SPELLING, NOUN PHRASE AND VERB PHRASE ERRORS IN THE WRITING OF LIBYAN LEARNERS OF ENGLISH: A CORPUS-BASED ANALYSIS

by

AHMEDA MOHAMMED ALTOATE

A thesis submitted to the University of Birmingham for the degree of

DOCTOR OF PHILOSOPHY

Department of English Language and
Linguistics
School of English, Drama and Creative
Studies
College of Arts and Law
The University of Birmingham
July 2021

Abstract

Within the broad field of learner error analysis, there is now a growing tradition of studies research focusing on errors produced by Arabic L1 learners of English as a second/foreign language. However, these previous studies have suffered from a number of important limitations. Many of them have focused on small numbers of features or even just single features (e.g., articles, tenses, auxiliaries, etc.). Many have also been based on very small datasets, and most have only been cross-sectional in perspective. Therefore, this thesis seeks to address these shortcomings by conducting a detailed and quasi-longitudinal analysis of spelling, noun and verb phrase errors produced by Arabic L1 undergraduate learners of English at Benghazi University, Libya. The study reported in this thesis adopts a corpus-based approach, applying computational tools and methods to the analysis of a large database of learner texts which I have called the *Libyan English as a Foreign Language Learners* (LEFLL) corpus. The aim of this thesis is to provide SLA researchers with a broader picture of the role of L1 in producing spelling, noun and verb phrase errors in the writing of Arab English learners and how the role of L1 influence develop across the three different university levels. The analysis revealed errors that can be attributed to the first language influence was found to be more prevalent and the most obvious in spelling errors followed by verb phrase errors. The analysis also showed that interlingual spelling errors followed a steady decline as from one level to another, but this was not the case in noun and verb phrase errors.

As well as providing a broader picture of the role of L1, the thesis also offers a critical appraisal of the two main approaches to classifying and quantifying errors that have been put forward in previous research. In an attempt to overcome the serious limitations of both of these approaches, this thesis introduces a new approach to error counting, which I have called the '*potential for error counting approach'.* It will be argued that this approach offers several advantages over other error counting approaches.

Dedication

First and foremost, I dedicate my work to my dearest parents. My mom who passed away on 24th March 2021 because of COVID-19 (may Allah, Subhana wa ta ala grant her Jannat-ul-firdos) and dad (may Allah, Subhana wa ta ala bless him with good health and righteous long life!). I also dedicate my work to my awesome family (my wife and children may Allah, Subhana wa ta ala bless them all!)

Acknowledgements

In the name of Allah, the Beneficent, the Merciful

I would like to start by praising and thanking Almighty Allah whose compassion and mercy have provided me with guidance and determination to complete this Ph.D. thesis.

Following my praise and thanks to Allah, I would like to express my sincerest gratitude to my lead supervisor Dr Nicholas Groom for his invaluable support and guidance throughout my thesis journey. His knowledge and dedication were inspirational to me, and he always motivated me to strive for the highest standards in my research. I sincerely thank him for his confidence in me despite the challenges I went through during my Ph.D. journey. Without his invaluable support, this thesis could never has been completed.

I would also like to thank my co-supervisor Dr Petra Schoofs, whose invaluable feedback has helped me in every chapter of my thesis. Although she was only relatively recently appointed as my co-supervisor, her dedication and insight were really helpful and appreciated. I am really grateful to her for her generous support.

I am also grateful to Dr Oliver Mason, who was my first lead supervisor when I started my Ph.D. His knowledge and insight, during the period I worked with him before leaving the University of Birmingham, has helped me to shape my thesis.

Finally, and most importantly, I would like to extend my deepest gratitude to my awesome family (my parents, my wife and children and my brothers and sisters) for their continuous love and support. Without them I could never have competed this thesis. Thanks also go to all my relatives and friends who have encouraged me and wished me the best. Big and special thanks go to my dearest friend Khaled Hassan Ghirani for his invaluable support.

Table of contents

List of Figures

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

Chapter 8

List of Tables

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

Chapter 8

Abbreviations/Acronyms	**Meaning**
BALC	BUiD Arab Learner Corpus
BAWE	British Academic Written English
BNC	British National Corpus
BUiD	British University in Dubai
CAWE	Chinese Academic Written English
CEA	Computer-aided Error Analysis
CEPA	Common Education Proficiency Assessment
CIA	Contrastive Interlanguage Analysis
CLAWS	Constituent Likelihood Automatic Word-tagging System
CVCV	consonant vowel consonant vowel
EAP	English for Academic Purposes
EFL	English as a Foreign Language
ELLs	English Language Learners
ELT	English Language Teaching
ESL	English as a Second Language
EXJA	Expert Journal Articles
FL	Foreign Language
FRIDA	French Interlanguage Database
ICLE	International Corpus of Learner English
IL	Interlanguage
ILV	Interlanguage Varieties
JEFLL	Japanese English as a Foreign Language
L1	First Language
L2	Second Language
LEFLL	Libyan English as a Foreign Language Learners
LOCNESS	Louvain Corpus of Native English Essays
NATO	North Atlantic Treaty Organization
NL	Native Language
NNS	Non-native speaker
NS	Native Speaker

PFEC	Potential for Error Counting
POA	Potential Occasion Analysis
POMAS	Phonological, Orthographical and Morphological Assessment of Spelling
POS	Part-Of-Speech
RLV	Reference Language Varieties
SL	Second Language
SLA	Second Language Acquisition
TCA	Traditional Contrastive Analysis
TEA	Traditional Error Analysis
TEC	Traditional Error Counting
TL	Target Language

CHAPTER 1: INTRODUCTION

1.1 Aims of the thesis

This thesis has three broad aims. The first is to carry out the first comprehensive and quasi-longitudinal error analysis study of the written English produced by Arab first language speakers. The study is based on a specially designed learner corpus consisting of 559 essays collected from Libyan undergraduate students at three different university levels. At 60,131 words tokens in total, this corpus is in one sense a rather small one by the standards of contemporary corpus linguistics. However, the corpus can be viewed as much larger and more ambitious than this word count might make it seem in the sense that the student essays within it were manually transcribed and have been annotated with 122,511 error tags for the purposes of detailed error analysis. This corpus, which I will call the *Libyan English as a Foreign Language Learners* (LEFLL) corpus, consists of three sub-corpora, each of which represents one academic year of undergraduate study. This quasi-longitudinal design means that the corpus can also provide insights into how learner error patterns change over time.

As well as providing an unprecedentedly detailed and thorough picture of the types of error committed by Arabic L1 English learners at different levels, the thesis also aims to contribute more broadly to the rehabilitation of language transfer perspectives in interpreting the causes of language learners' errors. As will be discussed in more detail in Chapter 2, first language influence on second language learning is a concept that has gone in and out of fashion over the last century. It was very much in fashion in the 1950s, but then fell out of favour in the 1970s, where it was replaced by approaches that saw learner errors exclusively as reflecting learning difficulties inherent in the target language grammar. More recently, the concept of language transfer has begun to make a comeback and has received a particularly strong boost from the advent of corpus linguistics, which has helped to reveal systematic differences between the learners' first and target languages. Of course, it is not possible to *prove* that any given language learner error has been caused by first language transfer, but this thesis will seek to argue for the plausibility of such an interpretation wherever it is possible to do so. Specifically, the thesis will draw on three types of support in making such claims. Firstly, it will make a reference to recognised formal or functional differences between relevant aspects of English and Arabic grammar. Secondly, it will provide quantitative evidence that indicates that the error in question is pervasive across the student body as represented by the corpus under analysis, and not just restricted to one or a small number of learners. Finally, first language transfer interpretations will also be supported wherever possible by reference to previous studies which have identified the same error as being typical of Arab L1 learners of English, and which have also identified first language transfer as the most likely cause of the error in question. The analysis chapters of the thesis will start by reporting on my analysis of spelling errors (as the structure of this thesis below will also show), which provides the most obvious evidence of transfer effects from Arabic to English. Another reason why this thesis includes an analysis of spelling errors is that spelling errors have received comparatively little research attention in the past, and the research that has been done has been somewhat piecemeal (see Chapter 3 for more discussion). The analysis of spelling errors reported in this thesis is far more thorough and there will be a separate literature review chapter for the analysis of spelling errors to shed light on the importance of spelling as a level of analysis for learner error research. Following the analysis of spelling errors, the thesis will move on and analyse noun and verb phrase errors.

The third and final aim of this thesis is methodological than substantive; specifically, the thesis will introduce and demonstrate a new approach to error counting, which I have called the '*potential for error counting approach'.* This new approach emerged out of my growing dissatisfaction with the approaches to error counting that were available from the previous literature, and which I had originally intended to use in my own research. It will be argued that the new approach that I have developed offers several advantages over these previous error counting approaches.

1.2 Research questions

As stated above, the main aim of this thesis is to provide SLA researchers with a broader picture of the role of L1 in the writing of Arab learners of English. Using the tools and methods of corpus linguistics, the thesis will carry out an unprecedentedly detailed and comprehensive analysis of the characteristic types of written errors produced by university-level Libyan Arabic learners of English. The study focuses on the identification and classification of errors at three levels of analysis – spelling, noun phrases and verb phrases – in a corpus of essays collected from 559 Libyan university undergraduate students across three cohort year groups. In the analysis chapters, the thesis seeks to address the following two main research questions:

1) To what extent does L1 influence affect spelling, noun phrase and verb phrase errors in the writing of Arab English learners?
2) Does this influence follow the same pattern (either an increase or decrease) as the learners proceed across the university academic levels?

This thesis is not the first study of learner English writing produced by Arabic first language speakers. On the contrary, in recent years a small but growing number of studies analysing different types of written errors produced by Arabic L1 learners of English have been published (e.g., Khuwaileh & Shoumali, 2000; AbiSamra, 2003; Alhaysony, 2012; Al-Shujairi & Tan, 2017; Altamimi et al., 2018; Ibrahim, 2018; Mohammed & Shwater, 2018; Morgan, 2018; Nuruzzaman et al., 2018; Qaddumi & Walweel, 2018; Khan, 2019; Thyab, 2019) and some have been corpus-based analysis studies (e.g., Randall & Groom, 2009; Crompton, 2011; Sawalmeh, 2013; Yildiz, 2017; Btoosh, 2019)

However, there are three significant problems with these previous research studies. First, many of these studies are focused on smaller numbers of features or even just single features. To take just one example, Al Alhaysony (2012) studied the errors produced by Saudi female students in their use of English articles (both definite and indefinite). This analysis is thorough, detailed and pedagogically useful, but is typical in that it only provides information about this very specific aspect of learner language. The problem here is that there is currently no account that links observations made about individual features such as this into a broader overarching picture of learner English errors among Arabic L1 speakers and writers.

A second problem with previous studies is that many of them are based on very small datasets. For instance, Qaddumi and Walweel (2018)'s study is typical in that it analyses writing produced by only of 22 Palestinian EFL learners. Finally, most these previous studies are cross-sectional; this means that it is impossible to see how error trends develop and change over time as the learners proceeded from one level to another. The aim of this thesis, therefore, is to consolidate and ultimately to supersede these previous studies by being far more ambitious in both size and scope. As mentioned above, the corpus upon which the current thesis is based is quasi-longitudinal, in that it consists of written work spanning three years of university undergraduate study. It is hoped that the results of my analysis will prove to be valuable to SLA researchers interested in L2 writing development, and to English teachers who work with Arabic L1 speakers, making them aware of errors that could be produced by L1 Arabic speakers.

As well as providing new substantive knowledge about learner errors, my thesis also introduces and showcases a new approach to learner error counting, which I have called the *potential for error counting approach.* In this approach, I call for empirical investigation in the learner corpus to verify which linguistic features (e.g., articles, subject pronouns for noun phrase errors, word lengths for spelling errors, etc.) have incorrectly been produced before calculating the percentages of each error category within its relevant environment of potential for error (e.g., the percentages of article errors out of the total number of articles that have potential for errors, the percentages of spelling errors out of the total number of tokens/words that have potential for error, etc.). The thesis will contrast this new approach with the two main previously existing approaches to error quantification, and will demonstrate that my approach has several crucial advantages over these previous approaches.

1.3 A Summary of the Structure of the Thesis

The overall structure of this thesis takes the form of nine chapters. Following the current introductory chapter, **Chapter 2** begins by reviewing the two traditional (i.e., non-corpus-based) approaches to error analysis approaches, namely the *traditional contrastive analysis approach* and the *traditional error analysis approach.* I will then show how these traditional approaches have been revitalised by the arrival of learner corpus methods. I will then discuss how the concept of L1 transfer has changed from its original formulation as found in the contrastive analysis approach, to its new corpus-based incarnation, the *contrastive interlanguage analysis approach.* As we will see, this approach has served to rehabilitate the concept of first language transfer, and thus makes it possible to regard L1 transfer as a valid concept for explaining the frequencies of some specific types of errors. In later chapters, I will draw parallels with Arabic where relevant as a possible explanation for some observed error types in the LEFLL corpus.

Finally, chapter 2 will highlight the limitations of previous computerised error analysis approaches. In particular, we will see that these approaches are unable to show us the proportions of the linguistic features (e.g., spelling, tenses, articles, etc.) that the language learners achieved/produced correctly.

Despite the obvious importance of spelling as a language skill and the pervasiveness of spelling errors in learner data, spelling has received surprisingly little research attention in the learner corpus research literature thus far. It is therefore necessary for **Chapter 3** to prepare the ground for the large-scale empirical investigation of spelling errors reported in chapter 6, by critically reviewing previous spelling error analysis research studies, and by presenting the two general classification systems of spelling errors namely the Linguistic Category and Surface Strategy Taxonomies proposed by Dulay et al (1982) and that are widely used in spelling error research and are also used in the analysis of spelling errors in this thesis in Chapter 6.

Chapter 4 is concerned with describing the corpus that was compiled for the purposes of the current study and the methodology used to analyse it. More specifically, the chapter presents the LEFLL corpus designed for this thesis, the procedures that were followed to collect and error tag the corpus files and retrieve the tagged errors, and the inter-rater reliability test that was conducted to establish the validity of my error classifications.

Chapter 5 aims to compare the three error counting approaches, namely: the *traditional error counting approach* (TEC)*, the potential occasion analysis approach* (POA) developed by Thewissen (2012; 2015), and the *potential for error counting approach* (PFEC) introduced by this thesis. The frequencies of errors of six different error categories (spelling, noun phrase, verb phrase, prepositional phrase, adjective phrase and adverb phrase) identified in the LEFLL corpus are calculated and compared based on these three error counting approaches. The chapter will show that the *potential for error counting approach* offers distinctive advantages over the other two error counting approaches. Specifically, it provides us with information regarding the proportions of both the language learners' errors and what the language learners were able to produce correctly in their writing. The chapter concludes by identifying spelling, noun phrase and verb phrase errors as the three error categories that will form the central focus of the rest of the thesis.

Chapter 6 comprehensively analyses the spelling errors identified across the three LEFLL sub-corpora. As mentioned above, this analysis is based on the Surface Strategy and Linguistic Categories Taxonomies presented in Chapter 3. The surface strategy taxonomy analysis reveals that omission and substitution spelling errors constituted the highest percentages of spelling errors in the LEFLL corpus and the three LEFLL sub-corpora. Based on the linguistic categories taxonomy, the analysis showed that the average percentage of interlingual spelling errors in the LEFLL corpus (i.e., errors likely to be due to L1 influence) is 68.96% and there is a steady decrease in the percentages of interlingual spelling errors versus an increase in the percentages of intralingual spelling errors (i.e., errors likely to be due to difficulties inherent in the target language system alone) across the three LEFLL sub-corpora. The steady changes in the percentages of interlingual vs intralingual spelling errors across the three sub-corpora were used to interpret the results obtained via the potential for error counting approach in Chapter 5, Section 5.3. The percentages of spelling errors, as a whole, follow a steady decline across the three sub-corpora, and it will be argued that this reflects a declining influence of the learners' L1 across the sub-corpora.

Chapter 7 comprehensively analyses the noun phrase subcategory errors in the LEFLL corpus and across the three LEFLL sub-corpora. Since this thesis strongly argues for a language transfer perspective on the causes of these errors, the chapter will look for plausible L1 transfer interpretations for the noun phrase subcategory errors identified in the LEFLL corpus. The analysis revealed that the average percentage of interlingual noun phrase errors (46.89%) is less than the average percentage of interlingual spelling errors observed in Chapter 6. Across the three LEFLL sub-corpora, the analysis showed that there are no steady changes in the percentages of interlingual noun phrase errors and subcategory errors.

Chapter 8 analyses the verb phrase subcategory errors in the LEFLL corpus and across the three LEFLL sub-corpora. The analysis reveals that the average percentage of interlingual verb phrase in the LEFLL corpus (52.06%) is less than the percentage of interlingual spelling errors but higher than the percentages of interlingual noun phrase errors. In the same way as observed in the analysis of noun phrase errors, the analysis revealed that the percentages of interlingual verb phrase errors across the three LEFLL sub-corpora do not follow steady changes.

Chapter 9 summarises the main findings of the thesis. First, it reminds the reader of the aims the thesis sought to achieve. Second, it provides a summary of the main findings. Third, it indicates the main limitations of the thesis, and identifies potentially fruitful directions for future research.

CHAPTER 2: APPROACHES TO THE ANALYSIS OF LEARNER ERRORS

Introduction

As established in Chapter 1, this thesis seeks to achieve two aims. The main aim of the thesis is to provide the researchers of SLA with a broader picture of the role of L1 in the writing of Arab English learners based on empirical investigation on a large learner corpus compiled by the researcher of this thesis. To achieve this aim, I will conduct a comprehensive, quasi-longitudinal analysis study of three error categories: spelling, noun phrase errors and verb phrase errors in the writing of Libyan Arab English learners using the tools and methods of corpus linguistics.

Corpus linguistics is now well established as an approach to empirical linguistic research that focuses on the computer-assisted analysis of very large collections of attested language data that have been collected in order to represent a language or language variety of some kind. Within corpus linguistics, there are two main traditions of research focusing on second language learners – one focusing on pedagogy (i.e., the '*data-driven learning*' tradition pioneered by Johns (1991; cf. Boulton, 2007; 2008; 2009) and the other focusing on issues in SLA. My thesis is mainly a contribution to the latter, although I hope that my findings will be of use or relevance to the former as well. This chapter will provide the background to my study by first presenting the history of error analysis, and then showing how this field has been revitalized by learner corpus methods. In this chapter, I will make the case for a language transfer perspective on interpreting language learners' errors, and I will go on to interpret some of the findings of my own analysis from this perspective in the analysis chapters of this thesis (i.e., in Chapters 6, 7 & 8). The chapter concludes by explaining how the current thesis aims to build on previous learner corpus research by introducing and demonstrating a new approach to error analysis that I believe offers several advantages over the current approaches.

2.1 The Traditional Approaches to Learner Error Analysis

Long before the emergence of corpus-based research studies, language learners' errors have attracted considerable interest among applied linguistics researchers, particularly those who are interested in second/foreign language (henceforth SL/FL) teaching and learning. Traditionally, two broad error analysis approaches have been distinguished: contrastive analysis and error analysis. Both approaches, as their names suggest, are based on the comparative analysis of the learners' first language, the target language they are learning, and the ever-changing learner 'interlanguage' that represents the learners' current state of knowledge as they progress in their learning.

This section briefly reviews both traditional error analysis approaches starting from their historical background before moving into a discussion of the shortcomings of each approach.

2.1.1 The Traditional Contrastive Analysis Approach

The Traditional Contrastive Analysis Approach (henceforth TCA) dates back to the 1950s and 1960s. This approach involves '… comparing systematically the language and culture to be learned with the native language and culture of the student' (Lado, 1958:vii). Proponents of TCA claimed that this type of comparison can help SL/FL teachers to predict the patterns that make SL/FL learning difficult and those that make SL/FL learning easy (Lado, 1957). The underlying motivation for this type of comparison is to help second language teachers to prepare the most efficient teaching materials (Fries, 1945). A basic assumption of this theory is that when a language learner wants to learn a particular SL/FL, he/she will find that some elements in the target language are easy to learn because they are similar to features that exist in his/her native language (NL). Conversely, he/she will find other features more difficult, either because they are very different from their equivalent forms in his/her NL, or because they do not exist in their NL at all (Fries, 1945). For instance, the fact that Chinese and Vietnamese learners of English experience problems in marking noun plurals would be attributed in TCA to the fact that Chinese and Vietnamese rarely mark plurals in their mother languages (Romaine, 2003).

The intuitive plausibility of TCA led to its rapid rise in popularity during the 1960s, and a number of major studies were published in this decade. For instance, Stockwell et al. (1965) constructed a pedagogically motivated hierarchy of difficulty for linguistic differences on the basis of a comparison of the grammatical systems of English and Spanish. Stockwell et al. (1965) concluded that 'the greater the linguistic difference between some aspect of the L1 and the L2, the greater the likelihood of interference.' (Stockwell et al , 1965, cited in Kellerman, 1995:126). Another early study by Hoang (1965) conducted a contrastive analysis of English and Vietnamese sound systems. The motivation for Hoang's study was to identify problematic areas for Vietnamese students in public high schools in Vietnam, when engaged in learning English pronunciation. The results revealed phonemic and structural differences between English and Vietnamese. Hoang claimed that the results should be used for English teaching materials design and could also be a basis for the comparative study of other languages.

By the early 1970s, TCA began to receive severe criticisms (Odlin, 1989). First of all, some researchers pointed out that cross-linguistic differences are by no means the only cause of language learning difficulties. For example, a large number of language learners' errors may simply be caused by bad teaching or irrelevant or inappropriate teaching materials (Lee, 1968).

Secondly, it was pointed out that cross-linguistic differences may not always lead to language learning difficulties in both directions. For example, the English verb *carry* corresponds to twenty-five verbs in the Tzeltal language (Lee, 1968). Whereas it might well be difficult for English native speakers to learn these twenty-five Tzeltal verbs and use them accurately in different contexts, it seems unlikely that Tzeltal native speakers would find it equally difficult to learn the one English verb *carry* and use it in different contexts (Lee, 1968).

Thirdly, critics pointed out that TCA was not able to predict or explain language learning difficulties that might not be due to the mother language interference (Odlin, 1989). For instance, despite the fact that Spanish and English both make similar use of copular verb forms, comparative studies of Spanish and English learners observed the omission of *be* forms in the speech of Spanish-native speaker English learners (Butterworth, 1978; Peck, 1978; Schumann, 1978; Shapira, 1978).

Fourthly, depending on the degree of similarity and difference between the two languages, TCA only proved in practice to be capable of predicting up to 30 per cent of errors that language learners are likely to produce due to transferring mismatching patterns from the learner's mother language into the target language (TL) (James, 1998). This meant that as much as 70 per cent of learner errors might not be identified by TCA.

Fifthly, critics have also argued that many of the predicted difficulties in a learner's TL are not evident in the interlanguage of the language learners themselves (Markham, 1985). Furthermore, language teachers whose mother language background matches that of their students may be aware of possible difficulties as a result of first language interference. The teachers' awareness of these possible difficulties may be raised from being native speakers of the same mother language as their students and their own journey as language learners before they became language teachers.

Sixthly, perhaps one of the most serious criticisms of TCA relates to ideas and evidence emerging in the 1970s about the acquisition order of certain grammatical features in the target language. As discussed above, the basic notion of TCA is based on the observation of similarities and differences between the learner's first and second languages, and the assumption that similar features will be easier for learners to acquire, and differences more difficult. Thus, TCA assumes that learners from different mother language backgrounds would not be expected to acquire those grammatical features in the same order.

However, in a series of studies focusing on the order of acquisition of eight grammatical morphemes (plural '*-s',* progressive *-ing,* copula *be,* auxiliary *be,* articles, irregular past tense, third person *-s* and possessive *-s*),Dulay & Burt (1973; 1974a; 1974b; 1975) found the acquisition order of these features to be similar irrespective of the learners' first language (e.g., English, Spanish or Chinese), their age, or the medium of the collected data (spoken or written). Dulay & Burt's findings were in line with Brown (1973), who also found that English speaking children learn grammatical morphemes in the same order.

Due to these criticisms, TCA fell out of favour and was gradually replaced by another approach, which will be reviewed in the next section. It should be noted here, however, that the decline of TCA did not entirely spell the death of the notion of CA, as we will see later in Section 2.2.2.

2.1.2 The Traditional Error Analysis Approach

As mentioned above, contrastive approaches to the analysis of learner errors fell out of favour in the 1970s. At the same time, a new paradigm emerged; in the discussion that follows, we will refer to this paradigm as The *Traditional Error Analysis* (TEA) approach. In this approach, the focus shifted from the comparative analysis of native and target languages, to the making of comparisons between the learners' interlanguage and the equivalent patterns in the target language (James, 1998).

The term 'interlanguage' was coined by Selinker (1972), and refers to the language (spoken or written) produced by the language learner during his/her journey of SL/FL learning. TEA was first suggested in earlier work by Corder (1967), and consists of five procedures: (1) The collection of data from language learners; (2) the identification of errors in the collected data; (3) the classification of these errors into different types; (4) providing explanations for these errors; and (5) evaluation of errors (Corder, 1971; Corder, 1974; Corder, 1981; Lennon, 1991; Ellis, 1994; Ellis, 2008). Methodological issues and considerations relating to each of these procedures will be discussed in more detail below.

With respect to the collection of data, it was recognised from the outset that the type of data collected for error analysis could have a significant impact on the results obtained from an error analysis (Ellis, 2008). LoCoco (1976), for example, compared three methods of data collection (Free Composition, Translation and Picture Description) for error analysis. The error analysis was aimed to verify the percentages of errors in grammatical and source categories. The analysis revealed that whereas most of the free composition and picture description results were similar, there was clear variation between some error categories in free composition and picture description on the one hand and translation on the other hand (LoCoco, 1976).

Regarding the second and third stages of the TEA procedure described above, identifying and classifying errors crucially depends on what counts as an 'error' in the first place (Lennon, 1991). On this point, it is important to note that the concept of 'error' in TEA is broader than just one of recognising deviations from formal correctness; crucially, it also takes contextual appropriacy into account. That is, it recognises that structures that seem to be correct in a certain context or situation might not be so in another context or situation (Corder, 1971). Ellis (2008:48), for example, points out that the sentence '*I want to read your newspaper*' is always grammatically correct, but could be pragmatically unacceptable if it is addressed to a stranger. This type of sentence is called '*covertly idiosyncratic'.* On the other hand, a sentence that is 'ill-formed' or grammatically incorrect is known in TEA as *'overtly idiosyncratic'* (Corder, 1981). For Corder (1974), therefore, identification of errors 'depends crucially upon the analyst making a correct interpretation of the learner's intended meaning in the context.' (p127). The fourth stage, explanation of errors, aims to analyse language learners' errors from a psycholinguistic perspective in order to provide an answer to the questions *how did the language learners produce these errors?* and *Why did they produce these errors?* (Corder, 1981). Finally, TEA examines the effect of language learners' errors on the people who perceive these errors (Ellis, 2008). That is, it considers how well the addressed people could comprehend these *'idiosyncratic* sentences' and/or how effectively they would be able react in response to these errors (Ellis, 2008).

A frequently cited example of a study that applied the TEA methodology described above is Robert et al. (1973). In their study, Robert et al. analysed the language errors produced by Mexican-American pupils in a bilingual school and a monolingual school. To perform this task, Robert et al. collected and analysed the speech samples from the pupils from both schools and classified the language errors into three categories: morphology, syntax and vocabulary. The analysis revealed that the source of these errors was: a) Spanish interference in English learning (interlingual errors), b) incorrect applications of standard English grammatical rules (intralingual/developmental errors), and c) errors due to the influence of non-standard English dialect forms.

In another study, Richards (1974) applied TEA methodology to analyse language errors collected from speakers of different first language backgrounds (Japanese, Chinese, Burmese, French, Czech, Polish, Tagalog, Maori, Maltese, and a number of major Indian and West African languages). This study took a different approach in that it only focused on the intralingual and developmental errors which cannot be predicted by TCA and ignored any language transfer errors. The intralingual errors were found to reflect over-generalization, incomplete application of the target language rules and failure to learn when the target language rules need to be applied, while the developmental errors were argued to be due to false system knowledge or assumptions that the learners had built up from their limited experiences gained from their classroom experience (Richards, 1974).

Although TEA proved to be a highly popular and influential approach to the study of learner errors, it was not without its problems. From a broad educational perspective, TEA has been criticised for what some scholars see as its generally negative orientation. As its name suggests, TEA focuses on language learners' *errors***,** and ignores what language learners are able to do correctly (Hammarberg, 1974).

As Dagneaux et al. (1998) suggest, the other main problems with TEA can be divided into two kinds: methodological problems, and problems of scope (cf. Schachter & Celce-Murcia, 1977).

The methodological problems are, firstly, represented in the heterogeneity of the compiled data. There are no clear design criteria for data collection, which may lead the researcher to obtain unsystematic or even potentially untrustworthy results. The second methodological problem is characterised in the ambiguity of error classification. That is to say, the *covertly idiosyncratic* VS *overtly idiosyncratic* classification system is very broad and underspecified, and in practice is thus too dependent on individual researchers' interpretations.

TEA's scope problems are exemplified in, firstly, its orientation towards errors only and ignoring other features that mark the learners' performance. The other problem relates to the static picture of learner interlanguage development that tends to result from TEA analyses. Most TEA studies are cross-sectional rather than longitudinal (Ellis, 1994). This makes it difficult to spot different errors that language learners produce as they progress through different proficiency levels.

2.2 The Corpus Revolution in Research on Learner Language

The limitations of both traditional error analysis approaches (contrastive analysis and error analysis) did not spell the death of the notion of comparative analysis, the core method of traditional error analysis approaches. On the contrary, analysis approaches have in recent decades come back into fashion as a result of the development of learner corpus research methods. Indeed, the emergence of learner corpus research as a new paradigm has effectively reinvented contrastive analysis and error analysis as Contrastive Interlanguage Analysis (CIA) (Granger, 1998) and Computer-aided Error Analysis (CEA) (Dagneaux et al., 1998) respectively. This section reviews the corpus-based 'revolution' in studies of learner language and shows how these new methods have given fresh impetus to both error analysis approaches.

2.2.1 Learner Corpus Research Studies

Learner corpus research studies are computer-based research methods that emerged in the late 1980s. The compiled data for learner corpus research studies is in the form of '…. electronic collections of natural or near-natural data produced by foreign or second language (L2) learners and assembled according to explicit design criteria' (Granger et al., 2015:1).

The distinctive features of learner corpus research studies are that they make use of computers and computer software programs for comparative analysis of large amounts of language learner data, i.e., up to millions of words/tokens. This offers SLA researchers a number of important advantages over traditional methods.

Firstly, the fact that we now live in a digital age means that it is now possible to compile and store very large sets of authentic data from a great number of language learners, thus making such data far more representative than was possible in the past (Granger et al., 2015). For instance, the 'Learner corpora around the world' webpage[1], which is maintained by the University of Louvain, lists more than 180 learner corpora in electronic format, representing language learners from a wide range of different mother language backgrounds. While many of these corpora focus on learners of English, a growing number of learner corpora for other target languages are now becoming available, including Spanish, Chinese, German, Arabic, Russian, Portuguese and Italian (Centre for English Corpus Linguistics, 2019).

Secondly, because the compiled data is available in electronic format, computers and computer software programs are utilized to perform learner corpus research analysis. This has massively reduced the time that would have been consumed in traditional manual analysis processes. Once the electronic format learner data is fed into the computers, performing large quantities of error analysis and retrieving different types of linguistic features via corpus linguistic tools automatically is possible within a short time period. For instance, fully automatic part-of-speech taggers (POS) such as CLAWS, (the Constituent Likelihood Automatic Word-tagging System) (Rayson & Garside, 1998), can be used to annotate learner corpora with codes that specify the grammatical class of each word in a learner corpus. Once annotated, it is possible to retrieve

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

and compare the frequencies of different word classes (e.g., preposition, adjective, adverb … etc.) and even carry out more syntactically-oriented searches in these corpora. It is also possible to compare the frequency of these linguistic features in one or more learner corpora with the frequency of the same linguistic features in a standard English native speakers' corpus (e.g., the British National Corpus) or a corpus of native English student writing (e.g., the LOCNESS corpus).

The comparison between learner corpora from different mother language backgrounds helps to identify (1) which linguistic features are shared between learners from different mother language backgrounds which subsequently may indicate developmental error source, and (2) which linguistic features are specific to a group of learners from the same mother language background, which may in turn provide indirect evidence of first language influence (Granger, 2002). The comparison between the learner corpus and the native speakers' corpus helps to uncover which linguistic features the learners use more 'overuse' or less 'underuse' than the native speakers do (De Cock et al., 1998).

Thirdly, adopting a computational approach also ensures that the results that come out of these types of learner corpus research studies are highly accurate. Whereas human analysts inevitably make coding errors, machines are entirely consistent in any classifying or counting task that they are asked to do. Above all, corpus methods provide researchers with new results which would be impossible to obtain by traditional manual methods, and which may also challenge or even refute previous research results obtained through traditional error analysis approaches, as we will see in further detail in Section 2.2.3 below.

Undoubtedly learner corpus research studies have opened new pathways for researchers of SLA. It is now possible to conduct empirical research, in the field of SLA, on massive amounts of authentic learner data and perform a series of large-scale contrastive analyses within a much shorter time period than has ever been possible before. It is these distinctive features that have rehabilitated the traditional error analysis approaches and brought them to their new versions *Contrastive Interlanguage Analysis* and *Computer-aided Error Analysis.* The next sections will review these two related but distinct traditions of analysis in more detail.

2.2.2 Contrastive Interlanguage Analysis

The criticisms that the traditional contrastive analysis (TCA) has received, as mentioned in Section 2.1.1, showed the need for a new approach to contrastive analysis, one that is concerned not so much with comparing the grammatical systems of two languages in general, but rather with "comparing/contrasting what non-native and native speakers of a language do in a comparable situation" (Pery-Woodley, 1990:143). Much of the initial impetus for developing such an approach using corpus methods came from the work of Sylviane Granger and her colleagues at the Centre for English Corpus Linguistics, Université Catholique de Louvain, as part of the International Corpus of Learner English (ICLE) project. On the basis of this work, Granger (1996) introduced a new version of the contrastive analysis approach known as Contrastive Interlanguage Analysis (CIA).

The distinctive feature of CIA is that it is by definition a corpus-based approach, i.e., it is conducted under the umbrella of learner corpus research. Computers and computer software programs are utilized to perform contrastive interlanguage analysis on compiled learner corpora in order to uncover patterns of overuse and underuse in learner interlanguage (Granger et al., 2007). In broad terms, there are two types of comparisons. The first type of comparison aims to contrast a comparable corpus of the native language (NL) that the language learners are seeking to learn, with the interlanguage (IL) of those language learners themselves. This type of comparison is known as NL/IL comparison (Granger, 1996) (it is also sometimes known as NS/NNS comparison (Granger, 2002) and L2 vs. L1 comparison (Granger, 2012)).

In addition to identifying the language learners' errors, NL/IL comparison helps to identify the underuse and overuse of linguistic features, ranging from structural features (e.g., grammatical morphemes) to lexical items (e.g., collocations).

The second type of comparison is when the interlanguages of language learners from two or more mother languages are contrasted with each other. This is known variously as IL/IL comparison (Granger, 1996), NNS/NNS comparison (Granger, 2002) or L2 vs L2 comparison (Granger, 2012). This type of comparison helps to identify 'language transfer errors' (cf. Benson, 2002; Odlin, 2005), i.e., the linguistic features that are specific to a group of language learners from the same mother language. For instance, Granger (2004) observed that French learners of English tend to overuse the word 'indeed' far more than do English learners from other L1 backgrounds. Her explanation for this observation is that French learners of English make 'a faulty assumption that there is a one-to-one equivalence between *indeed* and *en effet*' (Granger, 2004:135). Some studies have combined both of these approaches. For instance, Ringbom (1998) made use of seven western European learner corpora (771,278 total tokens) extracted from the *International Corpus of Learner English* (ICLE) and compared the frequency of the top 100 words in the seven learner corpora with the *Louvain Corpus of Native English Essays* (LOCNESS), thereby incorporating both IL/NS and IL/IL comparisons.

As an alternative approach to identifying language transfer errors, Granger (1996) suggested integrating contrastive analysis in its old version (TCA) and the new version of contrastive analysis (CIA) in such a way that TCA is used to predict the language transfer errors that may be expected due to the differences between the two languages, and CIA is used, empirically, to confirm or refute these predictions.

Since its emergence in 1996, CIA has been increasingly widely used to study learner interlanguage and has yielded new insights into lexical and grammatical acquisition phenomena. While most of the research conducted so far has focused on learners of English as a second or foreign language, there have also been studies focusing on learners of other languages. For instance, Granger (2003a) designed three-tiered error annotation system to annotate the *French Interlanguage Database (FRIDA)* corpus. After annotating 300,000 words portion of the FRIDA corpus, text retrieval software was used to extract detailed error statistics. In another study, Zinsmeister and Breckle (2012) investigated discourse-related phenomena produced by Chinese learners of German.

Despite its popularity as a new research method, the first iteration of Contrastive Interlanguage Analysis has not escaped criticism, particularly in its first type of comparison, i.e., NL/IL. The main criticism, as summarized by Granger (2015), is that the approach could be accused of falling prey to the "comparative fallacy" and the issue of what precisely is meant by 'the target language'. Each of these criticisms will now be considered in turn.

The term 'comparative fallacy' was coined by Bley-Vroman (1983), who argued that '…work on the linguistic description of learners languages can be seriously hindered or side-tracked by a concern with the target language' (Bley-Vroman, 1983:2). Since the aim of the first type of comparison (NL/IL), as discussed earlier, is to contrast a comparable corpus of the native language that the language learners are aiming to learn, with the interlanguage of those language learners, opponents of CIA consider taking the native language as a reference makes it '… difficult to avoid seeing the learner's IL as anything but deficient' (Larsen-Freeman, 2014:217) when it is continuously contrasted with the ‘perfect’ production of the native language speaker (Larsen-Freeman, 2014). For Selinker (2014), this is problematic as interlanguages ‘… are never perfect when measured in terms of the target language…’ (p223). Therefore, analysts should not compare the interlanguages either with the target language or with the language learners’ first languages. Instead, argues Selinker, they should be studied in their own right.

The second criticism identified above relates to the practice of comparing learner language against a ‘target language’ benchmark. The problem here is that this benchmark may not be as clearly defined as it might seem on the surface. For example, many if not most NL/IL comparisons focus on learners of English, but which variety of English (i.e., British, American, Canadian, Australian, etc.) is assumed to be the target in each case? This is often not specified by the researcher (Granger, 2015). In this respect, CIA has ignored the fact that, for an internationally widespread language such as English, the validity of the native vs. non-native dichotomy becomes questionable (Brutt-Griffler & Samimy, 2001).

Other critics have objected to some of CIA’s key concepts. Aston (2011:11), for example, criticises ‘the employment of the damning terms ‘overuse’ and ‘underuse’ of particular features in comparison with some reference group’ (Aston, 2011:11). In response to this, Granger (2015) argues that the two terms ‘overuse’ and ‘underuse’ have been misinterpreted, and maintains that they are merely neutral and descriptive, referring only to frequency of usage and not value judgements on the part of the researcher.

In response to these criticisms, Granger (2015) accepted that CIA needs to be updated, and announced the launch of a new version of Contrastive Interlanguage Analysis, called (CIA^2), which implements modifications to the approach intended to address the aforementioned criticisms.

The notion of CIA^2 is based on the concept of 'varieties' which can be classified into two types: 'Reference Language Varieties' (RLV) and 'Interlanguage Varieties' (ILV). The term Reference Language Varieties indicates the possibility of employing more than one reference language corpus to conduct NL/IL contrastive interlanguage analysis. For instance, to investigate which words may cause problems to Chinese university students when they write their dissertations, Lee and Chen (2009) compiled the Chinese Academic Written English (CAWE) corpus, which consists of dissertations written by Chinese undergraduate learners of English, and made use of two reference language corpora. The first reference corpus is an expert reference corpus known as the Expert Journal Articles (EXJA) corpus, which consists of academic articles collected from high standard linguistics journals. The second reference corpus is a native speaker reference corpus which is a sub-corpus of the British Academic Written English corpus (BAWE).

The concept of Interlanguage Varieties highlights the importance of taking into account the learner variables and task variables that potentially affect the learners' interlanguage (Granger, 2015). In addition to the mother language of the language learners, the learner variables could also include such factors as the learners' age, gender, any other languages spoken by the learner, and the duration of stay in a target language speaking country (Möller, 2013). The task variables include the medium of the corpus (spoken vs written), the genre (e.g., argumentative vs narrative written essays), the field of the corpus (general vs specialized corpora), the length of the task (e.g., the minimum number of words) and the time limit to perform the task (Granger, 2003b; Granger, 2004).

With respect to overuse and underuse, Granger (2015) suggested replacing these two terms with 'overrepresentation and underrepresentation', which she believes may convey less negative connotations.

2.2.3 Computer-aided Error Analysis

As mentioned earlier, the arrival of computational corpus-based methods has also served to revitalise traditional error analysis (TEA) as a general approach. In the era of learner corpus research, a new version of the error analysis approach known as Computer-aided Error Analysis (CEA) has come to prominence. Again, the main initial impetus for this approach was provided by the Centre for English Corpus Linguistics, Université Catholique de Louvain, and was originally formulated in a study where the research team analysed a 150,000-word English written corpus produced by speakers of French (Dagneaux et al., 1998).

The CEA approach consists of two main stages. In the first stage, the compiled learner corpus is corrected manually by a native speaker of the target language. Following the correction of the learner corpus, an error tagging code is assigned to each error by an editing software tool. At Louvain the research team has developed an MS Windows error editor (Dagneaux et al., 1998). This error tagging system is hierarchical and consists of nine major category codes: Formal, Grammatical, LeXico-grammatical, Lexical, Register, Word redundant/word missing/word order and Style. Each major code is followed by one or more sub-codes (Dagneaux et al., 1998).

In the second stage, errors are analysed by retrieving the tagged errors via text retrieval software tools such as AntConc. The analysis of these identified errors can then be performed by calculating the frequency of specific error type(s) and/or demonstrating these errors in Concordances to view the contexts where the errors have been committed.

As with CIA research, much of the research conducted thus far in the field of computer-aided error analysis has focused on learners of English, and has covered a wide range of learners from different mother language backgrounds, such as: Spanish (López, 2009), Indonesian (Merizawati, 2019), Chinese (Darus & Ching, 2009) and Arabic (Mohammed & Abdalhussein, 2015).

An example of a major study in this new tradition is Granger's (1999) analysis of verb tenses. In this study, Granger (1999) retrieved and analysed all verb tense errors from a corpus of English-language texts written by French-speaking student learners at both post-intermediate and advanced levels. The analysis was aimed to identify the potential difficulties of using tenses, and subsequently help ELT materials designers to produce more effective teaching materials. The analysis revealed that students make increasingly slow progress as they move into higher levels. It also showed that the students always use tenses accurately in sentences that contain time adverbials (e.g., *for, since, yet* for present perfect and *ago* for past simple) (Granger, 1999). Granger concluded that the reason behind these findings is that the French students are taught tenses at sentence level only. On the basis of these conclusions, Granger called for the teaching of English tenses to be carried out at discourse level as well.

2.2.4 The Impact of Learner Corpus Research

Earlier in this section it was claimed that learner corpus research has changed our understanding of the SL/FL learning process and helped us to obtain new results that may refute previous error analysis research findings. The following example is a case study that substantiates and illustrates this claim.

As discussed in Section 2.1.1 above, one of the major problems of TCA is its central notion that the similarities and differences between the learner's first and second languages will correlate with features that learners will find easy or difficult to acquire. If this is correct, it should be the case that the acquisition order of target language elements (e.g., grammatical morphemes) would be influenced by L1 and thus different for learners from different L1 backgrounds. As we saw, however, researchers such as Dulay and Burt (1973, 1974a, 1974b, 1975) found that learners seemed to acquire basic grammatical morphemes in the SL in the same order irrespective of their first language backgrounds.

This research went more or less unchallenged for over two decades, until Tono (2000) used corpus methods to replicate Dulay and Burt (1975). In his study, Tono investigated the acquisition order of the eight grammatical morphemes: plural *'-s',* progressive *-ing,* copula *be,* auxiliary *be,* articles, irregular past tense, third person *-s* and possessive *-s* among Japanese speakers learners of English as a FL. Tono used a Japanese EFL corpus to identify the frequency of both correct and incorrect grammatical morphemes produced by the Japanese EFL learners and compared the results with those of Dulay & Burt and other researchers who claimed that English as SL/FL learners acquire grammatical morphemes in the same order regardless of their first languages.

To perform this task, Tono used a subsection of the Japanese EFL learner Corpus (JEFLL) (c. 300,000 words) and automatically annotated the corpus with POS tags using the CLAWS tagger (Rayson & Garside, 1998). He then retrieved and calculated the morphemes that were accurately and inaccurately used. The results showed that the accuracy order of the grammatical morphemes is different from that observed in Dulay & Burt's research findings. For instance, whereas Dulay & Burt found that the students in their research had least difficulty with progressive *-ing,* plural -s and copula *be* and most difficulty with the third person *-s* and genitive, Tono's research findings, on the other hand, showed that the Japanese learners of

English had least difficulty with copula *be*, auxiliary *be,* possessive -*s* and progressive -*ing* where the accuracy rate in these four morphemes reached 90%. Tono's research findings also showed that the Japanese learners of English had most difficulty with the definite and indefinite articles *the/a.* Tono attributed the difficulty in acquiring the articles by the Japanese learners of English to their absence from the Japanese language.

2.3 The Limitation of Computerised Error Analysis Approaches

The main problem with computerised error analysis approaches is that they tend to focus on language learners' errors and ignore what learners can do correctly. In this respect, it could be argued that corpus-based error analysis has not yet overcome one of the main limitations of traditional error analysis as identified in Section 2.1.2 earlier, that is, its strongly negative orientation (Hammarberg, 1974).

However, it is worth mentioning that there have been attempts to rectify this problem by proposing an alternative approach that looks not only at the language learners' errors but also at what the language learners have said or written correctly. Most notably among these, '*potential occasion analysis'* (Thewissen, 2012:2) is claimed to be a promising approach which is based on '… counting the errors of a particular type out of the number of times it could **potentially** have been committed.' (Thewissen, 2012:140). Thus, for example, the percentage of noun phrase grammatical errors is calculated out of the total number of noun phrases in the learner corpus (Thewissen, 2012; 2015), including those which are free from error.

To understand to what extent the results obtained by applying the potential occasion analysis may differ from traditional error counting, consider this example from a pilot study by Thewissen (2005, cited in Thewissen, 2012). In this study, Thewissen calculated the percentage of modal auxiliary errors at three levels of English proficiency (B2, C1, C2). The calculation involved applying the two error counting approaches viz. traditional error counting and potential occasion analysis approaches. The results of this analysis showed that C1 learners made fewer errors than B2 learners when the traditional error counting approach was applied. One the other hand, the calculation based on the potential occasion analysis showed that there is no difference between learners across the three levels. This would seem to indicate that learners did not make progress in acquiring modal auxiliaries.

There is a major conceptual problem with Thewissen's approach, however. To understand this, let us consider another example from Thewissen (2012; 2015). In her study, Thewissen calculated the percentage of spelling errors out of the total tokens assuming each token has the potential for spelling error. But this assumption raises an important question: does each word have the *same* potential for spelling error? That is, are language learners equally likely to misspell three-letter words such as *big* or *far* as they are to misspell five-letter words such as *right* or *based*, let alone even longer words?

In short, Thewissen's approach ignores the fact that any compiled corpus consists of words of different lengths starting from one-letter words, such as: the indefinite article *a* and the first-person singular pronoun *I.* The corpus may also contain non-alphabetic words, such as cardinal numbers (4, 7, 10) and dates (1991, 2001, 2010, etc.) which may also be added to the total word count/tokens. Thus, it is impossible to misspell non-alphabetic words and the one letter-words. In addition, it is not expected that even elementary learners will misspell two-letter words (e.g., *on, at, an, etc.*) since they are mostly very high frequency grammatical words which are commonly used by all language learners at many different all proficiency levels. Furthermore, it needs to be acknowledged that advanced learners simply do not make spelling errors or, at least, the same spelling errors as beginner learners.

Therefore, this thesis seeks to propose an alternative approach that looks at both the language learners' errors and what language learners produced correctly, and to apply this approach in an empirical investigation of a Libyan EFL corpus (LEFL). I will aim to show how this new approach outperforms not only the traditional approach to error counting, but also Thewissen's potential occasion analysis approach. I will do this by running all three analyses on my corpus and evaluating the results of each analysis in comparison to the others. This comparative evaluative analysis will be presented in Chapter 5.

Conclusion

The aim of this chapter has been to provide a critical review of the main approaches to the analysis of learner errors that have been developed in the field of applied linguistics over the past half century. It was argued that both contrastive analysis and error analysis enjoyed brief periods of popularity before coming under sustained criticism and falling out of favour as a result of serious problems of methodology and scope.

It was then argued that the advent of corpus-based research methods in the late 1980s effectively rehabilitated and revitalized both of these traditional error analysis approaches, and brought them together under the general umbrella of *learner corpus research*, where they have been redefined as Contrastive Interlanguage Analysis (CIA) and Computer-aided Error Analysis (CEA) respectively. Researchers now utilize computers to perform CIA and analysis of language learners' errors in new ways and obtain new results that were not possible before the era of learner corpus research. Equally importantly, the 'corpus revolution' in error analysis has rehabilitated the concept of first language influence, or language transfer as it is now more frequently termed (Benson, 2002; Odlin, 2005), by making it possible to identify features that are, or seem to be, significantly or even exclusively associated with students from a single L1 background, and traceable to particular features in the student's L1.

Finally, it was suggested that Computer-aided Error Analysis research approaches still share with their traditional predecessor a tendency to focus on language learners' errors and to ignore what language learners can produce correctly. Therefore, as a second aim, this thesis seeks to propose a new approach to error analysis that offers to remedy this problem. This approach will be described in detail in Chapter 5.

Before moving on to discuss the methodology that has been that was adapted in this thesis, the next chapter will provide a literature review of spelling as an important language skill, and a survey of previous spelling error research. I believe this is a necessary step since, generally, spelling has relatively neglected in previous research on language learners' errors. The following chapter will also present the two classification systems of spelling errors that will be used in the analysis of spelling errors identified in the LEFLL sub-corpora, in Chapter 6.

CHAPTER 3: SPELLING ERROR ANALYSIS: A REVIEW OF THE LITERATURE

Introduction

On the face of it, there would seem to be no need to make a case for spelling as an important focus for learner corpus research. Spelling is very obviously a fundamentally important language skill that spans both writing and reading skills (Zhao et al., 2016), and spelling errors constitute a significant proportion of the errors found in any corpus of learner writing. However, despite its obvious importance as a language skill and its pervasiveness in learner data, spelling has received surprisingly little research attention in the learner corpus research literature thus far. It is therefore one of the broad aims of this thesis to help to redress this relative neglect.

In order to provide a context for the research, this chapter aims to provide an overview of spelling as a language skill, and a review of previous spelling error research. The chapter will first discuss the importance of spelling as a skill, and then provide an overview of spelling error research studies highlighting the limitations of previous work. Finally, the chapter will present the two general classifications of spelling errors that are the most widely used in spelling error research, and which will both be deployed in my own analysis: the *surface structure* and *linguistic category* taxonomies.

3.1 The Importance of Spelling Skills

Spelling is an important language skill that strongly influences and forms an integral part of both writing and reading skills. Spelling is an essential '…building block of writing development' (Lindner et al., 2020:1), and can have a significant effect on students' academic writing performance (Westwood, 2004). According to Stewart and Cegelka (1995, cited in Westwood, 2004), bad spelling reduces the intelligibility of a student's writing and can give the impression that the student is either careless or less intelligent than his/her colleagues.

As for reading skills development, it has been shown that reading ability improves as the students improve their spelling skills (Lindner et al., 2020). For example, Fender (2008) used a spelling task and standardized reading and listening tests to investigate the relationship between spelling knowledge and reading skills among a group of 16 intermediate-level Arab learners of English as a second language (ESL) and a comparable group of 21 intermediate-level non-Arab ESL learners (L1 Chinese, L1 Korean, and L1 Japanese) in an English for academic purposes (EAP) program. The spelling task was used to assess spelling knowledge and the standardized reading and listening tests were used to assess the general language processing and comprehension skills. Fender reported that no significant differences were observed in listening comprehension whereas in the spelling and reading comprehension tests, the Arab students group scored lower than the non-Arab students' group. Fender attributed these results to the strong relationship between spelling and reading skills, which does not exist between spelling and listening skills. In short, spelling is known to be an important aspect of knowledge in word recognition in reading development (Berninger et al., 2002a; Mehta et al., 2005; Chiappe et al., 2007; Fender, 2008).

It is equally uncontroversial to note that spelling poses a great challenge to SL/FL learners. In a number of research studies that analysed errors produced by English as SL/FL learners, it was reported that spelling errors constituted important figures among other types of language learners' errors. For instance, in an investigation of the most common types of errors in the English writing of 22 Palestinian students at Al-Istiqlal University, Qaddumi and Walweel (2018) reported that the most common types of errors the students committed were spelling errors and most students were not able to spell correctly. In another study that was aimed to analyse the most common types of errors produced by Saudi female students EFL learners at Majmaa'h University – Saudi Arabia, Khatter (2019) observed that spelling errors were the second most frequent types of errors among the six types of errors (punctuation, spelling, preposition, article, wrong verb tense and wrong word form errors) produced by the learners. Demirel (2017) observed that spelling errors represented the third most frequent types of errors produced by Turkish student EFL learners.

The importance and the wide spread of spelling errors in the writing of English as FL/SL has attracted the interest of English as FL/SL researchers where a number of research studies have focussed on spelling errors committed by English learners (e.g., Al-Jarf, 2005; Pacton et al., 2013; Alhaisoni et al., 2015; Alsaawi, 2015; Hameed, 2016; Al-Oudat, 2017; Al-Sobhi et al., 2017; Altamimi et al., 2018; Othman, 2018; Fitria, 2020). Unfortunately, most studies of spelling errors have been carried out on a small number of English learners and based on a small amount of data. For instance, Fitria (2020) examined the kinds of spelling errors in the writing composition of 24 Indonesian students whereas Alsaawi (2015) analysed spelling errors produced by 26 Saudi students' English learners where a spelling test of 25 words had been conducted.

Unfortunately, in the era of learner corpus research where it became possible to deal with huge amount of data collected from large numbers of language learners, a limited number of research studies have utilized the corpus research tools to analyse spelling errors produced by the large numbers of language learners. A notable example of corpus-based spelling error research is Bestgen and Granger (2011). In their study, Bestgen and Granger (2011) analysed spelling errors in 223 argumentative essays extracted from the International Corpus of Learner English (ICLE). The argumentative essays were taken from French (FR), German (GE) and Spanish

(SP) components and the purpose of the study was to achieve two main objectives. 1) To assess whether spelling errors could help to predict the learners' proficiency automatically. 2) To examine whether subcategorizing these spelling errors could improve the prediction of both manual and automatic detection. The study showed that both manually and automatically detected spelling errors are good predictors of the quality of the L2 texts and the prediction scores improve by subcategorizing the spelling errors.

Although the study has successfully resolved the problem with the previous spelling errors research (as they were based on limited numbers of students and small amount of data) by utilizing corpus tools and resources, it is unfortunate that this study did not include sub-corpora of L2 learners from different language competency levels. As is well known, the essays in ICLE were written by intermediate to advanced English as FL learners (Granger & Bestgen, 2014). A comparison between English learners at different language competency levels may make it possible, for instance, to assess whether both the manual and automatically detected spelling errors at lower levels can also predict the quality of L2 texts.

In another corpus-based spelling errors research study, Randall and Groom (2009) analysed the spelling errors produced by 16-year-old Arab English learners in the BALC corpus (the BUiD (*British University in Dubai*) Arab Learner Corpus). The corpus is a collaborative research project conducted by researchers at the British University in Dubai, the United Arab Emirates, and the University of Birmingham, UK. The corpus consists of 1,865 texts (287,227 word-tokens and 20,275 word-types) written by third year secondary school students (the last year of schooling) and first year university students. The texts were collected from three different sources: by MED students in secondary schools, retired first Year University test essays, and texts sourced from the Common Education Proficiency Assessment (CEPA) examinations.

Randall & Groom reported that the Arab student English learners encountered a major problem with vowels, a problem that they referred to as 'vowel blindness' (Randall & Groom, 2009:8). In contrast, learners were observed to have much fewer difficulties with consonants as fewer consonant spelling errors were observed. They attributed this reduction in consonant spelling errors to the fact that Arabic, as a 'consonantal script' (Randall & Groom; 2009:8) directs Arab students to focus on consonants rather than vowels. Among the different types of vowels spelling subcategory errors, Randall & Groom reported that orthographic spelling subcategory errors, particularly vowel diagraphs, constituted the most significant error type. A word such as *friend(s),* that contains the diagraph *'ie'* and initial and final consonant clusters, constituted a problem to Arab student English learners where it has been misspelled in different ways (e.g., *frend(s), fraind(s), freand(s), firend, etc.*).

Although this study did analyse spelling errors committed by learners at different levels (unlike Bestgen & Granger, 2011) by comparing third year secondary school students and first year university students, unfortunately, it is not possible to make inferences from Randall & Groom's analysis about the development of spelling errors from one level to another, as the two groups belong to two very different study environments (a secondary school and a university) and the texts were collected from three different sources as pointed out above.

More recently, Yildiz (2017) has identified and analysed misspelled words in the BALC corpus that exhibit consonant clusters at the phonological level. The analysis revealed that the two-member onset clusters (misspelled words that exhibit two consonants at the beginning) were the most frequent cluster errors (44%) followed by the two-member coda clusters (misspelled words that exhibit two consonants at the end) (35%) and the three-member coda clusters were the third most frequent cluster errors (17%) whereas the three-member onset clusters were the lowest (1%). The analysis also revealed *epenthesis,* the insertion of a vowel or consonant either at the beginning of the word or between two sounds (e.g., *'balack'* for *black*) as the most common repair strategy (49%) adapted by the Arab student English learners at both onset and coda positions, followed by the omission strategy (19%) mainly occurring at coda position. Yildiz attributed the high percentage of the two-member coda cluster errors (35%) to the influence of the learners' first language. Arabic syllable structures permit two-member coda clusters only, and Yildiz (2017) suggests that this expectation has been transferred to the learners' L2. She also ascribed the high frequency of epenthesis repair strategy to first language influence as it was commonly observed among Arab student English learners.

Unfortunately, and as conceded by Yildiz (2017), her study has only looked at a specific group of English learners at specific language competency level, meaning that there is at the current time still no overall survey in the research literature of spelling errors committed by Arabic L1 learners of English at points in their learning. This thesis attempts to address this directly, by offering a quasi-longitudinal analysis of spelling errors in a large corpus of written English collected from Arab student English learners at three different stages, at the English department, Benghazi University.

3.2 The Classification of Spelling Errors

A review of the analysis of spelling error research studies revealed that the classification of spelling errors is based on one or a combination of two of the following classification systems: the surface structure and the linguistic category taxonomies. The following two sections are dedicated to discuss both classification systems of spelling errors in more detail, as this thesis will use both classification systems in the spelling error analysis reported below. The purpose of using both classification systems of spelling errors in this thesis stems from the need to employ the surface structure taxonomy to interpret the linguistic categories of spelling identified in the LEFLL corpus, as will be seen in Chapter 6.

3.2.1 The Surface Structure Taxonomy

Classifying spelling errors based on their surface structure provides a deliberately 'minimal' approach to analysis, in that it shows the 'non-linguistic' (Cook, 1997:479) alterations that affect misspelled words. The classification system is simple and widely used in spelling error research studies (e.g., Alhaisoni et al., 2015; Al-Oudat, 2017; Al-Sobhi et al., 2017; Othman, 2018; Omar, 2019; Fitria, 2020). In this type of classification system, spelling errors are classified according to the following non-linguistic alterations: 1) *omission* of a single letter; such as the omission of *'i'* in '*universty*'*;* 2) *addition* of a single letter, such as the addition of *'r'* in *'morrning';* 3) *substitution* of a single letter by another, such as the substitution of *'u'* by *'e'* in '*Saterday';* and 4) *transposition* of two adjacent letters, such as the transposition of *'i'* and *'r'* in *'frist'.*

The surface structure taxonomy is also used in a combination with spelling error classification systems based on more detailed and complex sets of linguistic categories, as the non-linguistic alterations (*omission, addition, substitution* and *transposition*) can assist in the interpretation of these more fine-grained categories. It is to a consideration of these linguistic categories that we now turn.

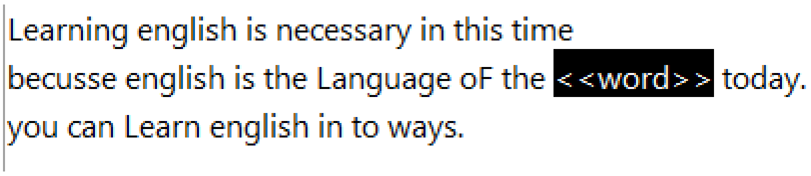
3.2.2 Classifying Spelling Errors According to Linguistic Categories

Whereas the surface structure taxonomy provides the minimal information of spelling errors by describing non-linguistic alterations (*omission, addition, substitution* and *transposition*) affecting the misspelled words produced by language learners, error classification taxonomies based on more formal linguistic categories, on the other hand, aims to provide maximal information about spelling errors, such as information about the causes of spelling errors and the possible influence of the phonological, orthographical and morphological systems of the learners' first language in learning how to spell in the second language. A typical example of this approach is Lindner et al. (2020), who study of spelling errors committed by Spanish-speaking English Language learners. Their analysis revealed that both the Spanish and English orthographies have influenced their spelling errors. The Spanish orthography is shallow, transparent and highly consistent, with few consonant-based inconsistencies, whereas the English orthography was found to be deep and opaque with many inconsistencies among both vowels and consonants. As a result, the proportions of both vowel-based and consonant-based spelling errors were almost equal (Lindner et al., 2020). Linder et al. attributed the high frequency of vowel-based spelling errors to the differences in vowel letter-sound correspondences between English and Spanish, whereas the high frequency of consonant-based spelling errors was attributed to the influence of Spanish orthography, as inconsistencies among letter-sound correspondences in Spanish were also present in consonants.

Depending on the researcher's interest, the classification of spelling errors based on linguistic categories taxonomies can vary substantially from one study to another. For instance, Martin (2017) classified the spelling errors of high-functioning dyslexics and typical students into: *phonologically simple, orthographically simple, phonologically complex* and *orthographically complex* whereas Protopapas et al. (2013) classified the spelling errors of Greek children with and without dyslexia into: *phonological* (*graphophonemic mapping*), *grammatical* (*inflectional suffixes*), *orthographic* (*word stems*), *stress assignment* (*diacritic*) and *punctuation*. Perhaps the most comprehensive classification system of spelling errors that may best suit the aims of the current thesis, as will be seen in Chapter 6, is the classification system of spelling errors proposed by Rimrott & Heift (2005; 2008). This classification system consists of the four following dimensions: a) a *linguistic subsystem* taxonomy; b) a *linguistic competence* taxonomy; c) a *language influence* taxonomy; and d) a *target modification* taxonomy. The following four subsections will discuss each of these four dimensions in more detail, supported by examples, where necessary, extracted from the LEFLL corpus.

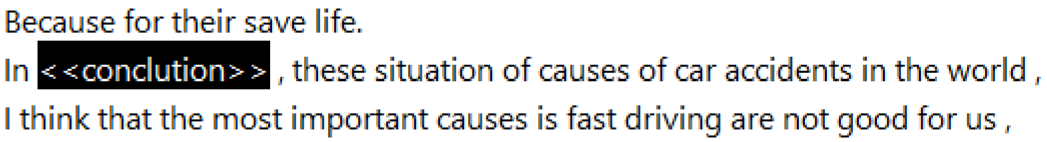
3.2.2.1 Linguistic Subsystem Taxonomy

In the linguistic subsystem taxonomy, spelling errors are divided into three subcategories: phonological, morphological, and orthographical spelling errors. The *phonological* spelling error occurs when the misspelled word is altered phonologically so that the misspelled word is pronounced differently from the intended word (Protopapas et al., 2013). In the following example, the learner misspelled the intended word *'world'* into *'word'*:



Learning english is necessary in this time
becusse english is the Language oF the <<word>> today.
you can Learn english in to ways.

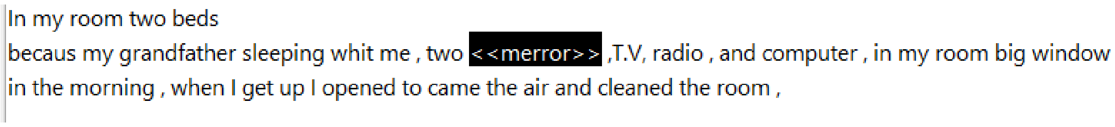
Morphological spelling errors affect word formation because of a problem with inflection or derivation (Protopapas et al. 2013). For instance, in the following example:



Because for their save life.
In <<conclution>> , these situation of causes of car accidents in the world ,
I think that the most important causes is fast driving are not good for us ,

This learner seems to have derived the word ‘*conclusion’* incorrectly from the original word ‘*conclude’*. As a result, he/she misspelled the target word as ‘*conclution’*.

Orthographical spelling errors occur when the learner alters the intended word but maintains the correct pronunciation (Protopapas et al. 2013). In the following example:



In my room two beds
becaus my grandfather sleeping whit me , two <<merror>> ,T.V, radio , and computer , in my room big window
in the morning , when I get up I opened to came the air and cleaned the room ,

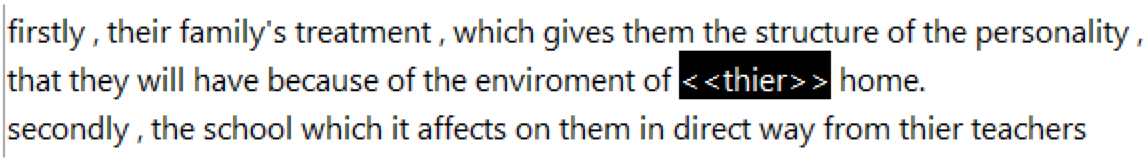
The learner has misspelled the intended word ‘*mirror’* by substituting the letter ‘*i*’ with ‘*e*’. However, the misspelled word may also be pronounced as /ˈmɪrə(r)/.

3.2.2.2 Linguistic Competence Taxonomy

In the linguistic competence taxonomy, Rimrott & Heift (2005; 2008) distinguish between performance (incidental) errors, which are produced by both native and non-native speakers and competence errors, which are expected to be produced by non-native speakers only.

a. **Performance Errors**

Both native and non-native speakers produce these types of spelling errors. They are incidental spelling errors and caused, for example, due to mistyping that may occur because of speed typing, or lack of concentration while the typing process proceeds. They do not reflect any limitation in the linguistic background of the writer, and it is assumed that the writer can correct these errors whenever he/she reviews his/her writing without any external support. Consider the following example:



firstly , their family's treatment , which gives them the structure of the personality , that they will have because of the enviroment of <<thier>> home.
secondly , the school which it affects on them in direct way from thier teachers

The language learner, in this example, has successfully spelled the word *'their'* in the first attempt and failed in the second and third attempts. Therefore, this type of spelling error could be marked as a performance spelling error. The alteration of the word form in performance spelling errors may only represent one of these four types of spelling errors: *omission, addition, substitution, and transposition* of the word letter(s) (see Section 3.2.1 above for more information about *omission, addition, substitution* and *transposition* spelling errors).

b. Competence Spelling Errors

Competence spelling errors mark any spelling errors that do not fall under the performance spelling error category. These types of spelling errors reflect problematic linguistic features which are either attributed to the role of the learner's first language, or produced as a result of incomplete knowledge of the target language spelling rules. The former are referred to as interlingual spelling errors, and the latter are referred to as intralingual spelling errors. These two error types will be further discussed below.

3.2.2.3 Language Influence Taxonomy

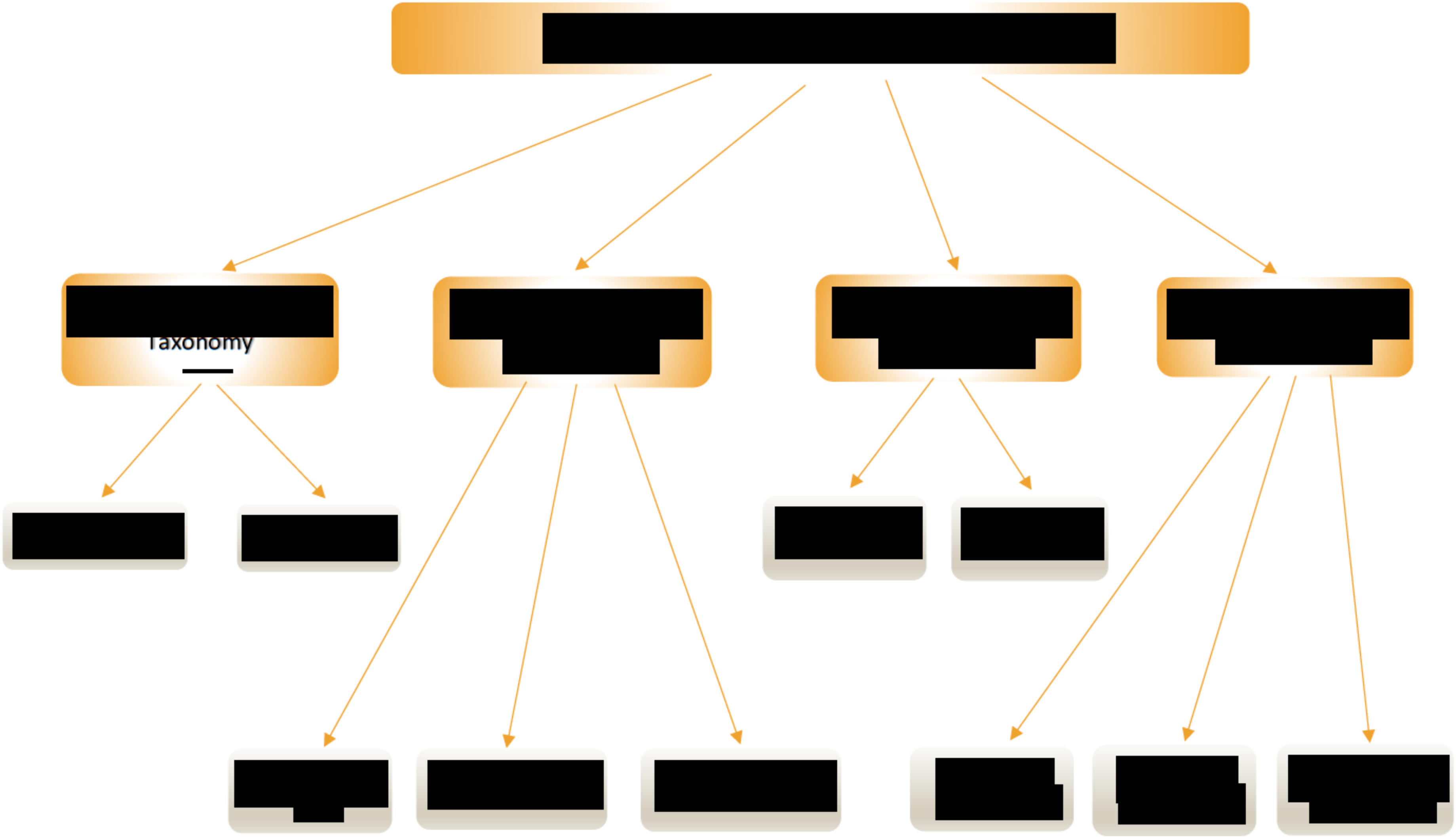
The language influence taxonomy also distinguishes between interlingual and intralingual spelling errors. Interlingual spelling errors reflect the role of the learner’s first language as the main source of competence spelling errors. These errors are assumed to occur when the SL/FL learner transfers patterns and processing routines from L1 to L2 (Rimrott & Heift, 2005; Saigh & Schmitt, 2012), with different rates depending on the level of the language learner (Weber et al., 2013). Intralingual spelling errors, on the other hand, reflect the language learner’s level of competence in the target language itself (Richards, 1974) and the strategies that the language learner uses to learn the language (Gafu et al., 2012).

3.2.2.4 Target Modification Taxonomy

The target modification taxonomy classifies the misspelled words into three categories: single, double and multiple edit misspellings (also known as one-edit distance, two-edit distance and multiple-edit distance) depending on the number of changes (additions, omissions, substitutions, and/or transpositions) that need to be applied to convert the misspelled word into the target one. The multiple edit misspellings may have three or more edit distances as shown in Table 3.1 below.

Edit distance	**Examples**
One-edit distance	I *prepar* (prepare) my books (addition of ‘*e’* is needed)
Two-edit distance	Restaurant or *caffee* (cafe) (omissions of ‘f’ and ‘e’ are needed)
Multiple-edit distance	Corss (course) (addition of ‘*u*’ and ‘*e*’ and omission of one ‘*s*’ are needed)

Table 3.1: Examples of spelling errors edit distance



[redacted] he linguistic spelling error categories developed by Rimrott & Heift (2005; 200[redacted] [redacted]

Conclusion

This chapter has provided an overview of the importance of spelling as an aspect of language learning and as a focus for learner corpus research. Spelling is an important but somewhat overlooked language skill that is crucial for enhancing both the writing and reading skills of second and foreign language learners. Spelling errors constitute a significant proportion of the total number of types of language learners' errors in any error analysis study. Unfortunately, most of the previous published spelling error research studies have been based only on limited numbers of learners and on small amounts of data. Even though we are now in the era of learner corpus research studies, where it became possible to analyse huge amounts of data collected from large numbers of English learners, there have at the present time been disappointingly few studies specifically focussing on spelling errors among learners of English. Furthermore, the few studies that have been carried out so far have been limited in that they have been purely cross-sectional, which makes it impossible to see how or whether spelling errors change from one level to another.

Thus, this thesis seeks to remedy these problems by analysing spelling errors identified in a large corpus of texts produced by Libyan university students learning English. In the remaining chapters of this thesis, we will report on and discuss the findings of this analysis in detail, starting, as mentioned above, at the level of spelling errors.

CHAPTER 4: METHODOLOGY

Introduction

This chapter describes and discusses the procedures that have been followed to compile and analyse the learner corpus I have designed, called ‘LEFLL’ (Libyan English as a Foreign Language Learners). The chapter starts by discussing how the LEFLL corpus was originally designed, and how and why this original design had to be adapted in the actual process of data collection and compilation. Following this is a section dedicated to discussing the annotation scheme. In this section, I will discuss the identification and classification of errors in the LEFLL corpus and the error tagging system that has been developed to annotate the LEFLL corpus. Next, I will present the corpus tools that have been used in the current research. These corpus tools are of two kinds: text annotation tools and text retrieval tools. Finally, the chapter concludes by reporting on interrater reliability tests that were carried out on my classification of the language learners' errors, in order to establish the validity and reliability of these error classifications.

4.1 The Learner Corpus Design

From the outset of the current research, the learner corpus design criteria proposed by Granger (2002) were taken into consideration. According to Granger, any corpus compilation project should address two sets of factors: language learners and task settings. With regard to the former, learner corpus researchers need to consider the mother language(s) of the language learners and whether the learners have previously learned other languages. The language learning context is also an important factor. That is, corpus designers need to consider whether the language learners are acquiring the target language as a second language (i.e., in a native

English-speaking country, e.g., in the UK or USA) or as a foreign language (i.e., in a non-native English-speaking country). Researchers also need to consider the age, gender, and the language learners’ competence level.

As regards task setting factors, researchers need to bear in mind the medium (spoken or written) of the data collection task and the genre of the task itself (i.e., descriptive writing, argumentative essays, etc.). Finally, researchers need to consider the task condition (e.g., whether the tasks were carried out under timed and invigilated exam conditions or at home, whether the students were allowed to consult dictionaries or grammars, whether a time limit was given to perform the tasks, etc.).

Based on these design criteria, compiling the learner corpus for the current thesis went through two data collection planning phases: a primary data collection plan and a modified data collection plan which developed during the data collection process. Each of these will now be discussed in turn.

