The influence of the grammatical structure of L1 on learners' L2 development and transfer patterns in ESL academic writing: A comparative study

(A case of Chinese and Czech speakers)

by

Anna P. Kosterina

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF ARTS

Major: Teaching English as a Second Language/Applied Linguistics
(Computer Assisted Language Learning)

Program of Study Committee:
Dan Douglas, Major Professor
Nick Pendar
Geoff Sauer

Iowa State University

Ames, Iowa

2007

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

CALL	**Computer assisted language learning**
CA	**Contrastive analysis**
CIA	**Contrastive interlanguage analysis**
CEA	**Computer-aided error analysis**
CL	**Classifier**
CLC	**Computer learner corpora**
DDL	**Data-driven learning**
EA	**Error analysis**
EFL	**English as a foreign language**
ELT	**English language teaching**
ESL	**English as a second language**
ESP	**English for specific purposes**
EXP	**Experiential aspect**
IL	**Interlanguage**
L1	**First language**
L2	**Second language**
LLC	**Longman Learner Corpus**
NLP	**Natural language processing**
NNS	**Non-native speaker**
NS	**Native speaker**
POS	**Part -of -speech**
PROG	**Progressive aspect**

PRT	**Sentence final particle**
SLA	**Second language acquisition**
TL	**Target language**

ABSTRACT

This study investigates whether the grammatical structure of learner's first language (L1) plays a role in English as a second language (L2) development and can result in transfer into L2 writing. The study aims to find patterns of language use and error in learners of English as a second language with respect to their native language by studying a corpus of writings produced by such learners. The main focus of the study is on examining the manually annotated part of the section C of the Longman Learners' Corpus (LLC) corpus (http://www.longman-elt.com/dictionaries/corpus/llcotn.html) which includes English writing samples from the native speakers of Czech and Chinese varieties for the evidence of transfer pertaining to several specific grammatical features. The selection of features for the statistical analysis was based on previous research and the typological differences between the two languages investigated. Czech and Chinese were selected to represent Indo-European and Sino-Tibetan language groups, which are typologically very different from each other and are appropriate for the research goal.

The results of the statistical analysis show that grammatical structure of learners' L1 does have an effect on learners' L2 writing and can result in developmental and persistent errors which are particular to each language group. More importantly, the results of the study provide empirical support in favor of the argument that one L1 background group is prone to use certain grammatical, lexical, and textual organization patterns more frequently in L2 environments in comparison to the other L1 background, which is an important finding for the field of second language acquisition (SLA) in terms

of its potential implications for the Intelligent Computer Assisted Language Learning (ICALL), material design, and teacher training.

INTRODUCTION

1.1 Purpose of the Study

This study sets out to provide solid empirical support in favour of the position which holds that the grammatical structure of learner's first language (L1) plays a role in learners' English as a second language (L2) development and can result in transfer into L2 writing, and to identify the error patterns as well as the linguistic patterns that might be influenced by the learners' L1s by employing descriptive and inferential statistical analysis. I will present the findings which provide empirical support in favour of this view.

The overall goals of this research are to (1) determine differences in the linguistic data from speakers of the Chinese varieties versus data from speakers of the Czech varieties, (2) assist in the development of a valuable resource for applied linguistic research that would be used for Intelligent Computer Assisted Language Learning (ICALL) and traditional material development, and (3) provide more insights into second language acquisition (SLA) in general.

SLA studies traditionally focused on exploring sequences of learners' L2 development with relation to specific features such as question formation, negation, tense, collocations, and complex clauses. SLA studies of sequences of learners' L2 development as well as contrastive interlanguage studies that aimed to explore the grammatical and syntactic patterns in L2 writing have had a limited impact up to this day, due to a number of factors which undermined the generalizability of their

findings (Pendar and Chapelle, In press). Using a large computerized and diverse learner corpus such as Longman Learner Corpus (LLC) carries a great potential in terms of its capability to avoid a number of methodological problems and can aid foreign and second language teaching.

The assumption that learners of English are heterogeneous—that they don't learn under the same conditions, under the same educational system, with the same amount and quality of exposure to the target language (TL) —naturally leads to the conclusion that the pedagogical approaches in language teaching in general are driven toward more learner-centred approaches. Another important difference in the heterogeneity of their L1 backgrounds, which becomes most apparent in English as second language (ESL) settings and usually carries less of an importance in English as foreign language (EFL) settings, is the variety of L1 backgrounds of the learners in many ESL classrooms, especially in the university settings. Using a computerized and diverse learner corpus such as LLC carries a great potential in terms of the possibilities that it offers for data-driven learning (DDL), ICALL, and material development. It becomes self-evident that linguistic research of this type is integral to the development of more adaptive technologies for language teaching and assessment (Granger, 2002).

1.2 Historical Overview

Corpus linguistics can be considered a relatively new subfield of applied linguistics, since it only started to develop rapidly with the availability of modern-day computer technology. One of the main reasons for the exploration of learner corpora

in the field is the ability to extract information about various aspects of learner language. Since computerized learner corpora (CLC) tend to be much larger than the other collections of SLA texts, they can provide researchers with a more reliable way of extracting linguistic data and making generalizations based on that data.

Earlier SLA studies based on learner corpora were able to make a considerably limited contribution in terms their findings, since computerized corpora were not readily available at that time. Consequently, the traditional error analysis (EA) was usually based on relatively small collections of language data, which often lacked diversity in terms of their linguistic characteristics and, as a result, the finding of the studies based of such linguistic data carried a low degree of generalizability (Granger, 1998). Recent technological progress completely turned around the reputation of EA in the fields of applied linguistics and SLA, paving the road to redefining and repurposing the outlook on other methodological approaches that were discarded as having a low degree of generalizability and being not particularly useful, such as CA and contrastive interlanguage analysis (CIA).

Applications of SLA research based on learner corpora were up to this day limited due to two major methodological problems: (a) small sample size undermines generalizability of the findings, and (b) findings in SLA research are often based on qualitative data analysis (Pendar and Chapelle, In press). More recent studies based on computerized learner corpora have not had any lasting impact, mainly due to the abovementioned methodological issues (Pendar and Chapelle, In press).

Due to recent technological advances such as the ubiquitous presence of various kinds of computerized corpora, natural language processing (NLP)

applications, and standard text retrieval software, contemporary EA is regaining popularity among today's methodological approaches in SLA. Error tagging became highly standardized and well-documented, with clearly defined error categories which leave very little room for the ambiguity associated with earlier error classification systems. Concordance tools which revolutionized language teaching and learning (Salaberry, 2001) allow for presenting any lexical item within the context of a phrase, a sentence, or entire text or paragraph (Granger, 2002). Availability of part-of-speech (POS) tagging systems and other software tools promise unlimited opportunities for numerous types of linguistic analysis.

1.3 Research Questions

The present study investigates whether grammatical structure of learners' L1 plays a role in learners' English L2 development and can result in transfer into L2 writing, and aims to find patterns of language use and error in learners of English as a second language with respect to their L1s through analyzing a corpus of writings produced by such learners. Specifically, there are differences in the types of text organization and specific grammatical and lexical patterns that one L1 background group is more inclined to use in the L2 environment in comparison to the other L1 background group.

This study is guided by a central research question: *Does grammatical structure of learners' L1 play a role in learners' English L2 development, and can it result in transfer into L2 writing?* Seven more specific research questions stated below explore aspects of this central question.

Czech and Chinese were selected to represent Indo-European and Sino-Tibetan language groups which are typologically very different from each other and, hence, are appropriate for the research purpose. Czech and Chinese are distinct from one another linguistically in terms of the linguistic structure and culturally in terms of their rhetorical traditions, and hence the writers from these two L1 backgrounds might be prone to exhibiting different grammatical and error patterns in their L2 writing.

The first grammatical difference between the two L1s to be investigated involves isolating languages versus inflectional languages. Chinese is a highly isolating language, whereas Czech is a prime example of a highly inflectional language. Because isolating languages do not mark words morphologically, they award a lot more importance to syntactic rules, which can become extremely complex (e.g., the sentence word order carries a lot more importance in English than in any of the Slavonic languages, and carries much more importance in Chinese than even in English). However, all Indo-European languages are more or less inflectional, and Slavonic languages (e.g., Czech, Slovak, and Russian) are highly inflectional, which is why they almost never imply a strict direct word order in a sentence (e.g., most words can be moved around in any given sentence in a Slavic language without obscuring the relationships among the words within the sentence). Hence, it was hypothesized that the overall number of sentence word-order errors would be higher in the English interlanguage (IL) of Czech and Slovak learners than in the English IL of Chinese learners for learners at the same overall level of proficiency in English. Research Question 1 follows:

> Research Question 1: *Will the overall number of* ***sentence word-order errors*** *be higher in the English IL of Czech and Slovak learners than in the English IL of*

Chinese learners for learners across levels of proficiency in English?

The second grammatical distinction to be investigated involves topic prominence. Languages usually get classified into four groups: (1) topic-prominent, (2) subject-prominent, (3) both topic- and subject-prominent, (4) neither topic- nor subject-prominent (Li and Thompson, 1976). Chinese is a clear example of a topic-prominent language and employs topic-comment relationships as one way to determine the relationship between words in the sentence.

Czech writers also tend to topicalize or at least make topicalization attempts in their writing. Czech and its varieties, as well as other Slavic languages, are highly inflectional which allows for a lot of syntactic movement within each given sentence. Due to this syntactic "flexibility," Czech learners are likely to make various topicalization attempts in their L2 writing, which could be attributed to the transfer of the L1 grammar. Therefore, it appears that both L1 groups might be prone to some topicalization attempts in their L2 writing due to the topic-prominence peculiar to both Czech and Chinese. However, Czech and Chinese are still typologically very different languages and, consequently, the specific ways in which they deal with topicalization can be quite dissimilar. Hence, it would be interesting to investigate topicalization in both L1 groups. Drawing on the typological considerations presented above, it was hypothesized that there should not be a significant difference in the number of preposition errors in the English IL of Czech and Slovak learners as compared with the English IL of Chinese learners across levels of proficiency in English. This leads to the Research Question 2:

> Research Question 2: *Will there be a difference in the* ***number of topicalization attempts*** *made by Chinese and Czech writers across levels of proficiency in*

English?

The third grammatical area to be investigated includes errors associated with expletive subject problems for the analysis within the Czech learner group, which has a very similar justification. Due to its highly inflectional nature, Czech and its varieties do not tend to use expletive subjects in their L1 grammatical constructions, simply because the Czech sentence structures do not require such grammatical category. Example *(i)* below shows an absence of expletive subject, as well as an absence of article.

(i) *Tam* *est'* *ŝkola.*
Over there is school.
'There is a school over there.'

English utilizes overt expletive subjects which are essentially non-referential noun phrases that are merely function words which fill a vacant subject position, as prescribed by the grammatical rules as in the following example:

(ii) *There is a school across the street.*
(iii) *It is windy today.*

As pointed out by Yip (1995), it has been argued by Li and Thompson (1976) that expletive subjects only exist in subject-prominent languages, such as English, and therefore it is questionable whether Chinese (as a topic-prominent language) lacks expletive subjects altogether, since the its structure does not necessarily require a presence of an overt subject category. However, Yip (1995), drawing on previous

research such as Gao et al. (1994) (as cited in Yip, 1995) recognizes that there are Chinese constructions which call for a non-referential null subject, as illustrated in the example below, adapted from Yip (1995):

(iv) PRO.EXP *lun* *ni* *xi* *wan.*
turn you wash dish
'It's your turn to wash the dishes.'

(v) *Zai* *xia* *yu* *le.*
PROG fall rain PRT
'(It's) raining.'

Furthermore, Czech and other Slavic languages, unlike other languages investigated; do not have to have direct word order in a sentence. This, however, does not mean that Chinese lacks subjects since, as stated in Yip (1995), in Chinese, subject always has a direct relationship with the verb. This is shown in the example illustrated in example below, adapted from Yip (1995, p. 75).

(vi) *Lisi* *wo* *jian* *guo* *Le.*
Lisi **I** **see** **EXP** **PRT**
'Lisi, I've already seen (him).'

Hence, it was hypothesized that the overall number of expletive subject errors would be higher in the English IL of Czech and Slovak learners than in the English IL of Chinese learners for learner across levels of proficiency in English, due to the highly inflectional grammatical structure of Czech compared to English and Chinese. Therefore, Research Question 3 follows:

Research Question 3*: Will the overall number of* ***expletive subject errors*** *be higher in the English IL of Czech and Slovak learners than in the English IL of Chinese learners for learner across levels of proficiency in English?*

Article errors are the fourth grammatical category investigated due to the fact that neither Czech nor any of its varieties include articles as a grammatical category. Chinese, on the other hand, utilizes word order, demonstratives, and classifiers to show definiteness (Yip, 1995). As illustrated in the example below, adopted from Yip (1995, p. 94), a noun phrase is definite if preceded by a classifier phrase that includes a demonstrative.

(vii) *zhe* *bun* *shu*
this CL book
'this book'

(viii) *nei* *zhang* *zhi*
that CL paper
'that sheet of paper'

Therefore, in Chinese, noun phrases which contain *zhe* ('this') and *nei* ('that') are definite. However, as illustrated in example below, adopted from Yip (1995, p. 94), a noun phrase which includes a classifier and a numeral but does not include a demonstrative is indefinite.

(ix) *yi* *ge* *ren*
One CL man
'a/one man'

(x) *san* *ke* *shu*
Three CL tree
'three trees'

In contrast to Chinese, Czech draws on the entire context on the sentence of discourse to infer definiteness, and it completely lacks the formal article category, as shown in the Example *(i)* above. Sometimes, however, definiteness is expressed by using certain determiners, such as equivalents of the English words *my* and *your*. Therefore, it is likely that L2 writers with Czech L1 background will have considerably greater difficulty with the use of articles compared to L2 writers with Chinese L1 background due to the fact that Czech has a 0 article category. Hence, it was hypothesized that the overall number of article errors would be higher in the English IL of Czech and Slovak learners than in the English IL of Chinese learners for learners across levels of proficiency in English. This leads up to the Research Question 4:

> Research Question 4*: Will the overall number of* ***article errors*** *be higher in the English IL of Czech and Slovak learners than in the English IL of Chinese learners for learners across levels of proficiency in English?*

Fifthly, Slavonic languages are usually characterized as languages that rely on extensive noun morphology. Czech and Slovak are more appropriate for the given research purpose other than any other highly inflectional language such as Russian

because they have, among other features, a fully developed case system that includes seven cases in comparison to Russian, for example, which is indeed typologically very close to Czech and Slovak but has a case system with six cases. The Czech case system includes nominative, vocative, accusative, genitive, dative, instrumental, and locative. It also includes three genders and a bipartite number system (Short, 1987). Every noun in Czech and Slovak is inflected by case, number, and gender.

As Slavonic languages are highly inflectional, the noun plural marking systems differ significantly from those of highly isolating Chinese. Table 1.1 below, adapted from Young (1993), provides an example of plural marking in Chinese and in two Slavonic languages.

Table 1.1: Final plural marking: a comparison of Chinese and Slavonic (i.e., Czech and Slovak) (adapted from Young, 1993, p. 84).

ENGLISH	CHINESE phonetic translation	CZECH	SLOVAK
a student	yīge xuésheng	student	Študent
two student**s**	*liǎnge* xuésheng	*dva* student***i***	*dva* študent***i***
many student**s**	*hěn duō* xuésheng	*mnoho* student***u***	*mnoho* študent***ov***
some student**s**	*yīxiē* xuésheng	*někteři* student***i***	*neketeri* študent***i***
Your student has arrived.	Nǐde xuésheng láile.	Tvuj student přijel.	Tvoj študent prišiel.
Your student**s** have arrived.	Nǐde xuésheng láile.	Tvuji student***i*** přijeli.	Tvuji študenti prišl***i.***

NB: Morphological marking of plural number is italicized

Young (1993) points out that "Chinese noun plurals are marked on only restricted classes of personal pronouns and vocatives and, in general, inflections in word or syllable final position are relatively rare in Chinese" (Young, 1993 p. 83). Hence, based on the explanation provided above and drawing on previous research (Young, 1988, 1991,

1993), it was predicted that—because in Czech noun suffixes carry a significant amount of meaning in contrast with Chinese, which never employs noun suffixes—the Chinese writers will have a significant difficulty with acquiring *s* plural inflectors compared to Czech writers. It was hypothesized that the overall number of errors with plural count nouns marked with *s* will be higher in the English IL of Chinese learners than in the English IL of Czech and Slovak learners for learners across levels of proficiency in English. This leads up to the Research Question 5:

> Research Question 5*: Will the overall number of errors with* ***plural count nouns marked with s*** *be higher in the English IL of Chinese learners than in the English IL of Czech and Slovak learners for learners across levels of proficiency in English?*

As discussed in the literature review, the results of a number of studies which analyzed the writing of ESL learners indicate that overpassivization appears to be among the top features indicative of language background (Cowan et al., 2003; Ju, 2000; Yip, 1995; Zolb, 1989); thus, this is the sixth grammatical category to be investigated in this study. Among the most well-known publications related to passivization and overpassivization by Chinese learners are the articles by Ju (2000) and the study by Yip (1995) that suggest that Chinese ESL writers tend to attempt to passivize unaccusative verbs for which transitive forms exist (such as *change* or *increase*) as well as unaccusative verbs for which transitive forms do not exist (such as *happen* or *appear*). Searching for the possibilities of L1 origin for this type of overgeneralization error, Zolb (1989) suggests that Chinese ESL learners somehow place a lexical rule within the same IL conceptual niche as the passive rule (as cited in Cowan et al., 2003). On the other hand, Yip (1995) suggests that this tendency, seen in Chinese ESL learners, can be

attributed to the possibility that Chinese learners somehow regard unaccusative verbs as underlyingly transitive. The latter explanation introduced by Yip (1995) as "transitivation hypothesis" has been given further support in a recent study by Cowan et al. (2003). Based on the array of studies exploring passivization in Chinese native speakers in their English production, it was hypothesized that the overall number of passivization attempts would be higher in the English IL of Chinese learners than in the English IL of Czech and Slovak learners for learners across levels of proficiency in English, which leads up to the Research Question 6.

