Relative frequency |
|---|---|---|---|---|
| 1 | Consonant character(s) | CC | 325 | 12% |
| 2 | Long Vowel character(s) | VL | 285 | 11% |
| 3 | Short Vowel character(s) | VS | 67 | 3% |
| 4 | Connections | CO | 32 | 1% |
| 5 | The Ezâfe Particle | EP | 87 | 3% |
| 6 | Dots | DT | 94 | 4% |
| 7 | Adjective | AJ | 92 | 3% |
| 8 | Noun-Plural | NP | 72 | 3% |
| 9 | Noun Other | NO | 175 | 7% |
| 10 | Pronoun | PR | 106 | 4% |
| 11 | Preposition | PP | 164 | 6% |
| 12 | Postposition (ra) | PO | 105 | 4% |
| 13 | Conjunction | CN | 52 | 2% |
| 14 | Verb agreement | VA | 81 | 3% |
| 15 | Verb tense | VT | 107 | 4% |
| 16 | Verb Other | VO | 233 | 9% |
| 17 | Adverb | AD | 32 | 1% |
| 18 | Word Order | WO | 245 | 9% |
| 19 | Word Selection | WS | 134 | 5% |
| 20 | Phrase selection | PS | 67 | 3% |
| 21 | Cohesion | CS | 65 | 2% |
| 22 | Unclear style | US | 56 | 2% |
| | Total | | 2,767 error tags | 100% |

The distribution of error type tags is illustrated in Figure 6.3 which provides a clear view of the types of error made by the Serbian learners of the Persian language in their texts. Based on the absolute frequencies, the most frequent error tag, with more than 300 tags out of 2,764 errors, was for consonant character and the least common for connection and adverb with only 32 error tags. In terms of the domain of the errors, both the high and less –frequent error types lie within the domain of orthography, which will be discussed in 6.3.2. As shown in Figure 6.3, the errors in long vowel character, word order, verb other, noun other and preposition are among the high frequent error types made by Serbian learners of the Persian language.

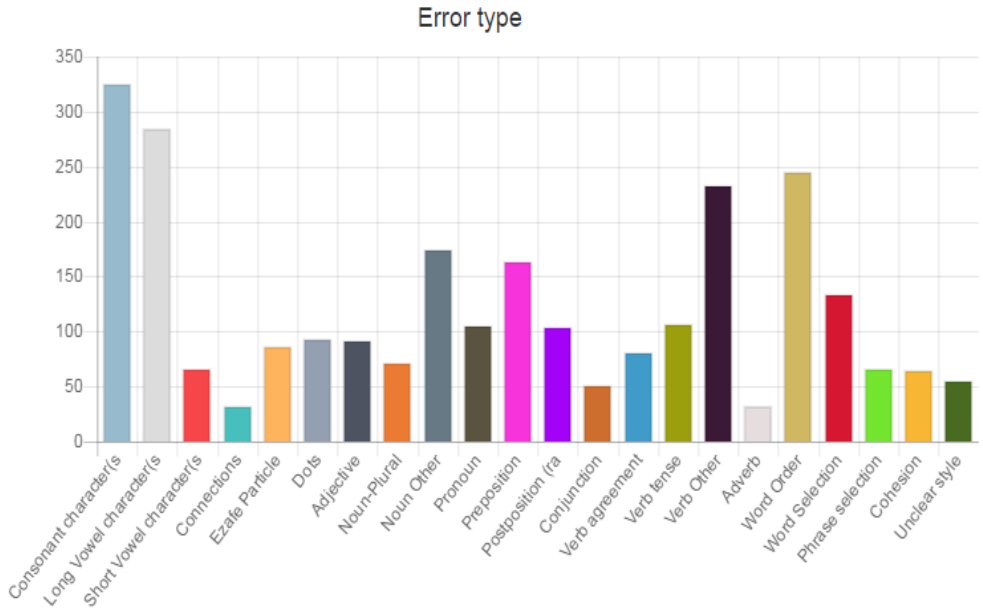



Figure 6.3: The distribution of error types

### 6.1.4 Overall Distribution of High-Frequency Errors in the SFLC

Based on the statistics, the 10 major error types of the Serbian learners of the Persian language in the SFLC are listed in Table 6.4. 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 6.4: The major error types in the SFLC

| | Error Type | Absolute frequency | Relative frequency | Error Domain |
|---|---|---|---|---|
| 1 | Consonant character(s) | 325 | 12% | Orthography |
| 2 | Long Vowel character(s) | 285 | 11% | |
| 3 | Word Order | 245 | 9% | Lexis |
| 4 | Verb Other | 233 | 9% | Syntax |
| 5 | Noun Other | 175 | 7% | Morphology |
| 6 | Preposition | 164 | 6% | Syntax |
| 7 | Word Selection | 134 | 5% | Lexis |
| 8 | Verb tense | 107 | 4% | Syntax |
| 9 | Pronoun | 106 | 4% | Morphology |
| 10 | Postposition (ra) | 105 | 4% | 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. The major errors will be discussed in 6.3 in details.

## 6.2 A Comparison of the Error Tag Distribution across Proficiency Levels

Based on the learner criteria (3.2.3), the SFLC contains 4 proficiency levels: A2, B1, B2 and C1.. Table 6.5, which repeats the data mentioned in 3.3.1.4, shows the distribution of the submitted documents and the number of words counted for each level.

Table 6.5: The distribution of the submitted documents in the proficiency levels

| Proficiency Levels | Submitted Documents | % of Total | Total Words |
|---|---|---|---|
| A2 | 62 | 21% | 5,575 |
| B1 | 81 | 27% | 7,284 |
| B2 | 101 | 33% | 9,082 |
| C1 | 56 | 19% | 5,035 |
| Total | 300 | 100% | 26,976 |

As Table 6.5 illustrates, the majority of the texts were submitted for level B2, with the smallest number for level A2. In this section we will compare the distribution of error tags in the surface structure, error domain and error type across proficiency levels.

### 6.2.1 Error Distribution in the Surface Structure based on Proficiency Levels

The error tag distribution in the surface structure based on the proficiency levels is shown in Table 6.6.

Table 6.6: Distribution of error tags in the surface structure for the proficiency levels

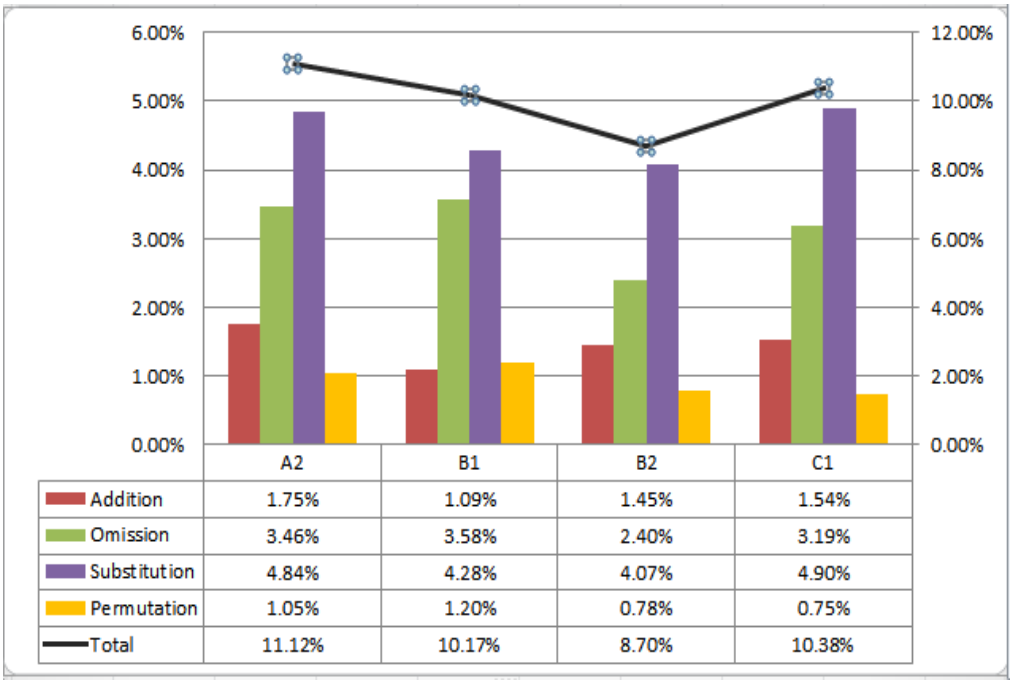
| Errors in the Surface Structure/ Levels | A2 | | B1 | | B2 | | C1 | |
|---|---|---|---|---|---|---|---|---|
| Total Words | 5575 | | 7284 | | 9082 | | 5035 | |
| | Total errors | % of total words | Total errors | % of total words | Total errors | % of total words | Total errors | % of total words |
| Addition | 98 | 1.75% | 80 | 1.09% | 132 | 1.45% | 78 | 1.54% |
| Omission | 193 | 3.46% | 261 | 3.58% | 218 | 2.40% | 161 | 3.19% |
| Substitution | 270 | 4.84% | 312 | 4.28% | 370 | 4.07% | 247 | 4.90% |
| Permutation | 59 | 1.05% | 88 | 1.20% | 71 | 0.78% | 38 | 0.75% |
| Total | 620 | 11.12% | 741 | 10.17% | 791 | 8.70% | 524 | 10.38% |

Based on the statistics and as illustrated in Table 6.6, errors in surface structure mostly decreased between level A2 and C1, thus indicating an improvement in language skills.

Errors in substitution are high in frequency and although they dropped from level A2 to B1 and B2, they increased at tC1 level which can be explained.

The information about the error domains obtained through the DST filter settings indicates that the majority of substitution errors tagged in A2 and B1 lie in the domain of orthography, while in B2 they pertain to lexis and for C1 the majority are related to syntax. It can be concluded that although errors in substitution are high at all levels and even record an increase at C1 level, the error domain is different and this should be noted when reviewing and studying the results.

Using a second vertical axis in the graph, Figure 6.4 shows the total errors in surface structure as well as the distribution of domain errors across the proficiency levels. The majority of errors in surface structure were tagged for substitution, followed by omission, addition and permutation respectively.




| | A2 | B1 | B2 | C1 |
|---|---|---|---|---|
| Addition | 1.75% | 1.09% | 1.45% | 1.54% |
| Omission | 3.46% | 3.58% | 2.40% | 3.19% |
| Substitution | 4.84% | 4.28% | 4.07% | 4.90% |
| Permutation | 1.05% | 1.20% | 0.78% | 0.75% |
| Total | 11.12% | 10.17% | 8.70% | 10.38% |

Figure 6.4: The total errors and distribution of tags in surface structure across the proficiency levels

### 6.2.2 Error Distribution in the Error Domains based on the Proficiency Levels

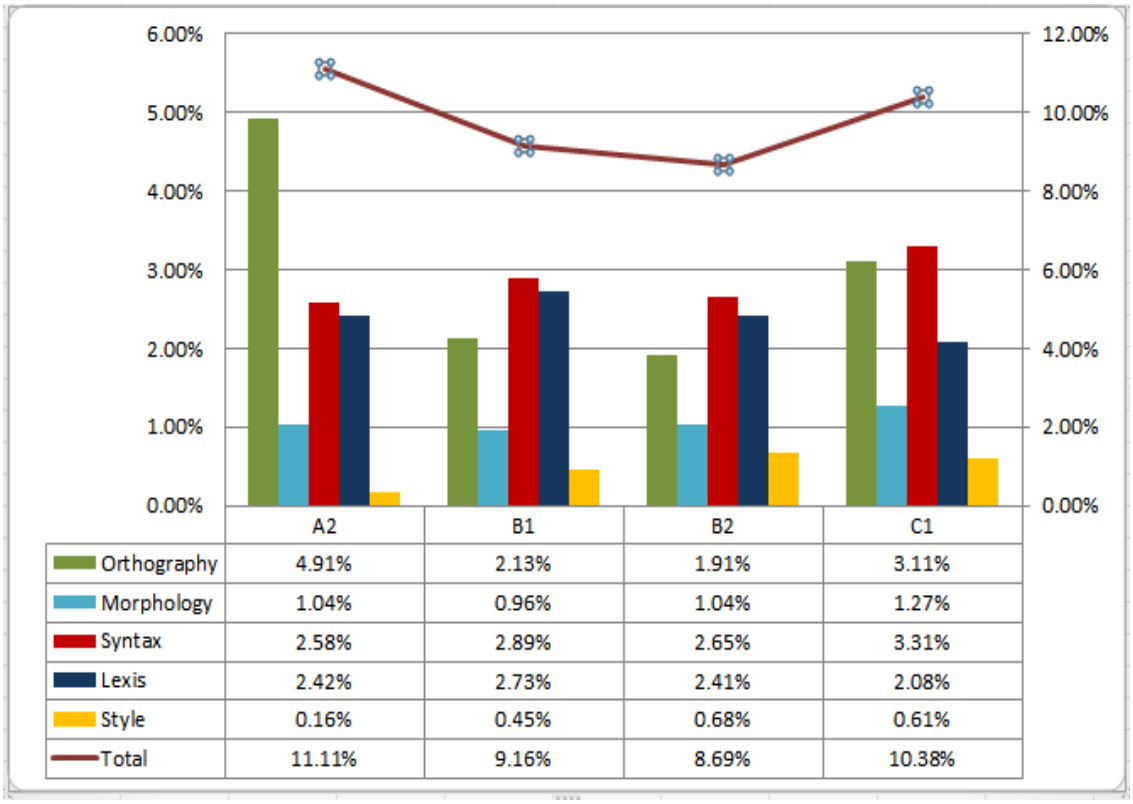
Table 6.7 shows the distribution of error tags in the error domains based on the proficiency levels. Errors dropped from 11.11% in level A2 to 9.16%, 8.69% and 10.36% in levels B1, B2 and C1 respectively.