4.1.1 The Primary Data Collection Plan

The primary data collection plan aimed to compile a representative quasi-longitudinal learner corpus of Libyan English learners. For the reader to be able to evaluate how well the thesis has achieved this aim, a brief outline of the Libyan education system, and of English language education in particular, is provided here. First, however, it should be noted that the current system is somewhat different to what is described here, as the whole education system has been subject to various and ongoing changes as a result of the civil war that has been ongoing in Libya for most of the period in which the current research was carried out. During the period of data collection for the current thesis, the education system in Libya consisted of four stages:

Primary, Preparatory, Secondary and *University.* The primary stage/school was from Year 1 to Year 6. In primary stage, children join schools when they are 6 years old.

Following the primary school, students started preparatory schools at Year 7 and finished at Year 9. They then started secondary school, where they chose specific subjects (engineering, medicine, business, etc.) and studied these for a further four years. Following secondary school, most students study undergraduate courses at Libyan universities, which were (and still are) usually for four years. English is a compulsory subject for all students from Year 5 until the end of secondary school. Therefore, all Libyan children have (in principle) studied English for at least nine years before going on to university or the workplace.

With respect to the mother language(s) of the language learners to be represented in my corpus, the picture is a complex one. The official language in Libya is Arabic. In addition to Arabic, however, there are also other languages spoken in Libya, the most prominent of which are Berber, Tedaga, Italian, and French (Oishimaya, 2017). Berber is an umbrella term covering several North African languages that are spoken by 15 to 20 million people (Kossmann & Stroomer, 1997). In Libya, there are a few Berber languages spoken in different parts of Libya. These include Awjilah, Tamazight, Nafusi, and Ghadames languages (Oishimaya, 2017). Tedaga is a Nilo-Saharan language spoken by a large population in the south of Libya known as the Tebu. It is also spoken in parts of northern Chad and eastern Niger (Oishimaya, 2017).

As for the Latin languages Italian and French, the Libyan dialect is rich with Italian vocabulary, and many people, especially elderly people who witnessed the invasion of Libya by Italy in the 20th century, speak Italian. French is a popular language among young people in Libya (Oishimaya, 2017). Many language centres in Libya also teach Italian and French. Thus, the participants in the primary data collection plan included both monolinguals (Arabic native speakers) and bi/multilinguals.

Initially, the primary data collection plan sought to obtain data from three universities and three language centres distributed across three cities in Libya: Benghazi, Jalu, and Kufra. The two universities in Jalu and Kufra are branches of Benghazi University and teach the same English modules and curriculums across academic years. Some academic lecturers from Benghazi University also teach at Jalu and Kufra. The three language centres were to be private EFL schools in the same cities. Based on initial contacts with the language centres, it was established that most of the language learners to be sampled from this group would be adults. The reason for aiming to collect data from both universities and language centres was to ensure that the participants came from a representative sample of different language learning environments, both academic and general.

After contacting these universities and language centres, however, it soon became obvious that the primary data collection plan was overly ambitious. Firstly, it did not consider the instability that dominated almost the whole country during that period of data collection. Some parts of Libya (e.g., Kufra) were even going through civil war, with the result that some educational institutions were closed during the period of data collection and would in any case have been inaccessible to the researcher in practice. Secondly, some institutions did not show any real willingness to contribute to this research and were either unwilling or unable to provide enough data for the current thesis.

The curriculums at the language centres chosen for the study were very different, leading to a concern on my part that these differences may distort the outcomes of the error analysis. Each language centre followed a different curriculum and used a different classification of language competence levels. For instance, the language competence levels in one language centre (e.g., the Modern School of Languages in Benghazi) consist of ten levels and the curriculums are adopted from the Cambridge University system for English learners, while in another language centre (e.g., Kufra College for Languages in Kufra city), there are six levels, and the curriculums are adopted from the Oxford University system for English learners. Furthermore, the former teaching system tends to focus on vocabulary acquisition, while the latter teaching system tends to focus more strongly on teaching grammar.

4.1.2 The Modified Data Collection Plan

Due to the above obstacles, the data collection plan was modified and restricted to Benghazi University, the main university campus in Benghazi City. Since the learner corpus, for the current thesis, will only be designed to study errors produced by Libyan English learners' resident in Libya, I will call this learner corpus 'the *Libyan English as a Foreign Language Learners*' (LEFLL) corpus.

The target participants were deemed to represent, at least, Libyan learners of English who speak Arabic as a first language. Unfortunately, it was not possible to obtain information about the English proficiency levels of the students as mentioned in Section 4.1.1 above. We only know that the students have studied English for a minimum of nine years including four years of study at the secondary schools specialized in English subject. They were supposed to be undergraduate students in Year 1, Year 2, Year 3, and Year 4 at the English department, Faculty of Languages, Benghazi University.

In several language learner error studies (e.g., Tono et al. 2014; Xu 2014), data collection was performed by setting writing tasks that were designed to elicit certain features that were expected to be interesting to the researchers. As discussed earlier, to ensure that corpora are comparable, researchers need to consider the writing task factors described in Section 2.2.2 above. With this in mind, the initial data collection plan for the current thesis envisaged setting up a writing task that would be administered to the students in the study. In this writing task, students at each of the four years of undergraduate study (Year 1, Year 2, Year 3, and Year 4) at the English department, Benghazi University, were supposed to write a free composition about the same topic "The International Intervention in Libya during the Arab Spring". This topic was chosen as the data collection process was conducted a few months after what is known as "The Arab Spring" and the intervention of NATO forces in Libya in 2011 to remove the Kaddafi Regime. It was assumed that this topic would be of great interest to Libyans in general and the students at the Libyan schools and universities in particular as writing tasks. In order to facilitate this, the Head Office of the English department at Benghazi University was contacted to set up the writing task.

After contacting the English department, it became obvious that it was not possible to collect the data via setting up the writing task as it had been hoped. The English department, at Benghazi University, stated that the university was closed during the revolution in Libya between February 2011 and August of the same year. As a result of this extended period of closure, the university was busy trying to catch up on the curriculum content that had been missed during the revolution. Instead, they suggested collecting data from previous final exam papers that the university was intending to destroy. Under these circumstances, I decided to accept this offer and base my corpus on these exam papers.

It is worth mentioning that Year 4 exam papers were excluded from the data collection process, and thus from the research reported in this thesis. This is due to accessibility issues during the period of data collection. The Exam Office at Benghazi University imposed restrictions on Year 4 exam papers because they were from the latest final exams, and thus still 'live' documents. The exam scripts for the students in Year 1, Year 2, and Year 3, in contrast, were already one year old and were due to be destroyed. Although it was somewhat disappointing not to have access to data representing a fourth-year group, I would nevertheless argue that a three-level learner corpus remains more than adequate for analysing the development of learner errors from one level to another.

In terms of what kinds of learning tasks and outcomes are represented in LEFLL, it is important to note that the pedagogic focus of the writing classroom changes across the three levels in the English departments at the universities in Libya. In year 1, students at Benghazi University study basic English writing skills. This includes mastery of punctuation marks and an introduction to different types of sentences structure (simple, compound, complex and compound complex sentences). In year 2, the students start learning and practising paragraph structure (topic sentence, supporting sentences and conclusion) and how to write different types of paragraphs (descriptive, narrative, persuasive, explanatory and illustration). In year 3, the students start learning how to write entire essays which should include at least three paragraphs (introductory, the main body and conclusion paragraphs). Thus, not only do the students learn to do more difficult things in English writing as they move from one year to another, but they are also engaged in different kinds of writing tasks. That is, they are not doing the same kind of task in each year group. Clearly, this needs to be borne in mind when looking at the results obtained from the analysis that will be reported in later chapters.

The final exam essays, in the three levels (Year 1, Year 2, and Year 3), collected for the LEFLL corpus consisted of different topics as follows:

In Year 1 final exam essays, the students were asked to write an essay about one of the following five topics:

- A person you care about
- Introducing yourself
- Your morning routine
- Your hometown
- Your sleeping habits (see Appendix (A) for a writing task sample, Year 1).

In the Year 2 final exam essays, the students were asked to write an essay on one of the following three topics:

- How you celebrate a special day
- A special occasion in your culture
- Describe your room (see Appendix (B) for a writing task sample, Year 2).

For the Year 3 final exam essays, the students had an open writing task. The students were asked to choose their own topics and write a well-structured essay. As shown in Appendix (C), the exam essays for the academic year 3 needed to include the following five elements: 1) a title, 2) a thesis statement/topic sentence with clear subtopics/ controlling ideas, 3) at least two body paragraphs, 4) a concluding paragraph, and 5) Transitional signals that link paragraphs together, providing overall unity and coherence.

The advantage of compiling this type of data for the current thesis is that it is consistent data, insofar as all texts were produced in essentially the same task environment and within the same limited period. Furthermore, in timed exams where students' overall proficiency levels were being judged, it is reasonable to assume that at least most of the students were writing to the best of their ability, and their work can thus be regarded as a more accurate reflection of their competence level than would be the case if the data was collected from voluntary, optional writing tasks where the personal stakes for the students are very low. This can also be observed in the collected questionnaires attached to the thesis as only 260 copies of questionnaires out of 1000 distributed copies were returned and a large portion of returned copies of the questionnaires were incomplete or incorrectly completed, as we will discuss in detail in Section 4.1.2.2 below. Finally, it is worth noting that in official exams, the rules and regulations are very strict, and are enforced by invigilation. This may reduce or prevent plagiarism, and thus contribute to the overall validity of the collected data.

4.1.2.1 Corpus Size

As was pointed out above, the collected data were hand-written final exam essays. Since the handwriting of a large portion of these exam scripts was not clear enough to be scanned and converted automatically into electronic readable texts, manual data transcription was the only option. Inevitably, transcribing many written exam essays was very time-consuming. This was aggravated by the fact that all types of language learners' errors (e.g., spelling, grammatical, etc.) needed to be carefully preserved and accidental corrections avoided.

Once transcribed and converted into electronic readable texts, the LEFLL corpus was manually annotated for errors. Manual annotation is also a painstaking and time-consuming process, but with the current levels of technology it is unavoidable if the aim is – as it should be – to achieve a high level of coding accuracy. It has been reported that the success rate of automatic error detection of specific errors produced by second language learners is between 25% - 35% (Granger 2003a). Therefore, a large number of researchers in many learner corpus research studies preferred to use manual annotation for error analysis (e.g., Abe, 2003; Granger 2003a; Mariko, 2007; Thewissen, 2008; Nagata et al., 2011; Bestgen et al., 2012; Thewissen, 2013; Demirel, 2017).

It is worth mentioning that since the process of manual annotation is time-consuming, the corpora compiled for these types of learner corpus research studies tend to be relatively small, whereas larger corpora are more feasible for studies that are based on the Contrastive Interlanguage Approach (CIA) (see Section 2.2.2 above), where the analysis can be performed largely automatically. Other factors that may influence corpus size includes the level of language learners from whom the learner corpus has been collected; inevitably, language learners at lower language competency levels do not write the same length of exam essays as language learners at higher language competency levels in the same exam conditions.

The corpus size, for the current thesis, thus needs to balance a range of theoretical and practical considerations. On the one hand, it should be big enough to ensure that we can obtain all types of errors the language learners may produce. On the other hand, it needs to be small enough to make transcription and manual error tagging feasible within the timespan of a PhD research project.

The compiled learner corpus should also consider the fact (as mentioned above) that students at different ability levels tend to write essays of different average lengths. To address this, the process of compiling the LEFLL corpus went through 4 stages. In each stage, 5000 words of exam scripts from each academic level (Year 1, Year 2, and Year 3) were transcribed, converted into electronic readable texts and classified for errors. Thus, in the first stage, a total number of 15,000 words from the three academic levels exam essays were transcribed and errors were identified and classified. The error classification at this stage was aimed to provide general descriptions of error types as we will see in Section 4.2 below.

Similarly, in the second, third, and fourth stages, 5000 additional words were collected from the exam essays for each year group, and errors were identified and initially classified and so forth, in an iterative process. The aim of compiling the learner corpus in 4 stages, is to ensure that data for all types of errors produced by Libyan learners of English were identified. In practice, the initial error analysis and classification showed that at the word count of approximately 60,000 words (almost 20,000 words in each level), no more new errors were observed. An overview of the final corpus composition is presented in Table 4.1 below.

	No. of essays	Overall tokens	Avg. token/essay	Overall word types
Year 1 Sub-corpus	238	20045	80 – 90	1482
Year 2 Sub-corpus	219	20035	90 – 100	2412
Year 3 Sub-corpus	102	20051	190 – 200	2858
LEFLL Corpus	559	60,131	100 – 110	5031

Table 4.1: The total number of exam essays, word tokens and types in the LEFLL corpus

4.1.2.2 The Questionnaire

To ensure that the LEFLL corpus is sufficiently homogeneous and based on clear design criteria as proposed by Granger (1998), questionnaires were distributed to 1000 students across the three academic years (Year 1, Year 2, and Year 3) at the English department, Benghazi University. The students were asked to complete and submit the questionnaires to their tutors after completion. A copy of the questionnaire is provided in Appendix D.

The questionnaire was designed to ensure that the participants at the three levels would be able to complete them on their own and within a short time period. The questionnaire consisted of five main questions. In addition to age and gender, the questionnaire sought to investigate the features that characterized learners at Benghazi University. Question 3 explores whether the students speak other languages in addition to Arabic. The purpose of this question was to ensure that the target participants are native speakers of Arabic only and exclude those students who are also native speakers of other languages such as Berber, Tedaga, French, etc. (See Section 4.1.1 above for more information about these other languages). The underlying purpose of this was to help to eliminate any consideration of influence from these other languages, thereby making it more feasible to interpret some errors as being caused by Arabic L1 influence. Question 4 aimed to investigate the learning context and practical experience of the students. This included information about whether the students learned English as a foreign language (FL) (in a non-native English-speaking country, e.g., Libya) or a second language (SL) (in a native English-speaking country, e.g., UK) before they joined the university and the period they spent in learning English as (FL) or (SL). Question 5 aimed to verify the practical experience of the students outside the university while they were studying. That is, it aimed to find out whether the students focused on a specific language learning skill (e.g., reading, writing, speaking, etc.) on their own or at the language centres and the number of hours per week the students spent in practicing that specific language skill.

Unfortunately, only 260 out of 1000 distributed questionnaires were collected from the students. Furthermore, some of the collected questionnaires suffered from major problems. For instance, some mandatory questions in the questionnaire were not answered. As can be seen in the example provided in Appendix D, the student did not provide an answer to question

4. Subsequently, we do not know the learning context and practical experience of the student. That is whether the student studied English as (FL) or (SL) before he/she started the university and the number of years/months he/she spent in learning English either as (FL) or (SL).

Despite these limitations, 260 completed questionnaires nevertheless arguably provide a large enough pool of information to be taken as broadly representative of the participants as a whole. The results showed that out of the total number of 260 completed questionnaires, there were 214 (82.31%) females and 46 (17.69%) males. 82 out of 260 (31.54%) students were less than 20 years old, while 176 (67.69%) were between 20 and 25 years old, 1 student was between 26 and 30 years old and 1 student was between 31 and 35 years old. All the students had already spent a minimum of 13 years at school (i.e., 6 years at the primary schools, 3 years at the preparatory schools and 4 years at the secondary schools) before they joined the university. Given that they started going to primary schools when they were 6 years old, it is reasonable to assume that the average age of the learners represented in LEFLL will be between 19 and 25.

Regarding other languages spoken by the participants, 214 participants (82.31%) were monolinguals who only speak Arabic as a first language. 26 students speak basic French, 4 students speak basic Italian, and 2 students speak basic German, having studied these languages for just one term at the university as optional modules. According to the questionnaire, most participants had learned English as a foreign language (EFL) in Libya and had never visited or lived in an Anglophone country.

Overall, the questionnaire results indicated that the LEFLL corpus is homogenous and representative of Libyan students studying English at the Faculty of Languages, Benghazi University.

4.2 Annotation Scheme

The annotation of the LEFLL corpus went through two stages. In the first stage, the annotation process was mainly geared towards the specific task of helping to determine an optimum corpus size for the current project. As mentioned in Section 4.1.2.1 above, to determine the optimum corpus size for this thesis, the corpus was compiled in four stages. In each stage, only 5000 words of each sub-corpus were manually transcribed, converted by Dexter Converter tool (see Section 4.3.1.1 below for more information about the Dexter Converter tool) into electronic readable texts and errors were identified via the Dexter Coder tool (see Section 4.3.1.2 below for more information about Dexter Coder tool) and classified by providing the general description of each error type (as we will see later in Figure 4.1, Section 4.3.1.2). Providing the general descriptions of errors in the first stage assisted to develop an error tagging system that has been used in the second stage of annotation. Thus, the first stage has two aims: (1) to determine the optimum size of the corpus that best suits the research interest, (2) to provide general descriptions of all types of errors identified in the corpus, which could be used to develop an error tagging code system for the LEFLL corpus. In the second stage of annotation, errors that were identified in the first stage, were replaced with error tagging codes, developed by the researcher. Thus, in the first stage of annotation, each grammatical and spelling error in each exam essay file was identified and a general description of each error type was provided (e.g., spelling, missing determiner, missing apostrophe, etc.) as shown in Table 4.2 below. Following the identification and general description of errors in the first stage, grammatical errors were classified into phrase errors (noun, verb, prepositional, adjective and adverb phrase errors) and subcategory errors based on the general description and word class of each error as shown in Table 4.2 below. For instance, the general description of this type of extracted from the LEFLL corpus '*In this essay << am>> going to analyze both'* is 'missing pronoun' indicates that this error occurred in a noun phrase.

Phrase Type	Phrase Error Type	Examples from LEFLL corpus
Noun Phrase	A problem with gerund	I don't like <<make>> a noise. (for *making*)
	A problem with regular and irregular nouns	I wash my face and brush my <<teeths>>. (for *teeth*)
	Arabic vocabulary	there is a very big tv with 42 <<bosah>>. (for *inches*)
	Missing apostrophe	<<The wall colour>> in my room is bink (for *The wall's colour…)*
	Missing determiner	love is<< important feature>> in … (for *an important feature*)
	Missing noun	I'm <<ninteen >> old (for *nineteen years,*)
	Missing pronoun	In this essay << am>> going to analyze both (for *I am*)
	Number agreement	I am <<20 year>> old (for *20 years*)
	Redundant apostrophe	I am <<twenty year's>> old (for *twenty years*)
	Redundant determiner	There are <<a many things>> in my room (for *many things*)
	Redundant noun	children who grow up with both of <<two>> parents are not the same who
	Redundant pronoun	The special day for me <<it>> is when I was with my father,
	Wrong determiner	I want to talk about how <<much>> sister with me in my room. (for *many*)
	Wrong noun	… and drink <<café>> and milk (for *coffee*)
	Wrong pronoun	I get up at 8:30 A,M <<it>> wash my face and … (for *I*)
	Wrong word class	In <<summarize>> my bedroom is very … (for *summary*)

	Wrong word order	I'm Sara I <<old nintean years>> (for *nineteen years old*)
Verb Phrase	A problem with active and passive voice	Secondly, chocolate <<was using>> not only in eating (for *was being used*)
	A problem with finite and non-finite verbs	we study, play, and <<lestining>> (listening) to music (for *listen*)
	A problem with infinitive or gerund	they are necessary <<to treating>> them … (for *to treat*)
	A problem with phrasal verbs	the bride wears whight dress and <<puts>> the make up (for *puts on*)
	Confusion between regular and irregular verbs	… and I <<haved>> TV (for *had*)
	Missing auxiliary verb	After I graduate I <<>> going to work as a translator. (missing *am*)
	Missing main verb	My bedroom <<>> very big (missing *is*)
	Redundant apostrophe	I <<won't to write>> about myself
	Redundant auxiliary verb	… and we should <<do>>n't stop playing …
	Redundant main verb	Finally, I <<am>> and my brother we want to share the room.
	Subject verb agreement	<<she have>> a big one
	Tense problem	do things or see things you <<do not do>> it before
	Wrong auxiliary verb	chocolate <<is>> not just give us, energy
	Wrong main verb	After the guests arrived, we <<produce>> to them orange Juice with a sweets
	Wrong word class	I <<wan’t>> to know more about it.
	Wrong word order	Feeling <<not should be>> in focus (for *should not be*)
prepositional Phrase	Missing preposition	This process consist <<>> two types of translation (missing *of*)
	Redundant preposition	I would like to introduce <<in>>

		myself
	Wrong preposition	First, I go <<in>> the university and I lived from her (for *to*)
	Wrong word class	
	Wrong word order	My alarm rings <<about at>> 7:00. (for *at about*)
Adjective Phrase	Adjective phrase plurality error	… this is my <<faviorts>> [*favourites*] thingis [*things*] (for *favourite*).
	Arabic vocabulary	… the sheets are <<bage>> (for *pink*)
	Missing adjective	I am 18 years <<>> I am study in … (missing *old*)
	Redundant adjective	I'm <<old>> 19 year's old
	Redundant apostrophe	They are <<old's>> tewnety yar's
	Wrong adjective	… the other for my <<smal>> sister …
	Wrong word class	The <<sencerly>> friend will remains with us
	Wrong word order	so, my Room is <<cleaning Always>>
Adverb Phrase	Missing adverb	I go to class from 8:00 to 12:30 after << >>I go to cafiteria with my friends
	Missing apostrophe	I do <<nt>> like football, …
	Redundant adverb	Translation process become <<more>> easier
	Redundant apostrophe	I get up at 5:00 <<o'clo'ck>>
	Wrong adverb	… where I feel <<much>> comfortabal.
	Wrong word class	… should speak english very <<good>> to communicate with forgin people …
	Wrong word order	which, <<turn in>>, points you of the paragraph organizing of the essay

Table 4.2: The general descriptions of all types of phrase errors in the LEFLL corpus

So far, this section demonstrated the classification of errors based on the general description of each type of error. Errors were first classified into spelling and grammatical errors. Grammatical errors were then classified into phrase errors and subcategory errors. In the following section, errors were classified based on both the Linguistic Category and Surface Strategy Taxonomies proposed by Dulay et al. (1982).

4.2.1 The Classification of Errors in the LEFLL Corpus

To develop an error tagging system of errors, identified in the LEFLL corpus, and following the classification of errors into spelling errors and phrase errors and subcategory errors, phrase errors and subcategory errors were reviewed and classified based on the Linguistic Category and Surface Strategy Taxonomies proposed by Dulay et al. (1982). The linguistic category taxonomy was adapted to classify the phrase errors, in the current thesis, into further subcategories, whereas the surface structure taxonomy was used to provide information about the alterations that affected the phrase errors. Spelling errors, on the other hand, were classified based on the Surface Structure Taxonomy only. Later in chapter 6, there will be further analysis and discussion dedicated to spelling errors where linguistic categories will be applied to spelling errors. The reason why phrase errors and subcategory errors will be classified based on both the linguistic category and surface strategy taxonomies, whereas spelling errors will be classified based on the surface structure taxonomy only is that the aim of error classification, at this stage, is to develop an error tagging system as we will see later in this chapter.

According to Dulay et al., linguistic category taxonomies classify language learners' errors on two levels. The higher level is focuses on language component(s), (e.g., phonology, syntax and morphology (grammar), semantics and lexicon (meaning and vocabulary), and discourse (style)), while the lower level focuses on constituents which represent elements within the language components (e.g., a subordinate clause error within the syntax level).

Thus, in the current thesis, the higher levels of grammatical errors represent the five types of phrases where the language learner's errors have been committed, i.e., noun, verb, prepositional, adjective, and adverb phrases. The lower level characterises the particular elements that have incorrectly been used (e.g., a determiner in the level of the noun phrase, an auxiliary verb in the level of verb phrase, etc.).

The surface strategy taxonomy provides information about 'the ways surface structures are altered' (Dulay et al., 1982:150). For Dulay et al., language learners' errors alter surface structures in four ways: the omission of necessary items, the addition of unnecessary items; the misformation of items; and the misordering of items. The following four sub sections will discuss the four types of alterations in the surface structure taxonomy in more detail.

4.2.1.1 Omission

Omission errors refer to the absence of important items that are necessary to construct a correct grammatical and meaningful sentence (Dulay et al., 1982). The missing items are either content words or grammatical words/morphemes. In the sentence:

> *The man went to the library in the morning*

The words: *man, went, library*, and *morning* are content words because they carry the main meaning of the sentence. Words such as *the, to*, and *in*, are grammatical words because they function as grammatical items. Thus, all prepositions, such as: *on, at, with*, *etc*. *definite* and *indefinite articles*, *verb* and noun inflections, such as: *-ing* in *are looking* and *-es* in *boxes* are also grammatical items. According to Dulay et al., language learners are more likely to omit grammatical words and morphemes more than content words. The omission of content morphemes usually occurs in the early stages of L1 acquisition. If it occurs in L2 learning, it is usually due to the limitation of vocabulary content. As a sign of this problem, the learner may use gestures to replace the missing vocabulary (Dulay et al., 1982).

4.2.1.2 Addition

An *addition* error is the existence of unnecessary items in a well-formed utterance. According to Dulay et al., it occurs in later stages of L2 acquisition when the learner has learned some target language rules. It occurs due to 'the all-too-faithful use of certain rules' (Dulay et al., 1982:156). There are three types of addition errors: *double markings, regularizations,* and *simple addition*.

Errors of double markings characterise the failure of deletion of some linguistic markers which are necessary for some linguistic structures but not in others (Dulay et al., 1982). In the following example extracted from Dulay et al. (1982), a simple past tense form, the verb *went* in the affirmative declarative sentence:

> *They went to lunch an hour ago*

is the tense marker since it signals the simple past tense. In negation and interrogative, the auxiliary '*did'* is needed as a tense marker instead. The verb *went* must take the infinitive form since only one past tense marker is needed:

> *they did not go to lunch an hour ago. (negation)*

Learners tend to use both tense markers, such as:

> *They did not went to lunch an hour ago.*

A regularization error occurs when the learner applies the rule of regular form and construction (i.e., regular verb form) on the irregular forms and constructions in a language. Regular past tense verbs are formed by adding '*-ed*'. Learners may apply the same rule on the irregular verbs such as: *go* → *goed, put* → *putted*. The same case is in the regular and irregular plural forms.

Simple addition errors are those errors that can neither be classified as double markings nor regularizations (Overgeneralization) as the following example may show where the verb *'is'* was unnecessarily added

I'm is nineteen years old

4.2.1.3 Misformation

In *misformation* errors, the learner supplies the incorrect morpheme or structure. Dulay et al. (1982) reported three types of misformation errors: *Regularizations, Archi-forms*, and *Alternating* forms.

Regularization errors occur when an irregular marker is replaced with a regular one, such as: *runned* for *ran*, *gooses* for *geese*

Archi-forms errors are the selection of one particular linguistic element to represent the rest of elements in the same class, say the selection of *this* to represent *this, that, those*, and *these*. The selection of the infinitive complement to represent the other complements, such as: *gerunds* and *that clause*. In the auxiliary verbs, *does* for example may be used to represent the rest of auxiliaries: *is, am, are* and *do*.

Dulay et al. (1982) observed that *archi-forms* errors occur in a learning stage when a learner acquired the rule of that particular structure. Therefore, the *archi-form* error of nominative-accusative case is observed in the early stage of L2 learning because the learner has already learned the rule of a nominative-accusative case in that L2 learning stage while the *archi-form* error of the past irregulars and past participles are observed in later language learning stage because the learner acquires the rule of past irregulars and past participles in that stage.

Alternating forms errors occur when the learner alternates between two members of the same class (Dulay et al., 1982), such as: the alternation between demonstratives:

Those dog and this cats

In pronouns:

Masculine for feminine (or vice versa), as in: *he* for *she*

Plural for singular (*They* for *he*)

Accusative for nominative (or vice versa), as in *her* for *she*

In the tense case, alternation may occur between participle form and past simple, as in: *I seen her yesterday / He would have saw them*

4.2.1.4 Misordering

Misordering errors occur when the learner misplaces single morphemes or a group of morphemes in an utterance, e.g.

He is all the time late

The underlined morphemes are misordered.

Both L2 and L1 learners produce misordering errors in the constructions that they have acquired before. They produce declarative sentence order in a direct question, e.g.

What Daddy is doing?

In a later stage and after they acquire the simple question order, they produce the simple question order in indirect question, e.g.

I don't know what is that

L2 learners also produce misordering errors due to word-for-word translations of their mother language structures, such as:

I met there some Germans

Or

Another my friend

4.2.2 Error Tagging Code System

Following the classification of phrase errors and spelling errors as detailed above, an error tagging code system was developed to annotate the phrase and spelling errors. The discussion will now focus on the phrase error tagging code system. Later in this section, we will discuss the tagging system developed for the categorization of spelling errors.

The error tagging code system was adapted from the hierarchical error tagging system developed by Dagneaux et al. (1998). Dagneaux et al.'s error tagging code system consists of seven major error category codes: Formal (F), Grammatical (G), LeXico-grammatical (X), Lexical (L), Register (R), Word redundant/word missing/word order (W) and Style (S). These major error category codes are followed by one or more subcategory codes that are aimed to provide more information about the error types. For instance: the error tagging code GA consists of the major category code (G) for Grammar and one subcategory code (A) to refer to an article error in the level of grammar whereas the GADJG consists of one major category code (G) for grammar and two subcategory codes. (ADJ) refers to an adjective and (G) denotes gender. Thus, the (GADJG) error tagging code signifies a grammatical error where an adjective is affected, and this involves gender.

For the current thesis, the error tagging code system developed to annotate the phrase errors consists of five major category codes: (NP) for noun phrases, (VP) for verb phrases, (PP) for prepositional phrases, (ADJP) for adjective phrases, and (ADVP) for adverb phrases. The major error category codes are followed by one or two subcategory codes. For instance, the error category code (VPT) consists of the major category code (VP) for a verb phrase and the subcategory code (T) for tense. Thus, the (VPT) error code indicates the verb phrase error where the tense has incorrectly been used. The error code (NPDETOM) consists of the major category code (NP) for a noun phrase and two subcategory codes, (DET) for determiner and (OM) for omission error (see Table 4.3 and Table 4.4 below for the Two-level and Three-level phrase error tagging codes respectively).

Major Code	Subcategory code (1): *surface structure alteration*	Examples
Noun Phrases (NP)	NO (Number/plurality error)	but there are <<two main type of friend ship>>
	UN (Uncountable noun confusion error)	start to grill <<the meats>>
	GER (Gerund)	<<Celebrate a special day>> is my birth
	GEN (Gender)	while <<the braid cat his hear and shave his beard>>
	LTR (Literal Transfer)	there is a very big tv with 42 <<bosah>> (for inch)
	WWC (wrong Word Class)	I think health and <<educated>> can be build very strong nation
	WWO (Wrong Word Order)	The oFFice was in << Center City>>
Verb Phrases (VP)	PV (Phrasal Verb misformation)	I stopped to buy some plaster . when I got oFFice <<put it in>> my Feet
	APF (Active/Passive Voice missformation)	the teachers who studying (for teach at) secondary school <<are obtained>> the bachelor degree
	T (Tense misformation)	help them when they are older and <<became>> a parent
	FIN (Finite/addition)	we study, play, and <<lestining>> to music in it.

	GERINF (Gerund/Infinitive Misformation)	small family helps country <<to controlling>> the popullation of the growth.
	REG (Regularization)	I <<chosed>> simple style
	WWC (Wrong Word Class)	Finally we never <<improvement>> unless we have a modren equipment
	WWO (Wrong Word Order)	Feeling <<not should be>> in focuse
prepositional Phrases (PP)	OM (Omission)	in studying << >>university the number of the students
	RED (Redundant/Addition)	Chocolet was discovered <<from>> 400,000 years ago by mexican
	W (Wrong)	Many parents <<on>> my country have many children they
	WWC (Wrong Word Class)	after that I waching TV, <<next>> that I go to prae, Finaly
	WWO (Wrong Word Order)	My alarm rings <<about at>> 7:00.
Adjective Phrases (ADJP)	LTR (Literal Transfer)	I made sure that the sheets are <<bage>> and so as the pillows
	WWC (Wrong Word Class)	then, to make <<health>> life without disease.
	WWO (Wrong Word Order)	so, my Room is <<cleaning Always>>,
Adverb Phrases (ADVP)	WWC (Wrong Word Class)	It's <<huge>> big
	WWO (Wrong Word Order)	Directions words title tell you what the types of the essay you must write, which, <<turn in>>, points you of the paragraph

Table 4.3: The two-level phrase error tagging codes

Major Code	Subcategory code (1): linguistic constituent	Subcategory code (2): *surface structure alteration*	Examples
Noun Phrases (NP)	DET (Determiner)	OM (Omission)	if you want to work in <<company>>
		RED (Redundnat)	After that I go to <<the>> shopping
		W (Wrong)	Ead Alfeter is <<a day>> comes after Ramadhan's mounth
	PRON (Pronoun)	OM (Omission)	in my room << >>have one bed, one TV,
		RED (Redundant)	and searching in the computer about the

			thing I don't know <<it>>.
		W (Wrong)	I have alot of make-up in my room because I love <<her>> so much
	N (Slimane et al.)	OM (Omission)	saterday is <<a busy >>for me.
		RED (Redundant)	my room is big <<room>> and beatifull.
		W (Wrong)	I have <<a big library>>, you can find any book (for *bookcase*)
	APOST (Apostrophe)	OM (Omission)	After that I drive <<my son car>>
		RED (Redundant)	and I hope to learn many <<language's>>.
	NO (Number)	REG (Regularization/Misformation)	we need all of them in our <<lifes>>,
Verb Phrases (VP)	AUX (Auxiliaries)	OM (Omission)	After I graduate I <<>> going to work as a translator.
		RED (Redundnat)	I <<am>> study in university of Benghazi.
		W (Wrong)	I <<am not sleep>> at afternoon
	V (Main Verb)	OM (Omission)	my father << >>a doctor,
		RED (Redundant)	Then I clean the house and <<clean>> my bedroom.
		W (Wrong)	I <<do>> a shower before any thing
	VN (Subject-Verb)	(CON) Concord	For example , << child sometimes cry>> for more caring
	APOST (Apostrophe)	RED (Redundant)	I <<a'm>> from Liban,
Adjective Phrases (ADJP)	AJ (Adjective)	OM (Omission)	I am 18 years <<.>>
		RED (Redundant)	I am 19 <<old>> years old,
		W (Wrong)	I'm not married Because I'm very <<small>>

		PL (Plurality)	and this is my <<faviorts>> thingis,
	APOST (Apostrophe)	RED (Redundant)	it must be bulid in a <<scientifc's>> study
Adverb Phrases (ADVP)	AD (Adverb)	OM (Omission)	people can cause traffic accidents because don't know<<>>fast they are driving their car
		RED (Redundant)	Translation process become <<more>> easier.
		W (Wrong)	five minutes <<just>>, and this is not enough time for students …
	APOST (Apostrophe)	OM (Omission)	drinking drivers they do<<nt>> feel they are drunk
		RED (Redundant)	in libya it is very hot and <<some time's>> come dusty

Table 4.4: The three-level phrase error tagging codes

In addition to the error tagging codes mentioned above and as a prerequisite for the potential for error approach proposed by the current thesis (see Section 5.3 in the following chapter for more discussion about the potential for error counting approach), the subcategory code (POTE), which denotes potential, has been added to the major category codes to form the code that will be used to tag the constituent in a phrase that has the potential for error. Thus, the NPPOTE tagging code will be used to annotate the constituent in a noun phrase that has the potential for error. Similarly, the tagging codes VPPOTE, PPPOTE, ADJPPOTE, and ADVPPOTE will be used to annotate the constituents in a verb phrase, prepositional phrase, adjective phrase and adverb phrase, respectively, that have the potential for errors (see Appendix E for a full list of the tagging codes).

In addition to inserting the tagging codes, the annotation process also involved inserting corrections, where necessary, to add more information about the language learners' errors (see

Figure 4.3 in Section 4.3.1.2 for examples of error corrections in the Dexter Coder tool). This inevitably increased the time needed to annotate the LEFLL corpus, but it was considered a useful step as it shows the reason why a particular error has been tagged with a specific error tagging code.

Turning now to the spelling error tagging code system, the hierarchical error tagging code system developed by Dagneaux et al. (1998) was also adopted to annotate spelling errors. As shown in Table 4.5 below, the spelling error tagging code system used in the current thesis consists of two levels, one general and one specific. At the general level, a major error tagging code (SP) simply denotes a spelling error, and at the more specific level a subcategory code (following the surface structure taxonomy) represents the nature of the alteration that affects the misspelled word.

Major Code	Subcategory code (1): *surface structure alteration*	Examples
Spelling (SP)	IN (insertion/addition)	<<Firstlly>> The food
	OM (Omission)	Smocking also could <<caus>> the death
	SUB (Substitution)	she can easily <<convay>> the meanings
	TRA (Transposition)	iF someone <<decied>> to mak party

Table 4.5: The two-level spelling error tagging codes

Concerning the alterations that affected the misspelled words, four subcategory codes were used to describe these alterations: 'IN' for Insertion/addition of an unnecessary letter, 'OM' for Omission of a necessary letter, 'SUB' for Substitution of a correct letter with incorrect one, and 'TRA' for Transposition between two adjacent letters. For instance, the spelling error tagging code (SPOM) denotes a misspelled word where a necessary letter has been omitted from the intended word. Table 4.6 below shows the distribution of phrase and spelling error tagging codes and corrections. This excludes the first stage of annotations which involved inserting the general description of language learners' errors.

The Main Tagging Code Category	No. of tagging codes + Corrections
Spelling errors and corrections	14958
Noun Phrase Errors and corrections	62645
Verb Phrase Errors and corrections	28864
prepositional Phrase Errors and corrections	9294
Adjective Phrase Errors and corrections	3864
Adverb Phrase Errors and corrections	2886
The Total number of tagging codes and corrections	122511

Table 4.6: The distribution of tagging codes and corrections in the LEFLL corpus

So far, this chapter has focused on designing and annotating the LEFLL corpus, the following will present the corpus tools used to annotate the LEFLL corpus and retrieve the tagged features in the LEFLL corpus.

4.3 The Corpus Tools

To conduct a Computer-aided Error Analysis (CEA) (see Section 2.2.3 above) on the LEFLL corpus, two sets of tools were used, namely: Text Annotation Tools and Text Retrieval Tools. The following two main sections will discuss those two sets of corpus tools in more detail.

4.3.1 Text Annotation Tools

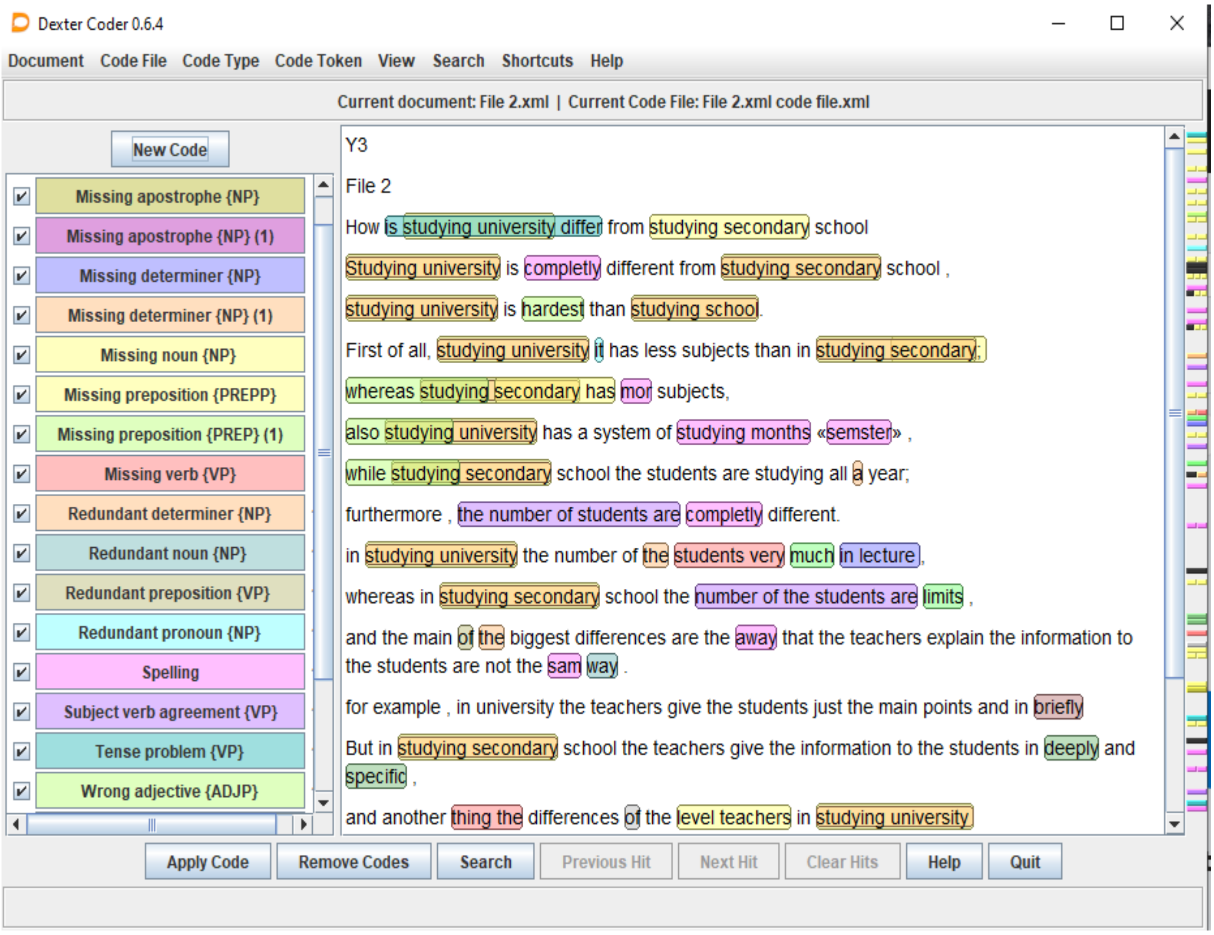
Annotation of the LEFLL corpus was performed via a corpus annotation software package called Dexter (Garretson, 2005; 2006). Dexter is a free corpus software program that can be used to annotate and analyse written and transcribed spoken texts. It consists of two sets of text annotation tools: Dexter Converter and Dexter Coder, and one text retrieval tool, known as Dexter Search. The Dexter Search tool will be discussed in more detail in Section 4.3.2. The following two sub-sections are dedicated to discussing the Dexter Converter and Dexter Coder tools.

4.3.1.1 Dexter Converter Tool

As a prerequisite for annotating the LEFLL corpus, it was necessary to convert the 559 manually transcribed plain text files (.txt), which form the corpus to (DeXML) files format, the format used by the Dexter Converter. A DeXML file format is a variant of the XML file format that can only be uploaded to and annotated by the Dexter Coder tool. To convert the files from .txt to DeXML, each transcribed exam essay file was carefully reviewed, compared with the original exam essay script, and cleared from any incidental mistakes that may have been made by the researcher during the transcription process. This was an essential step as it was not possible to correct or amend any mistakes after converting the files to the DeXML format.

4.3.1.2 Dexter Coder Tool

After converting the transcribed essay files from .txt to DeXML files format, each converted file was manually annotated using the Dexter Coder tool. As shown in Figure 4.1 below which illustrates the first stage of annotation (see Section 4.2 above), As the figure shows, an error tagging bubble was created for each type of error encountered as I manually searched through each text in the corpus and was marked with a different colour. Once each code has been added to the code list on the left of the application window, it can be re-used to annotate all subsequent instances in this and subsequent texts.


Dexter Coder 0.6.4
Document Code File Code Type Code Token View Search Shortcuts Help
New Code
☑ Missing apostrophe {NP}
☑ Missing apostrophe {NP} (1)
☑ Missing determiner {NP}
☑ Missing determiner {NP} (1)
☑ Missing noun {NP}
☑ Missing preposition {PREPP}
☑ Missing preposition {PREP} (1)
☑ Missing verb {VP}
☑ Redundant determiner {NP}
☑ Redundant noun {NP}
☑ Redundant preposition {VP}
☑ Redundant pronoun {NP}
☑ Spelling
☑ Subject verb agreement {VP}
☑ Tense problem {VP}
☑ Wrong adjective {ADJP}
Apply Code
Remove Codes
Search
Previous Hit
Next Hit
Clear Hits
Help
Quit

Current document: File 2.xml | Current Code File: File 2.xml code file.xml

Y3

File 2

How is studying university differ from studying secondary school

Studying university is completly different from studying secondary school ,

studying university is hardest than studying school.

First of all, studying university it has less subjects than in studying secondary;

whereas studying secondary has mor subjects,

also studying university has a system of studying months «semster» ,

while studying secondary school the students are studying all a year;

furthermore , the number of students are completly different.

in studying university the number of the students very much in lecture,

whereas in studying secondary school the number of the students are limits ,

and the main of the biggest differences are the away that the teachers explain the information to the students are not the sam way .

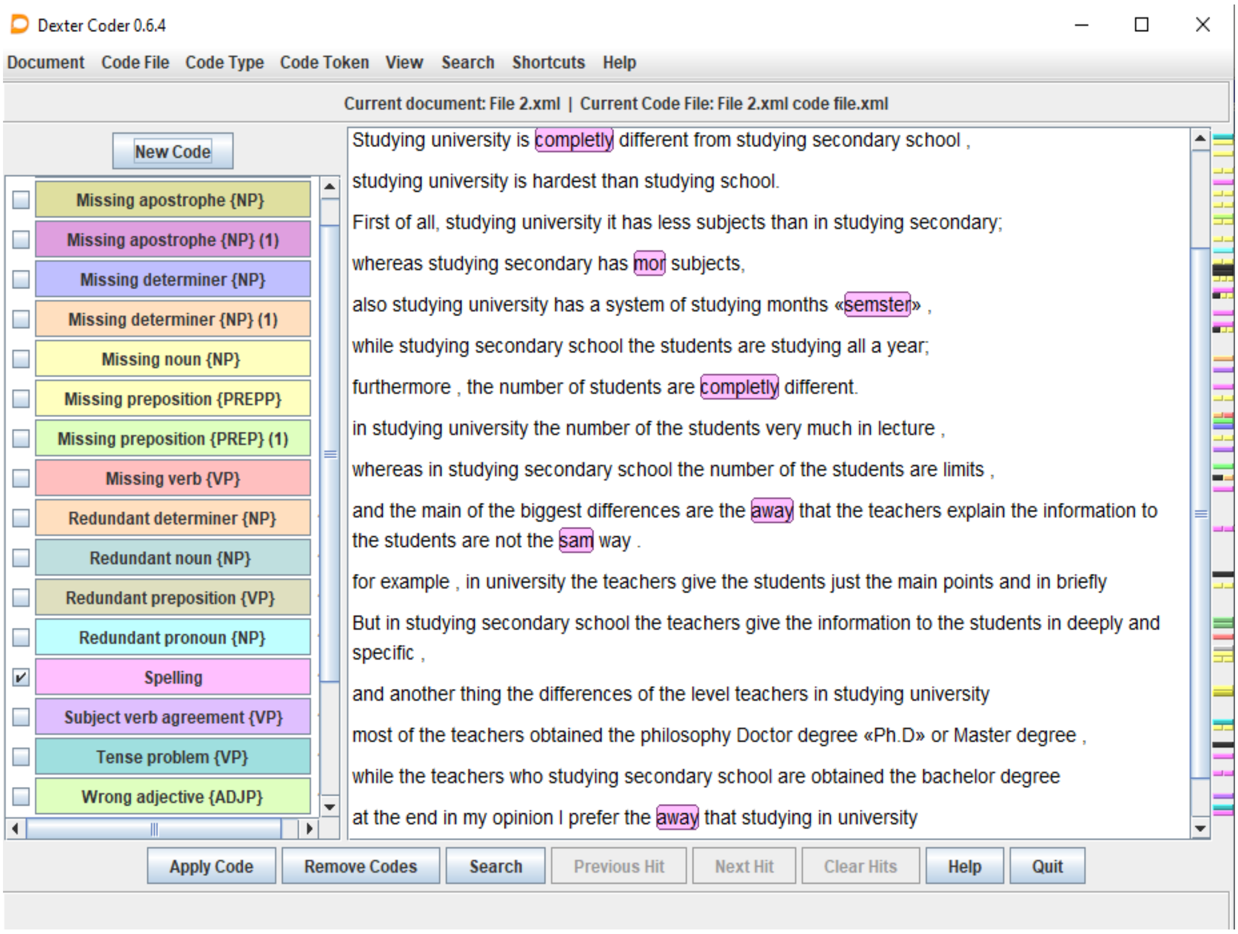
for example , in university the teachers give the students just the main points and in briefly

But in studying secondary school the teachers give the information to the students in deeply and specific ,

and another thing the differences of the level teachers in studying university

Figure 4.1: The description of the language learners' errors in Dexter Coder (first stage)

Once the coloured error tagging bubbles were created and the annotation system applied across the uploaded exam essay file, each error type, across the text, can then be viewed by selecting the error type from the coloured error tagging bubbles. Figure 4.2 below is given as an example showing spelling errors across an annotated exam essay file.


Dexter Coder 0.6.4
Document
Code File
Code Type
Code Token
View
Search
Shortcuts
Help
Current document: File 2.xml | Current Code File: File 2.xml code file.xml
New Code
□ Missing apostrophe {NP}
□ Missing apostrophe {NP} (1)
□ Missing determiner {NP}
□ Missing determiner {NP} (1)
□ Missing noun {NP}
□ Missing preposition {PREPP}
□ Missing preposition {PREP} (1)
□ Missing verb {VP}
□ Redundant determiner {NP}
□ Redundant noun {NP}
□ Redundant preposition {VP}
□ Redundant pronoun {NP}
☑ Spelling
□ Subject verb agreement {VP}
□ Tense problem {VP}
□ Wrong adjective {ADJP}
Apply Code
Remove Codes
Search
Previous Hit
Next Hit
Clear Hits
Help
Quit

Studying university is completly different from studying secondary school ,

studying university is hardest than studying school.

First of all, studying university it has less subjects than in studying secondary;

whereas studying secondary has mor subjects,

also studying university has a system of studying months «semster» ,

while studying secondary school the students are studying all a year;

furthermore , the number of students are completly different.

in studying university the number of the students very much in lecture ,

whereas in studying secondary school the number of the students are limits ,

and the main of the biggest differences are the away that the teachers explain the information to the students are not the sam way .

for example , in university the teachers give the students just the main points and in briefly

But in studying secondary school the teachers give the information to the students in deeply and specific ,

and another thing the differences of the level teachers in studying university

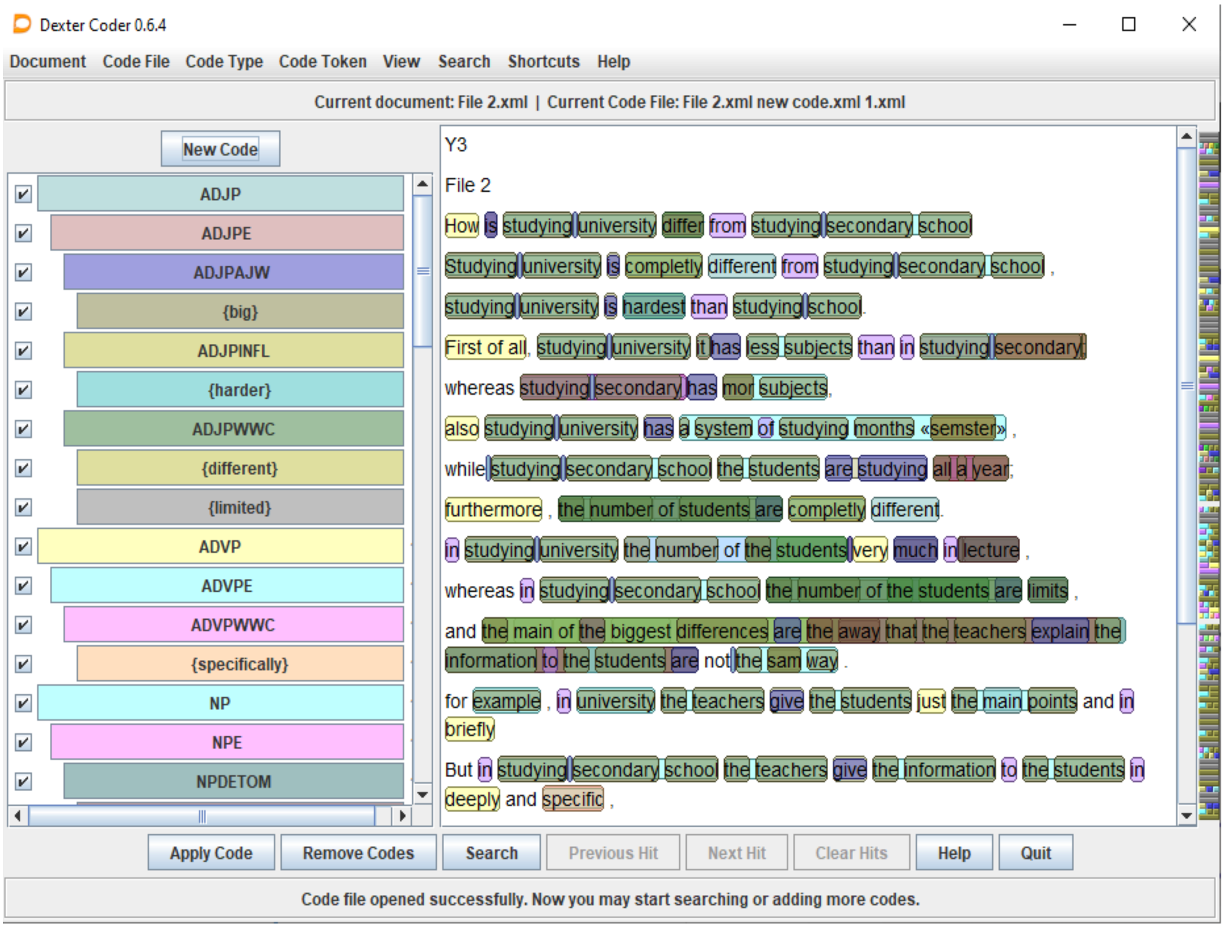
most of the teachers obtained the philosophy Doctor degree «Ph.D» or Master degree ,

while the teachers who studying secondary school are obtained the bachelor degree

at the end in my opinion I prefer the away that studying in university

Figure 4.2: Viewing coded spelling errors in Dexter Coder

In the second stage, after annotating the LEFLL corpus with the general descriptions of error types, the tagged errors were revised and classified into two main groups: phrase and spelling errors. Following the classification of errors into two main groups, an error tagging code system was developed (see Section 4.2.2 above) to replace the general descriptions of phrase and spelling errors as shown in Figure 4.3 below. Figure 4.3 also shows that where necessary, corrections were inserted to add more information about the learners' errors and may explain why a specific error tagging code was assigned to tag the error.



Current document: File 2.xml | Current Code File: File 2.xml new code.xml 1.xml

Y3

File 2

How is studying university differ from studying secondary school

Studying university is completly different from studying secondary school ,

studying university is hardest than studying school.

First of all, studying university it has less subjects than in studying secondary;

whereas studying secondary has mor subjects,

also studying university has a system of studying months «semster» ,

while studying secondary school the students are studying all a year;

furthermore , the number of students are completly different.

in studying university the number of the students very much in lecture ,

whereas in studying secondary school the number of the students are limits ,

and the main of the biggest differences are the away that the teachers explain the information to the students are not the sam way .

for example , in university the teachers give the students just the main points and in briefly

But in studying secondary school the teachers give the information to the students in deeply and specific ,

Code file opened successfully. Now you may start searching or adding more codes.

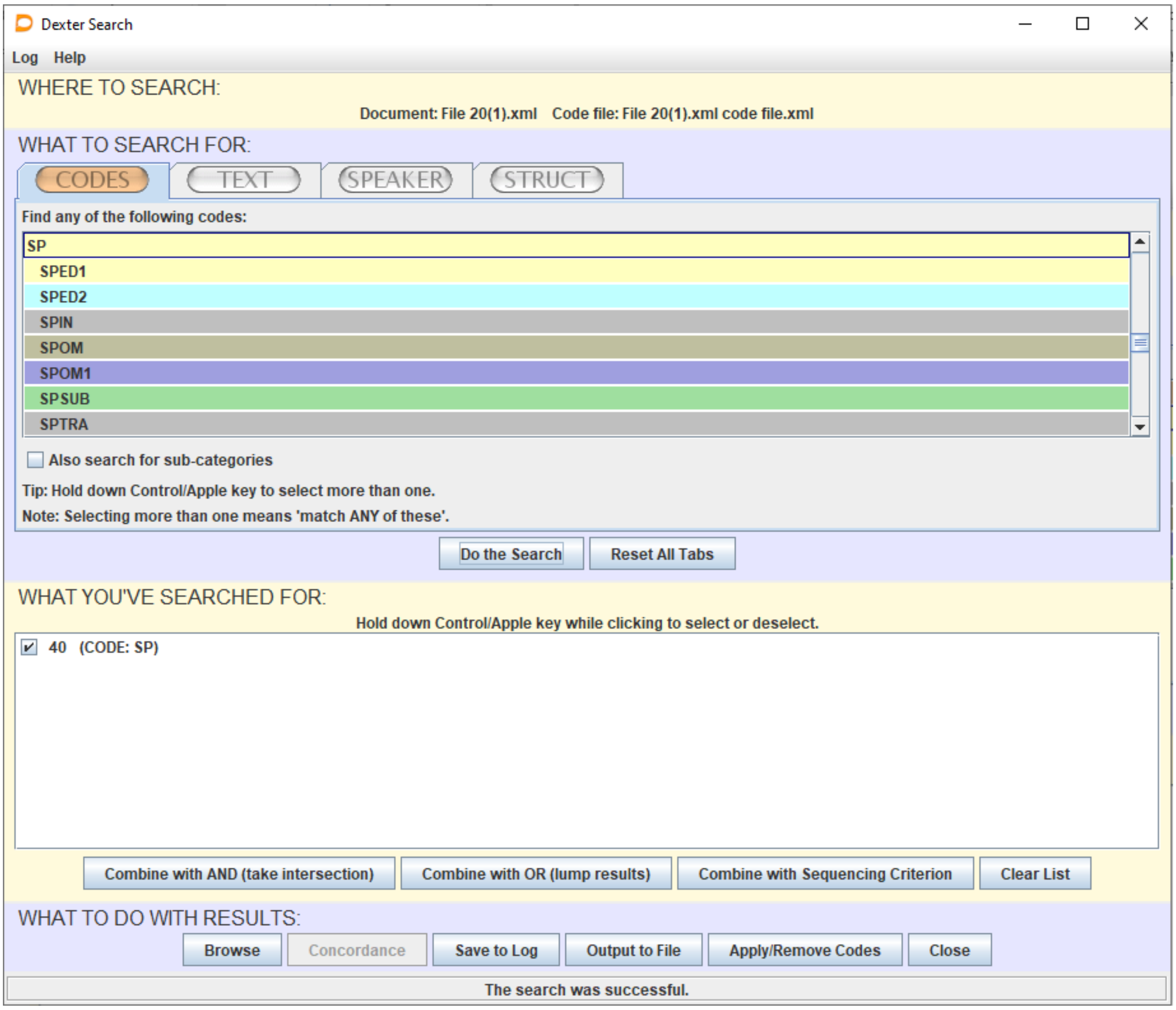
Figure 4.3: Phrase errors tagging codes and corrections in Dexter Coder (second stage)

4.3.2 Text Retrieval Corpus Tools

Having annotated the LEFLL corpus for spelling and phrase errors via Dexter Coder tool, the corpus was then passed to text retrieval programs that could retrieve the tagged features for error analysis. To perform this task, three sets of corpus tools were utilised: Dexter Search Tool, AntConc Corpus Analysis Toolkit and WordSmith Tools.

4.3.2.1 Dexter Search Tool

Dexter Search Tool is the third tool in the Dexter corpus toolkit. Once an exam essay file has been annotated in Dexter Coder, Dexter Search can be used to search and view the frequencies of the tagged features (e.g., spelling errors, the omission of determiners, etc.) in the annotated exam essay file. This is performed by selecting the tagging codes that represent the tagged features of interest in Dexter Coder (e.g., (SP) is the tag for spelling errors, (NPDETOM) is the tag for the omission of determiners in noun phrases, etc.). Figure 4.4 below shows the results provided by the Dexter Search Tool for the frequency of spelling errors in exam essay file 20.



Dexter Search

WHERE TO SEARCH:

Document: File 20(1).xml Code file: File 20(1).xml code file.xml

WHAT TO SEARCH FOR:

Find any of the following codes:

SP
SPED1
SPED2
SPIN
SPOM
SPOM1
SPSUB
SPTRA

□ Also search for sub-categories

Tip: Hold down Control/Apple key to select more than one.
Note: Selecting more than one means 'match ANY of these'.

WHAT YOU'VE SEARCHED FOR:

Hold down Control/Apple key while clicking to select or deselect.

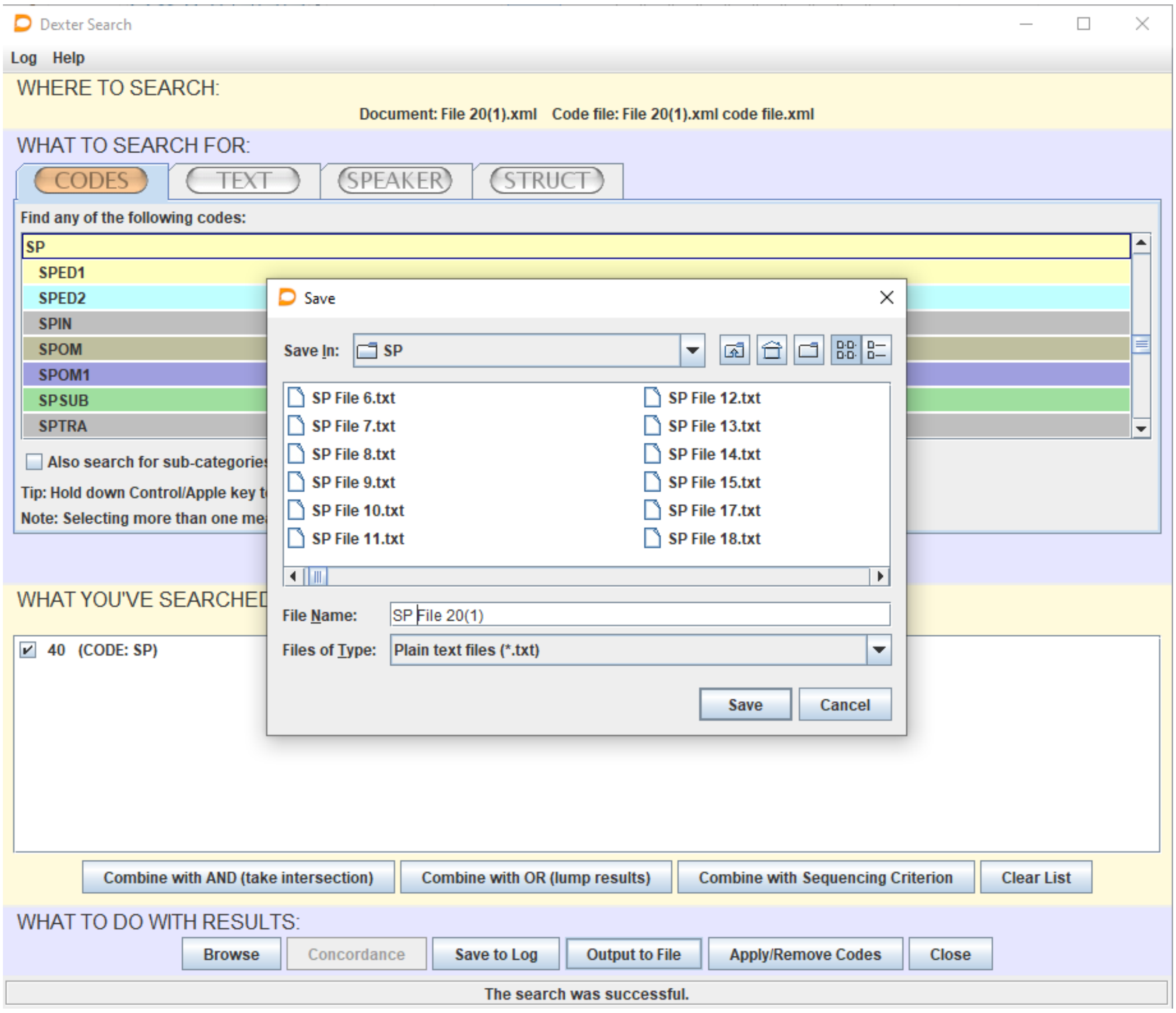
☑ 40 (CODE: SP)

WHAT TO DO WITH RESULTS:

The search was successful.

Figure 4.4: The frequency of an individual spelling error type in Dexter Search Tool

As discussed in Section 4.3.1.1 above, as a prerequisite and before starting to annotate the transcribed plain text exam essays files via Dexter Coder Tool, it was necessary to convert the plain text files from .txt files format to DeXML files format. Subsequently, the tagging codes in Dexter Coder and Dexter Search tools were also in the DeXML files format and could not be retrieved using AntConc. It was therefore necessary to convert the DeXML files back to .txt files. This was performed via the Dexter Search Tool. Each error tag in each annotated file was converted and saved separately as a .txt plain file format, see Figure 4.5 below as an example of converting the spelling tag SP from DeXML file format to the.txt file format.



Document: File 20(1).xml Code file: File 20(1).xml code file.xml

Find any of the following codes:

SP
SPED1
SPED2
SPIN
SPOM
SPOM1
SPSUB
SPTRA

□ Also search for sub-categorie

Tip: Hold down Control/Apple key t

Note: Selecting more than one me

Save In: SP

SP File 6.txt
SP File 7.txt
SP File 8.txt
SP File 9.txt
SP File 10.txt
SP File 11.txt
SP File 12.txt
SP File 13.txt
SP File 14.txt
SP File 15.txt
SP File 17.txt
SP File 18.txt

File Name: SP File 20(1)

Files of Type: Plain text files (*.txt)

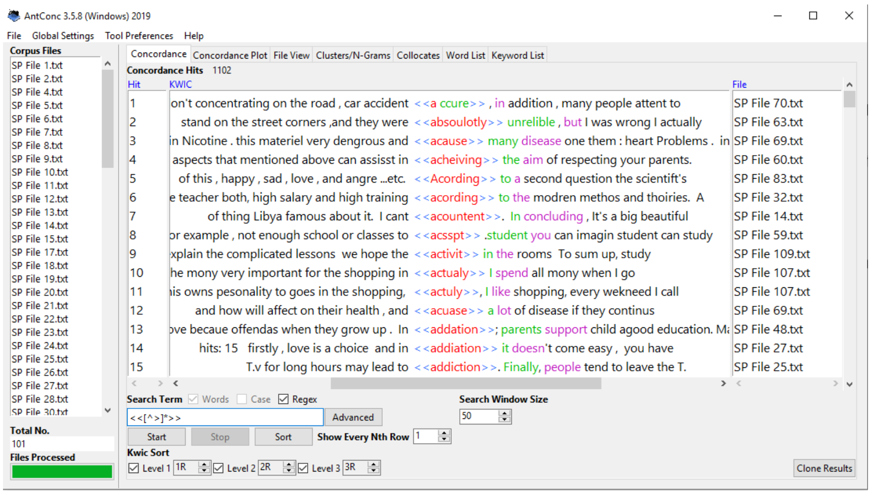
☑ 40 (CODE: SP)

The search was successful.

Figure 4.5: Converting an error tag from DeXML to .txt via Dexter Search

4.3.2.2 AntConc Corpus Analysis Toolkit

AntConc is a free corpus analysis toolkit that can be used for many different types of corpus linguistic analysis, e.g., generating the frequencies of words and keywords. The main reason for using AntConc in the current thesis is its ability to retrieve the converted tagged features (e.g., PPE tagging code (for *prepositional phrase errors*), NPDETOM tagging code (for *omission of determiner error in a noun phrase*)), from DeXML file format to .txt file format, and view them in concordance lines (see Figure 4.6 below for the concordance lines of spelling errors produced by year 3 students in LEFLL). The retrieved tagged features can then be contrasted with a standard native speaker corpus for error identification, or with a comparable corpus of texts written in the first language of the learners (in this case, Arabic) to identify evidence of possible L1 influence.



Corpus Files
SP File 1.txt
SP File 2.txt
SP File 4.txt
SP File 5.txt
SP File 6.txt
SP File 7.txt
SP File 8.txt
SP File 9.txt
SP File 10.txt
SP File 11.txt
SP File 12.txt
SP File 13.txt
SP File 14.txt
SP File 15.txt
SP File 17.txt
SP File 18.txt
SP File 19.txt
SP File 20.txt
SP File 21.txt
SP File 22.txt
SP File 23.txt
SP File 24.txt
SP File 25.txt
SP File 26.txt
SP File 27.txt
SP File 28.txt
SP File 30.txt

Total No. 101

Files Processed

Concordance Hits 1102

Hit	KWIC	File
1	on't concentrating on the road , car accident <<a ccure>> , in addition , many people attent to	SP File 70.txt
2	stand on the street corners ,and they were <<absoulotly>> unrelible , but I was wrong I actually	SP File 63.txt
3	in Nicotine . this materiel very dengrous and <<acause>> many disease one them : heart Problems . in	SP File 69.txt
4	aspects that mentioned above can assisst in <<acheiving>> the aim of respecting your parents.	SP File 60.txt
5	of this , happy , sad , love , and angre ...etc. <<Acording>> to a second question the scientift's	SP File 83.txt
6	e teacher both, high salary and high training <<acording>> to the modren methos and thoiries. A	SP File 32.txt
7	of thing Libya famous about it. I cant <<acountent>>. In concluding , It's a big beautiful	SP File 14.txt
8	or example , not enough school or classes to <<acsspt>> .student you can imagin student can study	SP File 59.txt
9	xplain the complicated lessons we hope the <<activit>> in the rooms To sum up, study	SP File 109.txt
10	he mony very important for the shopping in <<actualy>> I spend all mony when I go	SP File 107.txt
11	is owns pesonality to goes in the shopping, <<actuly>>, I like shopping, every wekneed I call	SP File 107.txt
12	and how will affect on their health , and <<acuase>> a lot of disease if they continus	SP File 69.txt
13	ove becaue offendas when they grow up . In <<addation>>; parents support child agood education. Ma	SP File 48.txt
14	hits: 15 firstly , love is a choice and in <<addiation>> it doesn't come easy , you have	SP File 27.txt
15	T.v for long hours may lead to <<addiction>>. Finally, people tend to leave the T.	SP File 25.txt

Search Term ☑ Words □ Case ☑ Regex
<<[^>]*>>
Search Window Size 50
Show Every Nth Row 1
Kwic Sort
☑ Level 1 1R ☑ Level 2 2R ☑ Level 3 3R

Figure 4.6: Concordance lines for spelling errors in Year 3 on AntConc

4.3.2.3 WordSmith Tools

WordSmith Tools is a suite of corpus analysis tools. Like AntConc, it contains powerful tools for carrying out various kinds of corpus-based analysis. In the current thesis, WordSmith Tools was mainly used to generate basic statistics of the words in the LEFLL corpus. The main reason for generating basic statistics of the words in the LEFLL corpus is provide the frequencies of different word lengths that form the LEFLL corpus (e.g., 2-letter words '*an, on, at, …*', 4-letter words '*with, from, four, …', etc.*). In the next Chapter, I will use the basic statistics of the words that compose the LEFLL corpus to show the advantage of the potential for error counting approach (proposed by this thesis) over the other two error counting approach, traditional error counting and the potential occasion analysis approaches.

To perform this task, the transcribed exam essays files were reviewed, and all spelling errors were corrected by the researcher. The reviewed files were saved as an additional corpus, apart from the original corpus that has been annotated via Dexter Coder. Following spelling corrections, the corpus was uploaded to WordSmith tools to generate the statistics of the words that form each sub-corpus (Year 1, Year 2, and Year 3) in the LEFLL corpus. The misspelled words were retrieved by AntConc, corrected and uploaded to WordSmith tools to generate the basic statistics. In the next chapter, the basic statistics of the misspelled words (in their correct forms) will also be used.

4.4 Inter-rater Reliability Test

As the previous section has shown, computer tools and digitized textual resources have made it possible to carry out error analyses of learner data on a scale and at a speed that would have been impossible to imagine a few decades ago. However, computers at the present time still cannot perform the central task of error analysis, which is to identify and classify learner errors in the first place. This task must still be carried out by the human analyst. Inevitably, this raises the possibility of subjectivity and inconsistency in the coding itself. The same researcher may classify the same error differently on different occasions, and different researchers may also classify errors differently. For instance. Dulay et al. (1982) classified the marked error in the sentence below as mis-formation:

*I *seen her yesterday*

This classification indicates that the writer has mis-formed the past tense verb form from the infinitive verb *'see'*. While this is a perfectly plausible interpretation, James (1998) subsequently objected and reclassified the above error as mis-selection, interpreting that the writer has used a dialect form that is inappropriate for the register in the above sentence. This can also be considered another possible classification, and there is in fact no easy way of deciding which of these two interpretations is ‘the correct one’.

Furthermore, in this research, since there are two main error categories - spelling and phrase errors - many observed language learners' errors could be classified under both categories. Consider the following example from the LEFLL corpus:

*I do many * thing in my Day*

In this example, the word *'thing'* is erroneous because it was preceded by the quantifier *'many'* which requires a plural countable noun. Therefore, the word *'thing'* should be replaced with *'things'*. On the one hand, this type of error could be classified as a spelling error assuming that the student has misspelled the word *'things'* by omitting the plural form marker '*s*'. On the other hand, it could instead be classified as a grammatical error (committed in a noun phrase) since it indicates a problem with number agreement between the quantifier *'many'* and the countable noun *'thing'*. One way of addressing this problem is to discuss such borderline cases explicitly and establish a clear and consistent policy for each case, but this is only a partial solution, which still does not provide sufficient assurance that researcher subjectivity has been controlled to an acceptable degree.

To overcome this problem and avoid the trap of subjectivity, it is now increasingly standard practice to carry out an inter-rater reliability (IRR) test, in which two or more observers carry out a coding task applying the same set of codes to the same dataset, and a statistical test is then applied to see whether and to what extent the two analysts have coded the data consistently with each other. If so, this can be considered as evidence that other observers would also be likely to agree with the researcher's observations and may thus understand and interpret the researcher's findings similarly. It is also an indication that the researcher's study and findings can be replicated (Carletta, 1996).

4.4.1 The Kappa Test

In the current thesis, Cohen’s *Kappa* was chosen as the statistical measure for the IRR test. Cohen’s *Kappa* provides a quantitative measurement of the degree of agreement between two observers who have classified a set of data using a set of categorical variables (Cohen, 1960). Kappa is widely recognized as a valid and robust measure of inter-rater reliability across the social sciences, and was chosen for the current thesis as there is a broad consensus as to how the results of this test are to be interpreted, as will be explained in more detail below.

Kappa is based on the difference between the observed agreement (the percentage of agreement) compared with the expected agreement (the percentage of the expected agreement by chance) (Brennan & Prediger, 1981; Byrt , 1996; Carletta, 1996; Randolph, 2005; Sim & Wright, 2005; Viera & Garrett, 2005; Simon , 2006; Warrens, 2010) and it is calculated as follows:

$K= (Po-Pe)/1-P\neg e$

Where Po is the observed agreement, Pe is the expected agreement.

Suppose that the two observers are asked to decide whether the number of randomly selected sentences are erroneous or correct. Both observers are asked to assign 'Yes' to the erroneous sentences and 'No' to the correct ones, as shown in Table 4.7 below.

		Observer 1 Result		
		Yes	No	Total
Observer 2 Result	Yes	**a**	b	m_1
	No	c	**d**	m_0
	Total	n_1	n_0	n

Table 4.7: Kappa Statistics Calculation

Where (a) and (d) are the number of times both observers agree on assigning (Yes) and (No) respectively to the examined sentences while (b) and (c) are the number of times both observers disagree on assigning (Yes) and (No) to the examined sentences.