> Research Question 6*: Will the overall number of* ***passivization attempts*** *be higher in the English IL of Chinese learners than in the English IL of Czech and Slovak learners for learners across levels of proficiency in English?*

Finally, Cowan et al. (2003) pose the argument that a number of preposition-related errors, such as using *interested at* in place of *interested in*, can be indicative of Korean or Chinese L1 background (i.e., Sino-Tibetan L1 backgrounds). It can also be argued that all ESL writers, regardless of their L1 background, usually make a significant number of prepositional errors, simply due to the absence of a clear-cut grammatical rule which would regulate the use of one preposition over another in English. Some of those errors are simply collocational as well. Unfortunately, no extensive studies exploring the issue were found, perhaps due to the same argument as in the sentence above. Based on this argument, it was hypothesized that there should not be a significant difference in the number of preposition errors in the English IL of Czech and Slovak learners compared with the English IL of Chinese learners across levels of proficiency in English, which lead up to the Research Question 7:

> Research Question 7*: Will there be any significant difference in the number of* ***preposition errors*** *in the English IL of Czech and Slovak learners comparing with the English IL of Chinese learners across levels of proficiency in English?*

1.4 Preview of the Study

Chapter 2, *Literature Review*, provides a historical overview of the development of EA and CA, and outlines the importance of Corpus Analysis and Learner Corpora and their use for Computer-Aided Error Analysis and Contrastive Interlanguage Analysis. Chapter 3, *Method*s *and Materials*, discusses the development of the annotation scheme and procedures and methods utilized to analyze the data. Chapter 4, *Results and Discussion*, presents the results of the statistical analysis and discusses the implications of the results. Finally, Chapter 5, *Conclusion*, summarizes the findings, presents the limitations, discusses the implications of the study, and gives suggestions for further research.

LITERATURE REVIEW

This chapter will provide a historical overview of the development of error analysis (EA) and contrastive analysis (CA). This chapter will also outline the importance of corpus analysis and learner corpora and their use for computer-aided error analysis and contrastive interlanguage analysis, and examine the implementation of these approaches in previous research as well as in contemporary studies in the field of applied linguistics.

2.1 Error Analysis

EA in its theoretical foundations relied on CA introduced by Lado (1957) and is a type of linguistic analysis which emerged in the applied linguistics field in the late 1960s. Although early EA evolved from CA, it was primarily concerned with discovering possible insights into leaner interlanguage system through finding the root of the error in a learner's native language—in contrast to CA which, at the time, solely focused on examining learner errors (Gass and Selinker, 2001). EA became one of the most prominent methodological approaches to linguistic analysis in second language acquisition (SLA) through the 1970s and into the 1980s.

Following the considerably lengthy period when CA and EA were very influential in the understanding and interpretation of second language acquisition, a period emerged wherein both approaches received much criticism due to their focus on erroneous forms that could result in overwhelming the learner with negative evidence without providing positive evidence (i.e., directing learner's attention to the instances of correct use of linguistic forms). Ellis (1994) argued that learner errors

are an initial attempt to systematize a target language (TL) rule and therefore should be seen as evidence of learning. Decontextualization of errors and the lack of standardization seen in many EA error typologies used at the time added to the negative perceptions that arose around EA and CA. Technology did not reach the point where computerized corpora became readily available. Hence, traditional EA was usually based on relatively small collections of language data which often lacked diversity in terms of its linguistic characteristics and, consequently, the findings of the studies based on such linguistic data carried a low degree of generalizability (Granger, 1998). These limitations, coupled with the inability to account for other linguistic factors such as avoidance, resulted in both approaches falling out of favor in SLA—but only for a short while, until the emergence of computer learner corpora (CLC) and computer-aided error analysis.

However, due to recent technological advances—such as the ubiquitous presence of various kinds of computerized corpora, NLP applications, and standard text retrieval software—contemporary EA is regaining popularity among today's methodological approaches in SLA. Error tagging became very standardized and well-documented, with clearly defined error categories which leave very little room for the ambiguity associated with earlier error classification systems. Concordance tools which revolutionized language teaching and learning (Salaberry, 2001) allow for presenting any lexical item within the context of a phase, a sentence, or entire text or paragraph (Granger, 2002). Availability of part-of-speech (POS) tagging systems and other software tools promise unlimited opportunities for numerous types of linguistic analysis.

Some of the well-structured error classification systems which are well-principled and avoid overlap in categories are being utilized in today's corpora analysis. For instance Corder (1974) developed a framework for traditional error analysis that included three basic stages: (1) effective recognition, (2) description, and (3) explanation of errors. This typology was later further expanded by Lennon (1991) to include 5 stages: (1) selection of a corpus of language, (2) identification of errors in the corpus, (3) classification of the errors identified, (4) explanation of the psycholinguistic causes of the errors, and (5) evaluation or error gravity ranking of the errors (Lennon, 1991).

Additionally, Corder (1974) argues that there is an important difference between spontaneous text production and controlled text production in terms of EA—the former being error-avoiding and the latter being error-provoking (Corder, 1974)—which in itself is an argument for utilizing computerized corpora of learner-written language for contemporary EA. Unquestionably, the existence of errors in L2 production is only one of the important indicators of the learner's linguistic competence. Needless to say, the absence of errors can be an indicator in itself of the learner's linguistic competence; however, it can also be just an indicator of an avoidance strategy which is commonly seen in beginner level learner writing. Inability to account for such common phenomena in learner language as avoidance was among the major criticisms of traditional EA.

Heift and Schultz (In press) point out that when utilizing contemporary error analysis with parsers, the above-described problem can be solved as follows:

> …if the parser does not only output an error analysis and feedback but also stores parsing results in a learner model. Consequently, the model can maintain a record of lexical and syntactic constructions used by the learner. Given standard frequency of certain constructions in a given type (e.g., genre) or language learning task, the student model can then flag overuse or (partial) avoidance of some constructions. (Heift & Schultz, In press, p. 162)

Recent technological advances completely turned around the reputation of EA in the fields of applied linguistics and SLA, paving the road for redefining and repurposing the outlook regarding other methodological approaches that had been previously discarded as having a low degree of generalizability and being not particularly useful, such as CA and contrastive interlanguage analysis (CIA).

2.1.1 Contrastive Interlanguage Analysis

Contrastive interlanguage analysis (CIA) usually presumes comparison of two or more groups of language uses or learners against each other. The two types of comparisons most commonly seen in CIA are native speaker language data compared with non-native speaker language data (NS/NNS), or non-native speaker language data compared with the language data produced by another population of non-native speakers (i.e., NNS/NNS comparisons) (Granger, 2002).

As pointed out by Granger (2002), NS/NNS data comparisons presume a comparison of linguistic features which are of interest to researchers through contrasting NNS linguistic data with NS linguistic data. NNS/NNS comparisons in CIA usually involve the comparison of language data elicited from two populations of non-native speakers who carry different learner characteristics, such as different L1 backgrounds. According to Granger (2002), these types of comparisons that involve two or more NNS groups with dissimilar L1 backgrounds provide researchers with

the opportunity to isolate linguistic features which are associated with only one of these NNS groups and not the other NNS group(s). These language-use features can be indicative of L1 influence. On the other hand, linguistic features seen in several NNS groups are likely to be developmental (Granger, 2002). Knowledge of learner interlanguage is one of the integral parts of the linguistic analysis in those types of NNS/NNS comparisons.

2.1.2 Error Analysis and Interlanguage

The term *interlanguage* is usually used to describe a certain target language variety that a learner uses at a specific point in time or to describe learner target language (TL) development over a certain time period (Ellis, 1994). The term was introduced by Selinker (1974) and later picked up by Corder (1981). Learner interlanguage in essence is a system of rules that positions itself somewhere between the native language and the TL of the learner on the so-called *learner language continuum* (Heift & Schultz, In press).

According to Selinker (1974), there are five fundamental processes underling interlanguage processes: (1) language transfer, (2) transfer training, (3) strategies of second language learning, (4) strategies of second language communication, and (5) overgeneralization of TL linguistic material. Selinker (1992) modifies the above-mentioned classification to include three main processes that underlie interlaguage development: (1) language transfer, (2) simplification, and (3) overgeneralization. Language transfer is the most noteworthy for this study since it "transfers" over the rules from L1 into the interlanguage constructed by the learner in at attempt to

decipher the given TL, whereas the first two interlanguage processes—overgeneralization and simplification—merely "modify" the rules from the TL system[1] .

As rightfully noted by Heift & Schultz, (In press), in certain cases, it becomes quite clear that more than one interlanguage process can be credited for causation of one specific error. Moreover, even when operating under the assumption that interlanguage is considerably systematic—which makes it possible to be utilized in parser-based error analysis (EA)—certainly there are instances that the same surface error can be attributed to different interlanguage processes.

In regard to language transfer, the situation gets even more complicated by the fact that transfer can occur not only between the TL and native language(s) but also between previously learned languages and the TL. This complication makes it even more difficult to classify a transfer process as a simplification or as an overgeneralization.[2] In a recent work, Heift and Schultz (In press), also drawing on Ellis (1994), arrive at the conclusion that "[s]implification and overgeneralization attempt to provide a psycholinguistic account of interlanguage phenomena. Hence, they are also meant to identify causes of deviations in the interlanguage grammar from the grammar of the target variety" (Heift & Schultz, In press, p. 165).

---

[1] According to Heift and Schultz (In press), "[s]implification refers to the writer ignoring certain rules in order to save processing time (in a psycholinguistic or cognitive sense)…" (p.165), whereas overgeneralization in linguistics is usually defined as an application of a general grammatical rule across all members of a grammatical class in case of an exception.

[2] Ellis (1994) provides an extensive discussion of the above-described problems and essentially comes to the conclusion that it sometimes might not be possible to establish the cause of error solely on the basis of these processes.

Despite the apparent fact that individual variation (in terms of deviation from the TL construction) cannot be completely deciphered through the attribution to interlanguage processes, the insights into underlying interlanguage processes can (1) provide evidence of the systematicity of the interlanguage, and (2) allow for exploration of various trends in the learner interlanguage processes, such as degree of variation or similarity between individual learners or groups of learners.

2.1.3 Computer-Aided Error Analysis

Although EA has been out of favor in the field of SLA research for a while, recently it has been drawing more attention from researchers in light of new technological developments such as the application of natural language processing (NLP) techniques in CALL and the use of computerized learner corpora, which have been discussed in detail above. It appears that error analysis has been resurrected under the new name of "computer-aided error analysis."

Emergence of computer-aided error analysis gave a new interpretation to EA as a branch of applied linguistics. Most of the downfalls of traditional EA, as previously discussed, were eliminated with the help of newly available technological advances in the field. The frameworks used for error classification and error tagging became much more standardized, and erroneous items are now being analyzed in the context in which they appear (i.e., in the context of a sentence, paragraph, or entire text in which they appear, side-by-side with the multiple instances of correct language use of the same and other linguistic items) (Granger, 2002).

Granger (2002) highlights two current methodological approaches to computer-aided error analysis. One frequently used method in computer-aided error analysis consists of retrieving a pre-selected linguistic item/feature (e.g., word/word category, syntactic structure, collocation) and scanning the entire corpus for the erroneous occurrences of this linguistic feature, usually with the help of text retrieval software. As Granger rightfully points out, this method—although not particularly time-consuming—is inherently limiting the research to those pre-selected linguistic features. The other method is less-commonly encountered due to its labor-intensiveness but can be much more rewarding due to its capabilities for supporting a considerably wider range of research opportunities. The alternative method presumes manual error tagging of the corpus for all the errors, or at least for all the errors which are of interest to the researcher. This process in certain cases can be aided with an automatic error tagger. Despite all of these obstacles, the second approach carries a much greater potential in terms of its applications.

Heift and Schultz (In press) draw attention to the fact that the notion of error is not by any means objective (i.e., there appears to be considerable variation in results among the grammaticality judgment tasks conducted even on native speakers). These differences can be attributed to dialectal differences and partially to the variation in socioeconomic and educational backgrounds among the native speakers performing the grammaticality judgment tasks; the same holds true when it comes to parsers that often overlook pragmatic and semantic errors.

Corder (1974) makes a distinction between errors, mistakes, and lapses—errors being the central category since, according to Corder's classification, only

"errors" are indicators of the possible gaps in the learner competence or misconceptions about the rules of L2 in the learners' interlanguage. Lapses and mistakes are performance-based and can be triggered by incidental influences of external conditions (e.g., due to lack of concentration, headache, etc.).

There have been several other error classifications developed in an exertion to address the specific research goals and needs of individual CALL projects. However, according to Heift and Schultz (In press), some of those classifications do not exhibit a high degree of systematicity in contrast with better-structured approaches, such as Corder's and later ones based on Corder's (1974) classification[3] .

As mentioned above, apart from the existing negative attitude pertaining to error analysis, a large number of CALL programs (particularly the ones which managed to successfully incorporate NLP techniques) are based on the underlying principle that error correction, whether explicit or implicit, carries a positive effect on learners' L2 development. A number of very reputable SLA studies supply concrete empirical evidence in support of it, such as studies by Nagata (1995, 1997, 2002). As pointed out by Cowan et al. (2003), the majority of the research studies related to the use of CALL examine the topic in regard to learners of low to intermediate proficiency, which explains the lack of evidence pertaining to the long-term effects of instructional approaches that embrace error correction. Hence, the long-term effects of error correction (L1 transfer errors and "persistent" errors) need to be explored further in SLA, traditional material and CALL development, and in teacher training.

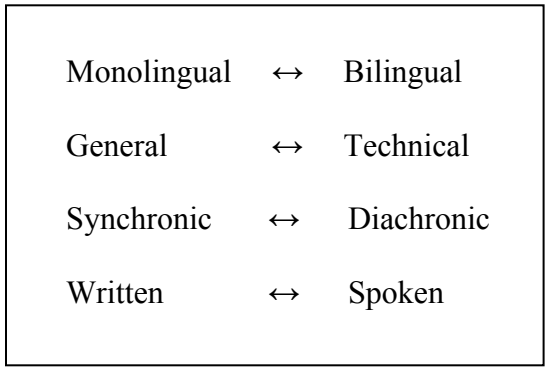
[3] Additionally, Heift and Schultz (In press) provide a comprehensive survey of a number of other error classification systems.

Apart from the criticism of EA discussed previously, EA continues to carry an important role in the field of SLA (Granger, 2002)—particularly in light of the rapidly developing corpus-based linguistics due to the technological opportunities which were unavailable even a decade ago, such as computerizing learner corpora and implementing NLP techniques in learner language analysis.

2.2 Corpora

Corpus linguistics can be considered a relatively new subfield of applied linguistics since it only started to develop rapidly with the availability of modern day computer technology. Any corpus is essentially a principled collection of linguistic data (Sinclair, 2004; Leech, 1998). Currently, there are two varieties of computerized corpora: (1) native speaker corpora, which are collections of language data produced by native language speakers, and (2) learner corpora, which can also be referred to as *interlanguage corpora* or *L2 corpora* (Granger, 2003). Corpora in general and learner corpora specifically are typically classified into dichotomous categories. As this study is focused on investigating learner corpora, Figure 2.1 below, adopted from Granger (2002), provides a simple and useful classification of learner corpora typology (Granger, 2002 p. 11). Granger also points out that monolingual, general, synchronic, and written corpora (i.e., features of the left side of Figure 2.1) tend to get more attention in recent CLC research.

Figure 2.1 Learner corpus typology (adapted from Granger, 2002, p. 11)



Despite the fact that research based on monolingual corpora is more commonly seen among the recent CLC publications, a number of studies have been conducted which give empirical illustration to the advantages of using bilingual corpora. Among others, Danielsson and Ridings (1996) utilized parallel corpora in educational training programs designed for translator training, and advocate for using bilingual corpora for translating and teaching translation (as cited in Nerbonne, 2003). The results of their studies suggest that the students who were studying translation benefited from being able to access a large resource of linguistic data which allowed them to locate atypical translation equivalences (Nerbonne, 2003).

One of the main reasons for using bilingual corpora is that, apart from providing authentic language data, it sufficiently increases the input comprehensibility for the learner by providing an L1 translation side by side with an L2 text. On the other hand, Nerbonne (2003) provides a word of caution by stating "that the use of bilingual corpora only makes sense if good software is available to support the sorts of searchers which instructors and students wish to conduct" (Nerbonne, 2003, p. 683). However, there is no doubt that bilingual corpora can

provide advanced students with the linguistic information that otherwise would be unavailable to them.

At the present time, corpora tend to be general rather than technical or genre-specific (Granger, 2002). Currently, ESP learner corpora are rarely encountered; the *Indiana Business Learner Corpus (IBLC)*—which was compiled by Connor et al. from the materials gathered from US-Belgian-Finnish writing project—was successfully used in a Connor et al. (2002) study of business English.

According to Granger's (2002) corpora classification provided above, synchronic corpora represent language use at a particular point in time. Granger (2002) defines diachronic corpora as "…corpora which cover the evolution of learner use…" (Granger, 2002 p. 11). Longitudinal corpora are not as frequently collected due to the obvious difficulties of compiling a corpus over an extended period of time.

Cowan et al. (2003) pose a convincing argument that, despite of the fact that there is no difference in the degree of validity between spoken and written corpora, the two main advantages of written corpora over spoken are as follows:

- L2 grammar errors in written learner corpora exhibit the ultimate level of learner competence since the learners are presumed to have had multiple opportunities for editing their writing. This contrasts with spoken corpora where a learner's utterance can be an accidental record of a learner's performance, which is known to not always be fully reflective of a learner's linguistic competence.
- The relative simplicity of the procedures employed to convert written corpora into electronically stored data allow for subsequent error

analysis through tagging, using concordancing programs and NLP techniques (Cowan et. al., 2003).

Needless to say, written corpora are much more common, due to the intrinsic difficulties associated with collecting spoken language data.

2.2.1 Corpus Analysis

In the light of recent technological advances, the use of electronic corpora offers great potential in terms of SLA research, due to their inherent features. For example, any electronic corpus is quickly and automatically searchable, providing great potential for language researchers and educators. Specifically, the "searchability" of electronic corpora brings an array of research opportunities that can result from various linguistic analyses that can be performed on it with a variety of currently available linguistic software tools that provide "for quick and efficient manipulation of the data through search, count, and sort functions and NLP programs which enrich the data with linguistic information" (Granger, 2003 p. 465). The authenticity of the language material that computerized corpora can supply was also noted by a number of researchers (Granger 1998, 1999, 2002, 2003; Nerbonne, 2003; Pravec, 2002). For instance, Nerbonne (2003) states that " [c]orpora are valued for providing access to authentic language use, unmediated by grammarians' theories, prescriptivists' tastes, pedagogy's traditions, or even lexicographers' limitations" (Narbonne, 2003, p. 681).

In contemporary applied linguistics, learner corpus analysis is undertaken from two different methodological standpoints, computer-aided error analysis and

contrastive error analysis. Computer-aided error analysis involves the application of computer applications to assist with data analysis. This method also might include standard text retrieval software for information retrieval. On the other hand, contrastive error analysis is usually concerned with carrying out comparisons between the linguistic data between two or more groups (i.e., NNS/NS comparisons and NNS/NNS comparisons) (Granger, 2002). This, however, does not imply that the two approaches are mutually exclusive.

On a cautionary note, it is important to point out that the pedagogical use of computerized learner corpora should be principled, drawing on SLA theory and educational psychology. Nerbonne (2003) notes that "[i]t is very clear that corpora can only be useful for advanced students—beginners would simply understand nothing they saw" (Narbonne, 2003, p. 681). Furthermore, it appears imperative that the learners need to achieve a certain level of proficiency in L2 before they are at the point where their L2 development can be aided by the guided exposure to corpus-based language learning activities.