Table 6.7: Distribution of error tags in the error domains based on the proficiency levels

| **Error Domain/ Levels** | **A2** | | **B1** | | **B2** | | **C1** | |
|---|---|---|---|---|---|---|---|---|
| Total Words | 5,575 | | 7,284 | | 9,082 | | 5,035 | |
| | Total errors | % of total words | Total errors | % of total words | Total errors | % of total words | Total errors | % of total words |
| Orthography | 274 | 4.91% | 228 | 2.13% | 174 | 1.91% | 157 | 3.11% |
| Morphology | 58 | 1.04% | 70 | 0.96% | 95 | 1.04% | 64 | 1.27% |
| Syntax | 144 | 2.58% | 211 | 2.89% | 241 | 2.65% | 167 | 3.31% |
| Lexis | 135 | 2.42% | 199 | 2.73% | 219 | 2.41% | 105 | 2.08% |
| Style | 9 | 0.16% | 33 | 0.45% | 62 | 0.68% | 31 | 0.61% |
| Total | 620 | 11.11% | 741 | 9.16% | 791 | 8.69% | 524 | 10.38% |

The distribution of the error domain tags, as illustrated in Figure 6.5, illustrates graphically that errors in orthography generally decreased within levels, from 4.91% in A2 to 3.11% in C1 and shows that the learners gradually developed their dictation skills. Errors in syntax rose from 2.58% at level A2 to 3.31% at level C1. The rise in syntactic errors between levels could be explained by comparing the error types at lower and upper levels, i.e. A2 and C1.The statistics obtained via the DST show that the top error type at level A2 in the domain of syntax is preposition (PP) while at C1 level, it relates to verb category (VO and VT). These changes may indicate the use of more complex structures at upper

levels. The same pattern, (i.e. an increase in errors from the lower to the upper levels) is repeated for errors in the domains of morphology and style, which generally increased from level A2 to C1. It is also notable that errors in lexis decreased within the levels, from 2.42% at level A2 to 2.08% at level C1.




| | A2 | B1 | B2 | C1 |
|---|---|---|---|---|
| Orthography | 4.91% | 2.13% | 1.91% | 3.11% |
| Morphology | 1.04% | 0.96% | 1.04% | 1.27% |
| Syntax | 2.58% | 2.89% | 2.65% | 3.31% |
| Lexis | 2.42% | 2.73% | 2.41% | 2.08% |
| Style | 0.16% | 0.45% | 0.68% | 0.61% |
| Total | 11.11% | 9.16% | 8.69% | 10.38% |

Figure 6.5: The total errors and distribution of tags in the domains across the proficiency levels

### 6.2.3 Error Distribution in the Error Types Based on the Proficiency Levels

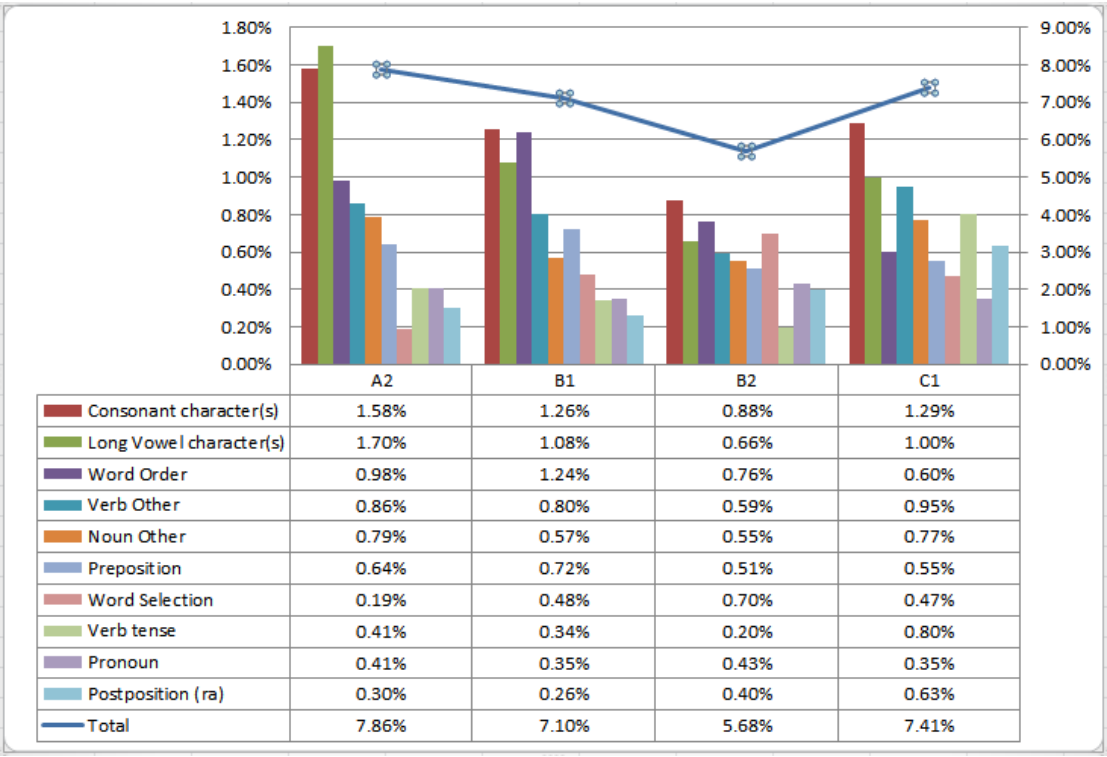
Table 6.8 shows the distribution of 10 major error types based on the proficiency levels. According to the statistics, the frequency of error tags decreased from 7.86% at level A2 to 7.10%, 5.68% and 7.41% at levels B1, B2 and C1 respectively, which may be

interpreted as improvements in the learners' language skills. Figure 6.6 illustrates the distribution of 10 major error tags across the proficiency levels.

Table 6.8: The distribution of the major error tags

| **Error Types/Levels** | **A2** | | **B1** | | **B2** | | **C1** | |
|---|---|---|---|---|---|---|---|---|
| Total Words | 5575 | | 7284 | | 9082 | | 5035 | |
| | Total errors | % of total words | Total errors | % of total words | Total errors | % of total words | Total errors | % of total words |
| Consonant character(s) | 88 | 1.58% | 92 | 1.26% | 80 | 0.88% | 65 | 1.29% |
| Long Vowel character(s) | 95 | 1.70% | 79 | 1.08% | 60 | 0.66% | 51 | 1.00% |
| Word Order | 55 | 0.98% | 91 | 1.24% | 69 | 0.76% | 30 | 0.60% |
| Verb Other | 48 | 0.86% | 59 | 0.80% | 78 | 0.59% | 48 | 0.95% |
| Noun Other | 44 | 0.79% | 42 | 0.57% | 50 | 0.55% | 39 | 0.77% |
| Preposition | 36 | 0.64% | 53 | 0.72% | 47 | 0.51% | 28 | 0.55% |
| Word Selection | 11 | 0.19% | 35 | 0.48% | 64 | 0.70% | 24 | 0.47% |
| Verb tense | 23 | 0.41% | 25 | 0.34% | 19 | 0.20% | 40 | 0.80% |
| Pronoun | 23 | 0.41% | 26 | 0.35% | 39 | 0.43% | 18 | 0.35% |
| Postposition (ra) | 17 | 0.30% | 19 | 0.26% | 37 | 0.40% | 32 | 0.63% |
| Total | 440 | 7.86% | 521 | 7.10% | 543 | 5.68% | 375 | 7.41% |

Figure 6.6 represents the data in Table 6.8 graphically. It clearly shows that errors at level B2 notably decreased compared to the other proficiency levels.




| | A2 | B1 | B2 | C1 |
|---|---|---|---|---|
| Consonant character(s) | 1.58% | 1.26% | 0.88% | 1.29% |
| Long Vowel character(s) | 1.70% | 1.08% | 0.66% | 1.00% |
| Word Order | 0.98% | 1.24% | 0.76% | 0.60% |
| Verb Other | 0.86% | 0.80% | 0.59% | 0.95% |
| Noun Other | 0.79% | 0.57% | 0.55% | 0.77% |
| Preposition | 0.64% | 0.72% | 0.51% | 0.55% |
| Word Selection | 0.19% | 0.48% | 0.70% | 0.47% |
| Verb tense | 0.41% | 0.34% | 0.20% | 0.80% |
| Pronoun | 0.41% | 0.35% | 0.43% | 0.35% |
| Postposition (ra) | 0.30% | 0.26% | 0.40% | 0.63% |
| Total | 7.86% | 7.10% | 5.68% | 7.41% |

Figure 6.6: The distribution of the 10 major error tags across the proficiency levels

### 6.2.4 The Overall Distribution of High-Frequency Errors Based on the Proficiency Levels

In this section the overall distribution of high-frequency errors, 5 major errors, is introduced separately for each proficiency level. The statistics provide a clear view of the error frequencies since they are organized according to the error type, error domain and errors in surface structure.

Table 6.9 shows the distribution of the major errors at A2 level. In total, almost half of the major errors are found in the domain of orthography (43%) which could have been predicted since the learners were developing their writing skills at the basic A2 level. The second major error domain is lexis (16%), followed by errors in the domain of syntax (12%) and morphology (4%) respectively.

Table 6.9: The distribution of major errors in level A2

| | Error Type | Absolute frequency | Relative frequency | Error Domain |
|---|---|---|---|---|
| 1 | Long Vowel character(s) | 95 | 15% | Orthography |
| 2 | Consonant character(s) | 88 | 14% | |
| 3 | Word Order | 55 | 9% | Lexis |
| 4 | Verb Other | 48 | 8% | Syntax |
| 5 | Noun Other | 44 | 7% | Lexis |

Figure 6.7 illustrates the error distribution at level A1 based on the error domains.

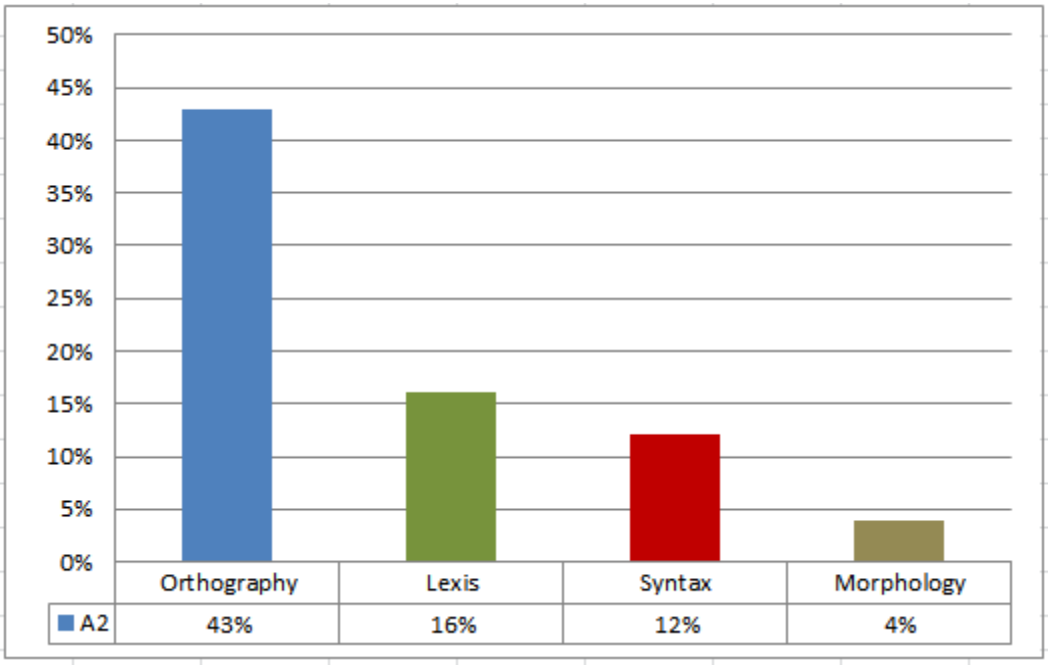



Figure 6.7: The distribution of the major error domains for level A2

Table 6.10 shows the distribution of major errors at level B1. Errors in orthography are still high in frequency accounting for 27% of the total major errors, however, such

errors fell in comparison with level A2 (43%). The second major error domain is lexis (23%) followed by those in the domain of syntax (19%) and morphology (4%).