(Po) is the sum of (a) and (d). When (Po) is 1, (b) and (c) will be zero. This means both observers 100% agreed on evaluating the sentences. When there is no agreement, (a) and (d) will be zero and therefore (Po) will be zero

(Pe) is calculated as follows:

$Pe=[(n1/n)*(m1/n)] + [(n0/n)*(m0/n)]$

Kappa ranges between -1 and 1. When Kappa is 1, this means there is a perfect agreement between the observers. When it is zero, this means the observed agreement exactly matches the expected one. When it is below zero, this means it is below the expected agreement range.

4.4.1.1 The Interpretation of Kappa Statistics

For this thesis, Kappa scores will be interpreted according to the popular rule-of-thumb interpretative framework proposed by Viera and Garrett (2005) as shown below:

If Kappa is < 0 this means the observed agreement is less than the expected agreement

If it is 0.01 – 0.20, there is a slight agreement

If it is 0.21 – 0.40, there is a fair agreement

If It is 0.41 – 0.60, there a moderate agreement

If it is 0.61 – 0.80, there is a substantial agreement

If it is 0.81 – 0.99, there is almost perfect agreement

4.4.1.2 The Administration and Results of the Kappa Test

For the current thesis, two raters were asked to participate in the IRR test. The first rater was an Arabic L1 speaking PhD student in linguistics, and the second rater was an MRes English literature researcher whose first language was English. The choice of one Arabic L1 and one English L1 rater was deliberate, as I wanted to establish whether my coding scheme was both plausible and usable from the point of view of raters from each language background.

There are two main options for the type of sample needed for an IRR test. The first option is based on the number of words, where 20% of the pilot study corpus will be extracted randomly. The problem with this option is that the distribution of all types of errors in each text may not be balanced. Some essays contain a large number of a specific type of errors whereas other essays may include hardly any instances of the same type of error. For instance, 40 spelling errors are present in exam essay File 17, whereas File 16 only contains 1 spelling error.

The IRR test data used in the current study consists of 300 errors which were displayed in concordances. The concordances will help the raters to identify the error from the context because some incorrect structures seem to be correct when they are viewed separately or used in another context. For example, there is nothing wrong with the word *'weak'* when it is used in the sentence:

He is too weak

But, in the following sentence:

First, I weak up in the morrning

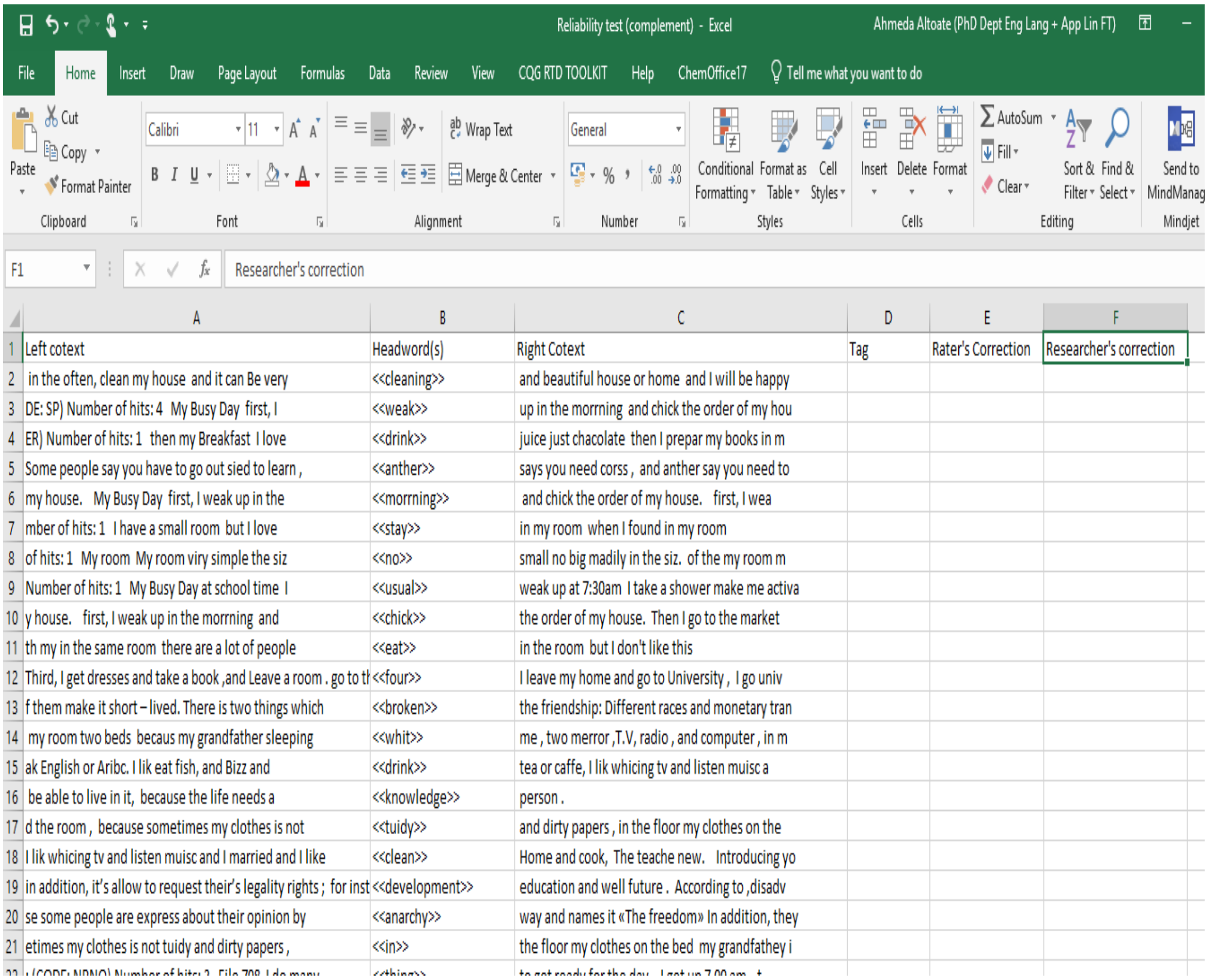
The word *'weak'* is either misspelled where the learner misspelled the correct word *'wake'* or has a lexical problem where the language learner has replaced the correct word *'wake'* with the wrong word *'weak'*. Thus, the examples above show (as discussed earlier in this chapter) that some errors are likely to fall under both grammatical and spelling error categories. It is the observer who is in the position to decide which error category to apply in each case. The two more raters were asked to classify the errors as either grammatical if the error indicates morphological or syntactical changes, lexical error if it shows substitution of the lexical item or spelling error if the error shows that one or more letters in the erroneous word is/are omitted, added or substituted, or if two letters are transposed. It is worth mentioning that the IRR test was conducted to determine the chance of agreement between the researchers and two raters when they classify errors as grammatical or spelling errors only. The finer categorisations (e.g., tense misformation, wrong subject pronoun, overuse of preposition, etc.) were excluded from

IRR test because we only need the frequencies of spelling and grammatical errors in the following chapter to show the advantages of the potential for error counting approach, proposed by this thesis, over the other two error counting approaches, the traditional error counting and the potential occasion analysis approaches. Therefore, it is important at this stage to ensure other researchers may classify the language learners' errors identified in the LEFLL corpus (either spelling or grammatical errors) similarly.

The total number of errors observed in the pilot study was 2905. Some of these errors could be classified as either spelling or grammatical errors, as in the example given above. Some other types of errors are more clearly grammatical, such as: the omission of the article in '*I've big table'*, the addition of unnecessary determiner in '*a people*', the omission of a verb as in '*when he or sh*' *sick'*. A third broad type of errors may be classified into both categories at the same time. e.g., *two merror* (for *two mirrors*) assuming that the learner misspelled the word by substituting the grapheme /i/ with /e/ and also had a grammatical error when he/ she omitted the plural marker '*s*'.

Before submitting the sample to the two raters, errors that are clearly grammatical were excluded, so only those errors that may fall in the grey area were extracted. The total number of the latter type of errors was 280. 943 instances of clear spelling errors were also extracted and added to the 280 'grey area' errors, giving a total set of 1223 errors overall. Due to the time-consuming nature of the task and also the degree of accuracy that the IRR test requires, only 20% of the targeted errors were given to the raters, i.e., a test sample of 300 errors.

The selection of the sample for the reliability test was based on a random selection of 300 errors out of 1223. If both types of errors are mixed and 20% of errors are randomly selected, there is a chance that the sample may contain a handful of grammatical errors of the kind which are not within the scope of the current thesis. The objective of the IRR test outcomes needs to be balanced, because any rater may classify the errors differently. The test must show that the researcher's classification of both grammatical and spelling errors is reliable in statistical terms. Therefore, half of the sample for the reliability test, 150 errors out of 280, were selected randomly from the grey area error types, and the second half, 150 out of 943, were also randomly selected from the spelling error data. The errors were extracted from AntConc and outputted into a spreadsheet in concordance format (see Figure 4.7 below). The first column contains the first part of the concordances that precedes the erroneous word, the second column only contains erroneous words between angle brackets as they were extracted from AntConc, and the third column contains the third part of the concordance.



	A	B	C	D	E	F
1	Left cotext	Headword(s)	Right Cotext	Tag	Rater's Correction	Researcher's correction
2	in the often, clean my house and it can Be very	<<cleaning>>	and beautiful house or home and I will be happy			
3	DE: SP) Number of hits: 4 My Busy Day first, I	<<weak>>	up in the morrning and chick the order of my hou			
4	ER) Number of hits: 1 then my Breakfast I love	<<drink>>	juice just chacolate then I prepar my books in m			
5	Some people say you have to go out sied to learn ,	<<anther>>	says you need corss , and anther say you need to			
6	my house. My Busy Day first, I weak up in the	<<morrning>>	and chick the order of my house. first, I wea			
7	mber of hits: 1 I have a small room but I love	<<stay>>	in my room when I found in my room			
8	of hits: 1 My room My room viry simple the siz	<<no>>	small no big madily in the siz. of the my room m			
9	Number of hits: 1 My Busy Day at school time I	<<usual>>	weak up at 7:30am I take a shower make me activa			
10	y house. first, I weak up in the morrning and	<<chick>>	the order of my house. Then I go to the market			
11	th my in the same room there are a lot of people	<<eat>>	in the room but I don't like this			
12	Third, I get dresses and take a book ,and Leave a room . go to tl	<<four>>	I leave my home and go to University , I go univ			
13	f them make it short – lived. There is two things which	<<broken>>	the friendship: Different races and monetary tran			
14	my room two beds becaus my grandfather sleeping	<<whit>>	me , two merror ,T.V, radio , and computer , in m			
15	ak English or Aribc. I lik eat fish, and Bizz and	<<drink>>	tea or caffe, I lik whicing tv and listen muisc a			
16	be able to live in it, because the life needs a	<<knowledge>>	person .			
17	d the room , because sometimes my clothes is not	<<tuidy>>	and dirty papers , in the floor my clothes on the			
18	I lik whicing tv and listen muisc and I married and I like	<<clean>>	Home and cook, The teache new. Introducing yo			
19	in addition, it's allow to request their's legality rights ; for inst	<<development>>	education and well future . According to ,disadv			
20	se some people are express about their opinion by	<<anarchy>>	way and names it «The freedom» In addition, they			
21	etimes my clothes is not tuidy and dirty papers ,	<<in>>	the floor my clothes on the bed my grandfathey i			
22	; (CODE: NPNO) Number of hits: 2 File 708 I do many	<<thing>>	to get ready for the day I get up 7.00 am t			

Figure 4.7: Error tagging codes extracted for reliability test

In the test, the raters were asked to rate each error as spelling, grammar, or both if they believe that it contains both error categories. They were also asked to provide the correct form of each error. This latter information made it possible to check for similarities and differences between the researcher's and the raters' corrections. It will also show whether the researcher identified the grammatical errors correctly in terms of phrase types (e.g., Noun, verb, adjective phrase errors, etc.) and phrase subcategory types.

The test consists of two stages. In the first stage, each rater was asked to code the test data individually. The Kappa statistic was then calculated. In the second stage, the researcher and the two raters had a group discussion regarding their error classifications and corrections and everyone was asked to re-rate each error following the discussion. The Kappa statistic was then re-calculated.

4.4.1.3 Kappa Results – Stage One

In the first stage, the agreement between the researcher and Rater-1 was 85.67% (257 errors out of 300) whereas the agreement between the researcher and Rater-2 was 79.67% (239 errors out of 300). Both percentages may, initially, indicate that the agreement between the researcher and the Raters is very low, but the Kappa results showed that there was a considerable agreement. With Rater_1 for example, the actual agreement between the researcher and Rater_1 is 138 which is almost twice as high as the chance of agreement, and in the case of spelling and grammar error categories, the actual agreement was 9 whereas the random chance of agreement was 5. With Rater_2, the actual agreement between the Researcher and Rater_2 was almost the same as with Rater_1 in the case of spelling errors, but decreased in the spelling and grammatical category to 9 cases of actual observed agreement versus 7 in the expected chance of an agreement.

The review of both Raters' work showed that some errors were not clear to the Raters because the two parts of the concordances (before and after the erroneous words) did not provide sufficient information that would enable the Raters to classify the errors confidently. The review process also identified some errors that were marked differently by both Raters. Some of the Raters' corrections and comments to some errors indicate that the Raters had mistakenly classified some errors.

Interestingly, some errors were identified by Rater_1 only, such as:

Gote <<anheal>> tree

and

there is a very big tv with 42 <<bosah>>

These two errors indicate language transfer. Since Rater_1 is L1 Arabic, she was able to recognize, as she mentioned in the group discussion, that this error had been caused by language transfer – specifically, by borrowing lexical items from his/her L1. The word '*anheal'* is an Arabic word which means '*palm'* in English and the word '*bosah'* is a unit of measurement which corresponds to '*inch'* in English. Rater_2 suspected that the two errors could be language transfer errors but he was not certain due to his unfamiliarity with Arabic. Both Raters agreed on 224 errors out of 300, and the disagreements were found to be largely due to one or other of the two issues discussed above, i.e., the failure of some concordance lines to disambiguate particular errors, or the transfer errors that Rater_1 noticed but Rater_2 failed to identify. This necessitated a group discussion between the Researcher and both Raters to discuss both Raters' interpretation of the errors.

Before the group discussion, the researcher went through both raters' markings of the IRR test sample and marked the raters' categories that did not agree with the researcher's categories. The researcher also marked the corrections that did not match his corrections as a preparation step for a group discussion.

4.4.1.4 Kappa Results – Stage Two

During the group discussion, the researcher and the two raters went through the ambiguous errors that both raters were not able to identify and the classifications and corrections of their errors that did not match the researcher's. The review process showed that some errors had been accidentally classified into a specific error category. Some error categories agreed with the researcher's categories but their corrections differed from the researcher's. When they were given access to complete learners' essays in the LEFLL corpus, the raters often decided to change their corrections. For example, the error *smalle* in:

> *sometimes I think the <<smalle>> room It's nice for me*

was, initially, marked as 'B' (stands for 'both' spelling and grammatical error) by Rater_1 and only classified as spelling error by the researcher and the Rater_2. Both raters' corrections were the comparative form of the adjective 'smaller' while the researcher's correction was 'small' (i.e., removal of the letter 'e'). In the group discussion, Rater_1 explained that she had accidentally classified the error into 'B' and confirmed that it was merely a spelling error assuming that the learner has only omitted the letter 'r'. However, when given access to the full text, Rater 1 could see that the same error 'smalle' occurred several times in the text, with even the essay title being written as 'A smalle room'. This casts doubt on the validity of interpreting 'smalle' as a comparative form of the adjective 'small' where the learner has only dropped the letter 'r'. The raters concluded that the learner was referring to a small room with no comparative meaning intended.

Some errors were controversial in terms of corrections and which error correction is possible, e.g., for the error:

Every one of us has <<know>> a lot of teachers during his / her studying time

the researcher and Rater_2 only corrected the verb <<*know*>> using the past participle form *'known'*, but Rater_1 claimed that the correction should be made to the tense by substituting the present perfect tense with simple past tense *'knew'*. Rater_1 argued that the present perfect form cannot be used to show the state that one or more people knew one or more other people. By consulting British National Corpus (BNC), the researcher was able to provide attested examples of native speakers of British English using the present perfect form in similar contexts that express the state where one or more people have known others. After the group discussion, the Kappa score was recalculated, and the results indicated a considerable increase of agreement compared to the first stage.

Overall, the group discussion stage provided clear support for the researcher's claim that the set of error categories developed for the current research is both reliable and replicable in further research. The group discussion was also useful in that it provided some suggestions for specific issues that needed to be taken into consideration. For spelling errors, for example, one of the suggestions to emerge from the discussion was that the researcher would need to check whether erroneous words occurred only once in a text or were repeated several times in the same essay. In other words, in coding the full dataset the researcher would need to look beyond the immediate sentence or paragraph where the error occurred.

Conclusion

This chapter began by describing the planning and compilation of the Libyan English as a Foreign Language Learners (LEFLL) corpus. The primary data collection plan was aimed to compile a representative learner corpus of the Libyan English learners from different parts in Libya. Due to unforeseen problems, the data collection plan was subsequently modified and the data collection process itself was ultimately restricted to Libyan English learners at the English department, Faculty of Languages, Benghazi University.

The chapter then went on to describe the two sets of corpus tools that have been used to annotate the LEFLL corpus and retrieve phrase and spelling errors tagged codes. The chapter also discussed the phrase and spelling error classification systems and the error tagging codes that were developed for the current thesis. To ensure the reliability of the phrase and spelling classifications that have been made for the current thesis, an inter-rater reliability (IRR) test called the '*Kappa statistics'* was conducted. The results obtained showed that the classification of phrase and spelling errors was reliable, and discussions with the two external raters provided valuable feedback on specific issues of difficulty that might otherwise have been overlooked.

The following chapter will also be somewhat methodological in orientation, in that it will compare three approaches to error counting, traditional error counting approach, the potential occasion analysis approach proposed by Thewissen (2012; 2015) and the potential for error counting approach proposed by this thesis. This comparison will show us the advantages the potential for error counting approach offers over the other two error counting approaches, and will also serve to pave the way for the remainder of the analysis by providing basic quantitative information about error frequencies in the LEFLL corpus.

CHAPTER 5: QUANTIFYING LANGUAGE LEARNERS' ERRORS IN THE LEFLL CORPUS: A COMPARISON OF THREE ANALYTICAL APPROACHES

Introduction

This chapter compares three error counting approaches, namely: The Traditional Error Counting Approach, the Potential Occasion Analysis Approach developed by Thewissen (2012; 2015) and the Potential for Error Approach proposed by this thesis. The comparison aims to evaluate each error counting approach and establish what each approach can tell us.

Thus, the chapter will analyse the language learners' errors in the LEFLL corpus by applying the three error counting approaches. Firstly, the traditional error counting approach will be used to calculate the percentage of each error category (e.g., spelling, noun phrase errors, verb phrase errors, etc.). Following this, traditional error counting approach will then be applied to quantify and compare between the percentages of spelling errors of different word lengths: 2-letter words (e.g., *in, on, an,* etc.), 3-letter words (e.g., *our, the, you,* etc.), etc. In the Potential Occasion Analysis Approach, in contrast, the percentage of each error category will be calculated within a relevant '…environment of potential occasions for error' (Thewissen, 2012:3). Thirdly, in the Potential for Error Approach, the percentage of each error category in the LEFLL corpus will be calculated within a relevant environment of potential for error. The Potential for Error Approach will also be applied on the level of spelling errors of different word lengths by looking at the percentages of spelling errors of different word lengths. The reason why the percentages of spelling errors will be calculated based on the Potential for Error Counting approach is to provide a practical example of what the Potential Occasion Analysis approach failed to consider. Finally, the chapter will discuss the results obtained via the three error counting approaches and highlight the limitations of each approach.

5.1 Error Analysis Based on the Traditional Error Counting Approach

In accordance with the Traditional Error Counting (henceforth TEC) approach, the percentage of each type of error (e.g., spelling errors, noun phrase errors, prepositional phrase errors, etc.) produced by the learners in each LEFLL sub-corpus was calculated out of the total number of all types of errors in the same LEFLL sub-corpus. As discussed in Chapter 4, following the manual annotation of the LEFLL corpus using the Dexter Coder, the tagged language learners' errors were retrieved and analysed using the AntConc suite of corpus tools. Thus, Figure 5.1 below shows the distribution of the language learners' errors (spelling and phrase errors) across the LEFLL sub-corpora (Year 1 sub-corpus, Year 2 Sub-corpus and Year 3 Sub-corpus). As the Figure shows, the Year 3 sub-corpus scored the lowest percentage (31.73%) of all types of errors (spelling, noun phrase, verb phrase, prepositional phrase, adjective phrase and adverb phrase errors) out of the total number of all types of errors in the LEFLL corpus.

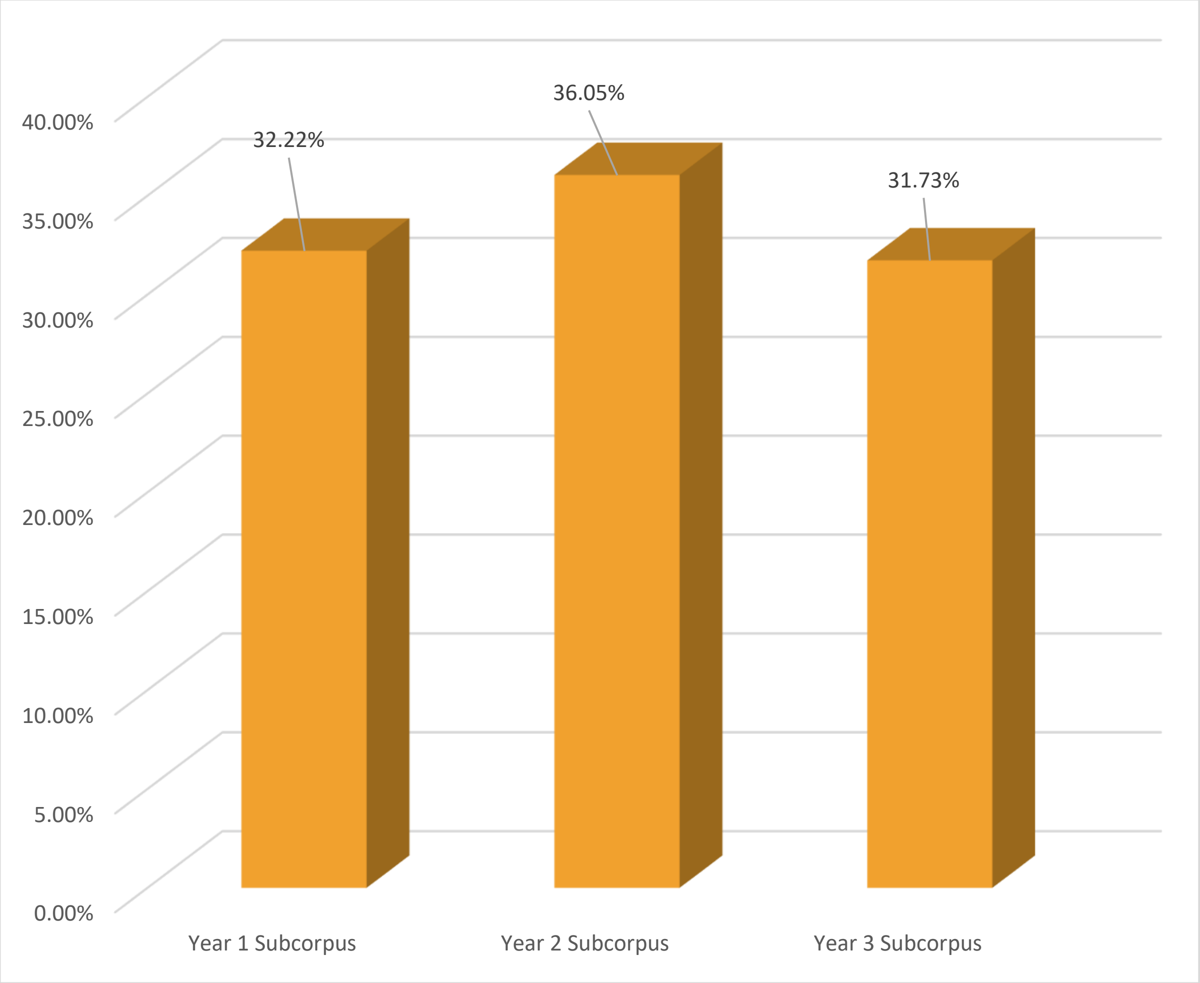


Figure 5.1: The percentages of errors in the LEFLL corpus

As can be seen, this analysis finds no steady change (either in increase or decrease) in the percentages of language learners' errors as the analysis moves from one level to another, despite the fact that the three sub-corpora are almost of the same size, and given that the students went through essentially the same language learning experience (they are Arabic native speakers, were taught by the same tutors at the same university and studied the same curriculums at least in Years 1 and 2). In fact, the Year 2 sub-corpus represents an anomalous stage between the Year 1 sub-corpus and Year 3 sub-corpus in that it scored by far the highest percentage of errors (36.05%) out of the total number of errors in the LEFLL corpus. That is 3954 out of all types of errors in the Year 2 sub-corpus out of 10968 total number of errors in the

LEFLL corpus. A possible reason for this deviation in the Year 2 sub-corpus could be the degree of topic flexibility in the writing task, which differs from one level to another. For Granger (1998), topic is one of the most important factors in the learner corpus design criteria because it affects lexical choice in obvious and fundamental ways. At first glance, this seems to offer a plausible explanation for the results presented in Figure 5.1. As discussed earlier in Section 4.1.2, in year 2 exam essays (which form the Year 2 sub-corpus), the students had the fewest options of the topics for the writing task. The students were given three topics only and asked to choose and write about one of these topics only: *how to celebrate a special day, a special occasion in your culture* or *describe your room.* In contrast, the students in Year 1 had five writing task topic options: *A person you care about, Introducing yourself, Your morning routine, your hometown* and *your sleeping habits.*

On closer inspection, however, topic variability seems a less likely explanation. There are two points to note here. Firstly, the students in year 3 were given an entirely open writing task in which they were allowed to choose their own topic, which strongly suggests that task variability seems an unlikely explanation for the results obtained overall. Furthermore, the type-token ratio data are not consistent with what we would expect to observe if topic were a critical factor. Given that learners in year 1 had more writing task topic options than did learners in year 2, it would be reasonable to expect that the learners in year 2 would have fewer opportunities to use different word types compared to the learners in year 1. This in turn should be reflected in the type/token ratios for the Year 1 and Year 2 sub-corpora. Specifically, we should expect the type/token ratio in the Year 1 sub-corpus to be higher than the type-token ratio in the Year 2 sub-corpus. However, the analysis finds that the opposite is the case: not only is the type/token ratio in the Year 2 sub-corpus higher than that of the Year 1 sub-corpus, but there is also an increase in the ratio as the learners proceed from one level to another, as shown earlier in Table 4.1

Year 1 sub-corpus	Year 2 sub-corpus	Year 3 sub-corpus	LEFLL Corpus
3534	3954	3480	10968

Table 5.1: The total number of errors in the LEFLL sub-corpora

So far, the results showed in Figure 5.1 and Table 5.1 above represented a mixture of spelling and phrase errors produced by the learners in the LEFLL corpus. By breaking down these results into spelling and phrase error categories, we can see, as shown in Figure 5.2 below, that the same deviation observed in the Year 2 sub-corpus (as shown in Figure 5.1 and Table 5.1 above) is clearly evident in almost all error categories. Apart from the prepositional and adjective phrase errors, the remaining error categories in the Year 2 sub-corpus represent an anomalous stage between the Year 1 sub-corpus and Year 3 sub-corpus. With the exception of the noun phrase error categories, the percentages of error categories (spelling, verb phrase and adverb phrase errors) declined in the Year 2 sub-corpus before increasing in the Year 3 sub-corpus. For instance, the percentage of spelling errors declined from 34.47%, in the Year 1 sub-corpus, to 28.50% in the Year 2 sub-corpus before increasing to 31.64% in the Year 3 sub-corpus.

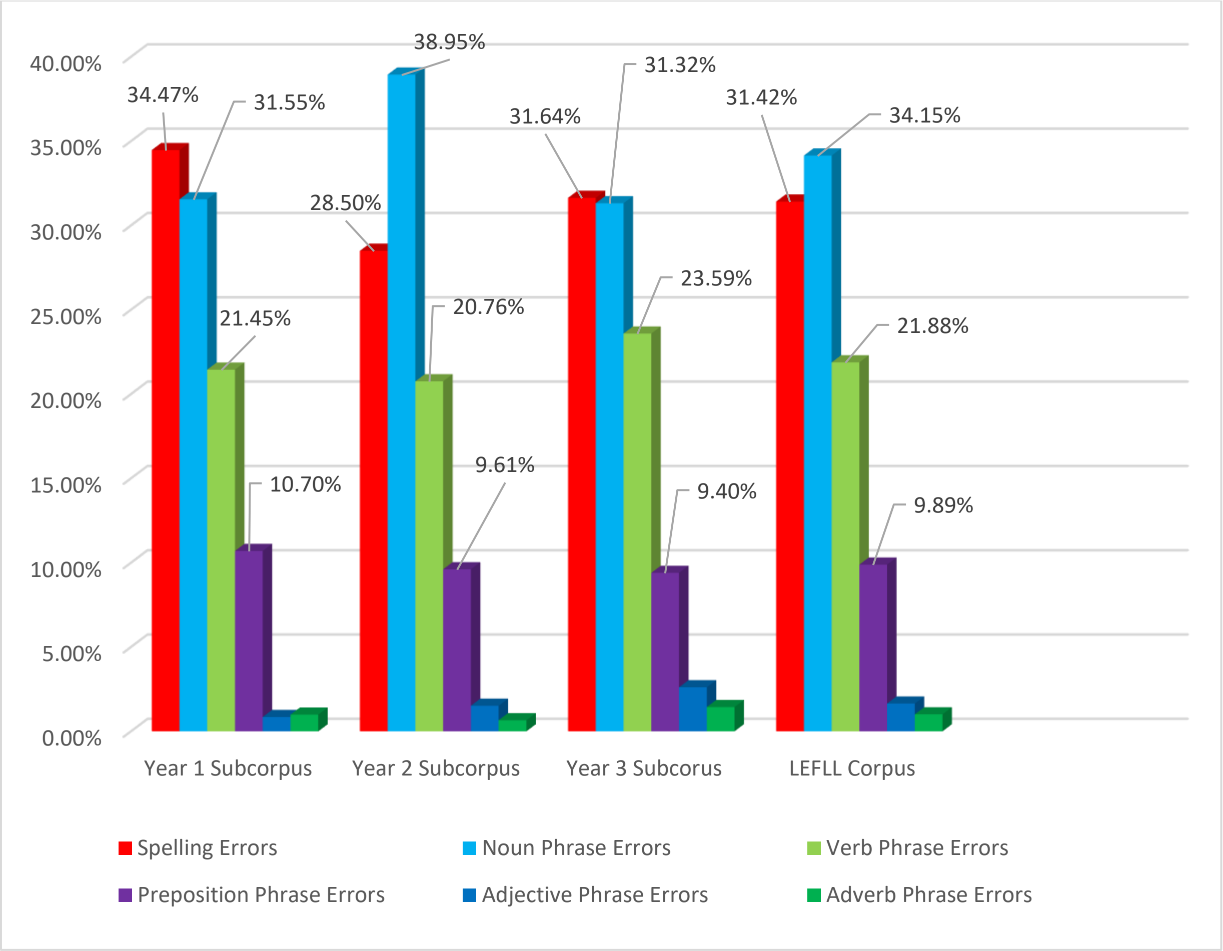


Figure 5.2: The percentages of spelling and phrase errors in the LEFLL corpus – TEC approach

Figure 5.2 shows that there is a steady but slight progress in the prepositional phrase category as the percentages of prepositional phrase errors start decreasing (10.70%, 9.61% and 9.40% in the Year 1, Year 2 and Year 3 sub-corpora respectively). On the other hand, there is a slight deterioration in the adjective phrase category as the analysis moves from one level to another. The percentage of spelling errors in the Year 3 sub-corpus, as discussed above, also increased and returned to a value close to the one in the Year 1 sub-corpus.

Generally, spelling and noun phrase errors collectively and in each LEFLL sub-corpus represented the major part (almost two-thirds) of all language learners' errors, in each LEFLL sub-corpus, while the other four error categories (verb, prepositional, adjective and adverb phrase errors) constituted the remaining portion of errors (about one-third), in each LEFLL sub-corpus. This can also be observed in the frequency of errors in each error category (spelling, noun phrase, verb phrase, etc.) across the LEFLL sub-corpora as shown in Table 5.2 below. For instance, the total number of spelling and noun phrase errors in the Year 1 sub-corpus is 2,333 out of the total number of errors (3534) in the Year 1 sub-corpus.

Another important finding is that there is an alteration in the comparative frequencies of both spelling and noun phrase errors as the analysis moves from one level to another. As presented in Figure 5.2 above, the first-year undergraduate students, for example, produced more spelling errors (34.47%) than noun phrase errors (31.55%) while the second-year undergraduate students produced more noun phrase errors (38.95%) than spelling errors (28.50%). The third-year undergraduate students, on the other hand, almost produced the same percentages of spelling and noun phrase errors with a slight increase in the percentage of spelling errors (31.64% and 31.32% of spelling and noun phrase errors respectively).

The remaining types of language learners' errors (verb, prepositional, adjective and adverb phrase errors) constituted a small portion of errors out of all types of errors in the LEFLL corpus. In most cases, it was found that the percentages of errors were slightly increased among the learners in the highest level. This is most clearly evident in the verb, adjective and adverb phrase errors.

These results, based on the TEC approach, may indicate that the learners encounter more major problems with spelling and noun phrases than they do with any other categories (verb, prepositional, adjective and adverb phrase categories). Furthermore, the TEC analysis suggests that this seems to be a persistent problem since there is no significant progress as the learners move from one year to another in their undergraduate degree study. In some cases, this issue may even escalate as the learners proceed from one level to another. For instance, the percentage of spelling errors produced by the second-year undergraduate students was 28.50%. This went up to 31.64% in the Year 3 sub-corpus.

Observed Errors	**Year 1 sub-corpus**	**Year 2 sub-corpus**	**Year 3 sub-corpus**	**LEFLL Corpus**
SP	1218	1127	1101	3446
NPE	1115	1540	1090	3745
VPE	757	819	821	2397
PPE	379	379	327	1085
ADJPE	30	60	91	181
ADVPE	35	26	49	110
Total	3534	3951	3479	10964

Table 5.2: The frequency of language learners' errors in the LEFLL corpus

So far, this section has taken a brief look at the distribution of spelling and phrase errors in the three sub-corpora based on the TEC approach. As we have observed, TEC has provided us with the percentages of each error category out of the total number of errors in each LEFLL sub-corpus. Based on the percentages of spelling and phrase error categories, we can see that the spelling and noun phrase error categories represented the major problems to the learners represented in the LEFLL corpus. We can also see that, according to TEC, the language learning process does not follow a linear trajectory, and that year 2 represented an anomalous stage between year 1 and year 3. Finally, the TEC results indicated that the learners did not make significant progress when we compared the results obtained from the Year 3 sub-corpus with those obtained from the Year 1 sub-corpus.

Turning now to the level of spelling errors, after retrieving spelling errors from the LEFLL corpus, the misspelled words were manually corrected and uploaded to WordSmith Tools in order to obtain statistics for the corrected misspelled words, i.e., the frequencies of spelling errors of different word lengths in the LEFLL corpus in their correct forms. Word length is the measurement of the number of letters that form the word. Thus, the 1-letter words are the words that consist of one alphabetical letter only, such as: *I, a, etc.* the 2-letter words are the words that consist of two alphabetical letters only, e.g., *an, on, we, etc.* and so on.

Obviously, it would be wrong to assume that the word lengths for misspellings would be the same as for the correctly spelled words. For example, in my data, the 7-letter word '*because'* has been misspelled in a variety of different ways, such as *'becec'* (5 letters) and *'becuss'* (6 letters), among many other variants. The aim of identifying the lengths of misspelled words is to apply and evaluate TEC on different word lengths. Typically, as the words become longer, they are expected to become more difficult to spell and vice versa. For instance, it would be more difficult for an English learner to spell the word *'manoeuvre'*, which consists of 9 letters and has three syllables, than the word *'man'* which consists of 3 letters and has one syllable only. Therefore, it is expected that the percentages of spelling errors start increasing as longer words become more frequent in the students' writing.

As shown in Figure 5.3 and Table 5.3 below, the TEC results paint a more complex picture. Contrary to expectations, spelling errors in the Year 1 and Year 2 sub-corpora actually fall between 2-letter words and 13-letter words whereas in the Year 3 sub-corpus spelling errors also include 14-letter words (6 misspelled words 0.54%) and 15-letter words (4 misspelled words 0.36%). Surprisingly, the highest percentages of spelling errors in the three sub-corpora fell between the 4-letter and 7-letter words (as shown in Figure 5.3 and highlighted in Table 5.3) which start decreasing as the words became longer. One of the possible explanations for these observations is that as the words become longer, learners become less confident. Thus, they tend to avoid taking the risk of using longer words and use shorter words instead to reduce the chance of committing spelling errors. Later in this chapter, however, I will propose an alternative explanation, based on the concept of *potential* for error in Section 5.3 below.

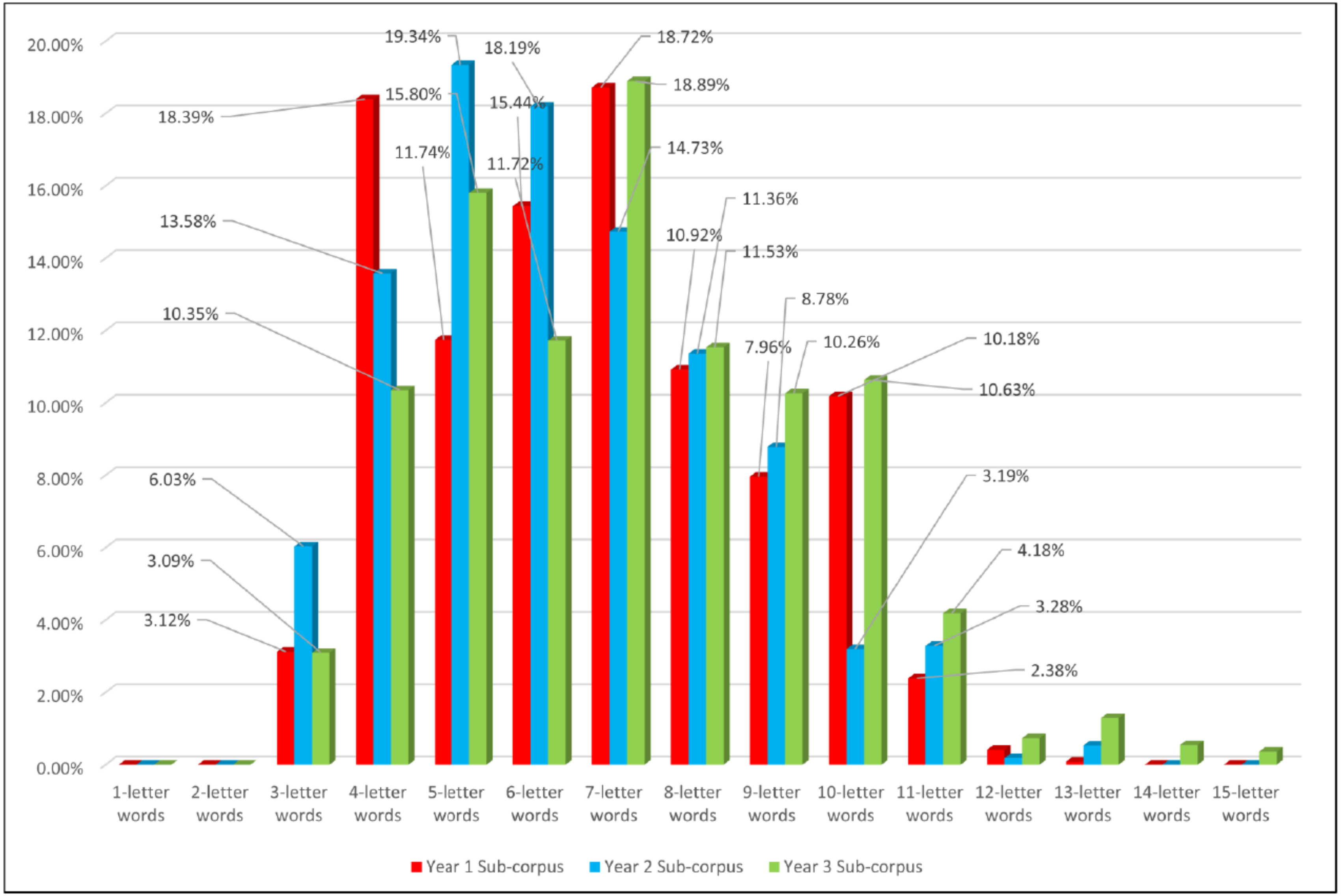


Figure 5.3: The percentages of spelling errors per word length in the LEFLL corpus – TEC approach

Once again, the Year 2 sub-corpus seems to be an anomalous stage between the Year 1 sub-corpus and Year 3 sub-corpus. For instance, the percentage of spelling errors of the 5-letter words produced by the first-year undergraduate students was 11.74%. This went up to 19.34% in year 2 and then went down to 15.80% in year 3. Conversely, the percentage of spelling errors of the 7-letter words produced by the first-year undergraduate students was 18.72% and then decreased to 14.73% in year 2 before increasing again to 18.89% in year 3.

Word Type	Year 1 Sub-corpus		Year 2 Sub-corpus		Year 3 Sub-corpus	
	SP	SP% (1)	SP	SP% (1)	SP	SP% (1)
1-letter words	0	0.00%	0	0.00%	0	0.00%
2-letters words	8	0.66%	9	0.80%	7	0.64%
3-letters words	38	3.12%	68	6.03%	34	3.09%
4-letters words	224	18.39%	153	13.58%	114	10.35%
5-letters words	143	11.74%	218	19.34%	174	15.80%
6-letters words	188	15.44%	205	18.19%	129	11.72%
7-letters words	228	18.72%	166	14.73%	208	18.89%
8-letters words	133	10.92%	128	11.36%	127	11.53%
9-letters words	97	7.96%	99	8.78%	113	10.26%
10-letters words	124	10.18%	36	3.19%	117	10.63%
11-letters words	29	2.38%	37	3.28%	46	4.18%
12-letters words	5	0.41%	2	0.18%	8	0.73%
13-letters words	1	0.08%	6	0.53%	14	1.27%
14-letters words	0	0.00%	0	0.00%	6	0.54%
15-letters words	0	0.00%	0	0.00%	4	0.36%
Total	1218		1127		1101	

Table 5.3: The percentages of spelling errors per word length in the LEFLL corpus – TEC approach

The analysis also revealed, as shown in Figure 5.3 and Table 5.3, that none of the 1-letter words were incorrectly spelled. This led to further investigation to retrieve the 1-letter words produced by the learners in the LEFLL corpus as shown in Table 5.4 below. The Table shows that there are twelve 1-letter words in the LEFLL corpus. Some of these 1-letter words are alphabetical letters attached to other words by apostrophes, such as *'s'* in *brother's* and *'o'* in *o'clock.* Another type of 1-letter word represents the abbreviation of words or phrases, such as the letters *'U'* and *'S'* in '*U.S'* (for '*US'* United States) and e.t.c (for *'etc'* Et cetera). The final type of 1-letter word are the grammatical words *'a'* as an indefinite article and *'I'* as the first person singular subject pronoun.

The word	description	Examples from LEFLL corpus
A	• as an indefinite article • as an abbreviation to refer to the time	- *I have a small room* - *get up at 7:00 A.M*
C	• in an abbreviation form	- e.t.c (for *etc.)*
D	• as an abbreviation form	- *Ph.D*
E	• in an abbreviation form	- *e.t.c (for etc.)*
I	• as the first person singular subject pronoun	- *I enjoy my job*
M	• as an abbreviation of the verb to be *'am'* • as an abbreviation to refer to the time	- *I'm 18 years old* - *I leave my room at 8:45 A.M / my second class starts at 12:30 P.M*
O	• to refer to the time	- *I get up at 8:00 o'clock*
P	• as an abbreviation to refer to the time	- *I go to the shopping at 6:30 P.M*
S	• as a possessive marker • as an abbreviation for the verb to be 'is' • as a letter in the abbreviation 'U.S'	- *my brother's room* - *He's fifty years old* - *Young people who live in the U.S have more freedom*
T	• as a letter in the short form '*n't'* • as a letter in the short form '*T.V'*	- *Most peole don't remember* - *First of all, watching T.v has many advanteges*
U	• as a letter in the abbreviation 'U.S'	- *In U.S, parents allow their children to move*
V	• as a letter in the short form '*T.V'*	- *watching T.v for long hours may lead to adiction*

Table 5.4: A list of 1-letter words in the LEFLL corpus

This may explain why all 1-letter words were correctly spelled by the learners. The fact that they are simple and commonly used by both the native speakers and English learners undoubtedly makes them easy to spell. Indeed, it would be hard to imagine even the lowest level elementary student mis-spell the words '*a*' or '*I*' in practice. Furthermore, according to the surface strategy taxonomy as discussed earlier in Section 3.2.1), mis-spellings are seen as falling into four kinds, *omission* of necessary letter(s), *addition* of unnecessary letter(s), *substitution* of correct letter(s) with incorrect one(s) and *transposition* between two adjacent letters – which are either highly implausible or actually impossible in the context of single letter words. For the grammatical words *'I'* and *'a'*, omission would lead us to reclassify this type of error as grammatical because it represents a syntactical error where the noun phrase has been affected (omission of the subject pronoun *'I',* omission of the indefinite article *'a'*). By contrast, the likelihood of adding unnecessary letter(s) to *'I'* and *'a'* seems to be remote since these words are widely used by even the most elementary of English learners. Equally implausible is the substitution of correct letter(s) *'I'* or *'a'* with incorrect ones, while transposition is by definition impossible as it requires a minimum of two adjacent letters.

With respect to the abbreviated forms, each letter in an abbreviated form is pronounced in the same way as the same letter in the alphabet. For instance, the letter '*T'* in TV is pronounced as /tiː/. Similarly, the letter '*V'* in TV is pronounced as /viː/. Therefore, it is unrealistic to imagine students misspell these letters in these abbreviated forms.

So far, this section has provided us with information about the percentages of spelling and phrase errors, based on the TEC approach. As was observed in this section, the analysis showed that spelling and phrase errors, collectively, represented almost two-thirds of language learners' errors, in the three LEFLL sub-corpora, whereas the other four error categories – verb, prepositional, adjective and adverb phrase errors – constituted the remaining portion (almost one-third). In the level of spelling errors, the findings were unexpected in that they showed that the highest percentages of spelling errors in the three sub-corpora fell between the 4-letter and 7-letter words, and that this started decreasing as the words became longer.

On the other hand, in spite of much information about the percentages of spelling and phrase errors across the LEFLL sub-corpora based on the TEC approach, it is not possible to see the percentages of what the learners have achieved correctly in spelling and phrase categories. Knowing the percentages of what the learners have achieved correctly in each category seems to be important. This may enable us to compare between the frequencies of spelling and phrase errors from another angle by looking at the proportion of what the learners achieved correctly in each error category. Therefore, this is what Thewissen (2012) has tried to solve by proposing what she refers to as the potential occasion analysis approach, as will be discussed in the following section.

5.2 Error Analysis Based on the Potential Occasion Analysis Approach – Thewissen's Approach

As discussed in Section 2.3 above, the potential occasion analysis approach (henceforth POA) was developed by Jennifer Thewissen, a doctoral researcher working under the supervision of Sylviane Granger at Louvain University. Thewissen's approach marks a significant break away from previous approaches to error analysis by 'counting the errors of a particular type out of the number of times it could potentially have been committed' (Thewissen, 2012:140). Thus, to calculate the percentage of spelling errors in a learner corpus, Thewissen called for calculating the number of spelling errors out of the total number of tokens in the corpus. The reason behind using the total number of tokens as a denominator to calculate the percentage of spelling errors in the corpus is that every token in the corpus, according to Thewissen, should be regarded as having the potential for spelling error. Crucially, she argued that this type of calculation provides a better insight not only into the percentages of the words that learners have incorrectly spelled, but also into the percentages of the words that they have correctly spelled.

In case of grammatical errors, the POA approach involves error tagging and part-of-speech tagging the learner corpus in question in order to calculate '…errors of a particular category out of the corresponding part-of-speech category' (Thewissen 2012:36). For instance, to calculate the percentage of singular-plural errors on nouns the learners have produced in a learner corpus, according to Thewissen (2012; 2015), we need to calculate the number of singular-plural errors out of the total number of nouns the learners have used in the corpus.

Consequently, to calculate the percentages of spelling errors produced by the students in the three LEFLL sub-corpora, we first need to calculate the number of tokens in each sub-corpus. This was done by loading each sub-corpus to WordSmith corpus tool to generate word lists and associated statistics.

	Total number of tokens
Year 1 Sub-corpus	20,655
Year 2 Sub-corpus	20,109
Year 3 Sub-corpus	20,002
LEFLL Corpus	60,766

Table 5.5: The total number of tokens in the LEFLL corpus

As Table 5.5 shows, the number of tokens in each sub-corpus is higher than the number of tokens in each sub-corpus that was given in Section 4.1.2.1 earlier (e.g., there were 20045 tokens in the Year 1 sub-corpus in Section 4.1.2.1). This discrepancy is due to differences in how words are defined and counted in different software packages (in Section 4.1.2.1, Microsoft Word files were used to calculate the total number of words). However, this discrepancy was minor (only 1.44%).

With respect to phrase errors, we first need to calculate the number of constituents that form each phrase type. Each phrase consists of one or more constituents. For instance, the prepositional phrase '*in the morning'* in the following example:

> *I weak up <<in>> the morrning*

consists of one constituent of the prepositional phrase *'in'* that has the potential for prepositional phrase error whereas the remaining part of the phrase *'the morning'* is a noun phrase. The verb phrase *'have been drinking'* in:

> *people can cause traffic accidents by driving After they <<have been drinking>> alcohel a lot*

consists of three verb phrase constituents (*have, been* & *drinking*), each of which has the potential for verb phrase error.

During the manual annotation of the LEFLL corpus files, each phrase constituent was given a potential for error tagging code that consists of two levels. The first level represents the phrase type whereas the second level denotes the abbreviation of the word potential 'POTE'. For example, the tagging code 'VPPOTE' signifies the constituent in a verb phrase that has the potential for error and it consists of two levels 'VP' and 'POTE'. The first level 'VP' refers to the verb phrase and the second level 'POTE' refers to the constituent that has the potential for error. Similarly, the tagging code 'NPPOTE' marks the constituent in a noun phrase that has the potential for error and it consists of the two levels 'NP' and 'POTE'. Table 5.6 below shows the distributions of the number of phrases and constituents that have the potential for errors in the LEFLL sub-corpora.

Phrase Type	Year 1 Sub-corpus		Year 2 Sub-corpus		Year 3 Sub-corpus		LEFLL Corpus	
	No. of phrases	No. of POTE	No. of phrases	No. of POTE	No. of phrases	No. of POTE	No. of phrases	No. of POTE
NP	7114	11024	6478	11173	6194	9920	19816	32117
VP	3595	4931	2840	3574	2899	4297	9334	12802
PP	2008	2012	1934	2166	1990	2137	5932	6315
ADJP	718	723	1097	1219	1265	1400	3080	3342
ADVP	691	816	776	876	1075	1205	2542	2897
Total	14126	19506	13125	19008	13423	18959	40674	57473

Table 5.6: Phrases and the constituents of phrases that have potential for errors in the LEFLL corpus

After calculating the total number of tokens in each LEFLL sub-corpus, the percentage of spelling errors in each LEFLL sub-corpus was calculated out of the total number of tokens in the same LEFLL sub-corpus as follows:

$$\text{The percentages of spelling errors in a LEFLL sub-corpus} = \frac{\text{the total number of spelling errors in the LEFLL sub-corpus}}{\text{the total number of tokens in the LEFLL sub-corpus}} \times 100$$

Correspondingly, the percentages of phrase errors were calculated ‘as ratios of each relevant word class’ (Milton, 2001:46). That is, the number of errors of a specific phrase type (e.g., prepositional phrase errors (PPE)) in a LEFLL sub-corpus was divided by the total number of constituents that form the same phrase (constituents of prepositional phrases (PPPOTE)) in the same LEFLL sub-corpus, as follows:

$$\text{The percentages of PPE in a LEFLL sub-corpus} = \frac{\text{the total number of PPE in the LEFLL sub-corpus}}{\text{the total number of PPPOTE in the LEFLL sub-corpus}} \times 100$$

Figure 5.4 below shows the distribution of the percentages of spelling and phrase errors based on the POA approach in the three LEFLL sub-corpora. Table 5.7 below provides more details about the total number of errors in each error category, the total number of constituents (in case of phrase errors) or the total number of tokens (in case of spelling errors) and the percentages of errors of each error category in the three LEFLL sub-corpora.

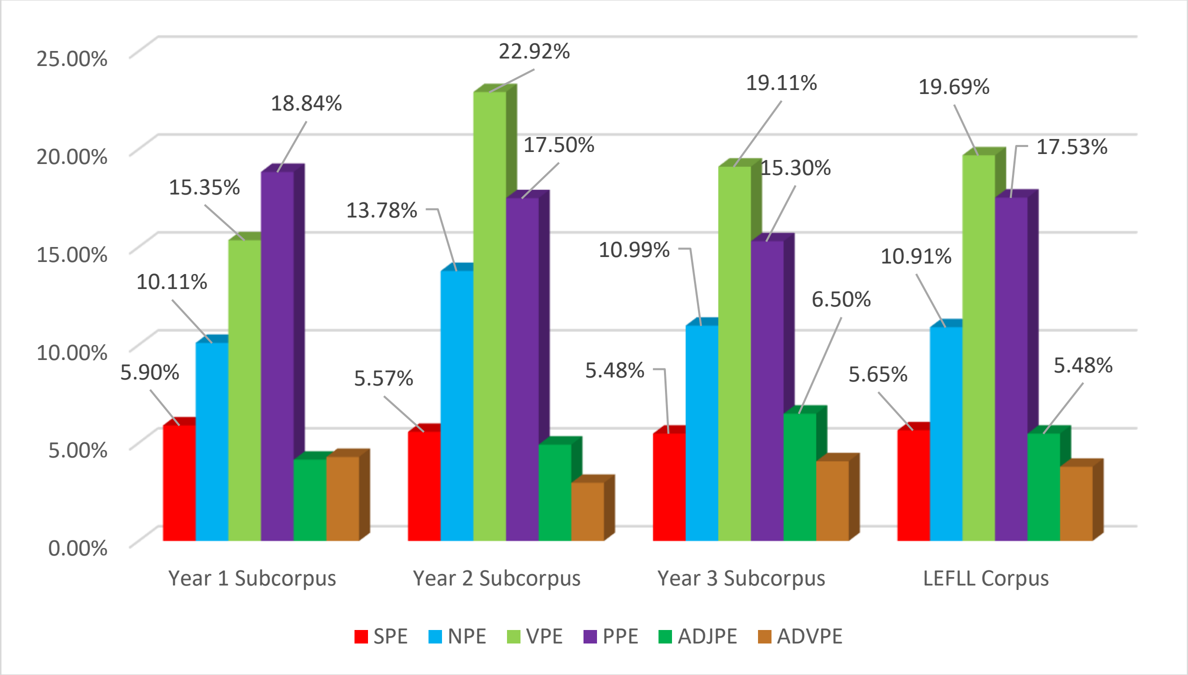


Figure 5.4: The percentages of spelling and phrase errors in the LEFLL corpus – POA approach

Both Figure 5.4 and Table 5.7 provide us with valuable insights. As the Figure and Table show, there is a significant difference between the results obtained via the POA approach and the TEC approach presented earlier in this chapter. Spelling and noun phrase structures do not represent the most significant problem areas for the learners, as they did in the TEC analysis. In fact, the results show that the learners generally performed well in the spelling category.

Error Categories	**Year 1 sub-corpus**			**Year 2 sub-corpus**			**Year 3 Sub-corpus**			**LEFLL Corpus**		
	No. of Errors	**No. of POTE**	**(%) of Errors**	**No. of Errors**	**No. of POTE**	**(%) of Errors**	**No. of Errors**	**No. of POTE**	**(%) of Errors**	**No. of Errors**	**No. of POTE**	**(%) of Errors**
SPE	1218	20655	5.90%	1127	20246	5.57%	1101	20108	5.48%	3446	61009	5.65%
NPE	1115	11024	10.11%	1540	11172	13.78%	1090	9920	10.99%	3745	32116	11.66%
VPE	757	4931	15.35%	819	3574	22.92%	821	4297	19.11%	2397	12802	18.72%
PPE	379	2012	18.84%	379	2166	17.50%	327	2137	15.30%	1085	6315	17.18%
ADJPE	30	723	4.15%	60	1219	4.92%	91	1400	6.50%	181	3342	5.42%
ADVPE	35	816	4.29%	26	876	2.97%	49	1205	4.07%	110	2900	3.79%
Total	3534	40161	8.80%	3951	39253	10.07%	3479	39067	8.91%	10964	118484	9.25%

Table 5.7: The percentages of spelling and phrase errors in the LEFLL corpus – POA approach

The mean percentage of spelling errors has dropped from 31.42% when the percentages of spelling errors were calculated based on the TEC to 5.65% only when they were calculated based on the POA approach. In other words, out of the total number of tokens (61009) that form the LEFLL corpus, only 3446 (5.65%) tokens were misspelled. This means that 57,563 (94.35%) tokens were spelled accurately. There is also a minor but steady improvement in spelling accuracy across the three university levels cohort groups. The percentages of spelling errors start decreasing from 5.90% in year 1 to 5.57% in year 2 to 5.48% in year 3.

The POA approach also finds that the gap between the highest and lowest percentages of errors in any LEFLL sub-corpus has significantly decreased. For instance, when we applied TEC to the Year 2 sub-corpus, noun phrase errors scored the highest percentage of errors (38.95%) whereas the percentage of adverb phrase errors was the lowest (0.66%), which shows a gap of 38.29%. This gap was reduced to 19.95% when we applied POA approach, as the percentage of noun phrase errors dropped to 22.92% while the percentage of adverb phrase errors increased to 2.97%. This may indicate that we have obtained more reliable results as we used a relevant denominator to calculate the percentage of each error category. The high frequency of noun phrase errors compared to adverb phrase errors obtained in Section 5.1 and based on the TEC approach could be attributed to the fact that the chance of committing noun phrase errors is higher than in case of adverb phrase errors. In other words, the total number of constituents that form noun phrases that have the potential for error (11172) is far higher than the constituents that form adverb phrases (876 only).

Unlike the results discussed in Section 5.1 above, the result presented in Figure 5.4 and Table 5.7 suggest that prepositional and verb phrase structures were the major problems encountered by the learners represented in the LEFLL corpus, with adjective phrases also becoming an increasing problem as these learners move from one level to another. The percentages of adjective phrase errors started increasing from 4.15% in the Year 1 sub-corpus to 4.92% in the Year 2 sub-corpus and it is even higher than the percentages of spelling errors in the Year 3 sub-corpus (6.50%). Again, this could also be attributed to the fact that we have used relevant denominators to calculate the percentage of each error category. Figure 5.4 and Table 5.7 also show that some error categories do not follow linear changes across the three university levels and year 2 is still the anomalous stage between year 1 and year 3. This is often the case with noun, verb and adverb phrase categories.

This section has employed the potential occasion analysis (POA) approach proposed by Thewissen (2012; 2015) to recalculate the percentages of spelling and phrase errors produced by the learners sampled by the LEFLL corpus. Interestingly, the results obtained via POA approach were very different from those obtained previously using TEC. In particular, the potential for error approach results indicated that spelling and noun phrase errors were not a major area of difficulty for the learners represented in LEFLL, and that the learners achieved steady progress in these categories overall. This is strikingly different from the TEC results, which found spelling and noun phrase errors to be the most problematic linguistic categories, with no clear linear development from year to year.

The distinctive feature of POA approach is that it offers the percentages for both what the language learners performed accurately and inaccurately. Subsequently, this may help to verify at which stage the language learners are in acquiring specific linguistic categories (e.g., spelling, tenses, prepositions, etc.). On these terms, it would seem that the perspective offered by the POA approach is likely to be more realistic and reliable than that provided by TEC.

However, there is one major limitation of the POA approach that casts substantial doubt on the validity of this approach as well. Specifically, it is based on the theoretical assumption that every token in the corpus has the potential for spelling error. Although this assumption is theoretically valid, it is not borne out by empirical observation. First of all, and as discussed previously, it makes little or no sense in practice to regard single letter words such as ‘a’ and ‘I’ as carrying the same potential for error as words such as ‘manoeuvre’ or ‘rhythm’ – or even has having any potential for spelling error at all. Furthermore, the assumption of equal potential for error does not consider the fact that learner interlanguage development is a dynamic rather than a static phenomenon. According to our POA results, there is an obvious progress across the three LEFLL sub-corpora. This was observed in Figure 5.4 and Table 5.7 as there is a small but steady decrease in the percentages of spelling errors as we move to the higher levels. What this means is that we may expect that the *potential* for spelling errors will progressively reduce as the learners progress from year to year. We may also expect that the learners will be more likely to misspell specific word types at different points in their development. For instance, it is not expected that advanced English learners will misspell short words, such as the 2-letter words, e.g., *on, in, an, etc.* the 3-letter words, e.g., *you, for, man, etc.* Therefore, for more accurate results, it is necessary to verify which word types the language learners have actually misspelled rather than assuming that they are likely to misspell any word as POA claims. In short, while there will remain the potential for error throughout the stages of a learner’s development, it would be wrong to assume (as POA seems to do) that this potential will remain the same at different stages. In other words, Thewissen’s approach rests on the faulty assumption that ‘potential for error’ can be conceptualised as a static and absolute baseline.

One further limitation of the POA approach is its reliance on automatic part-of-speech tagging. Once a corpus has been automatically tagged, the percentage of each type of error (e.g., the percentage of prepositional phrase errors) can be calculated for a given part-of-speech category (e.g., total number of prepositions). While this is of course a major boost in terms of speed and efficiency of processing, the major problem with part-of-speech taggers is that they have been developed on the basis of a native corpus and are affected by the errors produced by the language learners in the learner corpora (Granger, 2008; 2009). Subsequently, this may affect the results obtained, especially in low level learner corpora such as those studied by the current thesis.

Taking these criticisms into consideration, in the following section, the percentages of spelling and phrase error categories will be recalculated based on a new approach, which I have termed the *potential for error* approach. As we will see, this approach does not reject Thewissen's approach entirely, but instead proposes to build on this approach by correcting its shortcomings as identified above.

5.3 Error Analysis Based on the Potential for Error Counting Approach

The potential for error counting approach (henceforth PFEC) emerged from the empirical investigations reported in this thesis, and in particular from the process of manually annotating the LEFLL corpus for spelling and phrase errors. With respect to spelling errors and as shown in Figure 5.3 and Table 5.3 above, this analysis revealed that the learners misspelled words of different lengths starting from the 2-letter words upwards; that is, none of the 1-letter words were misspelled at all. In fact, there are good reasons to regard these 1-letter words as effectively impossible to misspell, as was argued in detail in Section 5.1. Therefore, to increase the accuracy of calculating the percentages of spelling errors based on the PFEC approach, this thesis proposes that the total number of 1-letter words should be excluded from the total number of tokens in the LEFLL corpus. The focus should be restricted to only those words that have an *empirically demonstrable* potential for spelling error. To perform this task, the LEFLL corpus files were reviewed to correct all misspelled words. Then, the corrected files were uploaded to WordSmith Tools to retrieve the statistics of the correct misspelled words in the LEFLL corpus.

In addition to the different word lengths, as shown in Table 5.8 below, a review of the LEFLL corpus showed that the corpus contains cardinal numbers (such as: 12, 1990, 24, etc) which are also included in the total number of tokens. They constitute different percentages out of the total tokens in each sub-corpus in the LEFLL corpus. As stated above, to calculate the percentages of spelling errors based on the PFEC approach, this approach calls to restrict our focus on the words that genuinely (i.e., empirically) have the potential for spelling error. Since these cardinal numbers were written as numbers rather than words, they are impossible to misspell but will nevertheless affect the accuracy of the percentages of spelling errors. Therefore, to obtain more accurate results they should also be excluded from the total number of tokens.

Word Length	Year 1 Sub-corpus			Year 2 Sub-corpus			Year 3 Sub-corpus		
	Word Count	SP	%	Word Count	SP	%	Word Count	SP	%
Numbers	588	0	0.00%	54	0	0.00%	45	0	0.00%
1-letter words	2554	0	0.00%	1308	0	0.00%	750	0	0.00%
2-letter words	4828	8	0.17%	4499	9	0.20%	3535	7	0.20%
3-letter words	2966	38	1.28%	4348	68	1.56%	4082	35	0.86%
4-letter words	3669	225	6.13%	4230	153	3.62%	3850	115	2.99%
5-letter words	2147	143	6.66%	2428	218	8.98%	2146	173	8.06%
6-letter words	1356	188	13.86%	1312	206	15.70%	1620	127	7.84%
7-letter words	1453	227	15.62%	1014	166	16.37%	1791	208	11.61%
8-letter words	813	133	16.36%	518	127	24.52%	977	127	13.00%
9-letter words	336	97	28.87%	312	99	31.73%	655	113	17.25%
10-letter words	309	124	40.13%	107	36	33.64%	417	117	28.06%
11-letter words	211	29	13.74%	91	37	40.66%	154	47	30.52%
12-letter words	5	5	100%	6	2	33.33%	67	8	11.94%
13-letter words	9	1	11.11%	13	6	46.15%	43	14	32.56%
14-letter words	2	0	0.00%	1	0	0.00%	13	6	46.15%
15-letter words	0	0	0.00%	0	0	0.00%	8	4	50%
Total	18129	1218	6.72%	18905	1127	5.96%	19358	1101	5.69%

Table 5.8: The percentages of spelling errors per word length in the LEFLL corpus – PFEC approach

As the Table shows, the total number of tokens in each sub-corpus declined in each sub-corpus after excluding the total number of cardinal numbers and 1-letter words from the total number of tokens. For example, the total number of tokens (that empirically have the potential for spelling error) in the Year 1 sub-corpus declined from 21242 to 18129 tokens. I would argue that these results constitute a more valid and meaningful set of percentages of spelling errors per word length across the LEFLL sub-corpora than was provided by either TEC or POA.

Unlike the results obtained via TEC, the results obtained via the PFEC approach show that there is almost a linear relationship between the word-lengths and the percentages of spelling errors. The percentages of spelling errors steadily increased as the words became longer as Table 5.9 above and Figure 5.5 below show. But for the longest words, e.g., 14-letter words (such as *responsibility*) and 15-letter words (such as *characteristics*), the percentages of spelling errors decreased. While it is difficult to explain this trend, it is worth noting that these words are very infrequent in the LEFLL data; for example, there are only nine 14-letter words. This invites the speculation that these words are spelled correctly because they are the only long words that the learners are confident about using in their writing. In other words, these words may be examples of what Hasselgren (1994) has influentially termed ‘lexical teddy bears’.

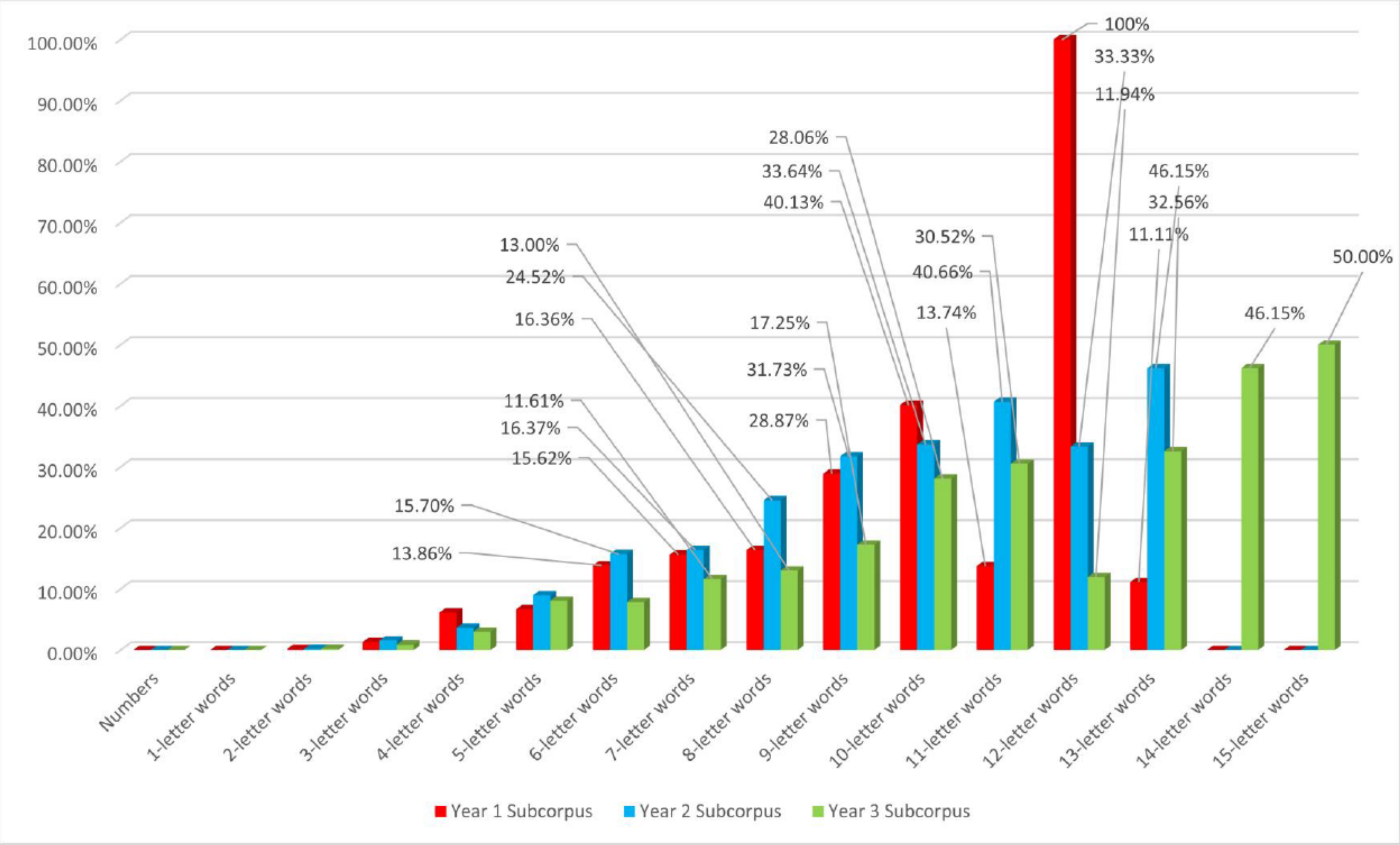
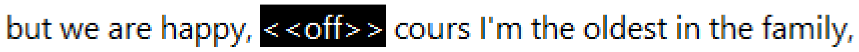


Figure 5.5: The Percentages of spelling errors per word length in the LEFLL corpus – PFEC approach

The results presented in Table 5.8 above may provide an explanation for the unexpected results presented earlier in Figure 5.3 and Table 5.3. The reason why the highest percentages of spelling errors fall between the 4-letter words and 7-letter words, when calculating via TEC, is that the total number of words per word length up to the 7-letter words is far higher than the total number of words per word length from 8-letter words onwards. For instance, in the Year 1 sub-corpus, the total number of 7-letter words is 1457, whereas in the case of 8-letter words there are 813 words only. As argued previously, it makes more sense here to assume that the likelihood for spelling errors in a set of 1457 words is likely to be higher than it is in a set of 813 words, than it is to assume equal potential for error in both cases, as would be the assumption in Thewissen's approach.

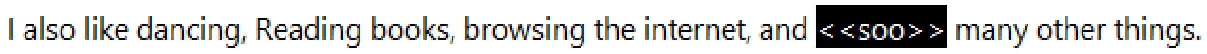
Another important finding in Table 5.8 that should be addressed from the outset concerns the limited numbers of spelling errors of 2-letter words in the three sub-corpora. The misspelled 2-letters words are very common and widely used by both the language learners and native speakers (e.g., *of*, *so, pm* and *us*). Consider the following examples:

'off' for *'of'*



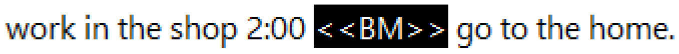
but we are happy, <<off>> cours I'm the oldest in the family,

'soo' for *'so'*



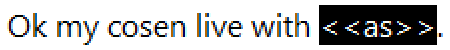
I also like dancing, Reading books, browsing the internet, and <<soo>> many other things.

'BM' for *'PM'*



work in the shop 2:00 <<BM>> go to the home.

'as' for *'us'*

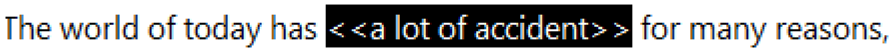


Ok my cosen live with <<as>>.

Since they are simple (consisting as they do of 2 letters only), commonly used and typically spelled correctly in the LEFLL corpus, the likelihood of observing an equal number of spelling errors in 2-letter words in an elementary level and an advanced level learner corpus (or even any errors in an advanced level corpus at all) seems to be very remote indeed in practice.

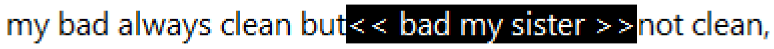
The Table also shows that the lowest percentage of spelling errors in the 3-letter words among the Year 3 undergraduate students was 0.86% only. Thus, it might not be possible to observe any spelling errors among the 3-letter words attested in advanced learner corpora. This means that as the learners move from one level to another, the minimum and/or average word lengths for misspelled words may increase. Thus, advanced learners may be more liable to misspell longer and less commonly used words than learners at low levels. This certainly seems to be the case in the LEFLL corpus. This in turn casts further doubt on Thewissen's claim that every token in a learner corpus has the potential for spelling error.

With respect to phrase errors, in contrast, the empirical findings reported above do support Thewissen (2012; 2015)'s assumption that each constituent of a phrase (e.g., determiner, pronoun, … in a noun phrase; a main verb, an auxiliary, … in a verb phrase, etc.) in a learner corpus has the potential for error, and that each erroneous phrase may contain more than one error. For instance, the highlighted noun phrase '*a lot of accident*' in the following example contains one noun phrase error only (number agreement error):



The world of today has <<a lot of accident>> for many reasons,

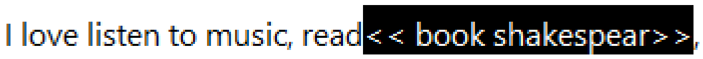
Whereas the noun phrase in following example:



my bad always clean but<< bad my sister >>not clean, Note: ‘*bad’* is the misspelled word for ‘*bed’*

contains two noun phrase errors*:* wrong word order and apostrophe omission errors. Thus, the correct noun phrase structure will be: *my sister’s bed*

It was observed that the maximum number of errors the learners have produced in any phrase type (noun, verb, preposition, etc.) in the LEFLL sub-corpora was less than or equal to the number of constituents that form the phrase when it is written in its correct structure. For instance, one of the possible correct structures for the highlighted incorrect noun phrase in the example below is ‘*the shakespear’s book’* which consists of three constituents if we exclude the apostrophe marker: *‘the’, ‘Shakespeare’s’* and ‘*book’*.



I love listen to music, read< < book shakespear> >,

In this example, there are three noun phrase errors: *omission of determiner* (e.g., *the, my*)*, wrong word order* and *omission of apostrophe* errors*.*

Since the corpus investigation proved that each constituent in a phrase has the potential for error, the total number of constituents of each phrase (e.g., noun phrase, verb phrase, etc.) was taken as the denominator to calculate the percentages of errors of each phrase type (e.g., noun phrase errors, verb phrase errors, etc.). Thus, Figure 5.6 and Table 5.9 provide a comparison between spelling and phrase errors based on the PFEC approach.

