# Error Analysis in a Learner-Written Corpus of Spanish Speakers EFL Learners: A Corpus-Based Study

Presented to the Graduate Program of the

University of Antioquia

A Dissertation Submitted in Partial

Fulfilment of the Requirements for the

Degree of Doctor of Philosophy

(Linguistics)

by

María Victoria Pardo Rodríguez

Supervisor: Prof. Dr. Gabriel Ángel Quiroz

Universidad de Antioquia

Facultad de Comunicaciones

Departamento de Lingüística

2019

# Error Analysis in a Learner-Written Corpus from Spanish Speakers EFL Learners: A Corpus-Based Study

El análisis de errores en un corpus de documentos escritos por hablantes de español aprendices de inglés como lengua extranjera: Un estudio de corpus en estudiantes universitarios

**DOCTORAL THESIS**

**Researcher:**

María Victoria Pardo Rodríguez

Bachelor in Modern Languages

Master in Translation

**UNIVERSITY OF ANTIOQUIA**

**PH.D. IN LINGUISTICS**

**MEDELLÍN**

**2019**

# Contents

*DOCTORAL THESIS* 2

*Researcher:* 2

*UNIVERSITY OF ANTIOQUIA* 2

*Contents* 4

*List of Tables* 8

*List of Figures* 12

*List of Annexes* 16

*Abstract* 17

*Acknowledgements* 18

*1. INTRODUCTION* 20

*2. LITERATURE REVIEW* 26

**2.1. Overview of Error Analysis** 27

**2.2. Research in the Field of Error Analysis (EA)** 31

**2.3. Criticism on Error Analysis** 34

**2.3.1. Limitations on methodological procedures and theoretical problems** 35

**2.3.2. Limitations in the scope of EA** 36

**2.4. Error Analysis Using Computerized Learner Corpora: A Counter-Argument Supporting EA** 37

**2.5. The Perception of Error in Linguistics, the Teaching of Languages and in Translation** 40

**2.6. The First Step into Error Analysis: The Diagnosis of Errors** 44

**2.7. Accounting for gender, social stratum and language development in Error Analysis** 51

**2.8. Overview of Corpus Linguistics (CL)** 54

**2.8.1. Main Features of a Written Corpus** 58

**2.8.2. Corpus annotation** 60

**2.9. Learners' Interlanguage Corpora** 60

**2.9.1. Overview of Learner Corpora (LC)** 61

**2.9.2. Dimensions that shape Learner Corpora** 62

**2.9.3. Environment, tasks and learner variables in a Learner Corpus (LC)** 63

**2.9.4. Corpus collection and process features** 64

**2.9.5. Annotation of Learner Corpora** 66

**2.9.6. Learner corpus research and the acquisition of a second or a foreign language** 68

**2.10. Common European Framework of Reference (CEFR) Writing Descriptors** 69

***3. METHODOLOGY*** *79*

**3.1. Error Analysis (EA) Procedure** 81

**3.2. Corpus Design** 85

**3.2.1. Corpus approach** 85

**3.2.2. Collection of data** 86

**3.2.3. Elicitation procedures** 89

**3.2.4. Corpus annotation** 91

**3.2.5. Data extraction software** 92

**3.2.6. Annotation system** 93

**3.3. Interface Design** 94

**3.3.1. The CLEC description** 95

**3.3.2. Interface methodology** 95

**3.3.3. Resulting product** 98

**3.4. Instrument Design. Methodology for Collection of Sociolinguistic Aspects** 104

**3.4.1. Variables of study** 107

***4. ANALYSIS*** *109*

**4.1. Verification or Proof of Previously Established Hypotheses** 109

**4.2. Description of research outcomes. Sociocultural variables** 118

**4.2.1. Population description. Parts one and two of the survey** 118

**4.2.2. Results of socio-cultural factors that could have incidence in the development of EFL. Part three of the survey** 135

**4.3. Description of Research Outcomes According to Research Objectives** 143

**4.3.1. Complete list of errors** 144

**4.3.3. Dispersion of errors in the corpus** 147

**4.3.3.1. Error dispersion in Pre-Intermediate level (B1.1)** 149

**4.3.3.2. Error dispersion in Intermediate level (B1.2)** 152

**4.3.3.3. Error dispersion in Intermediate II level (B1.3-B2.1)** 156

**4.3.3.4. Error dispersion in Upper-Intermediate level (B2.2-B2.3)** 160

**4.3.4. Overview of errors by gender in levels B1 and B2** 164

**4.3.4.1. Errors in level B1 and B2** 164

**4.3.4.2. General analysis of errors by gender** 172

**4.3.5. Errors and Common European Framework of Reference (CEFR)** 174

**4.3.6. Overview of errors by strata** 177

**4.3.6.1. Analysis of errors by strata in level B1** 180

**4.3.6.2. Analysis of errors by social strata in level B2** 182

**4.4. Incidence of errors by category: Overview of eight error categories** **193**

**4.4.1. Analysis of Grammatical errors (G)** 197

**4.4.2. Analysis of Lexis Errors (L)** 202

**4.4.3. Analysis of Word errors (W)** 206

**4.4.4. Analysis of Form errors (F)** 209

**4.4.5. Analysis of Punctuation errors (Q)** 212

**4.4.6. Analysis of Style errors (S)** 214

**4.4.7. Analysis of Lexico-Grammar errors (X)** 216

**4.4.8. Analysis of Infelicities (Z)** 218

***5. CONCLUSIONS AND RECOMMENDATIONS*** ***221***

**5.1. Synthesis of results** **222**

**5.2. Scope and limitations** **228**

**5.3. Problems found in the process** **229**

**5.4. Recommendations** **230**

***REFERENCES*** ***233***

***Annexes*** ***242***

## List of Tables

Table 1 Equation between error and mistake vs. acquisition and learning ..... 29

Table 2 Intralingual errors: learning-strategy-based errors ..... 47

Table 3 Intralingual errors: communication strategy-based errors ..... 48

Table 4 Induced errors prompted by teachers ..... 49

Table 5 Other induced errors ..... 50

Table 6 Dimensions that distinguish learner corpus ..... 63

Table 7 Common reference levels B1-B2 Global Scale ..... 70

Table 8 Writing self-assessment grid for levels A1 to C2 ..... 71

Table 9 Descriptors on overall written production levels B1 and B2 ..... 72

Table 10 Writing descriptors for creative writing, reports and essays in levels B1 and B2 .... 72

Table 11 CEFR learning, teaching, assessment for vocabulary range ..... 73

Table 12 CEFR learning, teaching, assessment for vocabulary control ..... 73

Table 13 CEFR learning, teaching, assessment for grammatical accuracy ..... 73

Table 14 CEFR learning, teaching, assessment for orthographic control ..... 74

Table 15 CEFR learning, teaching, assessment for sociolinguistic appropriateness ..... 74

Table 16 CEFR learning, teaching, assessment for coherence and cohesion ..... 75

Table 17 Errors that improve from A1 to B1 ..... 75

Table 18 Errors that improve from B1 to B2 ..... 76

Table 19 Error categories with their tags ..... 84

Table 20 Classification from Universidad del Norte according to the CEFR ..... 88

Table 21 Errors by categories and levels ..... 101

Table 22 Average time in each case ..... 101
Table 23 Summary of variables ..... 107
Table 24 Normality tests ..... 111
Table 25 Non-parametric test – U de Mann Whitney ..... 113
Table 26 Statistics of test[a] ..... 113
Table 27 Non-parametric test – K Kruskal-Wallis ..... 114
Table 28 *Tests [a,b] statistics 1* ..... 114
Table 29 Ranges ..... 117
Table 30 Testsa,b statistics 2 ..... 117
Table 31 Distribution of students by age ..... 119
Table 32 Distribution of students by socio-economic strata ..... 121
Table 33 Distribution of students by bachelor's degree ..... 122
Table 34 Distribution of students by semester of their BAs ..... 124
Table 35 Distribution of students by EFL level enrolled ..... 125
Table 36 Cross variable: semester of enrolment & English level ..... 126
Table 37 Students enrolled in language courses different from English ..... 127
Table 38 Time spent in an English-speaking country & trip purpose ..... 129
Table 39 Trip purpose ..... 130
Table 40 Institutions attended in elementary school ..... 131
Table 41 Institutions attended in secondary school ..... 132
Table 42 Location of institution where students attended elementary school ..... 133
Table 43 Location of institution where students attended secondary school ..... 134
Table 44 The most and least important reasons to study English ..... 137

Table 45 Time spent listening to music in English ..... 138

Table 46 Motivation to listen to music in English ..... 138

Table 47 Language established in PC as default ..... 139

Table 48 Is or not the environment a facilitator of the English learning process? ..... 140

Table 49 Will you become proficient in English while living in a Hispanic culture? ..... 141

Table 50 Importance of Interaction in English within a group of friends or classmates ..... 142

Table 51 Grammatical errors ..... 144

Table 52 Lexis errors ..... 145

Table 53 Word errors ..... 145

Table 54 Form errors ..... 146

Table 55 Punctuation errors ..... 146

Table 56 Lexico-Grammar errors ..... 146

Table 57 Style errors ..... 146

Table 58 Infelicities ..... 147

Table 59 Errors with most incidence in level B1.1 ..... 150

Table 60 Errors with most incidence in level B1.2 ..... 153

Table 61 Errors with most incidence in level B1.3-B2.1 ..... 157

Table 62 Comparative of the first 5 errors in three different levels ..... 157

Table 63 Errors with most incidence in level B2.2-B2.3 ..... 160

Table 64 Comparative of the first 5 errors in four different levels ..... 161

Table 65 Incidence of the first five type of errors in B1 and B2 levels ..... 168

Table 66 Comparative chart of errors in B1 to B2 ..... 175

Table 67 Errors that improve from B1 to B2 ..... 176

Table 68 Distribution of students that confirmed their strata ..... 177
Table 69 Classification of students by strata..... 178
Table 70 Errors and social strata from male students in B1 ..... 180
Table 71 Errors and strata from female students in B1..... 181
Table 72 Errors and strata from male students in B2..... 189
Table 73 Errors and strata from female students in B2..... 190
Table 74 Main five errors and quantities in B2..... 191
Table 75 Categories of errors..... 194
Table 76 Total of errors by categories with percentages and means in corpus..... 195
Table 77 Numbers of errors tags..... 196
Table 78 Five errors with most incidence in the grammar category..... 197
Table 79 Examples of GA errors and possible cause..... 199
Table 80 First three errors with most incidence in the lexis category..... 202
Table 81 Word errors with percentages and means ..... 206
Table 82 Form errors with percentages and means..... 210
Table 83 Punctuation errors with percentages and means ..... 212
Table 84 Style errors with percentages and means ..... 214
Table 85 Lexico-Grammar errors with percentages and means..... 216
Table 86 Infelicities with percentages and means..... 219

## List of Figures

Figure 1. Retrieved from Valanglia (González, 2017). ..... 87

Figure 2. Architecture MTV application. ..... 96

Figure 3. CLEC Interface login. Source TNT Research group. ..... 98

Figure 4. CLEC Interface search of level. Source TNT Research group. ..... 99

Figure 5. CLEC Interface search of level. Source TNT Research group. ..... 100

Figure 6. Graphic of roles. ..... 102

Figure 7. Graphic Interface for an invited user. ..... 103

Figure 8. Response page CLEC. ..... 104

Figure 9. Histogram standard deviation of errors from 515 files. Source: Henao et al. (2018). ..... 112

Figure 10. Distribution of observations by social strata. Source: Henao et al. (2018). ..... 116

Figure 11. Distribution of observations by social strata (1-6). Source: Henao et al. (2018). 116

Figure 12. Account of students according to age range. ..... 120

Figure 13. Account of students by gender. ..... 121

Figure 14. Account of students by strata. ..... 122

Figure 15. Account of students by bachelor's degree. Source from the author. ..... 123

Figure 16. Account of students by semester of enrolment. ..... 124

Figure 17. Account of students by CEFR level. ..... 125

Figure 18. Account of students by enrolment in courses different from English. ..... 128

Figure 19. Time spent in an English-speaking country ..... 130

Figure 20. Trip purpose to an English-speaking country. ..... 131

Figure 21. Institutions attended in elementary school. ..... 132

Figure 22. Institutions attended in Secondary School. ..... 133
Figure 23. Location of institutions attended in elementary school. ..... 134
Figure 24. Location of institutions attended in Secondary school. ..... 135
Figure 25. Importance of learning English. ..... 136
Figure 26. Reasons to study English. ..... 137
Figure 27. Time devoted to listen to music in English. ..... 138
Figure 28. Motivation to listen to music in English. ..... 139
Figure 29. Language established in PC as default. ..... 140
Figure 30. Environment as a facilitator of the EFL learning process. ..... 141
Figure 31. Environment as a facilitator of the EFL learning process. ..... 142
Figure 32. Importance of interacting in English in groups. ..... 143
Figure 33. Distribution of errors in the corpus. ..... 148
Figure 34. Graphic of errors in the corpus. ..... 148
Figure 35. Distribution of errors in level B1.1. ..... 150
Figure 36. Most relevant errors in level B1.1. ..... 151
Figure 37. Distribution of errors in level B1.2. ..... 153
Figure 38. Most relevant errors in level B1.2. ..... 155
Figure 39. Distribution of errors in level B1.3-B2.1. ..... 157
Figure 40. Most relevant errors in level B1.3-B2.1. ..... 158
Figure 41. Distribution of errors in level B1.1. ..... 160
Figure 42. Most relevant errors in level B2.2-B2.3. ..... 162
Figure 43. Overview of errors by gender in each level. ..... 164
Figure 44. Overview of errors by gender in level B1. ..... 165

Figure 45. Overview of errors by gender in level B2. ..... 166

Figure 46. Comparative of errors by gender in levels B1 & B2. ..... 167

Figure 47. Distribution of errors by gender in level B1.1. ..... 169

Figure 48. Distribution of errors by gender in level B1.2. ..... 170

Figure 49. Distribution of errors by gender in level B1.3-B2.1. ..... 171

Figure 50. Most frequent written errors found in level B1. ..... 174

Figure 51. Most frequent written errors found in level B2. ..... 175

Figure 52. Distribution of errors by strata in levels B1 and B2. ..... 179

Figure 53. Distribution of errors by strata in level B1. ..... 180

Figure 54. Distribution of errors by strata in level B2. ..... 182

Figure 55. Distribution of errors level B2 stratum 1. ..... 183

Figure 56. Distribution of errors level B2 stratum 2. ..... 184

Figure 57. Distribution of errors level B2 stratum 3. ..... 185

Figure 58. Distribution of errors level B2 stratum 4. ..... 186

Figure 59. Distribution of errors level B2 stratum 5. ..... 187

Figure 60. Distribution of errors level B2 stratum 6. ..... 188

Figure 61. Most relevant errors by type. ..... 194

Figure 62. Great total of errors in eight categories. ..... 195

Figure 63. Account of grammatical errors. ..... 198

Figure 64. Account of Lexis errors. ..... 203

Figure 65. Account of Word errors. ..... 207

Figure 66. Account of Form errors. ..... 211

Figure 67. Account of Punctuation errors. ..... 213

Figure 68. Account of Style errors. ..... 215

Figure 69. Account of Lexico-Grammar errors. ..... 217

Figure 70. Account of Infelicities. ..... 219

## List of Annexes


| | | Page |
|---|---|---|
| 1. | Error tagging manual | 241 |
| 2. | Writing assessment directions level B1.1 | 241 |
| 3. | Writing assessment directions level B1.2 | 244 |
| 4. | Writing assessment directions level B1.3-B2.1 | 247 |
| 5. | Writing assessment directions level B2.2-B2.3 | 249 |

## Abstract

This thesis was conducted to investigate the relationship between the main errors found in written compositions from students at university level and the socio-demographic factors that could have incidence in the development of the writing skills at ***Universidad del Norte*** **in Barranquilla, Colombia.**

The participants in the present research are 515 university students classified through a placement test in accordance with the parameters from the Common European Framework of Reference for Languages (CEFR). The data was compiled following the steps of Corpus Linguistics methodology, correspond to the written compositions of the participants, and was the final task for the course of English as a Foreign Language. The analysis of the data was based on Corder's ( 1981) theory about Error Analysis (EA) and following the error description guide given by James (1998). This thesis presents a comprehensive framework of error analysis that aimed to reveal the frequent error patterns in a leaner corpus of university students as well as the relationship of these errors with some socio-demographic aspects that may have an impact on the development of writing. 14,631 errors in eight categories were identified and analysed by categories according to the EA methodology. The errors were labelled using Louvain University Error Tagging Manual version 1.2 (E. Dagneaux et al., 2005) in order to obtain comparable results with similar work worldwide. The analysis finds that among the eight error categories proposed by Louvain's tagger grammatical errors are the most recurrent. The results highlight the difficulties the learners present in the use or omission of the definite article. A comprehensive account and description with examples of the main errors in the different categories along with possible causes for future pedagogical intervention are presented.

# Acknowledgements


## God

"¡Oh almas criadas para esas grandezas y para ellas llamadas!, ¿qué hacéis?, ¿en qué os entretenéis? Vuestras pretensiones son bajezas y vuestras posesiones miserias. ¡Oh miserable ceguera de los ojos de vuestra alma, pues para tanta luz estáis ciegos y para tan grandes voces sordos, no viendo que, en tanto que buscáis grandezas y glorias, os quedáis miserables y bajos, de tantos bienes hechos ignorantes e indignos!" **San Juan de la Cruz**

We have been called to do great things, but is God the One who shapes our paths to accomplish them. Thanks to God, for giving me this tremendous opportunity that very few can have in this country. I hope to enlighten those who want to learn what I know with humbleness.

## My family

In the beginning was the Word, and the Word was with God, and the Word was God…[14] And the Word became flesh and dwelt among us, and we beheld His glory, the glory as of the only begotten of the Father, full of grace and truth. **Gospel of John 1. 1. & 1. 14.**

Everything starts as an idea that one day becomes a reality. This idea began to take shape several years ago when I was looking to help my students to improve their learning processes. What I did not imagine was the magnitude of a process like this that started as my individual project but little by little became my family project, and I say family project because without the help of my family it would not have been possible to achieve this goal. My family lived my absences when present at home I was working without allowing interruptions. My family was always there despite everything. For that, thank you, dear husband, for loving me unconditionally. To my beloved daughter for understanding my short calls. To my beloved mother for her patience when she visited me despite not having much time. To my beloved philosopher, Arturo for being my support when I needed to understand the epistemology of everything. To my beloved niece Paulita for her help and unconditional company. To Alex for many times repairing my computer from the distance. To all my relatives, who with words of encouragement helped me.

## To my thesis advisor

Thank you Dr. Gabriel Quiroz for teaching me so much during these years. I appreciate your time and kindness.

## Academics

Thanks to my advisor during my internship in Lancaster University, UK. Dr. Vaclav Brezina.

Thanks to professor Antonio Tamayo, for his kind help and advice in the statistical development of this thesis.

Thanks to the professors and colleagues from the Ph.D. program for their suggestions and contributions during the different colloquia.

**Sponsor and supporter**

To COLCIENCIAS and Universidad de Antioquia for the scholarship during the last four years

To Universidad del Norte for allowing me to do the collection of data and the time to do this research.

# 1. INTRODUCTION

This thesis examines the written compositions from learners of English as a Foreign Language, university students aiming to complete their bachelor's degree (BA) in different fields of knowledge at *Universidad del Norte* in Barranquilla, Colombia. This thesis presents an elaborated Error Analysis framework that facilitates the investigation of written errors from Spanish speakers learning English in a foreign context. According to Bell (2011) English as a Foreign Language (EFL) refers to the study of English where it is not the dominant language, e.g. people who study English in a Spanish-speaking country. ESL, by contrast, refers to the study of English where English is the dominant language, e.g. immigrants learning English in Canada or the USA. For the present study and because the students are learning English in Colombia, we refer to them as EFL learners/students.

This introductory chapter starts with an overview of the research context. It outlines the motivation of the study, the research problem, the hypotheses, and the main objectives proposed for this research.

To graduate from a BA in Colombia, students must certify proficiency in a foreign language. Students registered at *Universidad del Norte* in Barranquilla, Colombia must hold a B2 proficiency certification, according to the Common European Language Framework of Reference (CEFR) (Europe, 2001), in one of the following languages: German, French, Portuguese or English. Most students prefer to be assessed in the English language because it has been recognized worldwide and it has become very relevant in business and most activities. As a global tendency, EFL studies have become highly important for the relevance in all professional

fields in Colombia. “A language achieves a genuinely global status when it develops a special role that is recognized in every country”(Crystal, 1997, p.3). In addition, that “global status”, in Colombia, was achieved several years ago. The importance of the English language in Colombia has been reflected in an increase of commercial agreements with the USA, the European Union, EFTA (European Free Trade Association), Korea, and Canada, among others. (Sánchez, 2013). Another reason for students to choose English is that when they come from English-Spanish bilingual schools, they find it easier to gain more proficiency or certify their knowledge of English, instead of starting the process of learning a third language.

In general, *Universidad del Norte* offers an appropriate environment to learn a foreign language where students have access not only to different materials that allow them to continue their language development after class, but they are also in contact with native-speaker teachers who provide them support in the process of learning the language. In contrast, the environment that surrounds the university is of and for Spanish speakers, so these learners have to make an additional effort if they want to improve their English level, or at least to keep the level they reach at the university. To have a more detailed profile from the population in this study, all participating students filled out a survey that is explained in chapter 4.

My personal interest in this research, stem from my practice as an English teacher for several years in Colombia. For the last eight years, I have taught English for adult students who either already held a bachelor degree (BA), or were pursuing that goal. For many of these learners, the level of English demanded at a professional position can make a difference in their salary or future promotions. It is important for them to have an accurate use of vocabulary and structures

when they communicate orally, or when they attend meetings. Written English is especially important for its use in daily mails and formal documents.

In my role as an educator, when checking student's work, I noticed some errors that kept recurring patterns in some levels across most texts. For that reason, I wanted to elucidate the most prevalent errors that hinder communication and their connection with some sociocultural aspects. Recurrent errors that prevail from basic to advanced levels hinder communication in every language level. In addition, the lack of statistical analyses of learner corpora that provide an understanding of the interlanguage from EFL university students in Colombia, makes of this study an asset to the research of EFL in this country and in Latin America. I consider important to analyse why learners continue making errors even though teachers make them aware of that. Only when teachers know the characteristics of their students' written output and their errors, they will be able to adapt materials or design and select better activities according to their students' needs. This analysis could give clues about either cultural or sociolinguistic features that interfere in the process of learning a language.

> There could be no reason to engage in Error Analysis unless it served one or both of two objects. Firstly, to elucidate what and how the learner learns when he studies a second language. Secondly, the applied object of enabling the learner to learn more efficiently by exploiting our knowledge of his dialect for pedagogical purposes. Corder (1981, p.27).

The main objective of the present study is to investigate the relationship between the main errors found in written compositions from students at university level and the socio-demographic factors that could have incidence in the development of their writing skills **at *Universidad del Norte* in Barranquilla, Colombia.**

There is a group of propositions set forth as an explanation for the occurrence of errors in the present study:

a. Males and females present statistically significant differences in the median of errors from written production of English as a Foreign Language (EFL) at university level.
b. There are statistically significant differences in the median of errors from written production of EFL university students in relation to the stratum classification given in Colombia (DANE, n.d.).

Specific objectives are the following:

a. To analyse how different in type and quantity are errors from male and female students
b. To analyse if there is a relationship between social strata and writing skills.

For the first hypothesis, it is necessary to recall previous studies on the effect of gender in Second Language Acquisition (SLA). Several authors have considered gender as one of the main variables in language acquisition due to the differences in learning styles from males and females. Aliakbari & Mahjub, (2010) considered gender as one of the main variables in language acquisition due to the differences in learning styles from males and females. Rúa & López (2006) analysed individual cognitive and neurological factors in language learning for both genders pointing to a relationship between genders in SLA. Aries, (1976) called attention to possible differences in interactional styles from males and females second language learners.

Regarding the second hypothesis about how social strata could affect the development of a second or a foreign language, there is “tangible evidence that there are serious gaps in the schooling of the poor” (Stromquist, 2004, p.95) in Latin America. Some theories affirm that inequalities in education tend to affect the less privileged. “Class is a significant influence on identity and social organization in the settings of second language learning” (Collins, 2006, p. 4).

Social strata are determining with respect to the options students have when they choose quality education and possibilities to travel. As a result, social strata could affect future performance in EFL. Social class can be understood as a hierarchical stratification of people in a given society according their social, economic, occupational and educational statuses (Mark & Engels, 1998). In the case of Colombia, social strata refer to the classification of residential property that should fulfil the minimum requests of public utilities. That stratification is differential in the fees paid for public utilities. The highest strata pay higher fees for public utilities and that remaining money is used to cover the gap left by the lowest strata in order to subsidise them (DANE, n.d.). In applied linguistics and specifically in language learning, social class is a marginalizing aspect for those who do not have access to travel abroad or get in touch with the target culture.

The present work focused on the detection, description, explanation (classification) of learners’ errors in written work from students of English as a Foreign Language (EFL) that are pursuing a bachelor’s degree in different disciplines. Following the methodology of Error Analysis (EA) in the process of classification and interpretation proposed by Corder (1981), and using the methodology of Corpus Linguistics (CL) in the assortment, tagging and organization,

the written work was collected and errors were tagged using the annotation system from Louvain University version 1.2 (Dagneaux, Denness, Granger, Meunier, & Neff, 2005). The goal using this tagger was to obtain comparable results with similar work worldwide. In each case, every error received an interpretation about what the learner tried to say, and the sentence was reconstructed. This part of the work was very critical because "The whole success of our description of errors hinges upon the correctness of our interpretation of the learner's intentions or meaning" (Corder (1981), p.37).

# 2. LITERATURE REVIEW

In this chapter, a background to the research of Error Analysis on written compositions of EFL undergraduate university students in *Universidad del Norte*, in Barranquilla, Colombia, is presented. This background contains a comprehensive review of the most relevant literature to the investigation as well as the discernment required to understand the theoretical and rational frameworks. The chapter is divided into ten main sections. In **Section 2.1. ,** an overview of Error Analysis (EA), a methodology to research learners' errors from its inception is traced. It is analyzed as a response to Contrastive Analysis (CA) from its theoretical beginnings. In **Section 2.2. ,** a brief overview of the research in EA in the last years with special focus on learner corpora from Spanish speakers learning English as a foreign language is presented. In **Section 2.3. ,** some criticism on EA when early corpora appeared is reviewed. In **Section 2.4. ,** a counter argument of EA as a response to the criticism where EA is analyzed from the perspective of modern technology is presented. In **Section 2.5. ,** the perception of error in linguistics, the teaching of languages and translation is introduced. In **Section 2.6. ,** a view of error diagnosis as the first step into EA is offered. In **Section 2.7. ,** a brief account of work and theory on gender, social stratum and language development is given. **Section 2.8. ,** gives a description of corpus linguistics and its main features. **Section 2.9. ,** is dedicated to Learner Corpus and its main characteristics. Section **2.10. ,** presents some relevant aspects from the CEFR in written output.

## 2.1. Overview of Error Analysis

In the late sixties, the prevailing theory about Second Language Learning (SLL) was behaviourist (Skinner, 1953a) and sustained that to learn a language the learner should acquire some language habits. Behaviourist theory, with one of the most influential American psychologists, Skinner, affirmed that behaviour is determined by its consequences, be they reinforcements or punishments. According to this theory, a new behaviour is learnt through operant conditioning (Pavlov, 1927). It means there is a change of behaviour by the use of reinforcement after the desired response (Skinner, 1953b). Skinner claimed that children learn language based on behaviourist reinforcement through the association of words with meanings. Learning a language was believed to be a process of acquiring a set of new language habits through input. For that reason, errors were considered the result of interference of the mother tongue habits into the new acquired language.