Table 6.10: The distribution of the major errors at level B1

| | Error Type | Absolute frequency | Relative frequency | Error Domain |
|---|---|---|---|---|
| 1 | Consonant character(s) | 92 | 12% | Orthography |
| 2 | Word Order | 91 | 12% | Lexis |
| 3 | Long Vowel character(s) | 79 | 11% | Orthography |
| 4 | Verb Other | 59 | 8% | Syntax |
| 5 | Preposition | 53 | 7% | |

Figure 6.8 illustrates the error distribution at level B1 based on the error domains.

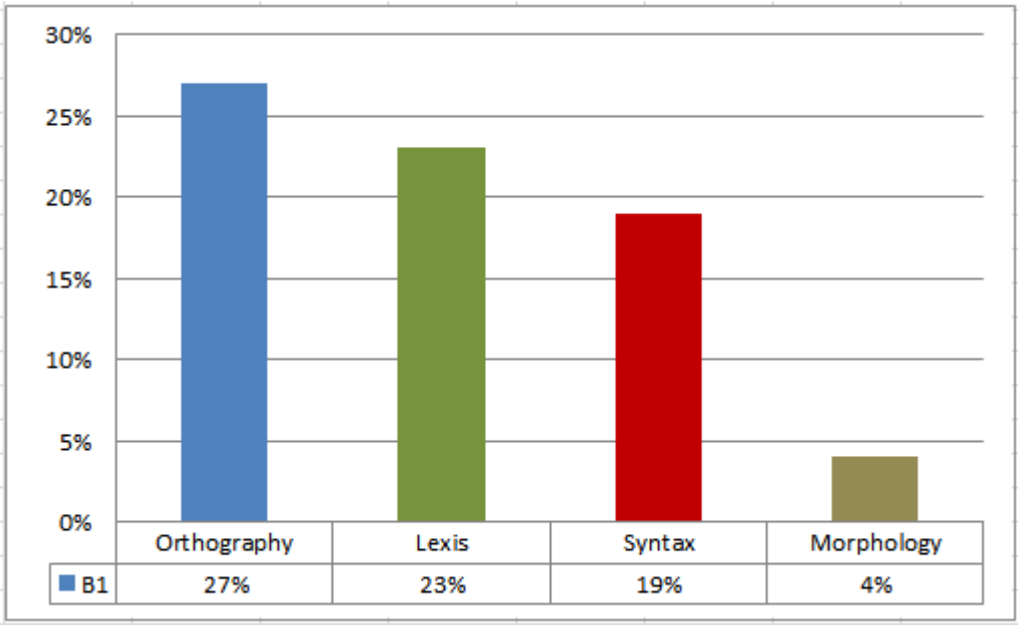



Figure 6.8: The distribution of the major error domains for level B1

Table 6.11 shows the distribution of the major errors at level B2. The major errors at this level are limited to 3 domains: syntax, lexis and morphology. Errors in syntax with 31% of the total major errors are predominant, which could indicate that the learners were developing their syntactic structures. The second major error domain is lexis (23%), followed by errors in the domain of morphology (18%).

Table 6.11: The distribution of the 10 major errors at level B2

| | Error Type | Absolute frequency | Relative frequency | Error Domain |
|---|---|---|---|---|
| 1 | Consonant character(s) | 80 | 10% | Orthography |
| 2 | Verb Other | 78 | 10% | Syntax |
| 3 | Word Order | 69 | 9% | Lexis |
| 4 | Word Selection | 64 | 8% | |
| 5 | Long Vowel character(s) | 60 | 8% | Orthography |

Figure 6.9 illustrates the error distribution at level B1 based on the error domains.

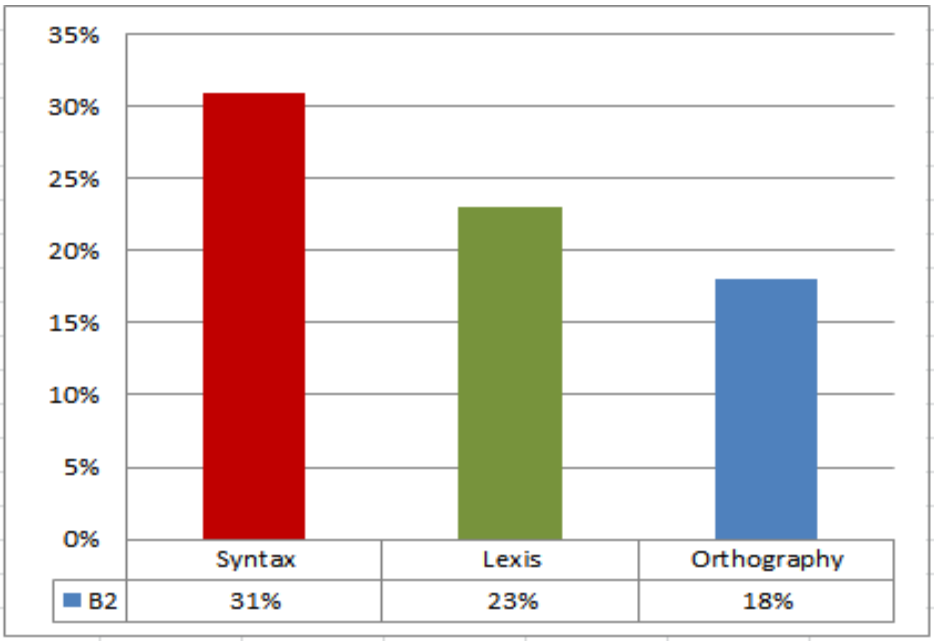



Figure 6.9: The error distribution at level B2 based on the error domains

Table 6.12 shows the distribution of the major errors at level C1. The major errors in this level lie in the domain of syntax, the same as for level B2, accounting for 28% of the total major errors. It is interesting to note that the second major error domain for this level, which is considered an advanced level, is orthography. However, by applying the DST and comparing the errors related to orthography across the levels, it is shown that orthography errors at level C1 differ in terms of surface structure errors. At level C1 orthography-related errors are mostly tagged as omission, while at other levels they were tagged in substitution. The third major error domain is lexis (11%) and errors in the domain of Morphology, just like for the other levels, are fewer, comprising only 7% of the total major errors at level C1.

Table 6.12: The distribution of the10 major errors T level C1

| | **Error Type** | **Absolute frequency** | **Relative frequency** | **Error Domain** |
|---|---|---|---|---|
| 1 | Consonant character(s) | 65 | 12% | Orthography |
| 2 | Long Vowel character(s) | 51 | 10% | |
| 3 | Verb Other | 48 | 9% | Syntax |
| 4 | Verb tense | 40 | 8% | |
| 5 | Noun Other | 39 | 7% | Morphology |

Figure 6.10 illustrates the error distribution at level B1 based on the error domains.

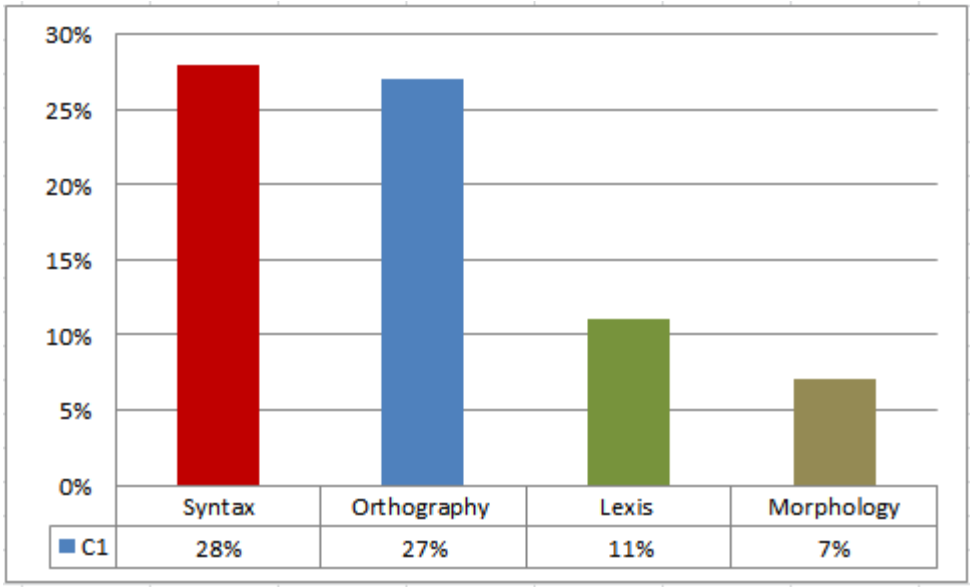



Figure 6.10: The error distribution at level C1 based on the error domains

## 6.3. The Results and Discussion

The frequency distribution of the error tags in the whole corpus and the error distributions based on the proficiency levels were introduced in previous sections (6.1, and 6.2). To discuss 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. Therefore, the major errors in the domain of syntax, lexis, morphology and style are discussed in the group of linguistic errors and those in the domain of orthography are discussed separately.

### 6.3.1 Linguistic Errors in the SFLC

The SFLC linguistic error tags consist of 4 domains and 17 error tags as introduced in 4.2.1. 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 6.13.

Table 6.13: 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 |

**(1) Word Order Errors**

In the SFLC, word order errors are the most frequent linguistic errors made by the Serbian learners of the Persian language. The errors are marked mainly for the domain of lexis, with 228 tags, and permutation, with 215 tags, in the surface structure which means that the learners misordered the lexical items. The specific linguistic error in the Persian language which can be detected based on these error tags (Permutation_Lexis_Word Order) is the Ezâfe construction.

The term Ezâfe, which means 'addition', is an enclitic which is mostly realized by the unstressed short vowel /-e/ or its allophone /je/ which is suffixed to the word that has a final vowel. The Ezâfe mostly appears between a noun and its modifier and is also repeated on subsequent modifiers. The Ezâfe construction is used to show different relations such as possession and addition. In such constructions, the primary noun comes first and is followed by the word modifying it, while the short vowel /-e/ serves to connect them. Ill-formed constructions occur when learners follow their native language structure of possessive or adjectival phrases which differs from Persian. Some evidence of violations of post-nominal modifier order in Ezâfe constructions have been provided as examples below.

(A) Misordering in the Ezâfe Construction (Noun + Noun)

Document Code: LIN091_T0302

* آولا برج <P_L_WO> ۲۰۵ متر ارتفاع دارد

*âvalâ bordj < P_L_WO> 205 mer ertfâ’ dârad.

*The error tag:* < P_L_WO> Permutation_Lexis_Word Order

Correct Form: [برج آولا]

Gloss:

*âvalâ bordj < P_L_WO> 205 metr ertfâ’ dâr=ad

Avala Tower [âvalâ bordj] < P_L_WO> [bordj-EZ âvalâ] 205 meter height have.PRS.3rd

“The height of the Avala Tower is 205 meters.”

*Error Description*: The Ezâfe construction (Noun + Ez + Noun) is incorrect. In the Persian word order system, the modifier usually follows the modified noun and when a noun precedes another noun, the first one receives the Ezâfe enclitic.

(B) Misordering in the Ezâfe Construction (Noun + Adjective)

Document Code: LIN012_T0046

* امروز تعطیلات مذهبی مقدس نیکولا < P_L_WO> است

*emruz ta’atilât mazhabi moqqadas nikolâ < P_L_WO> ast.

*The error tag:* < P_L_WO> Permutation_Lexis_Word Order

Correct Form: [نیکولای مقدس]

Gloss:

*emruz ta’tilât mzahabi moqaadas nikolâ < P_L_WO> ast.

Today holiday-EZ ADJ-religious [Saint Nikola] < P_L_WO> [nikolâ + EZ moqqadas] be-3sg.

“Today is the religious holiday in honour of Saint Nikola.”