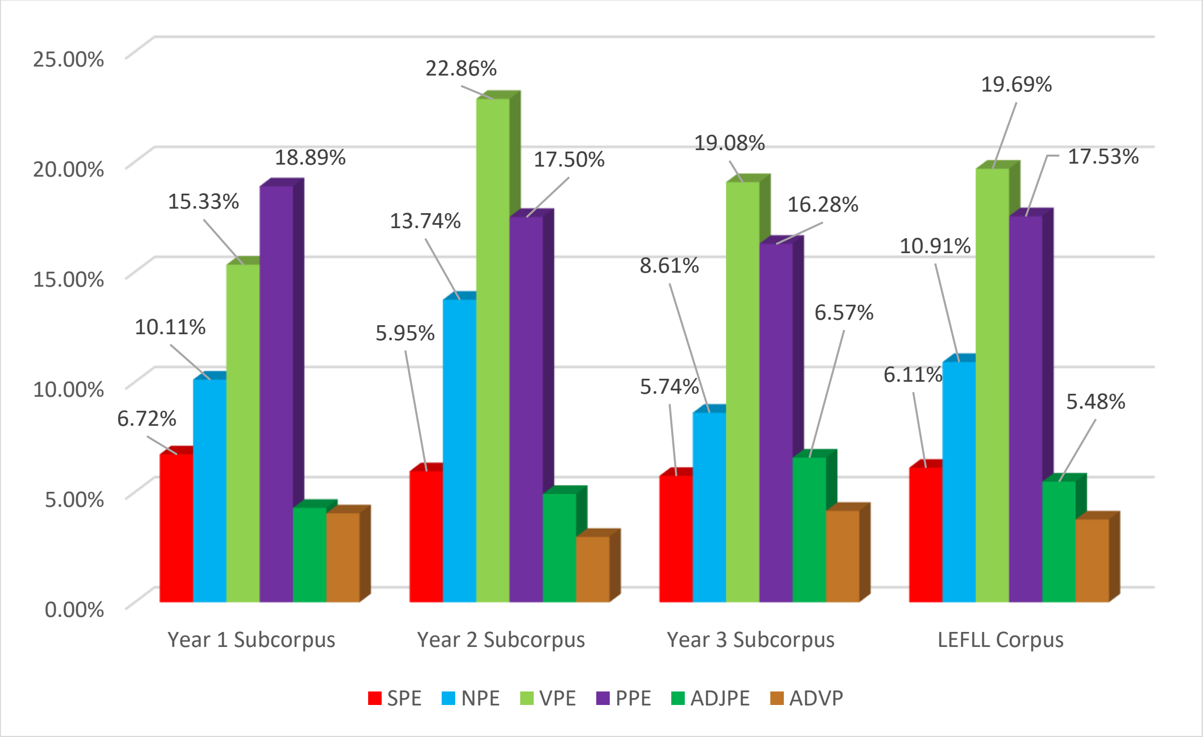


Figure 5.6: The percentages of spelling and phrase errors in the LEFLL corpus – PFEC approach

By comparing the results obtained via the three approaches to error counting presented in this chapter - TEC, POA and PFEC - it is evident that TEC was not able to provide a clear picture of which linguistic category or categories (e.g., spelling, noun phrases, verb phrases, etc.) are more problematic to the learners in the LEFLL corpus. TEC incorrectly showed that almost two-thirds of errors produced by the learners across the three LEFLL sub-corpora were spelling and noun phrase errors.

Error Categories	Year 1 sub-corpus			Year 2 sub-corpus			Year 3 Sub-corpus			LEFLL Corpus		
	No. of Errors	No. of POTE	(%) of Errors	No. of Errors	No. of POTE	(%) of Errors	No. of Errors	No. of POTE	(%) of Errors	No. of Errors	No. of POTE	(%) of Errors
SPE	1219	18129	6.72%	1125	18905	5.95%	1102	19358	5.74%	3446	56392	6.11%
NPE	1115	11024	10.11%	1535	11172	13.74%	854	9920	8.61%	3504	32116	10.91%
VPE	756	4931	15.33%	817	3574	22.86%	820	4297	19.08%	2393	12802	19.69%
PPE	380	2012	18.89%	379	2166	17.50%	348	2137	16.28%	1107	6315	17.53%
ADJPE	31	723	4.29%	60	1219	4.92%	92	1400	6.57%	183	3342	5.48%
ADVPE	33	816	4.04%	26	876	2.97%	50	1205	4.15%	109	2900	3.76%
Total	3534	37635	9.39%	3942	37912	10.40%	3266	38317	8.52%	10742	113867	9.43%

Table 5.9: The percentages of spelling and phrase errors in the LEFLL corpus – PFEC approach

This issue was resolved by the POA approach. The POA approach provided better insight regarding the percentages of different error categories. It offered both percentages: what the language learners performed inaccurately and accurately. Subsequently, this has helped to verify which linguistic categories may constitute major problems to the English learners, as was observed in Figure 5.6 and Table 5.9 above. However, the major issue with the POA approach is that it is based on the empirically dubious assumption that each token in the corpus has the potential for error. Thus, it calls for the counting of spelling errors, for example, out of the total number of tokens in the corpus.

This issue has been taken into consideration in the PFEC approach, where the LEFLL corpus was manually annotated for error analysis. In the case of spelling errors, the analysis revealed that some tokens in the LEFLL corpus do not empirically have the potential for spelling errors (e.g., cardinal numbers and 1-letter words). Instead, they affect the accuracy of the percentages of spelling errors if they are not excluded from the total number of tokens before calculating the percentages of spelling errors. It was therefore argued that they should be excluded from the total number of tokens before calculating the percentages of spelling errors.

As argued earlier in this section, excluding the total number of cardinal numbers and 1-letter words leads to a more valid set of percentage results than was obtained via Thewissen's approach. As was observed, this has changed the overall results once again. For instance, the overall percentage of spelling errors in the LEFLL corpus was 5.65% (almost the same rate as the adjective phrase errors 5.48%) when calculating via the POA approach as shown earlier in Figure 5.4 and Table 5.7. This percentage has increased to 6.11% when the PFEC approach has been applied, as shown in Figure 5.6 and Table 5.9. If this is the case in a small sized corpus such as LEFLL, such variation could be even higher in larger learner corpora. There would therefore seem to be a definite need for empirical investigation in the learner corpus to verify the tokens that realistically have the potential for spelling error. This concept could also be applied when calculating the percentages of other types of language learners' errors. Empirical investigation in the learner corpora may help to ensure that the results obtained are reliable.

Conclusion

This chapter has compared the three error counting approaches: the traditional error counting (TEC) approach, the potential occasion analysis approach (POA), and the potential for error counting approach (PFEC). These three approaches have been employed to calculate the percentages of spelling and phrase errors identified in the LEFLL corpus, and the results of each approach have been critically evaluated.

The TEC analysis suggested that the learners encountered major problems with spelling and noun phrase structures. The results showed that spelling and phrase errors, collectively, represented almost two-thirds of language learners' errors, in the three LEFLL sub-corpora, whereas the other four error categories – verb, prepositional, adjective and adverb phrase errors – constituted the remaining portion (almost one-third). The major problem with this approach is it ignores calculating the percentage of each error within a relevant error environment (e.g., tense errors within verb phrases, preposition errors within prepositional phrases, etc.). Therefore, it is not possible to see what percentage of each linguistic category (e.g., spelling, prepositions, adjectives, etc.) was performed accurately by the learners.

The POA approach provided a strikingly different set of results to those obtained from TEC. The verb and prepositional phrase errors scored the highest percentages compared to the other error categories (spelling, noun, adjective and adverb phrase errors). The POA approach offered the percentages for both what the learners performed accurately and inaccurately in the spelling and phrase categories. However, major problems were also identified with this approach. Firstly, it is based on the dubious theoretical assumption that each constituent (e.g., determiner in a noun phrase, a word token in the corpus) has the same potential for error but does not provide empirical evidence from the learner corpus under investigation to back this assumption up. Secondly, it relies on automatic part-of-speech taggers which have been developed on the basis on a native corpus and are affected by errors in learner corpora.

It was argued that the PFEC approach provided more accurate results and addressed the major problems with the POA approach. The former approach called for empirical investigation in the learner corpus to verify which constituent (e.g., token for spelling, a determiner in a noun phrase, auxiliary in a verb phrase, etc.) has the *empirically demonstrable* potential for error, rather than relying on the theoretical assumption that each constituent has the potential for error as the POA claims. The major problem with the PFEC approach is it is based on manual annotation of the learner corpus, and is therefore very laborious and time-consuming, as well as potentially prone to human coding errors. (The solution to this latter problem is to carry out an inter-rater reliability test, as was discussed and carried out in Chapter 4 of this thesis.)

Having achieved the second aim of this thesis, comparing the three error counting approaches to illustrate the advantage of the PFEC approach, proposed by this thesis, over the other two error counting approach, I will now move on to address the main aim of this thesis. As stated earlier in the introductory chapter, the main aim of this thesis is to provide the researchers of SLA with a broader picture of the role of L1 Arabic in the writing of Arab English learners by conducting a comprehensive and quasi-longitudinal analysis of spelling, noun phrase and verb phrase errors identified in the LEFLL sub-corpora. First, the analysis will be conducted on the spelling errors that have been identified in the LEFLL sub-corpora as we will see in Chapter 6.

CHAPTER 6: ANALYSIS OF SPELLING ERROR TYPES IN THE LEFLL CORPUS

Introduction

Spelling is a fundamentally important language skill that spans both writing and reading skills (Zhao et al., 2016). However, and as discussed in the chapter 3, despite its obvious importance as a language skill and its pervasiveness in learner data, spelling has received surprisingly little research attention in the learner corpus research. Therefore, this chapter seeks to remedy this deficiency by conducting a comprehensive and quasi-longitudinal analysis of spelling errors identified in the LEFLL sub-corpora. To perform this task, spelling errors were tagged and classified based on the *surface structure taxonomy* (as briefly presented in chapter 3, Section 3.2.1 and will further be discussed in Section 6.1) to describe the four types of non-linguistic alteration described in Chapter 3 (i.e., *omission, addition, substitution* and *transposition*). Following the error tagging process, the error tagging codes were then retrieved using AntConc. The spelling errors were then analysed and classified based on the linguistic category taxonomy developed by Rimrott & Heift (2005; 2008), as presented earlier in Section 3.2.2 (and will further be discussed in Section 6.2 below). As established in Chapter 3, Rimrott & Heift's spelling error classification system has four dimensions: a) *linguistic subsystem taxonomy*, b) *linguistic competence taxonomy*, c) *language influence taxonomy,* and d) *target modification taxonomy*.

In the current research, the linguistic subsystem analysis examined spelling errors linguistically by looking at the phonological, orthographical and morphological variations and the relationship between these variations and the phonological, orthographical and morphological routines in Arabic, the mother language of the language learners represented in the LEFLL corpus. The linguistic competence analysis classified spelling errors into performance and competence spelling error categories. This phase of the analysis aimed to identify and retrieve spelling errors that were produced due to target language knowledge deficiencies, and to ignore any accidental spelling errors. The language influence analysis aimed to classify the competence spelling errors retrieved from the linguistic competence analysis into interlingual spelling errors*,* which represent the language transfer phenomenon, and intralingual spelling errors, which reflect the target language development, as the learners move from one year of study to another. Finally, the target modification analysis was used to classify spelling errors into one-edit distance, two-edit distance and multiple-edit distance based on the alterations needed (i.e., omission, addition, substitution and transposition) to change the misspelled words to the correctly spelled ones. For reliability purposes, the results obtained from the target modification taxonomy were used to support the results obtained from both linguistic competence and language influence dimensions.

The analysis showed that omission and substitution spelling errors constituted the highest percentages of spelling errors out of all types of spelling errors (omission, addition, substitution and transposition) across the three LEFLL sub-corpora. The analysis also revealed that the average percentage of interlingual spelling errors in the LEFLL corpus is higher than the percentage of intralingual spelling errors (68.96% vs 31.04% respectively). However, there is a steady decrease in the percentages of interlingual spelling errors from one level to another.

6.1 Spelling Errors Based on the Surface Structure Taxonomy

As explained in Chapter 3, the aim of applying the surface structure taxonomy is to provide more information about the non-linguistic alterations that affected the misspelled words in the LEFLL corpus. Knowing the non-linguistic alterations that affected the misspelled words is important as they help to classify spelling errors based on the spelling errors' classification system developed by Rimrott & Heift (2005; 2008). They help to categorize spelling errors, based on the linguistic competency taxonomy, into performance and competence spelling errors. They also facilitate the categorization of spelling errors based on the target modification taxonomy, which classifies the misspelled words according to the surface structure changes (i.e., *omission, insertion/addition, substitution* and *transposition*) needed to change them into correctly spelled words. Therefore, the surface structure taxonomy provides '…a starting point for comparison purposes' (Cook 1997:479). The analysis shows that, structurally, spelling errors reflected the following types of alterations:

1) The omission of a letter, e.g., *anther* (for *another*)
2) The insertion/addition of a letter, e.g., *theire* (for *their*)
3) The substitution of a correct letter with incorrect one, e.g., *conclosion* (for *conclusion*)
4) The transposition between two adjacent letters, e.g., *thier* (for their).

The analysis shows that the percentages of these surface structure changes varied from one type of alteration to another among the learners in the three sub-corpora, as can be seen in Figure 6.1 and Table 6.1 below.

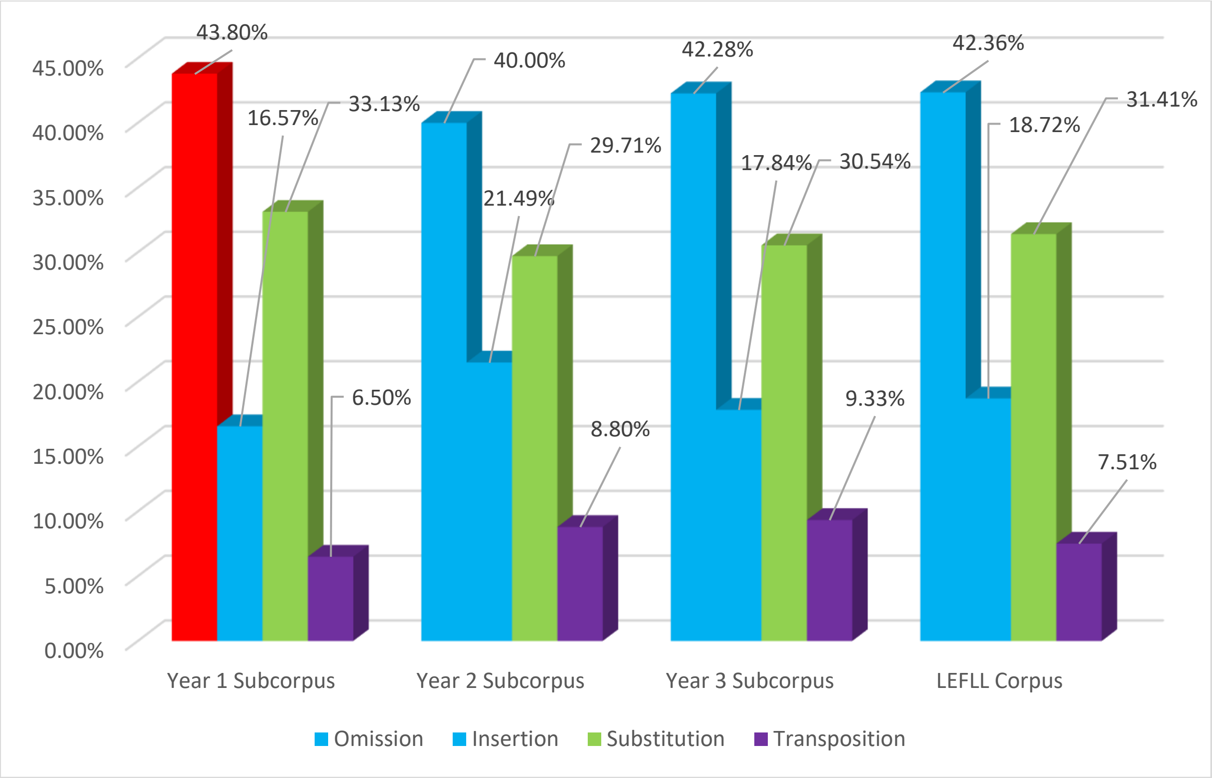


Figure 6.1: The percentages of spelling errors in the LEFLL corpus based on the surface structure taxonomy

Figure 6.1 above and Table 6.1 below show that almost two-thirds of spelling errors in the three sub-corpora omit a necessary letter and substitute a correct letter with an incorrect one. In fact, spelling errors that were produced as a result of omitting a necessary letter constituted the predominant phenomena, at an average of 42.36% out of the total number of all types of alterations that affected the misspelled words. The analysis revealed that omission affected both consonants and vowels, as the two examples below show:

then I <<chang>> dressed and going to the university with friends (the omission of ‘e’)

because the <<goverment>> is weak (the omission of ‘n’)

However, the rate of vowel omission spelling errors was much higher than that of consonants. This could be due to the language transfer from the learners’ first language (i.e., interlingual spelling errors) or incomplete knowledge or wrong application of the target language spelling rules (i.e., intralingual spelling errors). Later in this chapter we will see more clearly the likely causes of the high frequencies of omission and substitution spelling errors.

The Non-linguistic Alterations of Spelling Errors	Year 1 Sub-corpus	Year 2 Sub-corpus	Year 3 Sub-corpus	LEFLL Corpus
Omission	(n=735) 43.80%	(n=618) 40.00%	(n=616) 42.28%	(n=1969) 42.36%
Insertion	(n=278) 16.57%	(n=332) 21.49%	(n=260) 17.84%	(n=870) 18.72%
Substitution	(n=556) 33.13%	(n=459) 29.71%	(n=445) 30.54%	(n=1460) 31.41%
Transposition	(n=109) 6.50%	(n=104) 8.80%	(n=136) 9.33%	(n=349) 7.51%
Total	1678	1513	1457	4648

Table 6.1: The percentages of spelling errors in the LEFLL corpus based on the surface structure taxonomy

Overall, it seems that the learners encountered major problems with phonetics, especially with vowels. This is evidenced by the high rate of spelling errors of the substitution type, as shown in Figure 6.1 and Table 6.1 above. The students substituted the correct consonants and vowels with incorrect ones at an average rate of 31.41%, and a high rate of substitution spelling errors was observed in substituting the correct vowels with incorrect ones, e.g.,

After that, I go to <<cafetiria>> and Eating a breakFast at 10:30 (substituting '*e*' with '*i*')

Across the university courses and as we move across the three university levels, Year 2 remains an anomalous case as was observed in Chapter 5. This is the case in with errors of Omission, Insertion/addition and substitution. While the percentage of omission spelling errors decreased from 43.80% in the Year 1 sub-corpus to 40.00% in the Year 2 sub-corpus before increasing to 42.28% in the Year 3 sub-corpus, the percentages of insertion spelling errors showed a considerable increase of more than 4.5% as we move from year 1 to year 2 followed by a considerable decrease of more than 3.5% in year 3.

Generally, by comparing the results obtained from the Year 1 sub-corpus with the Year 3 sub-corpus, we can see that the percentages of omission and substitution spelling errors decreased as the learners reached the advanced level, e.g., spelling errors as a result of substitution decreased by almost 3% between Year 1 and Year 3. On the other hand, there is a considerable increase in the percentages of spelling errors due to insertion/addition and transposition alterations, e.g., the percentage of insertion/addition spelling errors increased by almost 1%.

The high percentages of omission and substitution spelling errors may reflect a systematic application of omission and substitution routines by the learners represented in the LEFLL corpus. This may be regarded as a systematic application in the sense that they are caused either due to the development of the language learning process, where the learners may have produced such types of spelling errors due to this development such as overgeneralization of phonological, orthographical or morphological spelling rules acquired during the English learning process, or by transferring some linguistic systems from Arabic to English. On the other hand, the low percentages of insertion and transposition spelling errors may be attributed to accidental misspellings which may not reflect any systematic language knowledge deficiency. These could therefore be marked as performance spelling errors. This interpretation is supported by the slight increase in the percentages of insertion and transposition spelling errors particularly among the first- and third-year students, which could reflect the exam conditions under which the texts were written, and the word count that the learners were required to reach by the task rubric in this case.

Nevertheless, it must be acknowledged that these are still just speculations and interpretations, and it is uncertain whether the high increase in certain types of non-linguistic alterations that affected the misspelled words may indicate more reliance on the first language spelling norms or be merely due to the incorrect use of spelling rules in the target language. Later in this chapter I will provide a stronger explanation regarding the likely causes that led to the high percentages of omission and substitution spelling errors and the low percentages of insertion and transposition spelling errors.

Having classified spelling errors, based on the *surface structure taxonomy*, and discussed the non-linguistic alterations that affected the misspelled words in the LEFLL corpus, I will now move on to discuss the linguistic alterations that affected the misspelled words. Thus, I will apply the classification system of spelling errors developed by Rimrott & Heift (2005; 2008), which, as discussed earlier in Section 3.2.2, consists of four dimensions: the linguistic subsystem taxonomy, the linguistic competence taxonomy, the language influence taxonomy, and the target modification taxonomy. The results for each of these dimensions will now be discussed in turn.

6.2 Spelling Errors Based on the Linguistic Subsystem Taxonomy

In the linguistic subsystem taxonomy, spelling errors were classified according to the standard POMAS (Phonological, Orthographical and Morphological Assessment of Spelling) spelling error categorization scheme (Silliman et al., 2006; Wood & Connelly, 2009; Bahr et al., 2012; Bahr et al. 2015). The analysis revealed that the spelling errors identified in the LEFLL corpus can be classified into six linguistic spelling error categories. Three were purely Phonological*,* Orthographical and Morphological, and three were a mix between two of the above linguistic spelling error categories: Phonological – Orthographical*,* Phonological – Morphological and Orthographical – Morphological*.* This section will discuss and provide examples of each type of

POMAS spelling error categories using the non-linguistic alterations discussed in Section 6.1 above.

1) *Phonological Spelling Errors*: This category describes the misspelled words where one or more consonant or vowel sound(s) is/are omitted or inserted, e.g., *drnk* (for *drink*), *eightee* (for *eighteen*)
2) *Orthographical Spelling Errors*: This category describes misspelled words that are affected orthographically, where a silent letter, such as *'–e'* at the end of some words, has been removed, e.g., '*graduat'* (for '*graduate'*), a consonant letter that has, mistakenly, been doubled, e.g., '*tallk'* (for *talk*), *a* consonant or vowel sound/letter that has been substituted with another consonant or vowel sound that may have close articulation, such as the substitution between /p/ and /b/ in '*pusy'* (for '*busy'*), /c/ and /s/ in '*nise'* (for '*nice'*).
3) *Morphological Spelling Errors*: This category describes misspelled words that were altered morphologically due to applying incorrect inflection, such as '*bigest'* (for '*biggest'*), incorrect derivation, such as *'finaly'* (for *finally*), or confusion between two words that have close or similar pronunciations, such as *'sit'* (for *set*).
4) *Phonological – Orthographical Spelling Errors*: This category describes misspelled words which contain both phonological and orthographical spelling error categories at the same time. The most common type of phonological – orthographical spelling error is *letter reverses/transposition* (see also (Bahr et al. 2015)), e.g., '*hoilday'* (for *holiday*)
5) *Phonological – Morphological Spelling Errors*: This category describes misspelled words that contain both phonological and morphological spelling error categories at the same time, e.g., *finly* (for *finally*).

   In the example above, for the first partition, phonological spelling error category, the vowel sound 'ə' has been omitted. With respect to the second partition, the morphological spelling error category, the learner has incorrectly derived an adverb from an adjective word class. As a result, the learner has only added *'y'* instead of *'ly'.*

6) *Orthographical – Morphological Spelling Errors*: This category contains a combination of both orthographical and morphological spelling errors, e.g., *regullarry* (for *regularly*).

In this example, for the first partition, orthographical spelling errors, the word contains two instances of orthographical spelling errors. Both the consonant letters '*l*' and '*r*' have been incorrectly doubled. With respect to the second partition, morphological spelling errors, the misspelled word contains derivational spelling error where the writer has only added the suffix '*y*' instead of '*ly'* to derive the adverb '*regularly'* from the adjective word class *'regular'*.

The analysis revealed, as showed in Figure 6.2 and Table 6.2 below, that the phonological and orthographical spelling errors scored the highest percentages both in the LEFLL corpus as a whole and in each of the three sub-corpora. The analysis also revealed that o*rthographical* spelling errors seem to be the predominant spelling error category among most of the learners. Following the orthographical spelling errors, the phonological spelling errors also constituted a prominent phenomenon.

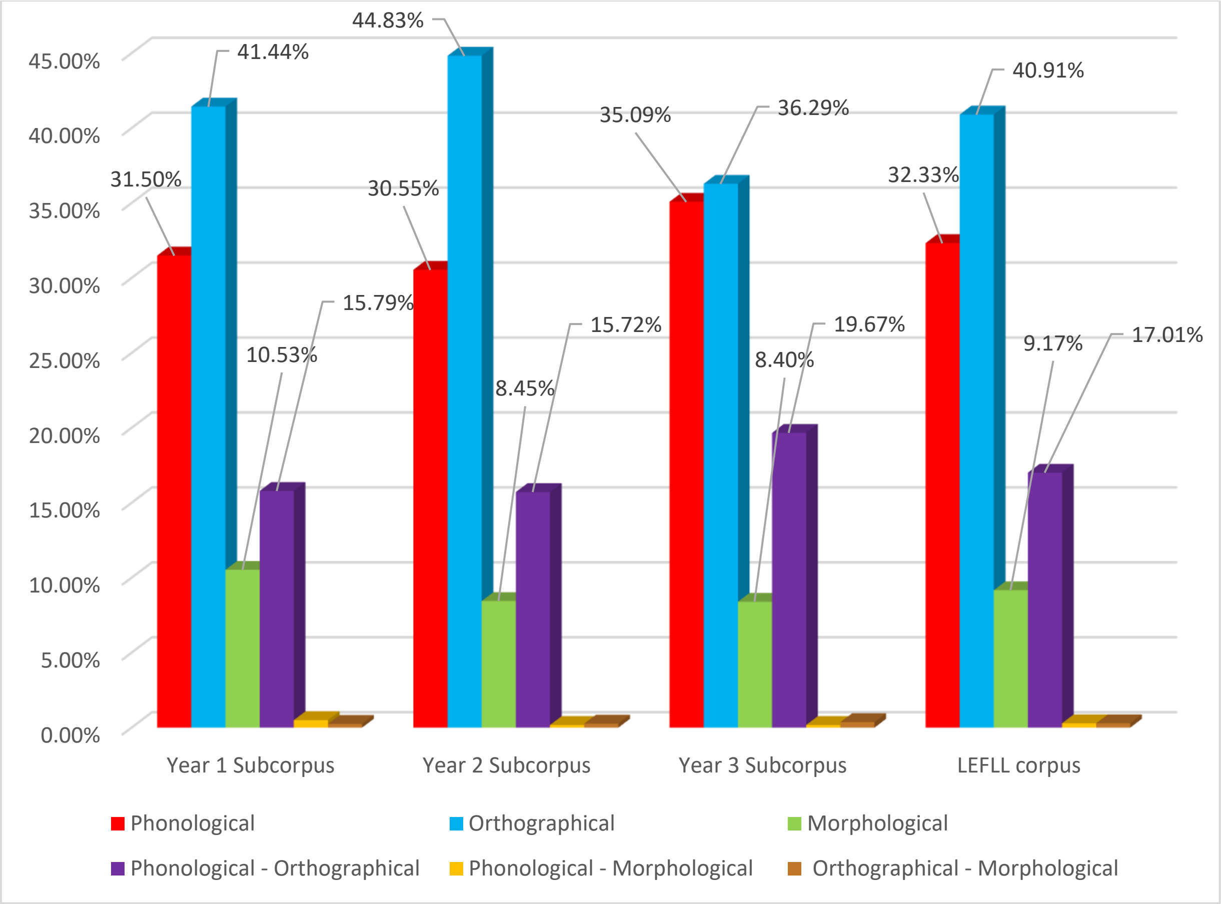


Figure 6.2: The percentages of linguistic spelling error categories in the LEFLL corpus

Interestingly, the gap between orthographical and phonological spelling errors starts decreasing as across the three levels. Whereas orthographical spelling errors were almost 10% higher than phonological spelling errors among the language learners in year 1, whereas in year 3, the percentage of phonological spelling errors were only 1.20% less than the orthographical spelling errors.

The Linguistic Spelling Error Categories	Year 1 Sub-corpus	Year 2 Sub-corpus	Year 3 Sub-corpus	LEFLL Corpus
Phonological	(n=377) 31.50%	(n=340) 30.55%	(n=380) 35.09%	(n=1097) 32.33%
Orthographical	(n=496) 41.44%	(n=499) 44.83%	(n=393) 36.29%	(n=1388) 40.91%
Morphological	(n=126) 10.53%	(n=94) 8.45%	(n=91) 8.40%	(n=311) 9.17%
Phonological-Orthographical	(n=189) 15.79%	(n=175) 15.72%	(n=213) 19.67%	(n=577) 17.01%
Phonological-Morphological	(n=6) 0.50%	(n=2) 0.18%	(n=2) 0.18%	(n=10) 0.29%
Orthographical-Morphological	(n=3) 0.25%	(n=3) 0.27%	(n=4) 0.37%	(n=10) 0.29%
Total (n)	1197	1113	1083	3393

Table 6.2: The percentages of linguistic spelling error categories in the LEFLL corpus

Figure 6.2 and Table 6.2 also show that the percentages of spelling errors that mark the morphological spelling error category (wholly or partially) tend to be low and showed almost no observable change particularly between the learners in years 2 and 3. Again, this may indicate the role of the mother language of the learners (i.e., interlingual spelling errors) or incomplete knowledge of the morphological spelling rules in the target language (i.e., intralingual spelling errors) or a combination of both interlingual and intralingual spelling errors.

The following section investigates each linguistic spelling error category more thoroughly, by breaking down each category into its components (linguistic subcategories). This will be carried out by demonstrating the surface structure changes that affected the misspelled words in the linguistic category domain. The subcategories of each linguistic spelling error category will then assist to classify spelling errors into two groups of spelling errors: interlingual vs intralingual spelling errors.

6.2.1 Phonological Spelling Errors

Based on the general description of the phonological spelling error category given above, eight types of phonological spelling errors were observed in the LEFLL corpus, as follows:

1) *Vowel Omission Spelling Errors*: These errors occur when a single vowel letter (*a,e,i,o* or *u*) is omitted from the target word, e.g., *universty* (for *university*)
2) *Consonant Omission Spelling Errors*: These errors occur when a single consonant letter (any English alphabet except the vowels given above) is omitted from the target word, e.g., *moring* (for *morning*)
3) *Consonant Insertion/addition Spelling Errors*: These errors occur when a single consonant letter is added (anywhere in the misspelled word except at the beginning of the word or between two sounds) to the target word, e.g., *universicty* (for *university*)
4) Epenthesis: Epenthesis, generally, is a well-known phenomenon among the language learners and defined as the addition of a vowel or consonant either at the beginning of the word or between two sounds (Richards and Schmidt 2010). e.g., *returen* (for *return*).
5) *Paragoge Spelling Errors*: These errors occur when a vowel is added at the end of the word which normally ends with a consonant sound (Sadhwani, 2005). This is represented in the addition of '*e*', e.g., *watche* (for *watch*).
6) *Digraph Omission Spelling Errors*: These errors occur when two letters that form one digraph are omitted. The omission of *'gh'* was the only omitted digraph observed in the analysis, e.g., *'throue'* (for *through'*). In this example, there are two phonological subcategory spelling errors. First, the digraph *'gh'* was omitted. Second, the vowel *'e'* was inserted which is marked as paragoge spelling error.

7) *Digraph Insertion Spelling Errors*: These errors occur when two letters that form one digraph are inserted. Similarly, the digraph *'gh'* was the only digraph insertion spelling error type observed in the LEFLL corpus, particularly in the Year 3 sub-corpus, e.g., *'undignifighed'* (for *'undignified'*)

As discussed in Section 6.1 above, the learners more frequently tend to omit a single letter, of both consonant and vowel letter types, from the target word. Figure 6.3 and Table 6.3 below show that the vowel omission errors reached the highest percentage of phonological spelling error category. In the Year 1 sub-corpus, for example, the omission of vowels is almost 37.5% higher than the omission of consonants. This gap declined in the Year 3 sub-corpus to 23.15%. Thus, the percentages of consonants omission spelling errors start increasing as the learners moved to the advanced levels vs a decrease in the percentages of vowel omission spelling errors. This may indicate that the learners rely less on their mother language phonological spelling system.

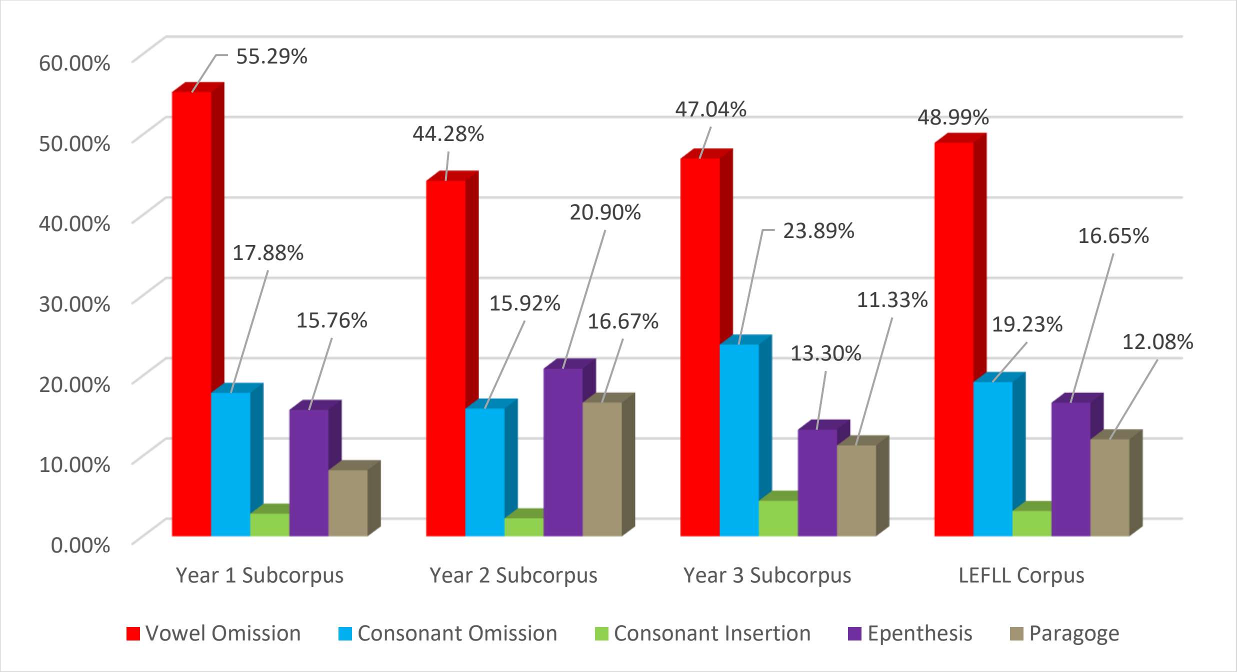


Figure 6.3: The percentages of phonological spelling subcategory errors in the LEFLL corpus

The English and Arabic phonological systems are very different from each other. English has 22 vowels and diphthongs to 24 consonants, whereas Arabic only has eight vowels and diphthongs (three short, three long and two diphthongs) to 32 consonants (Swan & Smith, 2001). Given this dissimilarity, it is inevitable that Arabs have significant problems with English vowel sounds particularly the short vowel sounds (Swan and Smith 2001). This may have led to the learners represented in the LEFLL producing vowel sound spelling errors where they employed the same spelling conventions in their mother language (this will be verified in the Discussion section of this chapter). As Figure 6.3 and Table 6.3 show, *Epenthesis,* the addition of a vowel or consonant either at the beginning of the word or between two sounds is slightly higher than *Paragoge,* the addition of a vowel at the end of the word which normally ends with a consonant, in the LEFLL corpus and across the three LEFLL sub-corpora.

Phonological Spelling Error Subcategories	**Year 1 Sub-corpus**	**Year 2 Sub-corpus**	**Year 3 Sub-corpus**	**LEFLL CORPUS**
Vowel Omission	(n=235) 55.29%	(n=178) 44.28%	(n=191) 47.04%	(n=604) 48.99%
Consonant Omission	(n=76) 17.88%	(n=64) 15.92%	(n=97) 23.89%	(n=237) 19.22%
Consonant Insertion	(n=12) 2.82%	(n=9) 2.24%	(n=18) 4.43%	(n=39) 3.16%
Epenthesis	(n=67) 15.76%	(n=84) 20.90%	(n=54) 13.30%	(n=205) 16.63%
Paragoge	(n=35) 8.24%	(n=67) 16.67%	(n=46) 11.33%	(n=148) 12%
Total (n)	425	402	406	1233

Table 6.3: The percentages of phonological spelling subcategory errors in the LEFLL corpus

In their study, Randall and Groom (2009) observed a high frequency of epenthesis spelling errors in their corpus of 16-year-old Arabic first language speakers. Furthermore, Arabs tend to employ epenthesis in their speaking (Sadhwani, 2005), therefore this could also be employed in spelling errors, and may thus be seen as evidence of a language transfer process. With respect to Paragoge*,* Sadhwani (2005) observed this type of spelling error to be prominent among Arab-Emirati college students, and proposed language transfer as a plausible explanation for this observation. A large percentage of consonant sound spelling errors, on the other hand, would not be so easily or plausibly attributed to language transfer phenomena. Only 5 out of 24 consonant phonemes of English do not have equivalents in Arabic which may cause problems to Arabs (Swan & Smith, 2001). This arguably provides further indirect support for the L1 transfer interpretation of vowel errors discussed above.

6.2.2 Orthographical Spelling Errors

Based on the general description of *orthographical* spelling error category given above, nine types of orthographical spelling errors were observed in the LEFLL corpus as following:

1) Vowel Substitution Spelling Errors occur when a vowel is substituted with another vowel due to sound confusion, e.g., '*bady'* (for '*body'*).
2) Consonant substitution Spelling Errors usually occur when a consonant is substituted with another consonant where both consonants have similar sounds in many cases, such as the confusion between '*b*' and '*p*' in '*pit'* (for '*bit'*).
3) Consonant Doubling Spelling Errors occur in two cases. The first case is when the learner erroneously doubles a single consonant by adding another consonant letter to the target word, such as: '*usefull'* (for '*useful'*). The second is when the learner removes a single letter from the doubled letter, e.g., '*alowed'* (for '*allowed'*).
4) Silent Letter Omission Spelling Errors occur when a silent letter is removed from the word, e.g., '*now'* (for '*know'*).

5) Silent [-*e*] Omission Spelling Errors occur when a silent [-*e*] is removed from the middle or the end of the target word, e.g., '*nintean'* (for '*nineteen'*) and '*mor'* (for *more*) respectively.
6) Consonant Digraph (ch/tch) confusion occurs when the learner substitutes '*ch*' with '*tch*' and vice versa, e.g., '*waching'* (for '*watching'*) or '*teatching'* (for '*teaching'*).
7) Consonant Digraph (ch/sh) confusion occurs when the learner substitutes the digraph '*ch*' with '*sh*' and vice versa, e.g., '*shild'* (for '*child'*) or '*wach'* (for '*wash'*).
8) Confusion between *'k'* & *'ck'* occurs, as the name suggests, when the learner replaces *'k'* with *'ck'* and vice versa. The only observed type of this orthographical spelling error category is when the learner inserted the letter *'ck'* instead of *'k'.* e.g., *'smocking'* (for *'smoking'*); and
9) Other: The *other* spelling error marks any orthographical spelling error that does not fall under the 8 orthographical spelling subcategory errors mentioned above.

As noted in Section 6.1 earlier, the substitution spelling errors scored the second highest percentage of all spelling errors (omission spelling errors were the highest as we observed in Section 6.1). Figure 6.4 and Table 6.4 below show that high prevalence of substitution spelling errors produced by all learners across the three university levels particularly affected the spelling of vowels. This may also indicate a state of confusion that the learners may experience in choosing the right vowel to form a word where there are two or more vowels that have similar sounds.

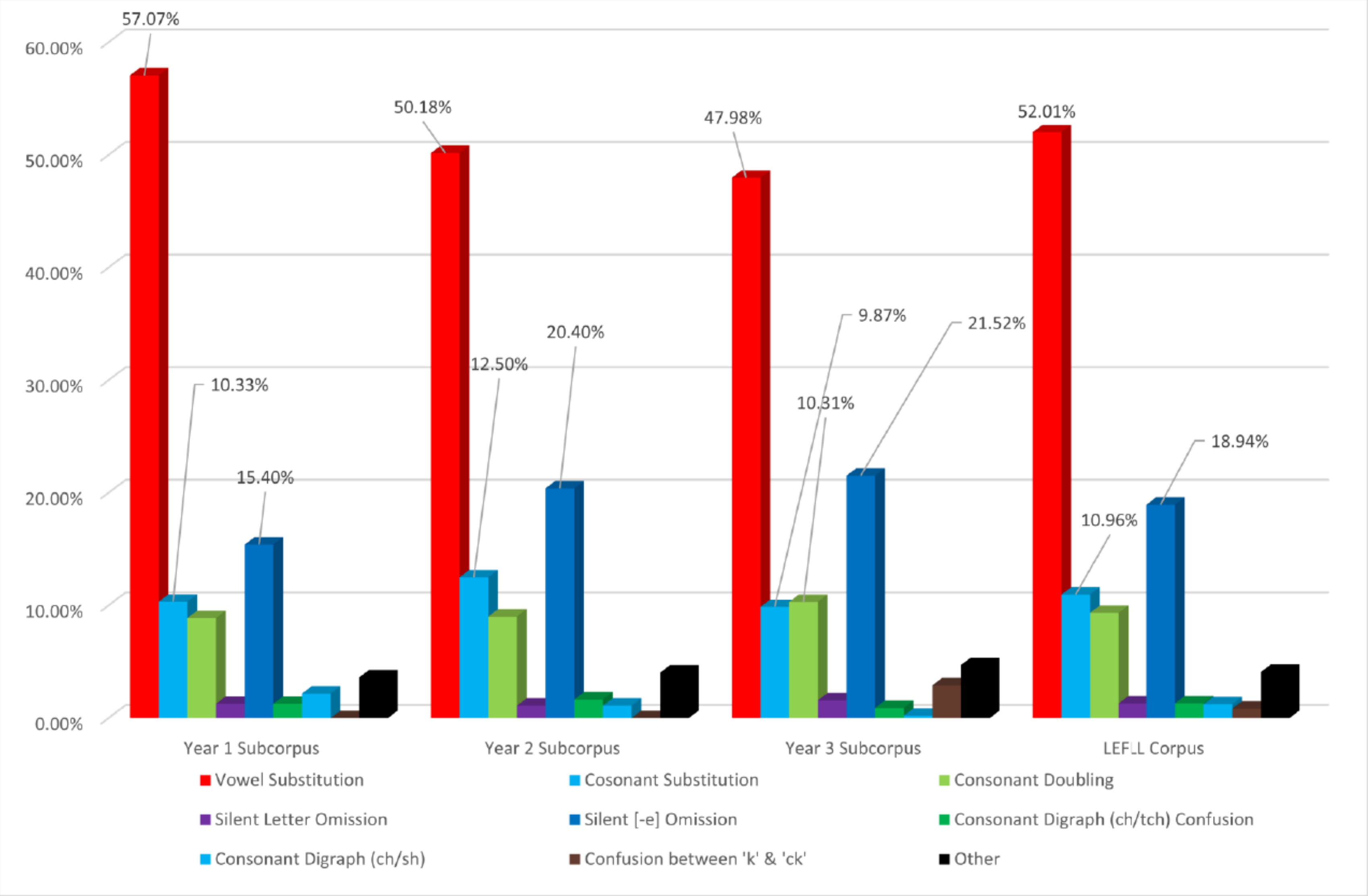


Figure 6.4: The percentages of orthographical spelling subcategory errors in the LEFLL corpus

However, as was observed for phonological spelling errors, it seems that the Libyan Arabic L1 learners represented in the LEFLL corpus start to overcome spelling problems affecting vowels across the three levels. The vowel substitution spelling errors for example show a slight decrease between year 1 and year 3.

On the other hand, there is almost a steady but slight increase in the percentages of spelling errors of the consonant doubling type (unnecessary doubling of a single letter or using a single instead of double letter) across the three-year groups. A plausible interpretation of this observation (originally suggested in a study by Bahr et al., 2015) is that it may reflect a growing

Orthographical Spelling Error Subcategories	**Year 1 Sub-Corpus**	**Year 2 Sub-Corpus**	**Year 3 Sub-Corpus**	**LEFLL Corpus**
Vowel Substitution	(n=315) 57.07%	(n=273) 50.18%	(n=214) 47.98%	(n=802) 52.01%
Consonant Substitution	(n=57) 10.33%	(n=68) 12.50%	(n=44) 9.87%	(n=169) 10.96%
Consonant Doubling	(n=49) 8.88%	(n=49) 9.01%	(n=46) 10.31%	(N=144) 9.34%
Silent Letter Omission	(n=7) 1.27%	(n=6) 1.10%	(n=7) 1.57%	(n=20) 1.30%
Silent (-*e*) Omission	(n=85) 15.40%	(n=111) 20.40%	(n=96) 21.52%	(n=292) 18.94%
Consonant Digraph (ch/tch) confusion	(n=7) 1.27%	(n=9) 1.65%	(n=4) 0.90%	(n=20) 1.30%
Consonant digraph (ch/sh)	(n=12) 2.17%	(n=6) 1.10%	(n=1) 0.22%	(n=19) 1.23%
Confusion between 'k' & 'ck'	(n=0) 0.00%	(n=0) 0.00%	(n=13) 2.91%	(n=13) 0.84%
Other	(n=20) 3.62%	(n=22) 4.04%	(n=21) 4.71%	(n=63) 4.09%
Total (n)	552	544	446	1542

Table 6.4: The percentages of orthographical spelling subcategory errors in the LEFLL corpus

awareness on the part of the learners of the double letter rule, but a growing uncertainty as to when or where they need to apply it.

6.2.3 Morphological Spelling Errors

Three main types of morphological spelling errors were observed in the LEFLL corpus, as follows:

1) *Homonyms* errors: These errors occur when the learner misspells a word, and in so doing producing another correctly spelled word whose pronunciation is the same or close to that of the target word. In other words, the learner uses a different word that can be pronounced the same as the target word, e.g., *to* (for *two*).
2) *Inflection* errors: These errors occur when the learner fails to apply an inflection rule correctly, such as his/her failure to double the last consonant letter due to the addition of some suffixes, e.g., *traveling* (for '*travelling*' due to the addition of '*-ing*'), inserting *'-d'* instead of *'-ed'* when inflecting the infinitive verb into past or past participle forms, e.g., '*happend*' (for '*happened*', *'happen'* is the infinitive), etc; and
3) *Derivation* errors: These errors occur when the learner misspells the word by changing the word from one-word class into another, such as the learner's failure to insert '*l*' when changing the word class from an adjective to an adverb word class, e.g., from the adjective final to the adverb *finaly* (for *finally*).

As shown in Figure 6.5 and Table 6.5 below, the students in years 1 and 3 did not show a significant difference in the percentages of homonym errors committed. On the other hand, there is a slight increase in the percentage of derivational spelling errors, particularly among year 2 undergraduate students. This could be attributed to an overgeneralization of the derivational linguistic rule which may have been acquired earlier during the language learning process (i.e., in or by year 1). If this is the case, these could be marked as development spelling errors because they indicate incomplete knowledge of the target language rules. Thus, this may indicate a positive development as the learners are seen here to be starting to apply the target language rules and distancing themselves from the influence of their first language.

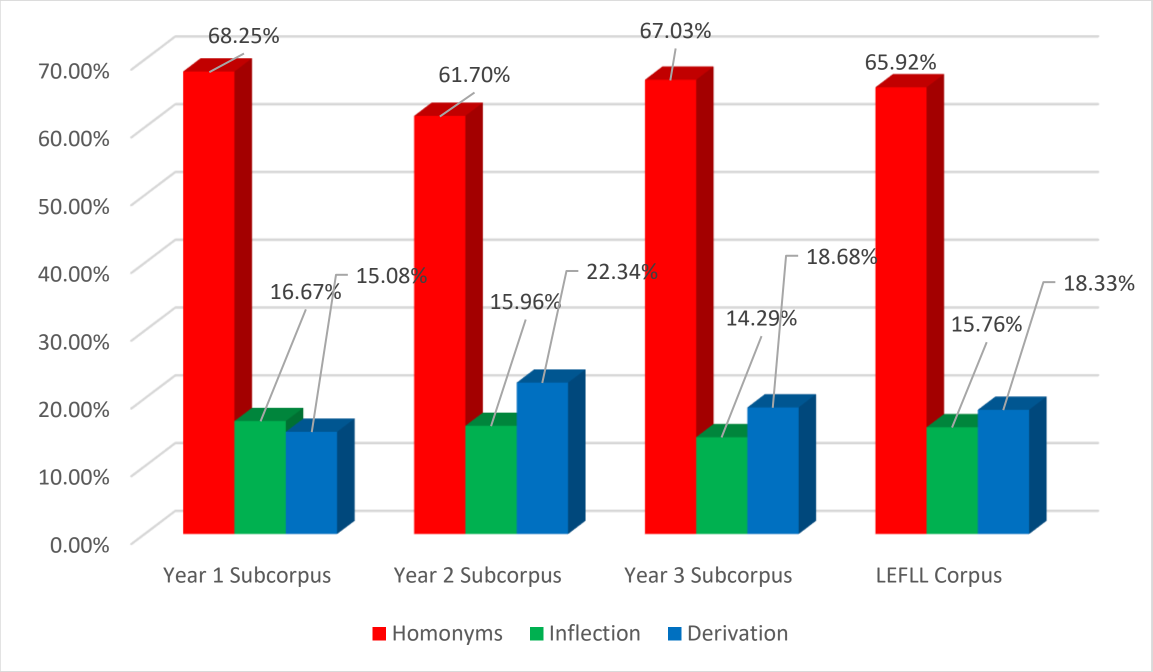


Figure 6.5: The percentages of morphological spelling subcategory errors in the LEFLL corpus

Morphological Spelling Error Subcategories	Year 1 Sub-corpus	Year 2 Sub-corpus	Year 3 Sub-corpus	LEFLL Corpus
Homonyms	(n=86) 68.25%	(n=58) 61.70%	(n=61) 67.03%	(n=205) 65.92%
Inflection	(n=21) 16.67%	(n=15) 15.96%	(n=13) 14.29%	(n=49) 15.76%
Derivation	(n=19) 15.08%	(n=21) 22.34%	(n=17) 18.68%	(n=57) 18.33%
Total (n)	126	94	91	311

Table 6.5: The percentages of morphological spelling subcategory errors in the LEFLL corpus

6.2.4 Phonological – Orthographical Spelling Errors

The *Phonological – Orthographical* spelling error category describes misspelled words that contain at least one phonological and one orthographical spelling error. The most obvious type of phonological – orthographical spelling error is *'Letter Reversed'* (see Bahr et al., 2015). Letter Reversed spelling errors show the transposition between two adjacent letters, e.g., '*cuases'* (for '*causes'*). The minimum number of edits that are required to convert a misspelled word of the phonological – orthographical spelling error type is one-edit distance for the *letter reversed*.

A single word may also contain two or more phonological and orthographical spelling errors where it needs two- or multiple-edit distances to modify the misspelled word and change it to the target one. Figure 6.6 and Table 6.6 below show the percentages of *letter reversed* spelling errors and the percentages of both phonological and orthographical spelling subcategory errors that were observed in the phonological – orthographical spelling error category.

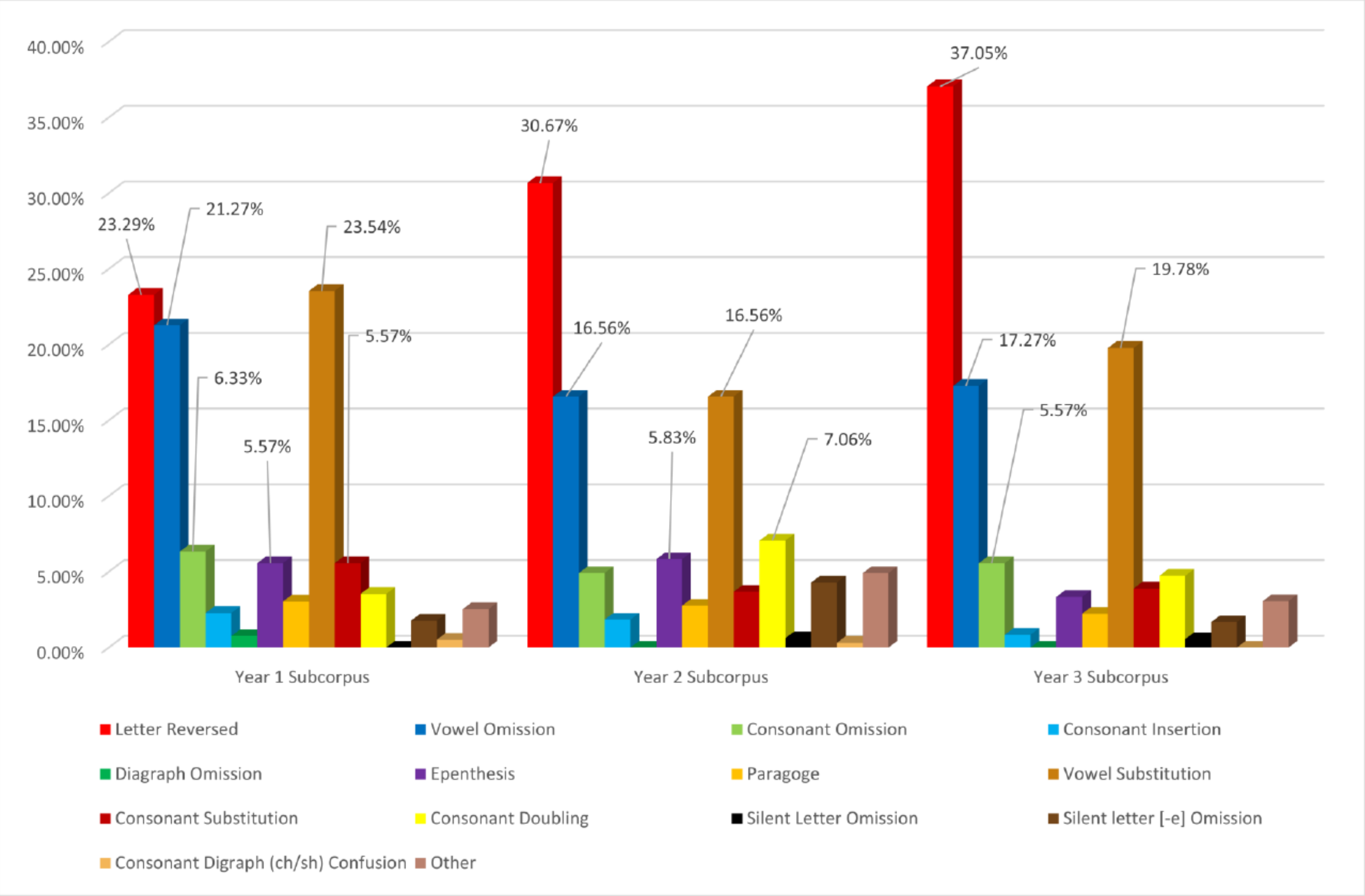


Figure 6.6: The percentages of phonological – orthographical spelling subcategory errors in the LEFLL corpus

As Figure 6.6 and Table 6.6 show, there is a steady increase of about 7% in the spelling errors of the *letter reversed* type as the analysis moves from one academic level to another. The *letter reversed* spelling error is a common phenomenon not only among the Libyan English learners in the current thesis but also in previous studies (e.g., Randall & Groom, 2009; Pacton et al., 2013). Thus, this may indicate that *letter reversed* spelling errors could be marked as instance of the intralingual spelling error type.

Phonological-Orthographical Spelling Error Subcategories	Year 1 Sub-corpus	Year 2 Sub-corpus	Year 3 Sub-corpus	LEFLL Corpus
Letter Reversed (Phono/Ortho)	(n=92) 23.29%	(n=100) 30.67%	(n=133) 37.05%	(n=325) 30.09%
Phonological spelling errors partition				
Vowel Omission	(n=84) 21.27%	(n=54) 16.56%	(n=62) 17.27%	(n=200) 18.52%
Consonant Omission	(n=25) 6.33%	(n=16) 4.91%	(n=20) 5.57%	(n=61) 5.65%
Consonant Insertion	(n=9) 2.28%	(n=6) 1.84%	(n=3) 0.84%	(n=18) 1.67%
Digraph Omission	(n=3) 0.76%	(n=0) 0.00%	(n=0) 0.00%	(n=3) 0.28%
Epenthesis	(n=22) 5.57%	(n=19) 5.83%	(n=12) 3.34%	(n=53) 4.91%
Paragoge	(n=12) 3.04%	(n=9) 2.76%	(n=8) 2.23%	(n=29) 2.69%
Orthographical spelling errors partition				
Vowel Substitution	(n=93) 23.54%	(n=54) 16.56%	(n=71) 19.78%	(n=218) 20.19%
Consonant Substitution	(n=22) 5.57%	(n=12) 3.68%	(n=14) 3.90%	(n=48) 4.44%
Consonant Doubling	(n=14) 3.54%	(n=23) 7.06%	(n=17) 4.74%	(n=54) 5.00%
Silent Letter Omission	(n=0) 0.00%	(n=2) 0.61%	(n=2) 0.56%	(n=4) 0.37%
Silent [–e] Omission	(n=7) 1.77%	(n=14) 4.29%	(n=6) 1.67%	(n=27) 2.50%
Consonant digraph (ch/sh) confusion	(n=2) 0.51%	(n=1) 0.31%	(n=0) 0.00%	(n=3) 0.28%
Other	(n=10) 2.53%	(n=16) 4.91%	(n=11) 3.06%	(n=37) 3.43%
Total (n)	395	326	359	1080

Table 6.6: The percentages of phonological – orthographical spelling subcategory errors in the LEFLL corpus

Following *letter reverse* spelling errors, spelling errors that are attributed to vowel confusion also scored high percentages among the learners. This is represented in the high percentages of both *vowel omission* in the phonological spelling error partition and *vowel substitution* in the orthographical spelling error partition. However, Figure 6.6 and Table 6.6 also show that among the advanced level essays, the percentages of these types of spelling errors declined. Again, it is not unreasonable to speculate that this may indicate a decreasing reliance on the mother language phonological and orthographical systems, in particular, the part that is concerned with vowel rules.

6.2.5 Phonological – Morphological Spelling Errors

The *Phonological – Morphological* spelling error category describes spelling errors that are shared between phonological and morphological spelling subcategory errors. The analysis showed that there is an overlap between phonological and morphological spelling errors. Misspelled words in this linguistic spelling error category contain at least one of the three morphological spelling subcategory errors (homonym, inflection and derivation) and one of the three types of phonological spelling error category (*vowel omission, paragoge, epenthesis*) as shown in Figure 6.7 and Table 6.7 below.

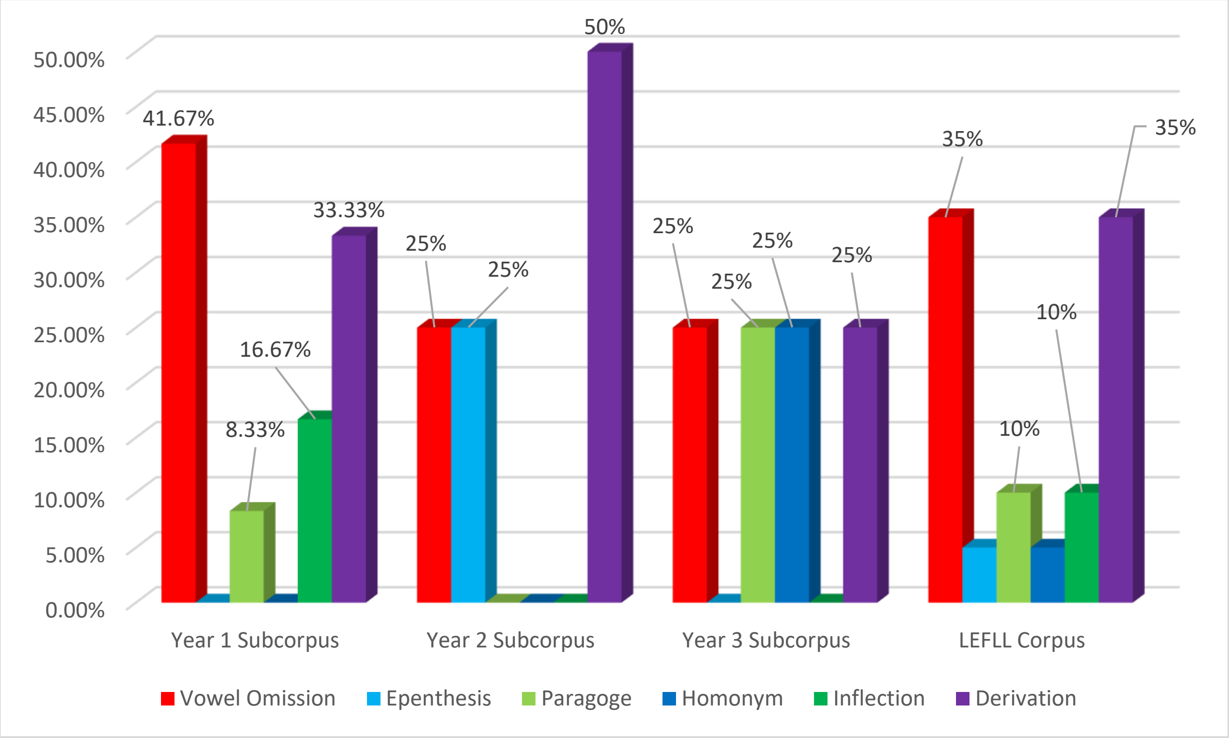


Figure 6.7: The percentages of phonological – morphological spelling subcategory errors in the LEFLL corpus

Generally, few instances of phonological – morphological spelling errors were observed in the LEFLL corpus, as can be seen in Table 6.7 below. Whereas *vowel omission* phonological spelling errors represent the highest percentage in the Year 1 sub-corpus, hardly any phonological – morphological spelling subcategory errors are observed in the year 2 sub-corpus and year 3 sub-corpus. The high percentage of *vowel omission* spelling error in Year 1 may be consistent with the claim in the previous sections that the vowel confusion may reflect a high degree of reliance on the mother language vowel system particularly in the lower levels.

Phonological – Morphological Spelling Error Subcategories	Year 1 Sub-corpus	Year 2 Sub-corpus	Year 3 Sub-corpus	LEFLL Corpus
Phonological spelling errors partition				
Vowel Omission	(n=5) 41.67%	(n=1) 25.00%	(n=1) 25.00%	(n=7) 35.00%
Epenthesis	(n=0) 0.00%	(n=1) 25.00%	(n=0) 0.00%	(n=1) 5.00%
Paragoge	(n=1) 8.33%	(n=0) 0.00%	(n=1) 25.00%	(n=2) 10.00%
Morphological spelling errors partition				
Homonym	(n=0) 0.00%	(n=0) 0.00%	(n=1) 25.00%	(n=1) 5.00%
Inflection	(n=2) 16.67%	(n=0) 0.00%	(n=0) 0.00%	(n=2) 10.00%
Derivation	(n=4) 33.33%	(n=2) 50.00%	(n=1) 25.00%	(n=7) 35.00%
Total (n)	12	4	4	20

Table 6.7: The percentages of phonological – morphological spelling subcategory Errors in the LEFLL corpus

6.2.6 Orthographical – Morphological Spelling Errors

The *Orthographical – Morphological* spelling error category includes misspelled words that contain both orthographical and morphological spelling subcategory errors. The analysis showed that there are only two morphological spelling subcategory errors (inflection and derivation) interfering with five orthographical spelling subcategory errors. As shown in Figure 6.8 and Table 6.8 below, there are few instances of these types of spelling errors.

Spelling errors that are related to vowel confusion (*vowel substitution*) spelling errors also seem to be an important figure in this type of spelling error.

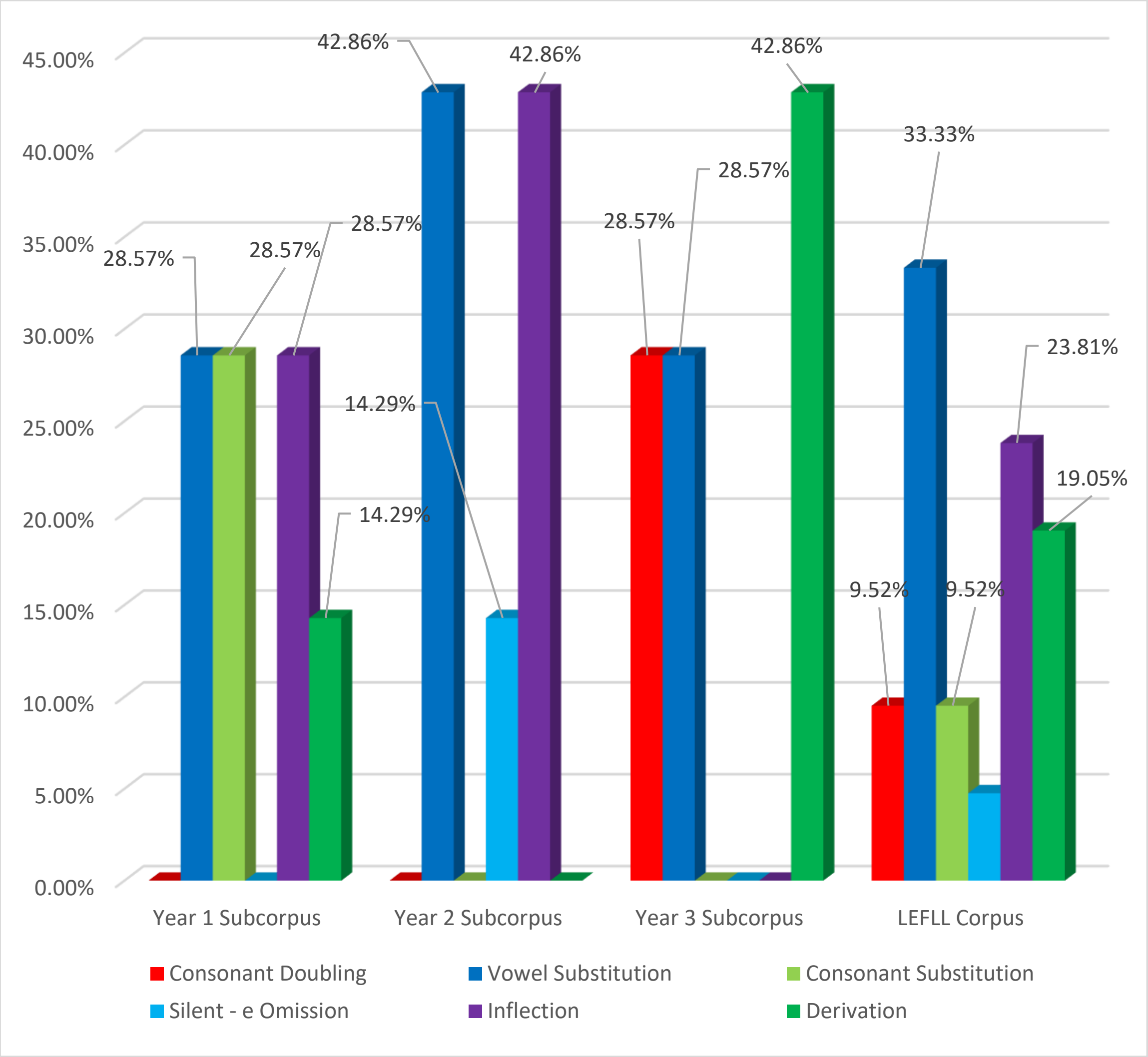


Figure 6.8: The percentages of orthographical – morphological spelling subcategory Errors in the LEFLL corpus

Orthographical – Morphological Spelling Error Subcategories	**Year 1 Sub-corpus**	**Year 2 Sub-corpus**	**Year 3 Sub-corpus**	**LEFLL Corpus**
Orthographical spelling errors partition				
Consonant Doubling	(n=0) 0.00%	(n=0) 0.00%	(n=2) 28.57%	(n=2) 9.52%
Vowel Substitution	(n=2) 28.57%	(n=3) 42.86%	(n=2) 28.57%	(n=7) 33.33%
Consonant Substitution	(n=2) 28.57%	(n=0) 0.00%	(n=0) 0.00%	(n=2) 9.52%
Silent – *e* omission	(n=0) 0.00%	(n=1) 14.29%	(n=0) 0.00%	(n=1) 4.76%
Morphological spelling errors partition				
Inflection	(n=2) 28.57%	(n=3) 42.86%	(n=0) 0.00%	(n=5) 23.81%
Derivation	(n=1) 14.29%	(n=0) 0.00%	(n=3) 42.86%	(n=4) 19.05%
Total (n)	7	7	7	21

Table 6.8: The percentages of orthographical – morphological spelling subcategory errors in the LEFLL corpus

To summarise the discussion so far, this section has discussed the classification of spelling errors based on the linguistic subsystem taxonomy into six categories, according to the POMAS (Phonological, Orthographical and Morphological Assessment of Spelling) spelling error classification, providing examples of each type from the LEFLL corpus. The analysis revealed that the learners encountered major problems with English vowels, which in turn led to the observation of high percentages of vowel sound spelling errors. These were observed in the form of vowel omission and substitution spelling errors, fundamentally, in the phonological and orthographical spelling error categories.

It was assumed, in most cases, that the mother language of the learners could be the main cause of spelling errors that affected vowels. This assumption is supported not only by standard knowledge regarding Arabic/English contrasts (e.g., Swan & Smith 2001) but also by the prevalence of these types of errors observed in previous studies of spelling errors produced by

Arabic L1 learners of English (e.g., Randall & Groom, 2009; Saigh & Schmitt, 2012; Yildiz, 2017; Altamimi et al., 2018; Ibrahim, 2018; Altamimi & Ab Rashid, 2019). It was also argued that some error trends observed across the three sub-corpora could be interpreted as positive signs of second language development.

The analysis also revealed that for some spelling error categories, such as phonological – morphological and orthographical errors, few instances of spelling errors were observed. With respect to the word length and as was observed earlier in Section 5.3, few instances of misspelled words in the 2-letter words group were observed in the LEFLL corpus. Given their rarity, these could be interpreted as accidental spelling errors and thus not necessarily reflecting either first language interference or target language development. They are nonetheless worth noting here if only for the reason that they may affect the percentages of both interlingual and intralingual spelling errors that will be discussed later in this thesis. Therefore, in order to ensure maximally valid results, they were removed before classifying spelling errors into interlingual and/or intralingual spelling error categories.

Thus, before proceeding to classify spelling errors into interlingual and/or intralingual type(s), the linguistic competence taxonomy will be applied to classify spelling errors into performance and competence spelling errors. Performance spelling errors are accidental, so they will be excluded before moving to the language influence taxonomy section. Competence spelling errors are systematic in that they do reflect either first language interference or target language development, as we will see below.

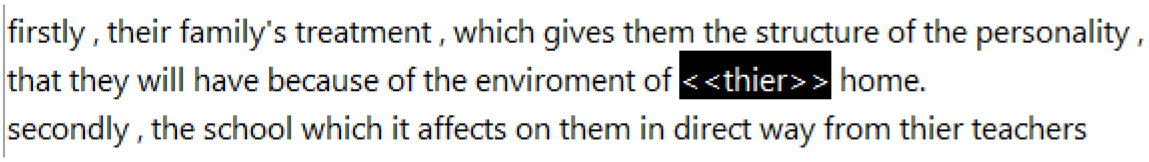
6.3 Spelling Errors Based on the Linguistic Competence Taxonomy

As discussed in the previous section, many of the spelling errors extracted from the LEFLL corpus contain linguistic features which may reflect first language interference (i.e., they are classifiable as interlingual spelling errors). Other high percentages of spelling errors were also assumed to be reflections of target language development (i.e., they are classifiable as intralingual spelling errors). The analysis also showed that some spelling errors may be best described as incidental/accidental spelling errors since they reflect neither first language interference nor target language development. Learners were found to be able to spell these words correctly in subsequent attempts in the same texts. Thus, this section is devoted to classifying spelling errors into two categories. Firstly, we will briefly review performance spelling errors, i.e., errors that are merely incidental/accidental. We will then turn to competence spelling errors, which mark serious linguistic problems, sub-classifying errors of this type into interlingual and intralingual spelling errors.

6.3.1 Performance Spelling Errors

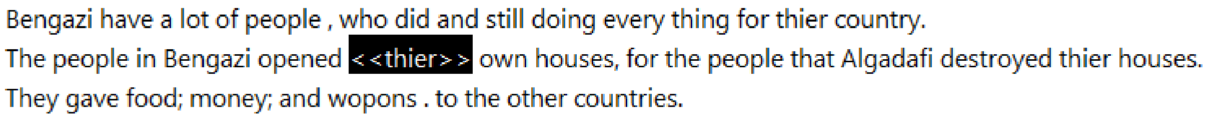
The *performance spelling error* category refers to incidental spelling errors where learners accidently misspell the target word, especially when typing their written work. They are accidental because the writers show themselves to be able to correct them whenever they are given a chance, or spell them correctly in subsequent attempts in the same essay. These types of spelling errors are common among both English native speakers and second/foreign language learners, and therefore do not reflect spelling difficulties of the kind we are concerned with in this thesis. In the current thesis, spelling errors were marked as performance/accidental when:

1. The language learner has successfully spelled the target word in the second attempt and vice versa, e.g., in the following example:



firstly , their family's treatment , which gives them the structure of the personality , that they will have because of the enviroment of <<thier>> home.
secondly , the school which it affects on them in direct way from thier teachers

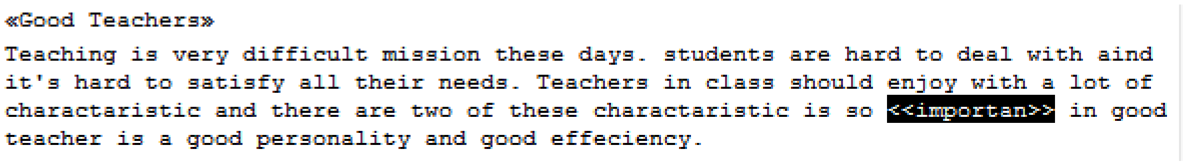
The language learner has successfully spelled the word ‘*their’* in the first attempt but failed in their second and third attempts. Therefore, it is marked as performance spelling error. As may be recalled, this was one of the recommendations that the researcher obtained from discussions during the inter-rater reliability test phase of the research (see Section 4.4.1.4). This criterion was applied in each text. In other words, the word ‘*their*’ was marked as a competence spelling error in other texts because the writers of those texts failed to spell the word *‘their’* correctly in the following attempts in the same texts as shown in the following example:



Bengazi have a lot of people , who did and still doing every thing for thier country.
The people in Bengazi opened <<thier>> own houses, for the people that Algadafi destroyed thier houses.
They gave food; money; and wopons . to the other countries.

In this example, the learner failed to spell the word *‘their’* correctly on all occasions. Therefore, any misspelled word of *‘their’* in the same text written by the same student was marked as a competence spelling error.

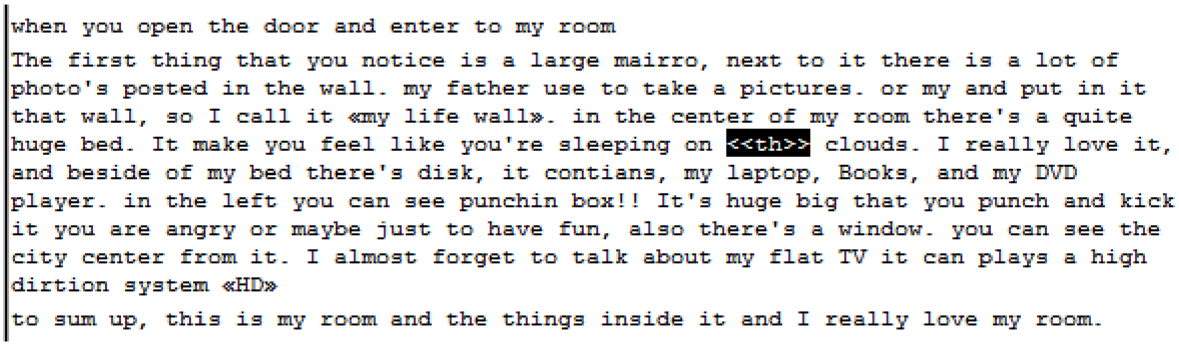
2. When there are very few instances of spelling errors of a common word verses high frequencies of correctly spelled ones. Consider the following example:



«Good Teachers»
Teaching is very difficult mission these days. students are hard to deal with aind it's hard to satisfy all their needs. Teachers in class should enjoy with a lot of charactaristic and there are two of these charactaristic is so <<importan>> in good teacher is a good personality and good effeciency.