From another point of view, Chomsky (1965) stated that acquiring tools to process an infinite number of sentences was not only dependent on language input. He proposed the theory of a Universal Grammar arguing that there were innate, biological grammatical categories that facilitate language development in children and language processing in adults. According to this theory, human beings are born with a Language Acquisition Device (LAD) that let them develop and manifest some grammar rules without being taught, which is why children are able to produce an infinite number of possible sentences without any formal training (Chomsky, 1965).

In the 1960s, researchers devoted a lot of time to compare and contrast languages. They wanted to contrast the mother tongue and the new language, searching for ways to predict and give reasons for learners' errors. Contrastive analysis (CA) appeared as a methodology for language teaching and learning because it deals with the comparison of structures across languages, therefore, CA became the main methodology to understand learners' errors. Nevertheless, CA left many gaps on its way because not all errors found in the learners' output could be linked to a particular language background. For that reason, researchers looking for answers about cognitive processes in SLA, started serious empirical research in the field. Psycholinguists influenced by Chomsky were searching if the processes that take part in first language acquisition were the same in second language acquisition.

In Corder's seminal paper "The significance of learner's errors" (1967) the author seriously questions the behaviouristic approach because in many cases using Contrastive analysis many errors were not predicted or explained. The following is a summary of his main remarks that advocate for an understanding of errors as part of a learning process that could give clues to improve the teaching practice.

a. Errors show the system of the language the learner is using. They give evidence of what the learner has taken in, which in many cases, is different from what teachers have given them, because **intake** is not equal to **input**
b. There is a "**transitional competence**" (Azevedo & Corder, 1983) shown in learner's errors made by first language (L1) and second language (L2) learners in which both develop an independent system of language

c. Errors are different from mistakes

d. Errors are meaningful because they tell the teacher what they need to teach or reinforce. They tell the researcher how learning takes place and they are means learners use to check and verify their hypotheses about the L2.

James (1998) defines a possible equation between **error/mistake** vs Krashen's (1982) distinction made about **acquisition/learning**. Table 1 explains the difference between mistake and error and when each one occurs.

Table 1

*Equation between error and mistake vs acquisition and learning*

| Mistake | Acquisition | Error | Learning |
|---|---|---|---|
| Acquiring a language (e.g. the case of a native speaker) is an unconscious process. Rules are acquired unconsciously therefore, the learner is not in a state of ignorance and will not make errors, but mistakes. | | Learning a language is the result of formal instruction. It requires consciousness about its rules (Krashen, 1982). In that case the learner makes errors but might be able to correct or avoid them using explicit knowledge. | |
| Mistakes are the result of a lack of attention. They are not systematic and are possibly self-corrected because the speaker knows the rule. | | Errors are systematic. They are the result of an ignorance of the rule; if the learner has not learnt a target language form, the result will be an error. | |

Source: data retrieved from James (1998).

Defined as, "the investigation of the language of second language learners" (Corder, 1971, p.14). Error Analysis (EA) appeared in the early 1970s as an alternative approach to CA. Corder (1967) was the first author to analyse the idea that second language learners generated an autonomous linguistic system that he called "**transitional competence.**" The author argued that learners gradually modify their native language rules towards target language rules probably

using a universal grammar or what he called a "**built-in syllabus.**" Furthermore, Nemser (1969) referred to the linguistic system students develop as "**approximative system**".

Even though those concepts were proposed before, **interlanguage** was the most accepted term to refer about the learners' linguistic system. The concept of interlanguage was first used by Selinker (1972) in reference to the linguistic system produced by foreign or second language learners when they attempt to communicate. Selinker affirmed learners'use of language was systematic at every level. It is an autonomous system different from their Native Language (NL) or the Target Language (TL). In order words, it is a transitional linguistic system with its own patterns and rules. This hypothesis maintains that the acquisition and use of an interlanguage are unconscious processes because learners are not aware of its linguistic characteristics. Another important aspect of interlanguage, that is relevant to EA, was the concept of **fossilization**. Selinker (1972, 209) assured that an interlanguage always fossilizes because it stops its developing process at some point before it is identical to the native speaker target language system. It generally occurs when some mistakes seem impossible to correct despite the ability or motivation learners could have. In other words, second language (SL) learners or English as a Foreign Language (EFL) learners will never produce the target language as accurately as native speakers do.

After all this process, Error Analysis (EA) rises as a methodology to analyse the interlanguage of language learners to describe it.

## 2.2. Research in the Field of Error Analysis (EA)

Even though by the end of the 1960's, EA was considered an acceptable alternative approach to Contrastive Analysis, until now, most of the work done through this methodology has been planned and produced in Europe and the USA. In Latin America, a few researchers have done work on Error Analysis and none of them has used the methodology of a linguistic computerized corpus in the compilation of the data because most of their work has been done in the analysis of one or few samples.

Nevertheless, there has been an increased interest about computerized corpus to carry out computational analysis of English learner corpora. The International Corpus of Learner English (ICLE) from Louvain University in Belgium is an example. It includes learners whose mother tongues are very varied: French, Dutch, German, Swedish, Chinese, Spanish, Finish, Japanese, Czech, among others, their corpora are not only of written work, but also, they include spoken corpora. By 2015 the University of Louvain maintained in their website 137 learner corpora, 82 of them (60%) were from L2 English and the rest from other languages (Granger, S, Gilquin, G, Meunier, 2015, p.2).

In the present research, some work related to the field of Error Analysis is presented, but it would be impossible to give full account of the work done around the world. A preference was given to work in Latin America or Spain, given the fact that the learners from this research share the mother tongue from learners in those countries, so that previous work is highly enlightening

and relevant for the current study. Every description uses mostly the corresponding abstract's words.

In Spain, MacDonald (2017) analyses errors in written argumentative texts of 304 Spanish university students of English taken from technical and humanities contexts. The findings show that grammar errors are the most frequent type of error, and that the linguistic competence of the learners has a lower than expected influence on the most types of errors coded in that corpus.

In Puerto Rico Morales-Reyes & Soler (2016) evaluates L2 learners' problems with English articles and its acquisition in 30 Spanish-speaking children learning English as a second language (SL). Their findings show that children transfer knowledge from their L1 in similar ways as adults do.

In Australia, McDowel (2016) presents a study that investigates the major error patterns in Japanese scientists' written English. Participants in the study are 13 Japanese scientists working in the field of materials science. The primary data are the participants' scientific research article manuscripts (i.e., the research articles before publication). An elaborated corpus-assisted Error Analysis (EA) methodology is employed, investigating error patterns through the lens of Systemic Functional Linguistics (SFL).

In Hong Kong, Lee et al. (2015) introduce a novel type of error-annotated learner corpus containing sequences of revised essay drafts written by non-native speakers of English. They conduct a case study on verb tenses using ANNIS, a corpus search and visualization platform.

In England, Weinberger (2002) explores an attempt to combine the principles of error analysis with the methodological devices of corpus linguistics in a computerized corpus based pilot study of errors by British learners of German. The corpus consists of approximately 28,000 words of written texts produced by first and final year students of German at Lancaster University, UK. The author designed an error taxonomy for the corpus.

In Argentina, Sánchez, Sevilla, and Bachrach (2014) did a study on a group of twenty/four native Spanish speakers on the production of gender and number agreement between head noun and predicative adjective in Spanish using an elicited-error paradigm with preambles that included either Control or Raising verbs. They analysed agreement error and omission patterns. According to their findings, the error rate was not different across the two syntactic conditions. The two verb classes were associated with feminine agreeing adjectives.

In Spain, Diez-Bedmar (2011) reports an overview of the main errors that Spanish students make when writing in English for an entrance university exam. The most interesting fact in this case is the use of *the Error Tagging Manual* version 1.1 from Louvain University, (1996). The author compares the findings from this study with previous work on EA and finds some common tendencies.

In Colombia, Vásquez (2008) analyses one written composition using four categories of error: omission, addition, misinformation and misordering. This analysis finds interference or transfer errors as well as developmental errors among others. The author concludes that L1 affects L2 learning process, syntax and meaning.

We may say that, even though there has been an increase in the work about Error Analysis, in every case researchers try to design their own methodology to analyse the types of errors they find. There is no agreement about the use of a collective error tagger for research purposes and that tendency is problematic because it is impossible to do comparative work with previous corpus tagged with different error categories and taxonomies. In Latin America, there is a big need of research on EA to gain a better understanding of the real needs and structural properties of the learners' interlanguage system to design material in agreement with their needs.

This thesis will contribute with an insight of the interlanguage of university Spanish speakers enrolled in an EFL course at a university level. It will make teachers aware of the student's main errors in written compositions and its relationship with socio-cultural factors such as stratum and gender. It is a standpoint that should be considered in the planning and designing of learners' materials and teaching strategies.

### 2.3. Criticism on Error Analysis

In this section, there is a detailed description about criticism in the early years of Error Analysis. Section 2.4 presents a counter argument supporting error analysis and a rationale on how computerized Error Analysis is nowadays, a reliable methodology.

The use of Error Analysis (EA) as a methodology to analyse learners' interlanguage has been criticized in several aspects. In the study of learner errors and EA, Ellis (1994), does a deep reasoning concerned about Error Analysis. This author makes an account of all work related to

EA and classifies its criticism in these categories: (1) Limitation on methodological procedures and theoretical problems, (2) Limitations in scope. An explanation of each aspect follows, and after that, an account of criticism from some other authors.

### 2.3.1. Limitations on methodological procedures and theoretical problems

Ellis (1994) mentioned that in the late 1960s and 1970s many EA studies did not pay attention to several aspects that are important to this kind of studies. There is a list of flaws in the procedures used in this methodology. The following is the list adapted from Ellis.

1. It was not clear the learner proficiency level (if the learner was in elementary, intermediate or advanced level)
2. Regarding the languages previously learned, there was no information
3. Those studies did not make clear the kind of learning experience (in a classroom, naturalistic or a mix)
4. There were not clear directions about the elicitation procedures used to collect the data.
5. They did not specify the medium used (written or oral), the genre (essay, letter or conversation), and the content (the topic)
6. Regarding the production, it was not certain if it was unplanned (natural, spontaneous), planned (under conditions that allow for detailed planning).

The author also discussed the distinction made by Corder (1981) about errors versus mistakes because in some cases learners use a correct target form and sometimes an incorrect non-target

form, so it cannot be concluded that the learner already knows the target language form. About covert and uncover errors, Ellis argues that some errors can be misleading because in some cases the learner may manifest target-like control of some constructions. For example, when he uses chunks, but fails to do the same when using his own utterances. It is believed that infelicities should not be considered erroneous. Regarding the description and evaluation of errors, the author argued that the description of errors is problematic. Even if the error has been identified, in some cases it is not easy to choose between two possible reconstructions. Another frequent problem by that time was the failure to quantify the different types of errors identified in a corpus because in some studies error frequencies were not given.

About the explanation of errors, Ellis (1994) argues that it is not easy to distinguish transfer or interlingual from intralingual errors. The author offers an account of work from several authors: Dulay, H; Burt, M; Krashen, (1982), Nash, Burt, & Kiparsky, (2006), among others, regarding the ambiguity of learner-errors classification and concludes that assigning an error a category is arbitrary and subject to individual biases. Concerning the evaluation of errors, Ellis argues that there are different criteria in assessing error gravity; for that reason, there are different judgements made by a native speaker (NS) and non-NSs.

### 2.3.2. Limitations in the scope of EA

One of the limitations mentioned is that EA fails to provide a complete picture of learner's language because its only focus is on errors. Bell calls EA a "pseudo-procedure" (Bell, 1974, p.39) arguing that there are other instances in which the language learner should be studied.

About this situation, Corder (1971) recognizes the importance of carefully observe the totality of learners' output. It is an option the researcher has when doing a study. Another limitation mentioned by Ellis is that most studies are cross-sectional with data collected at a single point in time, which gives a very static view of the SLA process.

In reference to other authors' critic point of view on EA, there is a group of articles edited by Startvik in 1972 and published in 1973 after a symposium on Error Analysis held in Lund, Sweden (Startvik et al. 1973). Articles in that compilation describe how in different studies researchers did EA and describe their results. In the article "The Insufficiency of Error Analysis", from that compilation, Hammarberg (1973) as previously pointed by Bell (1974), claims that EA is limited to the study of errors, but non-errors are not considered. The author also argues that it is problematic deciding whether an item is erroneous or not. Referring to the assumption made by Corder (1981), that errors reveal the teacher what to teach, Hammarberg (1973), says it is partially true because to know what to teach it is essential to analyse successful language cases or where errors tend not to occur.

### 2.4. Error Analysis Using Computerized Learner Corpora: A Counter-Argument Supporting EA

In the article: "Computer-aided error analysis" (Dagneaux, Denness, & Granger (1998) make an account of traditional error analysis with the former explained limitations and introduces a new tendency from the 1990s called Computer Learner Corpus (CLC) which are collections of natural language produced by L2 learners that are machine readable. Computerized learner data

brought sophistication to EA because it entailed the use of more complex software tools that provided automatic linguistic analyses and gave this methodology a more reliable approach based on real data.

Talking about Error Analysis in the 1970s is not the same as it is in 2018. First, previous corpora were manually compiled, and for that reason, they were composed by few samples of compositions that were not representative. Therefore, researchers could not draw completely valid conclusions. Regarding this aspect, Biber (1993) mentions “a corpus must be 'representative' in order to be appropriately used as the basis for generalizations concerning a language as a whole.” Representativeness in this case refers “to the extent to which a sample includes the full range of variability in a population” (p. 243). To guarantee corpus representativeness, in the last three decades, new technologies have given researchers invaluable help, not only for the quantity of data that can be processed, but also for the reliability in the results because statistics and cross variables are done using software that avoids human failure.

In 1999, Granger and Hung carried out the first international symposium on learner corpora in Hong Kong (Tono, 2003). By that time, most of the activities focused on the International Corpus of Learner English (ICLE) from Louvain University. From that year to 2018, there has been a growing interest in the field among researchers in SLA and EFL. The development of learner corpora was prompted by the need to have representative and authentic data samples and, this is possible using computerized corpus.

The use of authentic data representing learners' interlanguage is one of the most important features when building up a learner corpus for research purposes. Another important asset is having a representative sample to withdraw valid conclusions from its results. From the time Error Analysis started (late 1960s, early 1970s) until nowadays, much water has run under the bridge and new technologies have been in alliance to make of this methodology a reliable starting point in the analysis of learners' interlanguage.

Nowadays, elicitation procedures are designed as natural as possible and students are given several topics to choose from to obtain reliable output. Furthermore, current studies account for sociolinguistic aspects that may affect learner's input and it is important to report all conditions that surround the learning process in EFL classes.

About the use of technology, using software like *WordSmith* (Mike Scott, 2005) for example, it is quite simple to obtain statistics of error frequencies, concordances, trends, etc. Another useful tool in the analysis of collocations developed by researchers at Lancaster University is *LancsBox*. With free access online, that tool, builds collocation networks, it works with your own data, visualizes language data, automatically annotates data and you can obtain the graphics from your analysis, etc.

Regarding a biased point of view when doing error classification, it is true that in some cases, errors overlap more than one category, for that reason, this kind of analysis must be thorough and supported not only by a team of experts in the teaching of language, but also supported by native speakers with training in EFL. The development of new software to do error tagging is another

plus for EA as it is possible to have a total account of errors with the possibility to cross variables. In the 1970s, it was not possible to do that kind of work, at least not with metadata or great amount of data, but nowadays, we have the great possibility to do so.

More criticism refers to the static view of learners' interlanguage because it only shows a snapshot of the learners' output without a possibility to see a transition from different levels of performance. Once more, I would resort to new technologies that let us keep diachronic studies to see how EFL develops and evolves.

Finally, criticism pointing out that non-erroneous output should be considered; it is precisely with new technologies that the researcher has access to every error within the whole context of the learner's output.

### 2.5. The Perception of Error in Linguistics, the Teaching of Languages and in Translation

In applied linguistics, the general use of the term error refers to two main possibilities. First, referring to mistakes in spontaneous speaking, errors are related to difficulties with the timing or sequence of commands that lead to addition, deletion or substitution of sounds and morphemes.

According to the linguistic approach to this topic, there are some errors of production and perception and the border between both is not easily found. For that reason, "the term error should be used with caution especially in language acquisition studies, where it can be easily

confused with the pedagogical notion of 'error' -in the context of essay marking" (Crystal, 2008 p.173).

The second approach of errors from linguistics gives way to Error Analysis as "a technique used in language teaching and learning for identifying, classifying, and systematically interpreting the unacceptable forms produced by someone learning a foreign language, using any of the principles and procedures provided by linguistics" (Crystal, 2008, p.173). This approach, as well as the classifications given by several authors in the search of ways to systematically analyse learners' errors can be accounted as the standpoint from linguistics regarding the aspects of error in linguistics from the teaching and learning of languages. Those taxonomies facilitate the classification and subsequent counting of errors to obtain statistics, or simply, in the case of a class environment, find ways to help students avoid language flaws. Linguistics also assumes the distinction from error and mistake related to language learning where mistakes are performance limitations that the learner would be able to correct. With that point of view of errors in linguistics, we move now to the perception of errors from the teaching of languages.

In this thesis, we understand an error as "an instance of language that is unintentionally deviant and is not self-corrigible by its author" (James, 1998, p. 78), "a linguistic form or combination of forms which in the same context and under similar conditions of production would, in all likelihood, not be produced by the native speakers' counterparts" (Lennon, 1991, p.182). For the teaching of languages, errors cannot stand away from Error Analysis. They can reveal the interlanguage of students and give clues about how the acquisition process is being developed. Research on how the teaching of languages understands and perceive errors shows a

focus on the pedagogical implications and the possibility to do corrections. In this sense, according to Ribas & D'Aquino, (2003 p.79) correction refers to "the response from an educator when he perceives an error in the linguistic production from students." Studies on errors are carried out first, "to identify strategies learners use to learn a language, second, to identify the causes of learners' errors, and third, to obtain information on common difficulties in language learning" (Richards & Schmidt, 1985 p.184). Other aims from error analysis in the teaching of languages are to devise appropriate materials and improve teaching techniques that lead to the correction of errors and the avoidance of future ones.

The teaching of languages uses the concept of error in two basic aspects: firstly, to contrast the pragmatic criteria of language. Therefore, it is possible to identify how learners violate some rules of use such as cooperation, turns, silences, interruptions, overlaps. Those aspects are analysed and the level of awareness is risen, so that learners become conscious of language uses. Secondly, to contrast knowledge of cultural aspects related to the target language. Therefore, it is possible to understand physical expressions of eye contact, proximity, greetings and gestures. In this matter, the teaching of languages focuses on providing students with knowledge to avoid cultural misunderstandings. The teaching of languages perceive errors as part of the learning process. The aim is to avoid errors using pedagogical strategies, appropriate material and language awareness.

A third point of view of error is presented here in the field of translation. In this case, the perception of error is focused on how the transfer from source text (ST) to target text (TT) differs. Translation defined as "a communicative action carried out by an expert in intercultural communication (the translator), playing the role of text producer and aiming at some

communicative purpose" (Nord, 2001, p.151). The same as in any conversation, in translation, there are communicative situations, pragmatic conditions and cultural backgrounds that may differ from the ST to the TT.

A translation error is " a failure to carry out the instructions implied in the translation brief as an inadequate solution to a translation problem." (Nord, 2012. p. 184). According to Nord (2012) translation errors can be classified into four categories:

a. Pragmatic translation errors
b. Cultural translation errors
c. Linguistic translation errors
d. Text-specific translation errors" (Nord 2012, p. 184).

In functionalistic approaches and according to the *Skopos* theory introduced by Vemeer (Reib & Vemeer, 2013) "translation is an action that must have a purpose and that purpose is assigned by means of commission." (Du, 2012, p. 2191) the *skopos* in translation is defined in relation to the fulfillment of the target-text. Therefore, an error depends on the translation *Skopos* and acceptability differs depending on the theoretical orientation. Nord states that a translation error derives from a translation *Skopos* and the translation objective is shaped by the function the text accomplishes in the target culture. There should be an "intertextual coherence" defined as fidelity (Nord, 2012, p.36).

To sum up, the perspective of error in translation theories differs depending on the approach and the target expectations such as fidelity, loyalty, equivalence or norms. There are different classifications that are dependent on circumstances such as technicality, kind of revision, kind of

research process or idiomatic errors. Those classifications allow the ranking of errors that assess the quality of translations according to the purpose and situations. Therefore, the view of errors is shaped according to the perspectives from different fields. In poetry translation, for example, the translator finds a different language or "different dialect" (Corder, 1981, p.16), consequently, interpretation becomes difficult. In this case, some errors of interpretation will rise if the text is approached and analysed from the perspective of a bilingual comparison to the Standard English. In cases like this, it is necessary to approach translation according to the field of the source.

### 2.6. The First Step into Error Analysis: The Diagnosis of Errors

When we talk about diagnosis of errors, we refer to the possible sources of errors. This activity is different from the description of errors "the accurate description of errors is a separate activity from the task of inferring the sources of those errors." (Dulay, Burt, Krashen, 1982, p.145). However, description and diagnosis should not be considered separated because it is the source of errors what gives the form they take. Those are two sides of the same coin. The source is the cause and the form they (errors) take, the effect.

Searching in different sources for the recognition and description of errors in language learning, the author decided to follow the diagnosis of errors given by Corder (1981) and adapted by James (1998), since it is one of the most complete not only for the sources consulted, but also for the way it is described and organized. After errors are diagnosed, it is necessary to go in deep to its description which is maybe the most critical procedure in EA because it is at that moment when the researcher must decide in which category each error goes. In that case, if an error

overlaps two or more categories, the researcher is the one who decides in what category that error will be placed. The hardest part comes when there are hundreds of errors of the same type that overlap with others and the researcher must be consistent with every decision made in their classification. This description of errors has several purposes, firstly, it is necessary to show with evidence the existence of an error, secondly, after having the evidence, the next step is to label errors to be able to count them, and thirdly, it is necessary to create classifying categories.

For the present work, the classification of errors was done according to the error tagger from Louvain University, Error Tagging Manual Version 1.2. (Estelle et al., 2005). The classification process is described in detail in the methodology (chapter 3) from this thesis. In this section follows the diagnosis of errors. This description was done as a source of information regarding the cause of errors, but as the main focus from this thesis is on finding the most recurrent errors in the writing of EFL students in *Universidad del Norte*, therefore, there will be no analysis on the source of errors.

James, (1998) states ignorance as the main cause of errors and refers to subordinate reasons to justify the forms errors take with a specific learner or category of learner. The author goes on to explain that a "**primary diagnosis** simply explains why errors occur and a s**econdary diagnosis** ʻexplainsʼ the form that errors assume" (James, 1998, p. 177). Finding out the source of errors is crucial to go beyond the simple imputation of errors to the learners' ignorance of the TL form. Several sources give evidence about what causes errors in EFL. Let us see some of the sources given by James.

When a learner ignores an item and borrows it from the L1 the consequence will be a **transfer error.** If on the contrary, the learner knows the TL component, but is not successful to access it and decides to borrow an L1 substitute, there is an **interference mistake.** Another case is **avoidance**, it occurs when the learner ignores a TL item. In this case, the learner resorts to his L1 for a substitute item but realizes that the L1 item is not appropriate or the learner totally ignores the L1 equivalent; there is a state of double ignorance, so the learner avoids that item. In other cases, the learner could take an alternative way to express himself paraphrasing what he needs to say or instead could make a circumlocution. If the learner chooses to avoid the structure and uses a paraphrase it will become a **covert error**, but if he chooses to make a circumlocution there will be an error of **verbosity** or **vagueness**. One-way or the other, the source is the same.

James, (1998) also makes an account of the four major diagnosis-based categories of error: Interlingual, intralingual, communication-strategy, induced. When there are similar elements in L1 and L2 those items are easier to learn, therefore, in that case, learners will benefit from positive transfer, but if those items differ in meaning there will be a negative transfer or **interference errors**, classified as **interlingual errors**; the source in this case is the L1.

According to the **markedness differential hypothesis** (Eckman, 1977), it will be difficult to learn a TL form that is different from an L1 form. If for example, the TL form is marked (with special features from the TL language) while the L1 is not marked, the result will be an error. If, on the contrary, the L1 form is marked, and the L2 is not, learners will not make negative transfer to the TL, so there will not be this kind of interference error. All of this means that an

unmarked L1 form will be transferred to the TL having as a result an error. If, on the contrary, the TL has a different (marked) parameter a marked L1 form will not be transferred.

Another source of errors is caused by the target language. When learners ignore a TL form, they can use a learning strategy, or a communication strategy; in any case, both are sources of errors. Table 2 presents the different learning strategy-based errors adapted from James (1998). Every strategy is mentioned and briefly explained. All comments are adaptations from the original author.

Table 2

*Intralingual errors: learning-strategy-based errors*

| Source of error | Explanation |
|---|---|
| **False analogy (or cross-association) George's (1972)** | The learner assumes that the new item B behaves like A: the learner knows that the plural of *boy* is *boys* and assumes that child (B) behaves likewise, so pluralizes **childs.* |
| **Misanalysis** | When the learner has formed unfounded hypothesis about an L2 item. E.g. *They are carnivorous plants and *its (√ their) name comes from…*the false hypotheses here is that *its* is the s-pluralized form of it… |
| **Incomplete rule application** | This is the opposite of overgeneralization. In the example, the learners applied only two components of the interrogative formation and omitted to invert subject and verb. E.g. *Nobody knew where *was Barbie (√Barbie was)* |
| **Exploiting redundancy** | The tendency to over-elaborate the target language and to lapse into verbosity and babu. |
| **Overlooking co-occurrence restrictions** | An example of this is *I would enjoy *to learn (√learning) about America*, caused by ignorance of the |

| | |
|---|---|
| | fact that the verb enjoy should have a gerundial complement. |
| **Hypercorrection (monitor overuse)** | It is the result of the learners over-monitoring their L2 output, and attempting to be consistent, so it is like system simplification. *The seventeen year*s old girl* the learner assumes the need to pluralize year when in this case it is an adjective. |
| **Overgeneralization, or system-simplification** | This strategy leads to the overindulgence of one member of a set of forms and the underuse of others in the set. One example is the generalization of the relative pronoun that as in: *Bill,*that had a great sense of unconventional morality…*<br>*The observing qualities of Roach, *that was a great observer…*<br>These learners use that to the exclusion of *who.* |

Source: data retrieved from James (1998).

Table 3

*Intralingual errors: communication strategy-based errors*

| Source of error | Explanation |
|---|---|
| **Holistic strategies (approximation)** | Is the use of a near-equivalent L2 item previously learnt, for example the use of a near synonym. E.g. a French learner of English that substitutes the cognate **credibility* for the intended √*truth*. Another form is the use of a superordinate term **fruits* for √*blackberries*. Another option is to use an antonym or opposite *happy* for √*sad*. Another option is to **coin** a word. The previous strategies are TL-based, but there are also L1-based strategies**:** **Language switch** occurs when words from the L1 language are transferred into the TL. **Calque**: is the literal translation into L2 of the L1 |
| **Analytic strategies: Circumlocution** | These strategies express the concept indirectly, by allusion rather than by direct reference. E.g. The big…medical…thing… |

Source: data retrieved from James (1998).

There are other sources of errors, for example, **induced errors** (Stenson, 1983) refer to learner errors that are more the result of a classroom situation than from a learner's ignorance or incompetence in English grammar or L1 interference. These usually are the result of misleading information, definitions, examples, explanations etc. In many cases, this source of errors cannot be fully eradicated because many teachers know the TL imperfectly, and James (1998) affirms, this applies to native speakers and non-native speakers. These errors are different from **spontaneous** (the learner just make them) and **elicited** errors that occur when learners need to gather data. James makes an account of a variety of induced errors caused by teachers and in general derived from a classroom situation; after, he suggests some other induced errors: **materials-induced errors, teacher-talk induced errors, exercise-based induced errors; errors induced by pedagogical priorities; lookup errors; compound and ambiguous errors.**

In order to get a general picture of the whole topic, in the next table there is a snapshot of induced errors with a short explanation and examples from James, for a deeper analysis, please refer to James (1998, p. 189).