*Error Description*: The Ezâfe construction (Noun + Ez + Adjective) is incorrect. In the Persian word order system, adjectives usually follow the nouns

Although Persian follows the word order system in which adjectives usually follow nouns (as explained in example number 2), the word order for superlative adjectives is different and such adjectives are considered pre-modifiers; that is, the adjective precedes the noun without using the Ezâfe construction. This may cause errors since learners follow the general word order for Ezâfe constructions (Noun + Ez + Adjective). Here is an example:

(C) Misordering in superlative constructions (Superlative Adjective + Noun)

Document Code: LIN090_T0295

* خیابان زیباترین <P_L_WO> کنز میخایلو است.

*xijabân-e zibâtarin < P_L_WO> kenez mixâjilo ast.

*The error tag:* < P_L_WO> Permutation_Lexis_Word Order

Correct Form: [زیباترین خیابان]

Gloss:

* xijabân -e zibâtarin < P_L_WO> kenez mixâjilo ast.

Beautiful-EZ street [xijabân-e zibâtarin] < P_L_WO> [zibâtarin xijabân] be-3sg.

"The most beautiful street is the Knez Mihailova."

*Error Description*: The superlative adjective word order is incorrect. The superlative adjective is a pre-modifier and precedes the noun.

Besides the errors in the Ezâfe Construction, which are the most frequent errors in the corpus related to word order, another notable error related to this error tag is that concerning standard syntactic word order. Standard Persian word order follows Subject, Object, Verb (SOV) order. Since Serbian word order follows SVO order, it may cause some errors in learners' productions. The following example explains this

(D) Error in standard syntactic word order

*Document ID*: LIN044_T0133

* من دوست دارم گوشت و سیب زمینی <P_S_WO >

*man dust dâram gusht va sibzamini < P_S_WO >.

*The error tag*: < P_S_WO > Permutation_Syntax_Word Order

*Correct Form*: [من گوشت و سیب زمینی دوست دارم]

Gloss:

*man dust dâr=am V1 gusht va sibzamini < P_S_WO >.

I like [dust dâram] meat and potatoes < < P_S_WO > [dust dâram].

"I like meat and potatoes."

*Error Description*: The syntactic word order (SOV) is incorrect and the sentence does not follow standard Persian word order.

**(2) Verb (other) Errors**

The second frequent linguistic error type as mentioned in Table 6.13 is marked as Verb Other. It should be noted that verb errors are divided into three categories: verb agreement (VA), verb tense (AT) and verb other (VO) based on the SFLC error tagset (4.2.2). The Verb Other tag contains all errors related to the use of verbs, with the exception of issues of agreement and tense such as missing verbs, wrong conjugations, and missing or incorrect use of main verbs, light verbs or auxiliaries etc.

According to error tag frequency, the majority of errors in Verb Other are tagged in the domain of syntax, with 181 error tags, and substitution errors at the surface structure annotation level with a frequency of 124 tags. Omission also appears in the next frequency ranking in the surface structure with 96 error tags. A review of the errors in the corpus shows that the majority of errors which were marked as substitution, syntax, and verb other (S_S_VO) were made by the learners due to the incorrect selection of compound verbs. In the Persian language, compound verbs are constructed from a simple verb and a non-verbal

element, such as a noun, adjective, past participle, prepositional phrase, or adverb, and a verbal constituent (Dabir-Mogaddam, 1997).

The learners made errors by mis-selecting, i.e. substituting a simple verb instead of a compound verb or substituting the wrong element (verbal or non-verbal) of a compound verb. The following evidence of mis-selections in compound verbs was found in the corpus.

(A) Mis-selecting a simple verb instead of a compound verb

*Document ID:* LIN043_T0128

* موسیقی خوب شنیدم <S_S_VO >

*musiqi xub šenidam < S_S_VO >.

*The error tag*: < S_S_VO > Substitution_Syntax_Verb Other

*Correct Form*: [گوش کردم]

Gloss:

*musiqi xub šenid=am V1 < S_S_VO >.

Music I listen [šenid.PAST.am] < S_S_VO > [guš kard=am].

"I listened to the music."

*Error Description*: The simple verb (šenidam) is used instead of the compound verb (guš kardam).

(B) Mis-selecting the verbal element of a compound verb

Document ID: LIN058_T0161

* دیروز مریض بودم، سرما گرفتم <S_S_VO >

*diruz mariz budam, sarmâ gereftam < S_S_VO >.

*The error tag*: < S_S_VO > Substitution_Syntax_Verb Other

*Correct Form*: [سرما خوردم]

Gloss:

*diruz mariz bud=am, sarmâ gereft=am V1 < S_S_VO >.

Yesterday I sick be. Past.1st , sarmâ [get .PAST.1sg] < S_S_VO > [xord.PAST.1st]

"Yesterday I was sick, I caught a cold."

*Error Description*: The verbal element of the compound verb, i.e. xordan, has been misselected and the learner used the verbal element 'gereftan' instead of 'xordan' (sarmš xordan: to catch a cold).

**(3) Noun (Other) Errors**

The third frequent error type in the category of linguistic errors is Noun Other (NO). Based on the SFLC error tagset (4.2.2), errors related to nouns are divided into two groups: Noun-Plural (NP), errors regarding the plurality/singularity of nouns, and Noun Other (NO), which marks any errors regarding nouns except for those concerning plurality. The error tag 'NO' is mainly used to address those errors related to noun phrase structures. The structure of the Persian noun phrase consists of a head noun which is followed by the modifiers, however, some elements such as the determiner, the numeral construction and the quantifiers precede the head noun (Megerdoomian, 2000). According to Megerdoomian (ibid), the lack of overt morphology to mark boundaries, a relatively free word order and the optionality of the subject are some of the reasons which make this structure highly ambiguous, which may result in learner errors.

In the SFLC, 'Noun Other' errors were marked in the domains of lexis, and morphology and most of them were tagged for omission in the surface structure. Such errors are explained in the following examples.:

(A) **Noun error in the lexis domain**

Document ID: LIN054_T0152

*در <O_L_NO> پنجشنبه به رستوران می روم.

*dar <O_L_NO> panjšanbe be resturân miravam.

*The error tag*: < O_L_NO > Omission_Lexis_Noun Other

*Correct Form*: [روز پنجشنبه]

Gloss:

* dar [ ]<O_L_NO> [ruz-e] panjšanbe be resturân mi=rav=am.

In <O_L_NO> Thursday to restaurant go. PRES.Stm.1SG

"I will go to the restaurant on Thursday."

*Error Description*: The head noun [ruz] and the subsequent Ezâfe construction [ruz-e] have been omitted. In this example, the head noun [ruz: day] should proceed the other noun [panjšanbe: Thursday] to complete the noun phrase.

**(B) Noun error in the morphology domain**

The majority of error tags in Omission, Morphology, Noun (O_M_NO) include the omission of the indefinite enclitic [i] which attaches to the end of a noun. Definiteness in Persian is marked either by the numeral 'one' (i.e. 'jek') before the noun or by using the enclitic [i] after the noun. Such constructions are difficult for learners to follow and may result in errors. The following is an example where the indefinite enclitic has been omitted.

Document ID: LIN037_T0114

* صربستان کشور< O_M_NO > در جنوب شرقی اروپا است.

*serbestân kešvar << O_M_NO dar jonub-e šarqi orupâ ast

*The error tag*: < O_M_NO > Omission_Morphology_Noun Other

*Correct Form*: [کشوری]

Gloss:

* serbestân [ kešvar]<O_M_NO> [kešvari] dar jonub-e šarqi orupâ ast.

Serbia country < O_M_NO > in south east Europe be-3SG

'Serbia is a country in South East Europe.'

*Error Description*: The indefinite enclitics [i] has been omitted.

**(4) Preposition Errors**

The fourth major error type in the linguistic error category pertains to Preposition. Such errors are marked for the domain of syntax and were mainly tagged in substitution as well as omission and addition at surface structure annotation level.

According to Perry (2007), Persian has only eight primary (six etymologically primitive) prepositions in general use. These are [به /be] 'at, to, in, by' (dative, locative, directional, instrumental); [در /dar] 'in(to)'; [از /az] 'from, through, along'; [با/bâ] 'with' (commutative, instrumental, concessive); [تا /tâ] 'up to, until'; [چون /cun] 'like, as'; [جز/joz] 'except' (historically, be-joz-e, < Ar. juz' 'part'); and [برای/barâ (-ye)] 'for' which. They are used to express case relations in Persian.

The high frequency of preposition errors marked for substitution in surface structure indicate that the learners did not learn the correct usage of the prepositions due to the varieties or the influence of first language prepositions.

**(A) Mis-selection of preposition**

Document ID: LIN027_T0299

* به < S_S_PP > سال نو من سر کار بودم

*be < S_S_PP > sâl-e now man sar-e kâr budam.

*The error tag*: < S_S_PP > Substitution_Syntax_Preposition

*Correct Form*: [در]

Gloss:

* [be]< S_S_PP > [dar] sâl-e now man sar-e kâr bud=am.

PREP-be < S_S_PP > year-EZ ADJ-new at work be.PAST.1SG

'I was at work on New Year's Eve.'

*Error Description*: The learner has mis-selected the preposition.

The other frequent error tag for preposition is marked as omission at surface structure level. Such absence of prepositions could indicate a lack of knowledge of the

usage of prepositions in a well-formed Persian syntactic construction. The following sample is an example of preposition omission.

**(B) Omission of preposition**

Document ID: LIN039_T0118

* می خواهم < S_O_PP > آثار باستانی بازدید کنم

*mixâham < S_O_PP > âsâr-e bâstâni bâzdid konam.

*The error tag*: < S_O_PP > Substitution_Syntax_Preposition

*Correct Form*: [از]

Gloss:

* mi=xâh=am [ ] < S_O_PP > [az] âsâr-e bâstâni bâzdid kon=am.

DUR-want-PRE.Stm1SG < S S_O_PP > [PREP-from] Antiquities visit-PRE.1SG

'I want to visit the Antiquities.'

*Error Description*: The preposition 'az' has been omitted.

**(5) Word Selection Errors**

Word selection error type is marked for the lexis domain and mainly for substitutions at surface structure annotation level. The error tags indicate the mis-selection of the proper word thus resulting in semantically ill-formed constructions. Here is an example of incorrect word selection:

Document ID: LIN085_T0265

* تاریخ هنر ایران خیلی دور < S_L_WS > است

*târix-e honar-e irân xejli dur < S_L_WS > ast.

*The error tag*: < S_L_WS > Substitution_Lexis_Word Selection

*Correct Form*: [کهن/طولانی]

Gloss:

*târix-e honar-e irân xejli [dur ] < S_L_WS > [tulâni] ast.

history-EZ art-EZ iran INT-very [ADJ-far] < S_L_WS > be-3SG

'Iranian history of art is very old.'

*Error Description*: The adjective has been mis-selected, thus causing a semantically ill-formed sentence.

### 6.3.2 Orthographic Errors in the SFLC

Table 6.14 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 (see Table 6.3). These errors are mainly tagged for substitution and omission at surface structure annotation level.

Table 6.14: 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 (2.8.1.3) 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, as already mentioned in Table 6.7, they are the most frequent

in the whole corpus as well as at each level of proficiency. In this section, two major orthographic errors are discussed.

**(A) Mis-selecting the consonant character**

The Persian script contains multiple consonant forms (2.8.1.4) which allow some identical consonant phonemes to be represented by different letters. For instance, for the phoneme /s/ there are 3 full forms /س/, /ص/ and /ث/, which also have their short forms/ـسـ/, /ـصـ/ and /ـثـ/ for connecting to the next/previous letters. Such diversity in forms makes it difficult and complex for learners to choose the right letter for a word. For instance, in document LIN069_T0174, the word 'صربستان' (serbestân/ Serbia) is written as 'سربستان'and the learner made the error by mis-selecting the wrong form of /s/, i.e. /ـسـ/ instead of /ـصـ/. Such substitutions occurred frequently due to the fact that there are no specific rules for using different forms of a single phoneme and learners should memorize the correct form separately. Therefore, errors related to the misselection of consonant characters are mostly due to the multiple forms of Persian letters.

**(B) Omission of the long vowel character**

The second frequent error type tags regard the omission of the long vowel character which is mainly the omission of the vowel / â /. This vowel does not exist in Serbian phonemes and since it is very difficult for Serbian learners of Persian to pronounce it, most of the learners tend to pronounce it as / a / which is a short vowel (2.8.1).The complexity lies in the fact that the vowel /a /is represented as a diacritic in writing which means there is no specific letter for it in the middle position. For example, in document LIN22_T0086, the word 'فارسی' (Farsi/Persian) is written as 'فرسی', the learner omitted the long vowel / â / character.