The word *'important'* has only been misspelled once in the whole LEFLL corpus. As shown in this example, the language learner omitted the final consonant letter 't'. Since there are 92 correctly spelled instances of the word '*important'* in the three sub-corpora (66 instances of the word '*important'* in the Year 3 sub-corpus where *'importan'* was spotted) vs 1 spelling error of the word *'important',* it is reasonable to regard this single instance of *'importan'* as an incidental spelling error.

3. When grammatical words, such as *an, the, and,* etc. which are both very frequent and very infrequently misspelled, are found to have been misspelled. To illustrate, consider the misspelling of the definite article '*the'* in the following example:



```
when you open the door and enter to my room
The first thing that you notice is a large mairro, next to it there is a lot of
photo's posted in the wall. my father use to take a pictures. or my and put in it
that wall, so I call it «my life wall». in the center of my room there's a quite
huge bed. It make you feel like you're sleeping on <<th>> clouds. I really love it,
and beside of my bed there's disk, it contians, my laptop, Books, and my DVD
player. in the left you can see punchin box!! It's huge big that you punch and kick
it you are angry or maybe just to have fun, also there's a window. you can see the
city center from it. I almost forget to talk about my flat TV it can plays a high
dirtion system «HD»
to sum up, this is my room and the things inside it and I really love my room.
```

The definite article *'the'* is the most frequently occurring word in the LEFLL corpus (as it is in virtually all corpora of written English). The analysis shows that *'the'* was so infrequently misspelled that the most plausible explanation is simply that it was accidently misspelled on this occasion. We can also see that the word *'the'* was spelled correctly several times in the example above. This would also seem to indicate that *'th'* is best regarded here as an accidental spelling error.

6.3.2 Competence Spelling Errors

The *Competence spelling errors* include misspelled words that do not fall under the performance/accidental spelling error category. These types of spelling errors reflect linguistic features that are problematic for the language learners and are either interlingual (i.e., attributable to the role of the learner's first language), or intralingual/developmental (i.e., produced as a result of incomplete knowledge, incorrect application of the target language spelling rules or an irregularity in the L2 rule system).

The analysis showed that some of the most frequently misspelled words could be marked as competence spelling errors. As can be seen in Table 6.9 below, the word '*because'* was among the top 4 most frequently misspelled words in the LEFLL corpus. In other cases, some misspelled words were only shared between two sub-corpora, such as '*friends'* (the seventh most frequently misspelled word in the Year 1 sub-corpus) and '*friend'* (the seventeenth most frequently misspelled word in the Year 3 sub-corpus).

	Year 1 Sub-corpus		Year 2 Sub-corpus		Year 3 Sub-corpus	
	SP Word Type	(Freq.) & %	SP Word Type	(Freq.) & %	SP Word Type	(Freq.) & %
1	University	(n=78) 6.68	Colour	(n=41) 3.66	Their	(n=41) 3.78
2	Nineteen	(n=28) 2.40	Beautiful	(n=40) 3.57	**Because**	(n=16) 1.47
3	Then	(n=27) 2.31	Bed	(n=25) 2.23	Another	(n=14) 1.29
4	**Because**	(n=23) 1.97	**Because**	(n=24) 2.14	Foreign	(n=14) 1.29
5	Introduce	(n=21) 1.80	Finally	(n=17) 1.52	Chocolate	(n=12) 1.11
6	Finally	(n=17) 1.46	There	(n=17) 1.52	Different	(n=12) 1.11
7	**Friends**	(n=17) 1.46	Chair	(n=15) 1.34	Money	(n=12) 1.11
8	Language	(n=17) 1.46	Mirror	(n=14) 1.25	Addition	(n=9) 0.83
9	Family	(n=16) 1.37	Computer	(n=13) 1.16	Equivalent	(n=9) 0.83
10	Brother	(n=15) 1.29	White	(n=13) 1.16	Your	(n=9) 0.83
11	College	(n=15) 1.29	Pink	(n=11) 0.98	Healthy	(n=8) 0.74
12	Want	(n=14) 1.20	Second	(n=10) 0.89	Population	(n=8) 0.74
13	With	(n=14) 1.20	Table	(n=10) 0.89	Studying	(n=8) 0.74
14	Arabic	(n=13) 1.11	Window	(n=10) 0.89	Interesting	(n=7) 0.65
15	Mother	(n=13) 1.11	Clothes	(n=9) 0.80	Things	(n=7) 0.65
16	Lecture	(n=12) 1.03	Colours	(n=9) 0.80	Will	(n=7) 0.65
17	Like	(n=12) 1.03	Laptop	(n=9) 0.80	**Friend**	(n=6) 0.55
18	Cafeteria	(n=11) 0.94	Large	(n=9) 0.80	Arabic	(n=5) 0.46
19	Introducing	(n=11) 0.94	Before	(n=8) 0.71	Difficult	(n=5) 0.46
20	Listening	(n=11) 0.94	Inside	(n=8) 0.71	Disadvantages	(n=5) 0.46

Table 6.9: The top 20 most frequent misspelled words in the LEFLL corpus

On the other hand, some misspelled words showed high frequencies in one specific sub-corpus but were infrequently misspelled in the other two sub-corpora. An example for this phenomenon is the most frequently misspelled word in the Year 3 sub-corpus, '*their',* which constituted 3.78% of all misspellings (*'their'* was misspelled 41 times) in the Year 3 sub-corpus. In the Year 1 and Year 2 sub-corpora, *'their'* was misspelled only once and six times respectively. The variation between the three sub-corpora helped to mark the misspelled word *'their'* as a competence spelling error. A thorough investigation revealed that this variation was mainly due to the frequencies of the word '*their'* in the three sub-corpora. After retrieving both incorrectly and correctly spelled instances of *'their',* it became obvious that the frequency of *'their'* in the Year 1 sub-corpus, Year 2 sub-corpus and Year 3 sub-corpus is 10, 19 & 178 respectively. Subsequently, the variations between the frequencies of '*their'* in the three sub-corpora has affected its potential for spelling error. The chance of misspelling *'their'* in the Year 3 sub-corpus was the highest because it contains the highest frequency of this word, followed by the Year 2 sub-corpus. This concept was applied to all misspelled words which showed a notably higher frequency in one specific sub-corpus than the others.

No matter how many instances of misspelled words of the same word type there were in the three sub-corpora, the distribution of the same misspelled word type across two or three sub-corpora raises the chance of marking these types of spelling errors as competence rather than performance spelling errors. As shown in Figure 6.9 and Table 6.10 below, the classification of spelling errors according to the linguistic competence taxonomy indicates that the learners started producing more performance spelling errors as the analysis moved from one university level to another vs a decline in the percentages of competence spelling errors.

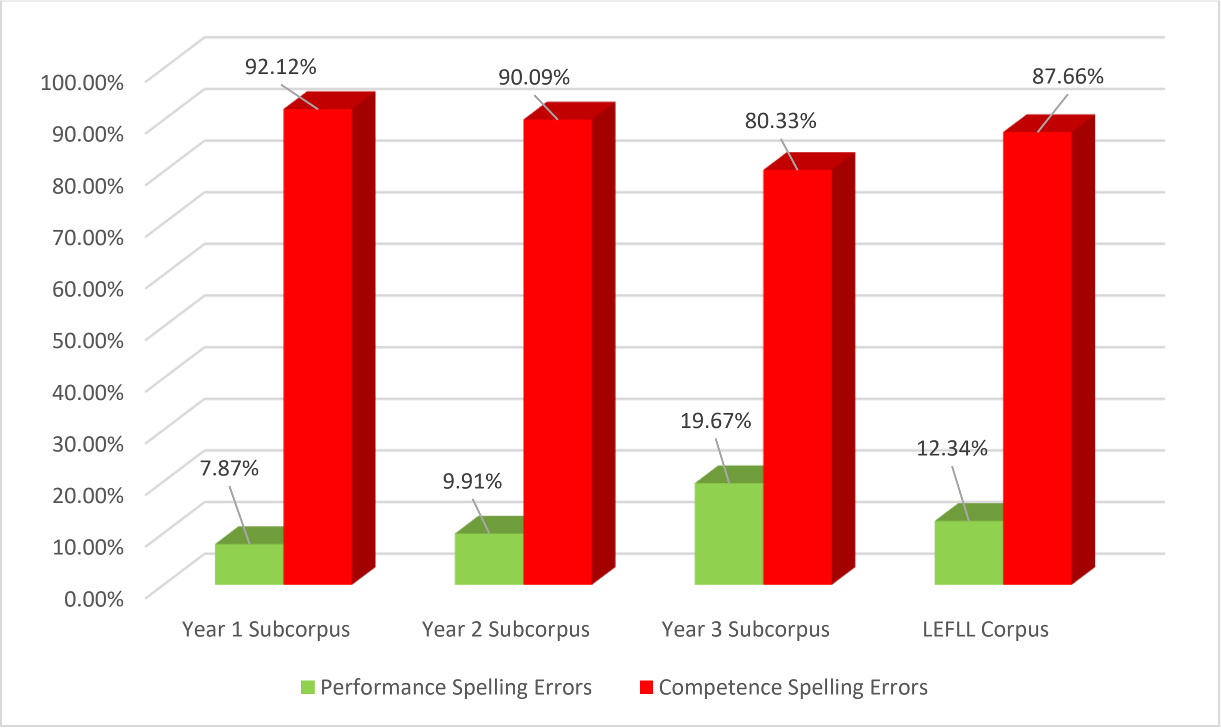


Figure 6.9: The percentages of performance and competence spelling errors in the LEFLL corpus

	Performance Spelling Errors		**Competence Spelling Errors**	
	Total	%	Total	%
Year 1 Sub-corpus	92	7.87	1077	92.12
Year 2 Sub-corpus	111	9.91	1009	90.09
Year 3 Sub-corpus	213	19.67	870	80.33
LEFLL Corpus	416	12.34	2956	87.66

Table 6.10: The percentages of performance and competence spelling errors in the LEFLL corpus

As a by-product of the linguistic competence taxonomy, the results obtained may constitute a further advantage of the PFEC approach (proposed by the current thesis) over Thewissen's POA approach. Whereas there was an increase in the percentages of performance spelling errors vs a decrease in the percentages of competence spelling errors, this was associated with an increase in the type/token ratio as we move from one level to another, as can be seen in Figure 6.10 and Table 6.11 below. This may indicate that the learners produce more varied word types at each successive year level. This may indicate that the year 3 undergraduate students may have successfully spelled the same word types that their colleagues in the lower levels (e.g., the year 1 undergraduate students) may have incorrectly spelled.

In cases where the year 3 undergraduate students still misspell the same word types that their colleagues in the lower levels have misspelled, an increase in the percentages of performance spelling errors may indicate that a large portion of misspelled words were accidental, and the students could spell them correctly if given a chance. Furthermore, the augmentation of performance spelling errors in the LEFLL corpus across the three sub-corpora may indicate (in contradiction to Thewissen) that there is an ever-decreasing chance of spotting spelling errors in simple and short words (e.g., 2-letter words, 3-letter words, etc.).

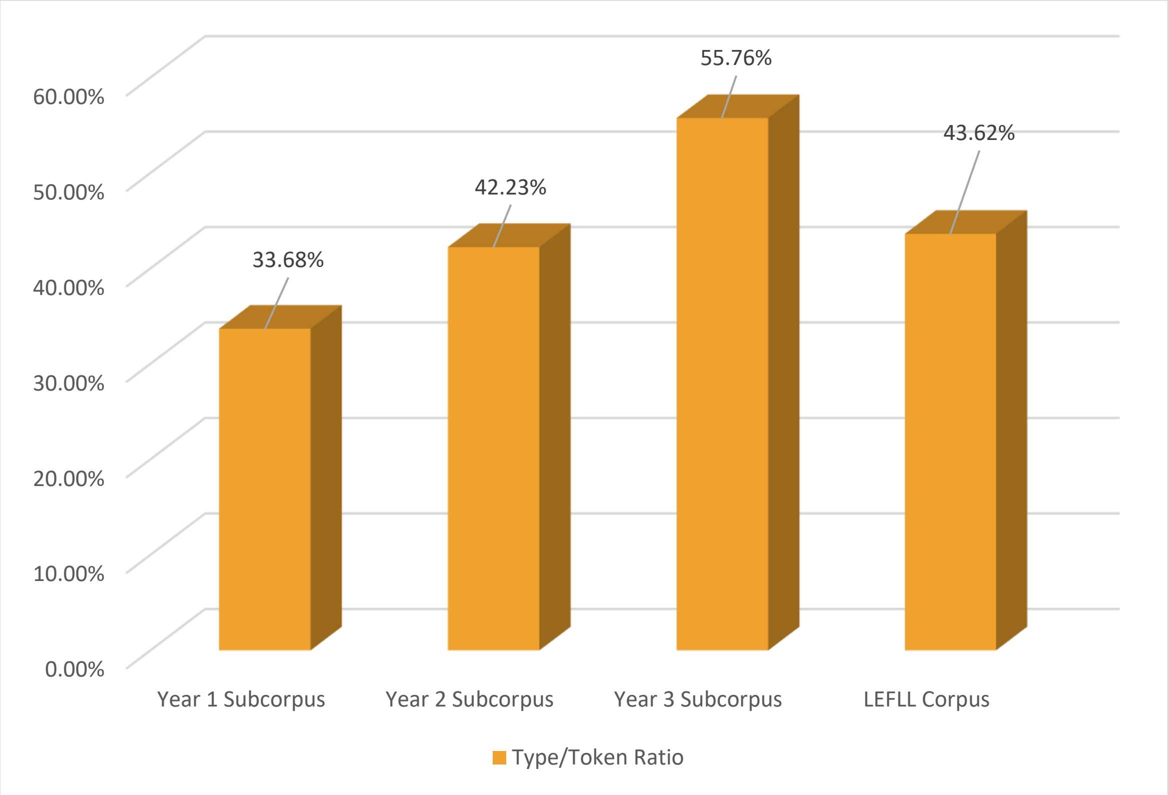


Figure 6.10: Spelling errors Type/Token ratio in the LEFLL corpus

Subsequently, this may support the claim that there is a need for practical investigation to verify which word types in any given learner corpus genuinely have the potential for spelling error. Accordingly, this thesis claims that it is necessary to verify at an early stage in any learner corpus analysis which words the language learners have actually misspelled in the learner corpus (as the PFEC approach claims) rather than relying on an untested assumption that every word in the learner corpus has the potential for spelling error (as the POA approach claims).

	Word Token	**Word Type**	**Type/Token Ratio**
Year 1 Sub-corpus	(n=1,167)	(n=393)	33.68%
Year 2 Sub-corpus	(n=1,120)	(n=473)	42,23%
Year 3 Sub-corpus	(n=1,085)	(n=605)	55.76%
LEFLL Corpus	(n=3372)	(n=1471)	43.62%

Table 6.11: Spelling errors Type/Token Ratio in the LEFLL corpus

So far, this section has applied the linguistic competence taxonomy to classify spelling errors into performance and competence spelling error types. This was an essential step in that it enabled us to exclude accidental spelling errors from the overall picture of spelling errors retrieved from the LEFLL corpus, allowing us to focus exclusively on competence spelling errors, which are of more theoretical and pedagogic relevance and importance. The following section will thus now proceed to classify competence spelling errors into interlingual and intralingual errors respectively. Spelling errors will be classified as interlingual if they can be plausibly explained as occurring as a result of language transfer. The intralingual spelling error category marks spelling errors that are more plausibly interpreted as being due to language development factors, i.e., when the learner, for example, overgeneralizes a spelling rule.

6.4 Spelling Errors Based on the Language Influence Taxonomy

The competence spelling errors retrieved in Section 6.3.2 above will now be further classified into, firstly, interlingual spelling errors, representing spelling errors that occurred due to the learners' first language transfer, and secondly, intralingual spelling errors, which are attributed to target language development.

The analysis revealed that at low levels, the learners relied heavily on their first language spelling rules and thus produced higher percentages of interlingual spelling errors as shown in Figure 6.11 and Table 6.12 below. As the analysis moves from one level to another, the role of the first language seems to decline, leading to an increase in intralingual spelling errors.

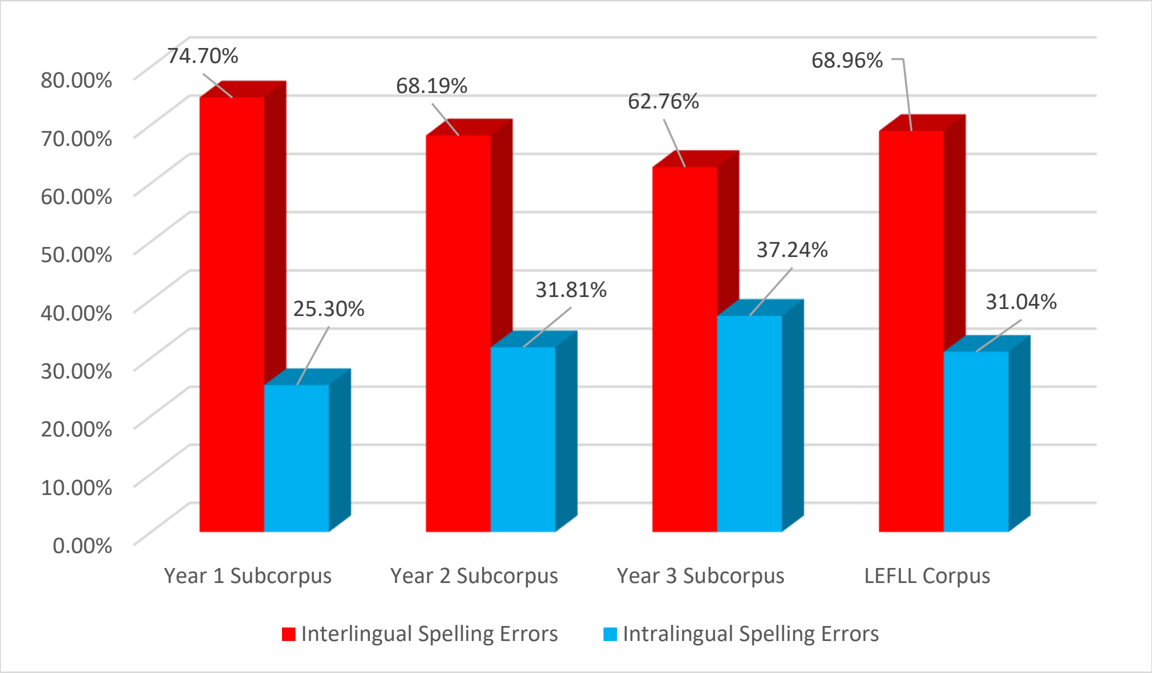


Figure 6.11: The percentages of interlingual vs intralingual spelling errors in the LEFLL corpus

Figure 6.11 and Table 6.12 show that there is a regular decline in the percentages of interlingual spelling errors vs an increase in the percentages of intralingual spelling errors as the learners as the analysis moves from one level to another. This shows that at the low levels, the learners rely heavily on their first language spelling strategies. This starts decreasing as the analysis moves from one level to another. This finding is consistent with a number of previous research findings. For instance, Figueredo (2006) reported in his study which reviewed 27 studies of first language influence on the development of spelling skills of the second language learners that these studies expressed two main key points. Firstly, the evidence of both positive and negative language transfer. Secondly, as the language learners proceed in the development of spelling skills in the second language, they tend to rely less on their first language knowledge.

	Interlingual Spelling Errors	Intralingual Spelling Errors
Year 1 Sub-corpus	(n=803) 74.70%	(n=272) 25.30%
Year 2 Sub-corpus	(n=688) 68.19%	(n=321) 31.81%
Year 3 Sub-corpus	(n=546) 62.76%	(n=324) 37.24%
LEFLL Corpus	(n=2037) 68.96%	(n=917) 31.04%

Table 6.12: The percentages of interlingual vs intralingual spelling errors in the LEFLL corpus

This may explain why learners at the lower levels, in the current thesis, produced more interlingual spelling errors than intralingual ones. The high frequency of interlingual spelling errors may be due to negative language transfer, caused by the vast differences between the spelling conventions of the Arabic and English writing systems. The steady decline in the percentages of interlingual spelling errors may also provide an explanation for why the percentages of spelling errors showed a small but steady decline when they were calculated based on the PFEC approach, as was discussed earlier in Section 5.3. The PFEC approach showed that the percentages of spelling errors were 6.72%, 5.95% and 5.74% in the Year 1, Year 2 and Year 3 sub-corpora respectively, unlike the results obtained via the TEC approach, where no steady change (either increase or decrease) was observed in the percentages of spelling errors, and where year 2 was an anomalous stage between years 1 and 3. This claim that there appears to be a systematic relationship between the steady decline in the percentages of interlingual spelling and the steady decline observed earlier in Figure 5.6 and Table 5.9 in Section 5.3 will be verified later in this thesis, by analysing the noun and verb phrase errors in chapters 7 and 8 respectively. As Figure 5.6 and Table 5.9 showed, there are no steady changes (either an increase or decrease) in the percentages of noun and verb phrase errors across the three LEFLL sub-corpora when they were calculated via the PFEC approach. Therefore, we would expect to see unsteady changes in the percentages of interlingual errors vs intralingual errors of both noun phrase and verb phrase errors as the analysis moves from one level to another.

So far, this section has presented the percentages of interlingual vs intralingual spelling errors in the LEFLL corpus and across the three LEFLL sub-corpora. In the following section I will present the subcategories of interlingual spelling errors based on the linguistic subsystem taxonomy described in Section 6.2 above. This analysis will aim to verify the percentages of interlingual spelling errors in the three linguistic categories of phonological, orthographical and morphological errors. Section 6.6 will discuss the reasons why the following linguistic spelling subcategory errors were classified as interlingual spelling errors.

6.4.1 Interlingual Spelling Errors

The analysis showed that interlingual spelling errors occurred in five out of the six linguistic spelling error categories (no interlingual spelling errors were observed at all in the morphological spelling error category) that were previously discussed in Section 6.2 above. An interlingual misspelled word may be classified as such if it contains only the linguistic features that clearly mark a language transfer effect. This was observed in the phonological and orthographical spelling error categories. In addition to interlingual spelling errors, some misspelled words may also contain the linguistic features that mark target language development processes (i.e., intralingual spelling errors). The latter type of misspelled words was observed in the phonological – orthographical, phonological – morphological and orthographical – morphological error types.

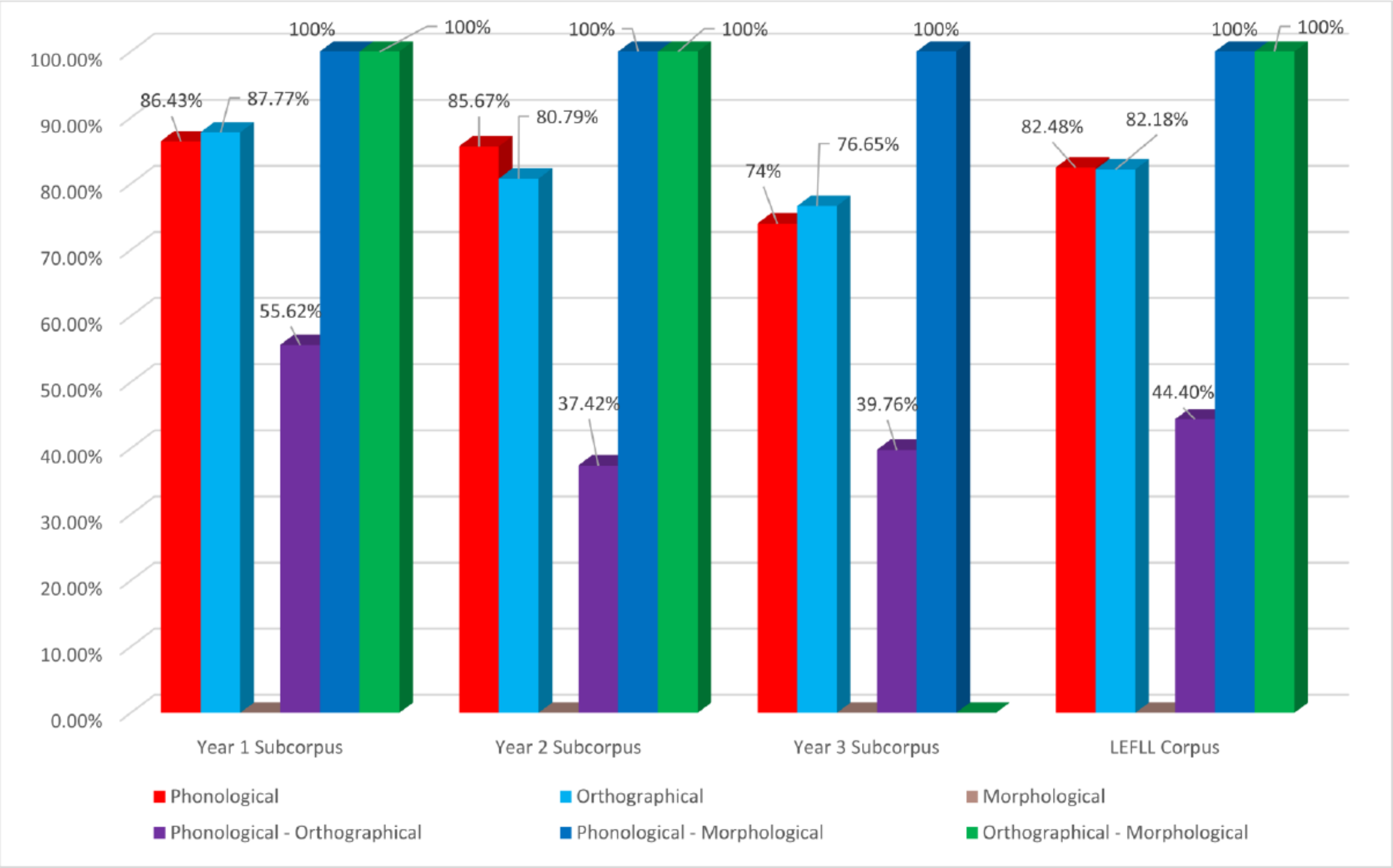


Figure 6.12: The percentages of interlingual spelling errors in each linguistic spelling error category in the LEFLL corpus

As Figure 6.12 above and Table 6.13 below show, in terms of frequencies, orthographical spelling errors seem to be the main source of interlingual spelling errors. There is also a largely regular decrease in the percentages of spelling errors of each linguistic spelling error category. While this global trend may be attributed at a very general level to a decrease in the role of the first language of the learners, the extent of first language influence may differ from one linguistic spelling error category to another, as will become clearer below, as we discuss each subcategory thoroughly in turn.

Linguistic Spelling Error Category Type	Year 1 Sub-corpus		Year 2 Sub-corpus		Year 3 Sub-corpus		LEFLL Corpus	
	Inter.	Intra.	Inter.	Intra.	Inter.	Intra.	Inter.	Intra.
Phonological	(n=293) 86.43%	(n=48) 13.57%	(n=257) 85.67%	(n=42) 14.33%	(n=222) 74%	(n=74) 26%	(n=772) 82.48%	(n=164) 17.52%
Orthographical	(n=402) 87.77%	(n=56) 12.23%	(n=366) 80.79%	(n=87) 19.21%	(n=256) 76.65%	(n=79) 23.35%	(n=1024) 82.18%	(n=222) 17.82%
Morphological	(n=0) 0.00%	(n=77) 100%	(n=0) 0.00%	(n=82) 100%	(n=0) 0.00%	(n=62)100%	(n=0) 0.00%	(n=221) 100%
Phonological – Orthographical	(n=99) 55.62%)	(n=79) 44.38%	(n=61) 37.42%	(n=104) 62.58%	(n=66) 39.76%	(n=100) 60.24%	(n=226) 44.40%	(n=283) 55.60%
Phonological – Morphological	(n=6) 100%	(n=0) 0.00%	(n=2) 100%	(n=0) 0.00%	(n=2) 100%	(n=0) 0.00%	(n=10) 100%	(n=0) 0.00%
Orthographical – Morphological	(n=3) 100%	(n=0) 0.00%	(n=3) 100%	(n=0) 0.00%	(n=0) 0.00%	(n=0) 0.00%	(n=6) 100%	(n=0) 0.00%

Table 6.13: The percentages of Interlingual vs intralingual errors in each linguistic spelling error category in the LEFLL corpus

So far, this section has provided a general overview of the distribution of interlingual spelling errors based on the linguistic subsystem taxonomy. The following section will provide more details about interlingual spelling errors, looking in detail at the five linguistic categories (phonological, orthographical, phonological – orthographical, phonological – morphological and orthographical – morphological) where the interlingual spelling errors were observed.

6.4.1.1 Phonological Interlingual Spelling Errors

A large percentage of interlingual spelling errors were attributed to problems with phonology, particularly where vowel sound spelling errors played a central role. This was mainly represented in the omission of vowels, which constituted the predominant phenomenon as shown in Figure 6.13 and Table 6.14 below. Interlingual spelling errors are marked in three phonological spelling subcategory errors (*vowel omission, epenthesis* and *paragoge*).

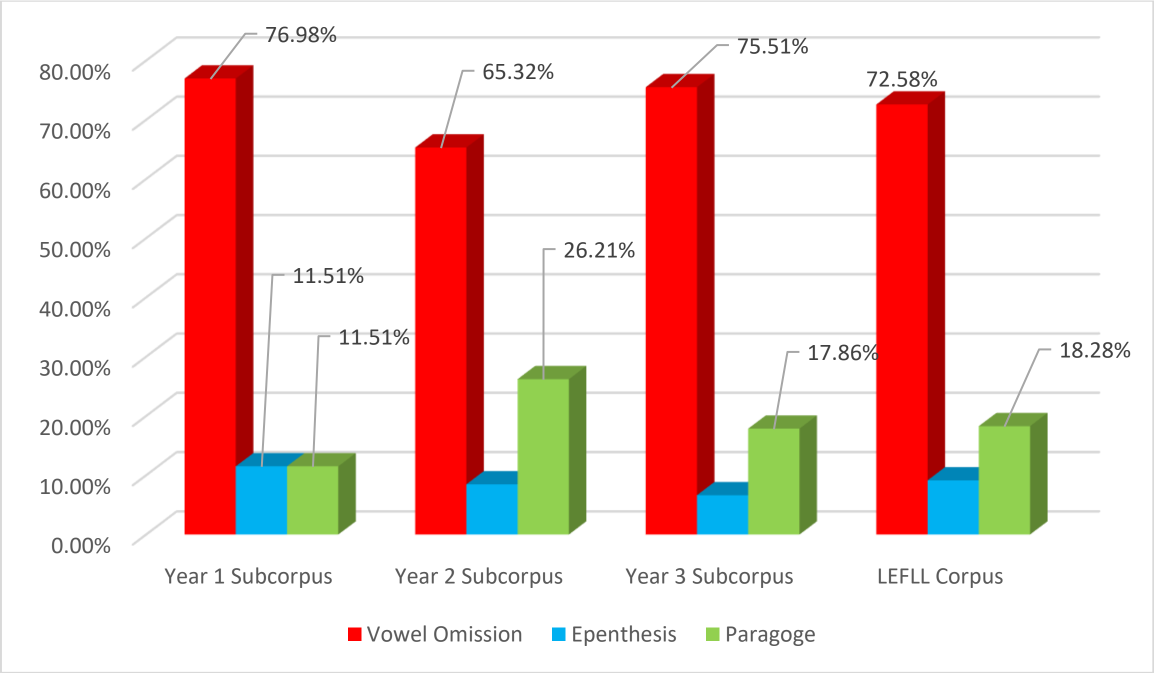


Figure 6.13: The percentages of phonological interlingual spelling errors in the LEFLL corpus

Phonological spelling categories	Year 1 Sub-corpus	Year 2 Sub-corpus	Year 3 Sub-corpus	LEFLL Corpus
Vowel Omission	(n=214) 76.98%	(n=162) 65.32%	(n=148) 75.51%	(n=524) 72.58%
Epenthesis	(n=32) 11.51%	(n=21) 8.47%	(n=13) 6.63%	(n=66) 9.14%
Paragoge	(n=32) 11.51%	(n=65) 26.21%	(n=35) 17.86%	(n=132) 18.28%
Total (n)	278	248	196	722

Table 6.14: The percentages of phonological interlingual spelling errors in the LEFLL corpus

6.4.1.2 Orthographical Interlingual Spelling Errors

Under orthographical spelling errors, six types of interlingual spelling subcategory errors were observed in the LEFLL corpus, as shown in Figure 6.14 and Table 6.15 below: 1) vowel substitution spelling errors, 2) consonant substitution spelling errors, 3) consonant doubling spelling errors, 4) silent letter omission spelling errors and 5) silent [-e] omission spelling errors.

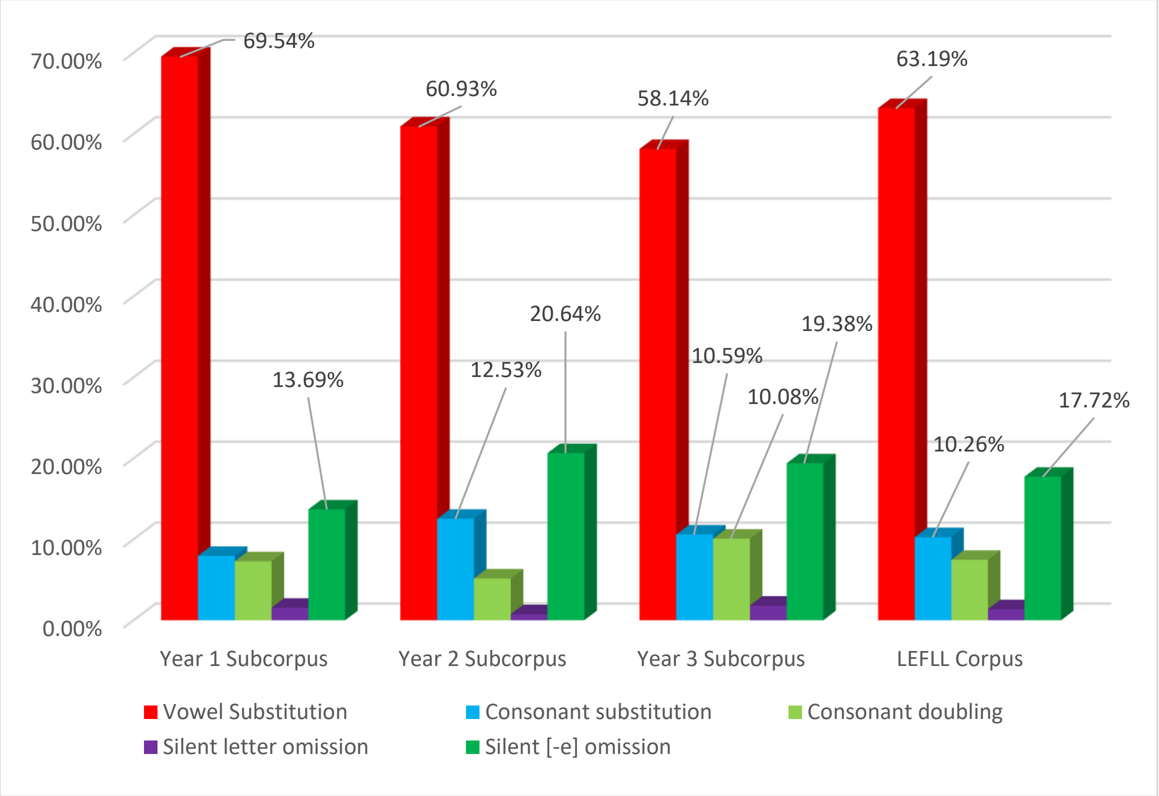


Figure 6.14: The percentages of orthographical interlingual spelling errors in the LEFLL corpus

Orthographical spelling categories	Year 1 Sub-corpus	Year 2 Sub-corpus	Year 3 Sub-corpus	LEFLL Corpus
Vowel Substitution	(n=315) 69.54%	(n=248) 60.93%	(n=225) 58.14%	(n=788) 63.19%
Consonant Substitution	(n=36) 7.95%	(n=51) 12.53%	(n=41) 10.59%	(n=128) 10.26%
Consonant doubling	(n=33) 7.28%	(n=21) 5.16%	(n=39) 10.08%	(n=93) 7.46%
Silent letter omission	(n=7) 1.54%	(n=3) 0.74%	(n=7) 1.81%	(n=17) 1.37%
Silent [–e] omission	(n=62) 13.69%	(n=84) 20.64%	(n=75) 19.38%	(n=221) 17.72%
Total (n)	453	407	387	1247

Table 6.15: The percentages of orthographical interlingual spelling errors in the LEFLL corpus

6.4.1.3 Phonological – Orthographical Interlingual Spelling Errors

Phonological-orthographical interlingual spelling errors are a combination between phonological and orthographical spelling errors, where at least one phonological and one orthographical interlingual spelling error are found in the same misspelled word. As can be seen in Figure 6.15 and Table 6.16 below, vowels represent a major problem to learners in this regard. As a result, learners tend either to omit or to substitute them.

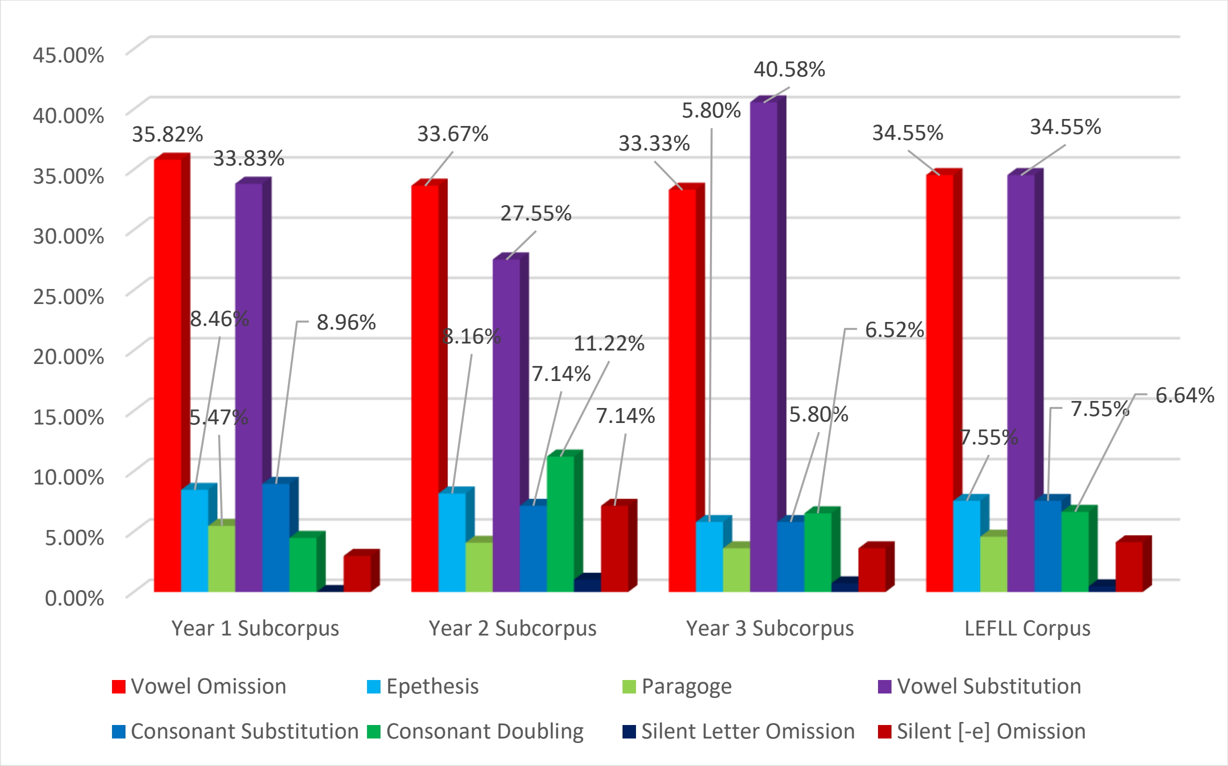


Figure 6.15: The percentages of phonological – orthographical interlingual spelling errors in the LEFLL corpus

Phonological – Orthographical spelling errors	Year 1 Sub-corpus	Year 2 Sub-corpus	Year 3 Sub-corpus	LEFLL Corpus
Phonological spelling errors partition				
Vowel omission	(n=72) 35.82%	(n=33) 33.67%	(n=46) 33.33%	(n=151) 34.55%
Epenthesis	(n=17) 8.46%	(n=8) 8.16%	(n=8) 5.80%	(n=33) 7.55%
paragoge	(n=11) 5.47%	(n=4) 4.08%	(n=5) 3.62%	(n=20) 4.58%
Orthographical				
Vowel Substitution	(n=68) 33.83%	(n=27) 27.55%	(n=56) 40.58%	(n=151) 34.55%
Consonant Substitution	(n=18) 8.96%	(n=7) 7.14%	(n=8) 5.80%	(n=33) 7.55%
Consonant Doubling	(n=9) 4.48%	(n=11) 11.22%	(n=9) 6.52%	(n=29) 6.64%
Silent letter omission	(n=0) 0.00%	(n=1) 1.02%	(n=1) 0.72%	(n=2) 0.46%
Silent [–e] omission	(n=6) 2.99%	(n=7) 7.14%	(n=5) 3.62%	(n=18) 4.12%
Total (n)	201	98	138	437

Table 6.16: The percentages of phonological – orthographical interlingual spelling errors in the LEFLL corpus

6.4.1.4 Phonological – Morphological Interlingual Spelling Errors

Phonological – morphological interlingual spelling errors are combinations between at least one phonological interlingual spelling error and any morphological spelling error. The highest percentage of phonological – morphological interlingual spelling errors, particularly among the year 1 undergraduate students, is a combination between the omission of vowels (once again a problematic area for learners in the LEFLL corpus) and inflectional or derivation morphological spelling errors.

It is worth mentioning here that very few instances of phonological-morphological interlingual spelling errors were observed in the LEFLL corpus, as can be seen in Table 6.17. This could be due to the fact that all morphological spelling errors observed in the LEFLL corpus were of the intralingual/developmental type as mentioned in Section 6.4.1 above. We will return to this type in more detail later in Section 6.4.2.1.

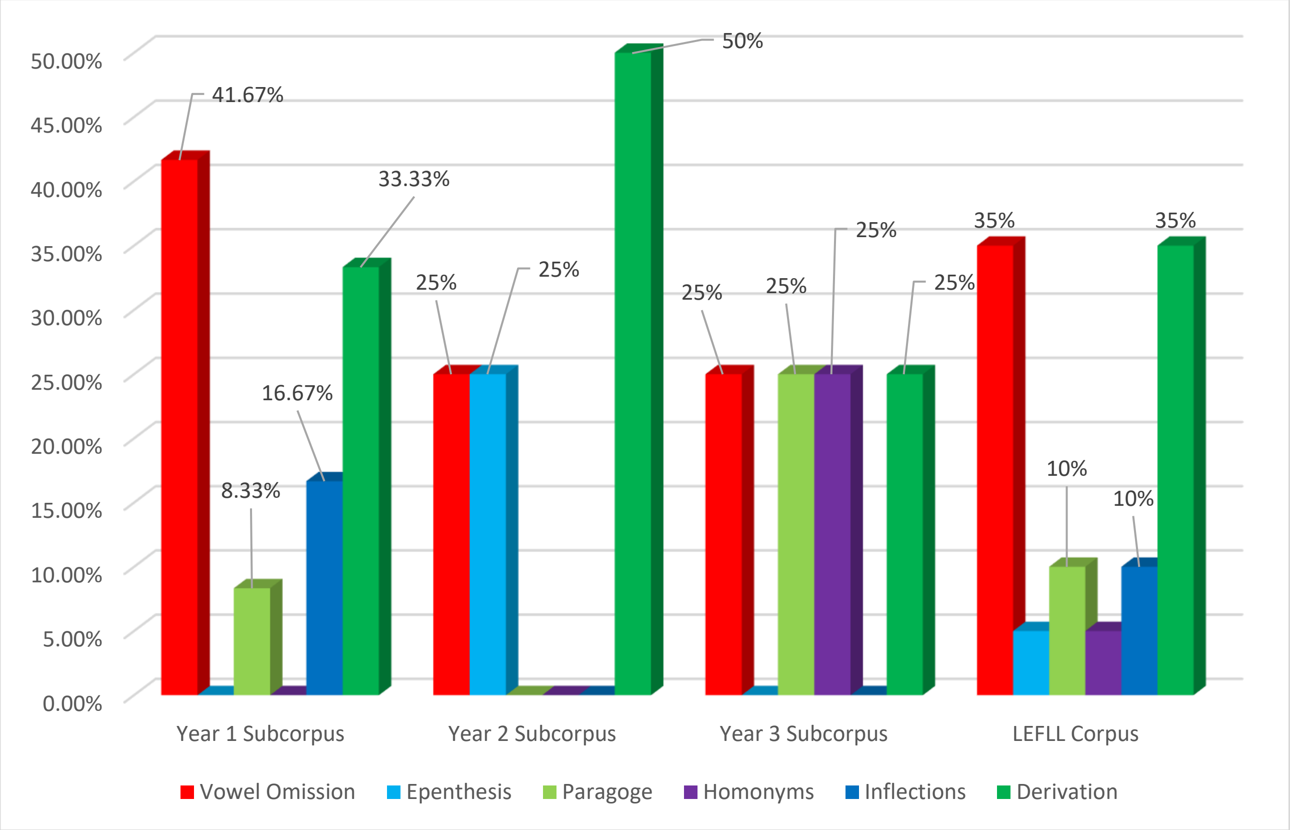


Figure 6.16: The percentages of phonological – morphological interlingual spelling errors in the LEFLL corpus

Phonological – Morphological spelling errors	Year 1 sub-corpus	Year 2 Sub-corpus	Year 3 Sub-corpus	LEFLL Corpus
Phonological				
Vowel omission	(n=5) 41.67%	(n=1) 25.00%	(n=1) 25.00%	(n=7) 35.00%
Epenthesis	(n=0) 0.00%	(n=1) 25.00%	(n=0) 0.00%	(n=1) 5.00%
paragoge	(n=1) 8.33%	(n=0) 0.00%	(n=1) 25.00%	(n=2) 10.00%
Morphological				
Homonyms	(n=0) 0.00%	(n=0) 0.00%	(n=1) 25.00%	(n=1) 5.00%
Inflections	(n=2) 16.67%	(n=0) 0.00%	(n=0) 0.00%	(n=2) 10.00%
Derivation	(n=4) 33.33%	(n=2) 50.00%	(n=1) 25.00%	(n=7) 35.00%
Total (n)	12	4	4	20

Table 6.17: The percentages of phonological – morphological interlingual spelling errors in the LEFLL corpus

6.4.1.5 Orthographical – Morphological Interlingual Spelling Errors

Orthographical – morphological interlingual spelling errors contain at least one orthographical interlingual spelling error and any morphological spelling error. As Figure 6.17 and Table 6.18 show, the highest frequency of orthographical – morphological interlingual spelling errors, particularly among learners in Year 2, features a combination of vowel substitution and inflectional or derivation morphological spelling errors. Table 6.18 also shows that the frequency of orthographical – morphological interlingual spelling errors across the three LEFLL sub-corpora is very low. This may be explained by the fact that all morphological spelling errors, observed in the LEFLL corpus, are intralingual*,* as mentioned in Section 6.4.1 above and will further be explained in more detail later in Section 6.4.2.1.

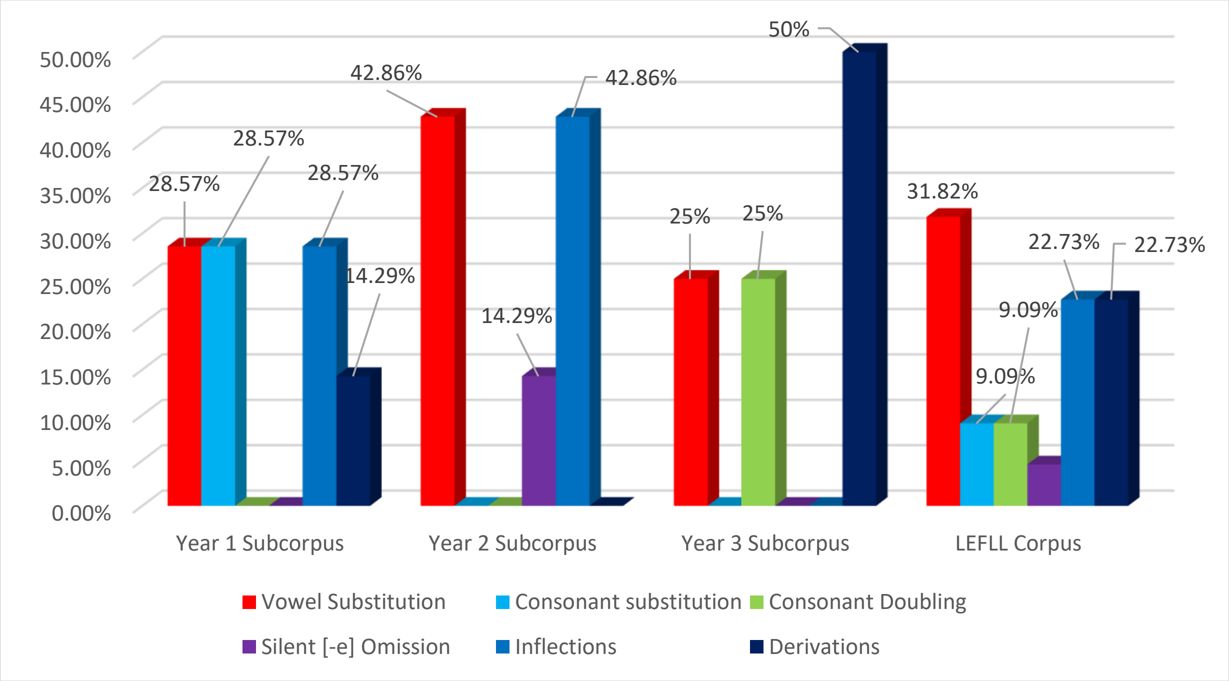


Figure 6.17: The percentages of orthographical – morphological interlingual spelling errors in the LEFLL corpus

Orthographical – Morphological spelling errors	Year 1 Sub-corpus	Year 2 Sub-corpus	Year 3 Sub-corpus	LEFLL Corpus
Orthographical spelling errors partition				
Vowel Substitution	(n=2) 28.57%	(n=3) 42.86%	(n=2) 25.00%	(n=7) 31.82%
Consonant Substitution	(n=2) 28.57%	(n=0) 0.00%	(n=0) 0.00%	(n=2) 9.09%
Consonant Doubling	(n=0) 0.00%	(n=0) 0.00%	(n=2) 25.00%	(n=2) 9.09%
Silent [–e] Omission	(n=0) 0.00%	(n=1) 14.29%	(n=0) 0.00%	(n=1) 4.55%
Morphological Spelling errors Partition				
Inflections	(n=2) 28.57%	(n=3) 42.86%	(n=0) 0.00%	(n=5) 22.73%
Derivation	(n=1) 14.29%	(n=0) 0.00%	(n=4) 50.00%	(n=5) 22.73%
Total (n)	7	7	8	22

Table 6.18: The percentages of orthographical – morphological interlingual spelling errors in the LEFLL corpus

6.4.2 Intralingual Spelling Errors

Intralingual spelling errors include all forms of non-interlingual spelling error. Whereas interlingual spelling errors are mainly marked phonological and orthographical categories, intralingual spelling errors, on the other hand, are mainly morphological, as shown in Figure 6.18 and Table 6.19 below. The Figure and Table also show that intralingual spelling errors occurred in four out of the six linguistic categories proposed by Rimrott & Heift (2005; 2008): phonological, orthographical, morphological and phonological – orthographical linguistic categories.

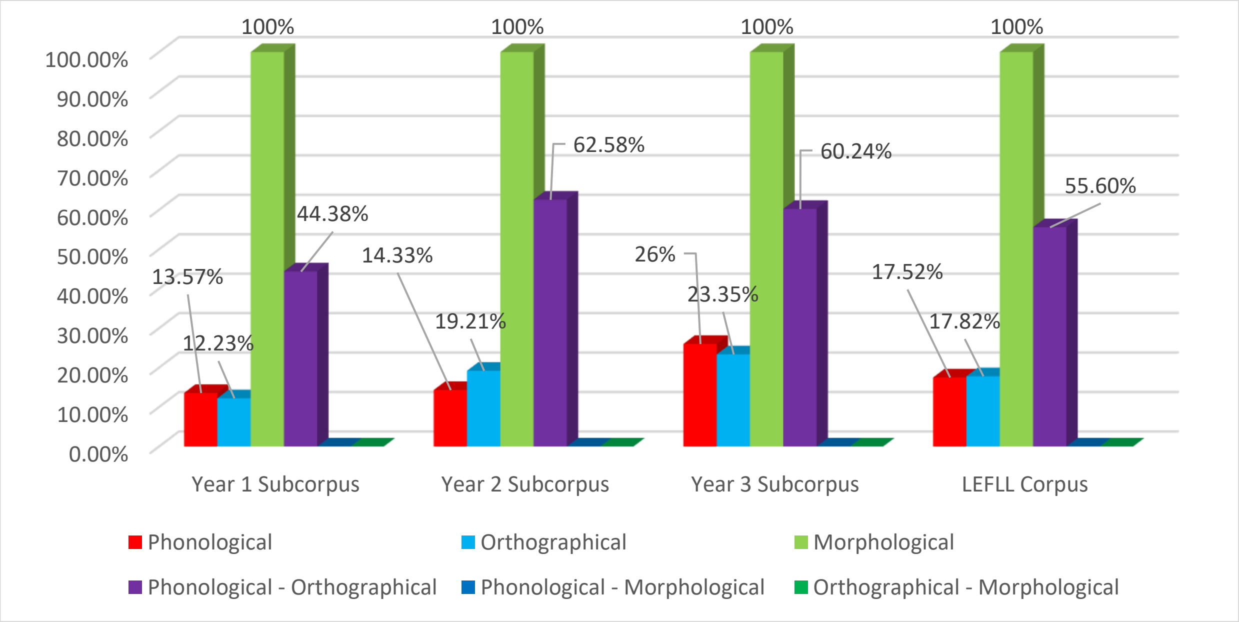


Figure 6.18: The intralingual errors in each linguistic spelling error category in the LEFLL corpus

Intralingual Spelling Errors	Year 1 Sub-corpus	Year 2 Sub-corpus	Year 3 Sub-corpus	LEFLL Corpus
Phonological	(n=48) 13.57%	(n=42) 14.33%	(n=74) 26%	(n=164) 17.52%
Orthographical	(n=56) 12.23%	(n=87) 19.21%	(n=79) 23.35%	(n=222) 17.82%
Morphological	(n=77) 100%	(n=82) 100%	(n=62) 100%	(n=221) 100%
Phonological – Orthographical	(n=79) 44.38%	(n=104) 62.58%	(n=100) 60.24%	(n=283) 55.60%
Phonological – Morphological	(n=0) 0.00%	(n=0) 0.00%	(n=0) 0.00%	(n=0) 0.00%
Orthographical – Morphological	(n=0) 0.00%	(n=0) 0.00%	(n=0) 0.00%	(n=0) 0.00%

Table 6.19: The percentages of intralingual spelling errors in each linguistic spelling error category in the LEFLL corpus

Figure 6.18 and Table 6.19 also show that there is a linear increase in the percentages of intralingual spelling errors in both phonological and orthographical spelling error categories. The Year 2 sub-corpus represents an anomalous stage only in the phonological – orthographical intralingual spelling error category but with a minor increase in the percentage of spelling errors compared to the Year 3 sub-corpus (an increase of only 2.34%).

Within each linguistic spelling error category (phonological, orthographical, morphological, etc.), intralingual linguistic spelling subcategory errors are different from the interlingual ones shown above. The following sections will present and discuss the linguistic subcategories of intralingual spelling errors in detail.

6.4.2.1 Morphological Intralingual Spelling Errors

In general, morphological intralingual spelling errors occur when learners are confused between the spellings of two different words that are pronounced similarly, such as: *'there'* and '*their'*. Choosing alternative words that are pronounced similarly to the target words may be an indication that the learners were aware of the target words that they were looking for, but due to incomplete knowledge, they conflated mixed between the target and erroneous words.

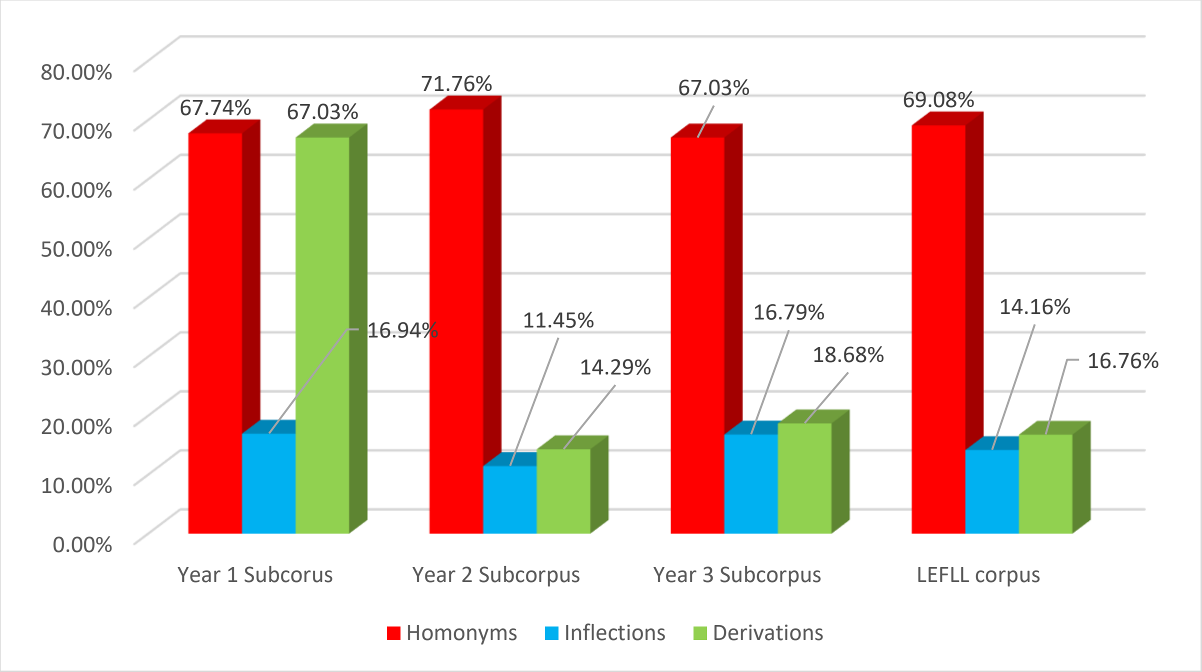


Figure 6.19: The percentages of morphological intralingual spelling errors in the LEFLL corpus

Morphological spelling errors	Year 1 Sub-corpus	Year 2 Sub-corpus	Year 3 Sub-corpus	LEFLL Corpus
Homonyms	(n=84) 67.74%	(n=94) 71.76%	(n=61) 67.03%	(n=239) 69.08%
Inflections	(n=21) 16.94%	(n=15) 11.45%	(n=13) 14.29%	(n=49) 14.16%
Derivation	(n=19) 15.32%	(n=22) 16.79%	(n=17) 18.68%	(n=58) 16.76%
Total (n)	124	131	91	346

Table 6.20: The percentages of morphological intralingual spelling errors in the LEFLL corpus

6.4.2.2 Phonological Intralingual Spelling Errors:

As regards phonological intralingual spelling errors*,* four types of spelling subcategory errors were observed in the LEFLL data. As shown in Figure 6.20 and Table 6.21 below, the learners tend to omit consonants. This is indicated in the omission of different types of consonants, such as: the omission of 'h' from the digraph *'sh', 'ch' or 'th'*. e.g., *'was'* (for '*wash*'), *'watcing'* (for *'watching'*) and the omission of 'y', e.g., '*health*' (for *'healthy'*). This may be due to an incomplete knowledge of the digraph rules; if so, there is evidence of a slight decrease in this error type from one year to the next. This may indicate that the learners do develop their knowledge of digraph rules as they move from one level to another.

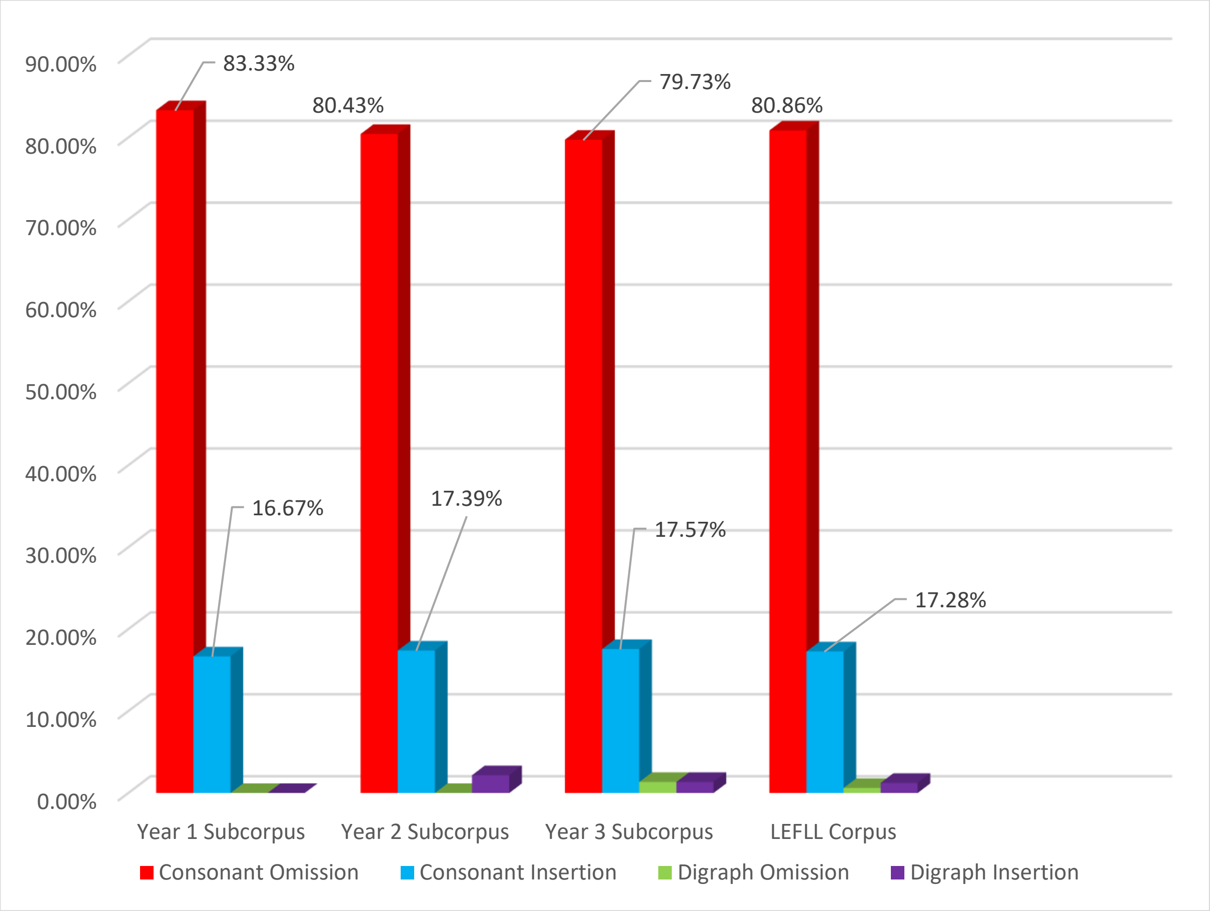


Figure 6.20: The percentages of phonological intralingual spelling errors in the LEFLL corpus

Phonological spelling errors	Year 1 Sub-corpus	Year 2 Sub-corpus	Year 3 Sub-corpus	LEFLL Corpus
Consonant Omission	(n=35) 83.33%	(n=37) 80.43%	(n=59) 79.73%	(n=131) 80.86%
Consonant Insertion	(n=7) 16.67%	(n=8) 17.39%	(n=13) 17.57%	(n=28) 17.28%
Digraph Omission	(n=0) 0.00%	(n=0) 0.00%	(n=1) 1.35%	(n=1) 0.62%
Digraph Insertion	(n=0) 0.00%	(n=1) 2.17%	(n=1) 1.35%	(n=2) 1.23%
Total (n)	42	46	74	162

Table 6.21: The percentages of phonological intralingual spelling errors in the LEFLL corpus

Consonant insertion, on the other hand, may be due to overgeneralization of orthographical knowledge. Examples of consonant insertion that maybe produced as a result of overgeneralization are: *'whant'* (for *'want'*) and *'whith'* (for *'with'*). In the previous two examples, the learners may be adopted the *'wh'* words (*what, where, when*) orthography where the letter *'h'* is needed when the word is started with the letter *'w'.*

6.4.2.3 Orthographical Intralingual Spelling Errors

Orthographical intralingual spelling errors are identified by a confusion between two digraphs (e.g., *sh* vs *ch*) or a digraph and one or three letter phonemes (e.g., *c* vs *ck; ch* vs *tch*). These types of spelling errors could be as a result of overgeneralization problems, for example, where the learners were not certain about which digraph or phoneme they had to use. They may also have produced errors of this type as a result of incomplete target language system knowledge, such as not being able to perceive a difference between two close phonemes. It is worth mentioning in this context that low frequencies of orthographical intralingual spelling errors were observed in the LEFLL corpus, as can be seen in Figure 6.21 and Table 6.22 below. On the other hand, there is an increase in the frequencies of orthographical interlingual spelling errors, as we have observed in Table 6.15 in Section 6.4.1.2 above.

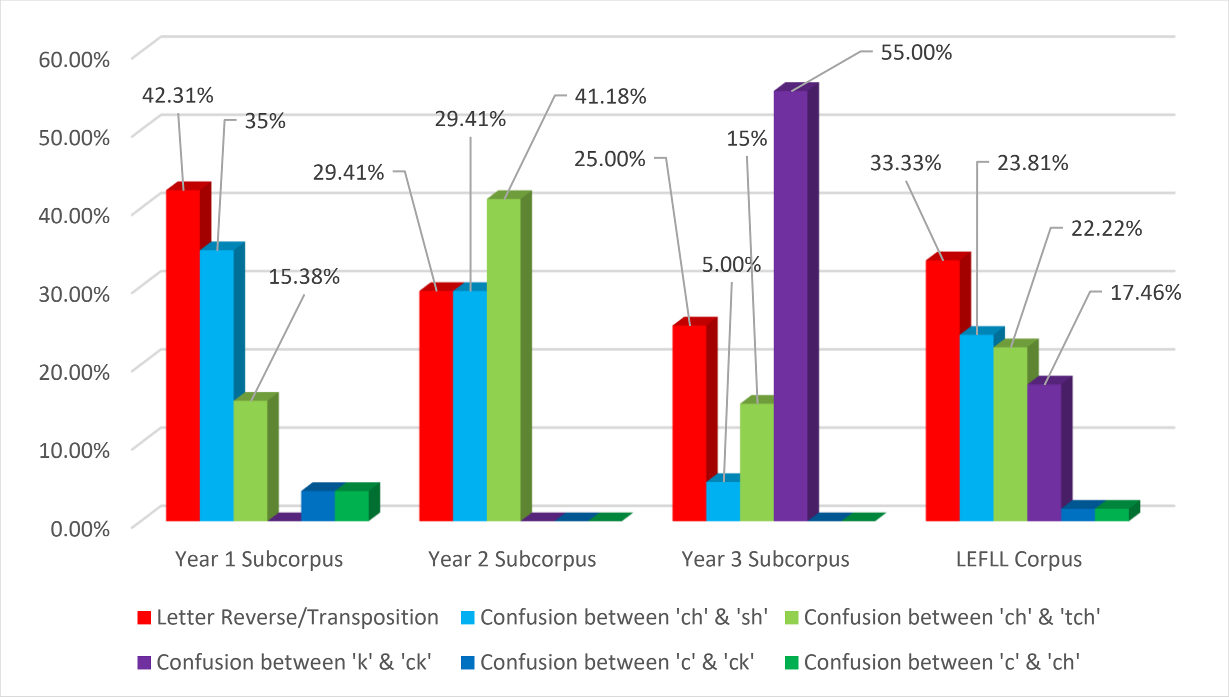


Figure 6.21: The percentages of orthographical intralingual spelling errors in the LEFLL corpus

Orthographical spelling errors	Year 1 Sub-corpus	Year 2 Sub-corpus	Year 3 Sub-corpus	LEFLL Corpus
Letter Reverse/Transposition	(n=11) 42.31%	(n=5) 29.41%	(n=5) 25.00%	(n=21) 33.33%
Confusion between ‘ch’ & ‘sh’	(n=9) 34.62%	(n=5) 29.41%	(n=1) 5.00%	(n=15) 23.81%
Confusion between ‘ch’ & ‘tch’	(n=4) 15.38%	(n=7) 41.18%	(n=3) 15.00%	(n=14) 22.22%
Confusion between ‘k’ & ‘ck’	(n=0) 0.00%	(n=0) 0.00%	(n=11) 55.00%	(n=11) 17.46%
Confusion between ‘c’ & ‘ck’	(n=1) 3.85%	(n=0) 0.00%	(n=0) 0.00%	(n=1) 1.59%
Confusion between ‘c’ & ‘ch’	(n=1) 3.85%	(n=0) 0.00%	(n=0) 0.00%	(n=1) 1.59%
Total (n)	26	17	20	63

Table 6.22: The percentages of orthographical intralingual spelling errors in the LEFLL corpus

6.4.2.4 Phonological – Orthographical Intralingual Spelling Errors

As can be seen in Figure 6.22 and Table 6.23 below, the phonological – orthographical intralingual spelling error category mainly consists of letter reverse errors, where the learners have transposed or mis-ordered two adjacent letters. According to Al-Sobhi et al. (2017), this type of error may be due to problems of phoneme-grapheme representation in English, where a grapheme may have several different phonemes. This may cause particular confusion to Arab student English learners since Arabic has an almost 1:1 phoneme-grapheme representation (Saigh & Schmitt, 2012). However, this explanation does not entirely satisfactorily account for the fact that letter reverse errors occur even more frequently in the year 3 sub-corpus than they do in the other two sub-corpora. An alternative explanation is offered by Carney (1994), who describes this type of error as an 'analogy error' (p84). That is, learners apply letter orders 'rules' that they have learned from previously acquired words to newly learned words, irrespective of whether these new words actually follow this order. If this is the case, it may at least partially explain why this problem seems to worsen as the learners progress to year 3: learners at this level are likely to have a larger stock of productive vocabulary, and thus a more highly developed (although still not always accurate) sense of what typical letter sequences exist in English orthography.

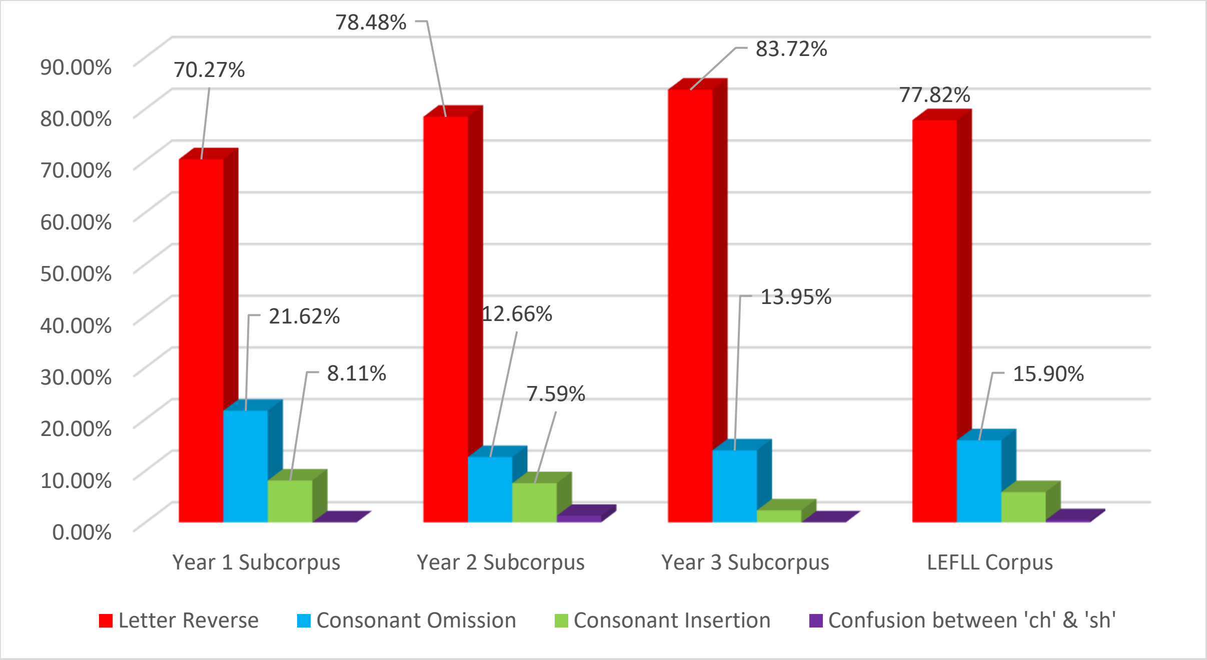


Figure 6.22: The percentages of phonological – orthographical intralingual spelling errors in the LEFLL corpus

Phonological – Orthographical spelling errors	Year 1 Sub-corpus	Year 2 Sub-corpus	Year 3 Sub-corpus	LEFLL Corpus
Letter Reverse	(n=52) 70.27%	(n=62) 78.48%	(n=72) 83.72%	(n=186) 77.82%
Phonological Spelling Errors Partition				
Consonant Omission	(n=16) 21.62%	(n=10) 12.66%	(n=12) 13.95%	(n=38) 15.90%
Consonant Insertion	(n=6) 8.11%	(n=6) 7.59%	(n=2) 2.33%	(n=14) 5.86%
Orthographical Spelling Errors Partition				
Confusion between 'ch' & 'sh'	(n=0) 0.00%	(n=1) 1.27%	(n=0) 0.00%	(n=1) 0.42%
Total (n)	74	79	86	239

Table 6.23: The percentages of phonological – orthographical intralingual spelling errors in the LEFLL corpus

Apart from letter reverse spelling errors, consonant omission spelling errors are also frequently attested in the LEFLL data. These may be interpreted more broadly as a symptom of incomplete knowledge of target language spelling rules, as explained earlier in Section 6.4.1.1.

6.5 Spelling Errors Based on the Target Modification Taxonomy

The *target modification taxonomy,* in the current thesis, is used for reliability purposes. As was discussed in Section 6.4 above*,* the highest percentage of spelling errors was attributed to language transfer factors. It was argued that this is likely to be due to the considerable variation between English and Arabic spelling systems, which seems to cause many learners to fall back on their mother language and apply its spelling rules.

It was also observed in Section 6.4 that the percentage of interlingual spelling errors decreased from one level to the next. This clearly indicates that at lower levels, learners are more likely to encounter problems with English spelling rules than are their colleagues at more advanced levels. This in turn suggests that, by year 3 of their studies, the students represented in the LEFLL corpus are beginning to rely less on first language transfer strategies when composing written texts.

The target modification taxonomy also aims to verify the results obtained in the linguistic competence taxonomy, where it was noted that performance spelling errors increased as from one level to another. In a similar pattern, the target modification taxonomy analysis is expected to show that fewer ‘edit distances’ are needed to change misspelled words to correctly formed ones. Thus, there should be an increase in the percentages of ‘single edit distance′ errors as the learners proceed.