Table 4

*Induced errors prompted by teachers*

| Source of error | Explanation |
|---|---|
| **Cross-association or overgeneralization** | The learner does the mental work to draw a false analogy. E.g. the teacher introduces *worship* as a general word for pray; they know *pray* selects the preposition *to* which they assumed also applied to *worship*. The result was *worshipping *to God.* |
| **Imprecise teacher explanations** | The teacher distinguishes the modals *should* and *must* claiming the former is "stronger" than the second. The error precipitated was We **should have worked in order to buy clothes, but we *must have worked in order to eat.* |

| | |
|---|---|
| | The learners attempt to convey the idea that it is less important to spend money on clothes than on food. |
| **Directed questions** | Involves the teacher (T) directing a student A to put a question (use an interrogative) to student B.<br>**T**: Where do you live?<br>**A**: I *do live in New Street.<br>**T**: Yes, you live in New Street. Where does Laura live? Ali, ask her where she lives.<br>**A**: Where *you lives?<br>**B**: I *lives in King Street. |

Source: data retrieved from James (1998).

Table 5

*Other induced errors*

| Source of error | Explanation |
|---|---|
| **Materials-induced errors** | In general, there are errors that can be found in different course books that make the learner make erroneous assumptions. In some other cases, they have wrong spelling. |
| **Teacher-talk induced errors** | As teachers provide models of the standard TL in class, in some cases, they expose learners to nonstandard and dialect forms of the TL. Even for NS it could be a problem if they are speakers of a local nonstandard. |
| **Exercise-based induced errors** | Occurs when teachers or textbooks' input prompts error from learners because of learners being required to perform certain manipulations on bits of language. |
| **Errors induced by pedagogical priorities** | In some cases, teachers prioritize accuracy, fluency or idiomaticity. If one of them is seen as paramount, the students will feel justified in de-emphasizing the others. |

| | |
|---|---|
| | That is the case with fluency in the communicative approach. Learners have been rewarded to keep communication, but accuracy has suffered. |
| **Look-up errors** | Refers to the misuse of reference aids as dictionaries or grammars. A common one is dictionary used to translate L1 in contexts where other words would be more appropriate, but students use them because they look as novelties. |
| **Compound errors** | In this case, the source of error is attributable to more than one cause that operates either simultaneously or cumulatively. In this case, the causes are complementary. |
| **Ambiguous errors** | In this case, the possible sources are two competing diagnoses. While in the compound error the causes were complementary, in this case two causes are competing. E.g. *having to explain* vs *having explained*. The ambiguity resides in the diagnosis. |

Source: data retrieved from James (1998).

James also emphasizes Moreira's remarks (1991) about the training of FL teachers need. "FL teachers need language skills adequate to the performance of four central roles. These are the roles of **classroom manager, instructor, spontaneous communicator** and **resource person"** (Moreira in James, 1998, p.192)**.**

### 2.7. Accounting for gender, social stratum and language development in Error Analysis

Since the aim from this research is to test two hypotheses, a brief description of relevant research on those topics will be presented. Let us remember the mentioned hypothesis:

- Males and females present statistically significant differences in the median of errors from written production of English as a Foreign Language (EFL) at university level.

- There are statistically significant differences in the median of errors from written production of EFL university students in relation to the stratum classification given in Colombia (DANE, n.d.).

Regarding the first hypothesis about how male and female differ in the type and quantity of written errors there is an initial difference in the use of the terms gender and sex. In the former, "Gender is not something we are born with, and not something we have, but something, we do" (Eckert and McConnell, 2003, p.10). Concerning the latter, the same author goes on to say: "sex is a biological categorization based primarily on reproductive potential, whereas gender is the social elaboration of biological sex" (Eckert & McConnell, 2003, p.10).

Several studies carried out to find out the impact of gender in SLA have pointed to possible gender differences (Lakoff, 2003); (Romaine, 2003); (Durán, 2011). It is believed that there is a relationship between an individual's gender and some features of language, e.g. males use more nouns related to certain social and economic activities related with a topic while females tend to focus on people involved in the given topics (Ishikawa, 2015). Aliakbari and Mahjub, (2010) analysed the analytical and intuitive differences in EFL learners based on gender effect. Their findings show that male students, compared with female students, adopted a more analytical approach, while female students adopted a more intuitive approach. Saeed, Ghani, and Ramzan (2011, p.1) report a better performance from females in the learning of languages. The authors affirm, "the female students committed fewer errors of L2 writing compared to male students.", and concluded that "females can be said better language learners than males."

Babayigid, (2015, p.33) reports a significant effect on gender in favour of girls. The author reports, “girls outperformed on all dimensions of written expression, except for organization". Summing up, according to previous research, the findings show that females are better language learners than males.

Regarding the second hypothesis on how social strata are determinant factors in the development of a foreign language, social stratification/class refers to a ranking system of people by hierarchy categories. In Colombia, socioeconomic stratification is a classification in strata of residential real estate that should receive public services. It is mainly carried out to charge differentially by strata the home public services allowing assigning subsidies and collecting contributions in this area. In this way, those who have more economic capacity pay more for public services and contribute so that the lower strata can pay their bills (DANE, n.d.). Therefore, the higher the status the greater the power and wealth in comparison to other groups. That classification entails an examination on how people locate themselves and others in the social structure of inequality.

Several studies about social strata and education, (Vandrick, 1995); (Arikan, 2010); (Morales, 2017) show how schools reproduce class differences and teachers view students according to those classifications that privilege high–class learners. In EFL learning the privileged students in a higher social stratum have more opportunities to travel abroad and practice the English language learned in class. This fact could influence their final results when learning a foreign language.

Social class status might affect experiences students live and therefore the results from a privileged class would be more successful. Privileged international students, for example “will return to their countries and step into positions of power, wealth and influence. They take their privilege for granted, as privileged people everywhere tend to do” (Vandrick, 1995, P. 375).

Persuading educators to understand the consequences of social class differences that play a role in schools and affirming that: “rules laid down by the privileged classes” (Lin, 1999, p.410). The author goes on to affirm that: “the middle-class students brought with them the right kind of habitus (i.e., cultural capital). They had the correct attitudes and interest and the correct linguistic skills” (Lin, 1999, p.407). Those characteristics could shape probabilities of succeeding not only in school but also in society. Therefore, in the current scenario from this research, students that can afford to travel for fun or take a short course are able to increase their skills in the foreign language and therefore will be more successful.

### 2.8. Overview of Corpus Linguistics (CL)

In the history of Corpus Linguistics, there have been different stages of great popularity and disenchantment. Before starting this section, it is important to remember that it is not the same to talk about CL in a period with no computers or technology compared to the new era plenty of technology where data can be compiled as a computational corpus for its analysis. In this case, not only the quantity of data changes, but also the results are more reliable because there is less chance for human error.

Before and in part of the 1940s, linguists used a methodological approach to linguistics based upon observed language use. Following a basic procedure to collect data, they kept records of utterances from language to have "material" to analyze. Their analyses were supported on this methodological approach to do bottom-up language analyses. During this period, the term Corpus Linguistics (CL) was not used in any texts or studies. CL as a methodology became popular during that decade due to the rise in work that included analyses in language acquisition, the study of child language, the study of language pedagogy, work on comparative linguistics, and work to analyze syntax and semantics among others.

After a period in the late 1950s, CL stopped developing and became very unpopular. The reasons for that sudden fall in popularity were the ideas from an eminent, influential linguist, Chomsky, who strongly criticized this empiricist approach. Chomsky (1965) considered that language should be analyzed from a rationalist perspective. Rationalist theories are based on the development of a theory of mind and that is the primary source of knowledge. Their main goal is to develop a theory of language that examines how the processing of human language takes place. Chomsky's theory is related to rationalist ideas found in nativism and innatism. On the other hand, empiricism is mainly based on the observation of naturally occurring data, in this case, data compiled in a corpus.

McEnery and Wilson (2001) clearly explained Chomsky's suggestions regarding the invalidity of a corpus as a tool for linguists, since the main goal of a linguist must be to model language competence instead of performance. Competence is understood here as the internalized knowledge of a language and performance as the evidence of language competence through

external means. For Chomsky, competence is the axis that distinguishes the knowledge the speaker has of the language. If we need to describe our knowledge of language, it should be done through competence, because, according to Chomsky, performance gives us a poor perspective of language.

The following are the main ideas that sum up Chomsky's position, they were taken and/or in some way adapted to the present text, from McEnery and Wilson (2001):

1. A corpus is a collection of externalized utterances as performance data
2. The main task of a linguist is the definition of a model of linguistic competence
3. Chomsky urged a move away from empiricism towards rationalism
4. Chomsky made some assumptions: Sentences of a natural language are finite and can be collected and enumerated. This statement leads us to conclude that, according to this view, language is finite
5. The goals of linguists are not enumeration and description of performance phenomena, but rather introspection and explanation of linguistic competence.

About Chomsky's position, McEnery and Wilson (2001) conclude that first, some of Chomsky's criticisms revealed powerful verities that shape the approach that CL takes today. The problems he highlighted were to the corpus itself, but not to the approach taken by corpus. Changing those interpretations of the findings in a corpus, we have a powerful tool for linguists. They finally conclude that it is not possible to use only empirical methods to describe a language as it is impossible to eschew introspection totally.

Other problems CL faced in the pre-computational era were related to pragmatic matters. In 1965, Abercrombie referred to CL as a pseudo procedure because it is a technique hard to use "it may not be literally impossible. would be so arduous and time consuming as a way of conducting an investigation that no one in their senses would ever set out to use it" (Abercrombie, 1965, P.114-115). Other authors referred to the time-consuming efforts to do analysis and the prone-error tendency from this approach. By that time, technology was not as advanced as it is nowadays.

For all the previous reasons CL was abandoned for a period, but with the boom in technology development in the 1980s, CL had a great growth, not only in the number of works, but also in its quality, proven to be scientifically tested. Nowadays, the scene of CL is very different. CL has continued developing, and naturally observed data is still the dominant source of evidence. Those sources of empiricist data, along with the tools from new technology have given CL a push to avoid criticism on pragmatic facts.

But, after all this prelude, what is Corpus Linguistics?

There are several definitions about CL, some of them are a compilation from the first points of view on the topic and a more modern position in the field according to the evolution and changes it has taken. In simple words, Corpus Linguistics is "the study of language based on examples of 'real life' language use" (Mc.Enery & Wilson, 2001, p.1).

Another definition points corpus as:

> "An area which focuses upon a set of procedures, or methods for studying language. Corpus deals with some set of machine-readable texts. The set of texts or *corpus* deal with, is usually of a size which defies analysis by hand and eye alone. It is the large scale of data used that explains the use of machine-readable text " (McEnery & Hardie, 2011, p.1-2).

The current work abides to the concept of CL as a methodology by the following definition: Corpus refers to " a collection of machine-readable authentic texts (including samples of spoken data) that is sampled to be representative of a particular natural language or language variety" (McEnery, Xiao, & Tono, 2006, p.147). In modern days, and thanks to the advancement in the IT field, CL is not a monolithic method since all its procedures explore and describe language objectively and not as a subjective speculation.

### 2.8.1. Main Features of a Written Corpus

There are some relevant features to keep in mind when building a corpus, but they depend on the researcher's goals and the available technology. They are very briefly explained, or more developed depending on relevance, since the aim from this part is to give a general idea about corpus linguistics before focusing on learner corpus that is the one used in the present study.

Sinclair, (2004) made some other recommendations that are also applicable when compiling a learner corpus:

1. Content selection: the criteria to choose the samples, nature, dimensions of the samples. Orientation to the language or variety to be sampled
2. Criteria: mode, type of text, domain, language varieties, location of texts, date of texts
3. Sampling: define its components: written or spoken and all the major criteria that define the corpus
4. Representativeness: the sample includes the full range of variability in a population
5. Balance: the proportions of different kinds of texts should correspond with informed and intuitive judgments
6. Topic: using external criteria define language patterns and vocabulary to be included
7. Size: depends on the kind of query and the methodology used to study data
8. Specialised corpora: can contain fewer words with characteristic vocabulary
9. Homogeneity: should be given by some features of its language, it's a criterion of acceptance of a text into a corpus.

According to Parody (2008, P.104) the following are some of the main characteristics to remember:

1. Extension: depends on the intended purpose
2. Format: electronic
3. Representativeness: "the sample includes a full range of variability in a population" Biber (1993, p.243).

4. Diversity: samples should include a range of genres, according to the corpus characteristics
5. Marking or tagging: standard codes added to a text to provide information
6. Background: should be documented about the contents and decisions regarding its compilation
7. Size of the samples: should aim to maintain homogeneity
8. Classification: should be done according to discipline, theme, etc.

### 2.8.2. Corpus annotation

Corpus annotation is "the practice of adding interpretative, linguistic information to an electronic corpus of spoken and/or written language data" (Leech, 2005, p.25). The additional information consists of tags that indicate the word class, as part-of-the speech (POS), for example indicating if the word in the context is a verb, adjective, noun etc.

Apart from POS annotation, there are several kinds of annotation e.g. phonetic, semantic, and pragmatic, discourse, lexical, stylistic, among others. Apart from the previous annotation categories, the type of annotation most relevant in this work is error annotation which is concerned with the tagging of errors according to categories and types. All sorts of annotation systems add the lemmas identity within a text (Leech, 2005, p.26).

## 2.9. Learners' Interlanguage Corpora

For this thesis, we will focus on corpora for language learners, which is the axis from the present work. Its main features will be explained in the following sections.

### 2.9.1. Overview of Learner Corpora (LC)

Defined as "electronic collections of natural or near-natural data produced by foreign or second language (L2) learners and assembled according to explicit design criteria" (Granger, S, Gilquin, G, Meunier, 2015, p.1). A learner corpus can contain written or spoken output that has previously been put into an electronic format to be analysed. In some other cases, spoken learner corpora are kept in their corresponding files or in the case of multimedia, files are kept in multimodal (audio-visual) files. All data from LC contain authentic collections of contextualized language use from SLA or EFL learners in electronic format.

Learner corpora (LC) emerged in the late 1980s (Granger, et al., 2015, p.1) as a valid scientific way to analyze learner's output and has the same characteristics attributed to other corpora with a difference in the data because in this case, the source is the output from language learners. The growth of LC in the late 1980s was in part to its potential for the investigation of authentic output from students. This methodology allows researchers to have access to great amounts of data samples to search for collocations, patterns and statistics. The rising of computational LC is due mostly to the great advancements in Information Technologies (IT), which facilitate searches unachievable by hand, or without having access to computers or technology. It is the kind of undertaking perfectly suited for computer analysis.

According to the Center for English Corpus Linguistics (CECL) there are 251 learner corpus in their data base including written and spoken corpus from learners with different mother languages from around the world (UC Louvain, 2018). The International Corpus of Learner English (ICLE) is the official name of a corpus created in the early 1990s that has more than 3,000,000 (three million) words of writing from learners with at least 21 different backgrounds and the same number of language varieties. The ICLE mainly collects essays from advanced-university students of English as a foreign language in their third or fourth year of study (Granger, 2006).

The initial focus of the ICLE was on finding stylistic and quantificational differences between native speaker English and the English produced by learners. Nevertheless, from its inception the Centre for English Corpus Linguistics (CECL) has developed different versions of error taggers that allow researchers to do comparative studies about the learners' interlanguage. Learner Corpora have become the main methodology to collect data and it is the starting standpoint when doing error analysis.

### 2.9.2. Dimensions that shape Learner Corpora

According to Gilquin (2015, p.12) there are several types of corpora that differ in one or more dimensions. Some of these dimensions are features from corpora in general and some others specifically belong to learner corpora.

Follows a table with every dimension adapted from this author.

Table 6

*Dimensions that distinguish learner corpus*

| Dimensions that distinguish Learner Corpus | |
|---|---|
| Medium | Is the learner corpus built up of written texts or are they transcriptions of spoken discourse? Does it consist of multimodal material? |
| Genre | What genres are present in the corpus? Is it a combination of genres? (Argumentative essays, research papers, summaries, letter, etc.). Is the language used for general purposes or for specific purposes (LSP)? |
| Target language they represent | English, French, German, etc.? |
| Learner's mother tongue | Is the learner corpus output from learners that have a single L1, or do learners have multiple L1s? |
| Was data compiled in one or several periods of time? | Synchronic or diachronic corpus? |
| Global vs local corpus | Global corpora are part of large-scale projects. For Local corpus, teachers collect Ss. work to identify learners' needs |
| Origin and purpose | Has the corpus a commercial purpose for specific learners? |
| Local learner corpora | It suits the researchers' purposes. The researcher can have access to his learner's interlanguage. |

Source: data retrieved from Guilquin (2015).

### 2.9.3. Environment, tasks and learner variables in a Learner Corpus (LC)

According to Gilquin (2015), other aspects regarding the environment, task and learner variables can affect the corpus itself and should be taken into account. Related to the **environment**, the use of the language as a **second language** greatly differs from the use of the language as a **foreign language**. In the former, it is part of the learner's environment, therefore, it can be used in different contexts or situations from everyday life whereas in the latter, it is

used in an educational setting (in the classroom) and in few natural settings e.g. when writing to a friend abroad.

Related to **tasks**, the author mentions different tasks that involve variables in time, availability of reference tools, access to secondary sources, use of electronic devices to write the assignment or if it was handwritten, time for preparation of task, access to written notes, and it is also important to know if the task was part of an exam.

Finally, **Learner variables** are those related to the learner individually such as age, gender, country/area, mother tongue. Other variables could have incidence and are specific and relevant to the learners' environment. For example, parents' native languages, languages spoken at home, learners' proficiency level, exposure to the target language (how long has the learner studied the target language), contact with the target language in normal activities, and stays in target language countries Guilquin (2015, p.17).

### 2.9.4. Corpus collection and process features

There are several options to collect a learner corpus. It can be collected as part of an academic activity in which all students participate e.g. as an exam with its corresponding permission for the use of data. Another option is to ask students to volunteer their work if they are willing to participate. In this second option, attention must be paid not to introduce a bias considering that the most successful students would be more willing to participate than those with a low performance, and that would compromise the balance and representativeness of the data.

Regardless of the way a corpus is collected, texts in a learner corpus do not occur strictly in a natural way because they are produced in a classroom context and are the result of activities designed to improve the learners' skills in the target language. In the present work, the output collected is the result of elicitation techniques that searched for the most natural output from students. The output resulted from questions that elicited students' information or opinions from current situations that affect their daily lives. Participating students were able to choose their own words to express their opinions in their compositions. Other distinctions correspond to whether the corpus is written or spoken. As the present research is on a written corpus, from a cross-sectional study the description will focus on its main characteristics.

A written corpus can start with handwritten or typed texts. On the one hand, in the case of handwritten texts, the researcher must make sure the transcription is accurate; therefore, in the process of typing it is essential to trace the texts for any involuntary addition or loss of data. When all texts are collected, they should be coded indicating a reference and information that make them traceable. Attention must be paid to quotations that do not belong to the learners' production. Guilquin (2015, p.19) recommends to "remove quotations (which do not represent the learner's own use of language and may therefore have to be excluded from the analysis of the corpus)". In the present work, quotations were not removed to keep the entire context from errors and because in some cases, removing quotations would mean to lose fundamental parts of the text indispensable to understand the context. On the contrary, they were kept, but close attention was paid in order not to analyse those parts. On the other hand, in the case of direct computerised

versions of learners' texts, they can be kept in files as TXT texts, to make sure they can be uploaded in the most appropriate software to do the tagging process.

### 2.9.5. Annotation of Learner Corpora

Learner corpora, as any other kind of corpora start as raw texts of electronic versions or transcribed texts from spoken learner output. Van Rooy (2015, p.79) mentions three advantages of using learner corpora to do research in language teaching: size, variability and automation. **Size** refers to the amount of data that can be processed (computerised corpus allow analyses of great amounts of data). **Variability** refers to the possibility to have more individuals and more text types to include in a corpus, this advantage is also linked to the possibility to have a computerised corpus. Finally, **automation** refers to some automatic aspects of data analyses that are possible thanks to the use of information technologies (IT).

As previously mentioned in section 2.8.2., corpus annotation is "the practice of adding interpretative, linguistic information to an electronic corpus of spoken and/or written language data" (Leech, 2005, p.25). The added information comes in the form of tags, which can be defined as single entities added to one part or parts of the speech. Tags are unique and can identify features of the analysed learner corpus. There are different types of annotation and they require different tags depending on the goal from the researcher, for instance, descriptive linguistic uses Part of Speech (POS) tags to obtain grammatical annotation in a corpus.

Another example is semantic annotation that requires assigning a word a semantic field used to do refined searches and classifications according to the research purpose. Another type of annotation is used when doing Error analysis (EA). As previously mentioned, EA is a methodology to analyse learners' errors "is the process of determining the incidence, nature, causes and consequences of unsuccessful language" (James, 1998, p.1). EA uses error tags to identify errors according to various categories and types.

To do the annotation of errors it is necessary to interpret learner corpus data based on the identity of each error, not based on the assumption of what the learner wanted to say. This entails the construction of one or several target hypotheses that the researcher must test. It is impossible not to interpret data. Only through interpretation, the researcher will find ways to unhide possible hypothesis to do an essential analysis. Assigning a tag to an error means that it was the interpretation given by the researcher and that interpretation is publicly available for the reader.

For that reason, when an error tag is assigned, though there could be other interpretations, the most important is to keep uniformity in the way the taggers are used. "The usefulness of error annotated corpora depends on the consistency on the annotation" (Ludeling & Hirschmann, 2015, p.148).

Once a learner corpus is annotated, it is easier to identify and extract data to do analysis because the data is organised and ready to be used with software that permit a further analysis. To extract relevant data, there are retrieval software programs like *WordSmith, LancsBox, and Antconc* among others. Those programs allow different types of searches and retrieve only the

list of tags relevant to the research. They can also give statistics and in some cases dispersions of data. In other cases, error annotated corpora can be used to do studies that focus on a linguistic feature. For this research, the learner corpus was tagged with a standarised error taxonomy that permitted the search and counting of errors doing an analysis within their context.

### 2.9.6. Learner corpus research and the acquisition of a second or a foreign language

Second Language Acquisition (SLA) and Foreign Language Acquisition (FLA) can greatly benefit from learner-corpus research. Once a learner corpus follows the standards of collection, tagging and analysis  according to the research purpose, the corpus itself is a source of information about the interlanguage from students representing a language community. SLA researchers can analyse the learners' areas that need support to reinforce and help students achieve their goals. Learner corpora can be used to observe the transition in different stages from language learning and SLA/FLA researchers can try to find answers to several questions regarding the different aspects of language acquisition.

As the purpose of SLA/FLA theory is to better understand how learners' language develops, learner corpora (LC) can provide well documented findings. In Error Analysis, for instance, LC can give a broad picture of areas where the learner is having difficulty, or how input is transforming into acquisition. Using LC to do research in SLA/FLA constitutes an accomplishment in the use of modern technologies in the research of language acquisition that assures reliable results.

The contribution of LC in the light of the Common European Framework of Reference (CEFR) is relevant for the work with authentic learner data. LC provides concrete examples of learners' performance and progress in different proficiency levels. The CEFR provides teachers with a useful framework to assess competence in a foreign language and though it is not the perfect operationalization tool according to proficiency levels, it constitutes a good guide about the parameters to keep in mind when analysing competence. Based on the three major areas that according to CEFR have incidence in language competence: linguistic, sociolinguistic and pragmatic areas, there have already been several contributions that LC have already made e.g. corpus-based work on lexis, syntax, verb patterns, concordances, role of age in instructed learners, cognitive organization of knowledge, among others.

## 2.10. Common European Framework of Reference (CEFR) Writing Descriptors

The Common European Framework gives the foundations to elaborate the different kinds of guiding material, course design, syllabuses, textbooks or any items that could contribute to the process of language learning. The CEFR "describes in a comprehensive way what language learners have to learn to do in order to use a language for communication and what knowledge and skills they have to develop, so as to be able to act effectively" (Council of Europe, 2001, p.1). According to the CEFR, language proficiency is organised in levels to assess learners' progress in each stage of the interlanguage.

Since the CEFR is a reference of what learners should achieve to be proficient in a language, it is relevant for this research to consider those stages, especially in the writing process and in the levels that are the focus of this research.

Tables 7 to 16 present the main descriptors of proficiency in the writing ability for levels B1 and B2.

Table 7

*Common reference levels B1-B2 Global Scale*

| | | |
|---|---|---|
| Independent user | B2 | Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options. |
| | B1 | Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics, which are familiar or of personal interest. Can describe experiences and events, dreams, hopes & ambitions and briefly give reasons and explanations for opinions and plans. |

Source: data retrieved from (Council of Europe, 2001).

As the present research focus on the analysis of written work, Table 8 presents the descriptors of proficiency in the writing ability for levels A1 to C1, to give the context of all stages, but the focus from this research is on level B1-B2 that is highlighted.

Table 8

*Writing self-assessment grid for levels A1 to C2*

| | **Writing** |
|---|---|
| A1 | I can write a short, simple postcard, for example sending holiday greetings. I can fill in forms with personal details, for example entering my name, nationality and address on a hotel registration form. |
| A2 | I can write short, simple notes and messages relating to matters in areas of immediate needs. I can write a very simple personal letter, for example thanking someone for something. |
| B1 | I can write simple connected text on topics, which are familiar, or of personal interest. I can write personal letters describing experiences and impressions. |
| B2 | I can write clear, detailed text on a wide range of subjects related to my interests. I can write an essay or report, passing on information or giving reasons in support of or against a particular point of view. I can write letters highlighting the personal significance of events and experiences. |

| | |
|---|---|
| C1 | I can express myself in clear, well-structured text, expressing points of view at some length. I can write about complex subjects in a letter, an essay or a report, underlining what I consider to be the salient issues. I can select style appropriate to the reader in mind. |
| C2 | I can write clear, smoothly-flowing text in an appropriate style. I can write complex letters, reports or articles, which present a case with an effective logical structure, which helps the recipient to notice and remember significant points. I can write summaries and reviews of professional or literary works. |

Source: data from CEFR (Council of Europe, 2001).

Table 9

*Descriptors on overall written production levels B1 and B2*

| | |
|---|---|
| B1 | Can write straightforward connected texts on a range of familiar subjects within his field of interest, by linking a series of shorter discrete elements into a linear sequence. |
| B2 | Can write clear, detailed texts on a variety of subjects related to his/her field of interest, synthesising and evaluating information and arguments from a number of sources. |

Source: data from CEFR (Council of Europe, 2001).

Table 10

*Writing descriptors for creative writing, reports and essays in levels B1 and B2*

| **Creative writing** | |
|---|---|
| B1 | Can write straightforward, detailed descriptions on a range of familiar subjects within his/her field of interest.<br>Can write accounts of experiences, describing feelings and reactions in simple connected text.<br>Can write a description of an event, a recent trip – real or imagined.<br>Can narrate a story. |
| B2 | Can write clear, detailed descriptions of real or imaginary events and experiences, marking the relationship between ideas in clear connected text, and following established conventions of the genre concerned. |
| | Can write clear, detailed descriptions on a variety of subjects related to his/her field of interest.<br>Can write a review of a film, book or play. |
| Reports and essays | |
| B1 | Can write short, simple essays on topics of interest.<br>Can summarise, report and give his/her opinion about accumulated factual information on familiar routine and non-routine matters within his/her field with some confidence. |

| | |
|---|---|
| | Can write very brief reports to a standard conventionalised format, which pass on routine factual information and state reasons for actions. |
| B2 | Can write an essay or report, which develops an argument systematically with appropriate highlighting of significant points and relevant supporting detail.<br>Can evaluate different ideas or solutions to a problem. |
| | Can write an essay or report which develops an argument, giving reasons in support of or against a particular point of view and explaining the advantages and disadvantages of various options.<br>Can synthesise information and arguments from a number of sources. |

Source: data from CEFR (Council of Europe, 2001).

Table 11

*CEFR learning, teaching, assessment for vocabulary range*

| | |
|---|---|
| B1 | Has a sufficient vocabulary to express him/herself with some circumlocutions on most topics pertinent to his/her everyday life such as family, hobbies and interests, work, travel, and current events. |
| B2 | Has a good range of vocabulary for matters connected to his/her field and most general topics.<br>Can vary formulation to avoid frequent repetition, but lexical gaps can still cause hesitation and circumlocution. |

Source: data from CEFR (Council of Europe, 2001).