# 7. Conclusions and Implications

**Chapter Summary**

This concluding chapter summarises the contributions presented in this thesis regarding the construction and development of the SFLC, the first error-tagged learner corpus of the Persian language. The thesis achievements are listed and the possible applications of the SFLC are introduced.

## 7.1 Summary of the Thesis and Achievements

The present research was primarily aimed at constructing and analyzing an error-tagged learner corpus of the Persian language using the data collected from Serbian learners of Farsi. The two main objectives of the present research as explained in 1.3 were (i) to construct and develop an error-tagged learner corpus of Persian and then (ii) to investigate the frequency and types of errors made by Serbian learners of the Persian language. To achieve this aim, the present thesis was organized into 7 chapters, including the present chapter, which provides a review of the main topics discussed in them.

**The first chapter** mainly presented an outline for the research by setting the specific objectives, describing the methodology and suggesting the stages and phases of the project development. The definition of some key terms in the research subjects such as learner corpus, error analysis, error-tagged learner corpus and the connections between learner corpus and EA and the SLA were briefly given and reviewed in this chapter.

**Chapter 2** reviewed the background and literature behind the research. The chapter consists of 7 sections and the topics were discussed in detail, with a special focus on the learner corpora research domain. The first 3 sections reviewed the topics and definitions of the corpora and corpus linguistics and discussed the learner corpus domain. Eight well-known types of learner corpora, introduced by different scholars, were reviewed and discussed separately. Since carrying out research in the field of corpus linguistics, like any other field of study, demands that certain specific stages be followed in order to obtain reliable results, section 4 reviewed the model of learner corpus research stages which were basically introduced by Granger (2012). Here topics such as choosing the appropriate methodological approach, data collection, data annotation and data analysis were discussed in detail. Section 5 provided a general review of learner corpora application and a comprehensive model of such applications introduced by Diaz-Negrillo & Thompson (2013) was discussed. Section 6 overviewed 10 well-known learner corpora projects around the world so as to identify a model for the corpus design criteria. Such a model consists of 3 types of main corpus criteria and includes 9 aspects in total. The selected 10 well-known learner corpora were reviewed based on these corpus criteria and aspects in order to

develop a model for designing the Persian language learner corpus for the project. Finally, section 7 reviewed the basic specifications of the Persian language in terms of phonological, morphological and syntactic characteristics.

**Chapter 3** focused on the design and development of the Salam Farsi Learner Corpus. In the first section of the chapter, a new model for the Learner Corpora Design Criteria was proposed by the researcher and was subsequently adopted as the basic model for designing the SFLC. Based on the proposed model, the specific outline for the SFLC was presented, covering all the basic design requirements for constructing a learner corpus, such as the specific corpus criteria, data criteria, learner criteria as well as the tagging type and metadata annotation. Further on in this chapter, the contents of the SFLC were discussed in detail: the data specifications (i.e. the type of materials and size of the corpus, the proficiency levels of the learners, text types, task types, genres) and the metadata specifications (i.e. the learner metadata and text metadata). In the last section of the chapter, the process of digitizing the SFLC raw data was explained, which mainly included scanning the hand-written texts and saving them in PDF format, manually transcribing the texts based on specific instructions and saving them into the corpus database.

**Chapter 4** consists of two sections. In the first section an overview of error analysis was presented and the topics of SLA Research and Error Analysis and Learner Corpora and Error Analysis were discussed in depth. Subsequently, Computer-aided Error Analysis was introduced as the new methodological model for EA which could help to overcome the limitations of traditional Error Analysis. In the second section, the system of SFLC error annotation and the SFLC error tagset were introduced. The SFLC adopted the descriptive classification of errors as introduced by Dulay et al. (ibid) as the main error taxonomy for the corpus. Based on such classifications the SFLC Error Taxonomy was introduced, consisting of errors divided into 3 levels: surface Structure, error domain and error type. Finally, the SFLC error tagset table was introduced, which makes it possible to mark the errors in the corpus by means of four-letter error tags.

**Chapter 5** was dedicated to the SFLC software interface and tools. This chapter consists of two sections: (i) the corpus webpage and interface, and (ii) the SFLC tools. In the first section, the main web pages of the corpus website were introduced in detail and

some specifications were discussed. In the second section, the four main tools for the SFLC, used for submitting, tagging, filtering, and searching the corpus data, were introduced. The tools developed originally and specifically for the SFLC are: the Data Submitting and Metadata Tagging Tool (DSMT), for storing data in the corpus database and marking with metadata tags; The Error Tagging Tool (ETT), which functions as a computer-aided error editor and facilitates error tagging; the Filter and Search Tool (FST), which includes different filters and enables searching for specific errors or words in the corpus; and finally the Data Statistics Tool (DST), which shows the numerical statistics related to the corpus.

**Chapter 6** provided statistical reports for the SFLC and discussed the results in two main sections: (i) the frequency of error distribution in the whole corpus and (ii) the comparison of error distribution across the proficiency levels. The statistics showed that the most frequent errors in the whole corpus are to be found in the domain of orthography and in terms of linguistic errors, the most frequent ones lie in the domains of syntax and lexis. Word Order is marked as the major frequent error type in the whole corpus. As for the distribution of errors across proficiency levels, the statistics showed that the total errors dropped from level A2 to C1, while errors in syntax increased from level A2 to C1 due to the use of complex and compound syntactic structures at upper levels.

## 7.2 Possible Applications of the SFLC

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, also provides such data resources, specifically on learner errors, which are expected be useful and provide helpful data sources not only for SLA researchers, but also for teachers, textbook and language material writers, lexicographers and even learners themselves. Some possible applications of the SFLC are listed below.

(1) Conducting Research into Second Language Acquisition

Research into Farsi as a second/foreign language may benefit from the SFLC data resources since the corpus not only provides authentic data gathered from learners at different proficiency levels, but also statistics regarding error tags and metadata. Researchers can either directly access the whole texts from the corpus to carry out research into different linguistic areas, or by using the SFLC tools, conduct specific research on learner errors.

(2) Investigating Learner Performance

The SFLC provides data sources for learner errors at different proficiency levels which may indicate the general trend of learning at each level. Moreover, since the submitted data in the SFLC database is saved identically for each learner, by assigning a unique ID for the learners called the LIN (Learner Identification Number), the trend of learning and language performance can be tracked for each proficiency level.

(3) Carrying Out Contrastive Interlanguage Analysis

According to Granger (2003), using learner corpora for CIA enables researchers to compare learners' data with native speakers' data and uncover a wide range of patterns of underuse, overuse, and misuse in learner lexis, (lexico-)grammar, and discourse. The SFLC tools, namely the Filter and Search Tool (FST) and the Data Statistic Tool (DST), enable SLA researchers to compare learners' errors according to different criteria such as proficiency levels, age, first language (which in the SFLC is Serbian by default), years of learning Farsi etc. Therefore, such research could be used in contrastive analysis for finding systematic errors and learner patterns.

(4) Improving Persian Language Teaching Skills

The SFLC data may be used for Persian language instructors as a methodology to improve teaching performance. According to Granger (2002:21), learner corpora 'open up interesting descriptive and pedagogical perspectives' with 'a profound and positive impact on the field of Foreign Language Teaching (FLT)'. To this aim, Data-Driven Learning

(DDL) is suggested, which according to Johns & King (1991:iii), refers to 'the use of computer-generated concordances in the classroom'. Such computer-generated concordances can include error-tagged learner corpora which provide access to frequent learning errors. Farsi instructors may use the SFLC data to set up concordance-based exercises and increase learner awareness of frequent errors.

(5) Syllabus Design and Developing Learning Materials

The SFLC data statistics provide a clear view of learner errors for educators and material writers. The errors could take on the role of a 'road map' for showing educators and material writers the linguistic weakness of learners at each proficiency level. Such weaknesses could be highlighted when creating educational contents such as textbooks, workbooks, grammar references, workshop materials, etc.

(6) Developing Learner Dictionaries

The SFLC could serve to list the most frequent learner errors based on different search options, using the FST, and even provides the contexts where such errors are made. These possibilities enable lexicographers to consider errors, i.e. errors in lexis, so as to enrich their entries when providing definitions and synonyms. The SFLC may also be used for developing a dictionary of common errors.

(7) Used by Language Learners

Farsi learners can access the SFLC directly to browse the texts and errors in the whole corpus. This direct access to the SFLC and the use of the Filter and Search Tool (FSL) enable them to see the usage of the words, phrases, and syntactic structure in authentic texts as well as the errors and corrections. According to Johns (2002:108), such access 'confronts the learner as directly as possible with the data' 'to make the learner a linguistic researcher', which may result in increased learning awareness and self-correction.

## 7.3 Recommendations for Future Research

The present thesis attempts to contribute to the field of Second Language Acquisition and teaching Persian as a foreign language by constructing and analysing an error-tagged learner corpus of the Persian language. The construction and development of different learner corpora for the Persian language have not been give due consideration to date and to the best knowledge of the researcher, the present thesis is the first attempt to develop an error-tagged learner corpus of Persian. In order to pave the way for and shed more light on future work, the following recommendations are made:

(1) Developing different 'design criteria' to boost Persian learner corpora may provide valuable new insights into corpus construction.

(2) For constructing the error-tagged corpora, new error taxonomies and error tagsets may be introduced to study different types of errors and to carry out different analysis.

(3) Compiling larger amounts of learner data (texts and metadata) from learners with different first language backgrounds and consequently building and investigating such data comparatively would provide valuable information on learner language skills and provide a clear view of the difficulties experienced when learning Persian within different groups of Farsi learners.

(4) New corpus tools and software need to be developed and introduced to facilitate data importing, tag inserting and exporting the reports and statistics.

(5) There is a need to develop specific Persian learner corpora with different types of annotations such as part of speech (POS), morphological, morphosyntacitc, syntactic, semantic etc. which will lead to the achievement of better results in

studying the language learning skills and learner difficulties involved in learning Persian as a foreign language.

In sum, this research has presented a method for building an error tagged learner corpus for the Persian language which can be used as a model to develop various learner corpora for the Persian language. It is hoped that the methodology used in this research will bring new insights and ideas to the field of learner corpus research and SLA and hopefully shed more light on future research and studies in developing and constructing more Persian learner corpora.

# References


Alfaifi, A and Atwell, E. (2013). Arabic Learner Corpus v1: A New Resource for Arabic Language Research. In proceedings of *the Second Workshop on Arabic Corpus Linguistics (WACL-2)*. Lancaster University, UK.

Alfaifi, A. (2015) *Building the Arabic Learner Corpus and a System for Arabic Error Annotation*. PhD thesis, University of Leeds.

Alfaifi, A., Atwell, E. and Abuhakema, G. (2013) Error Annotation of the Arabic Learner Corpus: A New Error Tagset. In: *Language Processing and Knowledge in the Web, Lecture Notes in Computer Science. 25th International Conference (GSCL 2013)*, 25-27 September 2013, Darmstadt, Germany. Springer, (9) 14 - 22.

Allen, W. S. (1956). Structure and system in the Abaza verbal complex, *Transactions of the Philological Society,* 127–176.

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

Bateni, M. R. (1975). zaban va Tafakkor. Entesharat-e Ketabe Tantl8in: Tehran.

Bennett, G. R. (2010). Using Corpora in the Language Learning Classroom: Corpus linguistics for teachers. Ann Arbor, MI: University of Michigan Press.

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

Biber, D. (2010). Corpus-based and corpus-driven analyses of language variation and use. In: Bernd Heine & Heiko Narrog (eds.), *The Oxford Handbook of Linguistic Analysis*. Oxford: Oxford University Press. pp.159-191

Biber, D., & Finegan, E. (1991). On the exploitation of computerized corpora in variation studies. In K. Aijmer & B. Altenberg (Eds.), *English corpus linguistics: Studies in honour of Jan Svartvik* (pp.204-220). London: Longman.

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

Brown, H. D. (1980). *Principles of Language Learning and Teaching.* New Jersey: Prentice-Hall Inc.

Burnard, L. (2005). Metadata for corpus work. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 30–46). Oxford, UK: Oxbow Books.

Buttery, P., & Caines, A. (2012). Normalising frequency counts to account for 'opportunity of use' in learner corpora. In Y. Tono, Y. Kawaguchi, & M. Minegishi (Eds.), *Developmental and cross-linguistic perspectives in learner corpus research* (pp. 187–204). Amsterdam, the Netherlands: Benjamins.

Callies M. & Paquot, M. (2015). Learner Corpus Research: An interdisciplinary field on the move. *International Journal of Learner Corpus Research*, (1), 1-6.