As Figure 6.23 and Table 6.24 below show, these predictions are largely borne out: there is a slight but steady increase in the percentages of single edit distance spelling errors (and a concomitant decrease in the percentages of multiple edit distances). These results also support the findings reported in the linguistic competence taxonomy and the language influence taxonomy sections discussed earlier. In the linguistic competence taxonomy, performance spelling errors increased as the learners moved from one level to another. On the other hand, there is a slight but steady decrease in the percentage of interlingual spelling errors from one level to another.

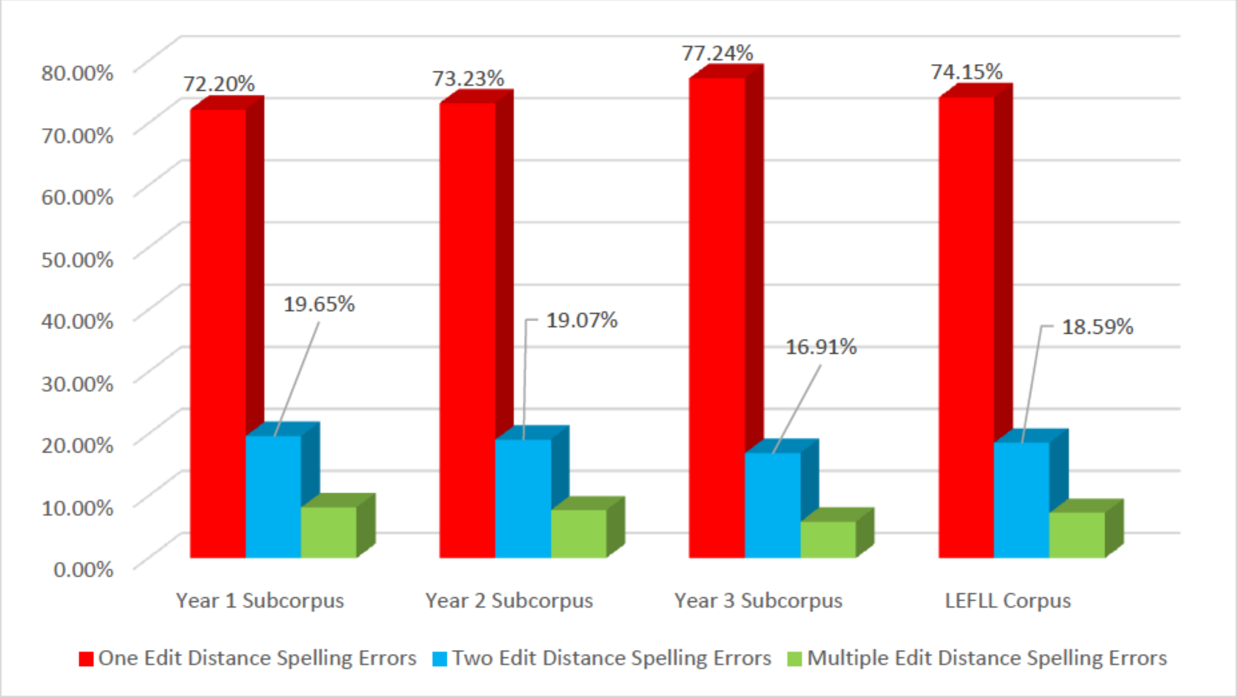


Figure 6.23: The percentages of one, two and multiple edit distances spelling errors in the LEFLL corpus

	Year 1 Sub-corpus	Year 2 Sub-corpus	Year 3 Sub-corpus	LEFLL Corpus
One Edit Distance Spelling Errors	(n=878) 72.20%	(n=818) 73.23%	(n=845) 77.24%	(n=2541) 74.15%
Two Edit Distance Spelling Errors	(n=239) 19.65%	(n=213) 19.07%	(n=185) 16.91%	(n=637) 18.59%
Multiple Edit Distance Spelling Errors	(n=99) 8.14%	(n=86) 7.70%	(n=64) 5.85%	(n=249) 7.27%
Total (n)	1216	1117	1094	3427

Table 6.24: The percentages of one, two and multiple edit distances spelling errors in the LEFLL corpus

6.6 Discussion

As was observed at the beginning of this chapter, the surface structure taxonomy revealed that omission and substitution spelling errors were more common than the other two types of non-linguistic alterations that affected the misspelled words in the LEFLL corpus (*insertion* and *transposition*). It was assumed that these observations could plausibly be attributed to first language influence, given the extreme dissimilarity between the English and Arabic phonological systems.

A review of previous spelling error analysis studies showed that researchers obtained different results when they analysed the spelling errors committed by the Arab L1 learners of English. For instance, Alhaisoni et al. (2015) obtained similar results to those reported in the current thesis when they analysed spelling errors in the writing of 122 students enrolled in an intensive English language program at the University of Ha'il in Saudi Arabia. Their analysis revealed that omission spelling errors scored the highest (39.6%), followed by substitution spelling errors (34.9%). Alhaisoni et al. attributed the prevalence of omission errors to the profound differences between English and Arabic orthography. In Arabic, words are written in precisely the same way they are pronounced, whereas in English words are often written quite differently from how they are pronounced (Alhaisoni et al., 2015). Similarly, Othman (2018) found substitution and omission spelling errors to be the most prevalent (36.7% and 33.3% respectively) of all types of spelling errors when he analysed spelling errors committed by English learners at the University of Tabuk, Saudi Arabia.

However, it must be conceded that some other studies of spelling error analysis of Arab student English learners have reported different results. For instance, Al-Oudat (2017) found that substitution and insertion spelling errors to be the most frequent (at 33.70% and 32.61% respectively) when she analysed spelling errors committed by 65 English-Major undergraduate students at Al-Balqa Applied University in Jordan. In another study, Ibrahim (2018) observed that omission and addition spelling errors accounted for the highest percentages of spelling errors (33.64% and 32.24% respectively) out of the total number of spelling errors committed by 100 English learners at Gezira University, Sudan.

The first language transfer argument is also somewhat weakened by the fact that some other studies in non-Arabic context have also reported similar results to the current thesis. For instance, Fitria (2020) reported that omission and substitution spelling errors constituted the highest percentages of spelling errors when she analysed spelling errors in the writing of Indonesian students English learners. Thus, the high percentages of omission and substitution spelling errors observed in the current thesis may not on their own constitute strong enough evidence for claiming that the learners’ first language plays a central role in the commission of spelling errors in English. As stated earlier, this section will discuss the reasons why different types of spelling errors (epenthesis, paragoge, vowel omission, vowel substitution, consonant substitution, consonant doubling, silent letter omission and silent [-e] omission) were classified as interlingual spelling errors:

6.6.1 Epenthesis

Epenthesis spelling errors refer to the addition of a vowel or consonant either at the beginning of the word or between two sounds. The findings of the current thesis for this category are consistent with the findings of several previous studies, which also reported epenthesis spelling errors in the English writing of Arab student English learners and attributed these types of spelling errors to the first language interference. For instance, Randall and Groom (2009) observed a great deal of *epenthesis* spelling errors in their 16-year old Arab learner corpus, particularly in the insertion of a vowel between two consonants in two clusters, reflecting the *consonant vowel consonant vowel* (CVCV) structure typical of Arabic. In another study, Sadhwani (2005) also noticed the appearance of *epenthesis* spelling errors among a cohort of

15 Arab student English learners of Emirati origin. According to Saigh and Schmitt (2012), first language transfer is by far the most plausible explanation for the wide spread of *epenthesis* phenomena among Arab learners of English.

6.6.2 Paragoge

Paragoge occurs when a vowel is added at the end of the word which normally ends with a consonant sound e.g., *watche* (for *watch*). The results of the current thesis for this error type are consistent with the findings of Sadhwani (2005). In his thesis, Sadhwani observed the frequent addition of vowels among Arab learners of English in their speech. For Sadhwani, errors such as *vowel omission, vowel lengthening, vowel monothongisations/dipthongisation, epenthesis* and *paragoge* all confirm L1 as a phonological barrier.

6.6.3 Vowel Omission

Vowel Omission spelling errors, the phonological system of Arabic contains eight vowels which are classified into three categories: three short (known as non-alphabetic signs (Wickens, 1980) or diacritical marks (Abu-Rabia & Sammour, 2013)), three long and two diphthongs vowel sounds (Swan & Smith, 2001). In written Arabic, short vowels are typically omitted altogether. In fact, it was observed that the presence of these three short vowels in Arabic reading tasks actually slowed down the reading speed of Arabic native speakers, as it caused them to spend more time focusing on each word in the texts which did feature the usually omitted diacritical marks (Roman & Pavard, 1987 cited in Hayes-Harb, 2006). Therefore, Arabic L1 speakers are known for having what is sometimes described as ‘vowel blindness’ (Randall & Groom, 2009:8). In English reading and writing, Arabic L1 speakers do indeed seem to apply this strategy, ignoring vowels and relying instead on consonantal segments (Ryan & Meara, 1991). Apparently, this was the strategy the learners adopted in the LEFLL corpus, as reflected in the fact that the percentages of omission spelling errors were higher than any of the other types of phonological interlingual spelling errors.

6.6.4 Vowel Substitution

Vowel Substitution spelling errors constituted almost two thirds of all interlingual spelling errors in the orthographical linguistic category. Spelling errors of this type occur as a result of confusion between vowel sounds in English. Due to the limited number of vowels in Arabic (which has only 8 vowels and diphthongs, compared with 22 vowels and diphthongs in English (Swan & Smith, 2001)), it is reasonable to assume that this may have caused some confusion among the learners represented in LEFLL. This finding is consistent with previous research studies, which have also observed vowel confusion spelling errors among Arab student English learners (e.g., Ibrahim, 1978; Haggan, 1991; Sadhwani, 2005; Saigh & Schmitt, 2012; Alhaisoni et al., 2015; Othman, 2018).

6.6.5 Consonant Substitution

Consonant Substitution spelling errors may be attributed to the absence of the target consonant phonemes from Arabic. In this case, Arabs borrow alternative consonants of close phonemes to the target ones and which have similar consonant sounds in Arabic. For instance, the substitution between the consonant letter '*p*' with *'b'* in the following example: *'back'* (for *pack*) could be due to the absence of an Arabic equivalent to the English letter *'p'*. In this example, the phoneme /b/ has an equivalence represented in the letter 'ب' (Ba) in Arabic whereas the phoneme /p/ does not exist in Arabic. The substitution between consonant letters was widely observed in the LEFLL corpus where a number of incorrect consonant letters have been used instead of the correct ones because they both have close phonemes, such as using 'g' instead of 'j' in *'engoy'* (for *enjoy*) and *'c'* instead of *'s'* in *'univercity'* (for *university*), etc. It was also observed that the learners tend to replace the digraph with a consonant letter, such as the replacement of *'ph'* with '*f*' in *'prefat'* (for *prophet*).

These findings are in line with those of other studies of spelling errors among Arab student English learners. For instance, In the analysis of spelling errors committed by Bahraini English learners, Allaith and Joshi (2011) observed the substitution of /p/ and /v/ phonemes with /b/ and /f/ respectively. Allaith & Joshi concluded that in the absence of some phonemes from the Arabic language, Arab learners of English tend to rely on the closest sounds to the target phonemes that do exist in their language. Similarly, Othman (2018) also observed that Saudi English-Major students at the University of Tabuk confused /b/ and /p/. He also attributed this confusion to the absence of the letter /p/ from Arabic. In another study of spelling errors among Saudi student English learners, Alhaisoni et al. (2015) also observed students substituting consonants, such as [c] for [s] and vice versa, e.g., 'nise' for 'nice' and 'sentar' for 'centre/center'.

6.6.6 Consonant Doubling

In written Arabic texts, consonant doubling is sometimes formed by placing a diacritic sign called '*Shedda'* over a single consonant instead of actually doubling the consonant. It seems that the learners in the LEFLL corpus applied a similar strategy, using single consonants instead of doubling them when required by English spelling rules, as in the following examples: *tenise* (for *tennis*), *wning* (for *winning*), *midel* (for *middle*), etc. Once again, these results match those of earlier studies of spelling errors among Arabic L1 learners of English. For example, in an analysis of spelling errors committed by Sudanese English learners, Ibrahim (2018) observed the omission of numerous consonants in words where the consonants should be doubled. Ibrahim attributed this type of spelling error to L1 interference. In another study, Alsaawi (2015) found consonant doubling to be the most common type of spelling error committed by Saudi student English learners. Alsaawi also attributed this type of error to first language interference, explaining that Arabic language does not have consonant doubling.

It is important to note here that consonant doubling has been found to be a common type of interlingual spelling error among English learners of other mother language backgrounds than Arabic. For instance, Bestgen and Granger (2011) observed that spelling errors featuring a single consonant instead of a double consonant are frequent in essays written by Spanish learners of English: in fact, over 80% of single letter errors were observed in the Spanish sub-corpus of ICLE. On the face of it, this could be seen as weakening the Arabic L1 transfer arguments that I have been putting forward in this section. On closer inspection, however, it turns out that Bestgen & Granger's findings actually support for the L1 transfer argument, albeit in a somewhat indirect way. Bestgen & Granger's (2011) study actually compared three subcorpora of ICLE: the French and German subcorpora as well as the Spanish one. Crucially, they observed that the failure to double consonants when required was observed only in the Spanish subcorpus; it was not an issue for French or German L1 speakers. Bestgen & Granger therefore argued that this type of error is likely to be due to L1 interference, explaining that the equivalents of these types of spelling errors in Spanish language are always words with a single consonant letter (e.g., *'communication', 'community', 'diference'* correspond to *comunicación*, *comunidad*, *diferencia* respectively). In other words, the fact that the same kind of spelling error may be observed in the writing of learners from a different L1 background does not necessarily undermine the argument that this error may be caused by first language influence.

6.6.7 Silent Letter Omission Error

Silent letter omission spelling errors may also be plausibly attributed to L1 vs L2 contrasts. Given that Arabic has an almost 1:1 phoneme-grapheme representation (Saigh & Schmitt, 2012), Arabic native speakers are thus accustomed to spelling words in such a way that each letter represents a sound. For learners from such a background, silent letters (e.g., *'k'* in '*know'*) are very likely to be problematic, and thus to be frequently omitted. As we have seen, this is borne out in the LEFLL data reviewed above. The findings of the current study are also in line with a number of previous spelling error studies, all of which found consonant omission errors to be typical of Arabic L1 learners of English (e.g., Al Jayousi, 2011; Al-Sobhi et al., 2017; Altamimi et al., 2018; Ibrahim, 2018).

6.6.8 Silent [–e] Omission Error

Along with the *silent letter* omission spelling errors, the *Silent* [*–e*] *Omission Spelling Errors* may also be plausibly attributed to the role of the learners' first language. Specifically, the 1:1 phoneme-grapheme representation strategy may have also been applied in the case of silent [–e] omission spelling errors. Sadhwani (2005) observed this phenomenon in his study of 15 Arab student English learners. The findings of the current study are also consistent with Othman (2018), who found the omission of the silent letter [-e] to be the most common type of omission spelling error committed by Saudi students. Othman attributed this type of error to the fact that Arabic is written the way it is pronounced, whereas this is often not the case with English. The finding is also in line with other studies which observed the omission of the silent letter [-*e*] among Arab speakers' English learners and attributed this to first language interference (e.g., Alhaisoni et al., 2015; Al-Oudat, 2017; Altamimi et al., 2018).

In summary, it is entirely reasonable to regard the interlingual linguistic spelling subcategory errors discussed above as strong evidence for the existence of Arabic first language influence on English spelling errors. The high percentages of interlingual spelling errors observed here may even support the claim that Arab learners of English may encounter more serious problems with English spelling than do English learners from any other first language background (Saigh & Schmitt, 2012), given the profound differences between English and Arabic phonological and orthographical systems.

The discussion also revealed that vowels may be considered the central problem area for Arab English learners, and the main cause of the high percentages of interlingual spelling errors reported in this chapter. The huge difference between English and Arabic vowel sounds (22 vowels and diphthongs in English versus 8 vowels and diphthongs in Arabic (Swan & Smith, 2001)) and the insignificance of the three short vowels in Arabic may seem to play a crucial role in the proliferation of vowel sound spelling errors observed in the current research. This high percentage of vowel spelling errors is also consistent with previous research findings on Arab L1 learners (cf. Randall & Groom, 2009; Saigh & Schmitt, 2012; Abu-Rabia & Rana, 2013).

Conclusion

This chapter aimed to conduct a comprehensive and quasi-longitudinal analysis of spelling errors identified in the LEFLL sub-corpora. To achieve this purpose, spelling errors were first classified based on the surface structure taxonomy proposed by Dulay et al. (1982). The aim of applying the surface structure taxonomy in classifying the spelling errors was to study the four types of non-linguistic alterations that affect misspelled words, namely: omission, insertion, substitution and transposition. The analysis revealed that omission spelling errors constituted the highest proportion of spelling errors in the LEFLL corpus and across the LEFLL sub-corpora, followed by substitution spelling errors. Transposition spelling errors, in contrast, were comparatively rarely attested in the current data. The analysis also showed that year 2 represented an anomalous stage between year 1 and year 3 in most spelling error types.

Following the classification of spelling errors based on the surface structure taxonomy, the spelling errors extracted from LEFLL were then classified into four dimensions, each based on a linguistically motivated taxonomy: these were the linguistic subsystem taxonomy, the linguistic competence taxonomy, the language influence taxonomy and the target modification taxonomy. In the linguistic subsystem taxonomy, spelling errors were classified according to the standard POMAS (Phonological, Orthographical and Morphological Assessment of Spelling) spelling error classification scheme. This analysis revealed that orthographical spelling errors constituted the highest proportion of spelling errors, followed by phonological spelling errors, whereas phonological – morphological and orthographical – morphological errors constituted the lowest proportions of spelling errors across the three LEFLL sub-corpora. This variation in the proportions of spelling error types across the three sub-corpora were attributed either to the crucial role of the learners' first language, Arabic, (in the case of interlingual spelling errors), or to developmental processes specific to the acquisition of the target language, English (i.e., intralingual spelling errors).

To verify which language (Arabic or English) has led to an increase in the percentages of orthographical and phonological spelling errors, spelling errors were first classified into performance and competence types. Performance spelling errors are accidental and do not reflect the role of the learners' first language or target language development factors.

Competence spelling errors, in contrast, are due to the role of the learners' first language and/or factors relating to target language development. The analysis of this category revealed that there is an increase in the percentages of performance spelling errors vs a decrease in the percentages of competence spelling errors from one level to the next. As a by-product of the linguistic competence taxonomy, an increase in the percentages of performance spelling errors vs decrease in the percentages of competence spelling errors provided further evidence of the shortcomings of Thewissen's approach to error counting, and supported the claim that there is a need for learner corpus researchers to conduct at an early stage in any analysis an empirical investigation to verify which words in the learner corpus the language learners have actually misspelled as the PFEC approach introduced in this thesis calls for.

The competence spelling errors extracted from LEFLL were then classified, using the language influence taxonomy, into interlingual (due to language transfer) and intralingual (due to the target language development) spelling errors. The analysis revealed that the interlingual spelling errors constituted the highest percentages of spelling errors in the LEFLL corpus and across the three LEFLL sub-corpora. This was mainly due to the variation between English and Arabic phonological and orthographical systems. The analysis also showed that there is a steady decrease in the percentages of interlingual spelling errors vs an increase in the percentages of intralingual spelling errors. This may indicate that learners at lower levels rely more heavily on their first language spelling strategies than do learners at more advanced levels. A thorough analysis indicated that the learners encountered major problems with English vowels due to the huge differences between English and Arabic vowel systems. It was observed that the learners tend to misapply the same spelling strategies they use in their first language, such as the *omission of vowels.*

For reliability purposes, the target modification taxonomy was used where spelling errors were classified into one edit distance, two edit distances and multiple edit distances. It was expected that there should be an increase in the percentages of one edit distance errors from one level to another, and that this may reflect an improvement in overall spelling ability. The analysis showed that, across the three-year groups represented by the LEFLL corpus, the percentages of one edit distance spelling errors did indeed start increasing, and that there was a corresponding decrease in the percentages of two edit distances and multiple edit distances spelling errors.

In the next chapter, the analytical focus of thesis turns from spelling to noun phrase errors in the LEFLL corpus. Throughout the chapter, I will argue for a significant role for L1 transfer influence at this level.

CHAPTER 7: ANALYSIS OF NOUN PHRASE SUBCATEGORY ERRORS IN THE LEFLL CORPUS

Introduction

In the field of error analysis, noun phrase errors constitute an important area of research. A number of studies have either focused on and analysed errors of a specific constituent in a noun phrase (e.g., articles, pronouns, possessive determiners, etc.) or have analysed errors of noun phrase constituents within a wider perspective. A good example of the former is Crompton (2011), which analyses article errors in the Arabic Learner English corpus, a sub-corpus of the International Corpus of Learner English (ICLE). As regards the latter, a typical example is Qomariana et al. (2019), who studied the role of the learners' first language in grammatical errors produced by students at Udayana University, Indonesia. Among the seven types of grammatical errors identified in their study, Qomariana et al. also identified pronouns, singular/plural and article errors.

This chapter occupies a middle ground between these narrow and broad perspectives, by carrying out a focused but comprehensive and quasi-longitudinal analysis of noun phrase subcategory errors identified in the LEFLL sub-corpora. The chapter aims to diagnose the possible causes of these types of errors, focusing on both interlingual and intralingual error types. To perform a comprehensive and quasi-longitudinal noun phrase error analysis, the LEFLL corpus was manually transcribed and error tagged by the researcher. The error tagging system is hierarchical and was adopted from Dagneaux et al. (1998) (see Chapter 4 for more information about the LEFLL corpus design and error tagging). Following the error tagging process, the noun phrase subcategory errors were retrieved and analysed using AntConc and classified into interlingual and intralingual error types. The analysis revealed that nearly 50% of all noun phrase errors in the LEFLL corpus are classifiable as interlingual.

7.1 Noun Phrase Error Analysis – the CEA Approach

Noun phrase errors were analysed based on the computer-aided error analysis approach (CEA), as described in detail in Chapter 4. Each noun phrase subcategory error (e.g., determiner omission error, pronoun addition error, wrong word class, etc) was error tagged by inserting an error tagging code that describes the alteration that affected the noun phrase subcategory error in question. The error tagging process was performed manually via Dexter Coder Tools (see Section 4.3.1 for an overview of Dexter)) and the error tagging system was adopted from Dagneaux et al. (1998) (see Section 4.2.2 in Chapter 4 for more information about this system). As discussed in Chapter 4, each noun phrase error tagging code consists of two or three levels depending on the information needed to describe the noun phrase subcategory error in question. The first level denotes the major category code, which describes the phrase type (in this chapter, noun phrases) followed by one or two subcategory codes. For example, the noun phrase error tagging code NPLTR consists of two levels only. The first level 'NP' represents the major category code 'Noun Phrase' followed by the subcategory code 'LTR', denoting "*Literal Transfer*" from the learners' L1. It was assumed that a two-level error tagging code would be sufficient to describe this type of noun phrase subcategory error. The subcategory code 'LTR' is reserved exclusively for describing instances of literal transfer from the learners' L1.

To take another example, the noun phrase error tagging code (NPDETRED) consists of three levels. The major category code 'NP' is followed by the subcategory codes 'DET', which denotes the noun phrase constituent '*determiner',* and 'RED' describes the alteration that affected the determiner (*redundant/insertion*) based on the surface strategy taxonomy proposed by Dulay et al. (1982). For the latter type of noun phrase subcategory error, it was necessary to use a three-level error tagging code as the *determiners* were affected in three different ways in the LEFLL corpus, namely, through *omission*, *addition* and *substitution*. It was therefore necessary in this and similar cases to add a third level of subcategory coding in order to describe the non-linguistic alteration (the surface structure change) that affected the determiner. In addition to the different types of noun phrase error tagging codes, the process of noun phrase error tagging also involved inserting the following:

- 'NP': The 'NP' tagging code has been used to tag each noun phrase in the LEFLL corpus.
- 'NPE': The 'NPE' tagging code has been used to tag the range/series of noun phrase constituents/words where a noun phrase subcategory error was spotted.
- 'NPPOTE': The 'NPPOTE' tagging code has been used to tag the noun phrase constituent that has the potential for error (see Section 4.2.2 in Chapter 4).
- Where necessary, corrections of the different types of noun phrase subcategory errors were provided.

Inevitably, the error tagging process was a time consuming one, as it involved the manual analysis of 559 exam essays (with a total word count of 60131 tokens). Each text was converted to DeXML file format (the file format that the Dexter Coder Tool requires for tagging corpus files) via the Dexter Converter Tool, and tagging codes and corrections were inserted via the Dexter Coder Tool (see Appendix E for a full list of the tagging codes that were used in the LEFLL corpus). The corrections were added in order to provide more detail about the reason why a noun phrase subcategory error had been classified under a specific noun phrase subcategory error.

In total, 62645 noun phrase error tagging codes and corrections were manually inserted into the LEFLL corpus for the noun phrase error analysis, as shown in Table 7.1 below. Following the error tagging process, each error tag in each tagged file in the LEFLL corpus was separately retrieved and converted back from DeXML to .txt format via the Dexter Search Tool (see Section 4.3.2.1 for a full discussion of the Dexter Search Tool), so that the data could be analysed using AntConc. AntConc was used to obtain concordances of the different types of noun phrase subcategory errors in the corpus. The Noun phrase errors were classified into 17 different subcategory types as shown in Table 7.1 below (see Appendix E for a full list of noun phrase tagging codes.

	Year 1 Sub-corpus	Year 2 Sub-corpus	Year 3 Sub-corpus	LEFLL Corpus
NP	7118	6490	6184	19792
NPE	1117	1537	1088	3742
NPPOTE	11029	11184	9892	32105
NPDETOM	466	645	378	1489
NPDETRED	74	105	87	266
NPDETW	20	14	21	55
NPGER	14	14	4	32
NPLTR	1	7	4	12
NPNO	106	137	158	401
NPNOM	17	26	16	59
NPNOREG	1	8	10	19
NPNRED	15	10	6	31
NPNW	35	53	64	152
NPPRONOM	208	252	141	601
NPPRONRED	28	82	50	160
NPPRONW	8	11	11	30
NPWWC	17	47	70	134
NPWWO	60	48	20	128
NPAPOSTOM	15	60	15	90
NPAPOSTRED	25	13	26	64
CORRECTIONS	969	1369	945	3283
TOTAL	21343	22112	19190	62645

Table 7.1: The distribution of noun phrase tagging codes and corrections in the LEFLL corpus.

Following the classification of the noun phrase errors into 17 noun phrase subcategory errors, each noun phrase subcategory error was classified either as an interlingual error (if it reflected the role of the learner's L1) or as an intralingual error (if it indicated the target language development). The following section will briefly present the interlingual noun phrase errors and subcategory errors. Section 7.4 below will then discuss the reasons why these noun phrase subcategory errors were classified as interlingual rather than intralingual types. The discussion will be supported by reference to previous research findings as well as by examples from the LEFLL data.

7.2 Interlingual Noun Phrase Errors and Subcategory Errors

The analysis revealed, as shown in Figure 7.1 and Table 7.2 below, that the average percentage of interlingual noun phrase errors in the LEFLL corpus is slightly less than 50%. The Figure and Table also show that there are no linear changes in the percentages of interlingual and intralingual noun phrase errors from one level to another. Year 2 once again represented an anomalous stage between the years 1 and 3: the percentage of interlingual noun phrase subcategory errors in the Year 2 sub-corpus is higher than that of either the Year 1 or Year 3 sub-corpora, whereas the percentage of intralingual noun phrase subcategory errors is the lowest in the Year 2 sub-corpus.

In all other respects, however, these results are different from the results obtained from the analysis of spelling errors. As we saw in Chapter 6, almost two thirds of spelling errors in the LEFLL corpus were interlingual and there was a steady decrease in the percentages of interlingual spelling errors associated with an increase in the percentages of intralingual spelling errors across the three LEFLL sub-corpora. These observations were interpreted as evidence that the learners rely less on their L1 as they move from one level to another.

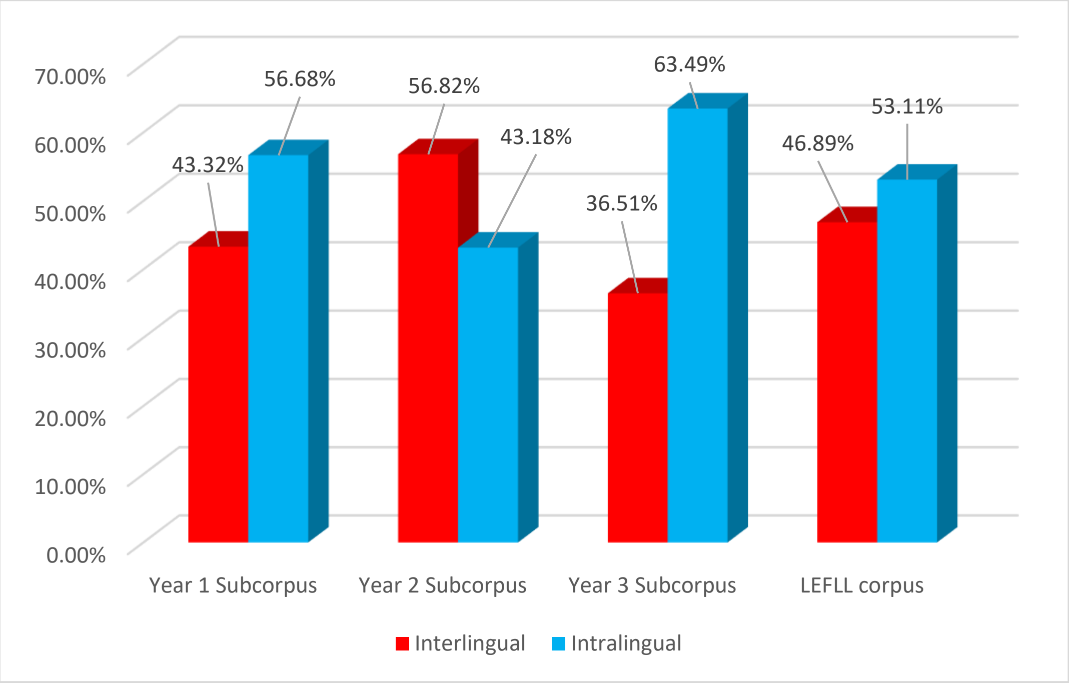


Figure 7.1: The interlingual vs intralingual noun phrase errors in the LEFLL corpus

The Figure and Table also show that the Year 2 sub-corpus deviated from the standard distribution of interlingual and intralingual noun phrase subcategory errors within the same sub-corpus. Whereas the percentages of interlingual noun phrase errors in the Year 1 and Year 3 sub-corpora (43.32% and 36.51% respectively) are less than the percentages of intralingual noun phrase errors in the same sub-corpora, the percentage of interlingual noun phrase errors in the Year 2 sub-corpus (56.82%) is higher than the percentage of intralingual noun phrase errors observed in the same sub-corpus. i.e., learners in the year 2 sub-corpus seem to rely more heavily on their L1 Arabic when constructing English noun phrases than do the learners in years 1 and 3. This may indicate that first language influence does not always follow steady changes (e.g., an increase or decrease in the percentages of interlingual errors) in different types of language learners’ errors as we have observed in the analysis of spelling error types.

This finding may also help to explain why Year 2 represented an anomalous stage between years 1 and 3 when the percentages of noun phrase errors were calculated based on the PFEC approach (as discussed in chapter 5, Section 5.3). There was no steady change (either increase or decrease) in the percentages of noun phrase errors from one level to another. As shown earlier in Figure 5.6 and Table 5.9 (Section 5.3), the percentage of noun phrase errors in the Year 2 sub-corpus (13.74%) was higher than the percentages of noun phrase errors in both the Year 1 and Year 3 sub-corpora (10.11% and 8.61% respectively), unlike the percentages of spelling errors obtained via the PFEC approach. It was observed that there was a small but steady decrease in the percentages of spelling errors from one level to another, at 6.72%, 5.95% and 5.74% in the Year 1, Year 2 and Year 3 sub-corpora respectively. This may support the claim, in Chapter 6, that the steady decline in the percentages of interlingual spelling errors has led to the steady decline in the percentages of spelling errors when they were calculated, based on the PFEC approach, as observed in Figure 5.6 and Table 5.9 in Section 5.3. On the other hand, this may alternatively be seen as indicating that the learners' first language influence does not necessarily follow a linear course in different error categories.

If we discount the Year 2 sub-corpus as an anomaly, however, there is a clear and steady decline in the percentages of interlingual noun phrase errors in the Year 3 sub-corpus (36.51%) compared to the Year 1 sub-corpus (43.32%) associated with an increase in the percentage of intralingual noun phrase errors in the Year 3 sub-corpus (63.49%).

	Inter.	Intra.	Total	Inter. %	Intra. %
Year 1 Sub-corpus	483	632	1115	43.32%	56.68%
Year 2 Sub-corpus	875	665	1540	56.82%	43.18%
Year 3 Sub-corpus	398	692	1090	36.51%	63.49%
LEFLL Corpus	1756	1989	3745	46.89%	53.11%

Table 7.2: The interlingual vs intralingual noun phrase errors in the LEFLL corpus

If we now look at the distribution of interlingual vs intralingual errors across the subcategory levels of noun phrase errors in the LEFLL corpus and across the three LEFLL sub-corpora as shown in Figure 7.2 below and in more detail in Table 7.3 towards the end of this section, we can see that the role of the learners' first language is evident in most noun phrase subcategory errors. The Figure and Table show that 13 out of the 17 noun phrase subcategory errors were either purely, predominantly or partially due to the role of the learners' L1, where the percentages of interlingual vs intralingual noun phrase subcategory errors varied from one noun phrase subcategory error to another. For instance, all apostrophe omission errors in the LEFLL corpus, are purely interlingual whereas 82.95% of wrong word order errors are interlingual and only 16.45% of using the wrong noun errors are interlingual (see Sections 7.4.10, 7.4.13 and 7.4.9 below for examples of apostrophe omission, wrong word order and using the wrong nouns errors respectively).

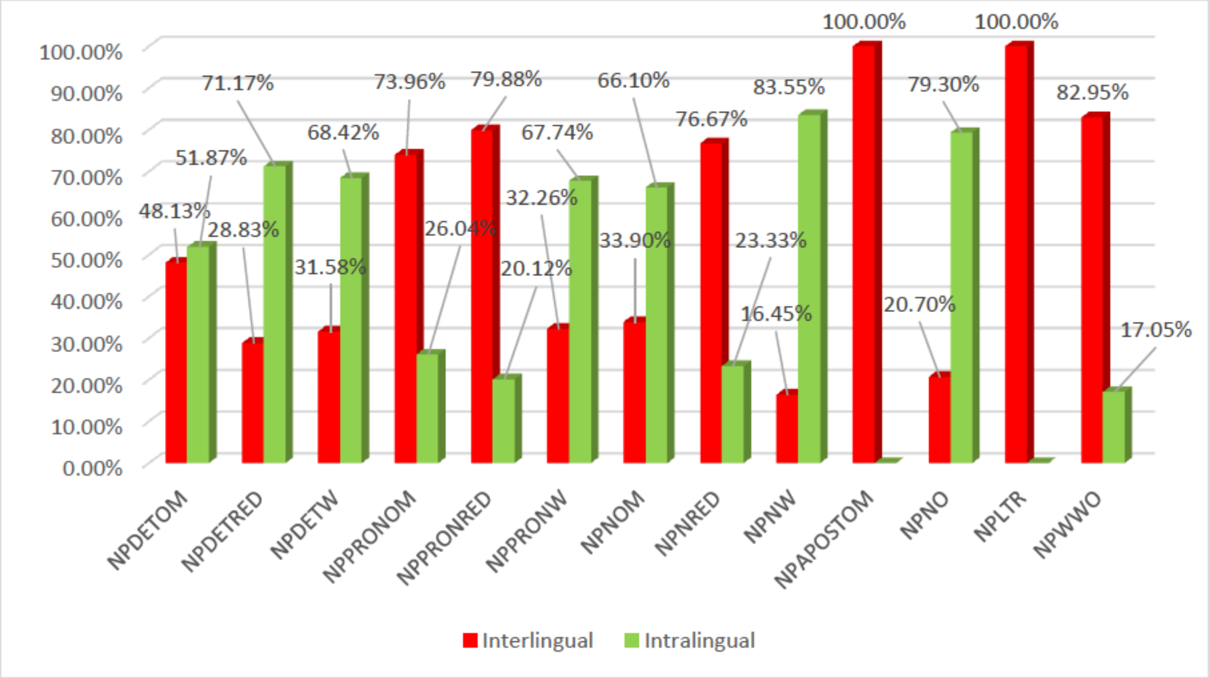


Figure 7.2: The interlingual vs intralingual noun phrase subcategory errors in the LEFLL corpus

Across the three LEFLL sub-corpora, Figures 7.3, 7.4 and 7.5 below show the distribution of interlingual noun phrase subcategory errors vs intralingual noun phrase subcategory errors in the Year 1, Year 2 and Year 3 sub-corpora respectively. A comparison between Figure 7.4 on the one hand and Figures 7.3 and 7.5 on the other shows that year 2 still constitutes an 'outlier' stage for most noun phrase subcategory errors. The percentages of interlingual noun phrase subcategory errors either increased in the Year 2 sub-corpus before decreasing in the Year 3 sub-corpus or decreased in the Year 2 sub-corpus before increasing in the Year 3 sub-corpus.

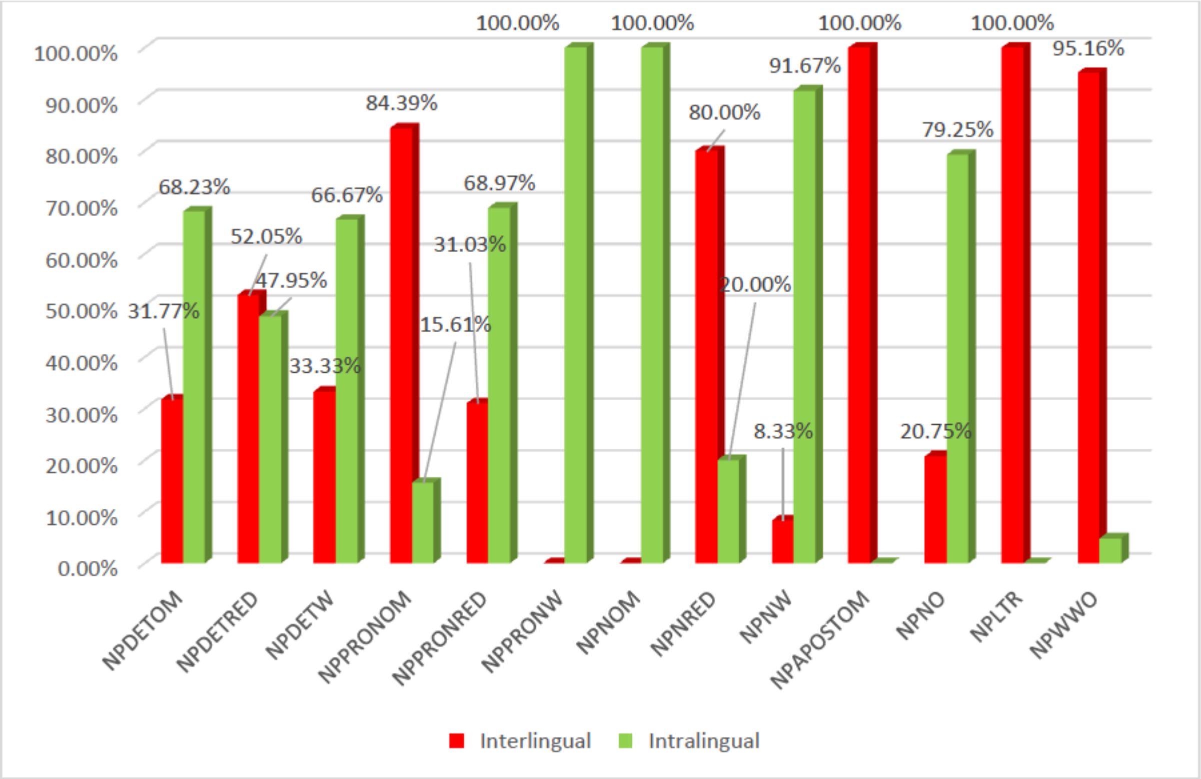


Figure 7.3: The interlingual vs intralingual noun phrase subcategory errors in the Year 1 sub-corpus

For instance, almost one third of determiner omission errors in the Year 1 sub-corpus (31.77%) were interlingual (nearly two thirds of noun phrase subcategory errors were intralingual). In the Year 2 sub-corpus, this proportion increased to two thirds before decreasing again to almost one third in the Year 3 sub-corpus (36.51%). On the other hand, the percentage of pronoun omission errors due to L1 transfer was 84.39% in the Year 1 sub-corpus. This declined in the Year 2 sub-corpus to 64.29% before increasing again to 76.06% in the Year 3 sub-corpus.

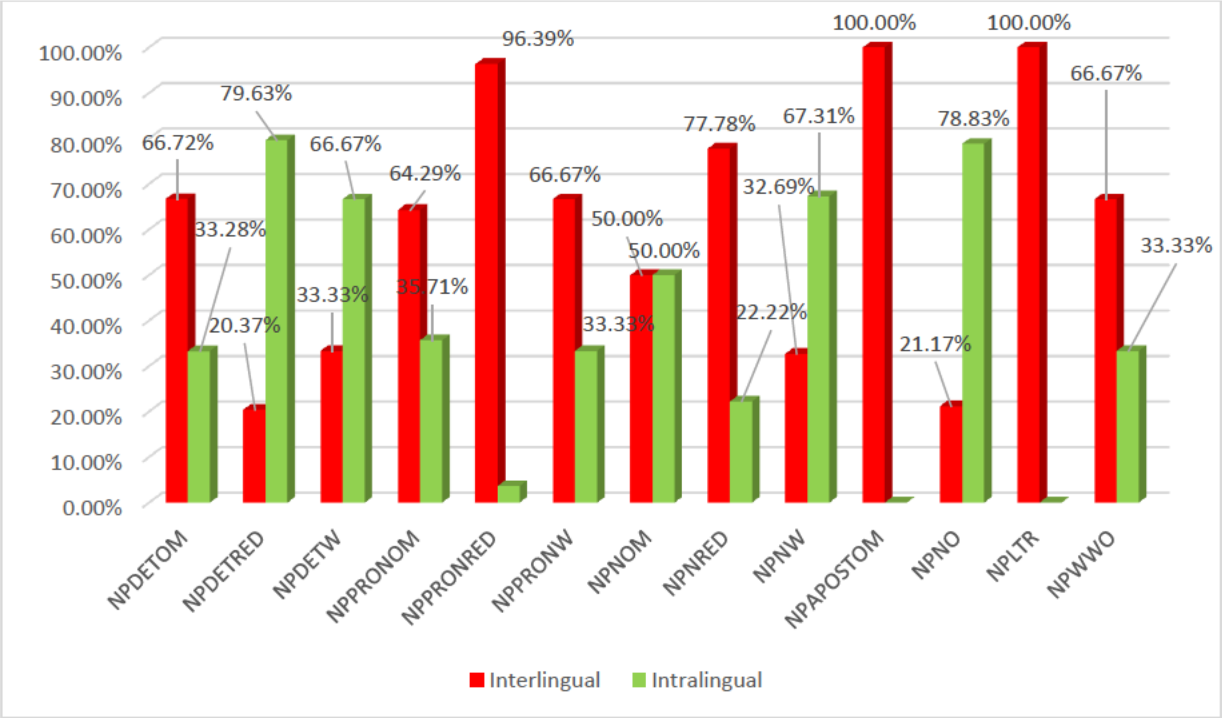


Figure 7.4: The interlingual vs intralingual noun phrase subcategory errors in the Year 2 sub-corpus

A comparison between Figure 7.3 and Figure 7.5 shows that there is a decrease in the percentages of interlingual errors vs an increase in the percentages of intralingual errors in most noun phrase subcategory errors among the learners in the year 3 sub-corpus. This may indicate that learners at a more advanced level start drawing more heavily on target language rules, and less on transferring first language knowledge to the production of second language target forms. The high percentages of intralingual noun phrase subcategory errors in the advanced levels would seem to indicate, however, that these learners still experience major problems in applying these target language rules.

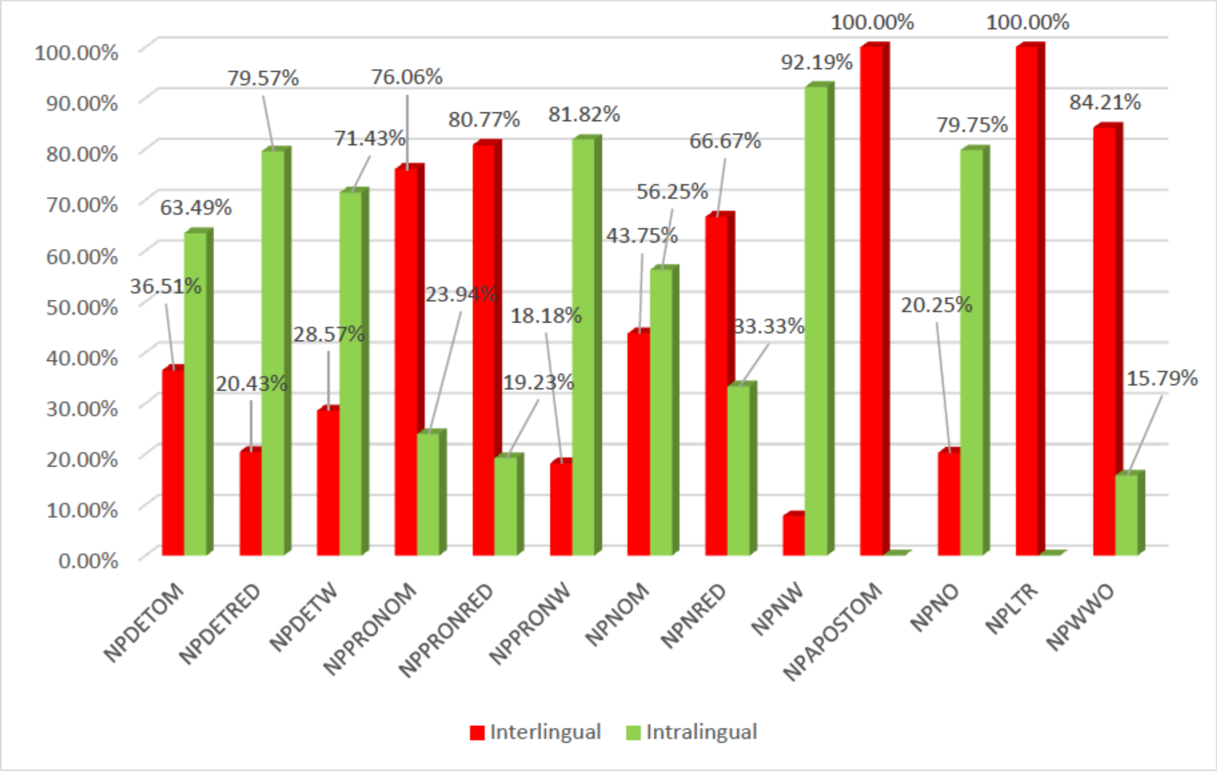


Figure 7.5: The interlingual vs intralingual noun phrase subcategory errors in the Year 3 sub-corpus

This in turn may relate to the observation of an increase in the percentages of interlingual errors of some noun phrase subcategory errors. This observation may indicate that some learners may fall back onto L1 rules on occasions when they are unable to access or apply target language rules. For instance, whereas the percentages of interlingual pronoun substitution and noun omission errors were 0% in the Year 1 sub-corpus, they increased to 18.18% and 43.75% respectively in the year 3 sub-corpus.

Error Type	Year 1 Sub-corpus				Year 2 Sub-corpus				Year 3 Sub-corpus				LEFLL Corpus			
	Inter.		Intra.		Inter.		Intra.		Inter.		Intra.		Inter		Intra.	
	Freq.	%	Freq.	%	Freq.	%	Freq.	%	Freq.	%	Freq.	%	Freq.	%	Freq.	%
NPDETOM	149	31.77	320	68.23	433	66.72	216	33.28	138	36.51	240	63.49	720	48.13	776	51.87
NPDETRED	38	52.05	35	47.95	22	20.37	86	79.63	19	20.43	74	79.57	79	28.83	195	71.17
NPDETW	7	33.33	14	66.67	5	33.33	10	66.67	6	28.57	15	71.43	18	31.58	39	68.42
NPPRONOM	173	84.39	32	15.61	162	64.29	90	35.71	108	76.06	34	23.94	443	73.96	156	26.04
NPPRONRED	9	31.03	20	68.97	80	96.39	3	3.61	42	80.77	10	19.23	131	79.88	33	20.12
NPRONW	0	0.00	8	100	8	66.67	4	33.33	2	18.18	9	81.82	10	32.26	21	67.74
NPNOM	0	0.00	17	100	13	50	13	50	7	43.75	9	56.25	20	33.90	39	66.10
NPNRED	12	80	3	20	7	77.78	2	22.22	4	66.67	2	33.33	23	76.67	7	23.33
NPNW	3	8.33	33	91.67	17	32.69	35	67.31	5	7.81	59	92.19	25	16.45	127	83.55
NPAPOSTOM	15	100	0	0.00	60	100	0	0.00	16	100	0	0.00	91	100	0	0.00
NPNO	22	20.75	84	79.25	29	21.17	108	78.83	32	20.25	126	79.75	83	20.70	318	79.30
NPLTR	1	100	0	0.00	7	100	0	0.00	4	100	0	0.00	12	100	0	0.00
NPWWO	59	95.16	3	4.84	32	66.67	16	33.33	16	84.21	3	15.79	107	82.95	22	17.05

Table 7.3: The percentages of interlingual vs intralingual noun phrase error categories in the LEFLL corpus

So far, this section has focussed on the distribution of interlingual noun phrase errors and subcategory errors in the LEFLL corpus and across the three LEFLL sub-corpora. The discussion above has attempted to understand the role of the learners' first language as the analysis moved from one level to another. As we did in the previous chapter, the results obtained in this section were used to interpret the unsteady changes in the percentages of noun phrase errors, across the three LEFLL sub-corpora, observed earlier in Figure 5.6 and Table 5.9, Section 5.3 when they were calculated based on the PFEC approach. The unsteady changes in the percentages of noun phrase errors observed in Figure 5.6 and Table 5.9 could be due to the unsteady changes in the percentages of interlingual noun phrase errors across the three LEFLL sub-corpora observed in this section. This may support the claim, as discussed in the previous chapter that the steady decline in the interlingual spelling errors observed in Chapter 6, Section 6.4 led to a steady decline in the percentages of spelling errors observed in Figure 5.6 and Table 5.9, Section 5.3. Section 7.4 will discuss and provide the reasons why the 13 noun phrase subcategory errors, in the current section, were classified as interlingual errors. The following section is dedicated to present and discuss the intralingual noun phrase subcategory errors.

7.3 Intralingual Noun Phrase Subcategory Errors

As noted above, just over half of all noun phrase errors were identified as being due to target language development factors, and thus as intralingual errors. These causal factors include such phenomena as incorrect application of target language rules and overgeneralization. Examples of these types of error from the LEFLL data will be provided in this section.

Whilst a rough 50/50 split was observed for interlingual and intralingual errors at a very general level, a more complex picture emerges when we look at the data for subcategories of these major types. These data are summarised in Table 7.3 above. Specifically, 13 out of 17 noun phrase subcategory errors were purely, predominately or partially interlingual, whereas only 4 out of 17 subcategory errors were purely intralingual, across the three LEFLL sub-corpora, as can be seen in Figure 7.6 and table 7.4 below.

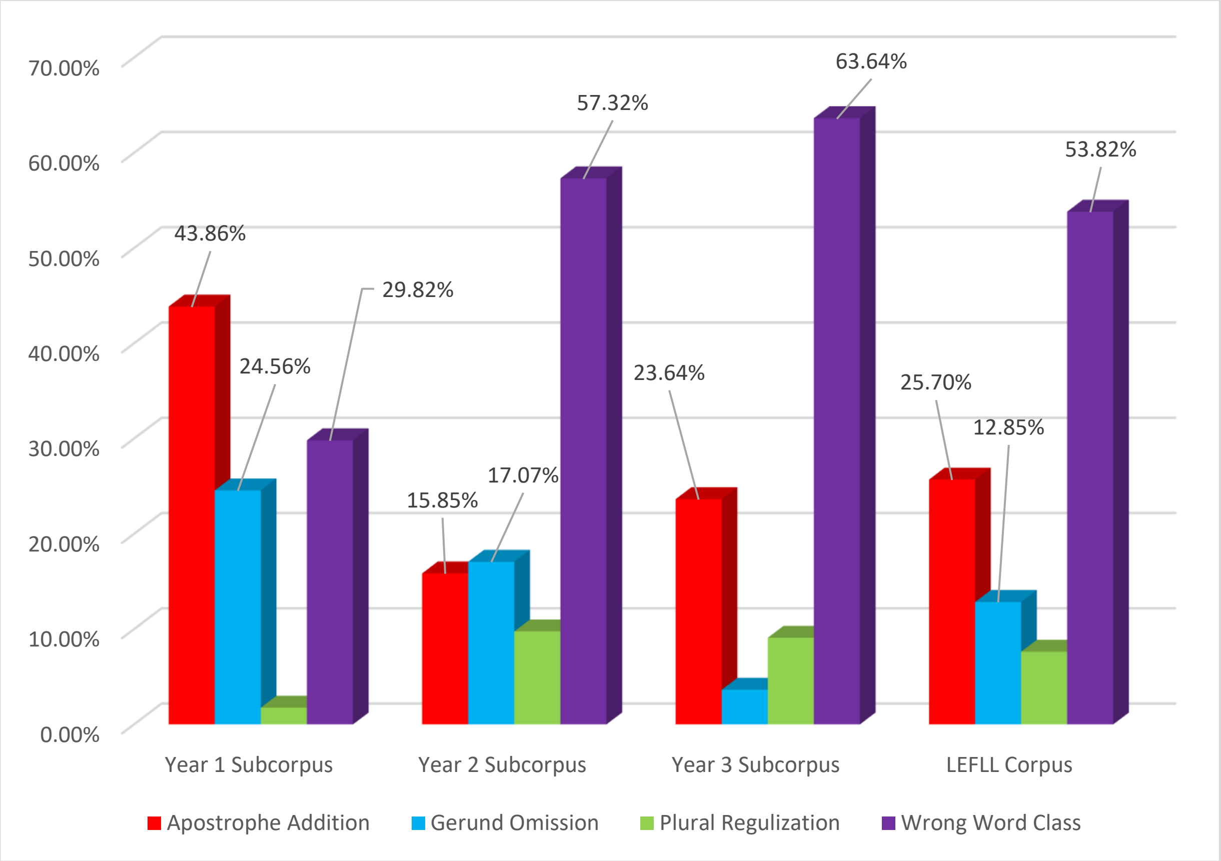


Figure 7.6: The intralingual noun phrase subcategory errors in the LEFLL corpus

Overall, noun phrase errors featuring wrong word class selections represented the highest proportion of intralingual noun phrase subcategory errors in the LEFLL corpus, at 53.82%. It seems that the learners represented in LEFLL have problems in distinguishing between different word classes, notably nouns, verbs and adjectives, when composing noun phrases. This can also be observed as we move across the three sub-corpora as the Figure and Table show. The percentages of intralingual noun phrase errors caused by wrong word class selections increases from one level to another, and is highest in the Year 3 sub-corpus. On the one hand, this steady increase may be at least partly attributable to students attempting to construct longer and more complex noun phrases as they progress in their English studies; the more (and the more complex) noun phrases they write, the greater the potential for error. Viewed from this perspective, the rising word class error rate trend could even be seen as a positive indicator overall, in that it indicates growing ambition on the part of the students. Nevertheless, these figures also strongly suggest that the students would benefit from additional teaching input on word class selection issues, and more explicit pedagogic attention to this area in general.

Intralingual Errors	Year 1 Sub-corpus	Year 2 Sub-corpus	Year 3 Sub-corpus	LEFLL Corpus
Apostrophe Addition	(n=25) 43.86%	(n=13) 15.85%	(n=26) 23.64%	(n=64) 25.70%
Gerund Omission	(n=14) 24.56%	(n=14) 17.07%	(n=4) 3.64%	(n=32) 12.85%
Plural Regularization	(n=1) 1.76%	(n=8) 9.76%	(n=10) 9.08%	(n=19) 7.63%
Wrong Word Class	(n=17) 29.82%	(n=47) 57.32%	(n=70) 63.64%	(n=134) 53.82%
Total (n)	57	82	110	249

Table 7.4: The percentages of intralingual noun phrase subcategory errors in the LEFLL corpus

A more obvious sign of positive development across the three levels can be observed in the percentages of gerund omission errors. As shown in Figure 7.6 and Table 7.4, these consistently decrease as we move across the year 1, year 2 and year 3 sub-corpora (24.56%, 17.07% and 3.64% respectively).

The following four subsections provides more details about the four types of intralingual noun phrase subcategory errors observed in the LEFLL corpus. These are supported by examples from the LEFLL corpus along with equivalent Arabic structures and the transliteration of these Arabic structures using the *Ijmes transliteration system for Arabic, Persian and Turkish* (see Appendix F).

7.3.1 Apostrophe Addition Error

In this type of intralingual noun phrase subcategory error, learners overgeneralise the rules for inserting an apostrophe marker (’). The analysis revealed that the learners overgeneralised two rules in particular. The first relates to possessive forms; the learners often unnecessarily insert an apostrophe marker to regular plural nouns where a plural s has been added, as in example (1) below:

1) I am <<twenty year's old>> (NPAPOSTRED File 738)

أنا (عمري عشرون عاما)

Ana (aᶜamrī ᶜshrūn ᶜāman

The second rule overgeneralisation occurs in contractions where one or more letters are removed from the original word and the apostrophe marker is added to indicate the missing letter(s). In example (2) below, the learner has inserted the apostrophe marker (’) between the letters (n) and (t) as a result of overgeneralising the rule for shortening the negative word ‘*not’* (*n’t* for *not*):

2) I’m <<studn't>> … (NPAPOSTRED File 774)

أنا (طالب)

Ana (ṭalib)

7.3.2 Gerund Omission Error

In this type of noun phrase subcategory error, learners exhibit uncertainty about when to use the gerund form ‘*ing*’. The following example is typical:

3) Finally <<Learn english>> is necessary For every bady. (NPGER File 1)

في النهاية (تعلم الأنجليزية) مهم لكل شخص

Fii al-nihāya (taᶜalum al-injilyzyah) muhim likul shakhṣ

7.3.3 Plural Regularization Error

In this type of noun phrase subcategory error, learners overgeneralise the grammatical rule for regular plural forms by adding the regular plural marker '*s*' to irregular plural forms, as in the following example:

4) our friends are an important part in our <<lifes>> (NPNOREG File 10)

أصدقائنا جزء مهم في (حياتنا)

ᵓaṣdiqaᵓna juzᵓa muhim fii (ḥayatuna)

7.3.4 Wrong Word Class Error

This noun phrase subcategory error type has already been briefly discussed above. The main additional point to note here is that, in the LEFLL data, this error typically takes the form of students incorrectly using a verbal, adjectival or adverbial form of a word where they should have used a noun form. For instance, in the following example, the learner used the verb word class '*conclude'* instead of *'conclusion'*:

5) In <<conclude>> , The room is the best place (NPWWC File 511)

في (الأستنتاج)، الغرفة هي أفضل مكان

Fii (al-istintāj), al-ghurfah hya ᵓafḍal makān

7.4 Discussion

As a reminder and as was pointed out in the Introduction of this thesis, this thesis aims to provide a broader picture of the role of L1 in the writing of Arab English learners. More specifically, the thesis studies spelling, noun phrase and verb phrase errors produced by Arab

English learners in their final exam English writing to determine errors that were produced due to L1 transfer and how they develop across the university academic levels. Thus, the thesis seeks to answer the following two main questions:

1) To what extent does L1 influence affect spelling, noun phrase and verb phrase errors in the writing of Arab English learners?
2) Does this influence follow the same pattern (either an increase or decrease) as the learners proceed across the university academic levels?

As was observed in the previous chapter, L1 has great influence on spelling errors in the writing of Arab English learners where almost two third of spelling errors in the LEFLL corpus were interlingual. The role of L1 in noun phrase errors, on the other hand, is less significant. As the analysis of noun phrase errors and subcategory errors above showed, there is a roughly 50/50 split in the current data between interlingual and intralingual noun phrase errors in the LEFLL corpus. Across the three sub-corpora (Year 1 sub-corpus, Year 2 sub-corpus and Year 3 sub-corpus), the analysis, in the previous chapter, showed that the percentages of interlingual spelling errors in the three sub-corpora were higher than the percentages of intralingual spelling errors but there was a steady decline in the L1 influence on spelling errors as the learners move from one level to another. In the analysis of noun phrase errors and subcategory errors, on the other hand, it can be seen that year 2 represents an anomalous stage between years 1 and 3. There is no steady change (either an increase or decrease) in the percentages of interlingual noun phrase errors across the three Sub-corpora. Whereas the percentage of interlingual noun phrase errors in year 2 was higher than the intralingual type in the same year (56.82% vs 43.18% respectively), the percentages of interlingual noun phrase errors in years 1 and 3 were higher than the percentages of intralingual types. A comparison between the Year

1 and Year 3 sub-corpora revealed that the percentages of interlingual noun phrase errors decreased in year 3 versus an increase in the percentages of intralingual noun phrase errors.

Noun phrase errors were then classified into 17 noun phrase subcategory errors. The analysis revealed that 13 out of the 17 noun phrase subcategory errors were either purely, predominately or partially interlingual error types. The following subsections will illustrate why these 13 noun phrase subcategory errors were classified as interlingual types. This will include examples for each type of the interlingual noun phrase subcategory error extracted from the LEFLL corpus along with equivalent Arabic structures and the transliteration of these Arabic structures using *Ijmes transliteration system for Arabic, Persian and Turkish* (see Appendix F).

7.4.1 Determiner Omission Error

Generally, the Error Analysis revealed that the learners encountered problems with the different types of determiners (definite/indefinite articles '*a, an* & *the'*, possessive determiners '*my, her, their*, etc, quantifiers '*every, many, much',* and determiners of differences '*other, another'*).

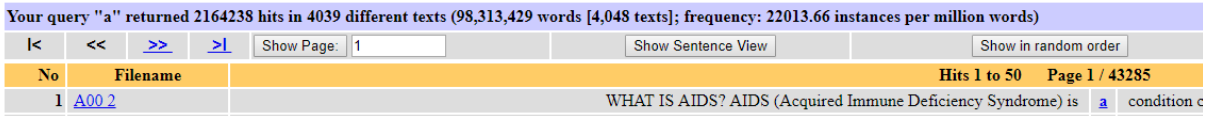
The analysis revealed that almost one third of determiner omission errors in the year 1 and year 3 sub-corpora and two thirds of determiner omission errors in the year 2 sub-corpus were due to first language interference. Most obviously, learners tend to omit determiners from noun phrase structures when the equivalent Arabic structures do not require determiners. Among the different types of determiner omission errors, the omission of the indefinite articles '*a, an'* represented the highest proportion of determiner omission errors out of all types of determiner omission errors, and a large percentage of the indefinite article omission errors were due to L1 interference. The reason for this type of error is that Arabic has no indefinite articles. Instead, it only has the definite article 'ال' '*al*', which is equivalent to the English definite article '*the*'.

Indefinite status in Arabic is marked by the absence of the Arabic definite article 'ال' '*al*' and it is often used for common nouns, e.g., 'طبيب' *'ṭabyb'* (*doctor*), 'طالب' *'ṭalib'* (student), etc. For these reasons, indefinite articles omission errors observed in the LEFLL corpus can plausibly be marked as interlingual. This coding policy is in line with numerous previous research studies which also reported the omission of determiners among Arab speaker English learners as a result of L1 transfer (e.g., Scott & Tucker, 1974; AbiSamra, 2003; Bataineh, 2005; Hourani, 2008; Crompton, 2011; Alamin & Ahmed, 2012; Alhaysony, 2012; Barry, 2014; Younes & Albalawi, 2015; Al-Shujairi & Tan, 2017; AMARA, 2018; Hamed, 2018).

The analysis also revealed that the omission of the indefinite article '*a*' is far higher than the omission of the indefinite article '*an*'. This finding echoes Alhaysony (2012), who also observed that the omission of the indefinite article '*a*' exceeds the omission of the indefinite article '*an*' in her analysis of English essays written by Saudi Arabian students . Alhaysony found 75 incidents of omission errors of '*a*', versus only 6 incidents of omission errors of '*an*' in her 100-essay corpus. Unfortunately. Alhaysony failed to provide an explanation as to *why* the students in her study omitted the indefinite article '*a*' more than they did with the indefinite article '*an*'. Instead, she focused on the overuse of the indefinite article '*a*' (29 incidents of additional error), observed in her study, which exceeded the overuse of the indefinite article '*an*' (7 incidents addition error). She attributed the latter variation to the complexity of '*an*', arguing that with this word there is a '… need to learn additional rules beyond those that apply to the article '*a*'' (Alhaysony, 2012:60). While this seems entirely reasonable in itself, it does not explain why Arabic L1 learners of English tend to omit 'a' more than 'an'.

Both '*a*' and '*an*' have the same linguistic function. They are used to introduce singular nouns that are previously unknown or mentioned for the first time. As Alhaysony (2012) points out, learning how to use the indefinite article ‘*an*’ requires ‘additional rules’ which are not required in learning how to use the indefinite article ‘*a*’. To put it simply, ‘a’ is used before singular nouns that start with consonants, while ‘*an*’ is used before singular nouns that start with vowels. The fact that there are far fewer vowels than consonants in English means that the chance or necessity of using the indefinite article ‘*a*’ is far higher than using the indefinite article ‘*an*’. In other words, it is possible to argue that ‘a’ has a greater potential for error than does ‘an’.

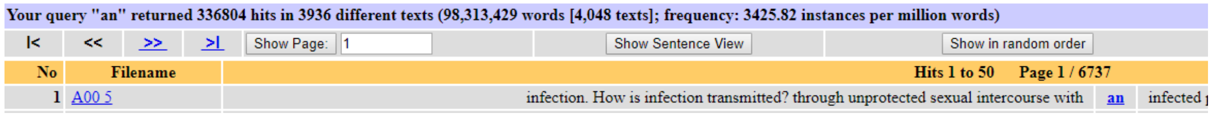
By consulting the BNC (British National Corpus), a quick search shows that the frequency of the indefinite article ‘*a*’ is 22013.66 instances per 1 million words, as shown in Figure 7.7 below, while the frequency of the indefinite article ‘*an*’ is only 3425.82 per 1 million words, as shown in Figure 7.8 below.



Your query "a" returned 2164238 hits in 4039 different texts (98,313,429 words [4,048 texts]; frequency: 22013.66 instances per million words)

No	Filename	Hits 1 to 50 Page 1 / 43285		
1	A00 2	WHAT IS AIDS? AIDS (Acquired Immune Deficiency Syndrome) is	a	condition c

Figure 7.7: The frequency of ‘*a*’ per 1 million words in the BNC



Your query "an" returned 336804 hits in 3936 different texts (98,313,429 words [4,048 texts]; frequency: 3425.82 instances per million words)

No	Filename	Hits 1 to 50 Page 1 / 6737		
1	A00 5	infection. How is infection transmitted? through unprotected sexual intercourse with	an	infected

Figure 7.8: The frequency of ‘*an*’ per 1 million words in the BNC

If we assume that all the indefinite articles ‘*an*’ in the BNC (3425.82 instances per 1 million words) were erroneously removed and only a quarter of the indefinite articles ‘*a*’ in the BNC (22013.66 instances per 1 million words) were erroneously removed from the BNC, there will be 5503.42 instances of the indefinite article ‘*a*’ omission errors per 1 million words vs 3425.82 instances of the indefinite article ‘*an*’ omission errors per 1 million words in the BNC. Thus, the variation between the frequencies of the indefinite articles' omission errors for both '*a*' and '*an*' may not necessarily reflect an indication of the difficulty the learners may encounter when they learn how to use either '*a*' or '*an*'. It is merely attributed to the chance of committing an error. The indefinite article '*a*' has a higher potential for error than does the indefinite article '*an*'.

This observation is a by-product finding of this chapter, but it is worth noting that it reinforces the more general arguments in favour of adopting the PFEC approach put forward by this thesis. As the discussion above shows, the PFEC approach enables us to better understand the probable reason why the language learners in the current data produced higher percentages of errors of a specific linguistic constituent (in this case the indefinite article) compared to other constituents.

If the above arguments are accepted, it then follows from that the high percentage of indefinite article omission errors over other types of determiners omission errors observed in the LEFLL corpus and sub-corpora cannot be considered as a measurement of the difficulty the learners may experience when they use the indefinite articles '*a*' and '*an*'. To verify which linguistic constituent of a noun phrase (e.g., definite/indefinite articles, possessive pronouns, subject pronouns, etc.) may constitute real problems to the learners over other types of linguistic constituents, we would need to apply the PFEC approach fully, by calculating the percentage of errors of each linguistic constituent. However, as this would be well beyond the scope of the current thesis, the following two examples (6) and (7) will have to suffice as illustrative examples of cases of omission of '*a*' and '*an*' due to L1 transfer:

The omission of (*a*)

6) After that, I have <<>> class later (NPDETOM File 710)

بعد ذلك، أنا عندي_درس مؤخرا

Ba'ada dhālika, ana 'andi_daress muākhran

The omission of (*an*)

7) Because she play <<>> important role for making our next generation (NPDETOM File 35)

لانها تلعب_دورا مهما في صناعة الجيل القادم

Le-anaha tal'ab dawran muhiman fii ṣna'at al-jeel al-qadim

As mentioned briefly above, the current analysis also found the omission of the definite article *'the'* to be another major type of determiner omission error in the LEFLL data. A review of Arabic grammar showed that the definite article '*the*' ('ال' '*al'* in Arabic) is very frequently used in Arabic. It can be used in Arabic structures where it is not required in the equivalent English structures, e.g., before the names of some countries or cities that consist of a single word, such as: 'العراق' *'<u>al</u>-ʿaraq'* (<u>*the*</u> *Iraq*), *'*القاهرة' *'<u>al</u>-Qāhirah'* (<u>*the*</u> *Cairo*), *etc.*, before proper nouns, as in 'هو يدرس الفيزياء' '*huwa yadrusū <u>al</u>-fīzya'* (*he studies the physics*), and – perhaps in most striking contrast to English – in the genitive case, such as: 'البدلة الجديدة''*<u>al</u>-badlah al-jadeedah'* (*the new <u>the</u> suit*), . This much more frequent usage has led previous studies of Arabic L1 learners of English (e.g., AbiSamra, 2003; Alhaysony, 2012) to attribute the omission of the English definite article *'the'* to language developmental rather than language transfer factors. Certainly, there are arguments in favour of this interpretation. The high frequency of the Arabic definite article 'ال' *'al'* in Arabic may indeed indicate that all English sentence structures that require the English definite article '*the'* do have equivalents in Arabic where the definite article *'*ال' *'al'* is necessary. Furthermore, there are some more Arabic structures that require the Arabic definite article 'ال' *'al'* which are not necessary in the equivalent English structures.

However, this is not always the case. In contrast to AbiSamra (2003) and Alhaysony (2012), the current thesis finds that the omission of the definite article *'the'* error in English structures may occur due to the absence of the Arabic definite article 'ال' *'al'* from the equivalent Arabic structures. Consider the examples below:

8) <<colore>> my room is grey and white (NPDETOM File 317)

لون غرفتي هو الرمادي والأبيض

Lawn ghurfatī hūwa al-ramady wa al-abyaḍ

Example (8) above shows one type of a genitive construction where the definite article '*ال*' 'al' (*the*) was removed. This is due to the presence of the possessive determiner '*ي*' '*ī*' (*my*) attached as a suffix to the noun '*غرفة*' '*ghurfa*' (*room*). Therefore, the noun '*غرفة*' '*ghurfa*' (*room*) is already defined and does not require the definite article '*ال*' '*al*' (*the*). In the following example:

9) I have <<biggest Televesion>>. (NPDETOM File 417)

أنا عندي أكبر تلفاز

Ana ʿandy akbar tilfaz

the definite article '*ال*' '*al*' (*the*) was removed. The definite article has been omitted here because the modified noun '*تلفاز*' '*tilfaz*' (*television*) was provided following the superlative form of the adjective '*أكبر*' '*akbar*' (biggest), which is not necessary in the equivalent Arabic structure. The word order of the noun phrase does not, also, require a definite article. The definite article '*ال*' '*al*' may be necessary in other structures of a similar meaning to the example above, e.g., in the absence of the noun '*تلفاز*' '*tilfaz*' (*television*), this will become (*أنا عندي الأكبر*) '*Ana ʿandy al-akbar*' (*I have the biggest*) or when the modified noun '*تلفاز*' '*tilfaz*' (*television*) precedes the superlative. This will be (أنا عندي التلفاز الأكبر) '*Ana ʿandy altilfaz al-akbar*' (*I have the biggest the television*). The omission of the definite article from English structures such as these was also observed by Diab (1996) among Lebanese students studying at the American University of Beirut. Diab attributed the omission of the definite article to the role of L1 Arabic interference, and it is this interpretation that will be preferred here.

7.4.2 Determiner Addition Error

As discussed in Section 7.4.1 above, just as the omission of determiner errors due to language interference often occurred with the English indefinite articles '*a*' & '*an',* errors of determiner addition often occurred with the English definite article '*the'.* A previous study by Al-Shujairi & Tan (2017) found that most of the errors produced by Iraqi postgraduates and undergraduates occurred in articles were of the addition error type, and that the addition of the definite article constituted 23% out of all addition errors in their data. Although the high percentage of errors of specific linguistic constituents may not necessarily reflect the strength or weakness of the language learners in those specific linguistic constituents as discussed above in Section 7.4.1, providing the percentages of errors of different linguistic constituents is nevertheless useful insofar as it provides a general picture of the extent to which a piece of writing produced by a language learner is distorted.

Since the definite article *'the'* is more commonly and more widely used in Arabic than it is in English, as discussed above, it is reasonable to assume that the addition of unnecessary definite article errors could be due to first language interference. This is certainly the view taken by Al-Shujairi & Tan (2017) , and by other studies of Arabic L1 learners of English as well (e.g., Crompton 2011; Hamed, 2018;). Consider example (10) for the unnecessary use of the definite article *'the'*:

10) … I go to <<the>> sleep (NPDETRED File 351)

أنا أذهب إلى النوم

Ana adhhib ilaa al-nawm

The current analysis also revealed that other types of determiners were unnecessarily added, again probably due to the same language transfer consideration. To illustrate, the following example shows the unnecessary addition of the possessive determiner '*my'* which is equivalent to the Arabic possessive pronoun 'ي' *'ya'* attached as a suffix.

11) … my brothers and <<my>> sisters. (NPDETRED File 792)

أخوتي وأخواتي

Akhwaty wa akhawāty

7.4.3 Incorrect Use of Determiner Error

The incorrect use of determiners observed in the LEFLL corpus could be due to various reasons. Of these, one of the most likely causes could be literal translation from Arabic to English of the function of the incorrect determiner in an equivalent Arabic structure. For instance, the Arabic determiner 'كثير' *'katheer'* can be used with both countable and uncountable nouns in Arabic. Therefore, the literal translation of 'كثير' from Arabic to English is both '*many'* and *'much'.* This may be the reason why the learners in LEFLL often incorrectly use these two determiners; they simply perceive them as the same. This is in line with Zughoul (2002), who found that Arab speaker English learners at the University of Texas Austin, consistently confused '*much'* with '*many'*. Zughoul also attributed this type of confusion to the fact that 'much' and 'many' have one equivalent in Arabic ('كثير' *'katheer'*). The following two examples from the LEFLL corpus are typical instances of this confusion:

12) when you want to make stistics for <<many money>> does the smoker pay … (NPDETW File 28)

عندما تريد أن تعمل إحصاء عن كمية المال الذي يصرفه المدخن

'andama turyd ann ta'amal iḥṣā 'an kamyat al-mal al-ladhy yaṣrifahu al-mudakhin

and

13) I want to talk about how <<much>> sister with me in my room. (NPDETW File 330)

أنا أريد أن أتحدث عن كم أخت معي في الغرفة

Ana auryd ann ataḥdath 'ann kam aukhit ma'ay fii al-ghurfa

Grammatically, all nouns (whether they refer to animate beings or inanimate objects), in Arabic, are classified either as masculine or feminine. In Arabic, both singular feminine inanimate objects and plural inanimate objects are treated as singular feminine beings. Therefore, the learners sometimes use the feminine singular possessive determiner to refer to an inanimate object, as in the following example:

14) My room is very simple the <<her>> color is pink (NPDETW File 431)

غرفتي صغيرة جدا لونها وردي

Ghurfaty ṣaghyra jadan lawnuha wardy

7.4.4 Pronoun Omission Error

There are two types of sentences in Arabic: *nominal* and *verbal sentences*. Nominal sentences start with a noun or pronoun while verbal sentences start with a verb. The subject of the verb in the verbal sentence is embedded in the verb, and the verb form itself may change depending on its grammatical subject. In the following example from the LEFLL corpus, the learner may have omitted the pronoun *"I"* due to language transfer, since the sentence is equivalent to a verbal sentence in Arabic:

15) Finally <<>> Go to my house to rest form (for *from*) … (NPPRONOM File 880)

في النهاية ( ) أذهب إلى منزلي لارتاح من ...