Table 12

*CEFR learning, teaching, assessment for vocabulary control*

| | |
|---|---|
| B1 | Shows good control of elementary vocabulary but major errors still occur when expressing more complex thoughts or handling unfamiliar topics and situations. |
| B2 | Lexical accuracy is generally high, though some confusion and incorrect word choice does occur without hindering communication. |

Source: data from CEFR (Council of Europe, 2001).

Table 13

*CEFR learning, teaching, assessment for grammatical accuracy*

| | |
|---|---|
| B1 | Communicates with reasonable accuracy in familiar contexts; generally good control though with noticeable mother tongue influence. Errors occur, but it is clear what he/she is trying to express. |
| | Uses reasonably accurately a repertoire of frequently used 'routines' and patterns associated with more predictable situations. |

| | |
|---|---|
| B2 | Good grammatical control; occasional ‘slips’ or non-systematic errors and minor flaws in sentence structure may still occur, but they are rare and can often be corrected in retrospect. |
| | Shows a relatively high degree of grammatical control. Does not make mistakes which lead to misunderstanding. |

Source: data from CEFR (Council of Europe, 2001).

Table 14

*CEFR learning, teaching, assessment for orthographic control*

| | |
|---|---|
| B1 | Can produce continuous writing, which is generally intelligible throughout.<br>Spelling, punctuation and layout are accurate enough to be followed most of the time. |
| B2 | Can produce clearly intelligible continuous writing, which follows standard layout, and paragraphing conventions.<br>Spelling and punctuation are reasonably accurate but may show signs of mother tongue influence. |

Source: data from CEFR (Council of Europe, 2001).

Table 15

*CEFR learning, teaching, assessment for sociolinguistic appropriateness*

| | |
|---|---|
| B1 | Can perform and respond to a wide range of language functions, using their most common exponents in a neutral register.<br>Is aware of the salient politeness conventions and acts appropriately.<br>Is aware of, and looks out for signs of, the most significant differences between the customs, usages, attitudes, values and beliefs prevalent in the community concerned and those of his or her own. |
| B2 | Can express him or herself confidently, clearly and politely in a formal or informal register, appropriate to the situation and person(s) concerned. |
| | Can with some effort keep up with and contribute to group discussions even when speech is fast and colloquial.<br>Can sustain relationships with native speakers without unintentionally amusing or irritating them or requiring them to behave other than they would with a native speaker.<br>Can express him or herself appropriately in situations and avoid crass errors of formulation. |

Source: data from CEFR (Council of Europe, 2001).

Table 16

*CEFR learning, teaching, assessment for coherence and cohesion*

| | |
|---|---|
| B1 | Can link a series of shorter, discrete simple elements into a connected, linear sequence of points. |
| B2 | Can use a variety of linking words efficiently to mark clearly the relationships between ideas. |
| | Can use a limited number of cohesive devices to link his/her utterances into clear, coherent discourse, though there may be some 'jumpiness' in a long contribution. |

Source: data from CEFR (Council of Europe, 2001).

Using the criteria from English Profile (UCLES/CUP, 2011), it was possible to have level descriptions for English according to CEFR. Examples of error types from levels helped to measure how learners acquire English.

The following are some error types from level A1-B1 according to the English Profile (UCLES/CUP, 2011)

Table 17

*Errors that improve on A1 to B1*

| | Error type |
|---|---|
| 1. | **Anaphor Agreement**<br>When the anaphor word is correct and the form of the anaphor is valid but wrong in the context because it does not agree grammatically with its coordinates, it is an Anaphor Agreement error. |
| 2. | **Form of Determiner**<br>When the articles 'a' and 'an' are confused. |
| 3. | **Missing Adjective**<br>When a sentence or construction requires an adjective for completeness and that adjective has been omitted, it is a Missing Adjective error. |
| 4. | **Missing Adverb**<br>When a sentence or construction requires an adverb for completeness and that adverb has been omitted, it is a Missing Adverb error. |
| 5. | **Missing Conjunction (Link Word)**<br>When a sentence or construction requires a conjunction / link word (or words) for completeness and that word has been omitted, it is a Missing Conjunction error. |
| 6. | **Missing Quantifier**<br>When a sentence or construction requires a quantifier for completeness and that quantifier has been omitted, it is a Missing Quantifier error. |
| 7. | **Inflection of Quantifier**<br>When the learner has created a feasible but non-valid inflected form of the quantifier. |
| 8. | **Replace Quantifier**<br>When a valid quantifier word in the language has been used and it is the correct part of speech but not the correct quantifier, it is a Replace Quantifier error. |
| | |

Source: table and data from (UCLES/CUP, 2011).

Table 18

*Errors that improve on B1 to B2*

| | Error type |
|---|---|
| 1. | **Derivation of Conjunction (Link Word)**<br>Where a conjunction / link word resembles, or includes the stem of, a valid word but has been incorrectly derived, usually because it has been given an incorrect affix, it is a Derivation of Conjunction error. |
| 2. | **Derivation of Determiner**<br>Where a determiner resembles, or includes the stem of, a valid determiner but has been incorrectly derived, usually because it has been given an incorrect affix, it is a Derivation of Determiner error. |
| 3. | **Form of Determiner**<br>When the articles 'a' and 'an' are confused. |
| 4. | **Inflection of Determiner**<br>When the learner has created a feasible but non-valid inflected form of the determiner, usually because of a mistaken belief that the determiner must agree in number with the noun which it precedes. |
| 5. | **Inflection of Quantifier**<br>When the learner has created a feasible but non-valid inflected form of the quantifier. |
| 6. | **Inflection of Verb**<br>When the learner has made a false assumption about whether a verb is regular or irregular and inflected it accordingly. Most commonly, the error is caused by putting regular inflections on irregular verbs. |

Source: table and data from (UCLES/CUP, 2011).

The errors described above, give an idea of how the interlanguage of students change that in elementary levels tend to be errors of omission and in the advanced levels errors in the use of structures.

Recapitulation chapter 2

This chapter started with an introduction to the most important aspects of Error Analysis (EA). It analysed its criticism and limitations. EA was examined using Computerised Learner Corpora. A framework for the diagnostic of errors was presented. This review was followed by

an overview of corpus linguistics as a methodological approach to do scientific linguistic analysis. Learner Corpus (LC) features were explained head-to-head with the acquisition of a Second or a Foreign language.

# 3. METHODOLOGY

Based on the theories from Error Analysis (Corder 1981), the research design and methodology from this thesis is presented. It is divided into three sections: **Section 3.1.** Describes Error Analysis methodology. **Section 3.2.** Gives an account of the steps followed in the corpus compilation and design. **Section 3.3.** Summarizes the steps followed for the interface design, the CLEC (Colombian Learner English Corpus). **Section 3.4** Gives an account on the design of the instrument to collect the socio-linguistic aspects from the learners.

Error analysis in Learner Corpora (LC) involves several steps that start with the choice of a research approach. On the one hand, there are deductive approaches (quantitative*)* designed to test a specific hypothesis from the learners' output. In this case, a hypothesis about the production or the source of errors can be rejected, confirmed or refined and re-tested. Results can be identified, classified, and quantified to obtain statistics. This approximation can lead to generalizations and easy research replications.

On the other hand, there are inductive approaches (qualitative) that do not search for any specific information, but examine corpora from two viewpoints: first, a hermeneutical point of view that gives sense and meaning to the findings from the learner's output. Second, a heuristic point of view, that looks for the designing of tools or strategies that contribute to the development in the field of EFL. In this approach, the observation of the learners' output with its description and interpretation is fundamental to reveal the different layers that underlie their interlanguage. In both approaches, deductive or inductive, the output should be produced in a

context as authentic as possible and should represent the most real communicative samples of the learner's interlanguage.

The present thesis in Applied Linguistics developed a deductive approach. Its main objective was to investigate the relationship between the main errors found in written compositions from students at university level and the socio-demographic factors that could have incidence in the development of the written skills **at *Universidad del Norte* in Barranquilla, Colombia.**

The data was collected using the methodology of a linguistic computational corpus, looking for a statistical profile of natural learner language from a computational perspective, and analysed according to the theories and methods of Error Analysis in order to locate, describe, count, classify, explain, interpret and reconstruct the different types of errors. There is a set of propositions set forth as an explanation for the occurrence of errors in the present study:
**1**. Male and female students greatly differ in the type and quantity of written errors. "girls outperformed on all dimensions of written expression, except for organization" (Babayigid, 2015, p.33). **2** There are statistically significant differences in the median of errors from written production of EFL university students in relation to the stratum classification given in Colombia.

To have a support in the testing process of the aforementioned hypothesis, a survey that searched the incidence of the sociolinguistic environment in students was designed. That information was qualitatively analysed and can be useful for future development of strategies or tools in EFL.

This descriptive, non-experimental research is a transversal study with data collected during the second semester of 2015. Following the methodology of corpus linguistics (CL), the written work was collected and errors were tagged using the annotation system from Louvain University version 1.2 (Estelle et al., 2005). The goal using this tagger was to obtain comparable results with similar work worldwide.

In each case, every error received an interpretation about what the learner tried to say, and the sentence was reconstructed. This part of the work was very critical as Corder mentions: " The whole success of our description of errors hinges upon the correctness of our interpretation of the learner's intentions or meaning" (Corder, 1981, p.37). In this way, interlanguage of learners can give an understanding of the learning process through the analysis of the whole learners' output.

### 3.1. Error Analysis (EA) Procedure

Corder (1967) was the first author to introduce the concept of EA. Subsequently, there have been other authors that have added new concepts to the term and claim "for a sound and systematic methodology of Error Analysis" (Dulay, Burt and Krashen, 1982, p.146). EA is a methodology that helps to obtain cross study comparisons and validation of results.

The following are the stages proposed by Corder (1981):

1. Recognition of idiosyncrasy. Or recognition of the "ill-formed" sentence
2. Accounting for a learner's idiosyncratic dialect

3. Explanation. (psycholinguistic stage) (Corder, 1981, p24).

These stages were slightly modified several years after, by James (1998) who follows the same concepts from Corder (1981), but changes the names given in each case. The present study followed those stages:

a. Recognition and reconstruction of errors
b. Description of errors
c. Explanation of errors (James, 1998, p.91).

The **first stage** refers to the identification of errors; it is to recognize their existence. You become aware of its presence, in this case, in a written text. After you detect an error, the next step is to identify its exact location. However, some errors are not easy to spot and that is the case for example of "**global errors:** Errors that significantly hinder communication and affect *overall sentence organization*" (Burt, 1975, p. 56). In this instance, whole sentences could be erroneous. Now, regarding **local errors**, parts of a sentence could be affected**:** these errors affect single elements (constituents) in a sentence and do not usually hinder communication significantly" (Burt, 1975, p. 57). Usually, these types of errors are easier to detect. In any situation, identification of errors should be done by reference to the target language (TL).

The **second stage** refers to the description of errors as a variety of one kind. For the present research, errors received a classification given with an annotation system that describes its linguistic level and its type. There are three purposes for describing errors according to James

(1998, p.96), the **first** one is to make explicit what might be tacit; in other words, we must justify our intuitions about errors that only take form when we label them as errors. The **second** purpose to describe errors is to make sure we can count them, so we can obtain statistics of error types. The **third** purpose to describe errors is to create categories that facilitate their study.

The **third stage** refers to the explanation of errors. It accounts for the explanation of the systems that are behind those errors. In this stage, errors receive an interpretation and sentences are reconstructed having in mind the possible sources of error. It also accounts for an explanation of possible processes underlying the stages of interlanguage.

For the description of errors, this research used the error tagger and the categorization of errors from Louvain University manual tagger version 1.2. (Dagneaux et al., 2005, p.4). This version of the tagging manual distinguishes between eight main categories. According to the manual, almost each of these categories are divided in sub-categories. In total, this tagger includes 56 error tags. In each case, the first letter of the tag shows the error category as it is shown in Table 19.

Table 19

*Error categories with their tags*

| Category | Code (letter that stands for error category) |
|---|---|
| 1. Form errors | F |
| 2. Grammatical errors, i.e. errors that break general rules of English grammar | G |
| 3. Lexico-grammar errors, i.e. errors where the morpho-syntactic properties of a word have been violated. | X |
| 4. Lexical errors, i.e. errors involving the semantic properties of single words and phrases. | L |
| 5. Word Redundant, Word Missing and Word Order errors. | W |
| 6. Punctuation errors. | Q |
| 7. Style errors. | S |
| 8. Infelicities. | Z |

Source: data retrieved from (E. Dagneaux et al., 2005).

The next letters in the tag give more detail about the type of error, they give details of each error's subcategory.

3567 and have some privileges like **(GVNF)** to spend $spending$

In the previous example, (GVNF) corresponds to the category of a grammar error affecting the verb and involves the use of non-finite/finite verb forms. The list of errors with their descriptions is part of the attachments from this work.

## 3.2. Corpus Design

As previously seen in **Chapter 2,** Corpus Linguistics is a methodology to collect linguistic data that follows several steps. Learner Corpus is a kind of corpus with specific, singular characteristics that require special attention in the choices of software for annotation, extraction and statistics. The following are the stages that guided this process:

1. Choice of corpus approach
2. Collection of data
3. Choice of appropriate corpus annotation
4. Choice of appropriate extraction software
5. Interpretation of data.

Every one of these stages correspond to a procedure to collect a specialised corpus, having thoughtful arguments about the criteria that governed this corpus design.

### 3.2.1. Corpus approach

Depending on the use and the kind of evidence needed, a learner corpus can be *corpus-informed;* in this approach, the corpus is used for general reference because it is a source of information Callies (2015). From a different perspective, *Corpus-based* and *corpus driven* approaches are found from a different perspective. Tognini-Bonelli makes a distinction between these two kinds of approaches:

> "Corpus-based studies typically use corpus data in order to explore a theory or hypothesis, aiming to validate it, refute it, or define it." On the other hand, corpus- driven rejects the characterization of corpus linguistics as a method; it claims that "the corpus itself should be the sole source of our hypotheses about language. It is thus claimed that the corpus itself embodies a theory of language" (2001, p.84-85).

Having in mind that the main objective from this thesis is to find out the relationship between the main errors found in written compositions of students at university level and the socio-demographic factors that could have incidence in the development of the written skills testing two hypotheses, this is a corpus-based research. This means that all the procedures follow a corpus-based approach.

### 3.2.2. Collection of data

After having the institution's permission to carry out the research, there were several stages needed to accomplish the collection process. To register in any language course at *Universidad del Norte*, students should take a placement test. This test consists on two parts: interview and online exam. The interview starts with a conversation between an EFL professor from the institution and the student. Using elicitation techniques, the interviewer starts asking basic questions about everyday life and gradually moves on to more difficult questions in structure and use. The interviewee is presented visuals and asked questions directed to find out his level and knowledge not only of structures, but also the use of English according to context (pragmatic use). This interview goes on, until the interviewee reaches his maximum oral production

threshold. It usually takes between one to 8 minutes. The second part of the exam consists of an online test that is supplied by Oxford University Press (Oxford University Press, 2017), and found in www.oxfordenglishtesting.com. After a brief registration and the introduction of a password, the student starts a one-hour test of about 100 questions that the system sorts out with different degrees of difficulty to find out the language level of the student. The online test along with the interview are used to decide in what level the student is placed. Students are classified according to their performance following the Common European Framework of Reference for Languages (Europe, 2001). Let us see the graph that accounts for the classification from the CEFR:

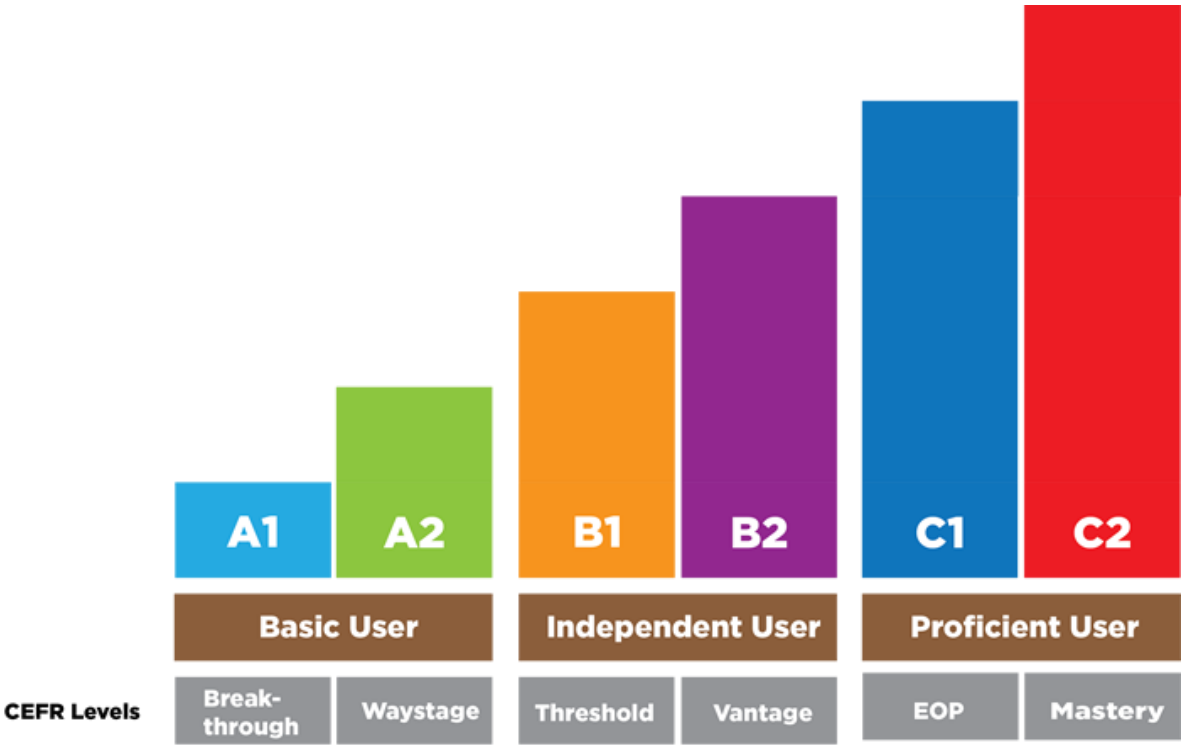



Figure 1. Retrieved from Valanglia (González, 2017).

In some cases, there are students that are between some levels; for that reason, the university has a more granular classification where students can be placed in those "gaps" to attend their

real needs. Table 20 shows the levels from the university and the corresponding CEFR classification. The table also shows number of students registered in levels B1.1 – B2.3 (highlighted) which are the levels analysed in the present research. It can be seen, a total of 2088 students were the initial population.

Table 20

*Classification from Universidad del Norte according to the CEFR*

| | Introductory Level | Level | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| U. Norte Levels | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| CEFR | A1.1 | A2.1 | A2.2 | B1.1 | B1.2 | B1.3 | B2.1 | B2.2 | B2.3 |
| # Students | 110 | 496 | 439 | 409 | 325 | 356 | 377 | 335 | 286 |
| | | | | Pre-intermediate | Intermediate | Intermediate II | | Upper-intermediate | |

**Source:** *Universidad del Norte*.

Having the placement test, the learners registered in the English Extension Program for undergraduate students. The first week of classes, according to the student's performance the teacher decides if the learner is well prepared to continue in that level or if he should be placed at a lower or a higher level. This part of the process is very important because, though the classification is a rigorous process, in some cases, due to human error or because the placement test only shows a part of the learners' abilities, students can be misplaced. Only about one in five hundred students are relocated in a new level. After that process, students start their courses that take place four hours a week for sixteen weeks, for a total of **64** hours per level, for every semester. Participating students in this study were registered in different semesters from several BA programs offered by the university: Architecture, Basic Sciences, Health Sciences, Law, Politic Sciences, International Affairs, Business School, Humanities and Social Sciences,

Engineering, Education Studies, and Mathematics. All participants share the same mother tongue: Spanish, and their average age is 23.

After all the registration process was ready and students classified, the teachers from the institution were informed about the project. In total 2088 students were registered in levels B1.1 to B2.3 and from that population 515 students signed the consent form to participate in the present research. Therefore, the present research examined in total a corpus of 515 written compositions from levels B1.1 to B2.3 according to the university classification. It is a corpus of 149,325 tokens, 12,164 types and 12,337 lemmas.

### 3.2.3. Elicitation procedures

Elicitation procedures are strategies that cause a learner to make a judgement about the adequacy of a grammatical form and make him generate a linguistic response. (Corder, 1981, p.61). The use of these procedures is necessary to obtain the state of the learner's grammar. According to Corder, (1981, p.69) there are two types of elicitation procedures: *clinical* (or open procedures) which involve the production of data of any sort using elicitation procedures such as interviews or by asking learners to write a composition. The other type of procedure is *experimental*; in this case, the researcher already has a hypothesis to prove, therefore, elicitation involves the use of special instruments designed to elicit structures or interlanguage he is interested to test. The ideal of an elicitation procedure is to generate the most natural, spontaneous language as possible.

In the case of this research, the elicitation was a clinical (open) procedure because the main objective was to find out the main errors in the written output from EFL students at university level. The elicitation and collection of data to do the analysis of errors from learners followed the procedures from a classroom environment. Based on the findings, the researcher established some propositions about the possible source of the learners' errors.

In every level from B1.1 to B2.3, students received a list of possible topics to choose from to write their compositions. Please see the attachments from this work to know about the topics, time limit, number of words per composition and the elicitation questions used to stimulate the learners' response. In all cases, the writing process was developed during the whole course within a semester period because students registered in this language courses also attended their regular classes from their Bachelor of Arts (BA). All courses were developed as seminars in which students were able to hand in their preliminary drafts for corrections. The papers analysed in this study are the output from the final work, so after several attempts and corrections during different classes, students wrote their final paper within a limited time and in some cases using some tools specified in the attachments according to the levels they were registered in.

Levels B1.1 to B2.1, in total 386 students, did the final task without an electronic device, so the papers were handed in handwritten. From levels B2.2 to B2.3, in total 129, students did the work using an electronic device and online dictionaries; they also had access to secondary sources through the Internet or databases from the university platform or through papers suggested or given by their teachers during the semester. In all cases, students had a limited amount of time to do the final task.

### 3.2.4. Corpus annotation

After the files were collected, they received different processes because they were in different formats. For instance, and because their final work was handwritten, for levels B1.1 to B2.1, the process started with the scanning followed by the typing of the texts. The typing of texts was assigned to students in their final year from the BA in languages at Universidad de Antioquia. They were given clear instructions regarding neither adding nor subtracting any words from the original handwritten compositions. After all texts were transcribed, they were thoroughly checked for mistakes and to make sure, they were exactly as the original. Next, they were converted into TXT texts to do error annotation. Students from levels B2.2 to B2.3 did directly the digital version, so those texts were immediately converted into TXT format for the error tagger. The handwritten files were in total 373 and the process of typing lasted approximately seven months. After all the previous preparation, all files were ready to start annotation.

There are two different types of annotation: Emendation and categorization (Rosen et al., 2012). In the first case, the researcher establishes one or more target hypotheses and does the correction according to the author's intention. After choosing a target hypothesis the researcher does an error categorization adding predefined tags from an error taxonomy. Error annotation relies on error taxonomies with categories for error classification given by error annotation systems. In this particular case, the researcher did each annotation following the categories from

Louvain University tagger (E. Dagneaux et al., 2005) When an error was detected, the error tag was placed just before the error and the correction followed the error in between two dollar signs: $*correction*$ as indicated by the manual. Let us analyse the following example.

Nowadays we have seen (GADJN) differents $different$ (this error corresponds to the category of grammar and it is a pluralization of an adjective).

This process was very slow because it was done manually, therefore, it needs a lot of attention to detail to first, give an appropriate error classification with the correct interpretation of errors and second, keep consistency in the error classification given in all cases. The researcher did in total three different revisions of the files, the **first** was a general annotation procedure, the **second** looked for possible missing errors that were not annotated and the **third** the researcher tried to look for consistency in all data. Next, using WordSmith software (M. Scott, 2008), all errors were extracted according to their categories, put in Excel sheets and sent to an external expert reviser to check the files for correctness and consistency in the tagging. The whole process from putting the 515 files in TXT format to have all errors tagged lasted approximately one year.

As shown before, the present corpus has 515 files, 150,262 tokens 12,871 lemmas and 12,681 types.

### 3.2.5. Data extraction software

The extraction software used in the present research is WordSmith (M. Scott, 2008). This software allows the search of patterns in language and facilitates statistics. Another software used in the extraction of data was LancsBox from Lancaster University (Brezina et al., 2015). This is a new generation software for the analysis of language data. It works with existing corpora giving statistics and the possibility to obtain graphics.

### 3.2.6. Annotation system

Choosing a classification system is not only a matter of following a theory, but a matter of methodology. The main objective for choosing a given classification is to follow a methodology that is suitable and applicable for the data. In addition, it is the only way to obtain a comparable work worldwide.

*Université Catholique de Louvain* designed the annotation system used in the present research: Error Tagging Manual Version 1.2. (E. Dagneaux et al., 2005). The criterion followed to classify errors was based on the eight error categories proposed in the manual. Those categories were mentioned in section 3.1. In some cases, some errors overlapped categories or there were doubts about the best choice to choose one classification over another. In those cases, the researcher had to decide for one of the possibilities and kept that in mind in order to be consistent in the classification of other similar errors. Those decisions could have affected the final score in some error categories, but they were necessary to keep consistency e.g.,

367 Commercials on TV aren't (GWC) honesty $honest$ for the (L.5)

In this case, the error was classified as GWC Grammar Word Class, but it could also be classified as FM Form Morphology. In those cases, the researcher tried to hold consistency tagging similar errors with the same criteria. In this way, the present research did the collection of data according to the Corpus Linguistics (CL) methodology and did the analysis of errors following the methodology of Error Analysis (EA) according to Corder (1981) in order, not only, to be consistent with the theories of EA, but also, to have a reliable system in the compilation of data.

### 3.3. Interface Design[1]

Even though there are several learner corpora and some of them from Spanish speakers learning English e.g. the Written Corpus of Learner English (WRICLE), the Santiago University Learner of English Corpus (SULEC), the Gachon Learner Corpus (GLC), the non-native Spanish corpus of English (NOSE), or the International Corpus of Learner English (ICLE), most of those corpora are not error-tagged and none of them has an interface free of charge for the academic public that allows searches within the data.

The TNT (Translation and New Technologies) research group from *Universidad de Antioquia*, in Colombia, has a collection of error-tagged files from learners of English as a Foreign Language (EFL), but it was necessary to create a tool for the academic public to, systematically, look for information and find an easy way to do searches in the error-tagged data.

[1] Description adapted from Henao, Ortega, & Tamayo, (2018)

With that purpose, the files were transferred into a database and a web application was designed to allow researchers and/or students access to the corpus and to give the possibility to look for relevant information screening out unimportant data. The main goal from this application was to do a systematic filtration of error tags and categories, classifying the tendencies from learners. The final product is a versatile friendly-user application able to present error types and possible corrections within a well-disposed and powerful system.

### 3.3.1. The CLEC description

CLEC stands for Colombian Learner English Corpus. It is a collection of 515 written compositions from university students taking their bachelor degree in *Universidad del Norte* in Barranquilla, Colombia. Participating students from the CLEC were placed in the following levels: B1.1, B1.2, B1.3-B2.1, B2.2 - B.2.3, according to the CEFR. All texts from the CLEC were collected in the second semester in 2015 and every text was error-tagged following the process explained in section 3.2.

### 3.3.2. Interface methodology

The interface design was developed using software design patterns to obtain a sturdy, friendly and supportable web application. The multi-tier architecture has several layers with an object-oriented design in which the service layer works with the programming language Python with the framework Django. This framework allows a quick and clean development with a practical design. It also includes three main strengths: speed, security and scalability. Besides that, Django

has a pattern of design: Model-View-Template (MVT) pattern, differently from Model-View-Controller (MVC) pattern. In the MVT pattern the layer processes which data will be displayed, but not how it will be displayed because the template layer will have the information on how the data will be shown.