2.2.2 Learner Corpora

As pointed out by Granger (1998), the primary goal in compiling a learner corpus is to gather authentic and objective L2 data that can aid in describing learner language. This type of learner language data derived from learner corpora also offers an enormous potential for further exploration of theoretical issues and can further lead to the development of pedagogical applications that can aid language learners. Apart from the ease of searchability, learner corpora can provide other advantages to

applied linguistics research which were not available with the L2 language data formats before computerized corpora were introduced. One of the main reasons for the exploration of learner corpora in the field is the ability to extract information about various aspects of learner language. Since CLC tend to be much larger than the other collections of SLA texts, they can provide researchers with a more reliable way of extracting linguistic data and making generalizations based on this type of language data (Granger, 2002).

The importance of learner corpora as opposed to native speaker corpora lies in the fact that they provide divergence from the standard or accepted form of a linguistic utterance in any given L1, when judged by native or native-like speakers of that language (Pravec, 2002). Granger (2002) argues for the use of learner corpora for investigating NNS errors. She notes that, despite the negative attitudes towards the traditional methods of error analysis, it cannot be disregarded in second language learning and teaching and is "a key aspect of the process which takes us towards understanding interlanguage development and one which must be considered essential within a pedagogical framework" (Granger, 2002, pp. 13–14). Additionally, Appendix A provides an overview of currently available learner corpora which can facilitate preliminary researchers' search for the learner corpora most appropriate for their research goals.

In addition to the advantages mentioned above, learner corpora can provide an unlimited amount of naturally occurring authentic language data, which welcomes numerous opportunities for research. Investigations of computerized learner corpora open up a window to the entire learner's interlanguage system as well as to learner

errors in general and, potentially, allow for the detection and classification of the patterns among those learner errors (Granger, 1998). In terms of more practical applications, corpus-based research has already lead to the creation of various EFL tools—such as *Electronic Language Learning and Production Environment* (http://www.longman-elt.com) developed by the Hong Kong University of Science and Technology (HKUST)—and a number of other skillfully implemented and well-known tools, such as the *Longman Dictionary of Contemporary English (LDOCE)* (http://www.longman-elt.com/dictionaries/research/dictres.html).

2.3 Learner Corpora and Error Analysis

Even during the period when error analysis fell our of favor, a number of prominent researchers such as Ringbom (1987) and Ellis (1994) continuously acknowledged its usefulness (a) from the perspective of the researcher, because it provided better insights into learners' interlanguage systems and their developments, and (b) from the perspective of language learners since, as pointed out by Granger (2003), "a detailed description of learner errors cannot but contribute to one essential FLT aim—that of helping learners to achieve a high level of accuracy in the language" (p. 466).

Apart from the criticism of EA discussed in the section devoted to it, EA continues to carry an important role in the field of SLA (Granger, 2002), particularly in the light of rapidly developing corpus-based linguistics due to the technological opportunities which were unavailable even a decade ago, such as computerizing learner corpora and implementing NLP techniques in learner language analysis. Heift

and Schultz (In press) point out that "[m]ost often, the errors made by language learners reflect hypotheses of linguistic norms that learners form about the L2" (pp. 154–155). By documenting errors in NNS-produced written texts, we can arrive at a deeper understanding of NNS's interlanguage processes; this is where contemporary EA carries tremendous potential, especially pertaining to the field of corpus linguistics. Table 2.1 below provides an overview of the studies most relevant to the discussion.

Table 2.1: Overview of the studies related to error analysis and learner corpora

Study/Author	Rationale/ Research Question(s)	Method	Result(s)/Conclusion
1. Investigating the Promise of Learner Corpora: Methodological Issues, by Pendar and Chapelle (In press).	Study explores methodological issues associated with learner corpora	ICLE corpus statistically analyzed based on a large number of predictors, including lexical and quantitative features, and explored issues such as identification of learner levels	Results suggest the need for a larger corpus with more systematically sampled subcorpora from across language groups and shows promise for the quantitative and lexical variables and machine learning statistical procedures
2. Four Questions for Error Diagnosis and Correction in CALL, by Cowan et al. (2003)	Whether persistent L2 errors can be corrected, and what types of computer feedback are most efficient for focusing students' attention on a task, and the evaluation of CALL programs focused on error correction	Contrastive interlanguage methodology; relatively small corpus of Korean English-learner writing collected; based on error counts performed, persistent errors identified; no statistical analysis	A large corpus of L2 learner errors is shown to be highly beneficial for identifying persistent L1 transfer
3. Error-Tagged Learner Corpora and CALL: A Promising Synergy, by Granger (2003)	Research aimed to produce a learner corpus-informed CALL program for learners of French	Manually annotated corpus was run through standard text retrieval software; error statistics were extracted; concordance-based analysis of specific error types was performed	The results were implemented in CALL exercises and were used to improve the error-diagnosis system integrated in the CALL program

Table 2.1 (continued)

Study/Author	Rationale/ Research Question(s)	Method	Result(s)/Conclusion
4. Modality in Advanced Swedish Learners' Written Language, by Aijmer (2002)	Compare modal forms, meanings, and uses in comparisons produced by NNSs and NSs	Contrastive interlanguage methodology; compared different types of interlanguage	Results show global overuse of modal auxiliaries by L2 writers which could be partly developmental and partly interlingual
5. Using Bilingual Corpus Evidence in Learner Corpus Research, by Altenberg (2002)	Compare model forms and their uses in comparisons produced by NNS and NS; Research Question: How can the Swedish learners' overuse of *make* be explained? Hypothesis: Overuse of causative *make* with adjective complements by Swedish L2 writers is due to L1 transfer	Cross-linguistic analysis/ translation	Results provide support for the hypothesis: the overuse is caused by an overgeneralization of the cross-linguistic similarity between *make* and *gÖra*, the most common unmarked equivalent of *make* in Swedish
6. A Corpus-Based Study of the L2-Acquisition of the English Verb System, by Housen (2002)	How learners acquire basic morphological categories of English; what stages of development can be seen in their acquisition; how L2 learners map these forms, and what stages can be observed in the development of these form-meaning relations	Cross-linguistic analysis/ translation	Results generally confirm the general order of emergence of the formal morphological categories posited by previous studies, but reveal significant variation at the level of individual learners and that formal variation precedes functional use
7. Overpassivization Errors of Second Language Learners: The Effect of Conceptualizable Agents in Discourse, by Ju (2000)	Do conceptualizable agents in the discourse play a role in English L2 overpassivization errors (by Advanced Korean learners of English)?	Contrastive interlanguage methodology; Advanced Chinese learners of English were given grammaticality judgment tasks which involved choosing active/passive form in a context of a sentence	Results indicate that learners transitivize unaccusative verbs before they passivize them and that the degree of transitivation varies depending on the presence of conceptualizable agents in the discourse
8. Interlanguage and Learnability: From Chinese to English, by Yip (1995)	Study explores overpassivization of English verbs by Chinese learners.	Contrastive interlanguage methodology; data collected through questionnaires that present grammaticality judgment tasks	Chinese ESL writers tend to attempt to passivize unaccusative verbs for which transitive forms exist (such as *change* or *increase*) as well as unaccusative verbs for which transitive forms do not exist

Table 2.1 (continued)

Study/Author	Rationale/ Research Question(s)	Method	Result(s)/Conclusion
9. Functional Constrains on Variation in Interlanguage Morphology, by Young (1993)	Do functional constrains affect variation in interlanguage morphology? (in the case of spoken English IL of learners with Czech/Slovak and Chinese L1 background	Contrastive interlanguage methodology; data collected through interviews with both Czech and Chinese and analyzed through type/token ratio and VARBRUL multi-variable procedure	The author concludes that functional constraints have little effect on variation in L1 morphology
10.The Interpretation of English Reflexive Pronouns by Non-Native Speakers, by Thomas (1989)	How do Chinese learners of English interpret reflexive pronouns?	Chinese L2 learners and NSs responded to a 30-item questionnaire to identify the antecedent of a reflexive pronoun	The Chinese L2 learners do not seem to transfer L1 grammar into L2, nor do they recapitulate the course of L1 acquisition
11. A comparison of Spanish-English and Japanese-English second language continuum: negation and verb morphology, by Stauble (1984)	Study attempts to compare six Spanish and six Japanese learners to establish a common English learning continuum and determine the extent to which native language differences affect this process	Study utilized cross-linguistic design to compare six Spanish and six Japanese learners of English and their development of negation across proficiency levels; no statistical analysis	Results suggest that a learner's negation characteristics can be used as a gross measure of his/her English verb morphology development

Aijmer's (2002) and Housen's (2002) studies utilize CA methodology; the first one is cross-sectional, and the second one longitudinal. Both of those studies, besides their relevance to the present study, are also important for being among the first examples in contemporary applied linguistics to reinterpret CA as a valid methodological approach, particularly in the context of computerized learner corpora (Granger, 2002).

In Thomas's (1989) study, Chinese L2 learners and native speakers responded to a 30-item questionnaire to identify the antecedent of a reflexive pronoun, in an attempt to examine the Chinese learners' interpretation of English reflexive pronouns.

The author arrived at the conclusion that Chinese L2 learners do not seem to transfer L1 grammar into L2.

In addition to the studies that examine errors cross-sectionally and contrastively, there is empirical evidence to support the view that some of the learner errors may be persistent (Cowan et al. 2003) (i.e., they will not disappear over time and with more exposure to more TL). Schwartz and Sprouse (1996) argue that one possible cause of L1 transfer errors is the structure of learners' interlanguage, which allocates values to different parameters during the very early stages of acquisition. There is also a common assumption that directing learners' attention to transfer errors can result in the short-term improvements at best, regardless of the fact that there is hardly any empirical evidence in its support.

In their study that addresses diagnosis and correction of persistent L2 learner grammatical errors based on the corpus collected from L2 Korean learners, Cowan et al. (2003)—drawing on Granger (1998, 2003)—develop a set of characteristics which would allow a corpus to properly investigate these types of grammatical errors. These characteristics are as follows:

> "It should (a) encompass different levels of proficiency, (b) consist of extensive samples of learner language that facilitate analysis of grammatical errors caused by phenomena beyond the boundaries of the sentence, (c) be labeled so that researchers and material developers can determine whether the total number of errors of a given type is produces by a small number of learners or by many different learners, and (d) be fairly large" (Cowan et. al., 2003, p. 452).

Cowan et al. (2003) point out that Yip's (1995) study, discussed in more detail in the subsection 1.3, *Research Questions*, of the first chapter, clearly demonstrates "that negative evidence can be effectively applied in certain cases but that other errors

will correct themselves if instruction simply supplies additional positive evidence of a grammatical construction in question" (Cowan et al. 2003, p. 456); this is because Yip's study suggests that some of the errors made by the Chinese writers could be attributed not only to overgeneralizations but also to pattern similarities with English constructions. Additionally, the results of a number of studies which analyzed the writing of ESL learners of English indicate that overpassivization along with errors related to transitivity appears to be among the top features indicative of Chinese and Korean (e.g., of Sino-Tibetan L1 group) language background (Cowan et al., 2003; Ju, 2000; Yip, 1995; Zolb, 1989).

Studies which examine learner language data from speakers of Czech, or any Slavic language for that matter, are not nearly as commonly encountered in SLA research. This might be attributed to the obvious demographic factors; prior to the breakdown of the Soviet Union in the early 1990s, virtually no native speakers of Slavic languages resided in English-speaking countries—most of which were included in the capitalist block, comprising Great Britain, Australia, and the United States. However, studies which compared learner language from native speakers of Spanish and Japanese (Stauble, 1984; Schumann, 1984) were much common during the same time period. Since Spanish and Japanese L1 background groups were selected for those studies as representatives of two typologically different L1 groups, they are relevant to the discussion. For example, Stauble's (1984) study utilized a cross-linguistic design to compare six Spanish and six Japanese adult learners of English and their development of negation across proficiency levels. Additionally,

Schumann (1984) focused on describing characteristics of Spanish learners of English in the early stages of L2 acquisition.

Young's (1993) study which compared learner language data of native speakers of Chinese with the learner language data of native speakers of Czech and Slovak is an exception rather than a norm. This study explored the effect of functional constrains on variation in interlanguage morphology. The data was collected through interviews with both Czech and Chinese native speakers and analyzed through type/token ratio and VARBRUL multi-variable statistical procedures. The author concludes that functional constrains have little effect on variation in L1 morphology. However, it is important to mention that this study was based on the data only from twelve subjects (i.e., only six subjects per language group) and examined spoken and not written L2 data.

In addition, according to the results of Hinkel (2003), ESL writers tend show an overuse of *it*- and *there*-existential, vague nouns, public verbs, and tentative verbs, in comparison with the native-speaker writing. Only some of the tendencies suggested by the results of Hinkel (2003) study were confirmed in a more recent study, by Pendar and Chapelle (In press).

2.4 Summary

This chapter discussed the development and evolution of EA and CA as methodological approaches in applied linguistics for the past 50 years. It outlined the importance of corpus analysis and learner corpora and their use for computer-aided error analysis, and examined the implementation of these approaches in previous

studies. The discussion provided above led to the selection of grammatical features for further analysis. The next Chapter, *Methods and Materials*, will describe the materials used and methods and procedures implemented to carry out the data analysis.

METHODS AND MATERIALS

This chapter will outline the development of the materials and the methods used in the present study and describe the analysis of the data. The overview of the participant characteristics will be given and the development of the mark-up scheme, sentence-tagging process, and the error-annotation process will be presented. The procedures utilized to collect and interpret the data will be provided and discussed.

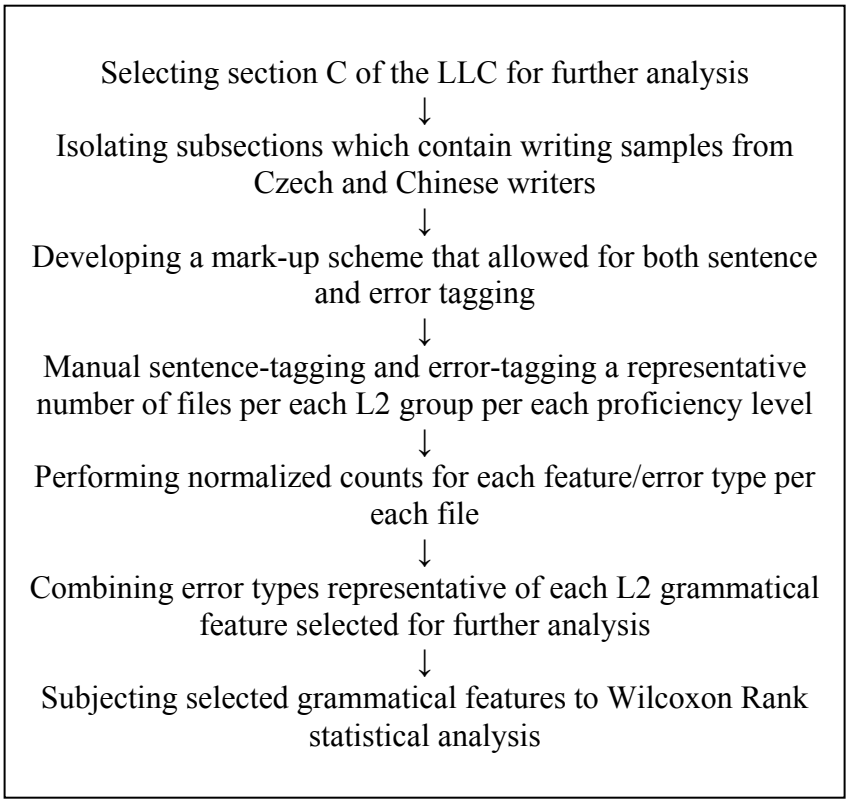
3.1 Procedure

The sequence of procedures implemented in the present study began with selecting LLC as a corpus from which the writing samples were drawn. From the array of currently available learner corpora[4] , LLC provided the most well-balanced and representative coverage of L2 learner writing in terms of L2 proficiency levels, L1 backgrounds, and the types of writing samples included (Pravec, 2002). As this research presumes a comparison of the L2 IL data drawn from two distinctly different linguistic groups, subsection C which includes the English writing samples from Chinese and Czech speakers was selected for further analysis. A mark-up scheme for sentence- and error-tagging the text files was developed and manually applied to 159 (71 Chinese and 88 Czech) files, including a representative number of files form each L1 group (Czech and Chinese) and each of the eight L2 proficiency levels. The counts of the occurrences of each mark-up category and domain per each file in the tagged subsection of the section C (totaling 36,237 words) were performed and

[4] The overview of currently available learner corpora is provided in Appendix A.

normalized for 100 words. Several grammatical features which could be extracted from the error-tagged data were selected for further statistical analysis using Wilcoxon Rank statistical test. Figure 3.1 shows the process flow chart. Each step of the procedure will be discussed in detail further in the chapter in section 3.3 .

Figure 3.1. Sequence of the procedures



3.2 Equipment and Materials

No printed materials were used in the study. The LLC corpus acquired by the English department of Iowa State University was used as the original data source. *CALLISTO* (2002) open source text annotation workbench was used to facilitate the

manual annotation process (http://callisto.mitre.org). *CALLISTO* (2002) is an annotation tool with a well-designed user interface, which was developed to support linguistic annotation of textual sources for any Unicode-supported language. *CALLISTO* (2002) allows for unique tag-set definitions and domain-dependent interfaces. Microsoft Excel was used to transfer, record, and perform data counts. R statistical software was used to perform non-parametric Wilcoxon tests. SAS statistical software was used to perform ANOVA on the data.

3.3 Data Collection

3.3.1 The Corpus

The investigated writing samples originated from the Longman Learners' Corpus (LLC) which consists of a collection of writing samples of learners of English as a second language. LLC contains approximately 8,000,000 words, which can be considered a relatively large collection of linguistic data. This word count also makes it a largest ESL learner corpus available. The writers who contributed to the LLC came from 70 different language backgrounds which include a total of 180 varieties from 16 source countries, and have been classified into eight language-proficiency levels. Appendix B provides a table with a detailed overview of LLC contributor characteristics.

Specifically, section C[5] of LLC was chosen for closer examination due to the fact that it includes two distinctly different linguistic groups, Chinese and Czech speakers, and a number of varieties of the above-mentioned languages (i.e., Chinese and its varieties, and Czech and its varieties). The abovementioned language groups were

[5] Entire section C of the *Longman Learners' Corpus* (LLC) contains approximately 2,600,000 words. The sections are organized in alphabetical order based on writers' L1.

selected due to the fact that these two language groups (Sino-Tibetan and Indo-European) are typologically very different from each other and, hence, are most appropriate for the research purpose. Czech and Chinese are distinct from one another linguistically in terms of the linguistic structure, and culturally in terms of their rhetorical traditions as well. Chinese is a highly isolating language, whereas Czech is a prime example of highly inflectional language; Czech is more appropriate for the given research purpose (i.e., to find error patterns in ESL writing with respect to writers' L1s through analyzing an ESL corpus of texts contributed by writers from two typologically different language groups) than other any other highly inflectional language (such as Russian or Turkish) because it has, among other features, a fully developed case system that includes seven cases in comparison to Russian, for example; Russian is also highly inflectional and is indeed typologically very close to Czech and Slovak, but its case system includes only six cases. Further, the Turkish case system includes just five cases.