Callies, M. & Götz, S. (2015). Learner corpora in language testing and assessment: Prospects and challenges. In: M. Callies & S. Götz (Eds.), *Learner Corpora in Language Testing and Assessment* (1-9). Amsterdam: John Benjamins.

Callies, M. (2015). Learner corpus methodology. In S. Granger, F. Meunier & G. Gilquin (eds.) *Cambridge Handbook of Learner Corpus Research* (35-55). Cambridge: CUP.

Callies, M., Díez-Bedmar, M. B. & Zaytseva, E. (2014). Using learner corpora for testing and assessing L2 proficiency. In P. Leclercq, H. Hilton & A. Edmonds (Eds.), *Measuring L2 Proficiency: Perspectives from SLA* (71–90). Clevedon: Multilingual Matters.

Callies, M., Paquot, M. (2015). Learner Corpus Research: An interdisciplinary field on the move. In: *International Journal of Learner Corpus Research*, 1: (1-6).

Casas-Pedrosa, A.V., Fernández-Domínguez, J. & Alcaraz-Sintes, A. (2013). 'Introduction: the use of corpora for language teaching and learning', *Research in Corpus Linguistics,* (1):1-5.

Castillejos Lopez, W. (2009). Error Analysis in a Learner Corpus. What Are the Learners' Strategies. *AELINCO*. Retrieved from http://www.um.es/lacell/aelinco/contenido/pdf/45.pdf.

Chuang F-Y. & Nesi, H. (2006). An analysis of formal errors in a corpus of Chinese student writing. *Corpora* 1, 251-271.

Corder, S. P. (1967). The significance of learners' errors. Cited in J.C. Richards (ed.) 1984 *Error Analysis: Perspectives on second language acquisition*, pp 19 – 27.

Corder, S. P. (1973). *Introducing Applied Linguistics*. (pp. 274-277). Harmondsworth: Penguin.

Council of Europe. (2001). Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR). New York, NY: Cambridge University Press.

Coxhead, A. (2000). A new Academic Word List. *TESOL Quarterly*, *34*, 213–238.

Dagneaux, E., Denness, S. & Granger, S. (1998). Computer-aided Error Analysis. *System: An International Journal of Educational Technology and Applied Linguistics* 26, 163-174.

Dash, N. (2003) Use of Language Corpora in Second Language Learning. *South Asian Language Review*. 13 (1): 1- 26.

Dash, N. (2010). '*Corpus linguistics: A general introduction'*. Proceedings of the workshop on Corpus Normalization, LDCIL, CIIL, Mysore, India, on 25th August 2010. Retrieved from www.ldcil.org/download/Corpus%20Linguistics.pdf

Davies, A. (2013). *Native Speakers and Native Users: Loss and gain*. Cambridge: Cambridge University Press.

Degand, L. & Perrez, J. (2004). Causale connectieven in het leerdercorpus Nederlands. *Tijdschrift n/f* 4. 115-128.

Diaz-Negrillo, A. & Thompson, P. (2013). Learner corpora: looking towards the future. In: N. Diaz-Negrillo, Ballier & P. Thompson (Eds.), *Automatic Treatment and Analysis of Learner Corpus Data. Studies in Corpus Linguistics* (pp. 9-30). John Benjamins, Amsterdam.

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.

Ellis, R. (1994). *The study of second language acquisition*. (p. 59). Oxford: Oxford University Press.

Flowerdew, L. (2004). The argument for using English specialised corpora to understand academic and professional language. In U. Connor & T. Upton (Eds.), *Discourse in the Professions* (pp. 11-33). Amsterdam: John Benjamins.

Flowerdew, L. (2012). *Corpora and Language Education*. New York: Palgrave MacMillan.

Gass, S. M., & Selinker, L. (2008). *Second language acquisition: An introductory course*. 3rd edition. New York and London: Routledge.

Gillard, P. & Gadsby, A. (1998). Using a learners' corpus in compiling ELT dictionaries. In S. Granger (Ed.), *learner English on Computer* (pp.159-171). London: Longman.

Gilquin, G. (2000/2001). The Integrated Contrastive Model. Spicing up your data. *Languages in Contrast 3(1)*, 95-123.

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.

Gilquin, G., Granger, S. & Paquot, M. (2007). Learner Corpora: The Missing Link in EAP Pedagogy. *Journal of English for Academic Purposes, 6*(4), 319-335.

Gilquin, G., Granger, S. & Paquot, M. (2007). Writing sections. In M. Rundell (Editor in chief) *Macmillan English Dictionary for Advanced Learners* (second edition) (pp. IW1- IW29). Oxford: Macmillan Education.

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. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg & M. Johansson (Eds), *Languages in Contrast. Text-based cross-linguistic studies* (pp. 37-51). Lund: Lund University Press,

Granger, S. (1998). The computer learner corpus: A versatile new source of data for SLA research. In S. Granger (Ed.), *Learner English on Computer* (pp.3-18). London: Longman.

Granger, S. (2002). A bird's-eye view of computer learner corpus research. In S. Granger, J. Hung, S. Petch-Tyson & J. Hulstijn (Eds.), *Computer learner corpora, second*

*language acquisition and foreign language teaching* (pp. 3-33). Amsterdam & Philadelphia: John Benjamins.

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

Granger, S. (2003d). The International Corpus of Learner English: A new resource for foreign language learning and teaching and second language acquisition research. TESOL Quarterly, 37(3), 538–546.

Granger, S. (2004). Computer learner corpus research: current status and future prospects. In U. Connor & T. A. Upton (Eds.), *Applied corpus Linguistics: A Multidimensional Perspective (pp. 123- 135).* Amsterdam: Rodopi.

Granger, S. (2008). Learner corpora in foreign language education. In: N. Van Deusen-Scholl & N. H. Hornberger, (Eds.), *Encyclopedia of Language and Education, Volume 4: Second and Foreign Language Education* (pp. 337–51). New York: Springer.

Granger, S. (2008). Learner Corpora. In A. Ludeling & M. Kyto (Eds.), *Corpus linguistics: An international handbook* (pp. 259–275). Berlin, Germany: Walter de Gruyter.

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In K. Aijmer (Ed*.), Corpora and language teaching*. (pp. 13-33). Amsterdam: Netherlands: John Benjamins.

Granger, S. (2012). How to use foreign and second language corpora. In A. Mackey & S. Gass (Eds.) *Research methods in second language acquisition: A Practical Guide* (pp. 7-29). Madlen, MA: Blackwell Publishing.

Granger, S., Gilquin, G., & Meunier, F. (2015). Learner corpus research past, present and future. In: S. Granger, G. Gilquin & F. Meunier, *The Cambridge Handbook of Learner Corpus Research* (1-5). Cambridge University Press: Cambridge.

Gries, S. & Berez, A. (2015). Linguistic annotation in/for corpus linguistics. In N. Ide & J. Pustejovsky (Eds.), *Handbook of Linguistic Annotation*. Berlin/New York: Springer.

Hammarberg, B. (1999). Manual of the ASU Corpus, a longitudinal text corpus of adult learner Swedish with a corresponding part from native Swedes. Stockholms universitet: Institutionen för lingvistik.

Hana, J., Rosen, A., Škodová, S. and Štindlová, B. (2010). Error-tagged learner corpus of Czech. In *Proceedings of the Fourth Linguistic Annotation Workshop* (11-19). Uppsala, Sweden: Association for Computational Linguistics.

Hasyim, S. (2002) *Error Analysis in the Teaching of English*.Vol.4, No.1, Jun. pp 42-50. Retrieved from: http://puslit2.petra.ac.id/ejournal/index.php/ing/article/viewFile/15485/15477

Hunston, S. (2002*). Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

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

Ife, A. (2004). The L2 learner corpus: reviewing its potential for the early stages of learning. In: M. Baynham, A. Deignan & G. White (Eds.), *Applied Linguistics at the Interface* (91-103). British Studies in Applied Linguistics 19. London: Equinox.

Jeremiás, Éva M. (2003). New Persian. In: *The Encyclopaedia of Islam*. Ed. by Supplement. Brill Publishers, pp. 426–448.

Johns, T. & King, P. (Eds.). (1991). *Classroom Concordancing* (Vol. 4). Birmingham: University of Birmingham.

Johns, T. (2002). Data-driven learning: the perpetual challenge. In B. Kettemann and G. Marko (Eds.), *Teaching and Learning by Doing Corpus Linguistics* (pp. 107-117). Amsterdam: Rodopi.

Karimi, S. (2003). On Object Positions, Specificity, and Scrambling in Persian. In: S. Karimi (Ed.), *Word Order and Scrambling* (pp. 91-124). Oxford/Berlin: Blackwell Publishers.

Kennedy, G. D. (1998). An Introduction to Corpus Linguistics. London; New York: Longman.

Lazard, G. (1992). *A Grammar of Contemporary Persian*. Translated into English by Shirley A. Lyon. Mazda Publishers.

Le´on, J. (2005), Claimed and Unclaimed Sources of Corpus Linguistics. In: *Henry Sweet Society Bulletin* 44, 36-50.

Lee, N. (1990). Notions of "Error" and Appropriate Corrective Treatment, *Hong Kong Papers in Linguistics and Language Teaching*, 14, pp.55-70.

Lee, S., Jang, S. & Seo, S. (2009). Annotation of Korean learner corpora for particle error detection. *CALICO Journal* 26(3): 529-544.

Leech, G. (1992). Corpora and theories of linguistic performance. In J. Startvik (Ed.), *Directions in corpus linguistics* (pp. 105-122). Berlin: Mouton de Gruyter.

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

Leech, G. (1998). Preface. In: S. Granger (Ed.), *Learner English on Computer* (xiv-xx), London: Longman.

Lennon, P. (1991). 'Error: Some Problems of Definition, Identification, and Distinction'. Applied Linguistics, 12, 180-95.

Lüdeling, A., Maik, W., Kroymann, E., Adolphs, P. (2005). Multi-level error annotation in learner corpora. *The Corpus Linguistics Conference Series 1, 1. Corpus Linguistics 2005*.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (2007). The TalkBank Project. In J. C. Beal, K. P. Corrigan & H. L. Moisl (Eds.), *Creating and digitizing language corpora* (pp. 163–180). Houndmills: Palgrave-Macmillan.

Maicusi, T., Maicusi, P. & Carrillo, M. (1999). The error in the Second Language Acquisition. *Encuentro. Revista de investigación e innovación en la clase de idiomas*. 11, 168-173.

McEnery, T. & Xiao, R. (2011). What Corpora Can Offer in Language Teaching and Learning. In E. Hinkel (Eds.), *Handbook of Research in Second Language Teaching and Learning* pp (364-380). London & New York: Routledge.

McEnery, T., Xiao, R. and Tono, Y. (2006). *Corpus-Based Language Studies: An Advanced Resource Book*. London: Routledge.

Megerdoomian, K. (2000a). A computational analysis of the Persian noun phrase. Memoranda in Computer and Cognitive Science MCCS-00-321, Computing Research Lab, New Mexico State University.

Meshkotod Dini, M. (1995). *Sound Pattern of Language: An Introduction to Generative Phonology*. Mashhad: Ferdowsi University Press.

Meyer, C. (2002). *English Corpus Linguistics: An Introduction.* Cambridge: Cambridge University Press.

Milton, J. (1998). WORDPILOT: enabling learners to navigate lexical universes. In S. Granger & J. Hung (Eds.), *Proceedings of the International Symposium on Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (97-98). The Chinese University of Hong Kong,

Myles, F. (2008). Investigating learner language development with electronic longitudinal corpora: Theoretical and methodological issues. In L. Ortega & H. Byrnes (Eds.), *The longitudinal study of advanced L2 capacities* (pp. 58–72). Hillsdale: Lawrence Erlbaum.

Nagata, R. Whittaker, E. & Sheinman, V. (2011). Creating a manually error-tagged and shallow-parsed learner corpus. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (pp. 1210-1219).* Portland OR, 19-24 June.

Nesselhauf N. (2004a), Learner corpora and their potential in language teaching. In: J. Sinclair (Ed.), *How to Use Corpora in Language Teaching* (125-152). Amsterdam: Benjamins.

McCarthy, M. J. & O'Keeffe, A.(2010). Historical perspective: What are corpora and how have they evolved? In: O'Keeffe, A.& McCarthy, M. J. (eds.), *The Routledge Handbook of Corpus Linguistics*. London: Routledge, p 3-13.

Osborne, J., Henderson, A., & Barr, R. (2012). *The Scientext English learner corpus*. Retrieved 14 June 2017 from http://scientext.msh-alpes.fr/scientext-site-en/?article19

Pravec, N. (2002). Survey of Learner Corpora. In: *ICAME Journal* 26: (81-114).