Figure 2 shows this process:

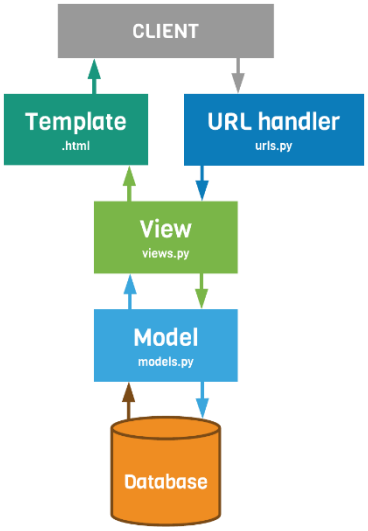



Figure 2. Architecture MTV application.

The persistence-data layer was worked using a data-based-model system (DBMS), this DBMS is focused on word documents due to the type of corpus from the present study. The DBMS allows an efficient access when doing any individual or simultaneous consultation or search to warranty the user a good experience. After having the software-patterns design, the DBMS and the documents' structure for the development, the next and main step was to do the data cleaning-up (formatting) and the indexing of data. Subsequently, a services layer was designed to allow searches and apply filters to narrow down any enquiry. From this point, a graphic interface was created considering a user-friendly design. This interface was created with

questionnaires that do basic queries required to do any search. A home page was designed to login after doing a registration. Registered-authenticated users will have several search options. Additional to this, a filter page was devised. It contains some elements from the questionnaire that allow the search of errors by level and categories or by types. The filter also allows to select options on how errors will be displayed e.g. number of tokens at each side of the error, and the quantity of errors per screen. Finally, a results page was designed to display the outcomes in a table that shows the type of error, the error in context and the correction within context.

The search filters were grouped as follows:

**Level**: the corpus was divided in 4 parts according to the English levels were students were placed. Those levels were displayed with an element that allows the possibility to select between:

- Pre-intermediate B1.1
- Intermediate B1.2
- Intermediate II B1.3-B2.1
- Upper-intermediate B2.2-B2.3

**Error type**: the eight error categories explained in section 3.1. , were put in the system creating a select-type element and a condition to display the checking boxes with the type of errors that changes according to the category. The former filters were conceived considering logic connections when displaying the checking boxes that included error sub-categories. Once a minimum product could find the whole collection stored in the database, the tests started to find possible weaknesses or fails that could affect the platform performance. When the FrontEnd and BackEnd stages finished, the final tests started using new data that simulated tests from final

users. Meanwhile, using the web server Apache the application was displayed where the modules for the framework Django and the static files were allocated using MongoDB as the database server. The mentioned tests were developed using a free tier server from AWS to allow access to both developers and tutor in real time to do the tests as soon as every function was completed.

Once the product had the approval from all members of the team, it was displayed in the server from the research group TNT from Universidad de Antioquia that will be the surrounding environment in its final production.

### 3.3.3. Resulting product

The result from the previous process, was a web responsive application that completely performs searches and does analysis on the tagged corpus of errors.

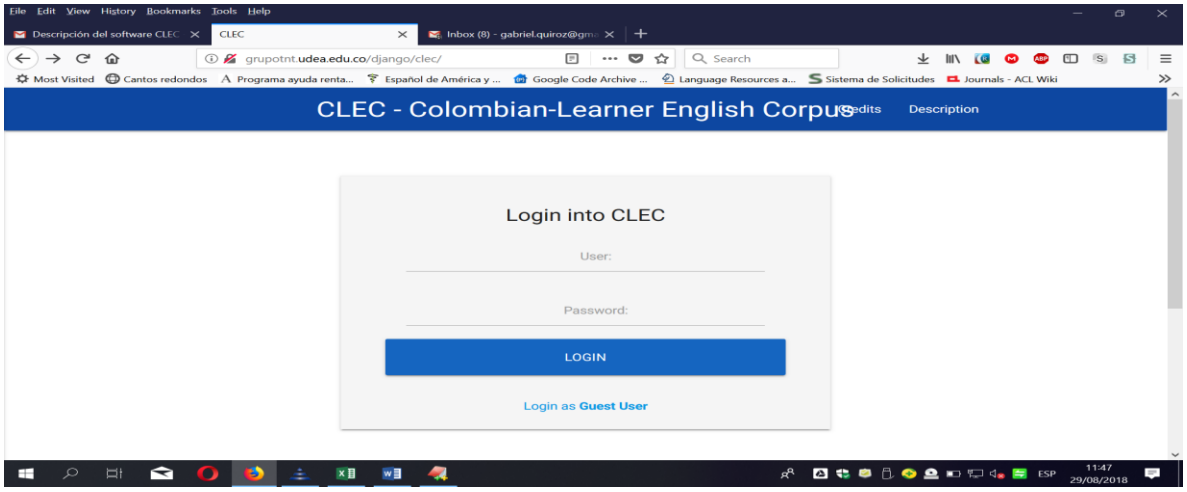



Figure 3. CLEC Interface login. Source TNT Research group.

CLEC can be found at: **URL <http://grupotnt.udea.edu.co/django/clec/>[2],** as well as help on how to use it for teaching or research purposes. CLEC[3] is fully available to the academic community under registration; guest users may have limited access to corpus results.

In Figures 4 and 5 it is possible to graphically observe the main functions from the developed application.

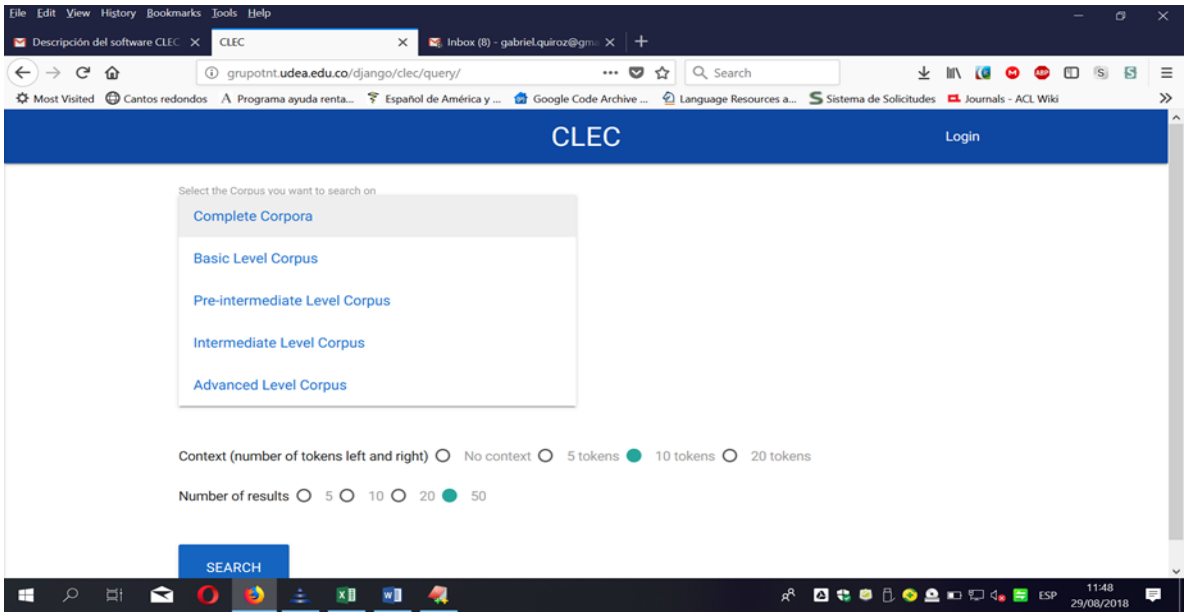


Figure 4. CLEC Interface search of level. Source TNT Research group.

Figure 4 shows the options of levels in the search menu, the number of tokens at left and right sides of the context and the number of results per hit. Errors can be found in the Corpus User Interface according to four proficiency levels: Basic level, Pre-intermediate, Intermediate, and Advanced.

[2] Sofware application was engineered by Nicolás Henao, Manuel Ortega, Antonio Tamayo.
[3] Use interface designed by Gabriel Quiroz and María Victoria Pardo

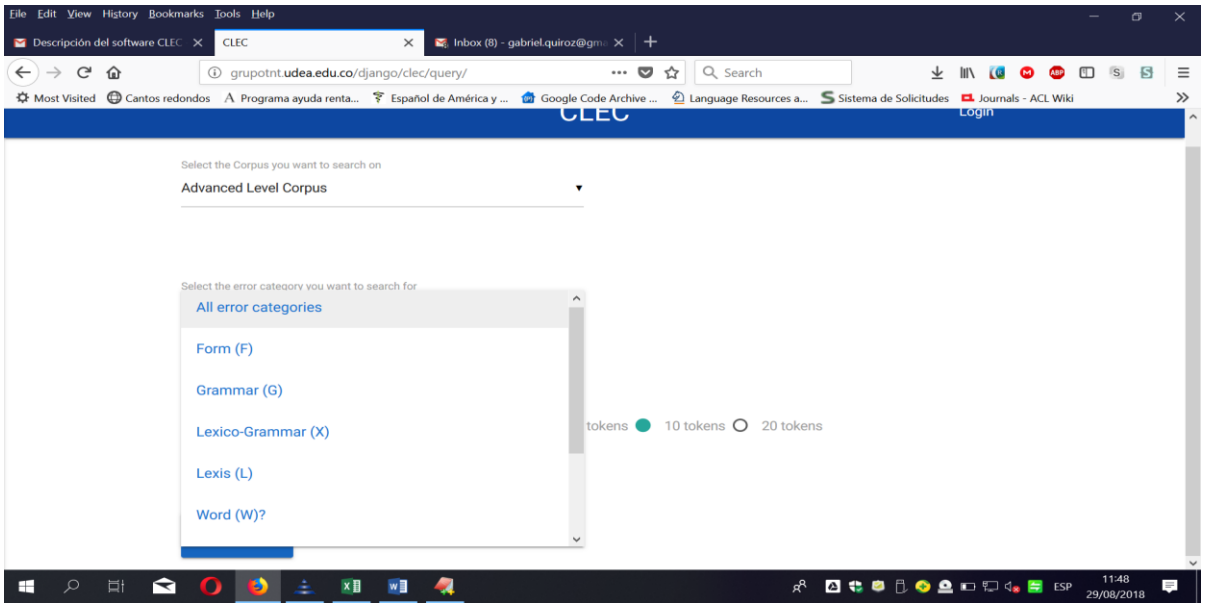



Figure 5. CLEC Interface search of level. Source TNT Research group.

Figure 5 shows how the researcher can choose the type of error to analyse. It is possible to retrieve errors from the eight error categories, individually by category, or by type of error. In order to be clear on the most useful type of database design according to the data, several sources in different web sites were consulted. It was necessary to choose between a relational and non-relational database; therefore, they recommended a possible different use for every type of database. In this case, the main consideration was regarding the way the data was presented because it was organized similarly to the way they are worked in Mongo which means, as documents. One of the concerns was to have the possibility to do searches without having to do subsequent processes with the data obtained. Using the aggregate MongoDB method was created a query that allowed the application of the required filters giving in return, separated one by one, every error as it was needed.

Several tests were carried out on the final product to try the efficiency from the performed queries and to obtain the general summary of the error frequencies in every category from the different levels.

Table 21

*Errors by categories and levels*

| | | Levels | | | | |
|---|---|---|---|---|---|---|
| | | **Pre-intermediate** | **Intermediate** | **Intermediate II** | **Upper-Intermediate** | **Totals per category** |
| **Error Categories** | F | 1043 | 553 | 119 | 220 | 1935 |
| | G | 2516 | 1347 | 261 | 2139 | 6263 |
| | X | 107 | 44 | 7 | 121 | 279 |
| | L | 893 | 504 | 79 | 1215 | 2691 |
| | W | 675 | 374 | 52 | 844 | 1945 |
| | Q | 245 | 308 | 26 | 381 | 960 |
| | S | 167 | 184 | 37 | 132 | 520 |
| | Z | 5 | 1 | 0 | 32 | 38 |
| | **Totals per level** | 5651 | 3315 | 581 | 5084 | 14631 |

In Table 21, it can be observed that the quantity of results in some searches are similar. That fact was used to carry out download-time tests using filters to achieve similar results and to assess the efficiency of the queries. The tests consisted of a repetitive search cycle, at least ten times, with 20 tokens of context and retrieving 50 errors per hit to obtain the download time from the result-page browser. Table 22 shows the average time in each case.

Table 22

*Average time in each case*

| Category | Level | Frequency | Response time (ms) |
|---|---|---|---|
| All categories | 2 | 3315 | 545.3 ms |
| L | All levels | 2691 | 585.5 ms |
| G | 1 | 2516 | 433.2 ms |
| All categories | All levels | 14631 | 1038.9ms |

Source: table from Henao et al. (2018).

Even though the filters were tried individually and together, the download times are similar. The tests also included the search without filters to show that despite having four times more results than previous tests the response time just double the average time.

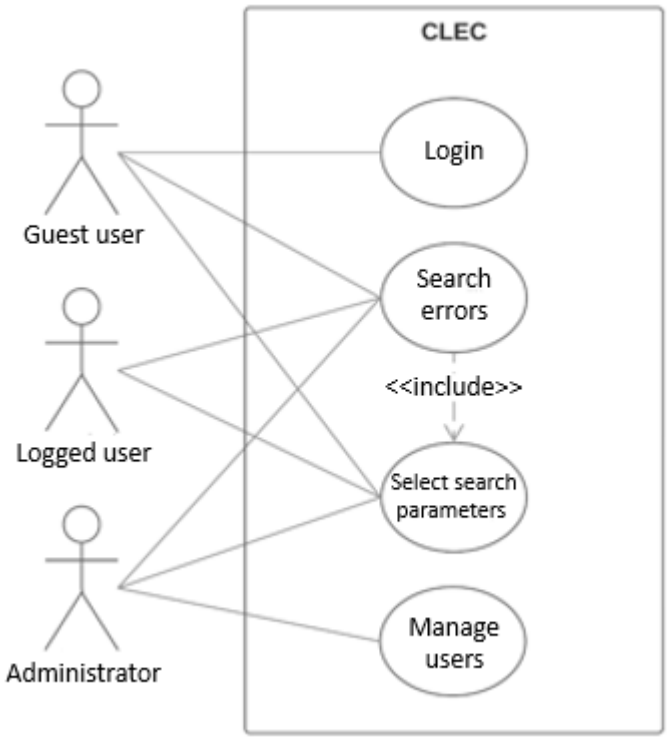



Figure 6. Graphic of roles.

Figure 6 shows the roles displayed in the application. There are three different roles: **invited** (role for users that use the platform without authentication), **user** (users who are authenticated) and finally, the **manager** (for authenticated users that belong to the CLEC team).

Additionally, to the functions observed in figure 6, the most relevant functions were considered to prioritize them accordingly in a list of developing tasks. One of the most important restrictions was on the searched context, that is, the quantity of context words on the left and on the right side.

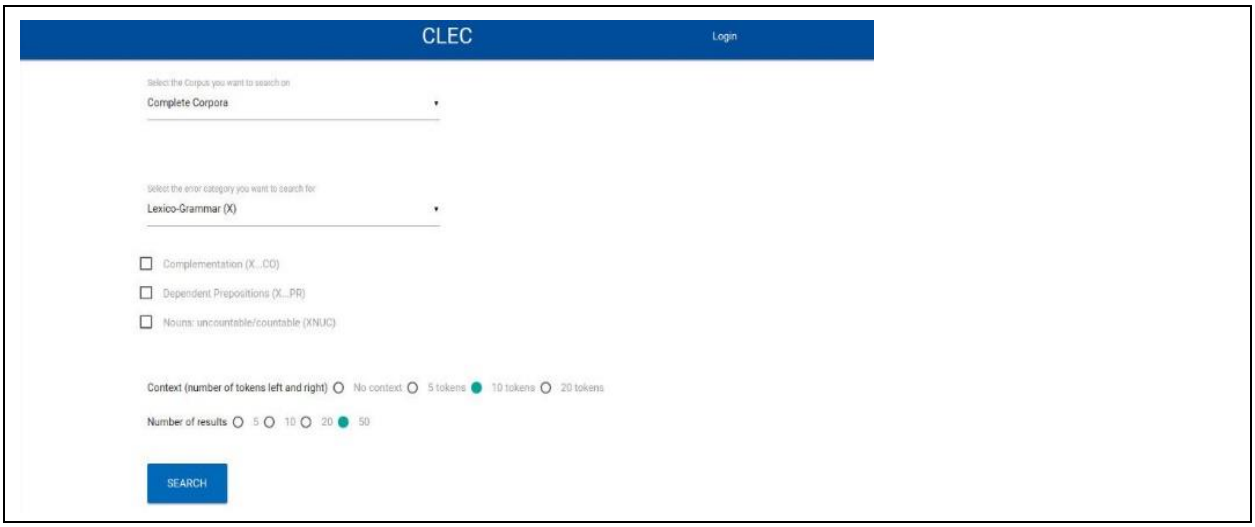



Figure 7. Graphic Interface for an invited user.

In figure 7, it is shown the interface for an invited user who would be a general user that does not need to login to use the application.

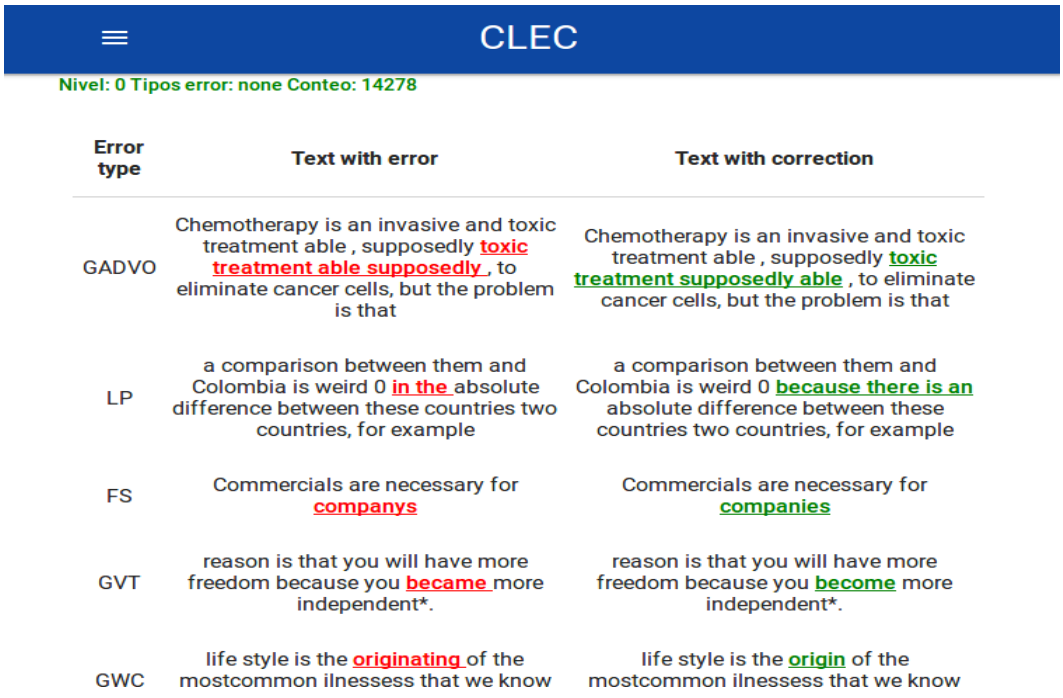



Figure 8. Response page CLEC.

In Figure 8, finally shows a sample from the response page. The top of the page shows a summary of the search that has level, type of error chosen, and the counting of results per hit. The results are presented on a table that has the type of error and the corrected context. Both the type of error and the context with the corrected error are highlighted.

### 3.4. Instrument Design. Methodology for Collection of Socio-demographic Aspects

The level of competence that English as a Foreign Language (EFL) or as Second Language (ESL) learners acquire can be explained by several factors. The role of social factors in the acquisition of proficiency can give an approach of the relationships between learning contexts and students' language acquisition. Sociocultural theory affirms that "people gain control of and reorganize their cognitive processes during mediation as knowledge is internalized during social

activity" (Spada & Lightbown, 2013, p.119). Therefore, cognitive processes are mediated by social activities that allow the mind to internalise new knowledge.

In order to establish the students' profile and the socio-demographic factors that could play a role in the learning of EFL, they answered a survey that explored their background and sociocultural and linguistic context. Context here refers to the definition given by Ellis concerning L2 learning: "the different settings in which L2 learning can take place –whether the setting is a 'natural' or an 'educational' one, and various subdivisions of these" (Ellis 1994, p.197). Social context can determine the learning opportunities that learners experience, for example, a learner that can travel abroad and practice the foreign language he is learning will have an individual experience that might influence his learning outcomes.

According to Ellis (1994), some social factors influence the levels of proficiency in an L2: age, sex, social class and ethnic identity are variables that have received the most attention in SLA research. They interact with one another and contribute to the learning process. To have a profile of the student's social aspects that could influence their learning, the instrumentation of the survey followed the criteria from SLA theorists about the most influential social factors that affect foreign language acquisition. The complete survey is available as part of the attachments from this work (please see Annex 1).

The instrument was designed and had the approval from experts in the field. It was uploaded in SurveyGizmo (*Survey Gizmo*, n.d.), which is an enterprise-level data collection platform that has intuitive software to do surveys and is user-friendly. With this platform, students had access

to the survey in any kind of electronic device they could have. Their teachers received instructions regarding the survey as well as how they should support the students in case of any difficulty.

As one of the questions asked about social strata, the teachers were told to raise students' sensitivity about the importance to truthfully answer this crucial question. The students were sent the survey that took approximately between ten to fifteen minutes to be answered. The survey was open and available to the students for two weeks.

In total, the survey included 25 questions that searched for the social context divided into three main parts:

a. Student profile (questions 1-3): Searched the profile of the students regarding their age, social stratum etc
b. Academic profile (questions 4-15): It searched all the academic learner background from elementary school. It also searched for institutions with whom students had interacted during their academic life
c. Socio-demographic aspects (questions 16-25): Searched for attitudes within the learners' culture, society, and aspects that could affect the learning process and practices of a foreign language.

In Colombia, there is a gap in the family income of students attending a private school and those in the public-school system. Education quality is different for students registered in

each case. Students from private schools usually have access to intensive EFL classes. In some cases, private schools are bilingual. According to Psacharopoulus, "private school students outperform public school counterparts on academic achievement" (Psacharopoulus, 1987, p.65).

According to sociocultural theories "the cognitive processes begin as an external socially mediated activity and eventually become internalized" (Spada & Lightbown, 2013, p.120). It is believed that the individual's internal cognitive processes are not the only ones that account for developmental process. In language learning, particularly, it is necessary to search the external social factors in the learner's environment.

### 3.4.1. Variables of study

Annex 1 considered all variables and scales used in the instrument. Table 23 presents a summary of the variables.

Table 23

*Summary of variables*

| | Variable | Nominal | Ordinal |
|---|---|---|---|
| 1. | Age | | X |
| 1. | Gender | X | |
| 2. | Strata | | X |
| 3. | Discipline of career | X | |
| 4. | Semester of enrolment | X | |
| 5. | English level | X | |
| 6. | Study of other languages | X | |
| 7. | Other languages studied | X | |
| 8. | Trips to English speaking countries | X | |

| | | | |
|---|---|---|---|
| 9. Time spent in an English-speaking country | X | | |
| 10. Purpose of trip | X | | |
| 11. Elementary studies | X | | |
| 12. Secondary studies | X | | |
| 13. Location of institution for elementary studies | X | | |
| 14. Location of institution for secondary studies | X | | |
| 15. Mother language | X | | |
| 16. Importance of English in Environment | X | | |
| 17. People in the family that speak English | X | | |
| 18. Practice of English at home | X | | |
| 19. Reason to study English | | X | |
| 20. Time devoted to listen to music in English | X | | |
| 21. Reasons to listen to music in English | X | | |
| 22. Language established in PC | X | | |
| 23. Favourable environment to learn English | X | | |
| 24. Beliefs about language learning | X | | |
| 25. Importance of English in environment | | X | |

**Recapitulation of Chapter 3**

This chapter presented an elaborated error analysis framework as the methodology followed in order to do Error Analysis. It showed how the learner corpus was collected and how the data was analysed. It described the type of software used for the extraction of data and the annotation system employed. Furthermore, it describes the interface design, which is an application that will be useful for future research. Finally, this chapter presents the results from the instrument designed to find out some aspects of the sociolinguistic aspects that could affect the acquisition of a foreign language.

# 4. ANALYSIS

The present chapter presents the following sections: **Section 4.1.** Verification of the established hypothesis. **Section 4.2.** Description of the research outcomes. **Section 4.3.** Description research outcomes **Section 4.4.** Incidence of errors by categories.

## 4.1. Verification or Proof of Previously Established Hypotheses

This research focuses on two situations:

**Situation 1:** To determine if females make less errors than males in the production of EFL written texts at university level as differences in EFL learning from male and female have been previously established by the literature.

A set of propositions were established as an explanation:

(**Null hypothesis**)

H0: Male and female, university students, do not present statistically significant differences in the median of errors from written production of English as a Foreign Language (EFL).

**(Alternative hypothesis)**

H1: Male and female present statistically significant differences in the median of errors from written production of English as a Foreign Language (EFL).

To do this test, the present research is constituted by a dichotomous variable (genre, 1-> male; 2->female) and one quantitative variable: written errors.

**Situation 2**: To determine if involvement in sociocultural factors such as travel abroad or access to private schools is related to the written production of English as a foreign language. Involvement in such leisure events is determined by the variable of socio-economic stratum.

A set of propositions were established as an explanation:

(**Null hypothesis**)

There are no statistically significant differences in the median of errors from written production of EFL university students in relation to the socio-economic strata classification given in Colombia.

**(Alternative hypothesis)**

There are statistically significant differences in the median of errors from written production of EFL university students in relation to the socio-economic stratum classification given in Colombia (DANE, n.d.).

To do this test, the present research is constituted by a qualitative variable, a component of six levels (strata from 1 to 6), and a quantitative variable: written errors.

First, it is necessary to analyse if the quantitative variable of errors follows a normal probability distribution to determine what type of test should be used to accept or reject the null hypotheses from each previous case. Table 24 presents the results from two normality tests for the variable **errors**.

Table 24

*Normality tests*

| Normality tests | | | | | | |
|---|---|---|---|---|---|---|
| | Kolmogorov-Smirnov[a] | | | Shapiro-Wilk | | |
| | Statistics | Gl | Sig. | Statistics | gl | Sig. |
| Errors | ,112 | 515 | ,000 | ,894 | 515 | ,000 |

a. Correction of significance of Lilliefors

Source: Henao et al. (2018).

Since the significance is less than 0.05 (p-value < 0.05), in both normality tests (Kolmogorov-Smirnov and Shapiro-Wilks), it is determined that the variable errors do not follow a pattern of normal distribution. For the previous reason, it is necessary to use a non-parametric test to accept or reject the null hypothesis in situations 1 and 2. Figure 9 presents the standard deviation of errors in the total corpus.