3.3.2 The Learners

As mentioned above, the data under the investigation will be drawn only from the materials contained in section C of the LLC. An overview of the contributor characteristics provided is provided in Table 3.1 below.

Table 3.1. The characteristics of the learner data groups

Learners' L1	**Czech varieties (CZS,CZC) representing Indo-European language group** CZC - Czech CZS - Slovak	**Chinese varieties (CHC,CHK,CHS, CHT, CHI, CHX) representing Sino-Tibetan language group** CHC - China CHI - Indonesia CHK - Hong Kong CHS - Singapore CHX - Unspecified
Source country	Czech Republic, Slovakia	China, Singapore, Indonesia
Learner level pertaining to TL (same for both language groups**)**	1. Beginner – BE* 2. Elementary – EL 3. Pre-intermediate – PI 4. Intermediate – IN 5. Upper Intermediate – UI 6. Advanced – AD 7. Proficiency – PR 8. Academic Studies –AS*	
Environment of writing sample production (same for both language groups)	▪ standardized examinations ▪ authentic letters and documentation ▪ business communication documents ▪ internal examinations ▪ homework ▪ in-class assignments	
Task type (same for both language groups)	▪ set essay ▪ free essay ▪ project essay ▪ exercise ▪ letter ▪ advertisement ▪ report ▪ diary	
Target language (same for both language groups)	British English, Australian English or American English; however, the LLC goal is to focus on American English.	

NB:*Subsection C of the LLC did not contain any beginner (BE) files produced by Chinese L2 writers, and it contained only nine samples of Czech academic prose (AS) files. All the other groups were equally well represented (i.e., contained approximately ten files per each of the eight proficiency levels and per each L1 group).

Each file in the Longman Learner Corpus is coded by first language of student, language level, source country, environment, task type, and language variety (e.g., British English, American English). However, student L2 proficiency levels[6] and L1s were considered the most important categories by the LLC. Hence, these categories are likely to carry a higher degree of consistency, particularly considering the research purpose—

[6] Unfortunately, Longman has not provided a theoretical basis no does it provide any explanation for how exactly the writing samples were classified into specific proficiency levels.

which is to investigate the grammatical differences in the English IL between the two distinct L1 writer groups—this study will also award more weight to those writer characteristics. Of the 159 files analyzed, 71 were contributed by Chinese writers of English and 88 by Czech writers. The two L1 language groups also include the varieties of each native language (i.e., the Chinese group includes written samples produced by writers from China, Singapore, Indonesia, while the Czech group includes writing samples produced by writers from Czech Republic and Slovakia). Other additional information included certain demographic factors—such as gender and age—that was sometimes provided in the file coding but did not appear in all the files. The LLC was coded so that all the grammatical and spelling errors were keyed in exactly as they were written by the learners.

3.4 Development of the Mark-up Scheme

Granger (2003) provides a brief discussion of several types of well-known descriptive error taxonomies, starting with the two provided by Dulay, Burt, and Krashen (1982). The first type of taxonomy is based on linguistic categories—from general (such as grammar, lexis, morphology) to less general (prepositions, auxiliaries, etc.). The second type of taxonomy is based on the surface structure and its alterations by learners (omission, addition, miss-ordering, etc.). Granger argues that this dichotomous classification is inherently limited to the levels of analysis that can be produced with the application of these taxonomies. Her study and several other recent studies such James (1998) attempt to integrate the two above-mentioned error taxonomies with one or more other dimensions, arriving at multi-dimensional taxonomies which allow for much deeper levels of error analysis.

Although I realize the above-described advantages of potentially utilizing a multi-dimensional error-tagging system for certain type of research, the error-tagging scheme used in this study was developed to facilitate the analysis of L2 grammatical development; hence, it includes several levels of error categorization but follows one central dimension (grammatical) which is most appropriate considering the research purpose (Kosterina and Haji-Abdolhosseini, 2006).

In order to achieve maximum objectivity, several major guidelines were taken into account: (1) this mark-up scheme is *not* simply an error analysis scheme; this research also accounts for language use patterns, not just the errors. All the files were fully tagged for sentence type (whether *declarative, imperative, interrogative*, or *exclamative*) and for word order (whether *canonical, cleft, pseudo-cleft, reversed pseudo-cleft*, or *topicalized*) before error tagging was applied. (2) The error-annotation scheme was designed specifically to determine differences in the linguistic data from speakers of the Chinese varieties versus data from speakers of the Czech varieties. (3) Only an utterance that a highly proficient speaker of English would consider "wrong/unacceptable" was considered and counted as an error. (4) The annotation scheme is goal-oriented (i.e., in order to identify the type of an error, the utterance was contrasted with the correct version). (5) Selection of the correct version is based on *the principle of minimal edit* (i.e., the fewest editing steps that yield an acceptable utterance were implemented to correct any given error) (Kosterina and Haji-Abdolhosseini, 2006).

In addition to the guidelines described above, Granger's (2003) advice—based on the error analysis system introduced by Dagneaux, Denness, and Granger (1998) for

English—was followed. Granger (2003) states that for the annotation system to achieve the optimal level of effectiveness, it should include the following steps:

> The annotation system should be
>
> > 1. *informative* but *manageable*: it should be detailed enough to provide useful information on learner errors, but not so detailed that it becomes unmanageable for the annotator,
> >
> > 2. *reusable*: the categories should be general enough to be used for a variety of languages,
> >
> > 3. *flexible*: it should allow for addition or deletion of tags at the annotation stage and for quick and versatile retrieval at the post-annotation stage, and
> >
> > 4. *consistent*: to ensure maximum consistency between the annotators, detailed descriptions of the error categories and error-tagging principles should be included in an error-tagging manual

In devising an error-tagging system for this project, the above stated requirements were taken as main guidelines.

3.5 Structure of the Mark-up Scheme

The mark-up scheme was utilized to tag the sentences for the sentence types and then to fully error-tag the writing samples. The error-tagging scheme for the annotation of section C of LLC corpus, which was developed and applied to 159 files totaling 36,237 words from section C of the LLC corpus, is provided in Table 3.2 below. A few minor modifications were made to the annotation scheme at the early stages of the annotation process, such as an addition of the *morpho-syntactic* category (Kosterina and Pendar, 2007). The error-tagging system developed to annotate the LLC corpus comprises several levels of annotation, keeping the central focus on

grammar. The number of the categorization levels is determined by the type of error. For example, Table 3.2 and Figure 3.6 below illustrate the relationship between error domains and categories. If the definite article *the* was substituted with the indefinite article *a,* an error would be classified as a *substitution* under *type*, and as an *indefinite-for-definite* under *specifics*. On the other hand, if the definite article *the* was just omitted from a noun phrase, this error would be classified as an *omission* under *type*, and the *specifics* category would be left blank. Additionally, Appendix D provides the entire structure of the annotation scheme.

Table 3.2. Structure of the error annotation scheme adopted from Kosterina and Pendar, 2007*

TEXT/UTTERANCE *The lexical item or a sequence of lexical items marked for correction*	
CORRECT	*the correct form using the principle of minimal edit*
LOCUS *The shortest (intended) constituent (preferably indicative of error)*	`s, np, vp, pp, adjp, advp, n, v, aux, det, adj, adv, p, part, inter, conj, pro, rel, wh-np, wh-det, wh-adv`
NOTE *any notes that might be relevant to a given error*	Any specific notes such as *passivization attempt*

Table 3.2 (continued)

SPECIFICS *error specifics*	**(apply to substitution, omission, addition)** n, v, aux, adj, adv, p, part, inter, conj, pro, rel, wh-np, wh-det, wh-adv **(applies to addition)** dummy-subj: The weather, it is bad. **(apply to morpho-syntactic)** added-marker misplaced-marker missing-marker **(apply to agreement)** acc-for-nom other-for-3sg pl-for-sg nom-for-acc 3sg-for-other sg-for-pl **(apply to substitution)** base-for-gerund indef-for-def gerund-for-base def-for-indef gerund-for-noun none-for-indef noun-for-gerund none-for-def indef-for-none wrong-inflection def-for-none active-for-passive cardinal-for-ordinal passive-for-active ordinal-for-cardinal
SUBSTITUTE *Lexical item or a sequence with which the intended correct item/sequence was substituted*	n, v, aux, adj, adv, p, part, inter, conj, pro, rel, wh-np, wh-det, wh-adv
TYPE *the most general error domain; in certain cases expanded to* specifics *category*	**addition** **aspect**:wrong aspect marking **morpho-syntactic**: misused marker: possessive, plural, infinitive **omission** **order** **other** **overgeneralization**: overuse of a rule **parallelism** **fragment** **run-on** **repetition** **spelling** **substitution** **agreement**: subj-verb, det-n **collocation** preposition: misused preposition **tense**: wrong tense marking on the verb **voice** **transitivity**: using a transitive verb as intransitive, etc.

*The list of abbreviations used to describe a mark-up scheme is provided in Appendix C.

3.5.1 Sentence Annotation

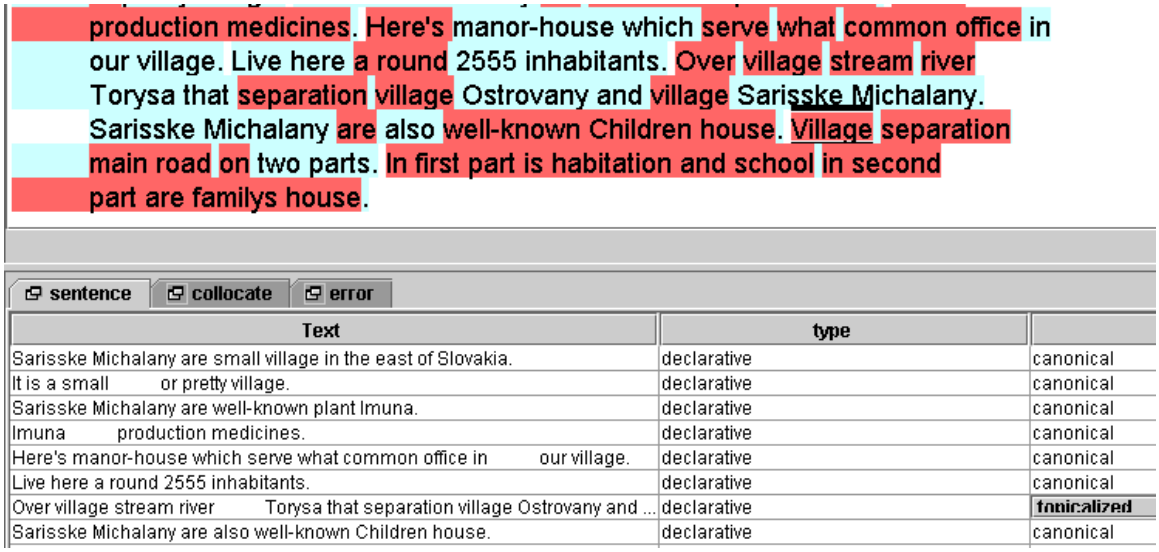
The sentences were tagged and classified by type (i.e., *declarative, imperative, interrogative*, and *exclamative*) and by word order (i.e., *canonical, cleft, pseudo-cleft, reversed pseudo-cleft*, and *topicalized*) as shown in the Table 3.3 below.

Table 3.3. Domains of the sentence annotation

SENTENCE	
TYPE	**WORD ORDER**
▪ **Declarative** John bought a car. ▪ **Imperative** Go, buy the car. ▪ **Interrogative** Have you bought the car? ▪ **Exclamative** What a nice car this is!	▪ **Canonical Word Order:** John bought a car. ▪ **Cleft:** It was a car that John bought. It was John who bought a car. ▪ **Pseudo-cleft:** What John bought was a car. Who bought a car was John. ▪ **Reversed pseudo-cleft:** A car is what John bought. John is the one who bought a car. ▪ **Topicalized:** John, he bought a car. A car, John bought.

Sentence fragments were error-tagged and marked as *fragments*. The intended fragments were left unmarked. The punctuation marks that clearly separated a clearly intended grammatical sentence into fragmented parts were ignored, and the entire sentence was sentence tagged. As shown in Figure 3.2 below, the second sentence from the top was classified as *canonical* under sentence type and *topicalized under* word order.

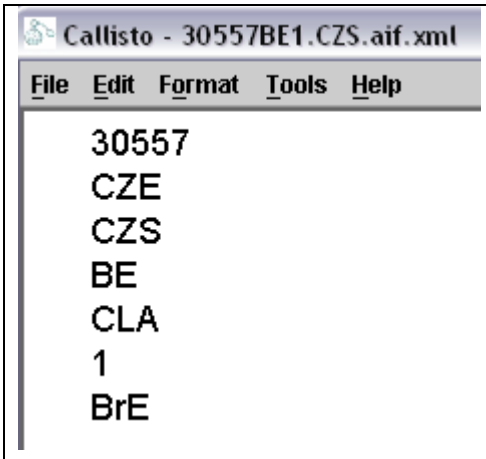
Figure 3.2. Sample of sentence-tagged text



The top left corner of the screen shot in Figure 3.4 is explained in more detail in Figure 3.3 below. Figure 3.3 illustrates the header which is included in each text file in LLC[7]. Line 1 (30557) refers to the file number in subsection C of the LLC. Line two refers to the L1 language group of the file origin (e.g., Czech [CZE]). Line 3 refers to the specific variety within the Czech 1 group (e.g., Slovak [CZS]). Line 4 refers to the proficiency level of the learner (e.g., beginner). Line 5 refers to the environment code (e.g., class work [CLA]). Line 6 refers to the task type (e.g., set essay [1]), and line 7 refers to the target language variety (e.g., British English [BrE]). Below is an example of the header 30557.BE1.CZS. Appendix E provides all the header tags encountered in the analyzed data.

[7] Longman provides a detailed file description and the description of all the coding applied to all the learner text files in the *Longman Learner Corpus (LLC)*http://www.longman-elt.com

Figure 3.3. Example and description of the header



← File name		
← Tool bar		
← Line 1	File reference number	30557
← Line 2	The source country code	Czech (CZE)
← Line 3	Source language	Slovak (CZS)
← Line 4	Proficiency level	Beginner (BE)
← Line 5	Environment	Class work(CLA)
← Line 6	Text type	Set essay (1)
← Line 7	Target language variety	British English (BrE)

3.5.2 Error Annotation

The annotation process began in August 2006 and was completed in March 2007. The selected subsection of section C of the LLC was fully tagged, first for the sentence types and then for all error types. *CALLISTO* (2002) open source text annotation workbench was used to facilitate the manual annotation process (http://callisto.mitre.org). The *CALLISTO* (2002) is an annotation tool with a well-designed user interface, which was developed to support linguistic annotation of textual sources for any Unicode-supported language. *CALLISTO* (2002) allows for unique tag-set definitions and domain-dependent interfaces. Figure 3.3 illustrates an example marked-up text as it appears in *CALLISTO* (2002). *CALLISTO* (2002) allows for extension with user interface components specific to a domain. Tag editing capabilities are shown in a highlighted text display and tag attribute tables. As domain-specific extension components were developed, they were integrated into the core of *CALLISTO* (2002) to provide a customized interface of necessary annotation components. The error-tagging appears in red, and the sentence tagging

appears in blue. The Available Actions tab was customized so that it would allow for sentence tagging (blue tab), error tagging (red tab), and tagging of collocation errors (green tab) which is a subtype of error tagging (Kosterina and Haji-Abdolhosseini, 2006). The bottom part of Figure 3.4 shows specific errors which have been categorized[8] .

The mark-up scheme is based on XML, a standard data transfer format which is both readable and easily processed by computers. Figure 3.5 provides an example of a fully tagged sentence as it appears in XML format.

```
1          <sentence type="declarative" word-order="canonical">
2          He
3          <error specifics="AUX"
4          type="addition"
5          correct="likes"
6          substitute=""
7          note=""
8          id=""
9          locus="VP">
10                 <error specifics=""
11                 type="tense"
12                 correct="likes"
13                 substitute=""
14                 note=""
15                 id=""
16                 locus="VP">
17                        was like
18                    </error>
19         </error>
20         <error specifics="indef-for-none"
21          type="substitution"
22          correct="cheese"
23          substitute=""
24          note=""
25          id=""
26          locus="NP">
27                    a cheese
28    </error>.
29   </sentence>
```

NB: He likes cheese = correct

[8] For a detailed explanation of error tagging and classification, refer to the *Correction* section of this chapter.

Figure 3.4 Example of marked-up text in *CALLISTO*

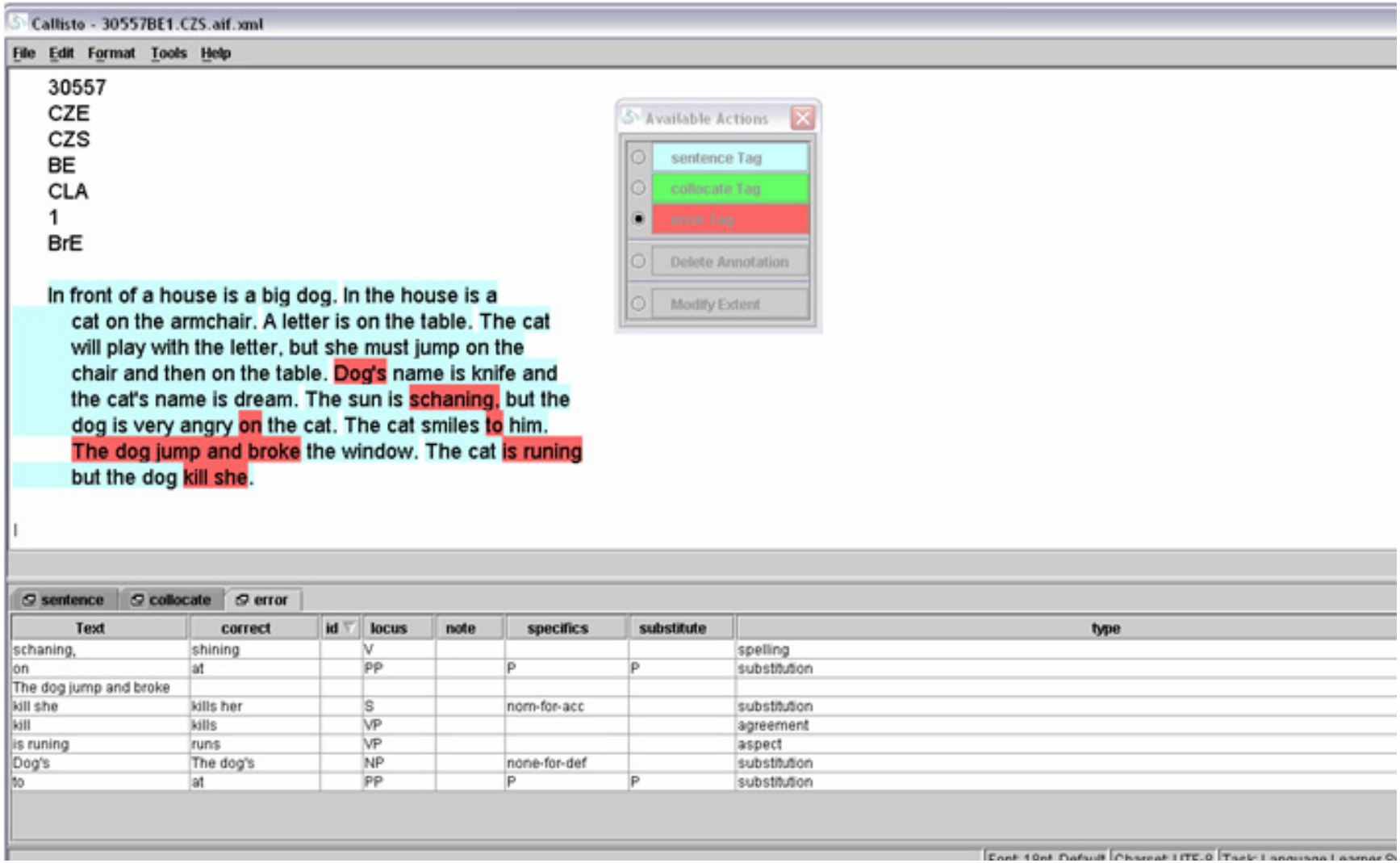


Figure 3.5 shows an example of a sentence which was tagged and classified by type as *declarative*, and by sentence order as *canonical*. It was inferred from the context[9] that the intended sentence was "He likes cheese," whereas the writer produced "He was like a cheese." This sentence contains two errors. The first one is tagged in lines 3 through 9 and is classified as follows. The locus (i.e., where the error occurred) is a verb phrase (VP); "was like" appears in the place of the correct form "likes." This error was corrected in two steps. It is a *tense* error and an *addition of an auxiliary* [was] error. The second error in the sentence in Figure 3.4 is coded between lines 11 and 17. The utterance, "a cheese," was used instead of the correct form "cheese," and is a *substitution* type error where no article should have been used and an indefinite article was used instead (i.e., categorized as an *indefinite-for-none* under the error specifics.