Perry, John R. (2007). Persian Morphology. In Alan S. Kaye (ed.) *Morphologies of Asia and Africa*, 975-1019. Winona Lake, Ind.: Eisenbrauns.

Price, N. (2013). The Gachon Learner Corpus. Retrieved 18 October 2016 from http://koreanlearnercorpusblog.blogspot.be/p/corpus.html

Ragan, P. H. (2001). Classroom use of a systematic functional small learner corpus. In M. H. Ghadessy & A. Roseberry (Eds.), *Small corpus studies and ELT (pp. 207-236)*. Amsterdam: Johan Benjamins.

Reder, S., Harris, K., & Setzler, K. (2003). The multimedia adult ESL learner corpus. *TESOL Quarterly*, 37(3), 546-557.

Richards, J. C. & Schmidt, R. (2002). *Dictionary of Language Teaching & Applied Linguistics*. Pearson Education Limited. London: Longman.

Richards, J. C. (1974). Error Analysis: Perspectives on second language acquisition. London: Longman.

Richards, J.C. (1971). A Non- Contrastive Approach to Error Analysis. *Journal of ELT*. 25, 204-219.

Rundell, M. & Granger, S. (2007). From corpora to confidence. *English Teaching Professional* 50: 15-18.

Safari, S. (2012). Designing and Developing a FFL Learner Corpus. Master's thesis. Allameh Tabatabai University, Tehran, Iran.

Samareh, Y. (1985). *Avashenasi-ye zaban-e Farsi*. Tehran: Tehran University Press.

Saville-Troike, M. (2006). *Introducing Second Language Acquisition*. Cambridge. Cambridge University Press.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics* 10. pp. 209-231.

Shih, R. H. H. (2000). Compiling Taiwanese learner corpus of English. *Computational Linguistics and Chinese Language Processing, 5*(2), 87–100.

Sinclair, J. (2005). Corpus and text - basic principles. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 1–16). Oxford, UK: Oxbow Books.

Sinclair, J. McH. (1991). *Corpus, Concordance, Collocations*. Oxford: Oxford University Press.

Stilo, D. (2004). Iranian as Buffer Zone Between the Universal Typologies of Turkic and Semitic. In E.A. Casto, B. Isaksson & C. Jahani (Eds) *Linguistic Convergence and Areal Diffusion: Case Studies From Iranian, Semitic, and Turkic* (35-63). London: Routledge/Curzon.

Stubbs, M. (1996). *Text and Corpus Analysis*. Oxford: Blackwell.

Taylor, C. (2008). *What is corpus linguistics? What the data says. ICAME Journal*, 32, 179-200

Tenfjord K., Meurer P., & Hofland K. (2004). The ASK corpus – a language learner corpus of Norwegian as a second language. Paper presented at the TALC 2004 conference, Granada – Spain, 6-9 July 2004.

Teubert, W. (2005). My version of corpus linguistics. *International Journal of Corpus Linguistics*,10 (1), 1–13.

Thompson, G. & Hunston, S. (eds.) (2006). *System and corpus: exploring connections*. London: Equinox.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam, Philadelphia: John Benjamins.

Tono, Y. (1999). Using Learner Corpora in ELT and SLA Research. Paper presented at the Symposium on the Roles of Corpora in Language Teaching and Language Engineering of the 12th World Congress of Applied Linguistics (AILA), Tokyo.

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.

Tono, Y. (2009). Integrating learner corpus analysis into a probabilistic model of second language acquisition. In P. Baker (Ed.), *Contemporary corpus linguistics* (pp. 184-203). London: Continuum International Publishing Group.

Turton, N. D., & Heaton, J. B. (1996). *Longman Dictionary of Common Errors*. Harlow, UK: Pearson Longman.

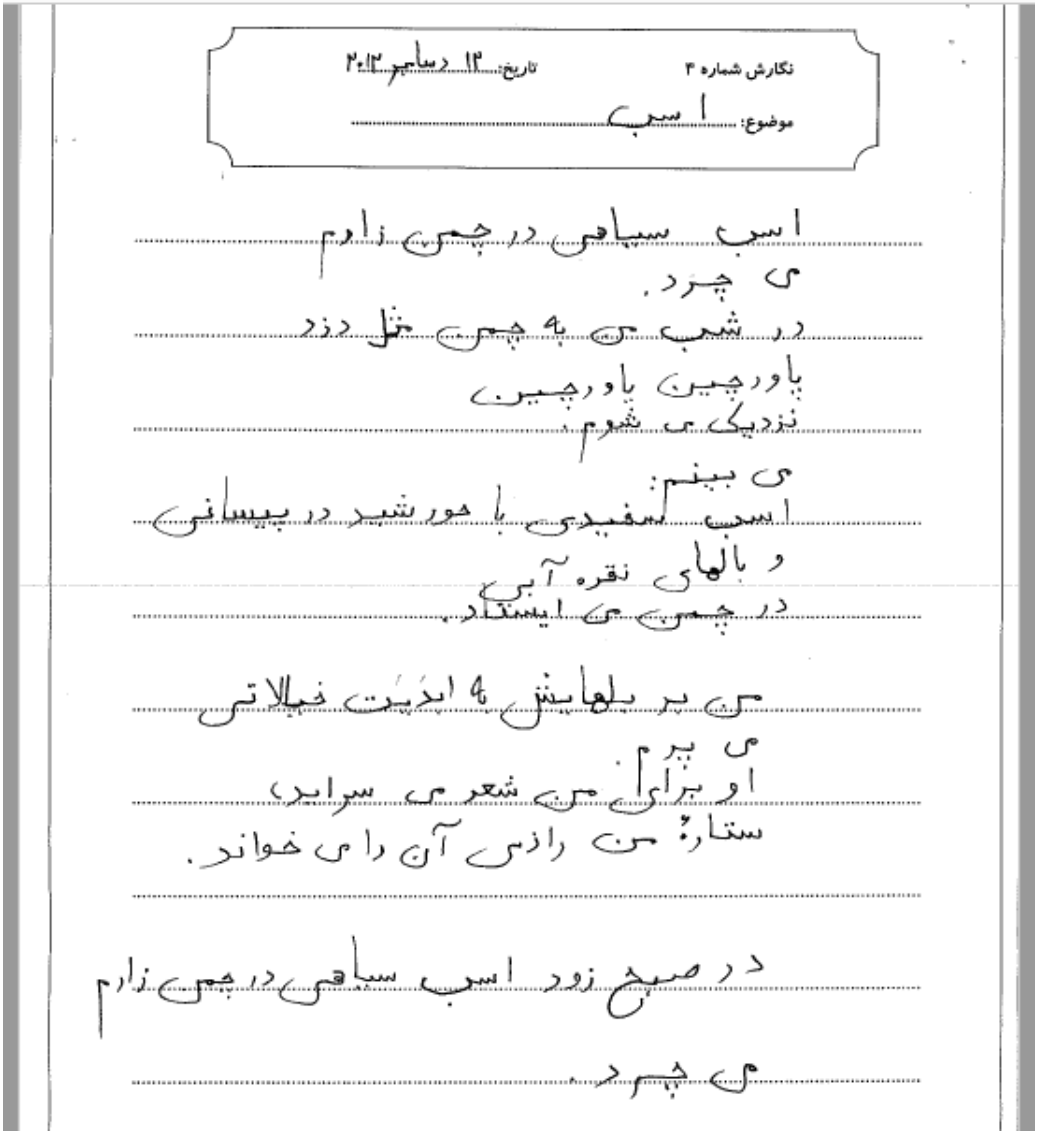
Xiao, Z. (2008). Well-known and influential corpora. In A. Lüdeling & M. Kyto (Ed.), *Corpus Linguistics: An International Handbook* (pp. 383-457. Berlin: Mouton de Gruyter.

http://www.learnercorpusassociation.org

https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

# Appendix

A. The SFLC file formats and annotations


نگارش شماره ۴ تاریخ: ۱۲ دسامبر ۲۰۱۲

موضوع: اسب

اسب سیاهی در چمن زارم
می چرد.
در شب می به چمن فرو رود
پاورچین پاورچین
نزدیک می شوم.
می بینم:
اسب سفیدی با خورشید در پیشانی
و بالهای نقره آبی
در چمن می ایستاد.
می پریم بالهایش را لمس می کنم
می پرم.
او برای من شعر می سراید
ستاره من رازی آن را می خواند.

در صبح زود اسب سیاهی در چمن زارم
می چرد.

Figure A 1: Example of hand-written text in PDF format

LIN005_T0010_NA_C1_sr_RS

Text:

اسب سیاهی در چمن زارم می چرد.
در شب من به چشمن مثل دزد پاورچین پاورچین نزدیک می شوم.
می بینم:
اسب سفیدی با خورشید در پیسانی و بالهای نقره آبی در چمن می ایستد.
من بر بلهایش به ابدیت خیالاتی می پرم.
او برای من شعر می سراید
ستاره من ارازی آن را می خواند.
در صبح زود اسب سیاهی در چمن زارم می چرد.

Figure A 2: Example of hand-written text transcription

Errors:

13 | چمن زارم | چمن زار | A_M_PR
30 | در شب من به چشمن مثل دزد پاورچین پاورچین نزدیک می شوم | | S_T_CS
42 | چشمن | چشمان | O_O_VL
107 | خورشید | خورشید | O_O_DT
117 | پیسانی | پیشانی | O_O_DT
165 | بلهایش | بالهایش | O_O_VL
175 | ابدیت خیالاتی | خیالات ابدی | P_S_AJ
221 | ستاره من ارازی آن را می خواند | | S_T_US

Figure A 3: Example of text error annotation

Data Criteria
Text type: Written
Task type: Free writing
Genre: Descriptive

Learner Criteria
First language: Serbian
Target language: Persian
Proficiency level: C1

Text Metadata
Text title: اسب
Year of production: 2012
Country of production: Serbia
City of production: Belgrade
Where produced: Home
Timing: N/A
References use: N/A
Grammar book use: N/A
Dictionary use: N/A

Learner Metadata
LIN: 005
Age: N/A
Gender: N/A
Nationality: Serbian
Number of languages spoken: 2
Number of years learning Farsi: 4
General level of education: BA
Major: N/A
Educational institution: ICC

Figure A 4: Example of text metadata annotation

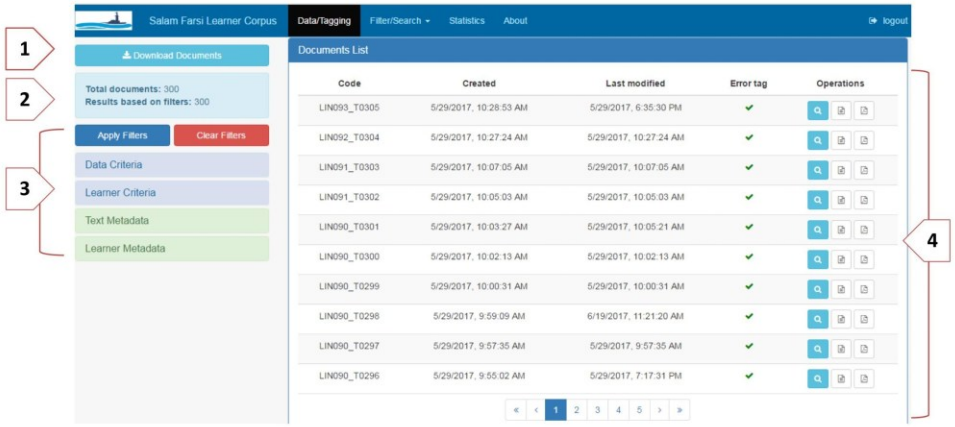
B. The SFLC user guide

FACULTY OF PHILOLOGY
UNIVERSITY OF BELGRADE

SFLC
SALAM FARSI
LEARNER CORPUS

# The SFLC User Guide

## Data/Tagging Page




1 Download Documents

All corpus documents (tagged data) can be downloaded in a compressed format (.zip).