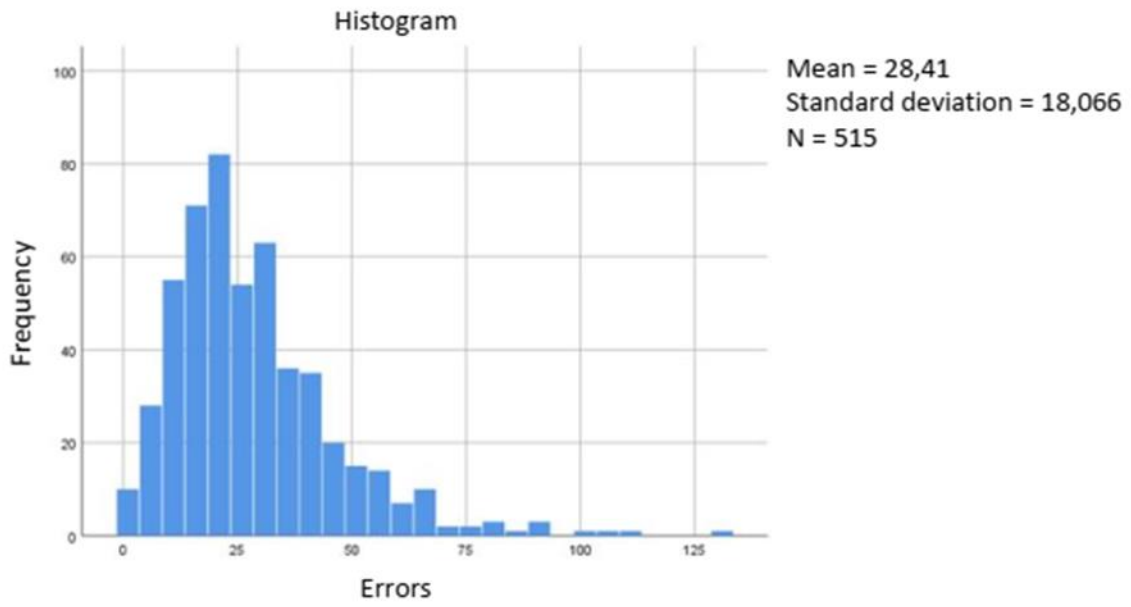



Figure 9. Histogram standard deviation of errors from 515 files. Source: Henao et al. (2018).

Note: Given that parametric tests are more reliable than non-parametric tests, the researcher tried to achieve normality in the data applying a logarithmic transformation of the variable errors, but it was not possible to achieve normality. For this reason, it is necessary to use a non-parametric test to accept or reject the null hypothesis in each case. As established before, the stated situations refer to independent samples; therefore, the U de Mann Whitney test will be used for situation 1 and the K kruskal-Wallis test for the situation 2.

**Hypothesis test of situation 1**

Since situation 1 refers to independent samples from a dichotomous variable (male-female), the U Mann Whitney test was applied. Tables 25 and 26 contain the results of the test.

Table 25

*Non-parametric test – U de Mann Whitney*

**Ranges**

| | Gender | N | Average range | Ranges addition |
|---|---|---|---|---|
| Errors | 1 | 238 | 252.21 | 60026.50 |
| | 2 | 274 | 260.22 | 71301.50 |
| | Total | 512 | | |

Source: table from Henao et al. (2018).

Table 26

*Statistics of test*[a]

| | Errors |
|---|---|
| U de Mann-Whitney | 31585.500 |
| W de Wilcoxon | 60026.500 |
| Z | -0,611 |
| Sig. asymptotic (bilateral) | 0.541 |

a. Variable of gender.

Source: table from Henao et al. (2018).

**Inference:**

Since the significance is 0.541 (p-value > 0.05) H0 is accepted. This study demonstrates, then, that there are no statistically meaningful differences in the median of written errors from male and female EFL university students.

**Hypothesis test of situation 2**

Since situation 2 refers to independent samples from six social strata, a qualitative polytomous variable, and written errors, a quantitative variable, the *K-Kruskal-Wallis* test was applied. Tables 27 and 28 contain the results of the test.

Table 27

*Non-parametric test – K Kruskal-Wallis*

**Ranges**

| | Strata | N | Average range |
|---|---|---|---|
| Errors | 0 | 324 | 240.78 |
| | 1 | 18 | 211.47 |
| | 2 | 31 | 275.02 |
| | 3 | 47 | 261.36 |
| | 4 | 51 | 310.48 |
| | 5 | 33 | 331.33 |
| | 6 | 11 | 315.73 |
| | Total | 515 | |

Source: Henao et al. (2018).

Table 28

*Tests [a,b] statistics 1*

| | Errors |
|---|---|
| H de Kruskal-Wallis | 22,553 |
| gl | 6 |
| Sig. asymptotic | 0.001 |

a. Kruskal Wallis test

b. Grouping variable: stratum

Source: Henao et al. (2018).

**Inference:**

According to the results from situation 2 where the significance is 0.001 (p-value < 0.05), the null hypothesis is rejected, and the alternative hypothesis is accepted. Therefore, it is claimed

that there are statistically significant differences in the median of errors from written production of EFL university students in relation to the stratum classification given in Colombia.

**Important clarification:** It is important to mention that 191 from 515 participants (77% from level B2 and 26% of level B1) reported their socio- economic strata to match their written work with the results from the survey on socio-demographic aspects previously done. For the previous reason, those assignments that missed the strata were given a number zero. This situation hinders the analysis process and, in some ways, contributes to the misinterpretation of the data and the testing of this hypothesis, not for the test itself, but because stratum zero does not exist, and therefore it does not correspond to the reality. Notwithstanding, it is necessary to consider that 324 out of 515 did not report their social strata.

**Exploratory analysis of the variable errors according to the socio-economic stratum**

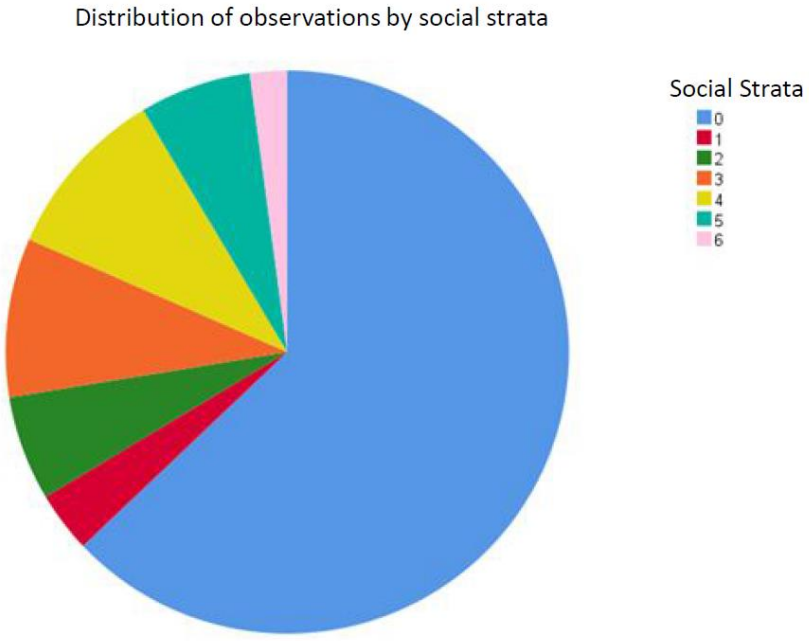


Figure 10. Distribution of observations by social strata. Source: Henao et al. (2018).

It is clearly established that there were several participants that did not report their stratum. The files tagged as stratum zero added up more than fifty percent, for this reason, this tendency might have an influenced the conclusion obtained in the K-kruskal-Wallis test. To avoid this effect, the same test is presented, but considering only the participants who confirmed their strata (in total 191 from level B2).

In Figure 11 shows the distribution of the 191 participants within the six socio-economic strata.

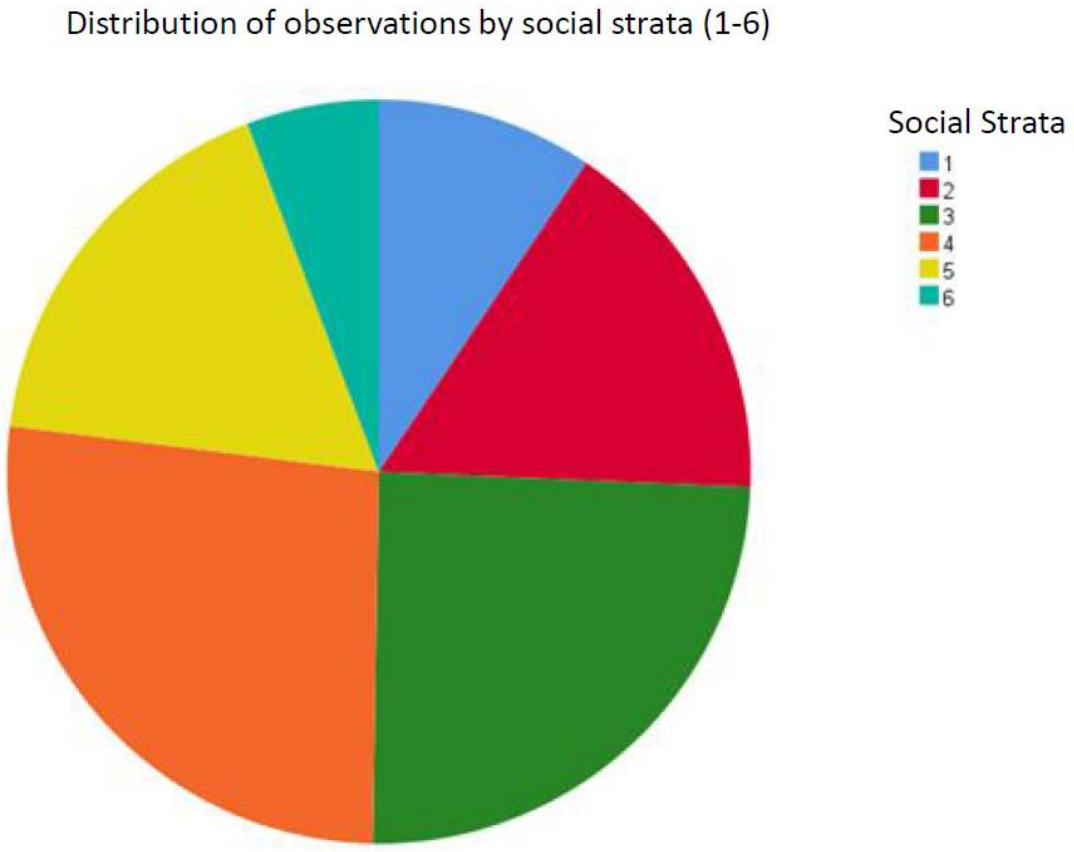



Figure 11. Distribution of observations by social strata (1-6). Source: Henao et al. (2018).

The present sample of 191 participants can be considered randomly assigned. There is evidence that all the strata levels are represented. Regarding the distribution of errors, when applying the Kolmogorov-Smirnov test to the new variable of errors with 191 participants the result is a new p-value < 0.05>, therefore, the result of the test does not reach the assumption of normality, for that reason, the non-parametric K kruskal-Wallis test was applied.

Tables 29 and 30 present the results of the non-parametric– K Kruskal-Wallis (with only the 191 individuals that confirmed their socio-economic strata)

Table 29
Ranges

| | Stratum | N | Average range |
|---|---|---|---|
| Errors | 1 | 18 | 70.11 |
| | 2 | 31 | 91.02 |
| | 3 | 47 | 87.18 |
| | 4 | 51 | 103.79 |
| | 5 | 33 | 111.38 |
| | 6 | 11 | 107.82 |
| | Total | 191 | |

Source: Henao et al. (2018).

Table 30
Tests [a,b] statistics 2

| | Errors |
|---|---|
| H de Kruskal-Wallis | 9.471 |
| gl | 5 |
| Sig. asymptotic | 0.092 |

a. Kruskal Wallis test
b. Grouping variable: social stratum

**Inference:**

Since the significance is 0.092 (p-value > 0.05), the null hypothesis is accepted. Therefore, it is claimed that there are no statistically significant differences in the median of errors from written production of EFL university students in relation to the stratum classification given in Colombia. It is important to remember that this conclusion is the result of a test in a subsample of 191 participants who confirmed their socio-economic strata, but not from the whole sample of 515 individuals that were considered for the analysis done with the U de Mann Whitney test for the two independent samples.

## 4.2. Description of research outcomes. Sociocultural variables

The participant students answered a survey with 27 questions that aimed to establish the learners' profile and the relevant sociocultural aspects that could affect their performance in the foreign language. This section informs those results.

### 4.2.1. Population description. Parts one and two of the survey

This part of the analysis describes the first two parts of the results from the survey mentioned in the methodology section (please see Annex 1). The survey was divided into three parts):

1. **Part one** searched for the students' profile that included aspects such as age, stratum, gender etc.

2. **Part two** searched for the academic profile history, status, achievements etc.
3. **Part three** searched for sociocultural aspects and beliefs that could affect performance of EFL.

During the first semester of 2015, there were 3133 students registered in EFL courses from levels A1.1 to B2.3 in *Universidad del Norte*, Barranquilla. From that group, the object of study were levels B1.1 to B2.3 that in total had 2088 individuals registered. From that group of students, 636 answered a survey about sociocultural aspects that could have incidence in the learning of EFL. Having the answers from the survey, the students' profile could be established to describe the population from this research.

**Age**: The first variable analysed refers to the age of students. The findings in the population show an even distribution of students mainly in two groups of age 15-19 and 20-25 representing 98% of the total. The distribution of the population in ages 15-19 at university level is mainly because education in Colombia has grown since 1960s. In average in Colombia students complete their Secondary School  at the age of 16.

Table 31

*Distribution of students by age*

| | Age range | Number of students | Percentage |
|---|---|---|---|
| | 15-19 | 374 | 58 |
| | 20-25 | 253 | 39 |
| | 25-30 | 9 | 0.014 |

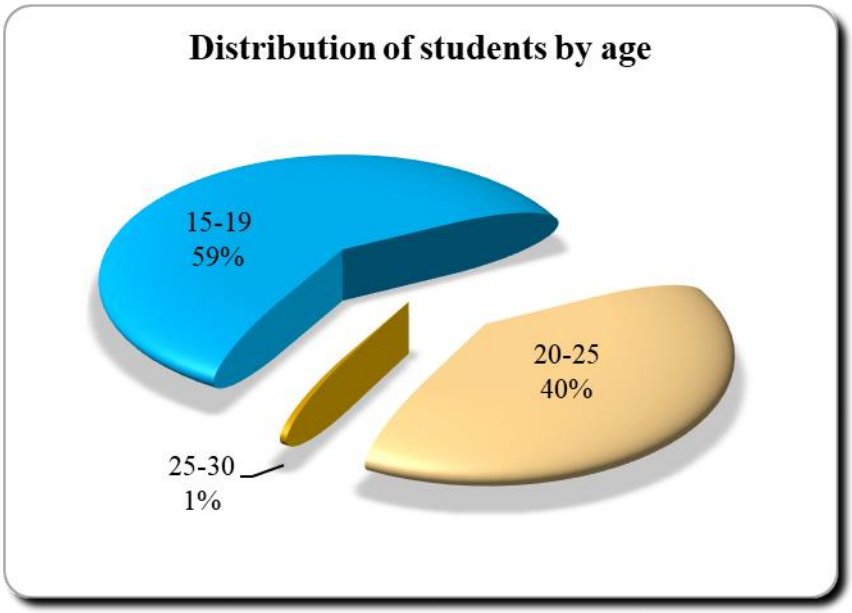



Figure 12. Account of students according to age range.

It can be seen that 59% of the students who answered the survey, that is, 374 participants were between 15 to 19 years of age. This means that most of the participants were still teenagers. In the age range from 20 to 25, 253 students made 39%. Therefore, the tendency is a population in the teen ages.

**Gender**: Regarding the gender variable females overpass males with 335, 53% in front of 301 males representing 47% of the total.

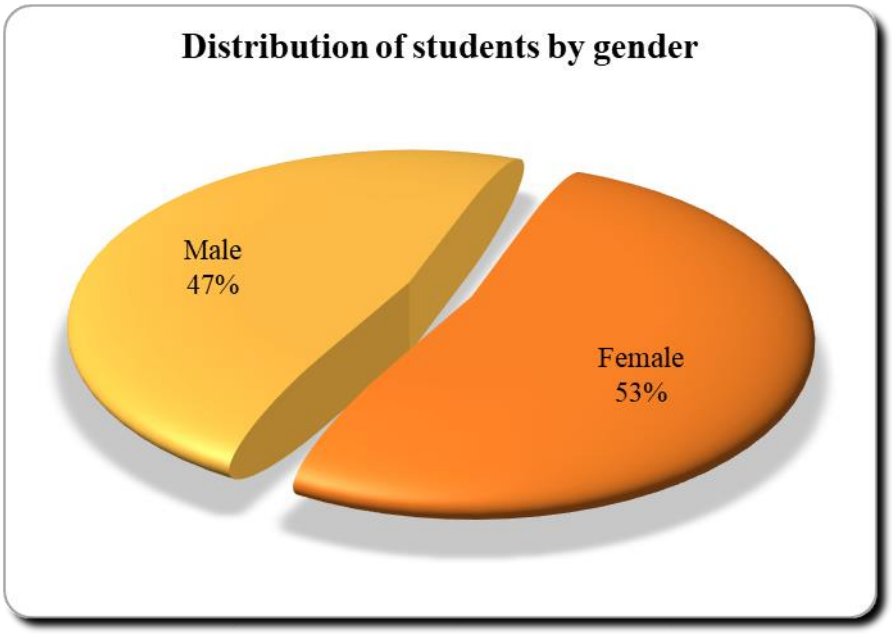



Figure 13. Account of students by gender.

**Socio-economical strata**: The population is distributed in six socio-economic strata:

Table 32

*Distribution of students by socio-economic strata*

| **Strata** | | **Number of students** | **Percentage** |
|---|---|---|---|
| Low-low | (1) | 29 | 0.045 |
| Low | (2) | 80 | 0.12 |
| Medium-low | (3) | 176 | 0.27 |
| Medium | (4) | 201 | 0.31 |
| Medium-high | (5) | 101 | 0.15 |
| High | (6) | 49 | 0.077 |

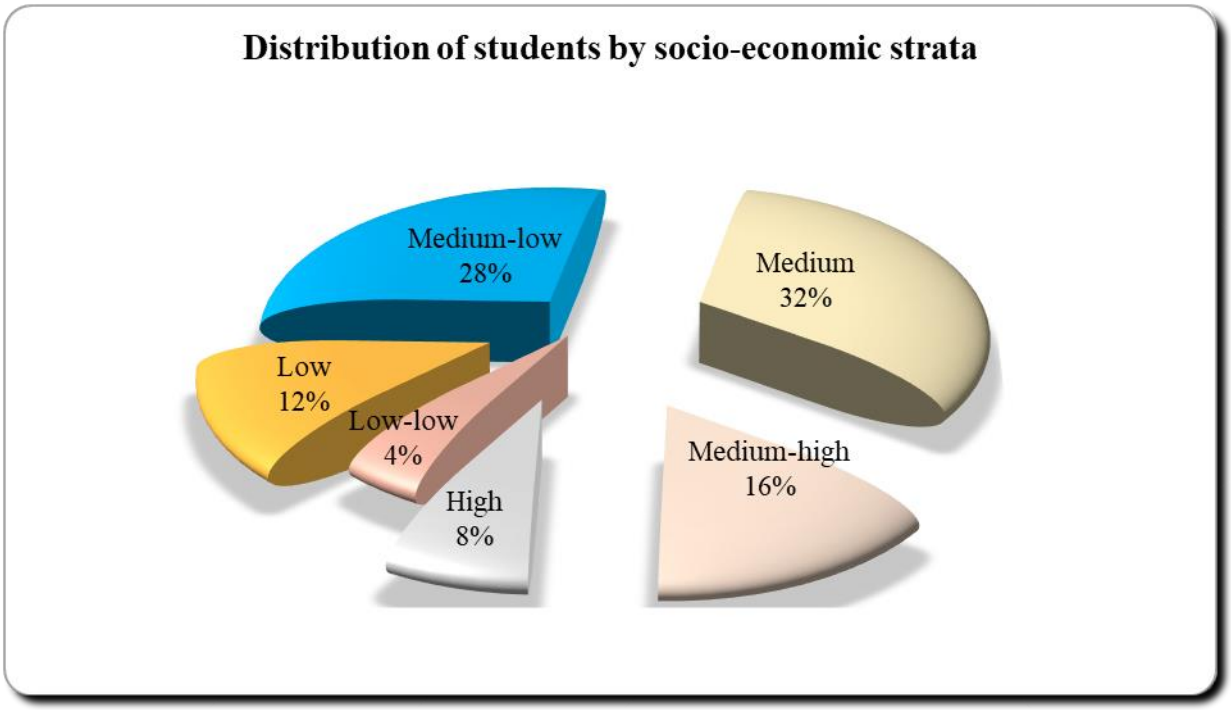



Figure 14. Account of students by strata.

In this case, most of the students, 201, that make 31% of the total, belong to stratum 4 (medium stratum) which is the middle class. After this group, comes the medium-low stratum with 176 students that make 27%, and next, the medium-high stratum with 101students that make 15%. Therefore, it can be established that the tendency is a population from the middle class that added in these three groups' accounts for 75% of the sample.

**Part 2 of the survey**: Regarding the academic aspects from the students, they were registered into nine different bachelor degrees.

Table 33

*Distribution of students by bachelor's degree*

| Bachelor's degree | Students registered | Percentage |
|---|---|---|
| Architecture, urbanism and design | 46 | 0.072 |
| Basic sciences | 1 | 0.0015 |
| Health sciences | 3 | 0.0047 |
| Law, political sciences and international relations | 53 | 0.083 |
| Institute of education | 21 | 0.033 |
| Business School | 115 | 0.180 |
| Humanities and Social sciences | 119 | 0.187 |
| Engineering | 272 | 0.42 |
| Music | 6 | 0.0094 |

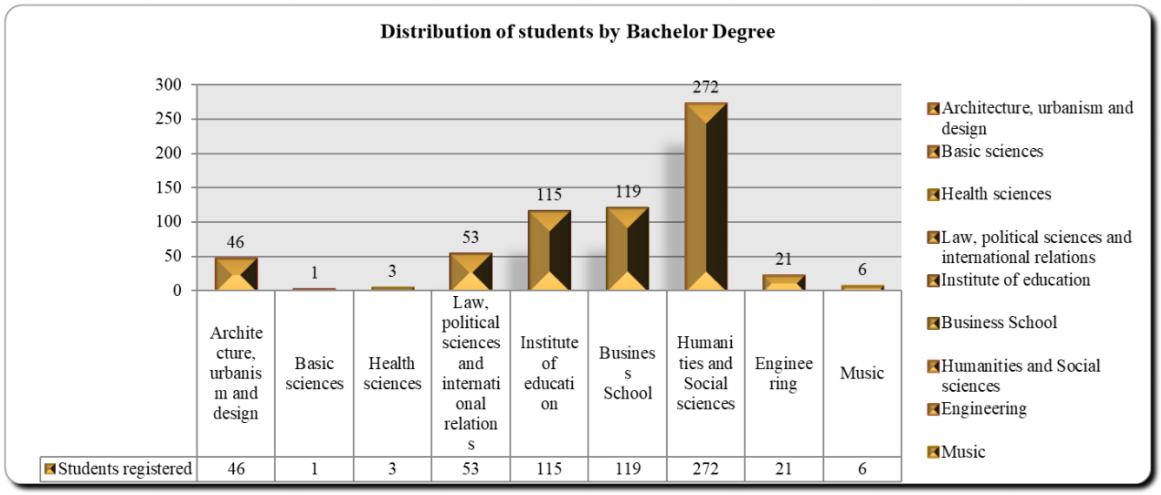



Figure 15. Account of students by bachelor's degree.

It can be observed that most of the population from the present research belongs to the Engineering programs with 42% followed by Humanities and Business with 18%, respectively.

The students from this research were enrolled in different stages of their bachelor degree (BA) from semester one to semester ten.

Table 34

*Distribution of students by semester of their BAs*

| Semester | Number of students | Percentage |
|---|---|---|
| 1 | 33 | 0.051 |
| 2 | 120 | 0.188 |
| 3 | 30 | 0.047 |
| 4 | 119 | 0.187 |
| 5 | 49 | 0.077 |
| 6 | 91 | 0.143 |
| 7 | 42 | 0.066 |
| 8 | 110 | 0.172 |
| 9 | 32 | 0.050 |
| 10 | 10 | 0.015 |

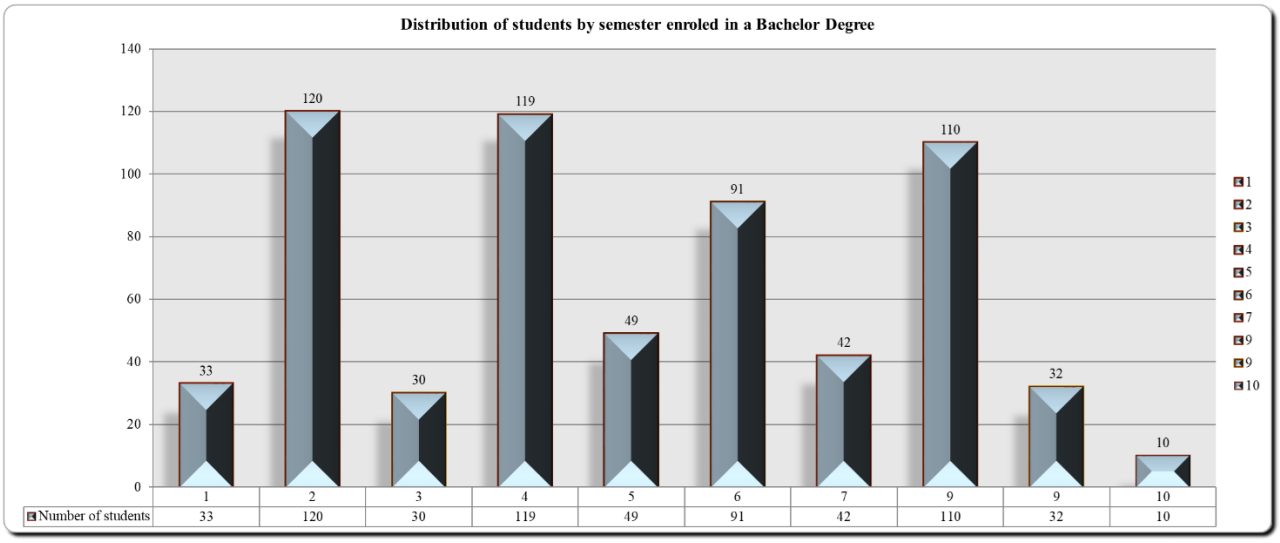



Figure 16. Account of students by semester of enrolment.

Most of the participants (120) were enrolled in the second semester of their BA followed by 119 students from semester fourth semester and 110 students from eight semester. Adding-up

students from second and fourth semesters the result shows that 37% of the students are in the first two years of their Bas.

Regarding the level of English in which students were registered, the following was established:

Table 35

*Distribution of students by EFL level enrolled*

| Level | CEFR | Number of Ss. | Percentage |
|---|---|---|---|
| Level 3 | B1.1 | 101 | 0.5 |
| Level 4 | B1.2 | 176 | 0.27 |
| Level 5 | B1.3 | 37 | 0.058 |
| Level 6 | B2.1 | 33 | 0.051 |
| Level 7 | B2.2 | 241 | 0.37 |
| Level 8 | B2.3 | 48 | 0.075 |

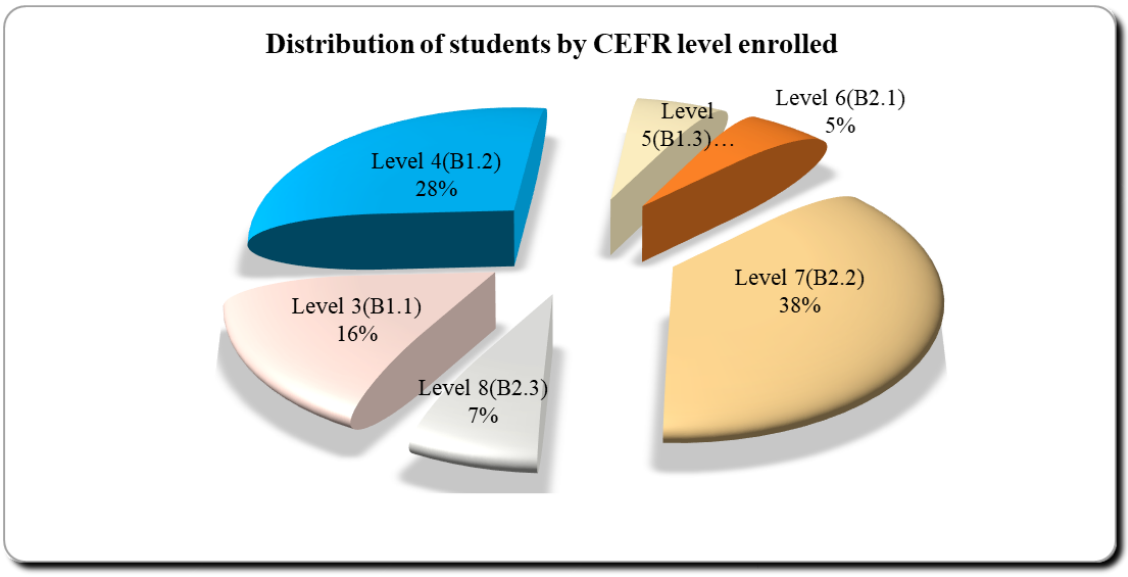



Figure 17. Account of students by CEFR level.