After corrections were inserted for each erroneous utterance, each correction corresponding to each erroneous utterance was saved[10]. In order to facilitate the tagging process, *CALLISTO* (2002), which was described in more detail at the beginning of the section 3.5.2 of this chapter, was used. It allows the editor to customize error tags, and insert the error tag by clicking on the appropriate tag from the error-tag menu. Figure 3.6 below shows a sample of fully-tagged text where the errors have been annotated.

The annotation of the utterance "not like", as seen from Figure 3.6 above, is explained in detail below. Additionally, Appendix C gives the list of the abbreviations used in the annotation scheme

[9] It is important to mention that although sometimes correction is a straightforward choice of a correct utterance, other times it is simply one of several potentially correct choices—especially when the researcher is forced to draw on the context of the entire text to infer the intended meaning of a given utterance.

[10] The annotated files were saved in XML format and later exported to SGML format for further quantitative analysis.

Figure 3.6. Sample of error-tagged text

15227AD.CZS

The rain is wet often in spring and autumm. In southern countries, it rain's in winter. I not like the rain in autumm. It's cold and we must carry umbrella but the rain in summer is warm and it's not cold.

<sentence "The rain is wet often in spring and autumm." type="declarative" word-order="canonical">

sentence | collocate | error

Text	correct	id	locus	note	specifics	substitute	type
not like	don't like		VP		AUX		omission
umbrella	umbrellas		N		sg-for-pl		substitution
wind	windy		ADJP		ADJ	N	substitution

.

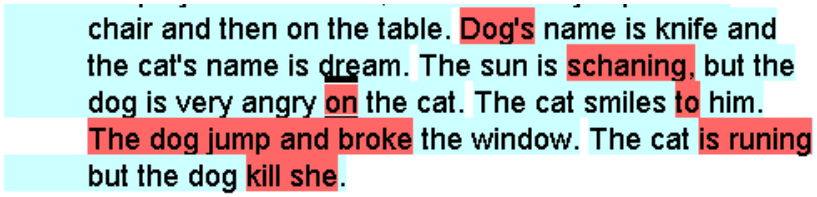
In this case, the utterance interpreted as erroneous, "not like", was error tagged and automatically pasted in the annotation table, which usually appears at the bottom of the screen in *CALLISTO* interface. Then the correct form was manually entered into the *correct* column. *Locus* represents the grammatical sequence where the error occurs, which is a verb phrase (VP) in this case. Since this is a considerably clear case, no *notes* were recorded. *Specifics* represents the lexical item or grammatical category which should have been present in place of error if the utterance were correct; in this case, if the auxiliary (AUX) "do" was present, the utterance would have been correct, granted that the negation marker may have been attached to that auxiliary. The auxiliary was not substituted with any other lexical item, which explains why the *substitute* category is empty; the *substitute* category is allotted for the items that were inserted in place of a correct item of utterance. This error was categorized as an omission *type* of error, since the needed auxiliary is missing. This item in essence represents a negation problem, commonly seen in the intermediate-level L2 writing of both L2 groups investigated in the study.

Another example of an error annotation can be also seen in Figure 3.6 above. In this case, the utterance interpreted as erroneous is "umbrella." Then the correct form was

manually entered into the *correct* column. *Locus* represents the grammatical sequence where the error occurs, noun (N) in this case, since the noun "umbrella" is missing a necessary final *s* plural marker. Since this is also a considerably clear case, no *notes* were recorded. *Specifics*, in this case, refers to the specific error category (i.e., a singular noun was used where a plural noun was needed (*sg-for-pl*). The noun was not substituted with any other lexical item, which explains why the *substitute* category is also empty; this category is assigned for the items that were inserted in place of a correct item in the utterance. This error was categorized as a substitution *type* error.

Additionally, Figure 3.7 below contains two prepositional errors. In line 3 at the bottom of the figure, the preposition "at*"* was substituted with the preposition "on*"*, and in line 9, the preposition "at*"* was substituted with the preposition "to*"*.

Figure 3.7 Sample of preposition error tagging



sentence | collocate | error

Text	correct	id	locus	note	specifics	substitute	
schaning,	shining		V				spelling
on	at		PP		P	P	substitution
The dog jump and broke							
kill she	kills her		S		nom-for-acc		substitution
kill	kills		VP				agreement
is runing	runs		VP				aspect
Dog's	The dog's		NP		none-for-def		substitution
to	at		PP		P	P	substitution

The third line from the top of the annotation table, "the dog was angry *at* the cat" is a substitution-type error in which the preposition "at*"* was substituted with the preposition "on*"*.

3.6 Analysis

The data collected for analysis includes 159 fully-tagged files written by speakers of Czech and Chinese and their varieties, totaling 36,237 words. Table 3.4 provides the number of written files analyzed, the average number of words per essay, and the total word counts per file per L1 group. Despite some variability in numbers of essays per L1 group (i.e., 71 Chinese and 88 Czech), and per each of the eight proficiency levels, a total of approximately 18,000 words per each of the two L1 groups was fully error-tagged and analyzed, and the final counts were normalized. The total number of words in the sample (18,263 for Chinese and 17,974 for Czech) only differ by 289 words between the two L1 groups and, hence, writing samples from both groups can be considered comparable. Features were tagged by hand over a period of seven months by the author, who consulted the principal investigator in the instances where the annotator experienced any uncertainty in classifying any particular error (Kosterina and Pendar, 2007). The principle investigator finalized all the of the tagging and error classification decisions. For this reason, it is believed that the tagging is reasonably reliable, though with only one annotator, no reliability estimates could be obtained.

Table 3.4. Writing samples and words in the sample by L1 groups of writers

L1 group	Number of writing samples analyzed	Average length (Mean number of words per file)	Total number of words per L1 group
Chinese	**71**	**259.34**	**18,263**
Czech	**88**	**204.25**	**17,974**
Total	**159**	**228.66**	**36,237**

NB: file=writing sample

As mentioned above, a total of 159 files total were manually annotated. The original goal of the annotation was to acquire ten files per each of the two L1 groups (i.e., Chinese or Czech) and per each of the eight L2 proficiency levels. However, after closer examination, it became clear that LLC does not contain any files in the Chinese beginner (BE) level, and contains only nine files in the Czech academic prose (AS) category. It also contains only five files in the Czech PR level. Hence, only the available files in these categories were annotated. Nevertheless, this did not undermine the methodology of the present study since it is concerned with the analysis of the two language groups through comparison between their L1s and not their L2 proficiency levels. Overall, the analyzed files were selected to give a representative picture of learner writing both across the language groups and proficiency levels, given the data available in the LLC corpus. Table 3.5 below provides information about the number of files, average length of the files per each proficiency level and L1 group, and the total word counts per each proficiency level and L1 group for all the analyzed files.

The frequency of occurrence of each error type and error specific were counted per each file and normalized for 100 words. The selected error types were converted into variables. Specifically, for the Research Question 1, the frequency of occurrence of all *word-order errors* were counted and normalized for 100 words. For the Research Question 2 all the instances annotated as *topicalization attempts* were counted and normalized for 100 words. For the Research Question 3 all the instances of error-annotations, classified as *expletive subject errors* were counted and normalized for 100 words. Research Question 4 investigated the occurrences of *article errors*. Hence, the article error variable was based on the sum of the normalized counts of all six article

[Table 3.5. Writing samples and words in the samples, categorized by L1 groups of writers and per proficiency levels

	CHINESE			CZECH			TOTALS per L1 group per proficiency level		
Proficiency level	**Number of files**	**Average length**	**Total number of words**	**Number of files**	**Average length**	**Total number of words**	**Number of files per level**	**Average length per level**	**Total number of words per level**
ZZ*	1	109.0	109	0	0.000	0	1	109.0	109
BE	0	0.0	0	14	132.143	1850	14	132.143	1850
EL	10	125.8	1258	15	127.13	1907	25	126.6	3165
PI	10	140.8	1408	14	163.64	2291	24	154.125	3699
IN	10	184.6	1846	12	250.41	3005	22	220.5	4851
UI	10	255.8	2558	9	287.2	2585	19	270.68	5143
AD	10	273.8	2738	10	256.4	2564	20	265.1	5302
PR	10	219.6	2196	5	312.6	1563	15	250.6	3759
AS	10	615.0	6150	9	245.4	2209	19	439.95	8359
Total	**71**		**18263**	**88**		**17974**	**159**		**36237**

NB: Files that were marked ZZ in the corpus are missing the proficiency level categorization.

error subtypes, comprising *indefinite-for-definite, definite-for-indefinite, none-for-indefinite, none-for-definite, indefinite-for-none, and definite-for-none* error subtypes. For the Research Question 5, the variable extracted for the analysis consisted of all the annotations of *errors with plural count nouns marked with s* (i.e., *substitution* type errors classified as *sng-for-pl* type). For the Research Question 6 which investigated the *passivization attempts*, the variable extracted for the analysis consisted of the sum of all the subtypes of the error types which were considered most indicative of passivization attempts drawn from the annotated files comprising *passive-for-active* and *auxiliary addition* error types. Research Question 7 investigated *preposition errors*, and therefore the variable extracted for the analysis consisted of the sum of all the subtypes of preposition errors drawn from the annotated files[11] .

3.7 Statistical Analysis

The data was analyzed through descriptive and inferential statistics, and Wilcoxon Rank test was selected as a consecutive measure of the differences between the L1 groups (i.e., Czech and Chinese). In order to establish a clear picture of how often each feature appeared in each annotated text, the computerized count of the number of words in each of the 159 files was obtained and then the number of occurrences of each tagged feature, including the features selected for further statistical analysis was calculated. For example, file 22315EL1.CZS is a writing sample produces by a Slavic (CZS) learner of an elementary (EL) L2 proficiency level, comprising 83 words. The frequency of occurrences of each feature selected for the statistical analysis were counted per each

[11] For the specific examples of annotation types extracted for the data analysis for each research question, refer to Chapter 3, the section 3.5.2, *Error Annotation,* of this document (i.e., pp. 58-65).

writing sample, and the counts were normalized for 100 words to account for the variability in the lengths of writing samples. For example, the file number 22315EL1.CZS contains approximately 1.37 topicalization attempts for 100 words.

As stated in Chapter 1, the present study sets out to investigate whether grammatical structure of learners' L1 plays a role in learners' English L2 development and can result in transfer into L2 writing. The study also aims to find patterns of language use and error in learners of English as a second language with respect to their L1s. Five out of the seven research questions posed in this study hypothesized that each error type selected for the investigation would appear at a significantly higher rate in the writing of one L1 group than in the other. The exceptions were Research Questions 2 and 7, which hypothesized no significant difference between the rates of occurrences of the features investigated by these two research questions between the two L1 groups. Specifically, Research Question 2 was formulated to investigate the existence or absence of a difference between the numbers of topicalization attempts made by each L1 group. Similarly, Research Question 7 also sets out to investigate the existence or absence of a difference between the numbers of preposition errors made by each L1 group. Consequently, to an understanding of the existence or absence of error types between the two L1 groups, ANOVA tests were initially chosen for the statistical analysis. Additionally, performing ANOVA tests has the advantage of eliminating the possibility of an I-type error, which is a common threat when performing multiple t-tests.

However, after obtaining graphic representations and performing further examination of the data, it became evident that two model assumptions which must be met for the parametric tests did not hold. First, the data has to be normally distributed.

The complete graphic representations for each data set before and after the data transformation are provided in Appendix F. It was evident form the histograms that the data are highly right skewed for each of the seven sub-data sets corresponding to each of the seven research questions. Furthermore, the means appeared to be higher than the standard deviations in each case, which is also indicative of high dispersion in the data. Second, for the appropriate administration of parametric tests it is assumed that the standard deviations between the groups tested against each other are similar, which is not consistent with the results obtained. This means that the constant variants assumption was also violated. A test of equal variances, Levene's test, was performed and the results confirmed that the variances are significantly different. For instance, in the case of the article errors, the *p*-value obtained from Levene's test equaled 0.0113. The log transformations did not resolve the distribution issues due to a high number of 0 (zero) observations.

Hence, due to the fact that the two modal assumptions required for the appropriate administration of parametric test were violated, the decision was made to administer non-parametric Wilcoxon tests on all of the data sets for each research question. In addition, the ANOVA tests were still performed to reaffirm the findings obtained from the Wilcoxon Rank tests. The decision was made to exclude comparisons of averages in order to avoid obscuring the distribution of frequencies in the sample. The alpha for achieving statistical significance was set at 0.05.

3.8 Summary

This chapter outlined the development of the materials and the methods used in the present study. The description and the analysis of the data were given. The procedures employed to collect and analyze the data were also presented. In the following chapter, *Results and Discussion*, the results obtained from the study employing ANOVA and non-parametric Wilcoxon tests will be presented and discussed.

RESULTS AND DISCUSSION

This study set out to investigate whether grammatical structure of learners' L1 plays a role in learners' English L2 development and can result in transfer into L2 writing. The study also aimed to find patterns of language use and error in learners of English as a second language with respect to their native language by studying a corpus of writings produced by such learners. The study examines the manually annotated part of section C of the LLC corpus which includes ESL writing data from the native speakers of Czech and Chinese varieties for the evidence of transfer pertaining to several selected grammatical features. In the previous chapter, I presented the research materials needed, the methods involved, and the procedure to analyze the data. Now, I present and discuss the results obtained in the study.

The study is guided by a central research question: *Does grammatical structure of learners' L1 play a role in learners' English L2 development, and can it result in transfer into L2 writing?* Seven more specific research questions, stated below, explore each grammatical feature selected for the analysis. **The overall results of the study show that learners' L1 plays a role in learners' English L2 development, and do result in transfer into L2 writing.** Specifically, the results of non-parametric Wilcoxon tests, as well as the results of ANOVA tests, provided empirical support for five out of seven original hypotheses.

The grammatical features selected for the final analysis were the sentence order errors, expletive subject errors, topicalization attempts, article errors, errors with plural *s* marking in count nouns, passivization errors, and prepositional errors. It was hypothesized that the Czech writers would tend to make significantly more sentence

order errors, expletive subject errors, article errors and topicalization attempts. It was also hypothesized that the Chinese writers would make significantly more passivization errors and errors with plural *s* marking in count nouns. Furthermore, it was predicted that there would be no significant difference in the number of preposition errors between the two L1 groups. The formulation of the research questions was grounded in the typological differences between the Czech and Chinese languages as well as previous research which were discussed in detail in the Chapter 1, section 1.3, *Research Questions*, and in Chapter 2, *Literature Review*. The remainder of the chapter will discuss the results of each research question in detail.

4.1 Research Question 1

The Research Question 1 investigated if the overall number of sentence word-order errors is higher in the English IL of Czech and Slovak learners than in the English IL of Chinese learners for learners across levels of proficiency in English. It was hypothesized that the overall number of sentence word-order errors would be higher in the English IL of Czech and Slovak learners than in the English IL of Chinese learners for learners across levels of proficiency in English. The results of the statistical tests are summarized in Table 4.1 below.

Table 4.1: Summary of the statistical results of Research Question 1

L1 Group	Number of files analyzed	Mean number of word order errors	Std Dev
Chinese	71	0.143	0.329
Czech	88	0.481	0.844

NB: Std Dev=Standard Deviation; p = 0.0050

The results of the Wilcoxon Rank test presented in Table 4.1 show that, overall, the Czech writers made significantly more sentence word order errors than did the Chinese writers. As evident from Table 4.1, the results of Research Question 1 are consistent with the original hypothesis that the overall number of sentence word order errors be higher in the English IL of Czech and Slovak learners than in the English IL of Chinese learners for learners across levels of proficiency in English.

4.2 Research Question 2

Based on the assumption that both L1 groups might be prone to some topicalization attempts in their L2 writing, Research Question 2 was introduced in an attempt to examine topicalization in both of the L1 groups and investigate whether there is a difference in the number of topicalization attempts made by Chinese and Czech writers across levels of proficiency in English. It was hypothesized that there should not be a significant difference in the number of topicalization attempts between the English IL of Czech and Slovak learners compared with the English IL of Chinese learners across levels of proficiency in English. The results of the statistical tests are summarized in Table 4.2 below.

Table 4.2: Summary of the statistical results of the Research Question 2

L1 Group	Number of files analyzed	Mean number of topicalization errors	Std Dev
Chinese	71	0.016	0.114
Czech	88	0.068	0.261

NB: Std Dev=Standard Deviation; p = 0.1030

Table 4.2 provides a summary of the descriptive statistics and the Wilcoxon Rank test results for the topicalization attempts made by both L1 groups. The results of the

Wilcoxon Rank test show that, although the Czech group made a larger number of errors than did the Chinese group, the difference was not statistically significant. There results are consistent with the original hypothesis, which was drawing support from the argument that both Czech and Chinese are topic-prominent languages; however, the ways in which they interpret topicalization in terms of their L1 language typology might differ. Hence, the results suggest that topicalization by itself might not be the ideal variable for eliciting L1 transfer in writing which is associated with topic-prominence of the learners' L1. Further examination of topicalization attempts in L2 writing made by learners of both L2 groups is called for.