2 **Total documents:** 300
**Results based on filters:** 300

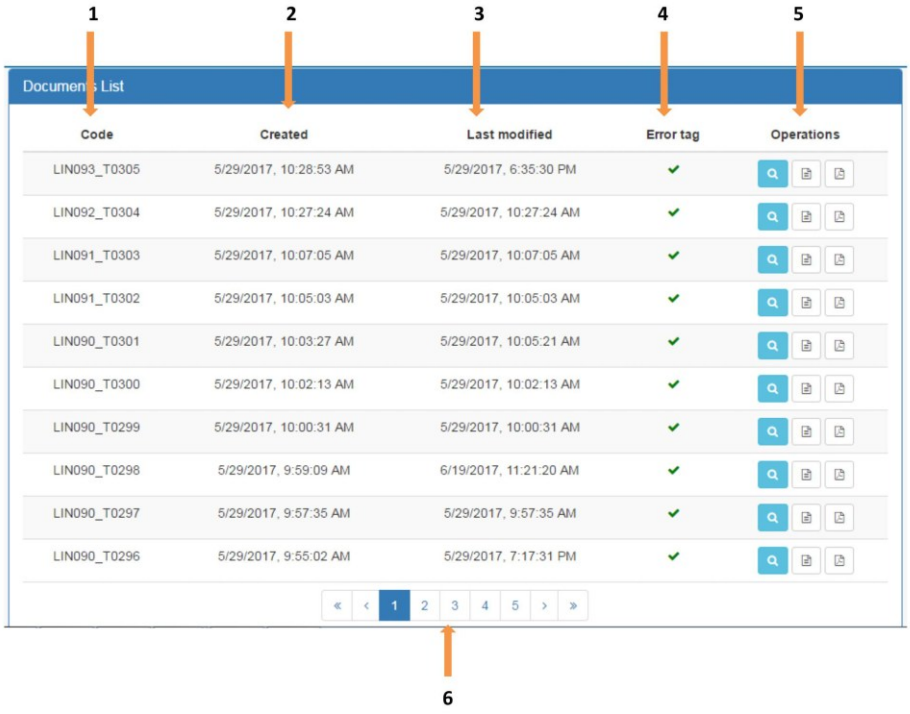
The **Data Box** shows total number of submitted documents. By selecting from the determinants (filters), the number of available texts will be shown in the "Results based on Filters".

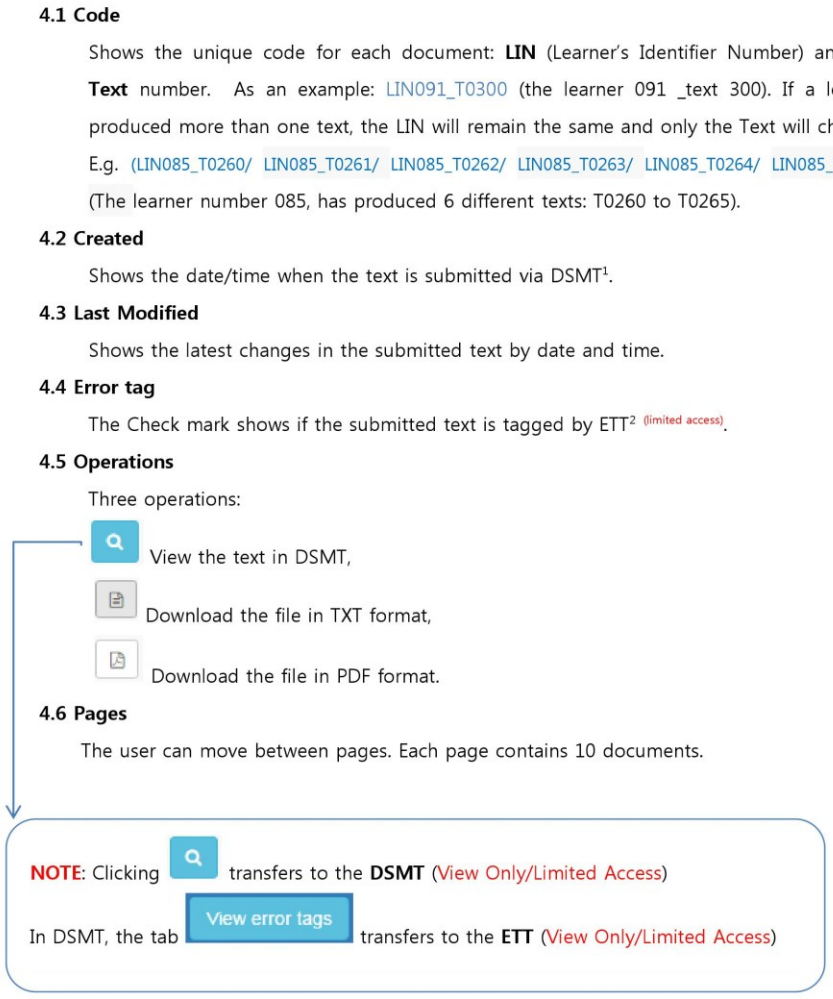
### 3 Metadata Filters


Select different variables in Metadata Filters for searching the corpus documents and to obtain specific results.

After selecting the variables, click Apply Filters to see the results in the **Document list** ▾ To Clear the determinants, click Clear Filters, doing this will reset the filters to the default mood and all documents will be shown in the Document List.

### 4 The Documents List





**4.1 Code**

Shows the unique code for each document: **LIN** (Learner's Identifier Number) and the **Text** number. As an example: LIN091_T0300 (the learner 091 _text 300). If a learner produced more than one text, the LIN will remain the same and only the Text will change. E.g. (LIN085_T0260/ LIN085_T0261/ LIN085_T0262/ LIN085_T0263/ LIN085_T0264/ LIN085_T0265) (The learner number 085, has produced 6 different texts: T0260 to T0265).

**4.2 Created**

Shows the date/time when the text is submitted via DSMT[1].

**4.3 Last Modified**

Shows the latest changes in the submitted text by date and time.

**4.4 Error tag**

The Check mark shows if the submitted text is tagged by ETT[2] (limited access).

**4.5 Operations**

Three operations:

View the text in DSMT,

Download the file in TXT format,

Download the file in PDF format.

**4.6 Pages**

The user can move between pages. Each page contains 10 documents.

**NOTE**: Clicking transfers to the **DSMT** (View Only/Limited Access)

In DSMT, the tab View error tags transfers to the **ETT** (View Only/Limited Access)

[1] Data Submitting and Metadata Tagging Tool (DSMT)
[2] Error Tagging Tool (ETT) the tool is only available for the corpus admins and annotators

Figure A 5: Example of the SFLC user guide

C. The specific notebook for collecting texts ( daftar-e negâresh)


Figure A 6: Example of the daftar-e negâresh

## BIOGRAFIJA AUTORA

Said Safari (Saeed Safari) rođen je 1976. godine u Iranu. Po završetku srednje škole i stručnih kurseva iz oblasti nastave persijskog jezika za osnovnu školu, 1994. godine počinje da radi za Ministarstvo prosvete Irana kao nastavnik persijskog jezika u državnim školama. Godine 1998. stiče diplomu više škole u oblasti nastave engleskog jezika kao stranog, pri Centru za obuku nastavnika Shahid Behshti (Shahid Behshti Teacher Training Center, Mashhad) u Iranu, a zatim 2003. stiče diplomu osnovnih univerzitetskih studija (BA) iz oblasti prevođenja na engleski jezik. U periodu od 2002. do 2012. godine paralelno je sticao radno iskustvo kao nastavnik i kaoprevodilac za više firmi. Godine 2012. završio je master studije (MA). Tokom master studija bio je angažovan na više projekata vezanih za nastavu persijskog jezika kao stranog, objavio je više naučnih radova i sarađivao je na izradi udžbenika i drugih materijala za nastavu persijskog jezika. Diplomirao je kao najbolji student u svojoj generaciji i kao dobitnik univerzitetske nagrade za studente. Na osnovu stečenog obrazovanja i iskustva u nastavi pozvan je od strane Veća zapromociju persijskog jezika (pri Ministarstvu prosvete Irana) da predaje persijski jezik u inostranstvu. Godine 2012. počeo je da drži nastavu persijskog na Filološkom fakultetuUniverziteta u Beogradu, a 2013. je postavljen za upravnika Centra za persijski jezik i Iranskog kulturnog centra u Beogradu. Tokom rada na Filološkom fakultetu u Beogradu napisao je prvi udžbenik persijskog jezika namenjen maternjim govornicima srpskog.

**Prilog 1.**

Izjava o autorstvu

Potpisani-a Saeed Safari

broj indeksa 13126 d

**Izjavljujem**

da je doktorska disertacija pod naslovom

**Constructing and Analysing an Error-tagged Learner Corpus of Persian**

- rezultat sopstvenog istraživačkog rada,
- da predložena disertacija u celini ni u delovima nije bila predložena za dobijanje bilo koje diplome prema studijskim programima drugih visokoškolskih ustanova,
- da su rezultati korektno navedeni i
- da nisam kršio/la autorska prava i koristio intelektualnu svojinu drugih lica.

**Potpis doktoranda**

U Beogradu, 13.10.2017.

**Prilog 2.**

Izjava o istovetnosti štampane i elektronske verzije doktorskog rada

Ime i prezime autora Saeed Safari

Broj indeksa 13126 d

Studijski program Jezik, književnost, kultura – modul Jezik

**Constructing and Analysing an Error-tagged Learner Corpus of Persian**

Mentor : dr Maja Miličević Petrović, vanredni profesor

Potpisani/a Maja Miličević Petrović

Izjavljujem da je štampana verzija mog doktorskog rada istovetna elektronskoj verziji koju sam predao/la za objavljivanje na portalu **Digitalnog repozitorijuma Univerziteta u Beogradu.**

Dozvoljavam da se objave moji lični podaci vezani za dobijanje akademskog zvanja doktora nauka, kao što su ime i prezime, godina i mesto rođenja i datum odbrane rada.

Ovi lični podaci mogu se objaviti na mrežnim stranicama digitalne biblioteke, u elektronskom katalogu i u publikacijama Univerziteta u Beogradu.

**Potpis doktoranda**

U Beogradu, 13.10.2017.

**Prilog 3.**

Izjava o korišćenju

Ovlašćujem Univerzitetsku biblioteku „Svetozar Marković“ da u Digitalni repozitorijum Univerziteta u Beogradu unese moju doktorsku disertaciju pod naslovom:

**Constructing and Analysing an Error-tagged Learner Corpus of Persian**

koja je moje autorsko delo.

Disertaciju sa svim prilozima predao/la sam u elektronskom formatu pogodnom za trajno arhiviranje.

Moju doktorsku disertaciju pohranjenu u Digitalni repozitorijum Univerziteta u Beogradu mogu da koriste svi koji poštuju odredbe sadržane u odabranom tipu licence Kreativne zajednice (Creative Commons) za koju sam se odlučio/la.

1. Autorstvo
2. Autorstvo - nekomercijalno
3. Autorstvo – nekomercijalno – bez prerade
4. Autorstvo – nekomercijalno – deliti pod istim uslovima
5. Autorstvo – bez prerade
6. Autorstvo – deliti pod istim uslovima

(Molimo da zaokružite samo jednu od šest ponuđenih licenci, kratak opis licenci dat je na poleđini lista).

**Potpis doktoranda**

U Beogradu, 13.10.2017.

1. **Autorstvo**. Dozvoljavate umnožavanje, distribuciju i javno saopštavanje dela, i prerade, ako se navede ime autora na način određen od strane autora ili davaoca licence, čak i u komercijalne svrhe. Ovo je najslobodnija od svih licenci.

2. **Autorstvo – nekomercijalno**. Dozvoljavate umnožavanje, distribuciju i javno saopštavanje dela, i prerade, ako se navede ime autora na način određen od strane autora ili davaoca licence. Ova licenca ne dozvoljava komercijalnu upotrebu dela.

3. **Autorstvo – nekomercijalno – bez prerada**. Dozvoljavate umnožavanje, distribuciju i javno saopštavanje dela, bez promena, preoblikovanja ili upotrebe dela u svom delu, ako se navede ime autora na način određen od strane autora ili davaoca licence. Ova licenca ne dozvoljava komercijalnu upotrebu dela. U odnosu na sve ostale licence, ovom licencom se ograničava najveći obim prava korišćenja dela.

4. **Autorstvo – nekomercijalno – deliti pod istim uslovima**. Dozvoljavate umnožavanje, distribuciju i javno saopštavanje dela, i prerade, ako se navede ime autora na način određen od strane autora ili davaoca licence i ako se prerada distribuira pod istom ili sličnom licencom. Ova licenca ne dozvoljava komercijalnu upotrebu dela i prerada.

5. **Autorstvo – bez prerada**. Dozvoljavate umnožavanje, distribuciju i javno saopštavanje dela, bez promena, preoblikovanja ili upotrebe dela u svom delu, ako se navede ime autora na način određen od strane autora ili davaoca licence. Ova licenca dozvoljava komercijalnu upotrebu dela.

6. **Autorstvo – deliti pod istim uslovima**. Dozvoljavate umnožavanje, distribuciju i javno saopštavanje dela, i prerade, ako se navede ime autora na način određen od strane autora ili davaoca licence i ako se prerada distribuira pod istom ili sličnom licencom. Ova licenca dozvoljava komercijalnu upotrebu dela i prerada. Slična je softverskim licencama, odnosno licencama otvorenog koda.