Most students (37%) were in level B2.2, but students from levels B1.1 and B2.2 made 47% which means that most of the population is between pre-intermediate and intermediate levels. When crossing this variable of EFL level with the semester of enrolment in the different careers, the evidence shows the following:

Table 36

*Cross variable: semester of enrolment & English level*

| | Current English level | | | | | | |
|---|---|---|---|---|---|---|---|
| | Level 3 (B1.1) | Level 4 (B1.2) | Level 5 (B1.3) | Level 6 (B2.1) | Level 7 (B2.2) | Level 8 (B2.3) | Total |
| First semester | 8 | 7 | 1 | 2 | 14 | 1 | 33 |
| Second semester | 34 | 41 | 4 | 1 | 32 | 7 | 120 |
| Third semester | 11 | 13 | 0 | 1 | 5 | 0 | 30 |
| Fourth semester | 29 | 62 | 5 | 1 | 18 | 4 | 119 |
| Fifth semester | 15 | 17 | 4 | 3 | 10 | 0 | 49 |
| Sixth semester | 3 | 23 | 15 | 9 | 34 | 7 | 91 |
| Seventh semester | 0 | 2 | 4 | 3 | 30 | 3 | 42 |
| Eighth semester | 1 | 11 | 3 | 11 | 73 | 11 | 110 |
| Ninth semester | 0 | 0 | 0 | 0 | 24 | 8 | 32 |
| Tenth semester | 0 | 0 | 0 | 2 | 1 | 7 | 10 |
| | 101 | 176 | 36 | 33 | 241 | 48 | 636 |

Regarding the students that were in different stages of their studies, it can be seen that most students from pre-intermediate level (B1.1) are in the second semester. In the case of intermediate (B1.2) most students are from fourth semester. In levels intermediate II (B1.3 and B2.1) most of the students are in fifth semester, and in the case of upper-intermediate (B2.2-B2.3) the majority of students were in eighth semester.

This evidence shows that students advance in their careers and in the level of English at the same time. Most of the students in levels B2.2 are between sixth to eighth semester and most of the students in lower levels of English are in second to fourth semester. Furthermore, the high number of students (84) in the upper-intermediate level that also are in eighth semester of the studies can be explained by the need to accomplish the mandatory levels to graduate. The data shows that when students start their BAs, most of them have not achieved the required level to graduate and for that reason, they need to take English classes at university.

Another aspect analyzed in the survey searched for other languages that students might be learning while they were learning English. From the 636 students surveyed, 45 were taking classes in other languages.

Table 37

*Students enrolled in language courses different from English*

| Language | Number of students |
|---|---|
| French | 15 |
| German | 12 |
| Portuguese | 11 |
| Mandarin | 1 |
| Italian | 6 |
| Total | 45 |

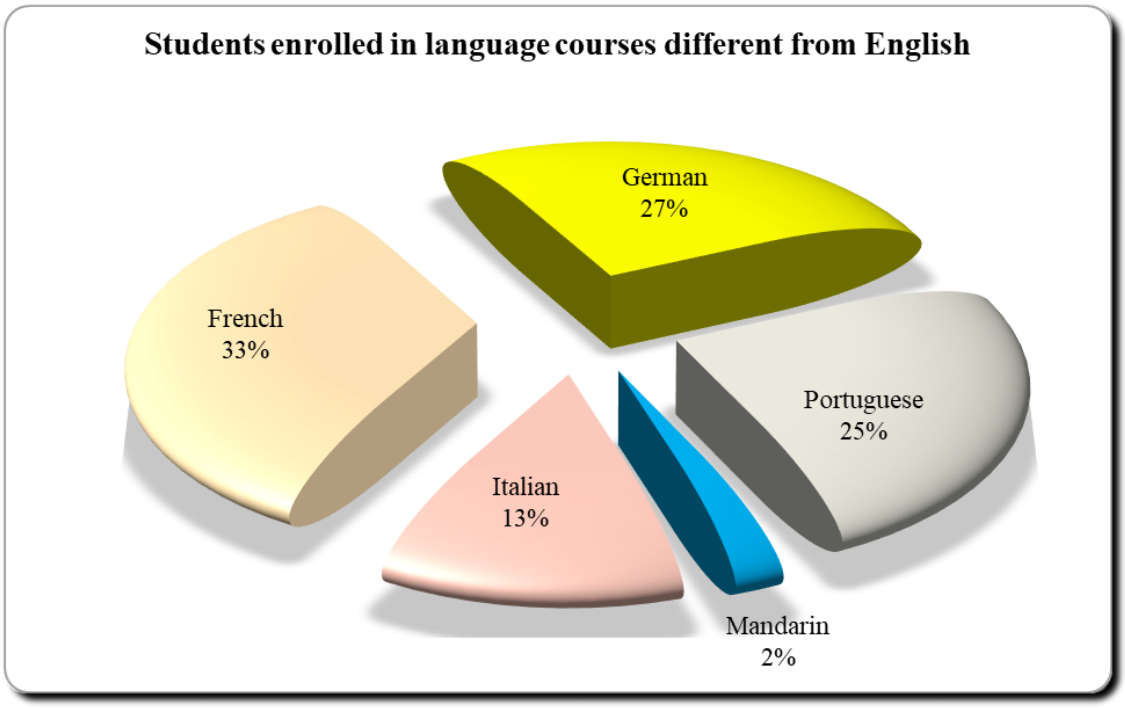



Figure 18. Account of students by enrolment in courses different from English.

Regarding the contact with the English language, students were asked if they had previously travelled to an English-speaking country, the time of stay and the reason to travel. 172 students, 27% had travelled to an English-speaking country. From that group, they stayed for different periods and for different reasons.

Table 38

*Time spent in an English-speaking country & trip purpose*

| | | Trip purpose | | | Total |
|---|---|---|---|---|---|
| | | Learn English | Other | Did not travel | |
| Time spent in an English speaking country | From 0 to 1 month | 1 | 108 | 0 | 109 |
| | From 1 to 3 months | 4 | 23 | 0 | 27 |
| | From 3 to six months | 15 | 9 | 0 | 24 |
| | From 6 months to 1 year | 7 | 5 | 0 | 12 |
| | Has not traveled to an English speaking country | 0 | 0 | 464 | 464 |
| Total | | 27 | 145 | 464 | 636 |

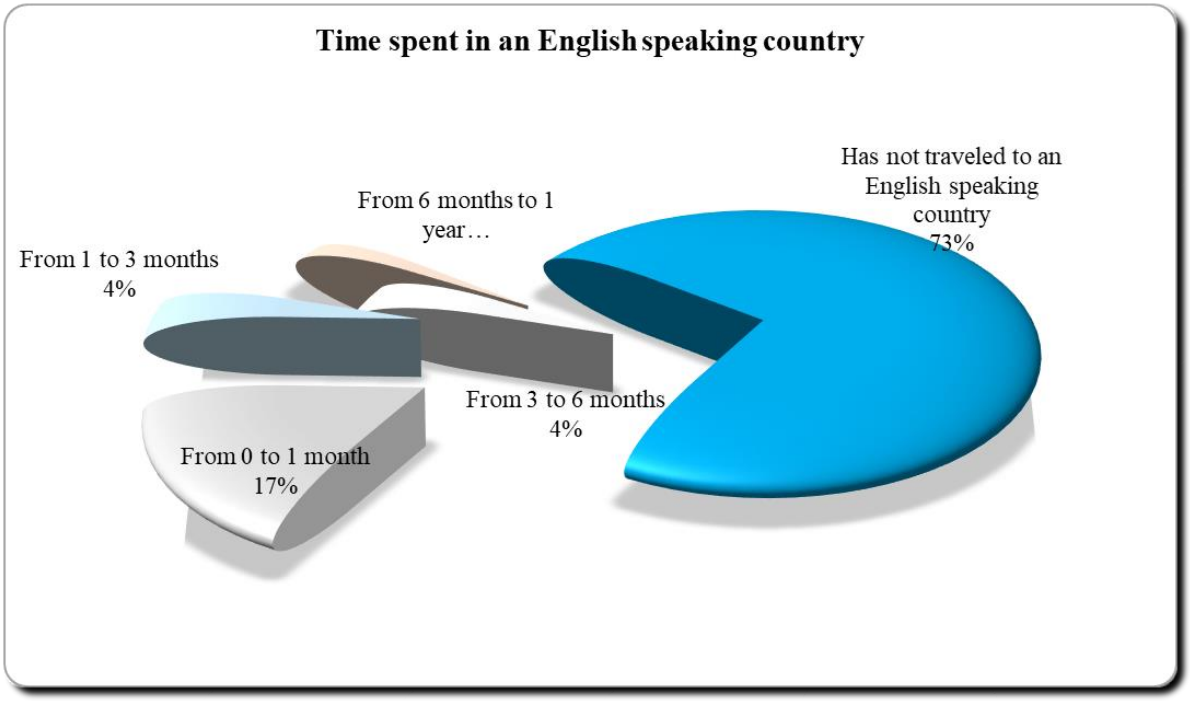



Figure 19. Time spent in an English-speaking country.

The second part of the question was the purpose to visit an English-speaking country.

Table 39

*Trip purpose*

| Trip purpose | | | | |
|---|---|---|---|---|
| | **Frequency** | **Percentage** | **Valid percentage** | **Accumulated percentage** |
| Learn English | 27 | 4.2 | 4.2 | 4.2 |
| Other | 145 | 22.7 | 22.8 | 27.0 |
| Did not travel | 464 | 72.5 | 73.0 | 100.0 |
| Total | 636 | 99.4 | 100.0 | |

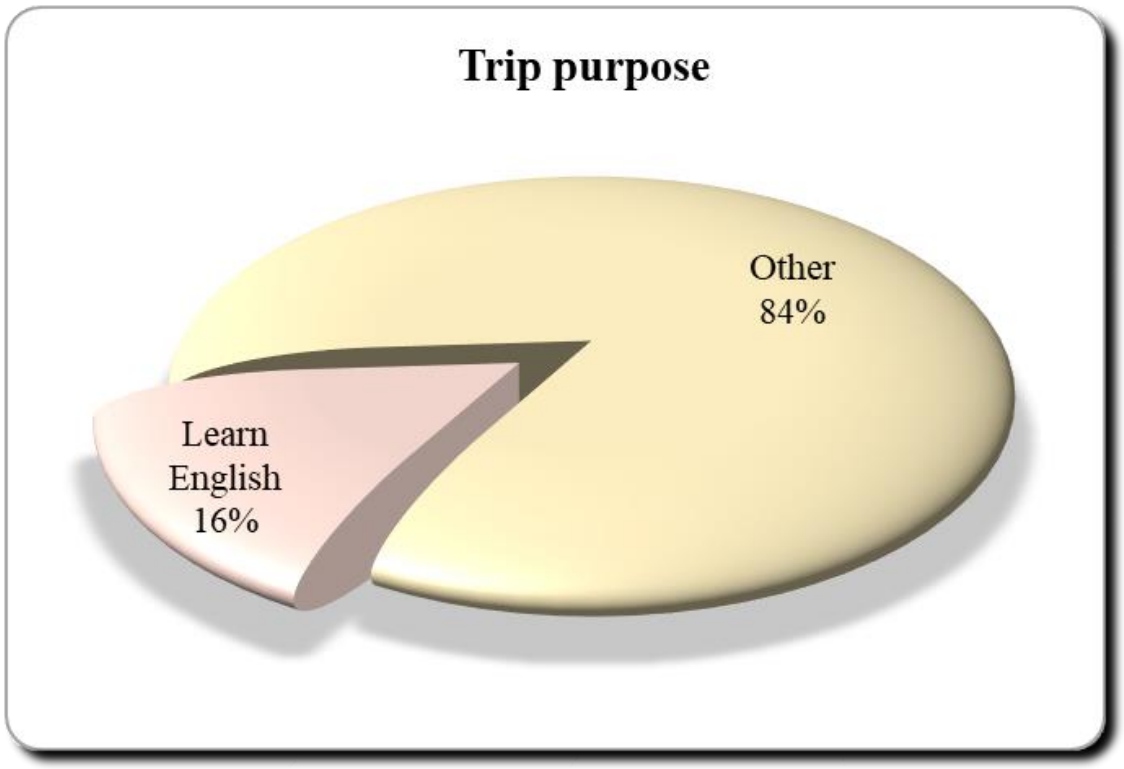



Figure 20. Trip purpose to an English-speaking country.

It can be observed that 27 from 172 travelled with the purpose to learn English, the remaining 145 had other purposes for their trips. From those 27 students 15 spent from 3 to 6 months in the English-speaking country while the 12 remaining students spent from 6 to one-year learning English in that country.

Regarding the type of institutions that students attended during primary and secondary school, the results showed that most of the students attended a private school during primary studies and a similar situation happened in the secondary school.

Table 40

*Institutions attended in elementary school*

| | Frequency | Percentage |
|---|---|---|
| Public institution | 101 | 15.9 |
| Private institution | 502 | 78.9 |
| Both, public and private | 33 | 5.2 |
| Total | 636 | 100.0 |

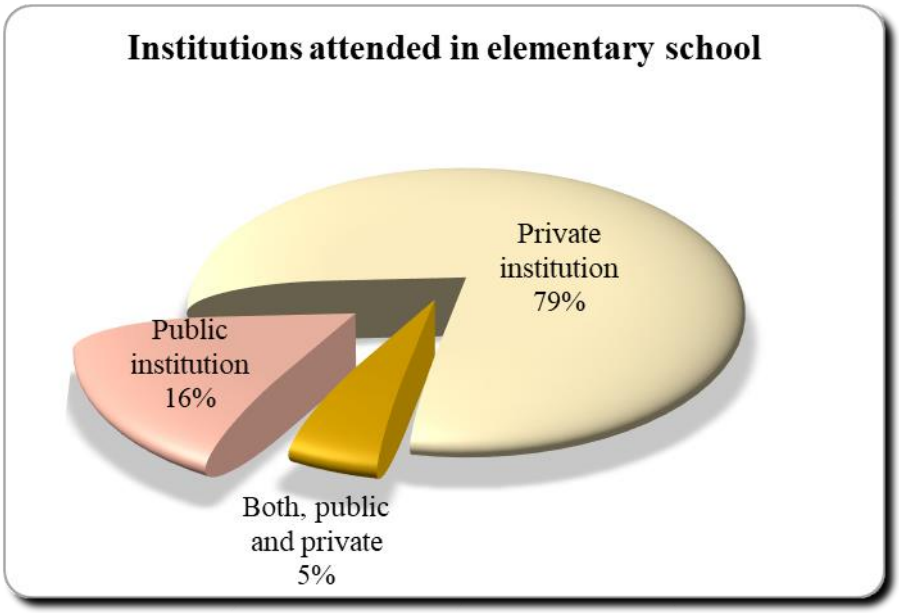



Figure 21. Institutions attended in elementary school.

Table 41

*Institutions attended in secondary school*

| | Frequency | Percentage |
|---|---|---|
| Public institution | 116 | 18.1 |
| Private institution | 489 | 76.4 |
| Both, public and private | 31 | 4.8 |
| Total | 636 | 99.4 |

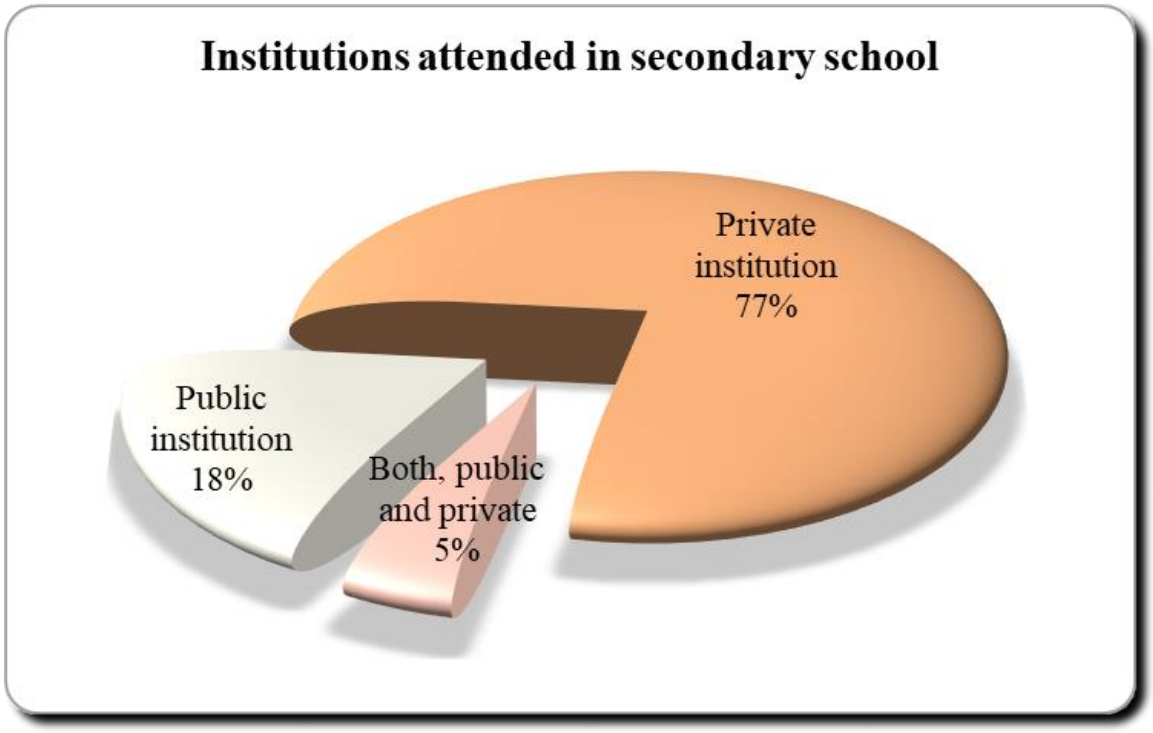



Figure 22. Institutions attended in Secondary School.

Regarding the location of the institutions attended during primary and secondary studies, students reported that most of them attended institutions in the Caribbean coast during both primary and secondary studies.

Table 42

*Location of institution where students attended elementary school*

| | Frequency | Percentage |
|---|---|---|
| Barranquilla | 368 | 57.5 |
| Santa Marta | 18 | 2.8 |
| Cartagena | 37 | 5.8 |
| Valledupar | 31 | 4.8 |
| Another city in the coastal area | 155 | 24.2 |
| Another city outside the coastal area | 27 | 4.2 |
| Total | 636 | 99.4 |

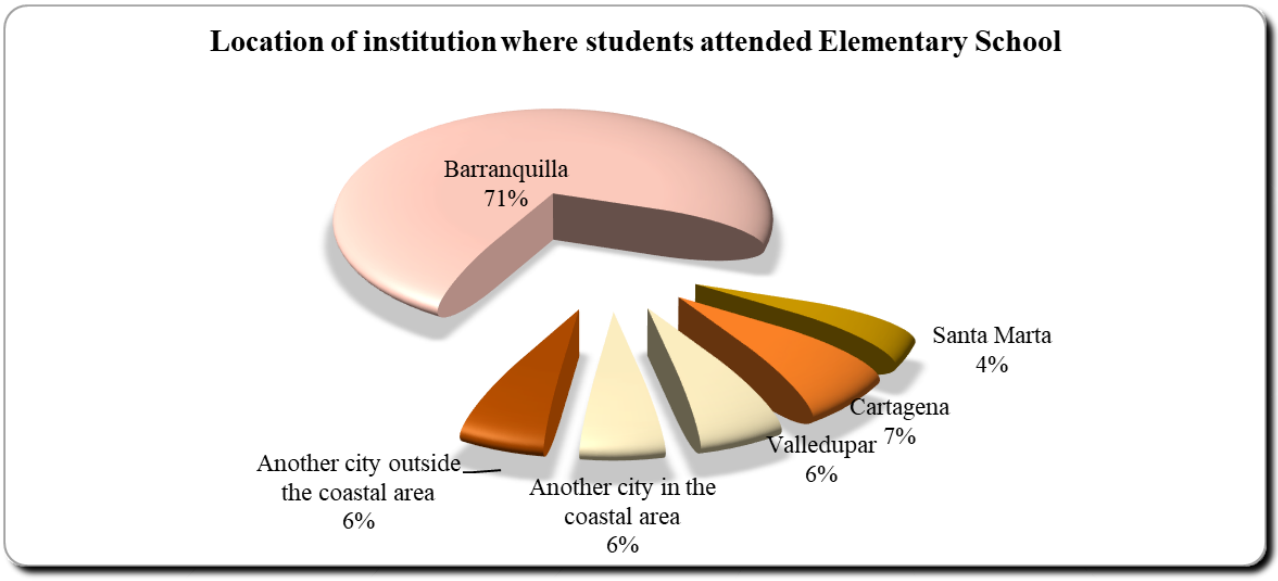



Figure 23. Location of institutions attended in elementary school.

Table 43

*Location of institution where students attended secondary school*

| | Frequency | Percentage |
|---|---|---|
| Barranquilla | 393 | 61.4 |
| Santa Marta | 21 | 3.3 |
| Cartagena | 27 | 4.2 |
| Valledupar | 36 | 5.6 |
| Another city in the coastal area | 134 | 20.9 |
| Another city outside the coastal area | 25 | 3.9 |
| Total | 636 | 99.4 |

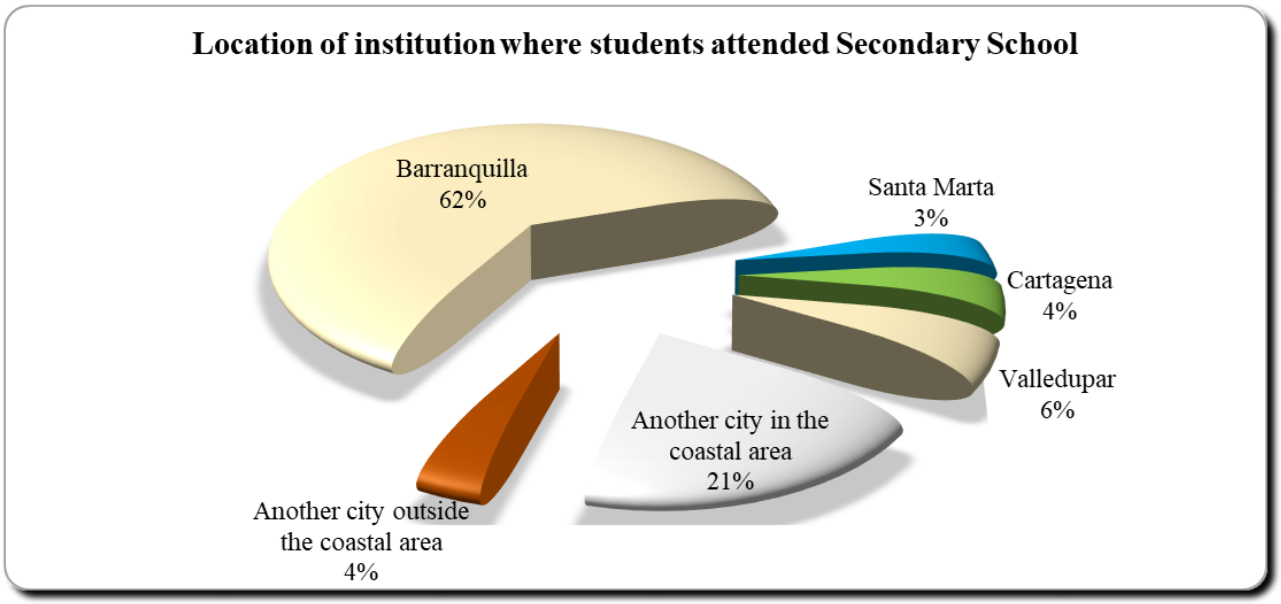



Figure 24. Location of institutions attended in Secondary school.

This evidence about the location of the institutions shows that in Elementary as well as in Secondary school most students attended schools in Barranquilla followed by other cities in the coastal area, Cartagena and Valledupar for Elementary and Valledupar and Cartagena for Secondary School.

Regarding the contact with other languages, all students from the present research reported that they had Spanish as the mother language.

### 4.2.2. Results of socio-cultural factors that could have incidence in the development of EFL. Part three of the survey

Regarding some sociocultural aspects that could influence the learning of EFL, the students answered several questions regarding their beliefs, behaviors, decisions and some situations that could affect EFL learning.

When asked if they considered that learning English was important, 635 reported that it was important, and 1 considered it was not.

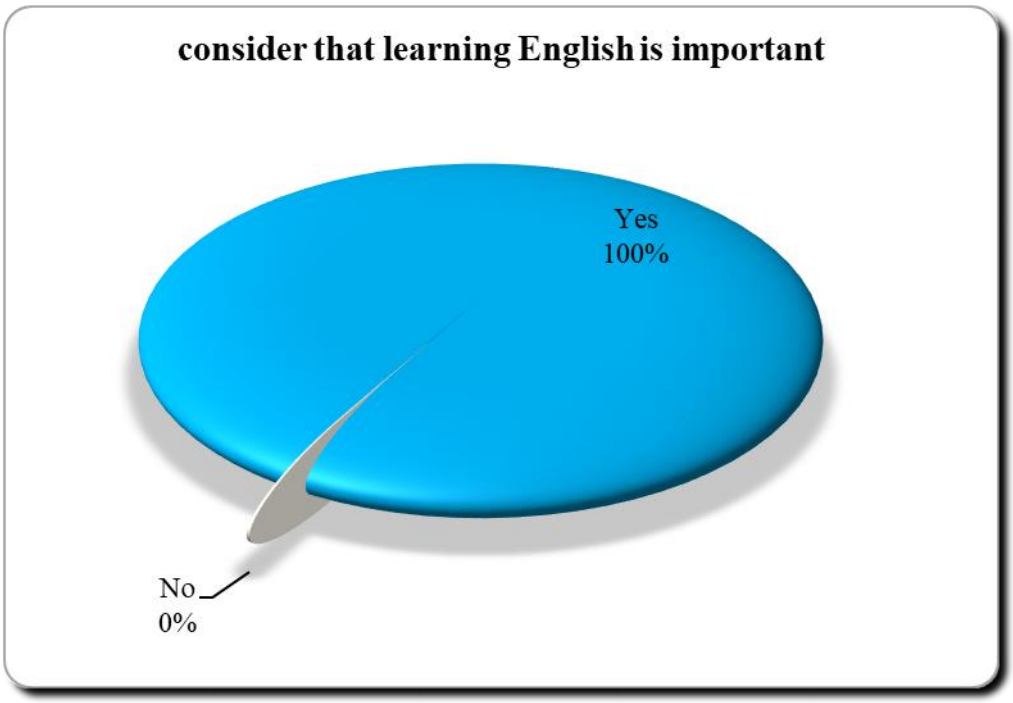



Figure 25. Importance of learning English.

When asked if any of their family members spoke the English language, 439 students reported an affirmative answer and 197 said no family members spoke the English language. From the students with family members who spoke the English language 174 practice with them every time they have the possibility, but 265, do not.