4.3 Research Question 3

Research Question 3 investigated whether the overall number of expletive subject errors was higher in the English IL of Czech and Slovak learners than in the English IL of Chinese learners for learners across levels of proficiency in English. It was hypothesized that the overall number of expletive subject errors would be higher in the English IL of Czech and Slovak learners than in the English IL of Chinese learners for learners across levels of proficiency in English. The results of the statistical tests are summarized in Table 4.3 below.

Table 4.3: Summary of the statistical results of Research Question 3

L1 Group	Number of files analyzed	Mean number of expletive subject errors	Std Dev
Chinese	71	0.000	0.000
Czech	88	0.448	1.986

NB: Std Dev=Standard Deviation; p = 0.0094

The results of the Wilcoxon Rank test presented in Table 4.3 show that overall the Czech writers made significantly more expletive subject errors than did the Chinese writers, which was also hypothesized based on the fact that Czech does not employ expletive subjects in its sentence structure. Under closer examination it can be seen that, whereas Czech writers made some expletive subject errors, there were very few of those; it does not appear that the Chinese writers made any expletive subject errors altogether.

4.4 Research Question 4

Research Question 4 investigated whether the overall number of article errors was higher in the English IL of Czech and Slovak learners than in the English IL of Chinese learners for learners across levels of proficiency in English. It was hypothesized that the overall number of article errors would be higher in the English IL of Czech and Slovak learners than in the English IL of Chinese learners for learners across levels of proficiency in English. The results of the statistical tests are summarized in Table 4.4 below. [12]

Table 4.4: Summary of the statistical results of Research Question 4

L1 Group	Number of files analyzed	Mean number of article errors	SD
Chinese	71	0.614	0.885
Czech	88	2.303	2.572

NB: SD=Standard Deviation; $p < 0.0001$

[12] The variable extracted for the analysis consisted of the sum of all the subtypes of article errors drawn from the annotated files (i.e., the sum of the following: *indefinite-for-definite, indefinite-for-none, definite-for-indefinite, def-for-none, none-for-indefinite, none-for-definite*).

The results of the Wilcoxon Rank test presented in Table 4.4 show that, overall, the Czech writers made significantly more article errors than did the Chinese writers. These results are also consistent with what was originally hypothesized.

4.5 Research Question 5

Research Question 5 investigated whether the overall number of errors with plural count nouns marked with *s* was higher in the English IL of Chinese learners than in the English IL of Czech and Slovak learners for learners across levels of proficiency in English. It was hypothesized that the overall number of errors with plural count nouns marked with *s* would be higher in the English IL of Chinese learners than in the English IL of Czech and Slovak learners for learners across levels of proficiency in English. The results of the statistical tests are summarized in Table 4.5 below.

Table 4.5: Summary of the statistical results of Research Question 5

L1 Group	Number of files analyzed	Mean number of plural count noun errors	Std Dev
Chinese	71	0.674	0.889
Czech	88	0.653	0.923

NB: Std Dev=Standard Deviation; p = 0.3243

The results of the Wilcoxon Rank test presented in Table 4.5 above showed no significant difference between overall number of plural count noun errors made by the Czech writers compared to the number of the same type of errors made by Chinese writers. Contrary to the original hypothesis, the results of the Wilcoxon Rank test did not show a significant difference between the overall number of errors with plural count

nouns marked with *s* between the writing of Chinese and Czech writers; this finding was not expected, since Chinese—a highly isolating language—almost never uses final noun morphology to convey meaning. Perhaps, the results can be partially attributed to the possibility that in English pluralization is employed differently on some level from pluralization in both Czech and Chinese. However, at the present time, it is not clear which factors can be attributed to the causation of these results.

4.6 Research Question 6

Research Question 6 investigated whether the overall number of passivization attempts was higher in the English IL of Chinese learners than in the English IL of Czech and Slovak learners for learners across levels of proficiency in English. It was predicted, based on an array of previous studies, that the overall number of passivization attempts would be higher in the English IL of Chinese learners than in the English IL of Czech and Slovak learners for learners across levels of proficiency in English. The results of the statistical tests are summarized in Table 4.6 below. [13]

Table 4.6: Summary of the statistical results of the Research Question 6

L1 Group	Number of files analyzed	Mean number of passivization errors	Std Dev
Chinese	71	0.205	1.165
Czech	88	0.144	0.942

NB: Std Dev=Standard Deviation; p = 0.4771

[13] The variable extracted for the analysis consisted of the sum of all the subtypes of the error types which were considered most indicative of passivization attempts drawn from the annotated files (i.e., the sum of *passive-for-active* and *auxiliary addition* error types).

Contrary to the original expectation, the results of the Wilcoxon Rank test, presented in the Table 4.6 above, showed no significant difference between overall number of passivization attempts made by the Czech writers and by Chinese writers. This, similar to the results of the Research Question 5, means that the results did not show that Chinese ESL writers tend to passivize much more frequently than Czech writers; this finding additionally contradicts the results of a number of previous studies (Cowan et al., 2003; Ju, 2000; Yip, 1995; Zolb, 1989). This might be attributed to the fact that the annotation scheme itself did not separate passivization into a separate category. Hence, the features extracted from the annotated data, which were selected as better predictors of passivization attempts such as additions of auxiliaries, perhaps did not fully account for all the passivization attempts encountered in the annotated data.

4.7 Research Question 7

Research Question 7 investigated whether, in fact, there was any significant difference in the number of preposition errors in the English IL of Czech and Slovak learners compared with the English IL of Chinese learners across levels of proficiency in English. It was hypothesized that there should not be a significant difference in the number of preposition errors in the English IL of Czech and Slovak learners compared with the English IL of Chinese learners across levels of proficiency in English; this hypothesis was based on the fact that ESL writers frequently make prepositional errors, regardless of their L1 background, simply due to the absence of a clear-cut grammatical

rule in English which would regulate the use of one preposition over another. The results of the statistical tests are summarized in Table 4.7 below.[14]

Table 4.7: Summary of the statistical results of the Research Question 7

L1 Group	Number of files analyzed	Mean of preposition substitution errors	Std Dev
Chinese	71	3.352	4.263
Czech	88	5.997	7.844

NB: Std Dev=Standard Deviation; p =0.1276

The results of the Wilcoxon Rank test presented in Table 4.7 above showed that, overall, no significant difference between overall number of preposition errors between the two L1 groups; the Czech writers made slightly more preposition errors than did the Chinese writers. As evident from Table 4.7, the results of Research Question 7 are consistent with the original prediction that there should not be a significant difference in the number of preposition errors in the English IL of Czech and Slovak learners compared with the English IL of Chinese learners across levels of proficiency in English. Cowan et al. (2003) pose an argument that certain preposition errors can be indicative of Chinese/Korean L1 background. Overall, the results of Research Question 7 provide substantial empirical evidence in support of the position that even when certain types of preposition errors might be peculiar to a specific L1 group (Cowan et al., 2003), the overall preposition errors are not specific to an L1 group.

[14] The variable extracted for the analysis consisted of the sum of all the subtypes of preposition errors drawn from the annotated files.

4.8 Summary

In this chapter, the results of the study obtained from employing non-parametric Wilcoxon statistical tests were presented and discussed. The study investigates whether grammatical structure of learners' L1 plays a role in learners' English L2 development and can result in transfer into L2 writing, and aims to find patterns of language use and error in learners of English as a second language with respect to their native language and their levels of English proficiency by studying a corpus of writings produced by such learners.

The results of the study show that learners' L1 plays a role in learners' English L2 development and does result in transfer into L2 writing. Specifically, the results of the non-parametric Wilcoxon tests provided empirical support for five out of seven original hypotheses. Additionally, all the results obtained from the Wilcoxon Rank tests were reaffirmed by administering additional ANOVA tests, which resulted in very similar *p* values to the ones obtained from the Wilcoxon Rank tests. In the following chapter, *Conclusion*, I summarize the findings, present the limitations of the study, discuss the implications of the study, and give suggestions for further research.

CONCLUSION

In the previous chapter, *Results and Discussion*, I presented and discussed the findings of the study. The primary research question of this study addressed the issue whether grammatical structure of learners' L1 plays a role in learners' English L2 development and can result in transfer into L2 writing, and aimed to find patterns of language use and error in learners of English as a second language with respect to their native language by analyzing a corpus of writings produced by such learners. In this chapter, I summarize the results, present the limitations of the study, discuss the implications of the study, and give suggestions for further research.

5.1 Summary of Results

The study investigated whether grammatical structure of learners' L1 plays a role in learners' English L2 development and can result in transfer into L2 writing, and set out to find patterns of language use and error in learners of English as a second language with respect to their native language by studying a corpus of writings produced by such learners. The study examined the manually annotated part of section C of the LLC corpus, which includes ESL writing data from the native speakers of Czech and Chinese varieties for the evidence of transfer pertaining to several selected grammatical features. In the previous chapters, I presented the research materials needed, the methods involved, and the procedure to analyze the data. Now, I present and discuss the results obtained in the study.

The results of non-parametric Wilcoxon tests show that, overall, Czech writers made significantly more sentence word-order errors, expletive subject errors, and article errors, as was hypothesized.

No significant difference between the number of preposition errors between the two L1 groups was found, which is also consistent with the original hypothesis; to be exact, Czech writers made slightly more preposition errors overall. This difference could be attributed to the fact that the original data, which was used as a basis for the statistical analysis, was missing Chinese beginner (BE) proficiency level files altogether, which could possibly explain the apparent difference between the number of preposition mistakes between the two L1 groups. Still, the results provide solid empirical support in favor of the argument which posits that overall preposition errors are not particular to writers of either L1 backgrounds.

Furthermore, the results of the Wilcoxon Rank test did not show a significant difference between the number of topicalization attempts between the writing samples of the two L1 groups, which is also consistent with the original hypothesis that was drawing support from the argument that both Czech and Chinese are topic-prominent languages; however, the ways in which they implement topicalization in terms of their L1 language typology might differ. Hence, the results suggest that topicalization by itself might not be the ideal variable for eliciting L1 transfer in writing which is associated with topic-prominence of the learners L1. Further examination of topicalization attempts in L2 writing made by learners of both L2 groups is called for.

On the other hand, contrary to the hypothesis, the Wilcoxon Rank test results did not show a significant difference between the number of passivization attempts between

the writing samples of the two L1 groups. This means that the results did not show that Chinese ESL writers tend to passivize much more regularly than Czech writers, which additionally contradicts the results of a number of previous studies (Cowan et al., 2003; Ju, 2000; Yip, 1995; Zolb, 1989). This might be attributed to the fact that the annotation scheme itself did not classify passivization into a separate category. Hence the features extracted from the annotated data, which were selected as better predictors of passivization attempts such as additions of auxiliaries, perhaps did not fully account for all the passivization attempts encountered in the annotated data.

Additionally, contrary to the original hypothesis, the Wilcoxon Rank test results did not show a significant difference between the overall number of errors with plural count nouns marked with *s* between the writing of Chinese and Czech writers. It was hypothesized that Chinese writers would have significantly more errors with plural count nouns marked with *s*, since Chinese—being a highly isolating language—almost never uses final noun morphology to convey meaning. As mentioned in section 4.5 of Chapter 4, it is possible that English on some level employs pluralization differently from pluralization in both Czech and Chinese. Still, at the present time it is not clear which factors can be attributed to the causation of these results.

The results of the study show that some aspects of learners' L1 play a role in learners' English L2 development and do result in transfer into L2 writing. Specifically, the results of non-parametric Wilcoxon tests provided empirical support for five out of seven original hypotheses. Furthermore, it is important to mention that all the results obtained from Wilcoxon Rank tests were reaffirmed by administering additional ANOVA

tests, which resulted in very similar *p* values to the ones obtained from the Wilcoxon Rank tests.

5.2 Limitations of the Study

The present study is far from ideal and there are many factors contributing to it, such as the structure of the annotation scheme itself, the inter-tagger reliability, and the lack of a sound theoretical basis for why a given file was classified as one specific proficiency level and not another. Out of the six main limitations of the study discussed in detail below, two are related to the characteristics of the LLC learner corpus, two are related to the error-annotation scheme developed for this research, and the last two can be classified as procedural and methodological limitations. Table 5.1 below presents suggestions that can be used to overcome the limitations of the present study.

5.2.1 Corpus-related Limitations

As already mentioned, the LLC corpus was chosen for this study due to the fact that it was the largest and the most representative English learner corpus in terms of L1 backgrounds and proficiency levels available at the time. However, from the very beginning of this research it was clear that one of the limitations of this study would have to do with the LLC corpus in itself.

First, Longman does not provide a theoretical basis nor does it provide an explanation for how exactly the writing samples were classified into specific proficiency levels. After the data was collected, an NLP application was applied to all annotated files. The results show that most of the Longmans' hierarchy of L2 proficiency levels

Table 5.1: Suggestions for limitations found in the present study

	Limitation type	Limitation	Suggestion
1	Corpus-related	Assessment of the L2 proficiency levels within the corpus	Theoretical basis for determining the assignment of the texts into L2 proficiency levels is needed. Quantitative examination of the annotated data for the reliability of the proficiency level setting can be suggested.
2		Lack of the representation of certain proficiency levels in each L1 groups in the corpus	Closer examination of the subsection of the corpus selected for further examination is necessary to ensure that the data from all the learner groups is equally representative of the target population.
3	Annotation scheme–related	The structure of the scheme itself	The annotation scheme should avoid overlap between the categories and be focused on the features which are of interest to the researcher
4		Subjectivity of the error annotation process	Subjectivity of the error annotation process could be reduced by using more than one annotator.
5	Methodological and procedural	Size of the annotated data sample	More than one annotator should annotate the data in order to arrive at a larger annotated data sample
6		No inter-annotator reliability could be obtained	More than one annotator should annotate the data in order to be able to establish inter-annotator reliability

does correspond to the difficulty levels of the writing samples; however, there is one major exception. The difficulty of the files assigned into upper intermediate (UI) proficiency levels quite often tends to be higher than that of the advanced (AD) proficiency level files, which is exactly the opposite of the hierarchy of these levels relative to each other as they appear in the LLC (Pendar, 2007). Obviously, this is something that could have somewhat obscured the results of the study if at least one of the research questions in this study involved a comparison of the data, not only by L1 group, but by the L2 proficiency level as well. Hence, a quantitative examination of the annotated data for the reliability of the proficiency level setting can be suggested to overcome this limitation. For example, an NLP application can be applied to all annotated files in order to determine if the file difficulty indeed corresponded to their assignment into L2 proficiency levels as they appear in the LLC.

Second, another limitation, which was not apparent until most of the data was already gathered, is rooted in the LLC corpus characteristics as well as its lack of the representation of certain proficiency levels in L1 groups. The original goal of the annotation was to acquire ten files per each of the two L1 groups (i.e., Chinese or Czech) and per each of the eight L2 proficiency levels. However, after closer examination, it became clear that LLC did not contain a minimum of ten files contributed by Czech and Chinese writers per all proficiency levels; LLC does not contain any files in the Chinese beginner (BE) level. It contains only nine files in the Czech Academic prose (AS) category, and five files in the Czech PR level. Consequently, only the available files in these categories were annotated. This obstacle resulted in missing an opportunity to collect the ideal data representation across all the proficiency levels per each L1 group as

the study design originally presumed, which turned out to be an additional limitation for this study. Hence, a closer examination of the subsection of the corpus selected for further examination is necessary to ensure that the data from all the learner groups is equally representative of the target population.

5.2.2 Annotation Scheme-Related Limitations

The structure of the error annotation scheme itself might have had an influence on the study and its results, since not all of the grammatical features which were later selected for the final analysis were always completely salient in the original structure of the scheme. For instance, it might have been one of the reasons why the results of Research Question 6 seem to be contrary to the original hypothesis. Research Question 6 investigated whether the overall number of passivization attempts was higher in the English IL of Chinese learners than in the English IL of Czech and Slovak learners for learners across levels of proficiency in English. However, the results of the Wilcoxon Rank test, summarized in Table 4.6, showed no significant difference between overall number of passivization attempts made by the Czech writers compared to the number of the same type of errors made by Chinese writers. This might be attributed to the fact that the annotation scheme itself did not separate passivization into its own category. Hence, the features extracted from the annotated data, which were selected as better predictors of passivization attempts such as additions of auxiliaries, perhaps might have not fully accounted for all the passivization attempts encountered in the annotated data, which presents an additional limitation to the study. Therefore, it might be suggested that the annotation scheme should avoid overlap between the categories and be focused on the features which are of interest to the researcher.

Additionally, the subjectivity of the error annotation process might be a limitation in itself, since in most cases the correct version inserted in place of an erroneous utterance is a clear choice, but in other cases it is simply one of the several equally acceptable corrections—especially in cases with multiple grammatical and sentence order errors occurring within one clause. Apart from employing *the principle of minimal edit* (i.e., the fewest editing steps that yields an acceptable utterance were implemented to correct any given error), there certainly were cases when the author had to rely on the context of the entire paragraph of text to infer the intended meaning of a given sentence. Using more than one annotator might slightly reduce the subjectivity of the annotation process.

5.2.3 Methodological and Procedural Limitations

As mentioned in Chapter 3, *Methods and Materials*, the data collection began in August 2006 and was completed in March 2007. The data collected for analysis included 159 fully tagged files written by speakers of Czech and Chinese and their varieties, totaling 36,237 words. However, since all of the files were manually tagged for sentence type and for word order and then manually tagged for each error type, the annotation process was extremely time-consuming and labor-intensive. The availability of only one annotator can also be considered a significant limitation for the data-gathering process, since it limited the size of the data sample to about 36,000 words when a larger data sample would have been preferable. Hence, it can also be suggested that more than one annotator should annotate the data in order to arrive at a larger annotated data sample.

Furthermore, all the tagging was done by the author, who consulted the principal investigator who finalized all the of the tagging and error classification decisions. Although it is believed that the tagging is reasonably reliable, no reliability estimates could be obtained with the use of only one annotator. This presents an additional limitation to the present study. Therefore, similarly to the suggestion for a limitation 4 and 5 above, it is suggested that more than one annotator should annotate the data in order to be able to establish the inter-tagger reliability.

5.3 Suggestions for Further Research

Findings of this study reaffirmed that certain L1 grammatical features do transfer into learners' L2 writing. These results supply valuable insights in planning and using computerized learner corpora as a research basis in L2 development through computer-aided error analysis of ESL writing. As it is, this study can be considered only the beginning of the inquiry, and further research is needed in regard to the effects of learners' L1 on their L2 writing. There is a definite need to explore the relationship between L1 and L2 errors in more depth utilizing contemporary error analysis methods. Specifically, further research is needed to determine how the L1 typology and rhetorical traditions influence learners' L2 production, not just in writing, but also in respect to other language skills—both in isolation and within the tasks that employ a combination of several primary language skills. The exploration of learners' L2 production with other L1 backgrounds is certainly of interest, as well as research exploring the influence of L1 and L2 typology on other language skills.