Fii al-nihayah ( ) adhib ila manzily liartaḥ men …

This is in line with Ridha (2012) who also observed the omission of the subject pronouns from verbal sentences in a study of English writing produced by Iraqi learners, and who also attributed this error type to first language interference, specifically to the above distinction between verbal and nominal sentences in Arabic.

7.4.5 Pronoun Addition Error

Whereas the omission of pronouns could be attributed to negative language transfer from Arabic to English due to the omission of pronouns from verbal sentences in Arabic, the addition/overuse of pronouns could also be attributed to the negative language transfer from Arabic to English. Errors of pronoun addition or overuse were observed in both subject and object pronouns. Example 16 below shows an overuse of the object pronoun "*it*" which refers to "anything" mentioned in the same example:

16) … and take anything we need <<it>> by the job … (NPPRONRED File 4)

ونأخذ كل شي نحن نحتاجهِ عن طريق العمل

Wa naʾkhudh kul shy nḥnu naḥtajuh ( ) aʿn ṭaryq ala‘amal

In English grammar, the relative pronoun replaces the noun/pronoun that the relative pronoun defines. In Arabic grammar, in contrast, repeating the defined noun/pronoun is obligatory, and is often used for rhetorical emphasis (Wickens, 1980). Therefore, this type of error may be interpreted as a case of negative language transfer. This finding is in line with Alotaibi (2016), who also observed the incorrect addition of pronouns in relative clauses by Kuwaiti EFL learners.

17) Sunday <<it>> is my busy day. (NPPRONRED File 714)

يوم الأحد (هو) يومي المشغول

Yawm al-aʾḥid (huwa) yawmy al-mashghwl

As shown in example (17) above, the pronoun "*it*" "هو" (*huwa*) has been unnecessarily added to the sentence referring to "*Sunday*". This finding agrees with Swan & Smith (2001) and Al-Hazzani & Altalhab (2018) who similarly observed the unnecessary use of pronouns by Arab English learners. Both Swan & Smith and Al-Hazzani & Altalhab attributed this type of error to Arabic interference in English learning.

7.4.6 Incorrect Use of Pronoun Error

Although Arabic subject pronouns agree with English subject pronouns in terms of function (they both replace nouns), Arabic pronouns tend to be more specific than their English equivalents. For instance, in English, the 2nd person subject pronoun *"you"* is addressed to the receiver no matter whether the receiver is singular or plural, masculine or feminine. Arabic, on the other hand, has five second person subject pronouns addressed to the receiver (أنتَ *"anta"*, for singular masculine, أنتِ *"anti"*, for singular feminine, أنتُما *"antuma"*, for dual masculine or feminine, أنتُم *"antum"*, for plural masculine and أنتُن *"antun"*, for plural feminine). Due to this variation between subject pronoun groups in both Arabic and English, Arab English learners may replace the correct pronouns with incorrect ones due to this distinction between Arabic and English subject pronouns. Consider example (18) below:

18) … there's adifficult things <<it>> will facing you. (NPPRONW File 97)

هناك أشياء صعبة هي سوف تواجهك

Hunak ᵓashya ṣaᶜabah hya sawfa tuwajihuk

The pronoun *'it'* in the above example is incorrectly used to refer to the noun *'things'*. A possible explanation for this misuse of *'it'* instead of *'they'* is that in Arabic, plural inanimate objects are treated as singular feminine human beings when a pronoun is called to replace plural inanimate objects (see Section 7.4.3 above). In this case, the subject pronoun *'she'* is often called to replace a plural inanimate object noun as a result of negative language transfer (see Swan & Smith, 2001).

The analysis also revealed that the learners used the pronoun *'it'* to replace plural inanimate objects. A possible explanation for this might be that the learner is still operating under the Arabic L2-influenced assumption that plural inanimate objects should be treated as singular feminine, but is aware that the pronoun *'it'* is reserved for singular feminine. This may therefore cause the learner to resort to *'it'* as an alternative possible option for plural inanimate objects, as in the following example:

19) I have a big window <<he>> opened in the sea (NPPRONW File 418)

أنا عندي شباك كبير (هو) يفتح على البحر

Ana a^{c}andy shubak kabīr (huwa) yaftḥ a^{c}ala al-baḥr

Unlike in English, the names of inanimate objects in Arabic are classified into masculine and feminine and they need to agree with the pronouns that are required to replace the nouns. One of the most common features that may distinguish feminine object from masculine object in Arabic is the last alphabet letter in singular nouns. Singular feminine objects ends with the *"tā marbūṭa"* (ة). The word "شباك" *"shubak"* (window) in example (19) above does not end with the "*tā marbūṭa"* (ة), therefore it is classified as masculine. For the third person singular masculine objects/human beings, the pronoun (هو) *"howa"* is required to replace the nouns. Thus, this type of error is likely to be due to negative language transfer.

7.4.7 Noun Omission Error

In some noun phrases, in Arabic, which contain the superlative form of adjectives, omission of the head noun is possible when the omitted noun has been previously mentioned or can be inferred from the context. The superlative form of an adjective in this type of phrase is defined by the definite article ‘ال’ ‘*al*’ (*the*), therefore the head noun should be removed (see Section 7.4.1. for more discussion). Clearly, this is markedly different from English, where no such omission is permitted. Given this contrast, it seems reasonable to code instances of noun omission in the LEFLL data as interlingual errors. To illustrate, consider the following example:

20) But, <<the most serious >> is car accidents (NPNOM File 15)

لكن أخطرها هي حوادث السيارة

Lakin akhṭaraha hya ḥawadith al-sayarah

This interpretation is endorsed by Alasfour (2018), who also observed the omission of head nouns in the writing of Arab student English learners at Portland State University (PSU) in the U.S.A and concluded that L1 transfer was the most likely cause of this error type.

7.4.8 Noun Addition Error

While the omission of head nouns from noun phrases may mark a language transfer from Arabic to English as was discussed in Section 7.4.7 above, the overuse of nouns, as shown in example (21) below, may also be interpreted as a form of language transfer error. This was also the conclusion reached by Mohamed-Sayidina (2010), who also observed the repetition of the same noun by Arab ESL students.

21) … the food should be light <<food>> and good for heatly (NPNRED File 328)

... الطعام يجب أن يكون (طعام) خفيف وجيد للصحة

Al-ṭa’am yajib an yakun (ṭa’am) khafyf wa jayd lilsaha

7.4.9 Wrong choice of Noun Error

This type of error may be best described as a semantic error caused by direct translation from Arabic to English. For instance, as example (22) below shows, when describing the furniture in his/her bedroom, the student used the word *'library'* to refer to the '*bookcase'* that he/she has in his/her bedroom:

> 22) on the right, you will find the <<library>> where I put my books in (NPNW File 405)
>
> على اليمين، سوف تجد (المكتبة) حيث أضع فيها الكتب
>
> *A'ala al-yamyn, saufa tajid (al-maktabah) ḥythu aḍa'a fyha*

For the following words: *library, bookstore/bookshop and bookcase,* there is only one equivalent Arabic word called '*مكتبة*' *'maktaba'*. Therefore, replacing the word '*bookcase'* with *'library'* in the example above could be due to L1 interference. Of course, it is also possible to interpret this error in more intralingual terms, seeing it as a simple matter of limited English vocabulary on this student's part: the learner may have used the word '*library*' here simply because they might not yet have learned the word '*bookcase'*. It could also be due to the higher frequency of the word *'library'* in English In the BNC web, the frequencies of *library, bookstore, bookshop* and *bookcase*, per million words, are 82.15, 0.28, 4.99 and 1.47 respectively. Thus, the student may simply have had more exposure to the word 'library' and has thus used it by default here. However, errors such as this are not limited to individual cases; one of the benefits of adopting a corpus-based approach to error analysis is that it shows the analyst which error types are prevalent across many speakers or writers, and that is the case here: wrong noun choices that reflect Arabic/English lexical contrasts are frequent and pervasive in the data, and as such it seems more realistic to interpret these errors as being due to L1 interference, and thus to code them consistently as interlingual.

7.4.10 Apostrophe Omission Error

Arabic has no apostrophe marker; therefore, the omission of apostrophes by Arabic L1 learners of English is likely to be due to negative language transfer. This type of error predominately occurred in the possessive cases, as examples (23) and (24) below show. The analysis showed that there are two types of apostrophe omission errors among noun phrase subcategory errors in the LEFLL corpus. The first type is the omission of both (') and (*s*) apostrophe markers. The second type is the omission of (') apostrophe marker as shown in examples (23) and (24) below respectively:

The omission of both (‘) and (s) apostrophe markers:

23) After that I drive <<my son car>> (NPAPOSTOM File 718)
بعد ذلك أنا أقود (سيارة أبني)
baᶜada dhālika ana ᵓaqwd (sayārit ᵓabnī)

The omission of the apostrophe marker only

24) … or gift in <<valantines day>> (NPAPOSTOM File 94)
أو هدية في (يوم عيد الحب)
ᵓaw hadyah fii (yawm ᶜayd al-ḥub)

The possessive case in Arabic (which is known as "*إضافة*" "*Idaafa*" (genitive) in Arabic) is often formed by adding the Arabic definite article "*ال*" "al" (the). e.g., *كتاب الطالب* '*kitāb altalib',* literally '*the book the student*'. Therefore, the apostrophe omission errors committed by the Arab English learners may be as a result of the first language interference. This finding agrees with Al-Shujairi & Tan (2017), who also observed the omission of apostrophe of the possessive case in the writing of Iraqi pre-university students, and who attributed this type of error to L1 influence.

7.4.11 Number Agreement Error

Number agreement in Arabic is very different from that of English. In Arabic, quantity or number is only grammatically marked as plural until the number 11, after which the noun in question is treated as singular. Thus, Arabic has 'كتب 10' '*10 kutub*' (*10 books*) but 'كتاب 11' '*11 kitāb*' (11 book), 'طالب 23' '*23 ṭālib*' '23 student', and so on. Given this difference, it is highly plausible to attribute number agreement errors such as that illustrated in example (25) below to negative language transfer. This coding policy agrees with that of a previous study by Chaaraoui (2017), who observed Arabic L1 foundation students learning English at King Abdul Aziz University in Saudi Arabia producing errors where the nouns that follow number 11 and above are in the singular form due to L1 interference.

25) I am <<19 year>> old. (NPNO File 921)

أنا عمري 19 سنة

Ana 'aumry 19 sanah

The next example, (26), also raises an issue with number agreement between the demonstrative adjective *"this"* and the modified noun *"advantages".* This type of error has a very similar profile to the error in Section 7.4.6 (the incorrect use of pronouns) in the sense that nouns that refer to non-human beings are treated as singular and feminine. In this example, since the noun "*المزايا*" *"al-majāyā"* (*advantages*) is plural and does not mark the name of a human being, the demonstrative adjective required (in an equivalent Arabic structure) is "*هذه*" "*hadhhi"* (*this*). Therefore, this thesis regards this type of error as likely to be due to language interference.

26) Through <<this advantages>> … (NPNO File 13)

من خلال هذه المزايا

Men khilal hadhhi al-majāyā

This error type (*number agreement*) also includes a problem with uncountable nouns. The LEFLL analysis revealed that some uncountable nouns that were incorrectly written in plural forms in English would be correct if translated directly into Arabic. Consider the following example:

27) second in getting <<informations>> and even learning through it, (NPNO File 95)

ثانيا في الحصول على المعلومات وحتى التعلم من خلالها

Thānyān fii alḥuṣūl 'alá alm'alūmāt wa ḥattá alta'alum men khilālihā

This finding agrees with Chaaraoui (2017), who observed Saudi English learners pluralizing uncountable English nouns which are countable nouns in Arabic. As the author puts it, Arab English learners '… tend to pluralize uncountable nouns as it suits their L1 system' (p95). Further support for this interpretation comes from Sabbah (2015), who reported that many uncountable nouns in English are countable in Arabic, and that Arab English learners tend to pluralize them.

7.4.12 Literal Transfer Error

This type of error may be considered to be the most prominent language interference error type of all. In this error type, learners clearly transliterate the Arabic words in English alphabet. This error type was therefore consistently marked as an L1 interference error. Consider the following example:

28) there is a very big tv with 42 <<bosah>> (for '*inches*') … (NPLTR File 352)

هناك تلفاز كبير 42 بوصة

Hunak tilfaj kabyr 42 būwṣah

In the example above, the learner transliterated the Arabic word “بوصة” “*būwṣah”,* which means *“inch”* in English. Of course, intralingual factors are involved here, too: it may be that this learner resorted to using a transliterated Arabic word simply because they have not yet learned the English word “inch”, so in this sense this error could be seen as a language development limitation. However, the position of this thesis is that it would be strange to code the clear use of an Arabic-derived substitute as a case of intralingual error; on the contrary, an interlingual classification makes much more sense in this context. And in this particular case, further support for an interlingual coding comes from the fact that the equivalent Arabic terminologies for the English measurements: *kilometre, meter* and *centimetre* are almost the same: “كيلومتر” *“kilomitr”,* “متر” *“mitr”* and “سنتيمتر” *“santimitr”* respectively. Therefore, the learner may have assumed that the equivalent English word for “بوصة” *“būwṣah*” will be a similar word “*būwṣah”*. If this is the case, this error type should clearly be classified as an intralingual error rather than an interlingual one.

Finally, a review of example (28) shows that the word “بوصة” *“būwṣah”* is in the singular form and followed the number “42”. This raises the same problem with number agreement as was discussed in Section 7.4.11 above. In Arabic, as we saw, plurals from 11 in number upwards are given in their singular forms. Accordingly, it is therefore likely that the word *“būwṣah*” is an interlingual noun phrase subcategory error as a result of both literal translation and number agreement error.

7.4.13 Wrong Word Order Error

In Arabic, the word order of phrases and sentences can sometimes be very different from equivalent English phrases and sentences. For instance, adjectives in English proceed the nouns they describe, whereas in Arabic they come after the nouns. In the analysis of noun phrase errors in the current thesis, the learners were often found to mis-order compound nouns in ways that appear consistent with L1 interference, as example (29) below shows.

> 29) The oFFice was in << Center City>> … (NPWWO File 17)
>
> المكتب كان في (وسط المدينة)
>
> *Al-maktb kāna fii (wasṭ al-madīna)*

This finding is in line with Ababneh (2019), who also reported wrong word order errors (e.g., *section women, centre operation, level women,* etc.) committed by Saudi English learners at both the University of Tabuk and the University of Hafr Al Batin when translating words and phrases.

Conclusion

This chapter aimed to determine the extent to which L1 influence affected noun phrase errors in the writing of Arab English learners and how this influence developed across the three university academic levels. To achieve this aim, a comprehensive and quasi-longitudinal analysis was conducted on noun phrase errors and subcategory errors that were identified in the LEFLL corpus. The analysis intends to offer plausible (if not definitively proven) accounts of the most likely causes of these types of errors. At its most general level, the analysis revealed that 13 out of the 17 noun phrase subcategory errors identified in the current research were either purely, predominately or partially interlingual.

The average percentage of interlingual noun phrase errors in the LEFLL corpus was 46.89%. Across the three LEFLL sub-corpora, the Year 2 sub-corpus represented an anomalous stage between the Year 1 and Year 3 sub-corpora. Whereas the percentages of interlingual noun phrase errors in the Year 1 and Year 3 (43.32% and 36.51% respectively) were less than the percentages of intralingual noun phrase errors in the same sub-corpora, the percentage of interlingual noun phrase errors in the Year 2 sub-corpus (56.82%) was higher than the percentage of intralingual errors in the same sub-corpus. This may indicate that first language (Arabic) influence in learning English as foreign language does not always follow a steady rate of change (e.g., an increase or decrease in the percentages of interlingual errors) as the analysis moves from one level to another. This is consistent with the same general trends that were observed in the analysis of spelling errors reported in Chapter 6. The steady decline in the percentages of interlingual spelling errors across the three LEFLL sub-corpora observed in Chapter 6, Section 6.4 influenced the percentages of spelling errors reported in Chapter 5, Section 5.3 when the frequencies of spelling errors in the three LEFLL sub-corpora were calculated based on the PFEC approach. As was reported in Chapter 5, Section 5.3, the percentages of spelling errors showed a steady decline as the analysis moved from one level to another. On the other hand, this may indicate that the learners' first language does not necessarily follow linear changes in different error categories. This point will be discussed further in the following chapter, which deals with verb phrase subcategory errors.

In the following chapter, verb phrase errors produced by the Libyan English as foreign language learners in the LEFLL corpus will be analysed. Once again, the chapter will take a particular interest in evaluating whether and to what extent the learners' first language may have played a decisive role in producing different types of verb phrase errors identified by the corpus analysis.

CHAPTER 8: ANALYSIS OF VERB PHRASE SUBCATEGORY ERRORS IN THE LEFLL CORPUS

Introduction

As with noun phrase errors, and unlike spelling errors, verb phrase errors and the constituents of verb phrase errors (e.g., auxiliary errors, tense errors, copula errors, etc.) have attracted considerable interest among error analysis researchers (cf. Wee et al., 2010; Masruddin, 2019).

This chapter aims to build on these previous studies by carrying out a comprehensive and quasi-longitudinal analysis of verb phrase subcategory errors identified in the LEFLL corpus. The chapter aims to investigate the causes that led to these types of verb phrase errors (i.e., whether errors are most plausibly categorised as interlingual or intralingual), and to study in particular detail the development of any interlingual verb phrase errors that might exist in the LEFLL corpus. Before proceeding to the results, however, the chapter will begin with a brief summary of the methodology that was used to extract and classify the errors under analysis here.

8.1 Verb Phrase Error Analysis – the CEA Approach

To conduct a computer-aided error analysis (CEA) for verb phrase error analysis, each verb phrase subcategory error (e.g., auxiliary omission error, tense error, wrong word class, etc) was tagged by inserting an error tagging code that describes the alterations that affected the verb phrase subcategory error. This was performed manually via Dexter Coder Tools (see Section 3.3.1 for a technical overview). The error tagging code system that was used to error tag the LEFLL corpus was adopted from Dagneaux et al. (1998). As discussed earlier in Section 4.2.2, this tagging system is hierarchical, with each verb phrase error tagging code consisting of two or three levels depending on the information needed to describe the verb phrase subcategory error in question. The first level marks the major error category which describes the phrase type. This is followed by one or two subcategory codes. For instance, the verb phrase error tagging code ‘VPT’ consists of two levels. The first level of the error tagging code ‘VP’ refers to the major category code ‘*Verb Phrase’* and the second level of the error tagging code ‘T’ represents the ‘*Tense’* error type. Thus, the error tagging code ‘VPT’ means ‘tense confusion error occurring in the verb phrase’. To take a more complex example, the verb phrase error tagging code ‘VPAUXOM’ consists of three levels. The major category code ‘VP’ is followed by the two subcategory codes: ‘AUX’, which represents ‘Auxiliary’, and ‘OM’, which denotes the alteration type ‘*Omission*’ that affects the verb phrase subcategory error. In addition to the two and three levels of verb phrase error tagging codes, the process of error tagging also involved inserting the following tagging codes:

- ‘VP’: The ‘VP’ tagging code has been created to tag each verb phrase in the LEFLL.
- ‘VPE’: The ‘VPE’ tagging code has been created to tag the range/series of verb phrase constituents where the verb phrase subcategory error was identified.
- ‘VPPOTE’: The ‘VPPOTE’ tagging code has been created to tag the verb phrase constituent that has the potential for error (see Section 4.2.2 in Chapter 4 for detailed discussion).
- Corrections: where appropriate, corrections for each of the verb phrase subcategory errors were provided. Providing corrections for some verb phrase subcategory errors makes it easier for the analyst to understand why a verb phrase subcategory error has been classified in a particular way.

Inevitably, the process of error tagging was time consuming. In total, 28,864 tagging codes and corrections for verb phrase errors were manually inserted into the LEFLL corpus. A more detailed overview is shown in Table 8.1 below. Following the process of error tagging, each tagging code (e.g., VP, VPPOTE, VPAPF, etc) in each of the 559 tagged files was manually and separately retrieved and converted from the DeXML file format (the only file format that the Dexter Coder Tool requires to tag the corpus files) to .txt file format using the Dexter Search Tool. This was an important procedure as it allowed the retrieved tagging code files to be uploaded to AntConc for the purposes of concordance analysis of the different types of verb phrase subcategory errors. The following tagging codes: 'VP' 'VPE' and 'VPPOTE' were retrieved for the PFEC approach as discussed earlier in Chapter 5, Section 5.3.

	Year 1 Sub-corpus	Year 2 Sub-corpus	Year 3 Sub-corpus	LEFLL Corpus
VP	3593	2839	2899	9331
VPPOTE	4899	3573	4293	12765
VPE	756	817	550	2123
VPAPOSTRED	4	0	1	5
VPAUXOM	4	14	24	42
VPAUXRED	58	13	12	83
VPAUXW	3	2	5	10
VPGERINF	208	62	79	349
VPPV	1	2	2	5
VPAPF	1	24	58	83
VPREG	1	9	1	11
VPT	125	164	177	466
VPVOM	193	328	115	636
VPVRED	20	13	26	59
VPVW	62	21	36	119
VPWWC	8	12	11	31
VPWWO	1	0	2	3
VPFIN	0	3	2	5
VPVNCON	67	150	263	480
CORRECTION	696	793	769	2258
Total	10700	8839	9325	28864

Table 8.1: The distribution of annotated codes of verb phrases and verb phrase errors in the LEFLL corpus

As Table 8.1 above shows, verb phrase errors in the LEFLL corpus were classified into 16 subcategories. Following the classification of the verb phrase subcategory errors, instances of each subcategory type were contrasted with standard reference works on Arabic grammar, and with the author’s own knowledge as an Arabic native speaker and as a teacher of English with substantial experience of working with Arabic L1 learners. The aim of contrasting each verb phrase subcategory error with the learners’ L1 was to determine whether it should be classified as interlingual (i.e., as reflecting influence from the learner’s L1) and/or intralingual (i.e., as reflecting target language-internal developmental processes). In the next section, we will discuss and compare the distribution of interlingual verb phrase subcategory errors in the LEFLL corpus and across the three LEFLL sub-corpora.

8.2 Interlingual Verb Phrase Errors and Subcategory Errors

As with the noun phrase errors reported in the previous chapter, the verb phrase analysis found that verb phrase errors in the LEFLL corpus were roughly equally distributed between interlingual and intralingual types overall. Whereas the average percentage of *interlingual* noun phrase errors (observed in chapter 6) was slightly less than 50%, the average percentage of verb phrase errors, as shown in Figure 8.1 and Table 8.2 below, is slightly higher than 50%. This can also be observed in the Year 1 and Year 2 sub-corpora, where the percentages of interlingual verb phrase errors in the Year 1 and Year 2 sub-corpora are 53.71% and 57.35% respectively. On the other hand, the common feature in both noun phrase errors and verb phrase errors is that there is no steady rate of change (either an increase or decrease) in the percentages of interlingual errors.

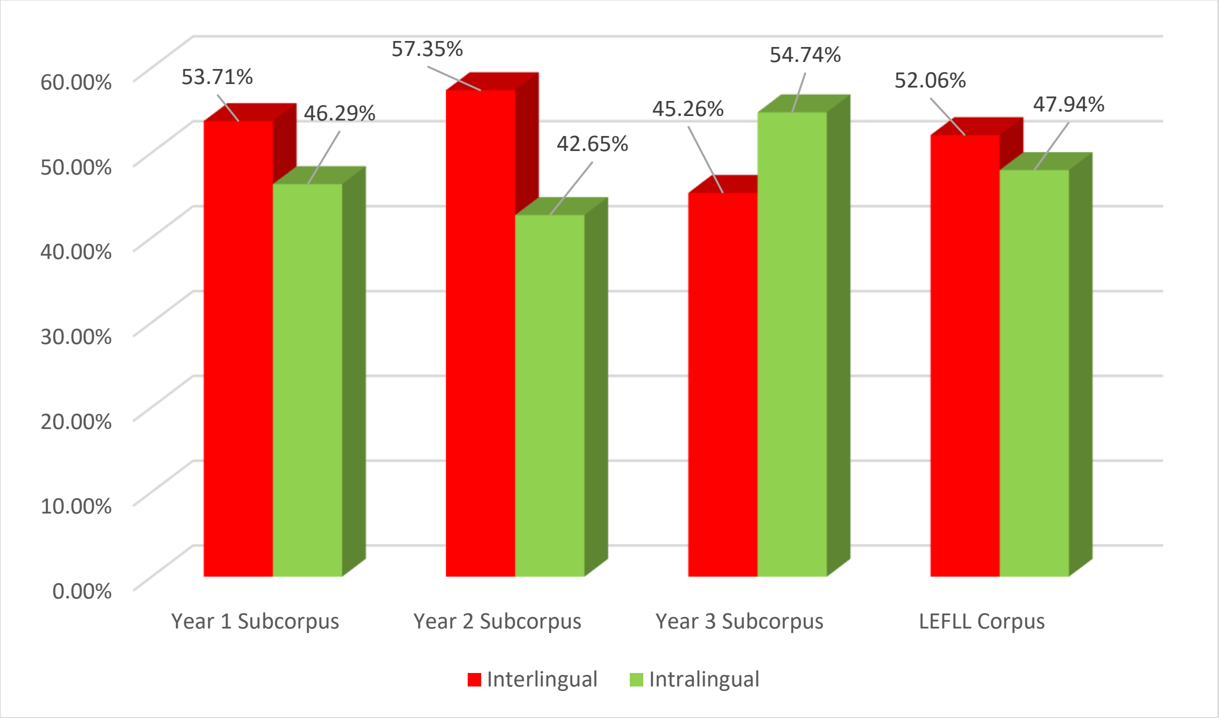


Figure 8.1: The distribution of interlingual vs intralingual verb phrase errors in the LEFLL corpus

As was observed in the analysis of noun phrase errors in Chapter 7, Figure 8.1 and Table 8.2 may also provide an explanation for why the percentage of verb phrase errors in year 2 (shown in Figure 5.6 and Table 5.9 in Section 5.3), was the highest. The percentages of verb phrase errors (as shown in Chapter 5, Section 5.3) in the Year 1, Year 2 and Year 3 sub-corpora were 15.33%, 22.86% and 19.08% respectively. As Figure 8.1 above and Table 8.2 below show, learners in year 2 seem to rely more heavily on their mother language when composing verb phrase structures: in total, 57.35% of verb phrase errors were deemed to be most likely due to L1 transfer.

	Inter.	Intra.	Total	Inter. %	Intra. %
Year 1 Sub-corpus	405	349	754	53.71%	46.29%
Year 2 Sub-corpus	464	345	809	57.35%	42.65%
Year 3 Sub-corpus	368	445	813	45.26%	54.74%
LEFLL Corpus	1237	1139	2376	52.06%	47.94%

Table 8.2: The distribution of interlingual vs intralingual verb phrase errors in the LEFLL corpus

Turning now to the subcategory level of the verb phrase errors across the LEFLL corpus, Figure 8.2 below (and presented in more detail in Table 8.3 towards the end of this section) shows that 7 out of 16 verb phrase subcategory errors were purely, predominately or partially due to L1 interference. More specifically, all auxiliary omission errors are purely interlingual, 88.52% of main verb omission errors are interlingual whereas 20.21% of verb-noun concord (subject-verb agreement) errors are interlingual (see Sections 8.4.1, 8.4.2 and 8.4.4 below for examples of auxiliary omission, main verb omission and verb-noun concord errors).

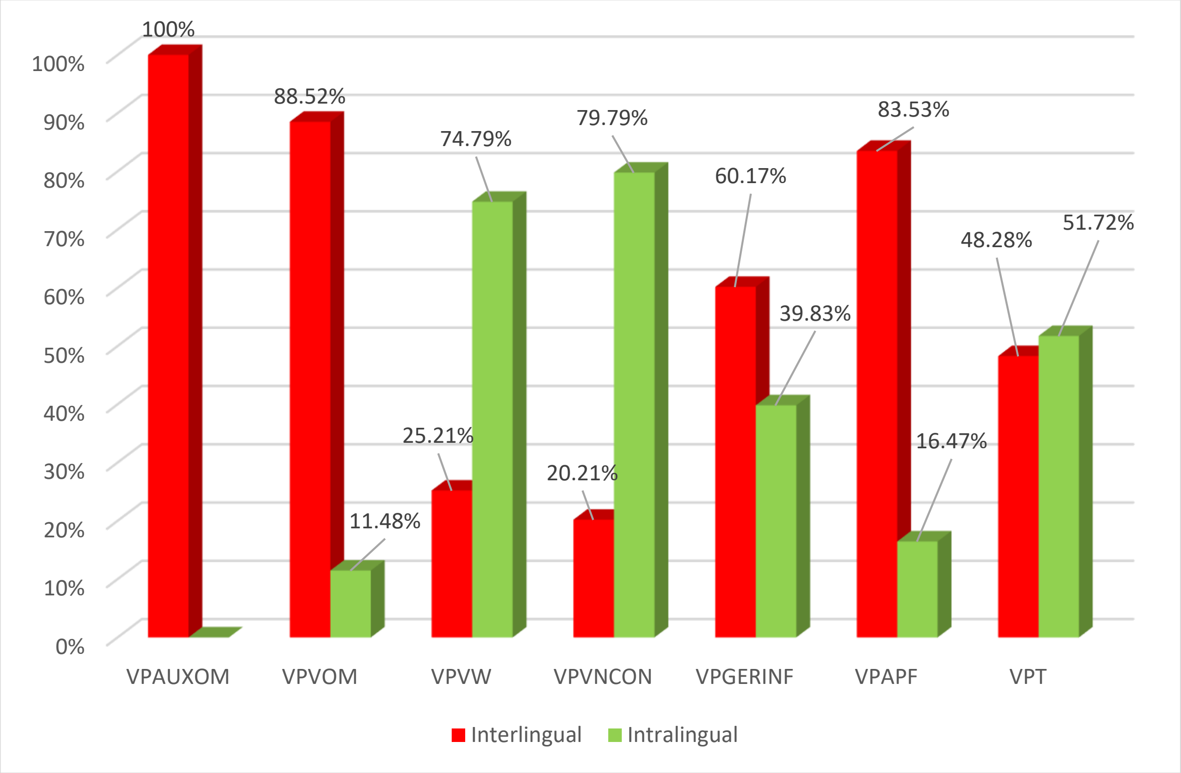


Figure 8.2: The interlingual vs intralingual verb phrase subcategory errors in the LEFLL corpus

Across the three LEFLL sub-corpora, Figures 8.3, 8.4 and 8.5 illustrate the distribution of interlingual vs intralingual verb phrase subcategory errors in the Year 1, Year 2 and Year 3 sub-corpora respectively. A comparison between the three sub-corpora shows that, with the exception of the auxiliary omission errors (in each LEFLL sub-corpus), which are completely interlingual, the percentage of interlingual errors of each verb phrase subcategory error varied from one level to another. Yet again, year 2 proved to be an anomalous stage between Year 1 and Year 3 in each verb phrase subcategory error.

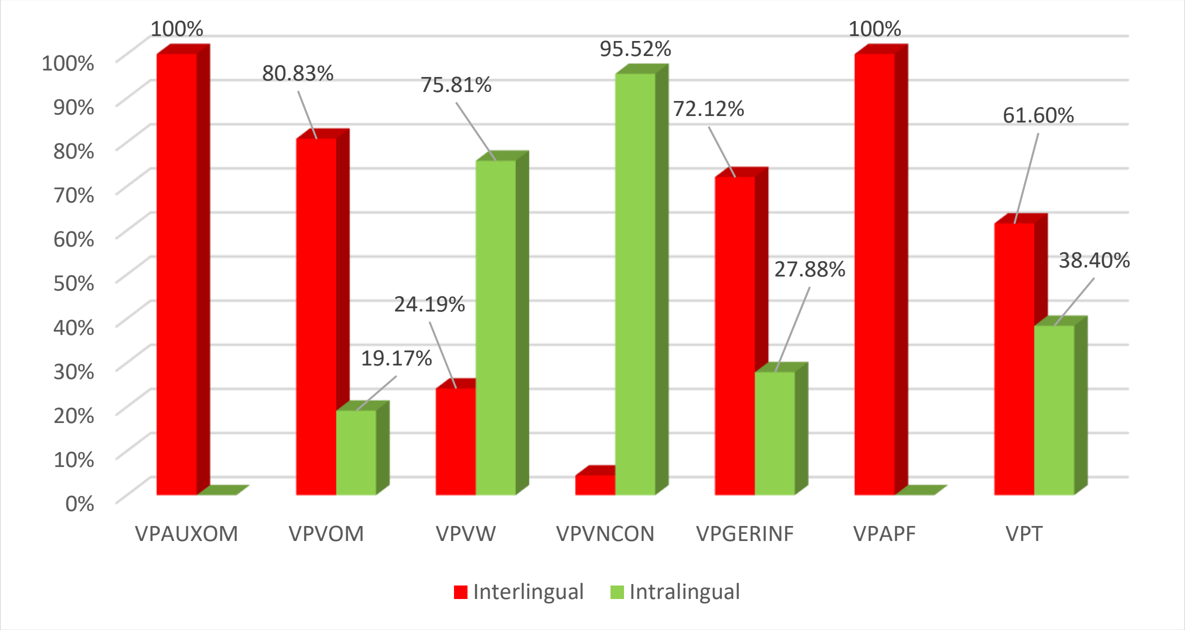


Figure 8.3: The interlingual vs intralingual verb phrase subcategory errors in the Year 1 sub-corpus

For instance, the percentage of interlingual verb-noun concord errors in the Year 1 sub-corpus was 4.48%. This increased to 32.67% in the Year 2 sub-corpus, before decreasing again to 17.11% in the Year 3 sub-corpus. Conversely, the percentage of interlingual tense errors fell from 61.60% in the Year 1 sub-corpus to 26.83% in the Year 2 sub-corpus, before increasing again to 58.76% in the Year 3 sub-corpus. These trends may indicate that the role of the learners' L1 differs from one linguistic constituent of a verb phrase (e.g., *auxiliary, main verb, tense, etc.*) to another as well as from one level to another.

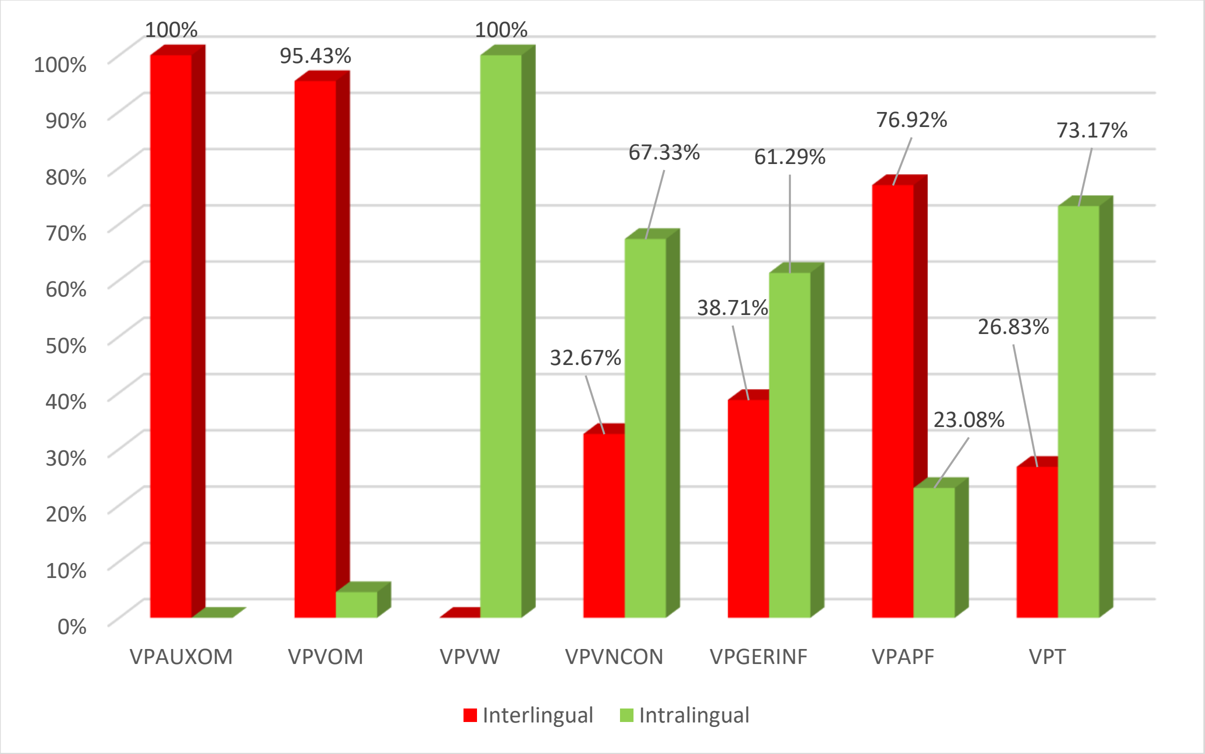


Figure 8.4: The interlingual vs intralingual verb phrase subcategory errors in the Year 2 sub-corpus

By contrasting the variations in the percentages of interlingual verb phrase subcategory errors in the same LEFLL sub-corpus and across the three LEFLL sub-corpora with the frequency of verb phrase subcategory errors given in Table 8.1 above, the picture that emerges is one in which the learners represented by LEFLL appear not to be making steady progress in each constituent in a verb phrase. It also seems that the role of the learners' first language differs from one constituent of a verb phrase (e.g., *auxiliaries, main verbs, tenses*, etc.) to another across the three levels.

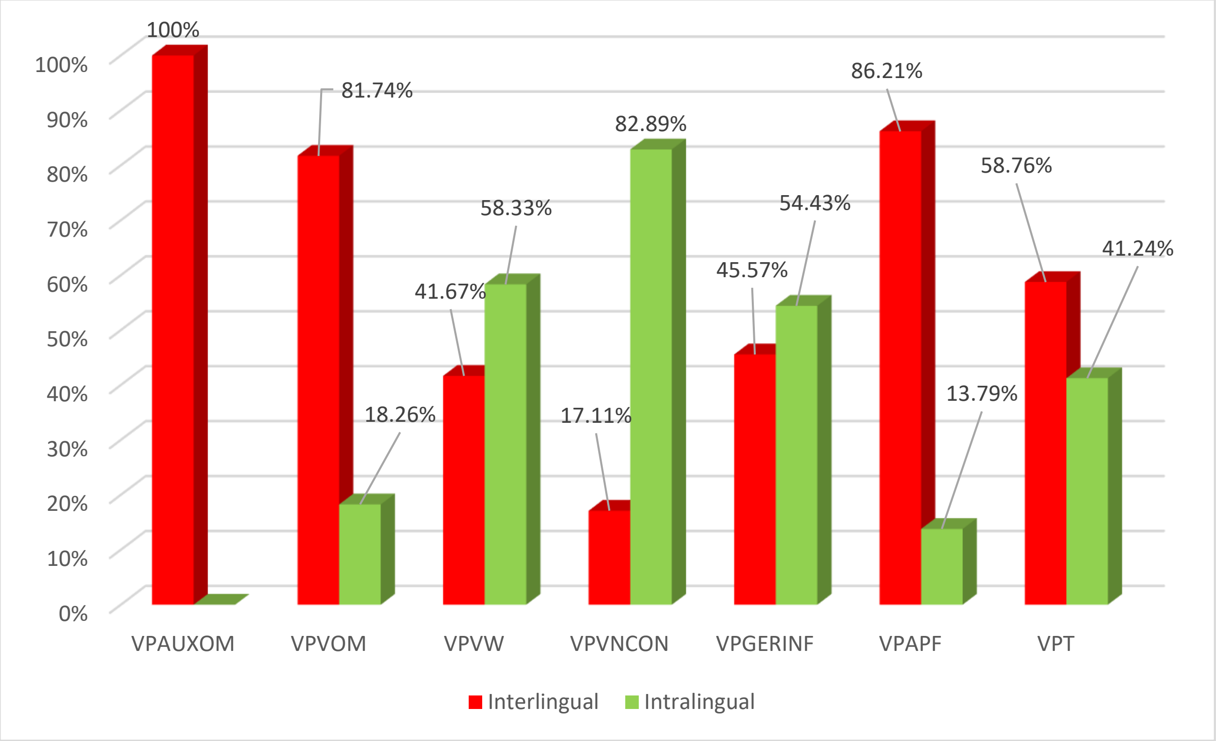


Figure 8.5: The interlingual vs intralingual verb phrase subcategory errors in the Year 3 sub-corpus

These observations are of clear pedagogic relevance. In particular, they suggest that English teachers working with these students (and perhaps other students from an Arabic L1 background) may need to work on ensuring that their learners have fully learned and are able to apply the appropriate rules for forming each constituent of a verb phrase. These findings also suggest that learners may benefit from teaching input that explicitly highlights the differences between verb phrase formation in English and Arabic.

Error Type	Year 1 Sub-corpus				Year 2 Sub-corpus				Year 3 Sub-corpus				LEFLL Corpus			
	Inter.		Intra.		Inter.		Intra.		Inter.		Intra.		Inter.		Intra.	
	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%
Auxiliary Omission	3	100	0	0.00	14	100	0	0.00	24	100	0	0.00	41	100	0	0.00
Main Verb Omission	156	80.83	37	19.17	313	95.43	15	4.57	94	81.74	21	18.26	563	88.52	73	11.48
Wrong Verb	15	24.19	47	75.81	0	0.00	21	100	15	41.67	21	58.33	30	25.21	89	74.79
Verb-Noun Concord	3	4.48	64	95.52	49	32.67	101	67.33	45	17.11	218	82.89	97	20.21	383	79.79
Gerund/Infinitive Confusion	150	72.12	58	27.88	24	38.71	38	61.29	36	45.57	43	54.43	210	60.17	139	39.83
Active/Passive Voice Confusion	1	100	0	0.00	20	76.92	6	23.08	50	86.21	8	13.79	71	83.53	14	16.47
Tense	77	61.60	48	38.40	44	26.83	120	73.17	104	58.76	73	41.24	225	48.28	241	51.72

Table 8.3: The percentages of interlingual vs intralingual verb phrase subcategory errors in the LEFLL corpus

To summarise the discussion so far, this introductory overview section has described and briefly discussed the overall distribution of interlingual verb phrase errors and subcategory errors in the LEFLL corpus and across the three LEFLL sub-corpora. The aim of this presentation has been to compare the current results with the corresponding results obtained from the analysis of spelling and noun phrase error analysis chapters (Chapter 6 and 7 respectively). Our aim has also been to consider the development of interlingual verb phrase subcategory errors across the three levels represented by the LEFLL corpus. In order to validate these general findings, Section 8.4 will discuss the reasons why these 7 verb phrase subcategory errors were classified as interlingual errors in the current analysis. The following section will discuss the intralingual verb phrase errors and subcategory errors.

8.3 Intralingual Verb Phrase Errors and Subcategory Errors

As was observed in Section 8.1 above, the average percentage of intralingual verb phrase errors in the LEFLL corpus is slightly less than 50%, and hovers slightly below or above 50% across each of the three LEFLL sub-corpora (46.29%, 42.65% and 54.74% in Year 1 Year 2 and Year 3 respectively). These errors were marked as intralingual errors because they do not exhibit any clear or obvious signs of first language influence. Whereas 7 out of 16 verb phrase subcategory errors in the LEFLL corpus and across the three LEFLL sub-corpora were purely, predominately or partially interlingual, the remaining (9 out of 16) verb phrase subcategory errors were found to be purely intralingual. The verb phrase subcategory errors that are predominately and partially interlingual (see Section 8.2 above) also contain intralingual verb phrase errors. However, the discussion in this section will be on the verb phrase subcategory errors that are purely intralingual only. It is worth mentioning that the main purpose of this chapter, as stated in the introduction section above, is to focus on and discuss the interlingual verb phrase subcategory errors for pedagogical purposes. Nevertheless, given the preponderance of intralingual errors in the verb phrase data extracted from LEFLL, it is important to begin by providing a survey some of intralingual verb phrase subcategory errors and why they have been classified as intralingual rather than interlingual types. As well as being useful and interesting in its own right, this discussion may also serve indirectly as an indicator of the robustness of the analysis in general. That is, it is hoped that the reader will see that the analyst was not intrinsically biased towards classifying errors as interlingual from the outset. On the contrary, it should become clear from the following discussion that decisions regarding the interlingual or intralingual status of each error in the corpus were made on a uniformly rigorous and consistent application of the same sets of guiding principles.

Before proceeding any further, it is worth noting here that although intralingual verb phrase errors substantially outweigh interlingual verb phrase errors in the current data, the analysis also showed that the frequency of interlingual verb phrase errors is still high and in some specific cases is even higher than the frequency of intralingual verb phrase errors. For instance, in the Year 1 sub-corpus where the percentages of interlingual vs intralingual verb phrase errors, as shown in Figure 8.1 above, were 53.71% vs 46.29% respectively, the frequencies of interlingual vs intralingual verb phrase errors in the same sub-corpus, as shown in Table 8.2 above, were 405 and 349 respectively.

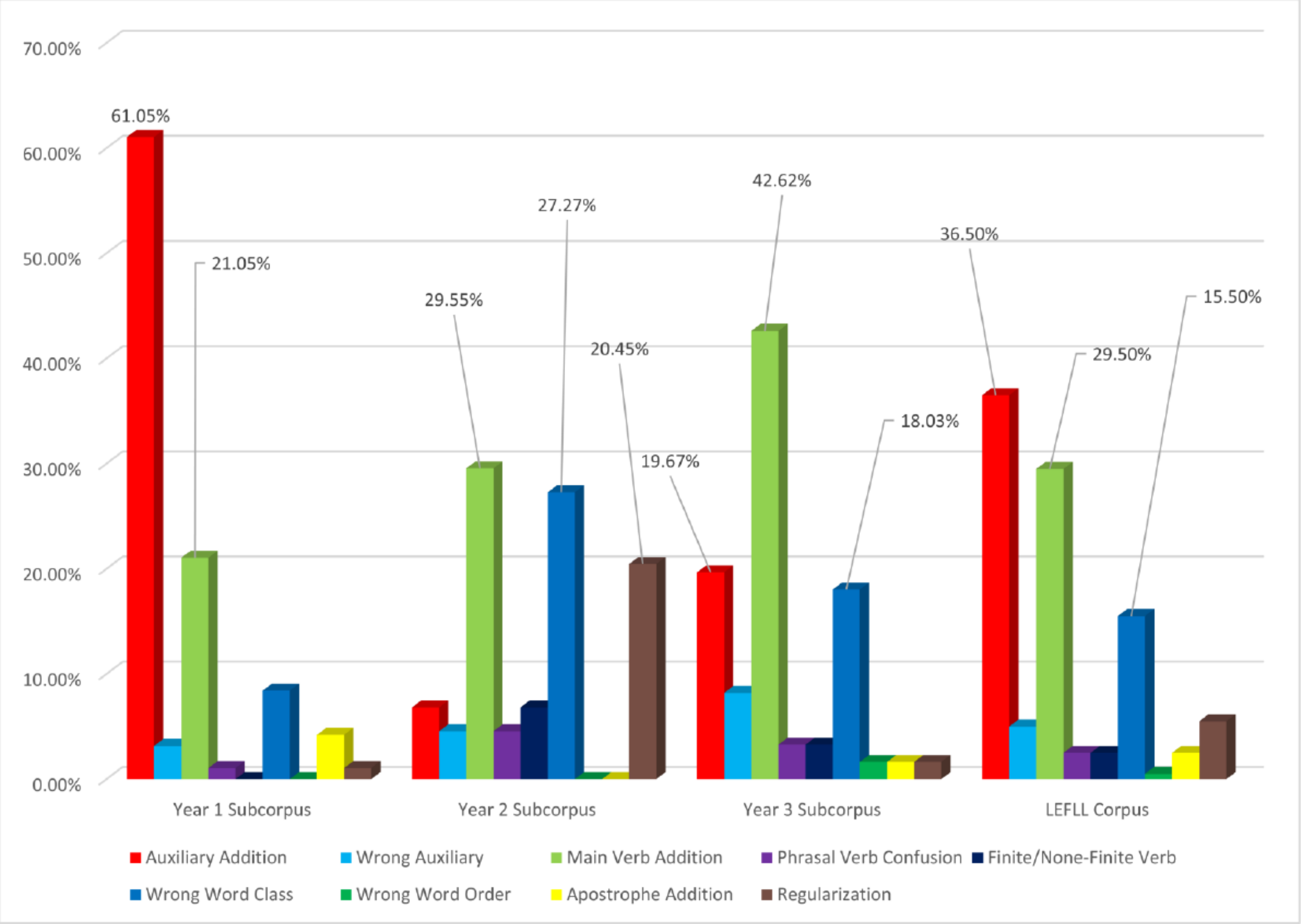


Figure 8.6: The intralingual verb phrase subcategory errors in the LEFLL corpus

When examining verb phrase subcategory errors, in the LEFLL corpus and across the LEFLL sub-corpora, that are purely intralingual as Figure 8.6 above and Table 8.4 below show that, overall, the auxiliaries and main verbs addition errors represented the highest percentages in the LEFLL corpus and in the Year 1 sub-corpus. Across the three LEFLL sub-corpora, there is a steady increase in the percentages of main verb addition errors. Another common feature that can be observed in the Year 2 and Year 3 sub-corpora is that the percentages of wrong word class errors, in both LEFLL sub-corpora, are relatively high. This could be linked to a corresponding increase in the type-token ratio in the Year 2 and Year 3 sub-corpus compared to the Year 1 sub-corpus, as was observed earlier (see Table 4.1 in Chapter 4, Section 4.1.2.1).

	Year 1 Sub-corpus	Year 2 Sub-corpus	Year 3 Sub-corpus	LEFLL Corpus
Auxiliary Addition	(n=58) 61.05%	(n=3) 6.82%	(n=12) 19.67%	(n=73) 36.50%
Wrong Auxiliary	(n=3) 3.16%	(n=2) 4.55%	(n=5) 8.20%	(n=10) 5.00%
Main Verb Addition	(n=20) 21.05%	(n=13) 29.55%	(n=26) 42.62%	(n=59) 29.50%
Phrasal Verb Confusion	(n=1) 1.05%	(n=2) 4.55%	(n=2) 3.28%	(n=5) 2.50%
Finite/Non-Finite Verb	(n=0) 0.00%	(n=3) 6.82%	(n=2) 3.28%	(n=5) 2.50%
Wrong Word Class	(n=8) 8.42%	(n=12) 27.27%	(n=11) 18.03%	(n=31) 15.50%
Wrong Word Order	(n=0) 0.00%	(n=0) 0.00%	(n=1) 1.64%	(n=1) 0.50%
Apostrophe Addition	(n=4) 4.21%	(n=0) 0.00%	(n=1) 1.64%	(n=5) 2.50%
Regularization	(n=1) 1.05%	(n=9) 20.45%	(n=1) 1.64%	(n=11) 5.50%

Table 8.4: The intralingual verb phrase subcategory errors in the LEFLL corpus

So far, this section has presented the distribution of purely intralingual verb phrase subcategory errors in the LEFLL corpus and sub-corpora. The following subsections will discuss these intralingual verb phrase subcategory errors in more detail, supported by examples from the LEFLL corpus along with equivalent Arabic structures and the transliteration of these Arabic structures using the *Ijmes transliteration system for Arabic, Persian and Turkish* (see Appendix F).

8.3.1 Auxiliary Addition Error

In this type of verb phrase subcategory error, learners unnecessarily add an auxiliary verb. The most common type of auxiliary addition error is the unnecessary addition of the verb *to be* in the simple present tense. This may be attributed to apparent confusion on the part of the learners between the present simple and present progressive, as the following example may show.

1) I<<'m stay>> in my room with my brother. (VPAUXRED File 523)

   أنا (أسكن) في غرفتي مع أخي

   ʾana (ʾaskun) fii ghurfatī maʿa ʾakhy

There is nothing in Arabic grammar that could plausibly be seen as licensing this type of error; therefore, the only possible option for the current analysis is to classify this error type as intralingual. This concept has been applied to all verb phrase subcategory errors that have been classified as intralingual error types.

8.3.2 The Incorrect Use of Auxiliary Error

In the current data, incorrect auxiliary use most frequently takes the form of the verb *to Be* being used instead of the verb *to do*), as in the following example:

2) I <<am not sleep>> at afternoon, (VPAUXW File 900)

   أنا (لا أنام) عند الظهر

   ʾana (la ʾanām) ʿanda al-ẓuhur

8.3.3 Main Verb Addition Error

In addition to the unnecessary addition of auxiliaries, learners were also found to add unnecessary verbal elements to the main verb of a clause. In this type of verb phrase subcategory error, learners most frequently add unnecessary auxiliaries (e.g., *is, am, are*, etc.), seemingly intending them to function as main verbs. Learners also add main verbs where they should not be added, e.g., in simple sentences that contain coordinate conjunctions. Typical examples of this error type are provided in (3) and (4) below:

3) Secondly, we stud for get some thing <<is>> important of our life, (VPVRED File 4)

ثانيا، نحن ندرس لنحصل على شي (يكون) مهما في حياتنا

Thānyan, naḥnu nadrus linaḥul ᶜala shaī (yakunu) muhiman fii ḥayātuna

4) I like watching TV and I <<lik>> listing (for *listening*) muic (for *music*) (VPVRED File 825)

أنا أحب مشاهدة التلفاز وأنا (أحب) سماع الموسيقى

Ana ᵓauḥibu mushāhadat al-tilfāz wa ana (ᵓauhibu) samaᶜ al-mūsīqa

8.3.4 Phrasal Verb Confusion Error

Few instances of this error were observed in the LEFLL corpus. In this type of verb phrase subcategory error, learners encounter problems in using a phrasal verb, e.g., missing the particle (an adverb or preposition) that is needed to make the phrasal verb or adding an unnecessary element (such as a verb). For instance, in example (5), the preposition *'on'* was omitted, whereas in example (6) *'to'* and *'get'* were unnecessary added:

5) … the bride wears whight (for *white*) dress and <<puts>> the make up (VPPV File 366)

العروس ترتدي ثوب أبيض (وتضع) المكياج

Al̠-ʿarūs tartadī thob abyaḍ wa (taḍʿ) almikyāj

6) … my alarm clock <<wakes to get up>> 6:30. (VPPV File 741)

منبهي (يرن لأستيقظ)

Munabihy (yarin liastyqẓ)

8.3.5 Confusion between Finite and Non-Finite Verbs

Finite verbs change their forms to show the tense, person and number (e.g., *I walk, she walks, they talked, etc.*) whereas non-finite verbs do not change their forms as they do not show tense, person or number. Non-finite forms are of three types: *infinitive with or without 'to'* (e.g., *to go, go*), *gerund* (e.g., *going*) and *participle* (e.g., *gone*). In this type of error, learners were found to use non-finite verbs instead of finite verbs as the main verb of an independent clause, as the following example shows:

7) We study, play, and <<lestining>> to music in it. (VPFIN File 525)

نحن ندرس، نلعب و(نستمع) إلى الموسيقى فيها

naḥnu nadrus, nal^{c}ab wa (nastam^c) ilaa al-musīqa fiiha

8.3.6 Wrong Word Class Error

In this type of verb phrase error, learners use different word classes (e.g., nouns, adjectives, adverbs, etc.) instead of the verbal form of a given word. For instance, in the following example, it seems that the learner is confused between *'improve'* (as a verb) and *'improvement'* (as a noun).

8) Finally we never <<improvement>> unless we have a modern equipment (VPWWC File 71)

في النهاية نحن أبدا لا (نتحسن) مالم تكن لدينا معدة/جهاز حديث

Fii al-nihāyah ᵓabadan laa (nataḥsan) mālam takun ladyna mu^{c}ida/jihāz ḥadīth

A possible explanation for this type of verb phrase subcategory error is that the learners may have problems with word derivation rules in English, or that they have not yet fully understood the concept of parts of speech as having different grammatical roles in a clause. In any case, there is nothing in Arabic grammar that would licence or explain this type of error (Arabic makes equally clear morphological distinctions between different parts of speech). Accordingly, this error type can confidently be classified as intralingual.

8.3.7 Wrong Word Order Error

In the LEFLL data, wrong word order errors occur in verb phrases when the learner mis-orders the constituents that form the verb phrase, as shown in example (9) below:

9) … feeling <<not should be>> in focuse (VPWWO File 27)

... الشعور (يجب أن لا) يكون محل أهتمام

… al-shuᶜūr (yajib ᵓan laa) yakun maḥal ᵓatimam

This type of error cannot easily be seen as anything other than a sign that the learner in question has not yet fully mastered the word order rules governing English verb phrase construction.

8.3.8 Apostrophe Addition Error

The erroneous addition of an apostrophe may occur as a result of overgeneralizing of the short form rules specific to English, where one or more letters are removed from a word or a phrase. For instance, in the following example, the addition of the apostrophe between the letters '*n'* and *'t'* is most likely to be a result of the learner misapplying the English short form rule for the adverb (*n't*) for (*not*) to the lexical verb 'want'.

10) I <<wan't>> to know more about it. (VPAPOSTRED File 826)

أنا (أريد) أن أعرف الكثير عنها

Ana (ᵓaurīd) ᵓan ᵓaᶜrif al-kathīr ᶜanhā

8.3.9 Regularization Error

In the LEFLL data, regularization errors occur when the learner misapplies the English regular past tense or past participle verb form rule (the addition of '*d'* or *'ed'*) to a verb that has an irregular conjugation in English. In example 11 below, the learner has incorrectly regularised 'chosen' as 'chosed':

11) I <<chosed>> simple style (VPREG File 466)

أنا (أخترت) أسلوب بسيط

Ana (ʾakhtartu) ʾaslūb basīṭ

Once again, this is clearly an intralingual error, reflecting as it does an incomplete and thus still developing grasp of aspects of English grammar on the part of this student, with no discernible evidence of first language transfer involved.

8.4 Discussion

As was observed in Section 8.2 above, the average percentage of interlingual verb phrase errors in the LEFLL corpus is 52.06%, which is slightly higher than the percentage of interlingual noun phrase errors (46.89%) observed in Section 7.2 above and less than the percentages of interlingual spelling errors (68.96%) observed earlier in Section 6.4. Across the three LEFLL sub-corpora, the percentages of interlingual verb phrase errors did not show a steady change (either an increase or a decrease), as was observed in the percentages of interlingual spelling errors across the LEFLL sub-corpora. In fact, this is a common feature between noun phrase and verb phrase errors; in both cases there is no steady change in the percentages of interlingual errors, and in both cases the Year 2 data constitute an anomalous stage between Year 1 and Year 3.

This may support the claim that the steady decline in the percentages of interlingual spelling errors across the three LEFLL sub-corpora observed earlier in Figure 6.11 and Table 6.12 in Section 6.4 has affected the percentages of spelling errors when they were calculated via the PFEC approach as was observed earlier in Figure 5.6 and Table 5.9 in Section 5.3. As was mentioned in Section 5.3, the percentages of spelling errors showed a small but steady decline, unlike the results obtained via the TEC approach. This was not the case with the percentages of noun and verb phrase errors (when they were calculated via the PFEC approach) as they both showed unstable patterns of change across the three LEFLL sub-corpora, with year 2 again being the anomalous stage between years 1 and 3.

The analysis of verb phrase errors above also showed that verb phrase errors can be classified into 16 verb phrase subcategory errors. It was observed that 7 out of the 16 verb phrase subcategory errors were either purely, predominately or partially interlingual. While this suggests that first language interference may be much less of a driver of verb phrase errors than it seems to be of spelling and noun phrase errors, it nevertheless remains the case that there are areas where L1 transfer issues do cause significant and consistent difficulties for the learners represented by the LEFLL corpus, and thus perhaps also for other Arabic L1 learners of English as well.

Having discussed the L1 influence and it affected the verb phrase errors in the LEFLL corpus and sub-corpora, the following subsection will illustrate why, specifically, 7 verb phrase subcategory errors out of 16 verb phrase subcategory errors were either purely, predominantly or partially classified as interlingual verb phrase error types. This will also be supported by examples from the LEFLL corpus along with equivalent Arabic structures and the transliteration of these Arabic structures using the *Ijmes transliteration system for Arabic, Persian and Turkish* (see Appendix F).

8.4.1 Auxiliary Omission Error

This type of error occurred where the learners were found to omit the auxiliaries (e.g., *is, am, was, have, etc.*). It was marked as an interlingual type because Arabic and English contrastive analysis showed that Arabic does not have auxiliaries. As a result, this could be transferred to English learning where auxiliaries were removed as the following two examples may show.

12) After I graduate I <<going>> to work as a translator (VPAUXOM File 824)

بعد أن أتخرج أنا (سأذهب) لاعمل كمترجم

baᶜda ān atakharaj ana (saᵓadhihab) li-aᶜmal kamutarjim

13) That <<not mean>> that you have to treat your friend the way they treat you (VPAUXOM File 77)

هذا (لا يعني) أنه يجب عليك أن تعامل صديقك بالطريقة التي يعاملونك بها

Hadha (lā yaᶜnī) anahu yajib ᶜalyka ᵓan tuᶜamal ṣadīquka bilṭarīqah al-laty yuᶜāmilwnaka biha

In example (12), the auxiliary *'am'* was missing, and in example (13), the student failed to provide the auxiliary *'does'*. Auxiliary omission errors among Arab speakers' English learners are very common in the current data and have been observed by numerous previous studies. (e.g., Sabbah, 2015; Younes & Albalawi, 2015; Alasfour, 2018; Qaddumi & Walweel, 2018; Alahmadi & Lahlali, 2019; Aljohani, 2019; Btoosh, 2019; Khan, 2019; Khatter, 2019). Accordingly, it is plausible in most cases to code this error type as interlingual.

8.4.2 Main Verb Omission Error

In addition to the omission of auxiliaries due to the L1 influence, learners were also found to omit the main verb as a result of Arabic interference in cases where the equivalent Arabic structures do not require main verbs. The common type of main verb omission is the omission of the copula (i.e., the linking verb) as shown in the following example, where the learner has removed the copula *'is'*:

14) but this room << >> not for me only

لكن هذه الغرفة ليست لي فقط

Lakin hadhhi al-ghurfah lysit lī faqṭ

The interpretation here is that the learner has committed this error because the copula is typically absent in the equivalent Arabic structure (provided above). This finding agrees with previous research findings, which also observed the typical omission of the English *copula* in writing produced by Arab learners of English. For instance, Alasfour (2018) investigated the first language influence in 50 academic papers written by Arab ESL learners, and observed that all types of main verb omission errors were copulas. She attributed this to first language interference. In another study, Alshayban (2012) investigated the omission of the English *copula* by 100 Saudi EFL learner, and also concluded that this type of error is most likely to have occurred as a result of first language transfer.

8.4.3 Wrong Choice of Main Verb Error

In the current data, errors featuring a wrong choice of main verb seem to be committed as a result of direct translation from Arabic to English. The analysis revealed that the main verbs that were incorrectly chosen and used in the English structures would always be correct and appropriate choices in equivalent Arabic structures. To illustrate, consider the following example:

15) Then I <<put my Phone in>> charger. (VPVW File 747)

ثم أنا أضع نقالي في الشحن

Thuma Ana aḍa'a naqalī Fii al-shaḥn

As the example above may show, the learner has used the verb '*put … in*' instead of *'connect … to'* to refer to connecting his/her mobile phone to the charger. Support for an interlingual interpretation for errors such as this comes from a previous study by Naba'h (2011), who observed frequent instances of wrong verb choices in collocations in English exam papers written by Jordanian students. Naba'h attributed some of these wrong verb choice types to students formulating sentences in Arabic and then translating them into English. While there is no easy way of verifying this claim in the current data, the consistent and persistent nature of this error means that it is plausible to code it as interlingual.

8.4.4 Verb-Noun Concord Error

Verb-noun concord (or subject-verb agreement) error occurs when the subject and verb do not agree with each other in number. In Arabic sentences structure, subjects and verbs must agree with each other in number (singular, dual or plural), gender (masculine or feminine) and person (first, second or third). This may not correspond with English sentences structure as shown in the following example:

16) <<my hobbies is>> riding horses and surfing (VPVNCON File 908)

(هواياتي تكون) ركوب الخيل والتصفح على الأمواج

(hwīāty takwnu) rukūbu alkhīl wa altaṣfḥ 'alá alamwāj

In example (16) above, the subject '*my hobbies'* was paired with the verb to be *'is'* instead of *'are'.* As discussed earlier in Section 7.4.6, in Arabic, plural inanimate objects are treated as singular feminine human beings. Thus, a possible explanation of the subject-verb agreement error in example (16) is that the subject *'my hobbies'* was conceptualised by the student as a singular feminine referent, and thus as replaceable with the third person feminine subject pronoun '*she'.* This in turn caused the learner to use the verb for *'is'* instead of *'are'.* Systematic instances of verb-noun concord (subject-verb agreement) error were also observed by Dweik & Othman (2017) and Adila (2019) both of whom also concluded that Arab L1 English learners probably produced this type of error due to Arabic interference of the kind discussed above.

8.4.5 Gerund and Infinitive Confusion Error

Arabic does have a gerund form or an infinitive with *to*. Instead, the verb is used in its simple form. For example, for the English sentence *'I like playing football',* the equivalent Arabic sentence structure would be *'I like (that) (I) play football'.* Similarly, for the English sentence *'I would like to play football',* the equivalent Arabic sentence structure would be *'I would like (that) (I) play football'.* The following two examples from the LEFLL data are typical:

17) I <<would like introduce>> myself (VPGERINF File 913)
أنا أريد أن أقدم نفسي
Ana auryd ann auqadim nafsī

18) I <<like visit>> my friends (VPGERINF File 907)
أنا (أحب أن أزور) أصدقائي
Ana (ˀauḥibu ann azūr) ˀaṣdiqaˀy

This finding has not been noted in any previous error analysis studies of Arabic L1 English learner writing, but the intralingual interpretation advocated here has the backing of Swan and Smith

(2001), who state that gerund and infinitive confusion errors are characteristic of Arab English learners and attribute it to Arabic L1 interference.

8.4.6 Confusion between Active and Passive Voice

The active and passive forms in English and Arabic are different. Whereas the auxiliary '*be*' is necessary to form the passive forms in English, Arabic, on the other hand, does not have the auxiliary *'be',* as discussed earlier in Section 8.4.1. Indeed, there are no auxiliaries at all in Arabic. In many cases, the active and passive forms in Arabic are identical. In writing, the difference can be recognised from the context whereas in speaking it is realised by the pronunciation of the verb. Therefore, the clear confusion between active and passive in the example below could very plausibly be interpreted as being due to L1 Arabic interference.

19) In the war a lot of people <<took>> to the prison … (VPPAF File 12)

في الحرب الكثير من الناس (أخذوا) إلى السجن

Fii al-ḥarib al-kathīr men al-nass (ʾukhdw) ila al-sijin

The above finding agrees with those of previous studies (e.g., Mourtaga, 2004; Dweik & Othman, 2017) which have also observed confusion between active and passive voice forms and the omission of the auxiliary *'be'* in the writing of Arab student English learners. These previous studies also concluded that L1 Arabic interference was the most likely cause of error in these cases.

8.4.7 Tense Error

Tenses in English and Arabic are very different. Tenses in English are linked with time and aspect, and can be further classified into 14 types (e.g., present simple, present continuous, present perfect, past simple, etc.). Tenses in Arabic, in contrast, are not linked with time or aspect. Accordingly, there are only 3 tenses in Arabic: the Imperfect المضارع *al-mudārī,* which denotes incomplete action; the Perfect الماضي *al-mādī,* which indicates completed actions (Haywood & Nahmad, 1962; Wickens, 1980; Mace, 1998); and the future المستقبل *al-mustaqbal.* Given the greater simplicity of the Arabic tense system, it is perhaps not surprising to find that the Arabic L1 learners of English represented in the LEFLL corpus were often found to mix tenses in similar ways to the following:

20) I <<am studying>> in English department in Bengazi (VPT File 806)

أنا (أدرس) في قسم اللغة الأنجليزية في بنغازي

Ana (ᵓadrus) fii qsim al-lghah al-injilyzyah fii Banghazi

21) All my families <<are starting to prepare>> the improtant (for *important*) things … (VPT File 401)

كل أفراد عائلتي (بدأو في تجهيز) الأشياء المهمة

Kul ᵓafrad ᶜāᵓilaty (badᵓau fii tajhīz) al-ᵓashīāᵓ al-muhimah

This finding is in line with previous research studies (e.g., Alamin & Ahmed, 2012; Muftah & Rafik-Galea, 2013; Hamed, 2018), all of whom reported tense confusion errors among Arab speaker English learners, and attributed these types of error to L1 Arabic interference.

Conclusion

This chapter aimed to determine the extent to which L1 influence affected verb phrase errors in the writing of Arab English learners and how this influence developed across the university academic levels. To achieve this aim, a comprehensive and quasi-longitudinal analysis was conducted on verb phrase errors and subcategory errors in the LEFLL corpus. The analysis revealed that, somewhat in contrast to the findings obtained for spelling and noun phrase errors, intralingual errors were more prevalent than interlingual errors in the LEFLL data.

The average percentage of interlingual verb phrase errors in the LEFLL corpus was 52.06% and was slightly higher than the percentage of interlingual noun phrase errors (46.89%) observed earlier in Chapter 7, Section 7.2 and less than the percentage of interlingual spelling errors (68.96%) observed earlier in Chapter 6, Section 6.4. Across the LEFLL sub-corpora, a more detailed analysis revealed that the percentages of interlingual verb phrase errors did not follow a clearly linear pattern of change (either an increase or decrease), and year 2 was found to be an anomalous stage between years 1 and 3, as was the case for noun phrase errors.

The findings in this chapter have potential pedagogical implications. At the very least, they should make English teachers working with Arabic L1 learners aware of the possible verb phrase errors that their students may produce, either due to the intrinsic complexities of English grammar or due to L1 Arabic interference. In either case, it would seem useful for teachers to point out the differences between the rules of verb phrase constituents in both English and Arabic.

CHAPTER 9: GENERAL CONCLUSION

9.1 Restatement of Research Aims

As mentioned in the introductory chapter, the main aim of this thesis was to provide the researchers of SLA with a broader picture of the role of L1 in the writing of Arab English learners. More specifically, the study focused on the identification and classification of errors at three level of analysis – spelling, noun phrases and verb phrases – in a corpus of essays collected from 559 Libyan University undergraduate students across three cohort year group. In the analysis chapters, the thesis sought to address the following two main research questions:

1) To what extent does L1 influence affect spelling, noun phrase and verb phrase errors in the writing of Arab English learners?
2) Does this influence follow the same pattern (either an increase or decrease) as the learners proceed across the university academic levels?

To achieve this aim, the tools and methods of corpus linguistics were used to carry out a comprehensive and quasi-longitudinal analysis of three error categories (spelling, noun and verb phrase errors) in the writing of university-level Libyan Arab English learners. In addition to this main aim, the thesis also aimed to introduce and advocate a new approach to error analysis, which looks not only at the language learners' errors in themselves but also what the language learners have performed correctly. This approach, which I called the '*potential for error counting approach'* (PFEC), calls for empirical investigation into the learner corpus under study in order to verify which constituents (e.g., tokens for spelling, determiners in noun phrases, auxiliaries in verb phrases, etc.) have the *empirically demonstrable* potential for error. This can only be achieved by error tagging the learner corpus, and then retrieving and analysing the language learners' errors on a case-by-case basis. Following this preparatory work, the percentage of each type of error can then be calculated within its relevant error category environment (e.g., the percentage of spelling errors within the total number of words that definitely have the potential for spelling error). Although proposing the PFEC approach was the second aim of this thesis, it was achieved first before moving to achieve the main aim of the thesis, providing the researchers of SLA with a broader picture of the role of L1 in the writing of Arab English learners via a comprehensive and quasi-longitudinal analysis of spelling, noun phrase and verb phrase errors in the LEFLL corpus.

9.2 A Summary of the Main Findings

9.2.1 The Role of L1 in the Writing of Arab English learners.

To achieve the main aim of this thesis, providing the researchers of SLA with a broader picture of the role of L1 in the writing of university-level Libyan Arab English learners. A comprehensive and quasi-longitudinal analysis of spelling, noun and verb phrase errors in the writing was conducted using the tools and methods of corpus linguistics. Errors in the LEFLL sub-corpora were identified, analysed and classified. The spelling errors were analysed and classified based on two error classification systems: the *surface structure* and the *linguistic category* taxonomies proposed by Rimrott & Heift (2005; 2008). The aim of classifying the spelling errors based on the surface structure taxonomy was to provide the minimal information about in what way spelling errors were altered. These alterations are known as 'non-linguistic' alterations. They did not provide any linguistic information but they helped to interpret the linguistic categories.

Using the surface structure taxonomy, spelling errors were found to have been misspelled in four different ways, namely: *omission, addition, substitution* and *transposition.* The analysis revealed that the omission spelling errors occupied the highest proportion of spelling errors in the LEFLL corpus and across the three LEFLL sub-corpora, followed by substitution spelling errors, whereas transposition spelling errors were the lowest. It was assumed that the high percentages of omission and substitution spelling errors could be attributed to the learners' L1 influence as previous spelling errors research studies conducted on Arab students' English learners also reported similar results (cf. Alhaisoni et al., 2015; Othman, 2018).

Following the classification of spelling errors based on the surface structure taxonomy, spelling errors were classified based on the linguistic category taxonomy. This analysis revealed that the interlingual spelling errors constituted the highest proportion of spelling errors in the LEFLL corpus (68.96%) and across the three LEFLL sub-corpora as assumed in the surface structure taxonomy above. It was observed that orthographical spelling errors occupied the highest percentages of interlingual spelling errors, followed by phonological spelling errors. It was also observed that the interlingual spelling errors followed a steady decrease as the learners proceeded from one university level to another vs a steady increase in the percentages of intralingual spelling errors. Moreover, the analysis revealed that the learners encountered major problems with English vowels, and that this could in many cases be plausibly attributed to the huge differences between English and Arabic vowel systems. It was observed that the learners tend to apply the same spelling strategy that they use in their first language, such as the omission of vowels.

Following the analysis of spelling errors, noun phrase subcategory errors were also comprehensively and quasi-longitudinally analysed and classified. This analysis revealed that 13 out of 17 noun phrase subcategory errors in the LEFLL corpus were either purely, predominantly or partially attributable to the L1 Arabic influence. This represented an average of 46.89% out of all noun phrase subcategory errors in the LEFLL corpus. Across the three LEFLL sub-corpora, the analysis revealed that there are no linear changes in the percentages of interlingual noun phrase subcategory errors and the Year 2 sub-corpus represented an anomalous sub-corpus between the Year 1 and the Year 3 sub-corpora.

Following the analysis of noun phrase subcategory errors, verb phrase subcategory errors were analysed and classified in the same way. The analysis showed, somewhat in contrast to the two preceding analyses, that only 7 out of 16 verb phrase subcategory errors identified in the LEFLL corpus either purely, predominantly or partially reflected the first language influence. Nevertheless, this still constituted an average of 52.06% out of all verb phrase subcategory errors in the LEFLL corpus. Across the three LEFLL sub-corpora, the analysis revealed that the interlingual verb phrase subcategory errors did not follow a linear change pattern, and the Year 2 sub-corpus was once again anomalous in comparison to the Year 1 and Year 3 sub-corpora.