Furthermore, studies aiming to find other language use patterns and error in learners of English as a second language with respect not only to their native language, but also to learners' L2 proficiency levels are integral for subsequent progress in ELT and teacher training. It is unquestionable that the results of such studies would inform the development of the next generation of ELT tools and materials, specifically in allowing for more learner autonomy and learner-centeredness of these teaching tools and techniques.

5.4 Implications

This study provided solid empirical support that L2 writers transfer at least some of the grammatical patterns of their L1 into their ESL writing. It has also reaffirmed the view that contemporary error analysis—especially based on the larger, diverse, and representative computerized learner corpora—carries a great potential and should be explored further in the SLA research, as it can aid in ELT and DDL and can facilitate the production of new CALL and ICALL tools and programs, as well as inform traditional material development in attaining a higher degree of learner-awareness.

As mentioned above, apart from the existing negative attitude pertaining to error analysis, a large number of CALL programs, particularly the ones which managed to successfully incorporate NLP techniques, are based on the underling principle that error correction (whether explicit or implicit) carries a positive effect on the learners' L2 development. A number of very reputable SLA studies such as those by Nagata (1995, 1997, 2002) supply concrete empirical evidence in support of it. As pointed out by Cowan et al. (2003), the majority of the research studies related to the use of CALL

examine the topic in regard to the learners of low to intermediate proficiency, which explains the lack of evidence pertaining to the long-term effects of the instructional approaches that embrace error correction. Hence, the long-term effects of error correction (L1 transfer errors and "persistent" errors) need to be explored further in SLA and teacher training as well as in traditional material and CALL development.

Furthermore, the fact that learners of English are heterogeneous—that they do not learn under the same conditions, under the same educational system, or with the same amount and quality of exposure to the target language (TL)—naturally leads to the conclusion that the pedagogical approaches in language teaching in general are driven toward more learner-centred approaches. Another important difference is the heterogeneity of their L1 backgrounds, which becomes most apparent in ESL settings with a variety of L1 backgrounds represented, which is the case for many ESL classrooms, especially in university settings. Accounting for these differences is important for the development and production of better learner materials. It becomes self-evident that linguistic research of this type is integral to the development of more adaptive technologies for language teaching and assessment (Granger, 2002).

REFERENCES

Aijmer, K. (2002). Modality in advanced Swedish learners' written language. In S. Granger, J. Hung & S. Petch-Tyson (Eds.) *Computer learner corpora, second language acquisition and foreign language teaching*, (pp. 77-116). Amsterdam: John Benjamins Publishing.

Allerton, D. J., Tschichold, C. & Wieser, J. (Eds.) (2005). *Linguistics, Language Learning and Language Teaching.* Basel: Schwabe.

Altenberg, B. (2002).Using bilingual corpus evidence in learner corpus research. In S. Granger, J. Hung & S. Petch-Tyson (Eds.) *Computer learner corpora, second language acquisition and foreign language teaching*, (pp. 37-54). Amsterdam: John Benjamins Publishing.

Chalhoub-Deville, M. (Eds). (2006). *Inference and Generalizability in Applied Linguistics: Multiple perspectives.* Philadephia: John Benjamins.

Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology.* Cambridge: Cambridge University Press

Connor, U. & Precht, K. (2002). Business English: learner data from Belgium, Finland and the U.S. In S. Granger, J. Hung & S. Petch-Tyson (Eds.) *Computer learner corpora, second language acquisition and foreign language teaching*, (pp. 175-194). Amsterdam: John Benjamins Publishing.

Corder, S. P. (1974). Error Analysis. In J. P. Allen & P. Corder (Eds.), *The Edinburgh Course in Applied Linguistics. Volume 3 – Techniques in Applied Linguistics* (pp. 122-131). London: Oxford University Press.

Corder, S. P. (1981). *Error Analysis and Interlanguage.* Oxford: Oxford University Press.

Cowan, R., Choi H. E., & Kim, D. H.. (2003). Four Questions for Error Diagnosis and Correction in CALL. *CALICO Journal. 20*(3), pp. 451-463.

Dagneaux, E. Denness, S., and Granger, S. (1998). Computer-Aided Error Analysis. *System 26*(2), pp. 163-174.

Dulay, H. C. & Burt M.K. (1974). You Can't Learn without Goofing. In J.C. Richards (Ed.), *Error Analysis: Perspectives on Second Language Acquisition* (pp. 95-123). London: Longman.

Dulay, H. C., Burt, M. K., and Krashen, S. D. (1982). *Language Two*. New York: Oxford University Press.

Ellis, R. (1994). *The Study of Second Language Acquisition.* Oxford: Oxford University Press.

Krashen, S. D. (1982). The fundamental pedagogical principle in second-language teaching. *Studia Linguistica 35* (1-2) pp.50-71

Gass, S. M., & Selinker, L. (2001). *Second language acquisition: An introductory course.* Mahwah, NJ: Erlbaum.

Granger, S., Hung, J. & Petch-Tyson, S. (ed.) (2002). *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: John Benjamins Publishing.

Gillard, P. & Gadsby, A. (1998). Using a learners' corpus in compiling ELT dictionaries. In S. Granger (Ed.), pp. 159-171

Granger, S. (ed). (1998). *Learner English on Computer*. London: Addison Wesley Longman Limited.

Granger, S. (1998). The computer learner corpus: A versatile new source of data for SLA research. In S. Granger (1998), pp. 3–18.

Granger, S.(1999). Use of tenses by advanced EFL learners: evidence from an error-tagged computer corpus. In H. Hasselgard & S. Oksefjell (Eds.), *Out of Corpora. Studies in Honour of Stig Johansson* (pp. 119-131).

Granger, S. (2002). A Bird's-eye View of Computer Learner Corpus Research. In Granger S., Hung, J. and Petch-Tyson, S. (Eds.) *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, (pp.3-33). Amsterdam: John Benjamins Publishing.

Granger, S. (2003). Error-tagged Learner Corpora and CALL: A promising Synergy. *CALICO Journal. 20*(3), pp. 465-480.

Granger, S., & Tyson, S. (1996). Connector usage in the English essay writing of native and non-native ESL speakers of English. *World Englishes, 15,* pp. 19-29.

Heift, T. & Schultz, M. (in press). *Errors and Intelligence in Computer-Assisted Language Learning: Parsers and Pedagogues.* New York: Routledge.

Hinkel, E. (2003). Simplicity without elegance: Features of sentences in L1 and L2 academic texts. *TESOL Quarterly*. *37*(2), pp. 275-301.

Hinkel, E. (2004). Teaching academic ESL writing: Practical techniques in vocabulary and grammar. New Jersey: Lawrence Erlbaum Associates.

Housen, A. (2002). A corpus-based study of the L2-acquisition of the English verb system. In S. Granger, J. Hung & S. Petch-Tyson (Eds.) *Computer learner corpora, second language acquisition and foreign language teaching*, (pp. 77-116). Amsterdam: John Benjamins Publishing.

James, C. (1998). *Errors in language learning and use. Exploring error analysis.* London & New York: Longman.

Ju, M. K. (2000). Overpassivization errors of Korean second language learners: The effect of conceptualuzable agents in discourse. *Studies in Second Language Acquisition, 22,* pp. 85-111.

Kosterina, A., and Haji-Abdolhosseini, M. (2006). Annotation of Learner Corpora: Developing a markup scheme, presentation at *Applied Linguistics Colloquium Series*, Iowa State University.

Kosterina, A., and Pendar, N. (2007). *Issues with Annotating Learner Corpora,* presentation at CALL club presentation series, Iowa State University.

Lado, R. (1957). *Linguistics Across Cultures*. University of Michigan Press, Ann Arbor.

Leech, G. (1998), Learner corpora: what they are and what can be done with them. In S. Granger (Ed.), *Learner English on Computer*, pp. xiv-xx.

Lennon, P. (1991). Error: Some Problems of Definition, Identification, and Distinction. *Applied Linguistics, 12*(2), pp. 180-196

Li, N. C. & Thompson, S.A. (1976): Subject and Topic: A New Typology of Languages, In C. N. Li, (Ed.) *Subject and Topic*, (pp. 457- 490). New York, San Francisco, London: Academic Press.

Nagata, N. (1995). An effective application of natural language processing in second language instruction. *CALICO Journal. 13*(1), pp. 47-67.

Nagata, N. (1997). An experimental comparison of deductive and inductive feedback generated by a simple parser. *System, 25*(4), pp. 515-534.

Nagata, N. (2002). BANZAI: An application of natural language processing to web-based language learning. *CALICO Journal. 19*(3), pp. 583-599.

Nerbonne, J.(2003). Natural language processing in computer-assisted language learning. In R. Mitkov (Eds.) *The Oxford Handbook of Computational Linguistics,* (pp.670-698). Oxford: Oxford University Press.

Pendar, N. (2007). Learner Language Characterization Using Generalized Instance Sets. Poster presented at AAA07.Costa Mesa, CA.

Pendar, N. and Chapelle, C. (in press). Investigating the Promise of Learner Corpora: Methodological Issues. *CALICO Journal. 24*(2)

Pravec, N. (2002). Survey of learner corpora. *ICAME Journal 26*, pp. 81-114.

Ringbom, H. (1987). *The role of the first language in foreign language learning.* Clevedon & Philadelphia: Multilingual Matters.

Rimrott, A., & Heift T., (2005). Language Learners and Generic Spell Checkers in CALL. *CALICO Journal, 23*(1).

Salaberry, M. R. (2001). The use of technology for second language learning and teaching: A retrospective. *The Modern Language Journal*, *85*(1), pp. 39-56.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

Schumann, J. H. (1984). Nonsyntactic speech in the Spanish-English basaling. In R., Andersen, (Ed.) *Second language: A cross-linguistic perspective,* (pp. 355- 374). Newbury House Publishers Rowley, Massachusetts.

Schwartz and Gubala-Ryzak, (1992). Learnability and grammar reorganization in L2A: Against evidence causing the unlearning of verb movement. *Second Language Research, 8,* pp. 1-38

Schwartz, B. D. & Sprouse, R. A. (1996). L2 cognitive states and the full transfer/full access model. *Second Language Research, 12*(1), pp. 40-72

Selinker, L. (1974). Interlanguage. In J.C. Richards (Ed.) *Error Analysis: Perspectives on Second Language Acquisition* (pp.31-54). London: Longman.

Selinker, L. (1992). *Rediscovering Interlanguage.* London: Longman.

Skehan, P. (1998). A *cognitive approach to language learning.* Oxford: Oxford University Press.

Short, D. (1987). Czech and Slovak in B. Comrie (ed.): *The World's Major Languages.* New York: Oxford University Press.

Sinclair, J. (Ed.) (1996). *How to Use Corpora in Language Teaching.* Philadelphia: John Benjamins Publishing.

Si-Quing, Chen, & Xu Luomai. (1990). "Grammar-Debugger: A Parser for Chinese EFL Learners." *CALICO Journal, 8* (2), pp. 63-75.

Straudle, A-M. (1984). A comparison of a Spanish- English and a Japanese-English second language continuum: negation and verb morphology. In R. Andersen (Ed.) *Second language: A cross-linguistic perspective* (pp. 323-354). Newbury House Publishers: Rowley, Massachusetts.

Thomas, M. (1989). The interpretation of English reflexive pronouns by non-native speakers. *Studies in Second Language Acquisition, 11*(3), pp. 281-302.

White, L. (1992). Long and short verb movement in second language acquisition. *Canadian Journal of Linguistics, 37* (1), pp. 273-286.

Wolfe-Quintero, K., Inagaki, S. & Kim, H-Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity*. Honolulu: University of Hawaii Press.

Yip, V. (1995). *Interlanguage and learnability: From Chinese to English*. Amsterdam: John Benjamins.

Young, R. (1988). Variation and the interlanguage hypothesis. *Studies in Second Language Acquisition, 10*, pp. 281-302.

Young, R. (1991). *Variation in Interlanguage Morphology.* New York: Peter Lang.

Young, R. (1993). Functional Constraints on Variation in Interlanguage Morphology. *Applied Linguistics, 14* (1), pp. 76-97.

Zobl, H. (1989). Canonical structures and ergativity. In S.M. Gass & M. Schachter (Eds.) *Linguistic perspectives on second language acquisition.* (pp. 203-221). New York: Cambridge University Press.

Websites discussed in this document

CALLISTO, Open-Annotation Workbench
http://callisto.mitre.org

Longman Learner Corpus (LLC)
http://www.longman-elt.com

Longman Dictionary of Contemporary English (LDOCE)
http://www.longman-elt.com/dictionaries/research/dictres.html

APPENDIX A

Characteristics Summary of Currently Available Learner Corpora

Characteristics Summary of Currently Available Learner Corpora

Corpus Name	Corpus Type (based on Figure 1)				Source Language	Corpus size - # of words	Origin/ Location	Availability/ Purpose	Specifics
	Monolingual (M) or Bilingual (B)	General (G) or Technical (T)	Synchronic (Sync) or Diasynchronic (D)	Written (W) or spoken (spk)					
CLC	M	G	sync	W	Multiple	> 10,000,000	Great Britain	Commercial	N/A
CYLIL	M	G	D	W		N/A	N/A	Academic	
HKUST	M	G	sync	W	Cantonese	> 25,000,000	University of Science & Technology, Hong Kong	Academic	N/A
IBLC	M	T/business	sync	W	Belgian, Finish	N/A	N/A	Academic	Technical/
ICLE	M	G	sync	W	Multiple	> 2,000,000	University of Louvain-La-Neure, Belgium	Academic	N/A
FRIDA			sync	W	French	~ 450,000	Center for English Corpus Linguistics, University of Louvain	Academic	2/3 were fully error-tagged
JEFLL	M	G	sync	W	Japanese	> 500,000	Meikai University, Japan	Academic	
JPU	M	G	sync	W	Hungarian	> 400,000	University of Pecs, Hungary	Academic	
LINDSEI	M	G	sync	Spk	Multiple	~ 100,000	Center for English Corpus Linguistics, University of Louvain	Academic	contains transcripts of 50 interviews
LLC	M	G	sync	W	Multiple	~ 8,000,000	Great Britain	Commercial	N/A
MELD	M	G	sync	W	Multiple	~ 100,000	Montclair State University, USA	Academic	N/A
PELCRA	M	G	sync	W	Polish	~ 93,000,000	University of Lodz, Poland	Academic	N/A
TSLC	M	G	sync	W	Cantonese	N/A	Hong Kong University, Hong Kong	Academic	N/A
USE	M	G	sync	W	Swedish	N/A	Uppsala University, Sweden	Academic	N/A

APPENDIX B

LLC Contributor Characteristics (adopted from Longman Learner Corpus, Longman)

Native Language	Over 70 different languages and 180 varieties
Source country	16 countries
Learner level	Beginner, Elementary, Pre-intermediate, Intermediate, Upper Intermediate, Advanced, Proficiency, and Academic Studies
Environment	pertaining to the task performed, including several standardized examinations, internal examinations, in-class assignments, homework, authentic letters and documentation, and business communication document
Task	set essay, free essay, project essay, exercise, letter, advertisement, report, speech and diary
Target language	British English, American English or Australian English; however, the goal is to focus on American English.

APPENDIX C

List of Abbreviations of the annotation scheme

S	**sentence**
NP	**noun phrase**
VP	**verb phrase**
PP	**prepositional phrase**
AdjP	**adjective phrase**
AdvP	**adverb phrase**
N	**noun**
V	**verb**
Aux	**auxiliary verb**
Det	**determiner**
Adj	**adjective**
Adv	**adverb**
P	**preposition**
Part	**particle**
Inter	**interjection**
Conj	**conjunction**
Pro	**pronoun**
Rel	**relative pronoun**
WH_NP	**wh-noun phrase (who, what...)**
WH_DET	**wh-determiner (which, what, whose...)**
WH_ADV	**wh-adverb (why, where, when,...)**

APPENDIX D

Structure of the Annotation Scheme

SENTENCE	
TYPE	**WORD ORDER**
▪ **Declarative** John bought a car. ▪ **Imperative** Go, buy the car. ▪ **Interrogative** Have you bought the car? ▪ **Exclamative** What a nice car!	▪ **Canonical Word Order:** John bought a car. ▪ **Cleft:** It was a car that John bought. It was John who bought a car. ▪ **Pseudo-cleft:** What John bought was a car. Who bought a car was John. ▪ **Reversed pseudo-cleft:** A car is what John bought. John is the one who bought a car. ▪ **Topicalized:** John, he bought a car. A car, John bought.

ERROR	
TEXT/UTTERANCE *The lexical item or a sequence of lexical items marked for correction*	
CORRECT	*the correct form using the principle of minimal edit*
LOCUS *The shortest (intended) constituent (preferably indicative of error)*	s, np, vp, pp, adjp, advp, n, v, aux, det, adj, adv, p, part, inter, conj, pro, rel, wh-np, wh-det, wh-adv
NOTE *any notes that might be relevant to a given error*	Any specific notes such as *overpassivization attempt*
SPECIFICS *error specifics*	**(apply to substitution, omission, addition)** n, v, aux, adj, adv, p, part, inter, conj, pro, rel, wh-np, wh-det, wh-adv **(applies to addition)** dummy-subj: The weather, it is bad.

	(apply to morpho-syntactic) added-marker misplaced-marker missing-marker **(apply to agreement)** acc-for-nom other-for-3sg pl-for-sg nom-for-acc 3sg-for-other sg-for-pl **(apply to substitution)** base-for-gerund indef-for-def gerund-for-base def-for-indef gerund-for-noun none-for-indef noun-for-gerund none-for-def indef-for-none wrong-inflection def-for-none active-for-passive cardinal-for-ordinal passive-for-active ordinal-for-cardinal
SUBSTITUTE *Lexical item or a sequence with which the intended correct item/sequence was substitutes*	n, v, aux, adj, adv, p, part, inter, conj, pro, rel, wh-np, wh-det, wh-adv
TYPE *the most general error domain; in certain cases expanded to* specifics *category*	**addition** **aspect:** wrong aspect marking **morpho-syntactic:** misused marker: possessive, plural, infinitive **omission** **order** **other** **overgeneralization:** overuse of a rule **parallelism** **fragment** **run-on** **repetition** **spelling** **substitution** **agreement:** subj-verb, det-n **collocation** preposition: misused preposition **tense:** wrong tense marking on the verb **voice** **transitivity:** using a transitive verb as intransitive, etc.

CORRECT	the correct form using minimal editing
ID	an id assigned to the incorrect portion of an collocation
SUBSTITUTE	the incorrect category used in a substitution n, v, aux, det, adj, adv, p, part, inter, conj, pro, rel, wh-np,

	wh-det, wh-adv
COLLOCATE id: refers	the error id to which the correct portion of a collocation

NB: * The list of abbreviations used to describe a mark-up scheme is provided in Appendix C.