Thus, to answer the first research question, “to what extent does L1 influence affect spelling, noun phrase and verb phrase errors in the writing of Arab English learners?”, it can be concluded from the comprehensive and quasi-longitudinal analysis of spelling, noun and verb phrase errors conducted in this thesis that the learners’ first language influence in English learning seems to differ from one linguistic category to another. Specifically, L1 influence was found to be pervasive in spelling errors (nearly 70% of all spelling errors in the current analysis), and less pervasive but still significant and systematic in the case of verb phrase errors (just over half of all verb phrase errors) and noun phrase errors (at just under 50% of all noun phrase errors). It was argued that the pervasiveness of L1 influence on spelling errors may plausibly be attributed to the huge differences between English and Arabic phonological and orthographical systems. Likewise, the thesis has argued that the slightly higher average proportion of interlingual verb phrase subcategory errors than that of interlingual noun phrase subcategory errors in the LEFLL corpus may reflect the fact that the differences between English and Arabic verb phrase construction are more significant than those between English and Arabic noun phrase construction. For instance, Arabic has three tenses – the Imperfect المضارع *al-mudārī,* which denotes incomplete action, the Perfect الماضي *al-mādī,* which indicates completed actions, and the future المستقبل *al-mustaqbal,* whereas tenses in English are linked with time and aspect and can thus be further classified into 14 types (e.g., present simple, present continuous, present perfect, past simple, and so on.).

To answer the second research question, "Does this influence follow the same pattern (either an increase or decrease) as the learners proceed across the university academic levels?", the analysis showed that the role of L1 influence may not necessarily follow a steady change pattern. This was particularly the case in the analysis of verb and noun phrase error categories, where year 2 represented an anomalous stage between year 1 and year 3. In contrast to this, spelling errors did follow a steady downward trend across the three year groups represented in the corpus.

9.2.2 The Potential for Error Counting Approach

In addition to the main aim To offer a new approach to error analysis that looks at both the language learners' errors and what the language learners achieved correctly, three error counting approaches were compared, namely: the *traditional error counting approach* (TEC) that is widely used in the analysis of learner errors, the *potential occasion analysis approach* (POA), developed by Thewissen (2012; 2015), and the *potential for error counting approach* (PFEC) proposed by this thesis. The comparison was conducted on six error categories identified in the LEFLL corpus and sub-corpora (spelling, noun, verb, prepositional, adjective and adverb phrase errors). The aim of this comparison was to evaluate each error counting approach and establish what each approach can tell us. The TEC approach revealed that spelling and noun phrase errors, collectively, represented almost two-thirds of errors in the LEFLL corpus, whereas the other four error categories – verb, prepositional, adjective and adverb phrase errors – constituted the remaining portion of language learners' errors (almost one-third only). Across the three LEFLL sub-corpora, the Year 2 sub-corpus was found to be the anomalous sub-corpus between the Year 1 and Year 3 LEFLL sub-corpora in almost all language learners' error categories. It was observed that the percentages of almost all error categories did not follow steady changes (either through an increase or a decrease in the percentages of each type of error) across the LEFLL sub-corpora. The only steady changes occurred across the sub-corpora was observed in the percentages of prepositional phrase errors; these showed a steady decline across the three LEFLL sub-corpora. The major problem with this approach is that it was not possible to see the percentages of what the language learners achieved accurately in each linguistic category (spelling, noun phrase structures, etc.).

The analysis via the POA approach provided different results. Using this method, spelling and noun phrase errors no longer constituted the highest percentages of errors out of all error categories, as had been observed in the TEC approach. Instead, the verb and prepositional phrase errors scored the highest. Across the three LEFLL sub-corpora, there were steady changes in the percentages of spelling and prepositional phrase errors from one university level to another. They both showed steady decline. The distinctive feature in the POA approach is that it offered us both the percentages of the language learners' errors and what the language learners achieved correctly, the latter perspective being one not offered by the TEC approach. However, the POA approach was also found to have significant shortcomings. Firstly, it is based on the theoretical assumption that each linguistic constituent (e.g., an auxiliary in a verb phrase, a word token in the corpus for spelling errors, etc.) has the potential for error, but fails to consider whether there is any empirical evidence from the corpus to back this assumption up. Secondly, it relies on automatic part-of-speech taggers which have been developed on the basis on a native corpus and whose accuracy is affected by errors in the learner corpora.

It was then argued that the PFEC approach developed for this thesis provides more accurate results and addresses the major problems with the POA approach. This approach requires an initial phase of empirical investigation in which the corpus is manually error tagged and errors are manually retrieved and analysed, in order to establish the actual, rather than the merely theoretical, error potential of any given lexico-grammatical feature. The results were in line with the POA approach with respect to the changes in the percentages of errors across the

LEFLL sub-corpora, with a steady decline being noted in the percentages of both spelling and prepositional phrase errors.

In contrast to the POA approach, the PFEC approach revealed that not all linguistic constituents have the potential for error. For instance, it was observed that spelling errors started at 2-letter words and none of the 1-letter words were misspelled. The analysis also revealed that the LEFLL corpus contains non-alphabetic words such as cardinal numbers (e.g., 1, 4,5, etc.) and dates (e.g., 1991, 2011, etc.) which are impossible to misspell but which nevertheless affect the accuracy of the percentages of spelling errors as they are added to the total word count of the LEFLL corpus and the sub-corpora. Therefore, for more accurate results they were excluded from the total word count before calculating the percentages of spelling errors. Following the retrieval and removal of 1-letter words and non-alphabetic words from the total word count of the LEFLL corpus and sub-corpora, it was argued that we had obtained more accurate results. The percentages of spelling errors in the LEFLL corpus and in each sub-corpus was increased in different rates. For instance, in the Year 1 sub-corpus, the percentage of spelling errors increased from 5.90% (via the POA approach) to 6.72% (via the PFEC approach) (up to 0.82%); whereas in the Year 3 sub-corpus the variation between the PFEC approach and the POA approach was only +0.26%. This variation (0.82% vs 0.26) was due to the fact that the total number of 1-letter words and non-alphabetic words in the Year 1 sub-corpus (3142) was far higher than the total number of these words in the Year 3 sub-corpus (795).

The analysis, via the PFEC approach, also revealed that despite their large percentages out of the total word count of the LEFLL corpus and sub-corpora, few instances of the 2-letter words were misspelled. For instance, there were only 8 misspelled words of the 2-letter words in the Year 1 sub-corpus which (the 2-letter words) constituted 22.72% out of the total number of tokens in the Year 1 sub-corpus. The analysis showed that the 2-letter misspelled words were simple, very common, and widely used by both language learners and native speakers, e.g., *of, so, us, etc*. Therefore, it is more likely that they were accidently misspelled; certainly, it is rare to observe such types of misspelled words in an advanced learner corpus. Therefore, the PFEC approach calls for conducting a practical investigation to verify which word length, indeed, has the potential for spelling errors. This should also be applied on other types of linguistic constituents (e.g., what type of determiners have the empirically demonstrable potential for errors?) to ensure that we only focus on the linguistic constituents that meaningfully have the potential for error.

9.3 Limitations of the Study

It is important to acknowledge here that although this thesis has successfully achieved the two aims outlined in the introductory chapter, it inevitably has certain limitations. Firstly, it is unfortunate that the thesis did not include participants from at least another Libyan university from a different part of the country. The data in the LEFLL corpus was collected from the university-level Libyan Arab English learners at Benghazi University in Benghazi city only. Collecting data from another Libyan university from a different part of the country (e.g., from Tripoli University in Tripoli city in the North West of Libya) may make it possible to have a comparable group of Libyan English learners and verify whether the results obtained from the Libyan undergraduates' English learners at Benghazi University can be representative to all Libyan undergraduates' English learners who study English as a foreign language. Secondly, the LEFLL corpus size is relatively small at only 60,131 tokens. For the representativeness of the data and generalizability of the results, a larger corpus would obviously have been preferable

(Granger, 2004). Thirdly, collecting the data for the LEFLL corpus and error tagging were performed manually. Not only were these arduous and time-consuming processes, but the fact that they were carried out manually means that they may have been liable to accidental human errors. It is, however, submitted that the IRR test reported in Chapter 4 indicates that this was not a major issue for the current study.

9.4 Suggestions for Future Research

The current thesis has only focused on Libyan English learners at Benghazi University to achieve the two aims set out in the introductory chapter. Therefore, it is not certain whether the results obtained in this thesis are applicable to Libyan English learners at the other Libyan universities, still less to Arabic L1 learners in other Arabic-speaking countries. Therefore, for the generalizability of the results and representativeness of the data, it is recommended to collect data from two or more Libyan universities from different parts of the country, and also to further extend the methods showcased in this thesis to learners in other Arabic-speaking countries. It is also recommended that future studies should carry out a comprehensive and quasi-longitudinal analysis on the other English error categories (prepositional, adjective and adverb phrase errors). This may help to study the role of the learners' L1 on the other error categories and how it develops across different proficiency levels.

Bibliography

Ababneh, I. (2019) Errors in Arabic-English Translation Among Saudi Students: Comparative Study between Two Groups of Students. *AWEJ for Translation & Literary Studies, 3*(4), pp. 118 - 129.

Abe, M. (2003) A Corpus-based Contrastive Analysis of Spoken and Written Learner Corpora: The Case of Japanese-speaking Learners of English. in *Proceedings of the 2003 Corpus Linguistics Conference*,Lancaster. pp. 1-9.

AbiSamra, N. (2003) *An analysis of errors in Arabic speakers' English writings*. Unpublished doctoral dissertation, American University of Beirut.

Abu-Rabia, S. and Sammour, R. (2013) Spelling Errors' Analysis of Regular and Dyslexic Bilingual Arabic-English Students. *Open Journal of Modern Linguistics, 3*(1), pp. 58 - 68.

Adila, W. (2019) A written grammatical error analysis of second year students of Arabic. *Journal of Arabic Studies, 4*(1), pp. 31-44.

Al-Jarf, R. S. (2005) The effects of listening comprehension and decoding skills on spelling achievement of EFL freshman students. *English Language & Literature Teaching, 11*(2), pp. 35-50.

Al-Oudat, A. (2017) Spelling errors in English writing committed by English-major students at BAU. *Journal of Literature, Languages and Linguistics, 17*, pp. 43-47.

Al-Shujairi, Y. B. J. and Tan, H. (2017) Grammar errors in the writing of Iraqi English language learners. *International Journal of Education and Literacy Studies, 5*(4), pp. 122 - 130.

Al-Sobhi, B. M. S., Rashid, S. M., Abdullah, A. N. and Darmi, R. (2017) Arab ESL secondary school students' spelling errors. *International Journal of Education and Literacy Studies, 5*(3), pp. 16-23.

Al Jayousi, M. T. (2011) *Spelling errors of Arab students: Types, causes and teachers' responses*. Unpublished Master of Arts Dissertation, The American University of Sharjah, UAE.

Alamin, A. and Ahmed, S. (2012) Syntactical and punctuation errors: An analysis of technical writing of university students Science College, Taif University, KSA. *English Language Teaching, 5*(5), pp. p2.

Alasfour, A. S. (2018) *Grammatical errors by Arabic ESL students: An investigation of L1 transfer through error analysis*. Unpublished MA dissertation, Portland State University.

Alhaisoni, E. M., Al-Zuoud, K. M. and Gaudel, D. R. (2015) Analysis of spelling errors of beginner learners of English in the English foreign language context in Saudi Arabia. *English Language Teaching, 8*(3).

Alhaysony, M. (2012) An analysis of article errors among Saudi female EFL students: A case study. *Asian Social Science, 8*(12), pp. 55 - 66.

Allaith, Z. A. and Joshi, R. M. (2011) Spelling performance of English consonants among students whose first language is Arabic. *Reading and Writing, 24*(9), pp. 1089 - 1110.

Alotaibi, A. M. (2016) Examining the learnability of English relative clauses: evidence from Kuwaiti EFL learners. *English Language Teaching, 9*(2), pp. 57 - 65.

Alsaawi, A. (2015) Spelling errors made by Arab learners of English. *International Journal of Linguistics, 7*(5), pp. 55-67.

Alshayban, A. S. (2012) *Copula omission by EFL Arab learners*. Unpublished PhD thesis, Colorado State University.

Altamimi, D. A. H. F., Ab Rashid, R. and Elhassan, Y. M. M. (2018) A review of spelling errors in Arabic and non-Arabic contexts. *English Language Teaching, 11*(10), pp. 88-94.

AMARA, N. (2018) Correcting students' errors: theory and practice. *Current Educational Research, 1*(05), pp. 45 - 57.

Aston, G. (2011) Applied corpus linguistics and the learning experience. in Viana, V., Zyngier, S. and Barnbrook, G., (eds.) *Perspectives on corpus linguistics*,Amsterdam and Philadephia: John Benjamins. pp. 1-16.

Bahr, R. H., Silliman, E. R., Berninger, V. W. and Dow, M. (2012) Linguistic pattern analysis of misspellings of typically developing writers in Grades 1–9. *Journal of Speech, Language, and Hearing Research, 55*(6), pp. 1587-1599.

Bahr, R. H., Sillimana, E. R., Danzak, R. L. and Wilkinson, L. C. (2015) Bilingual spelling patterns in middle school: it is more than transfer. *International Journal of Bilingual Education and Bilingualism, 18*(1), pp. 73-91.

Barry, D. (2014) *The impact of native Arabic on English writing as a second language,* Michigan: Debbie Barry.

Bataineh, R. F. (2005) Jordanian undergraduate EFL students’ errors in the use of indefinite article *The Asian EFL Journal Quarterly, 7*(1), pp. 56-76.

Benson, C. (2002) Transfer/Cross-linguistic influence. *ELT Journal, 56*(1), pp. 68-70.

Berninger, V. W., Abbott, R. D., Abbott, S. P., Graham, S. and Richards, T. (2002) Writing and reading: Connections between language by hand and language by eye. *Journal of Learning Disabilities, 35*(1), pp. 39-56.

Bestgen, Y. and Granger, S. (2011) Categorising spelling errors to assess L2 writing. *International Journal of Continuing Engineering Education and Life Long Learning, 21*(2-3), pp. 235-252.

Bestgen, Y., Granger, S. and Thewissen, J. (2012) Error Patterns and Automatic L1 Identification. in Jarvis, S. and Crossley, S. A., (eds.) *Approaching language transfer through text classification*,Bristol: Multilingual Matters. pp. 127-153.

Bley-Vroman, R. J. L. l. (1983) The comparative fallacy in interlanguage studies: The case of systematicity 1. *Language Learning, 33*(1), pp. 1-17.

Boulton, A. (2007) DDL is in the details... and in the big themes. in *Corpus Linguistics Conference: CL2007*. pp. XX.

Boulton, A. (2008) Looking (for) empirical evidence of data-driven learning at lower levels. in: Frankfurt: Peter Lang. Lodz Studies in Language.

Boulton, A. (2009) Testing the limits of data-driven learning: language proficiency and training. *Recall, 21*, pp. 37-54.

Brennan, R. L. and Prediger, D. J. (1981) Coefficient kappa: Some uses, misuses, and alternatives. *Educational and psychological measurement, 41*(3), pp. 687-699.

Brown, R. (1973) *A first language: The early stages,* Harmondsworth: Penguin.

Brutt-Griffler, J. and Samimy, K. K. (2001) Transcending the nativeness paradigm. *World Englishes, 20*(1), pp. 99-106.

Btoosh, M. A. (2019) Tense and aspect in the academic writing of Arab L2 learners of English: A corpus-based approach. *Journal of Language and Education, 5*(2), pp. 26-47.

Butterworth, G. (1978) A Spanish-speaking adolescent's acquisition of English syntax. in Hatch, E., (ed.) *Second Language Acquisition: A Book of Readings*,Rowley, Mass: Newbury House.

Byrt, T. (1996) How good is that agreement? *Epidemiology, 7*(5), pp. 561.

Carletta, J. (1996) Assessing agreement on classification tasks: The kappa statistic. *Computational linguistics, 22*(2), pp. 249-254.

Carney, E. (1994) *A survey of English spelling,* London: Routledge.

Centre for English Corpus Linguistics (2019) *Learner corpora around the world*, Available: https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html [Accessed 05/12/2019.

Chaaraoui, J. (2017) Grammar accuracy, language threshold level, and degree of bilingualism in the Saudi EFL learner's interlanguage. *International Journal of Language and Linguistics, 4*(3), pp. 86-101.

Chiappe, P., Glaeser, B. and Ferko, D. (2007) Speech perception, vocabulary, and the development of reading skills in english among Korean-and English-speaking children. *Journal of Educational Psychology, 99*(1), pp. 154.

Cohen, J. (1960) A coefficient of agreement for nominal scales[1]. *Educational and psychological measurement, 20*(1), pp. 37-46.

Cook, V. J. (1997) L2 Users and English Spelling. *Journal of Multilingual and Multicultural Development, 18*(6), pp. 474-488.

Corder, S. P. (1967) The significance of learner's errors. *IRAL-International Review of Applied Linguistics in Language Teaching, 5*(1-4), pp. 161-170.

Corder, S. P. (1971) Idiosyncratic dialects and error analysis. *International Review of Applied Linguistics in Language Teaching, 9*(2), pp. 147 - 160.

Corder, S. P. (1974) Error Analysis. in Allen, J. P. B. and Corder, S. P., (eds.) *Techniques in Applied Linguistics. The Edinburgh Course in Applied Linguistics*,Oxford: Oxford University Press. pp. 122-154.

Corder, S. P. (1981) *Error analysis and interlanguage,* Oxford: Oxford University Press.

Crompton, P. (2011) Article errors in the English writing of advanced L1 Arabic learners: The role of transfer. *Asian EFL Journal, 50*, pp. 4-34.

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

Darus, S. and Ching, K. H. (2009) Common errors in written English essays of form one Chinese students: A case study. *European Journal of social sciences, 10*(2), pp. 242-253.

De Cock, S., Granger, S., Leech, G. and McEnery, T. (1998) An automated approach to the phrasicon of EFL learners. in Granger, S., (ed.) *Learner English on computer*,New York: Addison Wesley Longman. pp. 67-79.

Demirel, E. T. (2017) Detection of common errors in Turkish EFL students' writing through a corpus analytic approach. *English Language Teaching, 10*(10), pp. 159-178.

Diab, N. (1996) The transfer of arabic in the english writings of lebanese students*. *The ESP, 8*(1), pp. 71-83.

Dulay, H. and Burt, M. (1975) Creative construction in second language learning and teaching. in Burt, M. and Dulay, H., (eds.) *On TESOL '75: New Directions in Second Language Learning, Teaching and Bilingual Education*,Washington, D.C.: TESOL. pp. 21-32.

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

Dulay, H. C. and Burt, M. K. (1973) Should we teach children syntax? *Language Learning, 23*(2), pp. 245-258.

Dulay, H. C. and Burt, M. K. (1974a) Natural sequences in child second language acquisition 1. *Language Learning, 24*(1), pp. 37-53.

Dulay, H. C. and Burt, M. K. (1974b) A new perspective on the creative construction process in child second language acquisition 1. *Language Learning, 24*(2), pp. 253-278.

Dweik, B. S. and Othman, Z. A. (2017) Lexical and grammatical interference in the translation of written texts from Arabic into English. *Academic Research International, 8*(3), pp. 65-70.

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

Ellis, R. (2008) *The Study of Second Language Acquisition,* Oxford: Oxford University Press.

Fender, M. (2008) Spelling knowledge and reading development: Insights from Arab ESL learners. *Reading in a Foreign Language, 20*(1), pp. 19-42.

Figueredo, L. (2006) Using the known to chart the unknown: A review of first-language influence on the development of English-as-a-second-language spelling skill. *Reading and Writing, 19*(8), pp. 873-905.

Fitria, T. N. (2020) Spelling error analysis in students' writing English composition. *Getsempena English Education Journal, 7*(2), pp. 240-254.

Fleiss, J. L. (1971) Measuring nominal scale agreement among many raters. *Psychological bulletin, 76*(5), pp. 378-382.

Fries, C. C. (1945) *Teaching and learning English as a foreign language,* University of Michigan: University of Michigan Press.

Gafu, C., Badea, M. and Iridon, C. (2012) Errors in the acquisition of Romanian as second language: A case study. *Procedia - Social and Behavioral Sciences, 69*, pp. 1626-1634.

Garretson, G. (2005) *Dexter: Tools for Analyzing Language Data*, Available: http://www.dextercoder.org/about.html [Accessed 20/06/2012.

Garretson, G. (2006) Dexter: Free tools for analyzing texts. in *Actas de V Congreso Internacional AELFE*: Citeseer. pp. 659-665.

Granger, S. (1996) From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. in Altenberg, B. and Johan, M., (eds.) *Languages in Contrast. Text-based cross-linguistic studies*,Lund: Lund University Press. pp. 37-51.

Granger, S. (1998) The computer learner corpus: A versatile new source of data for SLA research. in Granger, S., (ed.) *Learner English on computer*,New York: Addison Wesley Longman. pp. 3-18.

Granger, S. (1999) Uses of tenses by advanced EFL learners: evidence from an error-tagged computer corpus. *LANGUAGE AND COMPUTERS, 26*, pp. 191-202.

Granger, S. (2002) A Bird's-eye view of learner corpus research. in Granger, S., Hung, J. and Petch_Tyson, S., (eds.) *Computer learner corpora, second language acquisition and foreign language teaching*,Amsterdam / Philadelphia: John Benjamins Publishing Company. pp. 3-33.

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

Granger, S. (2003b) The International Corpus of Learner English: A new resourse for foreign language learning and teaching and second language acquisition research. *Tesol Quarterly, 37*(3), pp. 538-546.

Granger, S. (2004) Computer learner corpus research: current status and future prospects. in Connor, U. and Upton, T. A., (eds.) *Applied corpus linguistics: a multidimensional perspective*,Amsterdam: Editions Rodopi. pp. 123-145.

Granger, S. (2012) How to use foreign and second language learner corpora? *Research methods in Second Language Acquisition: A practical guide*, pp. 7-29.

Granger, S. (2015) Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research, 1*(1), pp. 7-24.

Granger, S. and Bestgen, Y. (2014) The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *IRAL-International Review of Applied Linguistics in Language Teaching, 52*(3), pp. 229-252.

Granger, S., Gilquin, G. and Meunier, F. (2015) Introduction: Learner corpus research–past, present and future. in Granger, S., Gilquin, G. and Meunier, F., (eds.) *The Cambridge Handbook of Learner Corpus Research*,Cambridge: Cambridge University Press. pp. 1-6.

Granger, S., Kraif, O., Ponton, C., Antoniadis, G. and Zampa, V. (2007) Integrating learner corpora and natural language processing: A crucial step towards reconciling technological sophistication and pedagogical effectiveness. *RECALL-HULL THEN CAMBRIDGE-, 19*(03), pp. 252-268.

Haggan, M. (1991) Spelling errors in native Arabic-speaking English majors: A comparison between remedial students and fourth year students. *System, 19*(1), pp. 45-61.

Hameed, P. F. M. (2016) A study of the spelling errors committed by students of English in Saudi Arabia: Exploration and remedial measures. *Advances in Language and Literary Studies, 7*(1), pp. 203-207.

Hammarberg, B. (1974) The insufficiency of error analysis. *IRAL: International Review of Applied Linguistics in Language Teaching, 12*(3), pp. 185-192.

Hasselgren, A. (1994) Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International journal of applied linguistics, 4*(2), pp. 237-258.

Hayes-Harb, R. (2006) Native Speakers of Arabic and ESL Texts Evidence for the Transfer of Written Word Identification Processes. *Tesol Quarterly, 40*(2), pp. 321-339.

Haywood, J. A. and Nahmad, H. M. (1962) *A new Arabic grammar of the written language,* Lund, Humphries.

Hoang, T.-Q.-H. (1965) *A phonological contrastive study of Vietnamese and English*. Unpublished Master of Art Dissertation, Texas Tech University.

Hourani, T. M. Y. (2008) *An analysis of the common grammatical errors in the English writing made by 3rd secondary male students in the Eastern Coast of the UAE*. Unpublished Master of Education Dissertation, the British University in Dubai.

Ibrahim, A. (2018) The first language influence on the EFL-learners' writing performance errors analysis and remedial perspective. *Journal of Education and Practice, 9*(14), pp. 152-165.

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

Johns, T. (1991) Should you be persuaded: Two samples of data-driven learning materials. *ELR Journal, 4*, pp. 1-16.

Kellerman, E. (1995) Crosslinguistic influence: Transfer to nowhere? *Annual review of applied linguistics, 15*, pp. 125-150.

Khan, M. (2019) Linguistic contrastive analysis of English and Arabic from a morphological and syntactical perspective. *International Journal of English and Education, 8*(1), pp. 257-271.

Khatter, S. (2019) An analysis of the most common essay writing errors among EFL Saudi female learners (Majmaah University). *Arab Society of English Language Studies, 10*, pp. 364-381.

Khuwaileh, A. A. and Shoumali, A. A. (2000) Writing Errors: A Study of the Writing Ability of Arab Learners of Academic English and Arabic at University. *Language, Culture and Curriculum, 13*(2), pp. 174-183.

Kossmann, M. G. and Stroomer, H. (1997) Berber phonology. *Phonologies of Asia and Africa, 1*, pp. 461-475.

Lado, R. (1957) *Linguistics across cultures: Applied linguistics for language teachers,* Ann Arbor: University of Michigan Press.

Larsen-Freeman, D. (2014) Another step to be taken - Rethinking the endpoint of the interlanguage continuum. in Han, Z. and Tarone, E., (eds.) *Interlanguage : Forty years later*,Amsterdam: John Benjamins Publishing Company. pp. 203-220.

Lee, D. Y. W. and Chen, S. X. (2009) Making a bigger deal of the smaller words: Function words and other key items in research writing by Chinese learners. *Journal of Second Language Writing, 18*(4), pp. 281-296.

Lee, W. R. (1968) Thoughts on contrastive linguistics in the context of language teaching. in Alatis, J. E., (ed.) *Mongraph Series on Languages and Linguistics*,Washington, D.C.: Georgetown University Press. pp. 185-194.

Lennon, P. (1991) Error: Some problems of definition, identification, and distinction. *Applied Linguistics, 12*(2), pp. 180-196.

Lindner, A. L., Wijekumar, K. and Joshi, R. M. (2020) English spelling performance in writing samples among Spanish-speaking ELLs. *Journal of Learning Disabilities*, pp. 1-9.

LoCoco, V. G.-M. (1976) A Comparison of Three Methods for the Collection of L2 Data: Free Composition, Translation, and Picture Description. *Working Papers on Bilingualism, 8*, pp. 59-86.

López, W. C. (2009) Error analysis in a learner corpus. What are the learners' strategies? In Paper Presented to the A survey of corpus-based research.

Mace, J. (1998) *Arabic grammar: A reference guide,* Edinburgh Edinburgh University Press.

Mariko, A. (2007) Grammatical errors across proficiency levels in L2 spoken and written English. *The Economic Journal of Takasaki City University of Economics, 49*(3), pp. 117-129.

Markham, P. (1985) Contrastive analysis and the future of second language education. *System, 13*(1), pp. 25-29.

Martin, K. I. (2017) The impact of L1 writing system on ESL knowledge of vowel and consonant spellings. *Reading and Writing, 30*(2), pp. 279-298.

Masruddin, M. (2019) Omission: common simple present tense errors in students' writing of descriptive text. *Ethical Lingua: Journal of Language Teaching and Literature, 6*(1), pp. 30-39.

Mehta, P. D., Foorman, B. R., Branum-Martin, L. and Taylor, W. P. (2005) Literacy as a unidimensional multilevel construct: Validation, sources of influence, and implications in a longitudinal study in grades 1 to 4. *Scientific Studies of Reading, 9*(2), pp. 85-116.

Merizawati, H. (2019) An error analysis in paper presentations: A case study of Indonesian EFL learners. *Linguists: Journal Of Linguistics and Language Teaching, 4*(2), pp. 71-103.

Mohamed-Sayidina, A. (2010) Transfer of L1 cohesive devices and transition words into L2 academic texts: The case of Arab students. *RELC Journal, 41*(3), pp. 253-266.

Mohammad, A.-K. and Shwater, B. (2018) The effect of 1st language on 2nd language acquisition: The acquisition of English preposition by Arabic native speakers. *International Journal of Literature, Language and Linguistics, 1*(1), pp. 34-45.

Mohammed, M. S. and Abdalhussein, H. F. (2015) Grammatical error analysis of Iraqi postgraduate students' academic writing: The case of Iraqi students in UKM. *International Journal of Education and Research, 3*(6), pp. 283-294.

Möller, V. (2013) How do educational settings at the secondary level impact on learners' use of the English passive?–Evidence from the secondary-level corpus of learner English (SCooLE). in *Learner Corpus Research Conference, Bergen, Norway*. pp. 27-29.

Morgan, G. (2018) The ineffectiveness of overt input on the problematic grammatical features of tense usage and verb conjugation for native Arabic speaking learners of English for academic purposes (EAP). *Advances in Language and Literary Studies, 9*(4), pp. 193-205.

Mourtaga, K. R. (2004) *Investigating writing problems among Palestinian students studying English as a foreign language*. Unpublished doctoral dissertation, The University of Mississippi.

Naba'h, A. A. (2011) Lexical errors made by in-service English language teachers in jordan. *Damascus University Journal, 27*(1), pp. 49-75.

Nagata, R., Whittaker, E. and Sheinman, V. (2011) Creating a manually error-tagged and shallow-parsed learner corpus. in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*,Portland, Oregon, USA: Association for Computational Linguistics. pp. 1210-1219.

Nuruzzaman, M., Islam, S. A. B. M. and Jahan Shuchi, I. (2018) An Analysis of Errors Committed by Saudi Non-English Major Students in the English Paragraph Writing: A Study of Comparisons. *Advances in Language and Literary Studies, 9*(1), pp. 31-39.

Odlin, T. (1989) *Language transfer: Cross-linguistic influence in language learning,* Cambridge: Cambridge University Press.

Oishimaya, S. N. (2017) *What languages are spoken in Libya?*, Available: https://www.worldatlas.com/articles/what-languages-are-spoken-in-libya.html [Accessed 08/08/2020.

Omar, J. A. (2019) Kurdish EFL learners' spelling error types and sources. *6*(2), pp. 111-120.

Othman, A. K. A. (2018) An investigation of the most common spelling errors in English writing committed by English-major male students at the University of Tabuk. *Journal of Education and Practice, 9*(1), pp. 17-22.

Pacton, S., Borchardt, G., Treiman, R., Lété, B. and Fayol, M. (2013) Learning to spell from reading: General knowledge about spelling patterns influences memory for specific words. *The Quarterly Journal of Experimental Psychology, 67*(5), pp. 1019–1036.

Peck, S. (1978) Child-child discourse in second language acquisition. in Hatch, E., (ed.) *Second Language Acquisition: A Book of Readings*,Rowley, Mass: Newbury House. pp. 383-400.

Péry-Woodley, M.-P. (1990) Contrasting discourses: contrastive analysis and a discourse approach to writing. *Language Teaching, 23*(03), pp. 143-151.

Protopapas, A., Fakou, A., Drakopoulou, S., Skaloumbakas, C. and Mouzaki, A. (2013) What do spelling errors tell us? Classification and analysis of errors made by Greek schoolchildren with and without dyslexia. *Reading and Writing, 26*(5), pp. 615-646.

Qaddumi, H. A. and Walweel, T. A. (2018) Analysis of English writing errors committed by students learning English as a Foreign Language at Al-Istiqlal University in Palestine. In Paper Presented to the The Ninth International Scientific Academic Conference Under the Title "Contemporary trends in social, human, and natural sciences".

Qomariana, Y., Puspani, I. A. M. and Rahayuni, N. K. S. (2019) Mother tongue interference on EFL The case of English department students in Udayana university. *Proceeding of the 65th TEFLIN International Conference, 65*(1), pp. 250-254.

Randall, M. and Groom, N. (2009) The BUiD Arab learner corpus: a resource for studying the acquisition of L2 english spelling. in *Proceedings of the Corpus Linguistics Conference (CL), Liverpool, UK*.

Randolph, J. J. (2005) Free-marginal multirater Kappa (multirater K [free]): An alternative to Fleiss' fixed-marginal multirater Kappa. in *Joensuu Learning and Instruction Symposium*.

Rayson, P. and Garside, R. (1998) The claws web tagger. *ICAME journal, 22*, pp. 121-123.

Richards, J. C. (1974) A non-contrastive approach to error analysis. in Richards, J. C., (ed.) *Error Analysis: Perspectives on Second Language Acquisition*,London: Longman. pp. 172-188.

Richards, J. C. and Schmidt, R. W. (2010) *Longman dictionary of language teaching and applied linguistics,* 4th ed ed*.,* Harlow: Pearson Education Limited

Ridha, N. S. A. (2012) The effect of EFL learners' mother tongue on their writings in English: An error analysis study. *ADAB AL-BASRAH,* (60), pp. 22-45.

Rimrott, A. and Heift, T. (2005) Language learners and generic spell checkers in CALL. *CALICO Journal, 23*(1), pp. 17-48.

Rimrott, A. and Heift, T. (2008) Evaluating automatic detection of misspellings in German. *Language Learning & Technology, 12*(3), pp. 73-92.

Ringbom, H. (1998) Vocabulary frequencies in advanced learner English: A cross-linguistic approach. in Granger, S., (ed.) *Learner English on computer*,London: Longman. pp. 41-52.

Robert, L. P., Ramirez, A., G and Ramirez (1973) An error analysis of the spoken English of Mexican-American pupils in a bilingual school and a monolingual school. *Language Learning, 23*(1), pp. 39-62.

Romaine, S. (2003) Variation. in Doughty, C. J. and Long, M. H., (eds.) *The Handbook of Second Language Acquisition*,Malden, MA: Blackwell Publishing Limited. pp. 409–435.

Ryan, A. and Meara, P. (1991) The case of the invisible vowels: Arabic speakers reading English words. *Reading in a Foreign Language, 7*, pp. 531-540.

Sabbah, S. S. (2015) Negative transfer: Arabic language interference to learning English. *Arab World English Journal (AWEJ) Special Issue on Translation,* (4), pp. 269-288.

Sadhwani, P. (2005) *Phonological and orthographic knowledge: An Arab-Emirati perspective*. Unpublished Doctoral Dissertation, The British university in Dubai (BUiD).

Saigh, K. and Schmitt, N. (2012) Difficulties with vocabulary word form: The case of Arabic ESL learners. *System, 40*(1), pp. 24-36.

Sawalmeh, M. H. M. (2013) Error analysis of written English essays: The case of students of the preparatory year program in Saudi Arabia. *English for Specific Purposes World, 40*(14), pp. 1-17.

Schumann, J. H. (1978) *The pidginization process: A model for second language acquisition,* Rowley, Massachusetts: Newbury House

Scott, M. S. and Tucker, G. R. (1974) Error analysis and English-language strategies of Arab students1. *Language Learning, 24*(1), pp. 69-97.

Selinker, L. (1972) Interlanguage *IRAL: International Review of Applied Linguistics in Language Teaching, 10*, pp. 209-231.

Shapira, R. (1978) The non-learning of English: Case study of an adult. in Hatch, E., (ed.) *Second Language Acquisition: A Book of Readings*,Rowley, Mass: Newbury House. pp. 246-255.

Silliman, E. R., Bahr, R. H. and Peters, M. L. (2006) Spelling patterns in preadolescents with atypical language skills: Phonological, morphological, and orthographic factors. *Developmental Neuropsychology, 29*(1), pp. 93-123.

Sim, J. and Wright, C. C. (2005) The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy, 85*(3), pp. 257-268.

Simon, P. (2006) Including omission mistakes in the calculation of Cohen's Kappa and an analysis of the Coefficient's paradox features. *Educational and psychological measurement, 66*(5), pp. 765-777.

Slimane, F., Kanoun, S., Hennebert, J., Alimi, A. M. and Ingold, R. (2013) A study on font-family and font-size recognition applied to Arabic word images at ultra-low resolution. *Pattern Recognition Letters, 34*(2), pp. 209-218.

Stewart, S. and Cegelka, P. (1995) Teaching reading and spelling. in Cegelka and Berdine, W. H., (eds.) *Effective instruction for students with learning difficulties. Boston: Allyn and Bacon*,Boston: Allyn and Bacon.

Stockwell, R. P., Bowen, J. D. and Martin, J. W. (1965) *The grammatical structures of English and Spanish,* Chicago: University of Chicago Press.

Swan, M. and Smith, B. (2001) *Learner English: A teacher's guide to interference and other problems,* Cambridge, United Kingdom: Cambridge University Press.

Thewissen, J. (2008) The phraseological errors of French-, German-, and Spanish speaking EFL learners: Evidence from an error-tagged learner corpus. in *Proceedings from the 8th Teaching and Language Corpora Conference.* pp. 300-306.

Thewissen, J. (2012) *Accuracy across proficiency Levels: Insights from an error-tagged EFL learner corpus*. Unpublished PhD thesis, UCL-Université Catholique de Louvain.

Thewissen, J. (2013) Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *The Modern Language Journal, 97*(S1), pp. 77-101.

Thyab, I. R. A. (2019) Most Common Grammatical Mistakes Made in English by L1 Arabic Learners. *Journal of Language Studies, 2*(4), pp. 49-62.

Tono, Y. (2000) A computer learner corpus based analysis of the acquisition order of English grammatical morphemes. in Burnard, L. and McEnery, T., (eds.) *Rethinking Language Pedagogy from a Corpus Perspective*,Frankfurt: Peter Lang Gmbh. pp. 123-132.

Tono, Y., Satake, Y. and Miura, A. (2014) The effects of using corpora on revision tasks in L2 writing with coded error feedback. *Recall, 26*(2), pp. 147-162.

Viera, A. J. and Garrett, J. M. (2005) Understanding interobserver agreement: The kappa statistic. *Family Medicine, 37*(5), pp. 360-363.

Warrens, M. J. (2010) Inequalities between Multi-rater Kappas. *Advances in data analysis and classification, 4*(4), pp. 271-286.

Weber, P., Kozel, N., Purgstaller, C., Kargl, R., Schwab, D. and Fink, A. (2013) First and second language in the brain: Neuronal correlates of language processing and spelling strategies. *Brain and language, 124*(1), pp. 22-33.

Westwood, P. (2004) *Spelling: Approaches to teaching and assessment,* London: David Fulton.

Wickens, G. M. (1980) *Arabic grammar: A first workbook,* Cambridge, UK: Cambridge University Press.

Wood, C. and Connelly, V. (2009) *Contemporary perspectives on reading and spelling,* Abingdon, UK: Routledge.

Xu, Y. (2014) Evidence of the accessibility hierarchy in relative clauses in chinese as a second language. *Language and Linguistics, 15*(3), pp. 435-464.

Yang, H.-J., Park, M.-j., Youn, S. Y., Yoo, S., Min, T. K., Jeon, Y. H., Lee, H. W., Lee, J. S. and Pyun, B. Y. (2014) Agreement between the Skin Prick Test and Specific Serum IgE for egg white and cow's milk allergens in young infant with atopic dermatitis. *Allergology International, 63*(2), pp. 235-242.

Yildiz, Y. (2017) Explaining the orthography–phonology interface in written corpora: an Optimality–Theoretic approach. *Corpora, 12*(2), pp. 181-205.

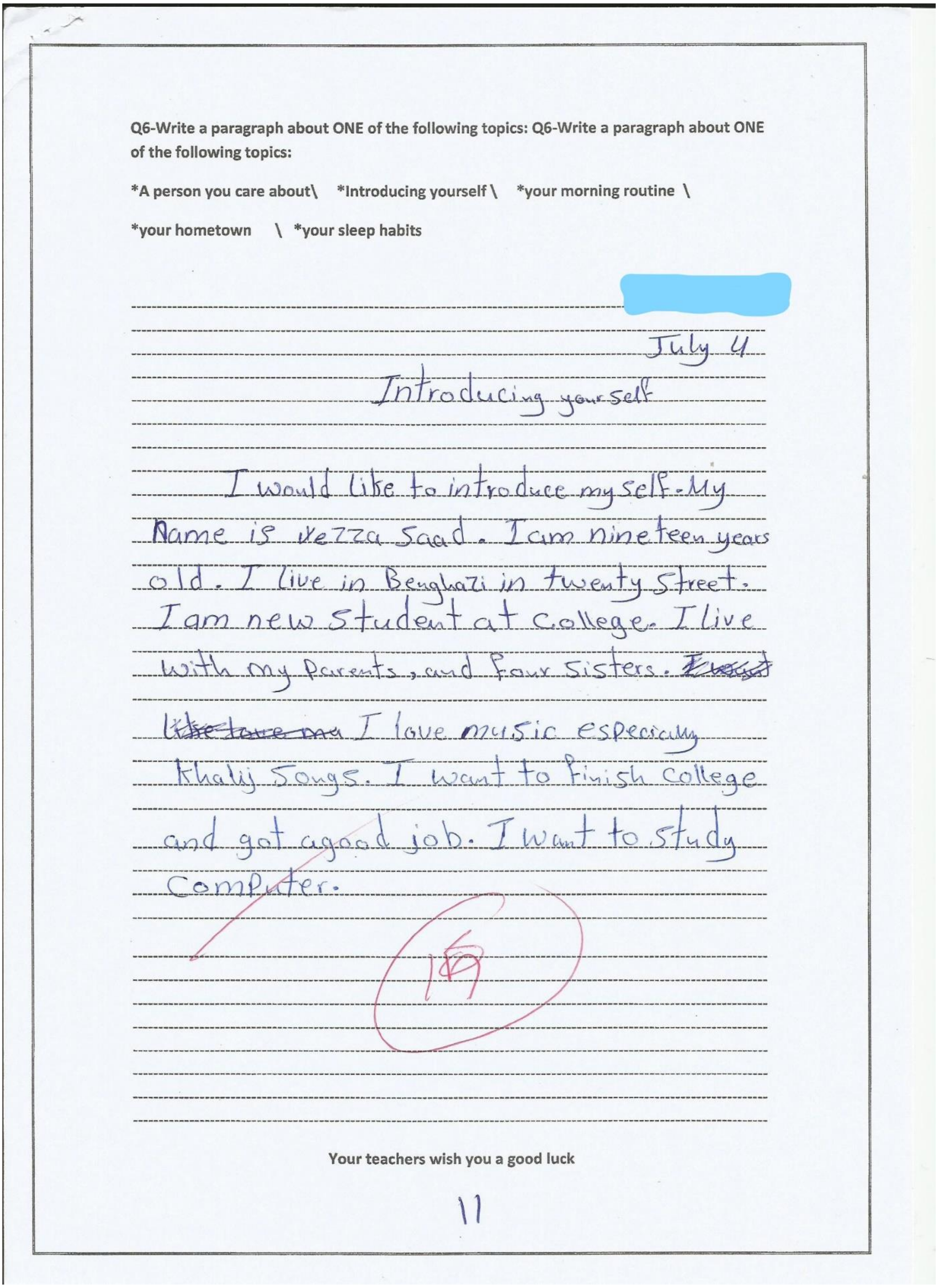
Younes, Z. B. and Albalawi, F. S. (2015) Exploring the most common types of writing problems among English language and translation major sophomore female students at Tabuk University. *Asian Journal of Basic and Applied Sciences, 3*(2), pp. 7-26.

Zhao, J., Quiroz, B., Dixon, L. Q. and Joshi, R. M. (2016) Comparing bilingual to monolingual learners on English spelling: A meta-analytic review. *Dyslexia, 22*(3), pp. 193-213.

Zinsmeister, H. and Breckle, M. (2012) The ALeSKo learner corpus. *Multilingual corpora and multilingual corpus analysis, 14*, pp. 71-96.

Zughoul, M. R. (2002) Interlanguage syntax of Arabic-speaking learners of English: The noun phrase. *Educational Resources Information Center (ERIC)*, pp. 1-24.

Appendix A: Year 1 Final Writing Exam



Q6-Write a paragraph about ONE of the following topics: Q6-Write a paragraph about ONE of the following topics:

*A person you care about\ *Introducing yourself \ *your morning routine \

*your hometown \ *your sleep habits

[redacted]

July 4

Introducing yourself

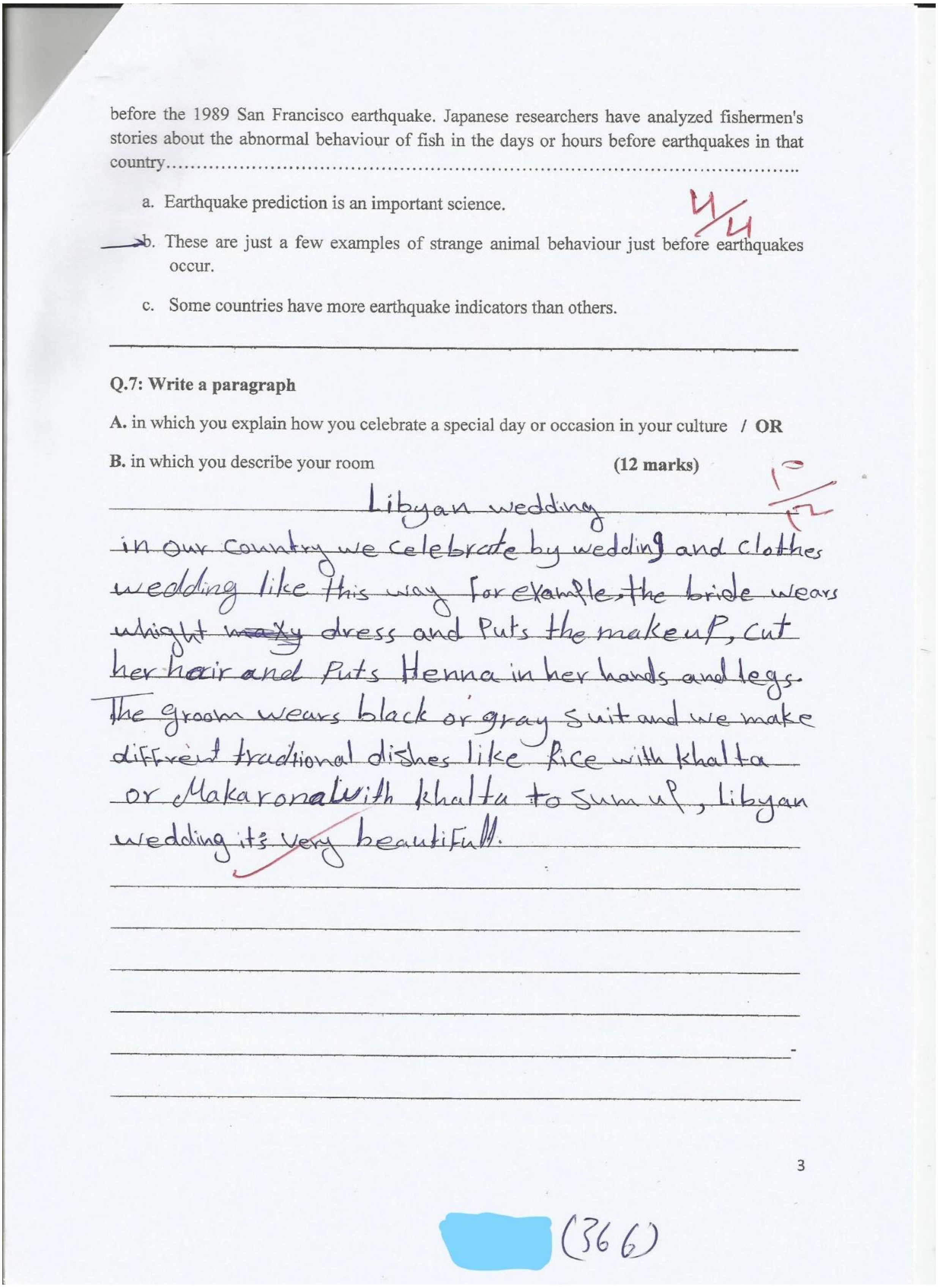
I would like to introduce my self. My Name is Nezza Saad. I am nineteen years old. I live in Benghazi in twenty street. I am new student at college. I live with my Parents, and four sisters. ~~I want~~ ~~like love mu~~ I love music especially khalij songs. I want to finish college and got agood job. I want to study computer.

14

Your teachers wish you a good luck

11

Appendix B: Year 2 Final Writing Exam



before the 1989 San Francisco earthquake. Japanese researchers have analyzed fishermen's stories about the abnormal behaviour of fish in the days or hours before earthquakes in that country………………………………………………………………………………………

a. Earthquake prediction is an important science.

4/4

b. These are just a few examples of strange animal behaviour just before earthquakes occur.

c. Some countries have more earthquake indicators than others.

Q.7: Write a paragraph

A. in which you explain how you celebrate a special day or occasion in your culture **/ OR**

B. in which you describe your room **(12 marks)**

10/12

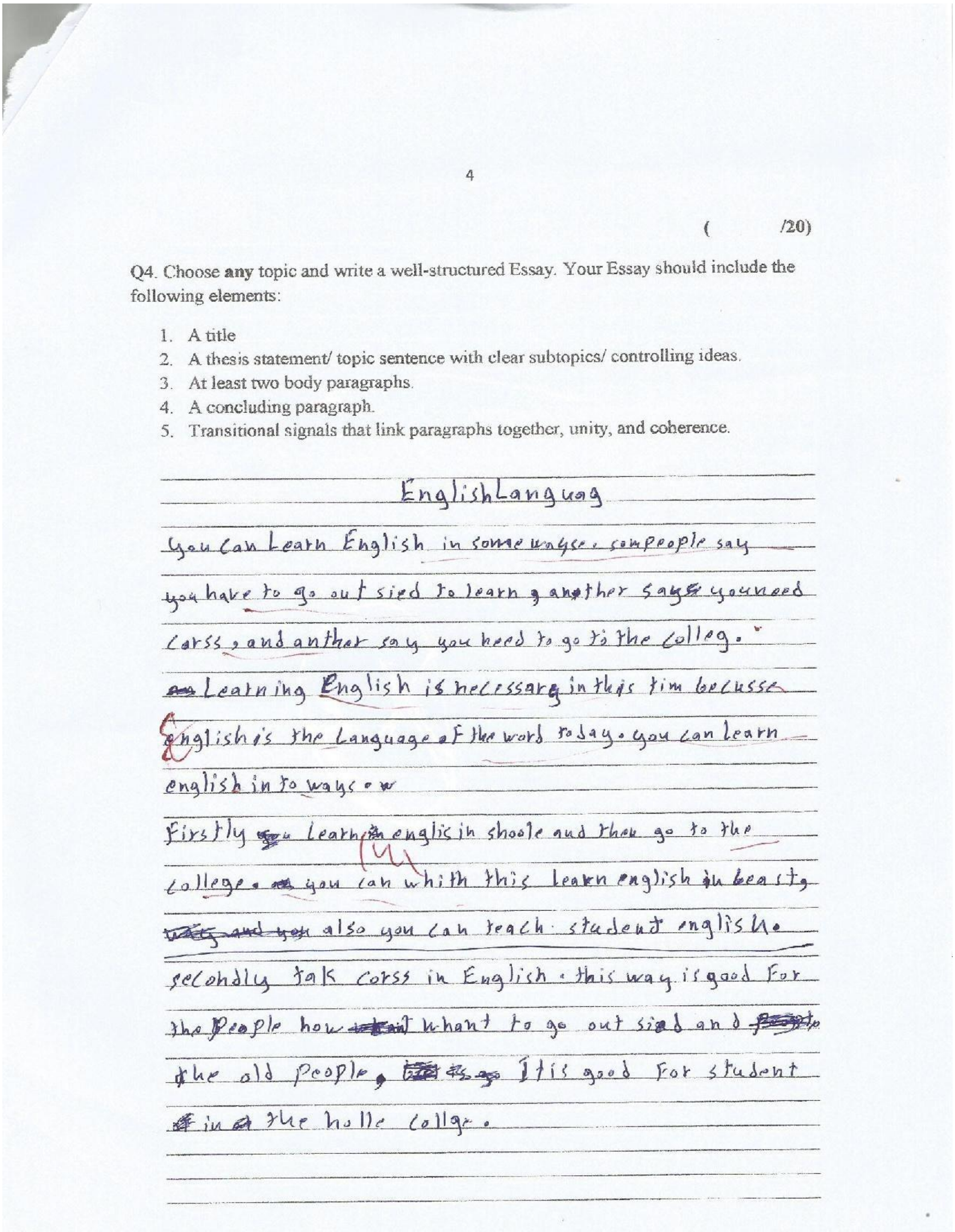
Libyan wedding

in our country we celebrate by wedding and clothes wedding like this way for example, the bride wears whight ~~waxy~~ dress and Puts the makeuP, cut her hair and Puts Henna in her hands and legs. The groom wears black or gray suit and we make diffrent tradtional dishes like Rice with khalta or Makaronawith khalta to sum uP, Libyan wedding it's very beautifull.

3

[redacted] (366)

Appendix C: Year 3 Final Writing Exam



4

(/20)

Q4. Choose **any** topic and write a well-structured Essay. Your Essay should include the following elements:

1. A title
2. A thesis statement/ topic sentence with clear subtopics/ controlling ideas.
3. At least two body paragraphs.
4. A concluding paragraph.
5. Transitional signals that link paragraphs together, unity, and coherence.

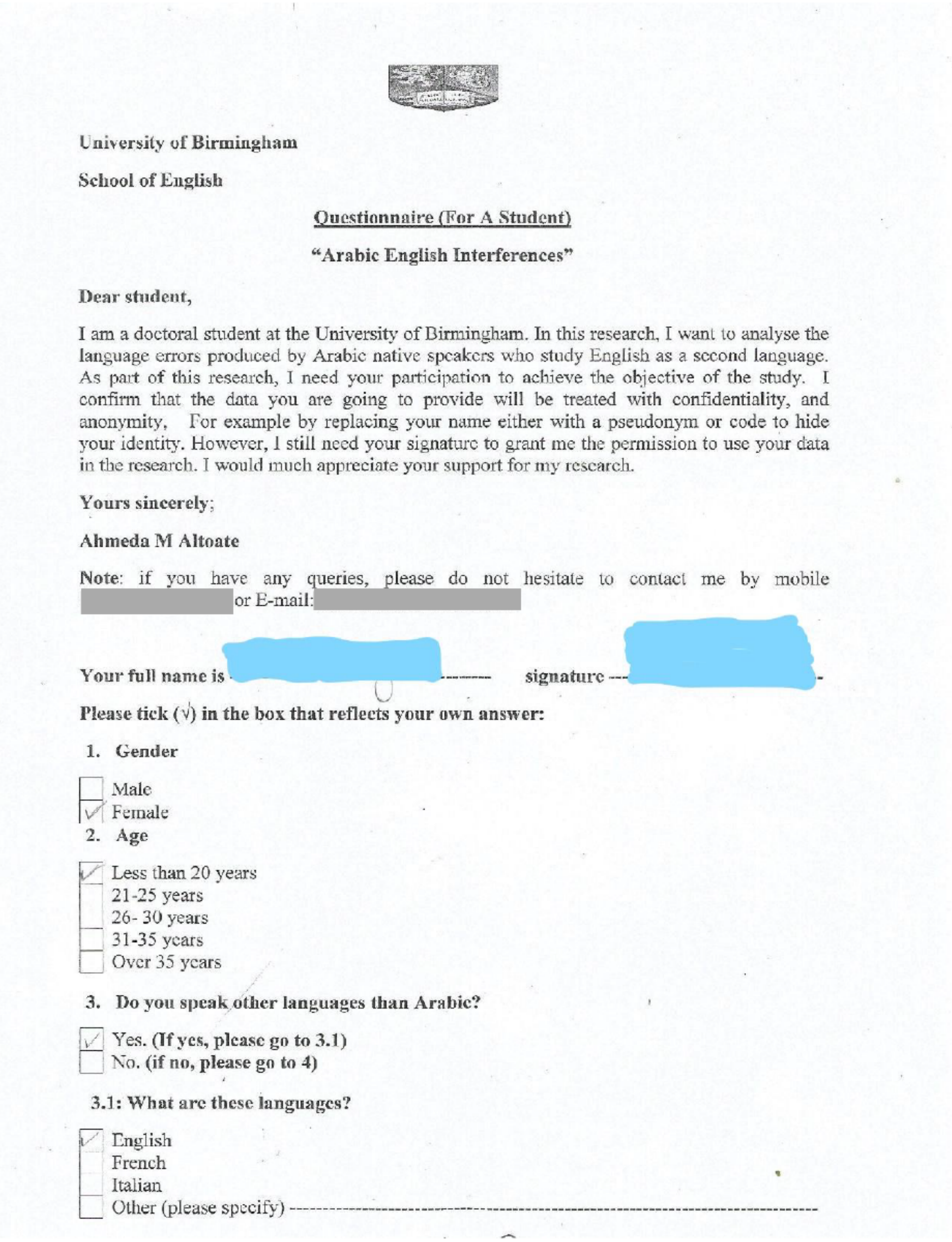
EnglishLanguag

You can Learn English in some wayse. compeople say you have to go out sied to learn, another says you need carss, and anther say you need to go to the colleg. Learning English is necessarg in this tim becusse english is the Language of the word today. you can learn english in to ways.

Firstly learn english in shoole and then go to the college. you can whith this learn english in beasty ~~way and you~~ also you can teach student english.

secondly tak corss in English. this way is good for the People how whant to go out siad and the old People, It is good For student in the holle collge.

Appendix D: The Questionnaires



University of Birmingham

School of English

Questionnaire (For A Student)

"Arabic English Interferences"

Dear student,

I am a doctoral student at the University of Birmingham. In this research, I want to analyse the language errors produced by Arabic native speakers who study English as a second language. As part of this research, I need your participation to achieve the objective of the study. I confirm that the data you are going to provide will be treated with confidentiality, and anonymity, For example by replacing your name either with a pseudonym or code to hide your identity. However, I still need your signature to grant me the permission to use your data in the research. I would much appreciate your support for my research.

Yours sincerely;

Ahmeda M Altoate

Note: if you have any queries, please do not hesitate to contact me by mobile [redacted] or E-mail: [redacted]

Your full name is [redacted]-------- **signature** ---[redacted]-

Please tick (√) in the box that reflects your own answer:

1. **Gender**

□ Male
☑ Female

2. **Age**

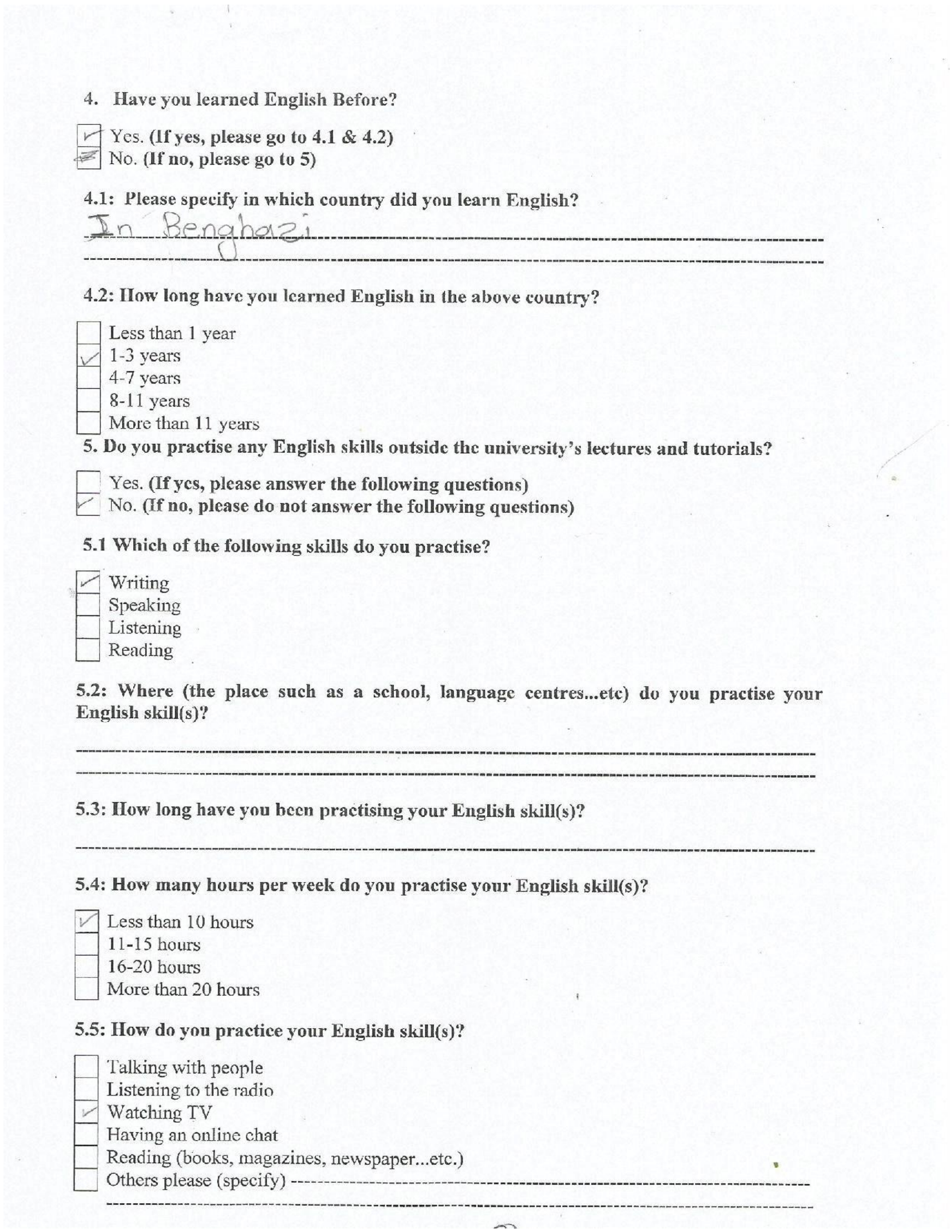
☑ Less than 20 years
□ 21-25 years
□ 26- 30 years
□ 31-35 years
□ Over 35 years

3. **Do you speak other languages than Arabic?**

☑ Yes. **(If yes, please go to 3.1)**
□ No. **(if no, please go to 4)**

3.1: What are these languages?

☑ English
□ French
□ Italian
□ Other (please specify) ---



4. Have you learned English Before?

☑ Yes. **(If yes, please go to 4.1 & 4.2)**
☒ No. **(If no, please go to 5)**

4.1: Please specify in which country did you learn English?

In Benghazi

4.2: How long have you learned English in the above country?

□ Less than 1 year
☑ 1-3 years
□ 4-7 years
□ 8-11 years
□ More than 11 years

5. Do you practise any English skills outside the university's lectures and tutorials?

□ Yes. **(If yes, please answer the following questions)**
☑ No. **(If no, please do not answer the following questions)**

5.1 Which of the following skills do you practise?

☑ Writing
□ Speaking
□ Listening
□ Reading

5.2: Where (the place such as a school, language centres...etc) do you practise your English skill(s)?

5.3: How long have you been practising your English skill(s)?

5.4: How many hours per week do you practise your English skill(s)?

☑ Less than 10 hours
□ 11-15 hours
□ 16-20 hours
□ More than 20 hours

5.5: How do you practice your English skill(s)?

□ Talking with people
□ Listening to the radio
☑ Watching TV
□ Having an online chat
□ Reading (books, magazines, newspaper...etc.)
□ Others please (specify) ---

Appendix E: A full list of tagging codes used in LEFLL corpus

Main Category	Tagging Code	Definition	Examples from LEFLL corpus
Spelling	SP	Spelling error	I will be a doctor in the <<universty>>
	SPED1	One-edit distance spelling error	... and <<anther>> say you need to go to the
	SPED2	Two-edit distance spelling error	To come to her and <<viste>> this sharming country
	SPEDM	Multiple-edit distance spelling error	arabic countries have a lot of thousand <<proficinal>> people
	SPIN	Insertion of unnecessary letter	<<Firstlly>> The food
	SPOM	Omission of necessary letter	Smocking also could <<caus>> the death
	SPSUB	Substituting a letter with incorrect one	she can easily <<convay>> the meanings
	SPTRA	Transposing two adjacent letters	iF someone <<decied>> to mak party
Noun Phrase	NP	Noun phrase	<<the number of students>> are completly different
	NPPOTE	The constituent potential for error in a noun phrase	<<the>> <<number>> of <<students>> are completly different
	NPE	Noun phrase error	Fainally,<< good teacher>> is file
	NPDETOM	Omission of a necessary determiner	if you want to work in <<company>>
	NPDETRED	Addition of unnecessary determiner (Redundant)	After that I go to <<the>> shopping
	NPDETW	Substituting a determiner with incorrect one	Ead Alfeter is <<a day>> comes after Ramadhan's mounth
	NPGER	Confusion with Gerund	<<Celebrate a special day>> is my birth
	NPLTR	Literally transfer from Arabic	there is a very big tv with 42 <<bosah>> (for inch)
	NPNO	Number agreement error	but there are <<two main type of friend ship>>
	NPNOREG	Plural Regulation error	we need all of them in our <<lifes>>,

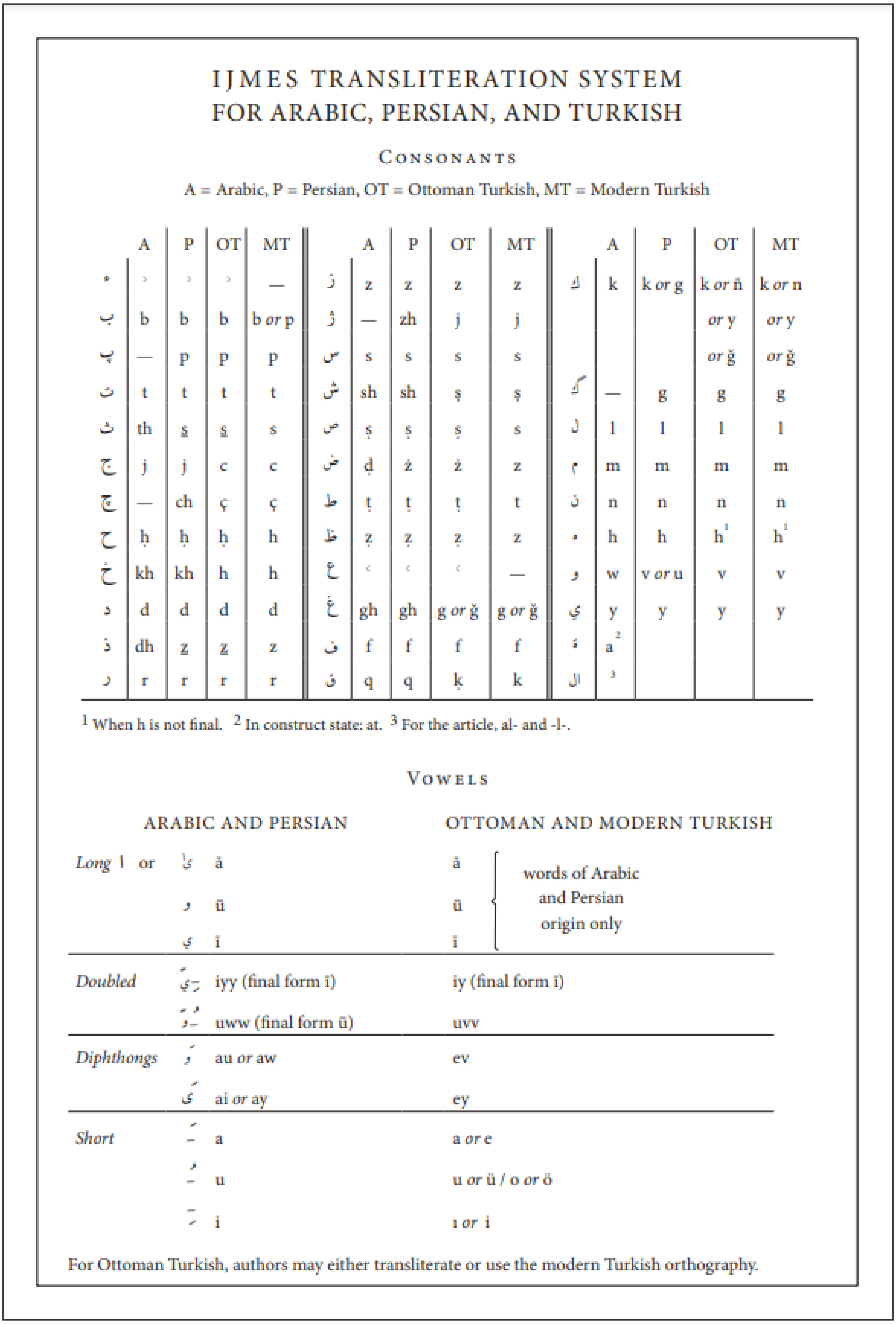
	NPNOM	A noun omission error	saterday is <<a busy >>for me.
	NPNRED	Addition of an unnecessary noun (Redundant	my room is big <<room>> and beatifull.
	NPNW	Substituting a noun with incorrect one	I have <<a big library>>, you can find any book (for *bookcase*)
	NPPRONOM	Pronoun omission error	in my room << >>have one bed, one TV,
	NPPRONRED	Addition of an unnecessary pronoun	and searching in the computer about the thing I don't know <<it>>.
	NPPRONW	Substituting a pronoun with incorrect one	I have alot of make-up in my room because I love <<her>> so much
	NPWWC	Wrong word class in a noun phrase	I think health and <<educated>> can be build very strong nation
	NPWWO	Wrong word order in a noun phrase	The oFFice was in << Center City>>
	NPAPOSTOM	Apostrophe omission error	After that I drive <<my son car>>
	NPAPOSTRED	Addition of unnecessary apostrophe	and I hope to learn many <<language's>>.
Verb Phrase	VP	Verb Phrase	You <<can Learn>> English in some ways
	VPPOTE	The constituent potential for error in a verb phrase	You <<can>> <<Learn>> English in some ways
	VPE	Verb phrase error	I <<Applaying>> For a few diFFerent post
	VPAPOSTRED	Addition of an unnecessary apostrophe in a verb phrase	I <<a'm>> from Liban,
	VPAUXOM	An auxiliary omission error	After I graduate I <<>> going to work as a translator.
	VPAUXRED	Addition of an unnecessary auxiliary	I <<am>> study in university of Benghazi.
	VPAUXW	Substituting an auxiliary with incorrect one	I <<am not sleep>> at afternoon
	VPGERINF	Gerund and infinitive confusion error	small family helps country <<to controlling>> the popullation of the growth.
	VPPV	Phrasal verb confusion error	I stopped to buy some plaster . when I got

			oFFice <<put it in>> my Feet
	VPAPF	Active and passive voice confusion error	the teachers who studying (for teach at) secondary school <<are obtained>> the bachelor degree
	VPREG	Verb conjugation regularity	I <<chosed>> simple style
	VPT	Tense error	help them when they are older and <<became>> a parent
	VPVOM	Main verb omission error	my father << >>a doctor,
	VPVRED	Addition of an unnecessary main verb	Then I clean the house and <<clean>> my bedroom.
	VPVW	Substituting a main verb with incorrect one (wrong verb)	I <<do>> a shower before any thing
	VPWWC	Wrong word class in a verb phrase	Finally we never <<improvement>> unless we have a modren equipment
	VPWWO	Wrong word order in a verb phrase	Feeling <<not should be>> in focuse
	VPFIN	Finite confusion error	we study, play, and <<lestining>> to music in it.
	VPVNCON	Noun-verb concord	For example , << child sometimes cry>> for more caring
prepositional Phrase	PP	prepositional phrase	<<According to>> statistics , more men are speeding than women
	PPPOTE	The constituent potential for error in a prepositional phrase	<<According>> <<to>> statistics , more men are speeding than women
	PPE	A prepositional phrase	49% <<From>> adult people are smockers
	PPPREPOM	A preposition omission error	in studying << >>university the number of the students
	PPPREPRED	Addition of an unnecessary preposition	Chocolet was discovered <<from>> 400,000 years ago by mexican

	PPPREPW	Substituting a preposition with incorrect one(wrong preposition)	Many parents <<on>> my country have many children they
	PPWWC	Wrong word class in a prepositional phrase	after that I waching TV, <<next>> that I go to prae, Finaly
	PPWWO	Wrong word order in a prepositional phrase	My alarm rings <<about at>> 7:00.
Adjective Phrase	ADJP	Adjective phrase	I will work in an <<international>> company
	ADJPPOTE	The constituent in an adjective phrase potential for error	I will work in an <<international>> company
	ADJPE	An adjective phrase error	The <<sencerly>> friend will remains with us in our trouble
	ADJPAJPL	An adjective phrase plurality error	and this is my <<faviorts>> thingis,
	ADJPAJOM	An adjective omission error	I am 18 years <<.>>
	ADJPAJRED	Addition of an unnecessary adjective	I am 19 <<old>> years old,
	ADJPAJW	Substituting an adjective with incorrect one (wrong adjective)	I'm not married Because I'm very <<small>>
	ADJPLTR	Literally transfer of an adjective from Arabic	I made sure that the sheets are <<bage>> and so as the pillows
	ADJPAPOSTRED	Addition of an unnecessary apostrophe	it must be bulid in a <<scientifc's>> study
	ADJPWWC	Wrong word class in an adjective phrase	then, to make <<health>> life without disease.
	ADJPWWO	Wrong word order in an adjective phrase	so, my Room is <<cleaning Always>>,
Adverb Phrase	ADVP	An adverb phrase	my boss <<oFten>> says that he would lost without me
	ADVPPOTE	The constituent in an adverb phrase potential for error	my boss <<oFten>> says that he would lost without me
	ADVPE	An adverb phrase error	you should speak english very <<good>> to communicate with forgin people

	ADVPADOM	An adverb omission error	people can cause traffic accidents because don't know<<>>fast they are driving their car
	ADVPADRED	Addition of an unnecessary adverb	Translation process become <<more>> easier.
	ADVPADW	Substituting an adverb with incorrect one	five minutes <<just>>, and this is not enough time for students …
	ADVPAPOSTOM	Apostrophe omission error	drinking drivers they do<<nt>> feel they are drunk
	ADVPAPOSTRED	Addition of unnecessary apostrophe	in libya it is very hot and <<some time's>> come dusty

Appendix F: Ijmes transliteration system for Arabic, Persian and Turkish



IJMES TRANSLITERATION SYSTEM FOR ARABIC, PERSIAN, AND TURKISH

Consonants

A = Arabic, P = Persian, OT = Ottoman Turkish, MT = Modern Turkish

	A	P	OT	MT		A	P	OT	MT		A	P	OT	MT
ء	ʾ	ʾ	ʾ	—	ز	z	z	z	z	ك	k	k *or* g	k *or* ñ	k *or* n
ب	b	b	b	b *or* p	ژ	—	zh	j	j				*or* y	*or* y
پ	—	p	p	p	س	s	s	s	s				*or* ğ	*or* ğ
ت	t	t	t	t	ش	sh	sh	ş	ş	گ	—	g	g	g
ث	th	s̱	s̱	s	ص	ṣ	ṣ	ṣ	s	ل	l	l	l	l
ج	j	j	c	c	ض	ḍ	ż	ż	z	م	m	m	m	m
چ	—	ch	ç	ç	ط	ṭ	ṭ	ṭ	t	ن	n	n	n	n
ح	ḥ	ḥ	ḥ	h	ظ	ẓ	ẓ	ẓ	z	ه	h	h	h[1]	h[1]
خ	kh	kh	h	h	ع	ʿ	ʿ	ʿ	—	و	w	v *or* u	v	v
د	d	d	d	d	غ	gh	gh	g *or* ğ	g *or* ğ	ي	y	y	y	y
ذ	dh	ẕ	ẕ	z	ف	f	f	f	f	ة	a[2]			
ر	r	r	r	r	ق	q	q	ḳ	k	ال	[3]			

[1] When h is not final. [2] In construct state: at. [3] For the article, al- and -l-.

Vowels

		ARABIC AND PERSIAN	OTTOMAN AND MODERN TURKISH	
Long	ا *or* ى	ā	ā	words of Arabic and Persian origin only
	و	ū	ū	
	ي	ī	ī	
Doubled	ـيّ	iyy (final form ī)	iy (final form ī)	
	ـوّ	uww (final form ū)	uvv	
Diphthongs	وَ	au *or* aw	ev	
	يَ	ai *or* ay	ey	
Short	ـَ	a	a *or* e	
	ـُ	u	u *or* ü / o *or* ö	
	ـِ	i	ı *or* i	

For Ottoman Turkish, authors may either transliterate or use the modern Turkish orthography.