APPENDIX E

The File Header Descriptions (adopted from Longman Learner Corpus, Longman)

The file names are constructed from information taken from the header. The first four or five characters are the < RF > number, the next two are the < LE > code and the last character is the < TT > code. The three characters that form the file extension are taken from the < LA > code.

The Header

Below is an example of a header for file
30557.BE1.CZS

< RF > 30557
< CO > CZE
< LA > CZS
< LE > BE
< EN > CLA
< TT > 1
< TV > BrE

File reference number	30557
The source country code	Czech (CZE)
Source language	Slovak (CZS)
Proficiency level	Beginner (BE)
Environment	Class work(CLA)
Text type	Set essay (1)
Target language variety	British English (BrE)

< RF > The document reference number (4. or 5 digits)

I

< CO > The source country code (3 digits), which shows where the script originated
e.g. CZE Czech Republuc and Slovakia (e.g. former Czechoslovakia)
CHE China

< LA > Student language code (3 digits). This refers to the student's mother tongue
e.g. CZS Slovak (Slovakia)

CZE Czech (Czech Republuc)
CHK Hong Kong (China)
CHC Chinese (China)

< LE > Student level
code (2 digits)

BE	beginners
EL	elementary
PI	pre-intermediate
IN	intermediate
UI	upper intermediate
AD	advanced
PR	proficiency
AS	academic studies

< EN > environment code (3 digits)
e.g.

CPE	Cambridge proficiency exam
CAE	Certificate of Advanced English
FCE	First certificate exam
PET	Preliminary English Test
CUE	CUEFL exam
INT	internal exam
CLA	classwork
HOM	homework
AUL	authentic letter '
AUD	authentic document
LON	Longman
BUS	business

< TT > Task type code (1 digit)

1	….set essay
2	.... free essay
3	.... project essay
4	.... exercises
5	letter/correspondence
6	advertisement
7	report
8	speech
9	diary

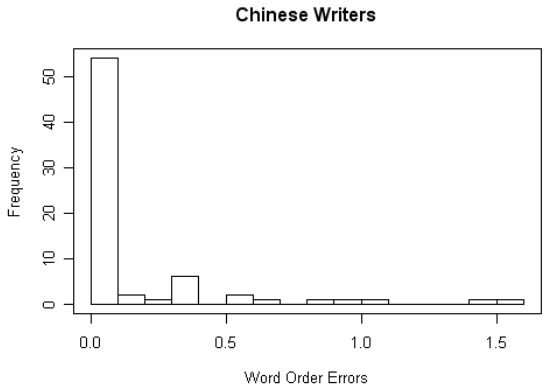
< TV > Target language variety
(3 digits)
e.g.

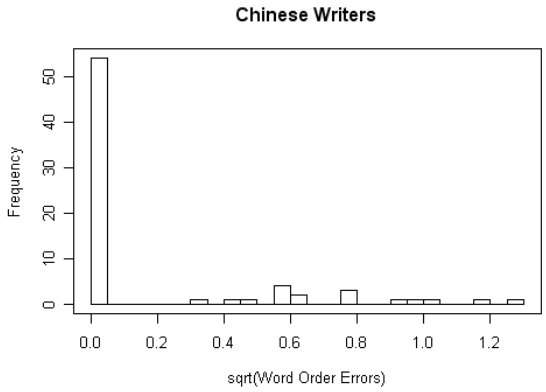
BrE	British English
AmE	American English
AuE	Australian English

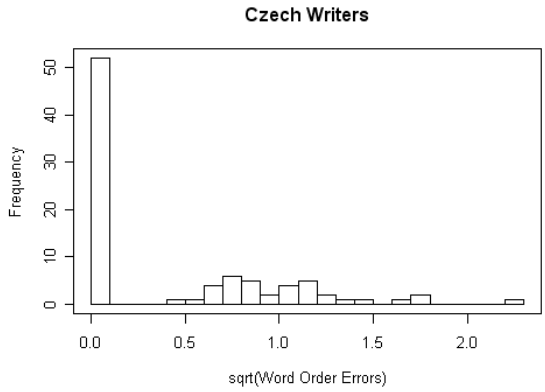
APPENDIX F

Graphic Representations of the Data for Each Research Question

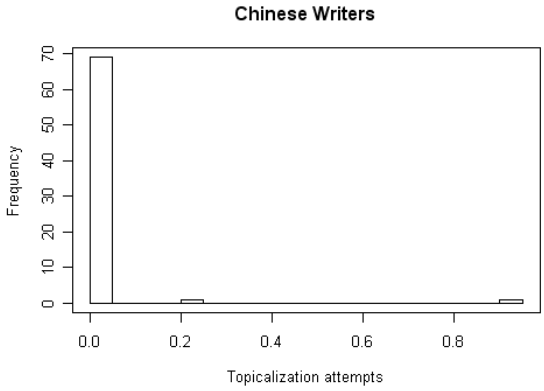
Research Question 1: Word order error

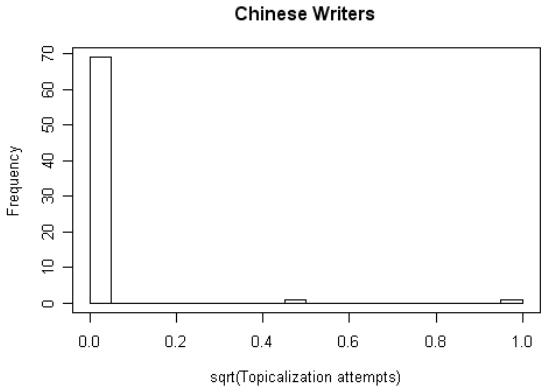


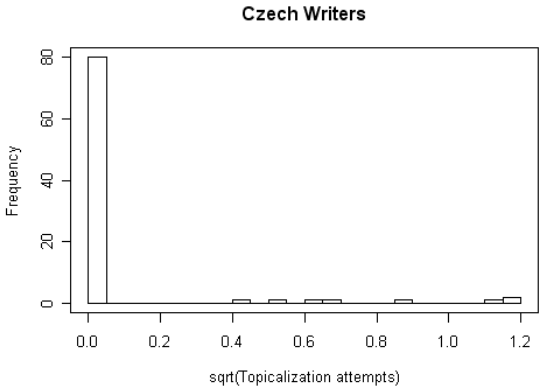




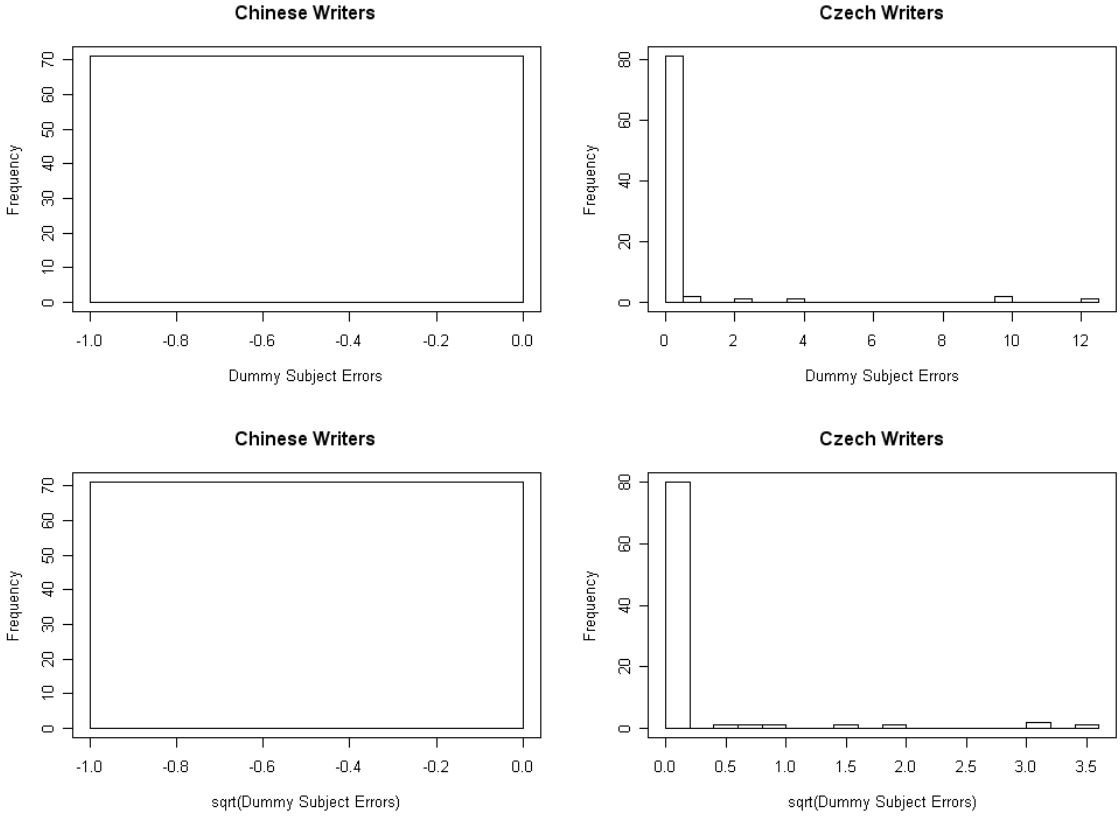
Research Question 2: Topicalization Attempts



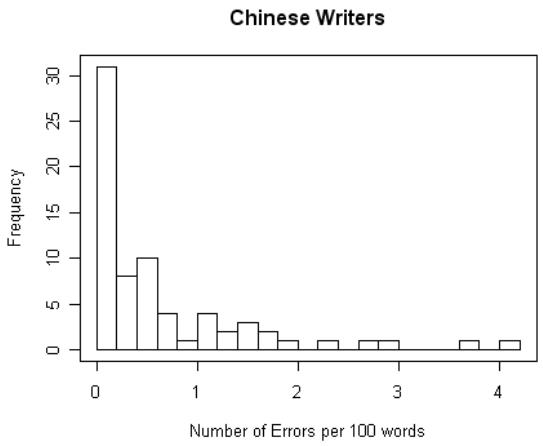


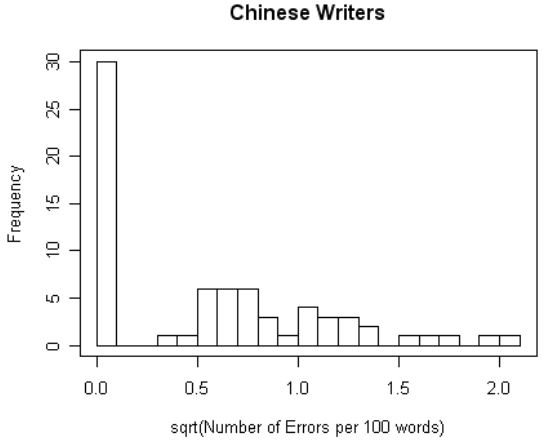


Research Question 3: Expletive-subject Errors



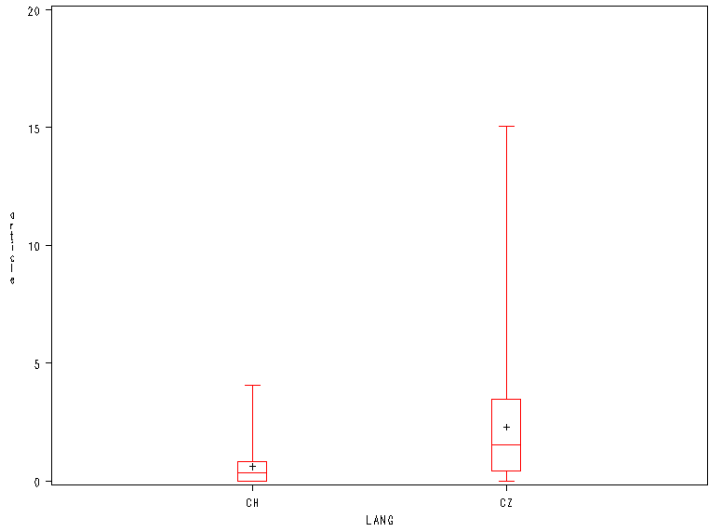
Research Question 4: Article Errors



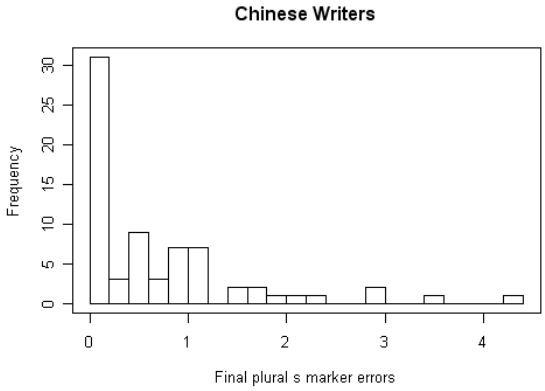


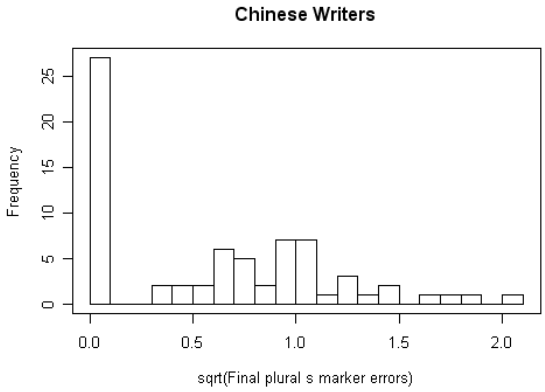
Research Question 4 (cont.)

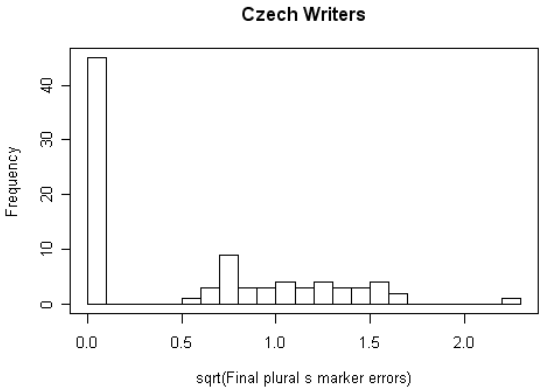
Boxplot for Article errors



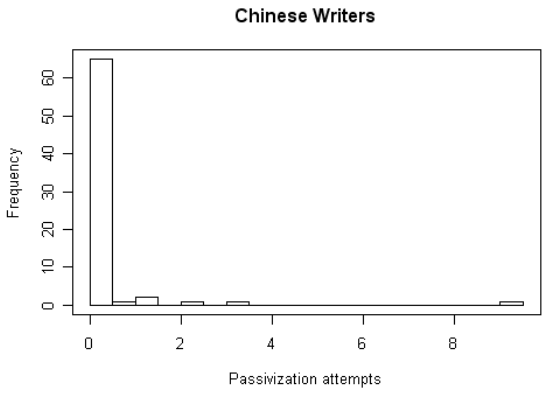
Research Question 5: Final Plural (s) Marker Errors

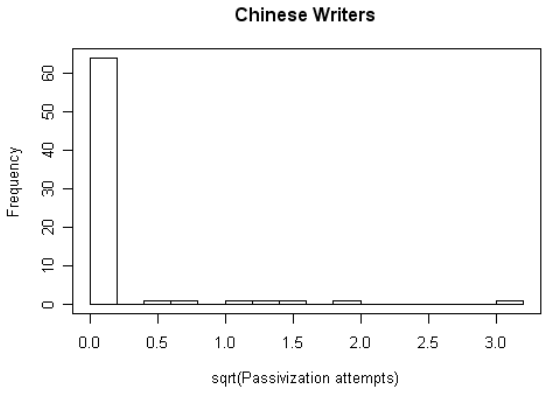




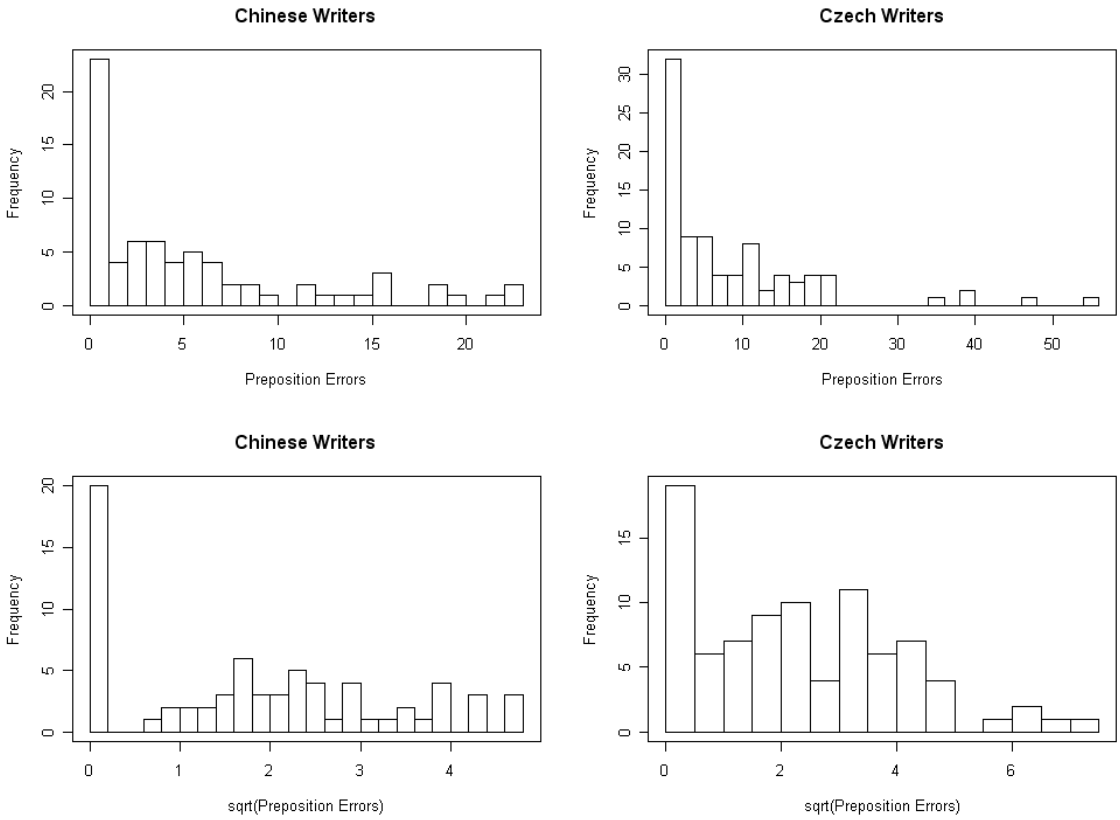


Research Question 6: Passivization Attempts





Research Question 7: Preposition Errors



AKNOWLEDGEMENTS

I am grateful to all of my committee members, my major professor, Professor Dan Douglas, Dr. Nick Pendar, and Dr. Geoff Sauer for keeping me on track to complete the thesis. I am indebted to Dr. Nick Pendar for his guidance and ideas. Finally, I would like to thank my husband for his encouragement and patience.