Regarding the most and the least important reason to attend the English course now (the time of the research) students reported the following:

Table 44

*The most and the least important reasons to study English*

| | Frequency | Percentage | Valid percentage | Accumulated |
|---|---|---|---|---|
| English is very important for my professional | 331 | 51.7 | 52.0 | 52.0 |
| English is necessary to graduate from my career | 157 | 24.5 | 24.7 | 76.7 |
| English is a subject that I like | 80 | 12.5 | 12.6 | 89.3 |
| English is one of the subjects I have to take | 37 | 5.8 | 5.8 | 95.1 |
| I study English to please my parents | 31 | 4.8 | 4.9 | 100.0 |
| Total | 636 | 99.4 | 100.0 | |

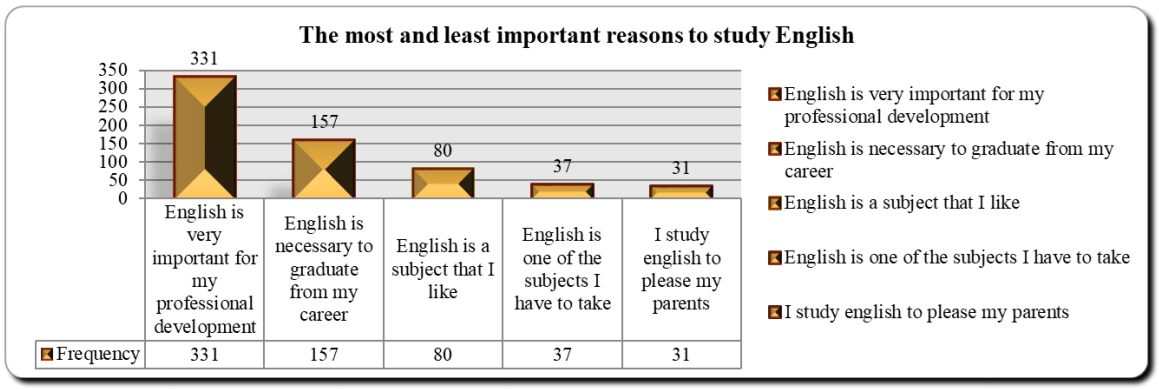



Figure 26. Reasons to study English.

In this case, students are motivated to learn English mainly to achieve professional development and on second place, because it is necessary to graduate from their studies. On the third place, students like English as a subject. On the fourth-place, students consider English as any other subject. We can see that the main motivation to study English is an instrumental one, but not because they like it.

Regarding the time devoted to listen to music in English and other languages students reported the following:

Table 45

*Time spent listening to music in English*

| Time spent listening to music in English | | | | |
|---|---|---|---|---|
| | **Frequency** | **%** | **Valid %** | **Accumulated %** |
| When listening to music, most of the time is in English, a few times in other languages or instrumental music. | 190 | 29.7 | 29.9 | 29.9 |
| When listening to music I spend the same time listening to music in English or other languages | 308 | 48.1 | 48.4 | 78.3 |
| Most of the time, I listen to music in other languages or instrumental music and a few times in English | 138 | 21.6 | 21.7 | 100.0 |
| Total | 636 | 99.4 | 100.0 | |

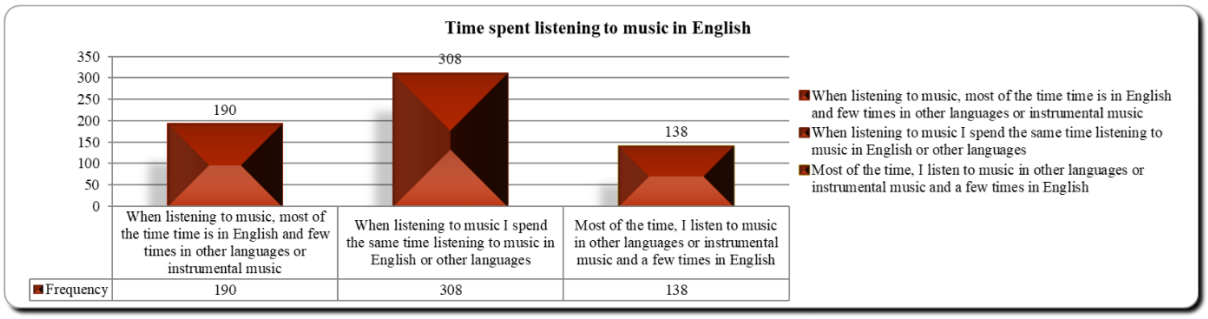



Figure 27. Time spent listening to music in English.

It can be observed that almost 30% of the students prefer to listen to music in English over other languages or instrumental music, but a higher percentage: 48% devote the same time to listen to music in English and other languages.

Another variable assessed in the survey was the motivation to listen to music in English.

Table 46

### *Motivation to listen to music in English*

| Motivation to listen to music in English | | | | |
|---|---|---|---|---|
| | **Frequency** | **Percentage** | **Valid percentage** | **Accumulated** |
| When listening to music in English, I do it for musical reasons without worrying about English pronunciation | 123 | 19.2 | 19.3 | 19.3 |
| When listening to music in English, I do it to enjoy music in English and incidentally pronounce English expressions | 383 | 59.8 | 60 .2 | 79.6 |
| When listening to music in English I enjoy it, but I do it mainly to pronounce English words and expressions | 130 | 20.3 | 20.4 | 100.0 |
| Total | 636 | 99.4 | 100.0 | |

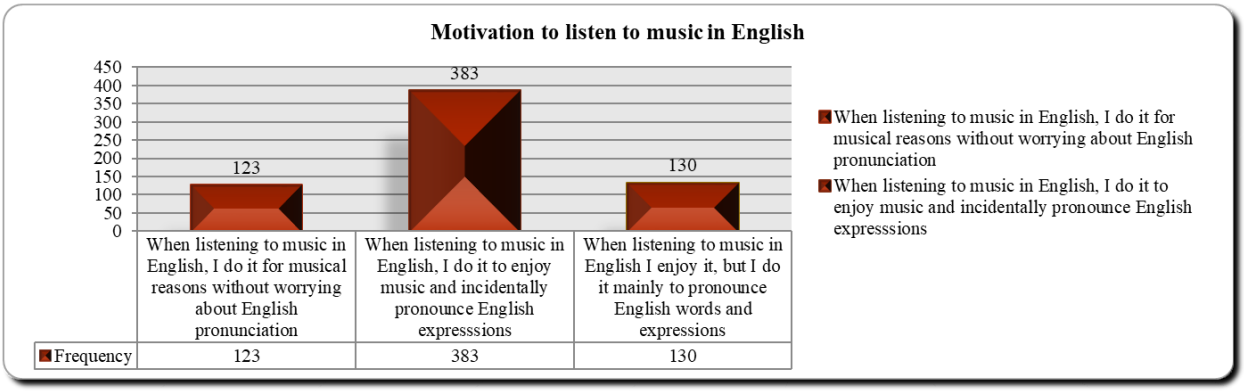



Figure 28. Motivation to listen to music in English.

In this case, most students do not have a purpose when listening to music in English. Sixty percent of students do it just to enjoy music and incidentally pronounce some words. Another 19% of students enjoy music without worrying about English pronunciation. Only 20% of students have the purpose to pronounce English words and expressions when listening to music in English.

Regarding the language stated as default in the students' PC:

Table 47

*Language established in PC as default*

| | Frequency | Percentage | Valid percentage | Accumulated percentage |
|---|---|---|---|---|
| English | 121 | 18.9 | 19.0 | 19.0 |
| Spanish | 511 | 79.8 | 80.3 | 99.4 |
| Other | 4 | .6 | .6 | 100.0 |
| Total | 636 | 99.4 | 100.0 | |

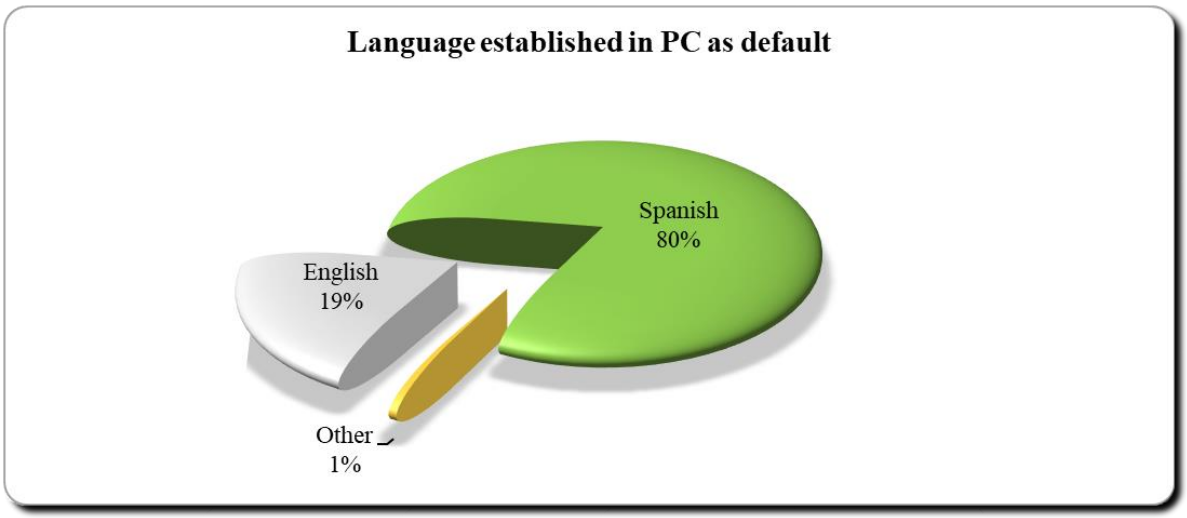



Figure 29. Language established in PC as default.

In this case, most students keep Spanish as the language of default in their PCs (80%).

Students were also asked if they considered their environment made their learning of the English language easier. In other words, if their environment was favorable for the learning of the English language.

Table 48

*Is the environment a facilitator of the English learning process or not?*

| | Frequency | Percentage | Valid percentage | Accumulated percentage |
|---|---|---|---|---|
| Yes | 191 | 29.8 | 30.0 | 30.0 |
| No | 445 | 69.5 | 70.0 | 100.0 |
| Total | 636 | 99.4 | 100.0 | |
| Total | | 640 | 100.0 | |

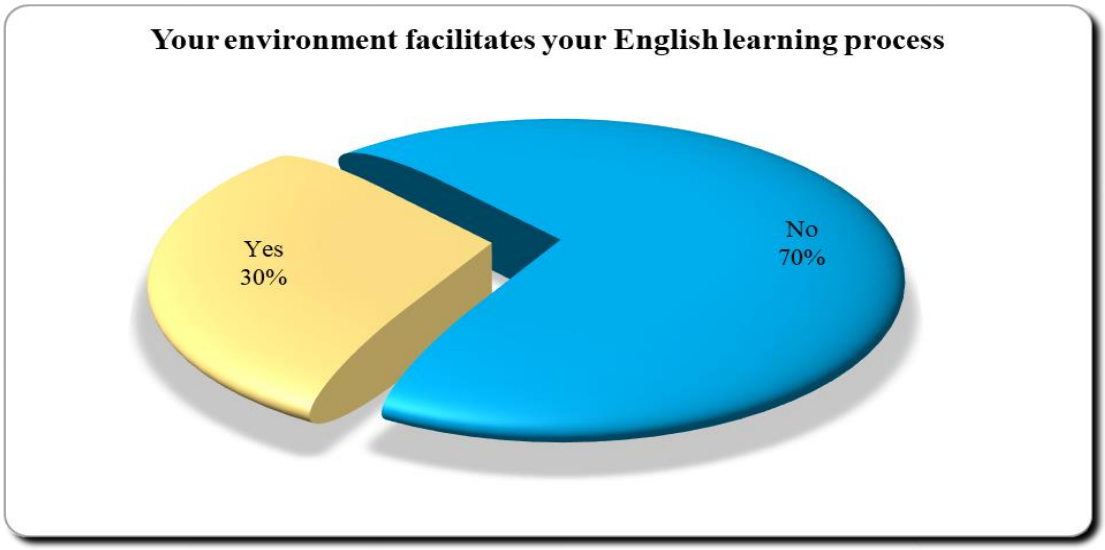



Figure 30. Environment as a facilitator of the EFL learning process.

Most students, 70%, consider their environment does not facilitate the learning of English.

Regarding the belief students had about the development of proficiency in the English language considering that they were living in a Hispanic culture, they answered the following:

Table 49

*Will you become proficient in English while living in a Hispanic culture?*

| | Frequency | Percentage | Valid percentage | Accumulated percentage |
|---|---|---|---|---|
| Yes | 317 | 49.5 | 49.8 | 49.8 |
| No | 319 | 49.8 | 50.2 | 100.0 |
| Total | 636 | 99.4 | 100.0 | |

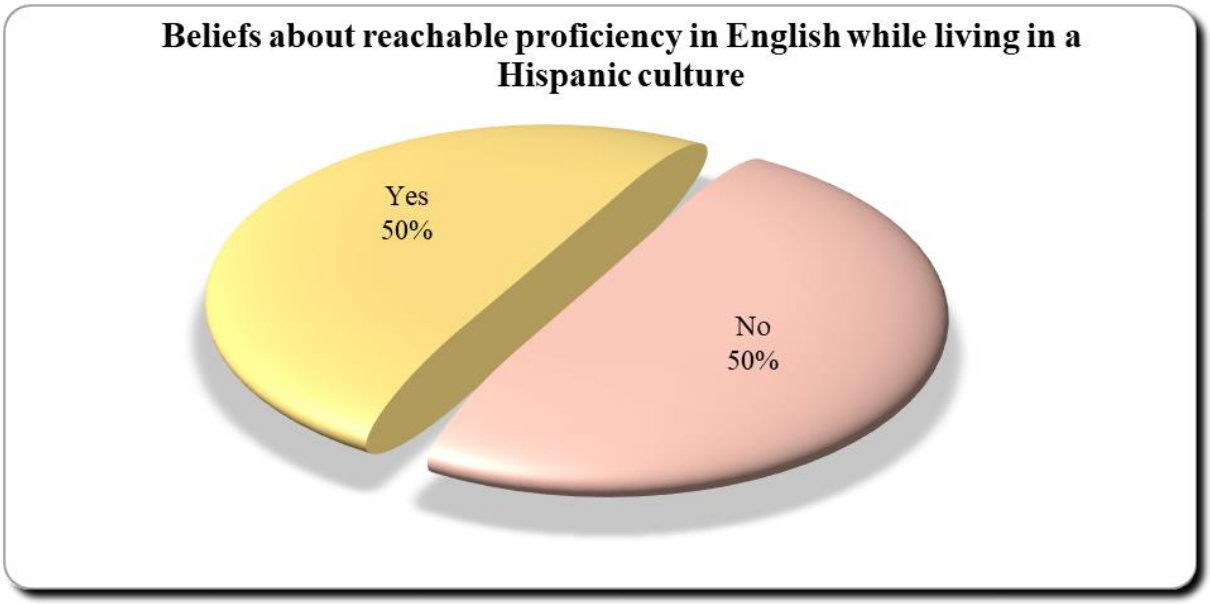



Figure 31. Environment as a facilitator of the EFL learning process.

There is a balance between the students who believe that proficiency in the English language is achievable even though they live in a Hispanic culture and those who believe that it is not possible.

The last variable to report is related to the importance of English interaction in two levels: one, in the circle of friends and two, with the classmates in the academic environment.

Table 50

*Importance of Interaction in English within a group of friends or classmates*

| Importance of interaction in English within a group of friends | | | | |
|---|---|---|---|---|
| | Frequency | Percentage | Valid percentage | Accumulated percentage |
| Very high | 103 | 16.1 | 16.2 | 16.2 |
| High | 284 | 44.4 | 44.7 | 60.8 |
| Average | 168 | 26.3 | 26.4 | 87.3 |
| Null | 81 | 12.7 | 12.7 | 100.0 |
| **Total** | 636 | 99.4 | 100.0 | |

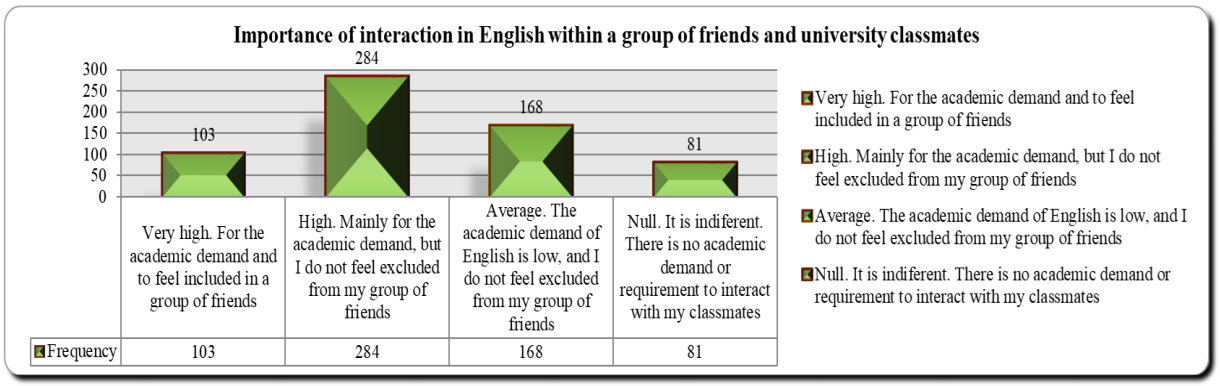



Figure 32. Importance of interacting in English in groups.

In this case, 44% of students consider they need to interact in English for academic reasons, but they do not feel excluded from their group of friends for not doing it in English. The group that considers having a high academic demand to interact in English, also considers they need to interact in English to feel included in their group of friends. That makes 16% of students.

## 4.3. Description of Research Outcomes According to Research Objectives

The main objective from the present study is to investigate the relationship between the main errors found in written compositions from students at university level and the socio-demographic

factors that could have incidence in the development of the written skills **at *Universidad del Norte* in Barranquilla, Colombia.**

### 4.3.1. Complete list of errors

Tables 51 to 58 present the list of 56 errors in eight categories.

Table 51 accounts for Grammar errors and the identifying letter is G.

Table 51

*Grammatical errors*

| | |
|---|---|
| **GDD** | Grammar, Determiner, Demonstrative |
| **GDO** | Grammar, Determiner, Possessive |
| **GDI** | Grammar, Determiner, Indefinite |
| **GDT** | Grammar, Determiner, Other |
| **GA** | Grammar, Articles |
| **GADJCS** | Grammar, Adjectives, Comparative / Superlative |
| **GADJN** | Grammar, Adjectives, Number |
| **GADJO** | Grammar, Adjectives, Order |
| **GADVO** | Grammar, Adverbs, Order |
| **GNC** | Grammar, Nouns, Case |
| **GNN** | Grammar, Nouns, Number |
| **GPD** | Grammar, Pronouns, Demonstrative |
| **GPP** | Grammar, Pronoun, Personal |
| **GPO** | Grammar, Pronoun, Possessive |
| **GPI** | Grammar, Pronoun, Indefinite |
| **GPF** | Grammar, Pronoun, Reflexive/Reciprocal |
| **GPR** | Grammar, Pronoun, Relative/ Interrogative |
| **GPU** | Grammar, Pronoun, Unclear reference |
| **GVAUX** | Grammar, Verbs, Auxiliaries |
| **GVM** | Grammar, Verbs, Morphology |
| **GVN** | Grammar, Verbs, Number |
| **GVNF** | Grammar, Verbs, Non-Finite / Finite |
| **GVT** | Grammar, Verbs, Tense |
| **GVV** | Grammar, Verbs, Voice |
| **GWC** | Grammar, Word Class |

Source: data retrieved from (E. Dagneaux et al., 2005).

Table 52 accounts for Lexical errors and the identifying letter is L.

Table 52

*Lexis errors*

| | |
|---|---|
| **LCC** | Lexis, Conjunctions, Coordinating |
| **LCLC** | Lexis, Connectors, Logical, Complex |
| **LCLS** | Lexis, Connectors, Logical, Single |
| **LCS** | Lexis, Conjunctions, Subordinating |
| **LP** | Lexical Phrase |
| **LPF** | Lexical Phrase, False friends |
| **LS** | Lexical Single |
| **LSF** | Lexical Single, False friends |

Source: data retrieved from (E. Dagneaux et al., 2005).

Table 53 accounts for Word errors and the identifying letter is W.

Table 53

*Word errors*

| | |
|---|---|
| **WM** | Word Missing |
| **WO** | Word Order |
| **WRS** | Word Redundant Single |
| **WRM** | Word Redudant Multiple |

Source: data retrieved from (E. Dagneaux et al., 2005).

Table 54 accounts for Form errors and the identifying letter is F.

Table 54

*Form errors*

| | |
|---|---|
| **FM** | Form, Morphology |
| **FS** | Form, Spelling |
| **FSR** | Form, Spelling, Regional |

Source: data retrieved from (E. Dagneaux et al., 2005).

Table 55 accounts for Punctuation errors and the identifying letter is Q.

Table 55

*Punctuation errors*

| | |
|---|---|
| **QC** | Punctuation, Confusion |
| **QL** | Punctuation, Lexical |
| **QM** | Punctuation, Missing |
| **QR** | Punctuation, Redundant |

Source: data retrieved from (E. Dagneaux et al., 2005).

Table 56 accounts for Lexico-Grammar errors and the identifying letter is L.

Table 56

*Lexico-Grammar errors*

| | |
|---|---|
| **XADJCO** | Lexico-Grammar, Adjectives, Complementation |
| **XADJPR** | Lexico-Grammar, Adjectives, Dependent Preposition |
| **XCONJCO** | Lexico-Grammar, Conjunctions, Complementation |
| **XNCO** | Lexico-Grammar, Nouns, Complementation |
| **XNPR** | Lexico-Grammar, Nouns, Dependent Preposition |
| **XNUC** | Lexico-Grammar, Nouns, Uncountable / Countable |
| **XPRCO** | Lexico-Grammar, Prepositions, Complementation |
| **XVCO** | Lexico-Grammar, Verbs, Complementation |
| **XVPR** | Lexico-Grammar, Verbs, Dependent Preposition |

Source: data retrieved from (E. Dagneaux et al., 2005).

Table 57 accounts for Style errors and the identifying letter is S.

Table 57

*Style errors*

| | |
|---|---|
| **SI** | Sentence, Incomplete |
| **SU** | Sentence, Unclear |

Source: data retrieved from (E. Dagneaux et al., 2005).

Table 58 accounts for Infelicities errors and the identifying letter is Z.

Table 58

*Infelicities*

| | | |
|---|---|---|
| **Z** | Infelicities | |

Source: data retrieved from (E. Dagneaux et al., 2005).

### 4.3.3. Dispersion of errors in the corpus

The total of errors in the corpus was 14,631 with a mean of *M=28.40* errors per student.

Figure 33 presents the normal distribution of errors based on 54 error types.

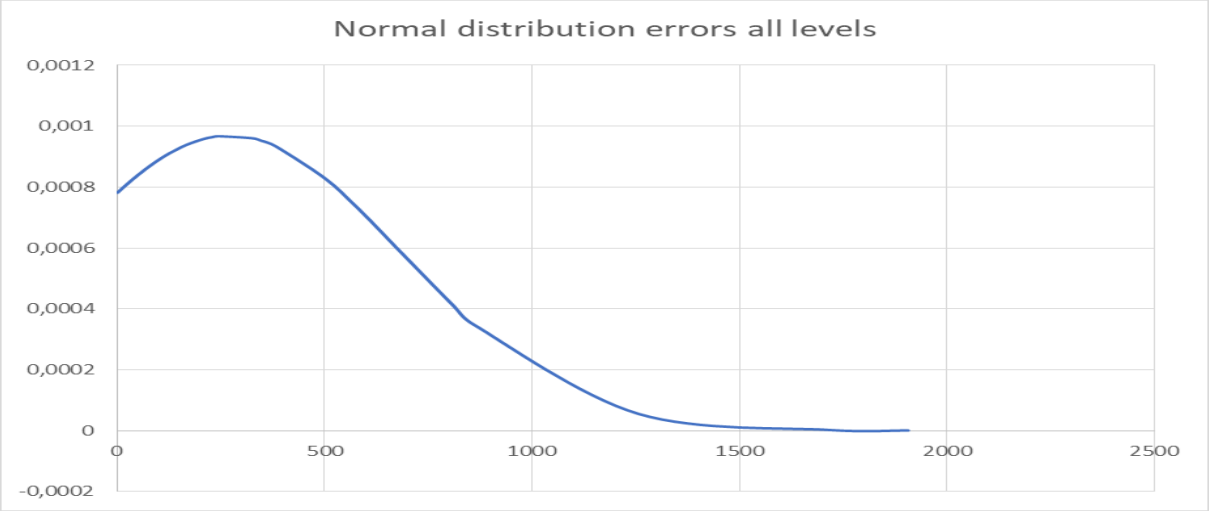



Figure 33. Distribution of errors in the corpus.

Even though the tagger accounts for 56 error types there were two error types not found in the corpus: XCONJCO (Lexico-Grammar, Conjunctions, Complementation) and XADJCO (Lexico-Grammar, Adjectives, Complementation). The dispersion will be presented as a total dispersion of errors in all levels and as an individual dispersion of errors per every level.

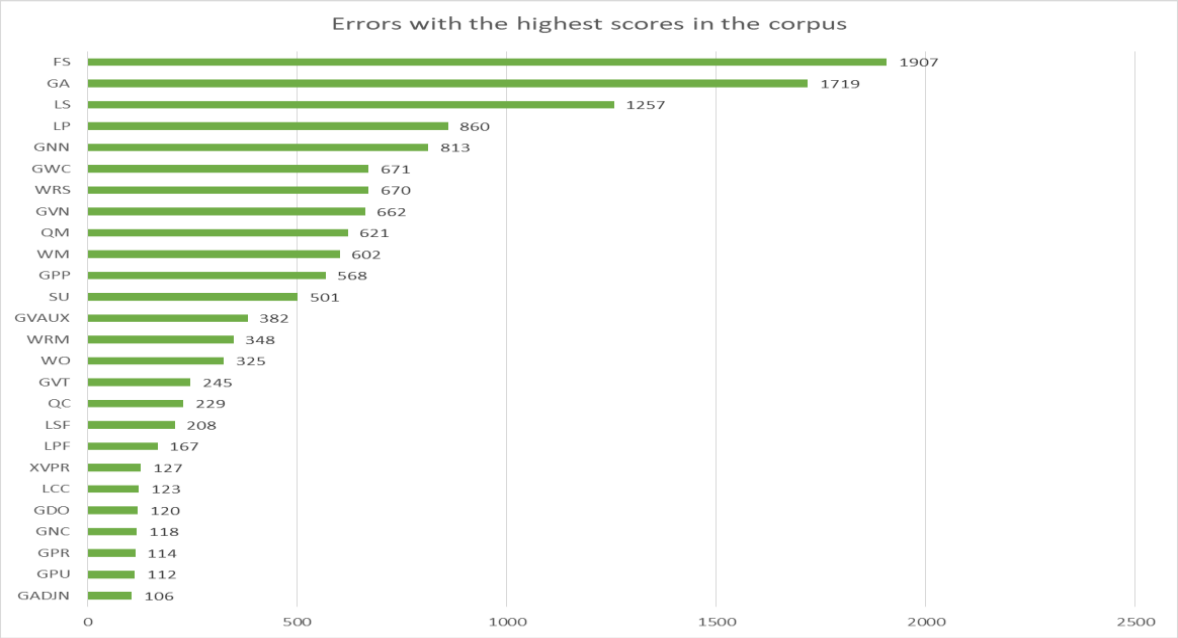



Figure 34. Graphic of errors in the corpus.

According to the classification from *Universidad del Norte* there are 4 levels:

- Pre-Intermediate: B1.1
- Intermediate: B1.2
- Intermediate II: B1.3-B2.1
- Upper-Intermediate: B2.2-B2.3
- Follows the analysis per every level.

#### 4.3.3.1. Error dispersion in Pre-Intermediate level (B1.1)

The total of errors in this level was 5,651 with a mean of *M=23,54* errors per student. The average of words per text in this level was 196.

Figure 35 shows the normal distribution of errors based on 52 error types. There were two more error types not found in this level: LCLS (Lexis, Connectors, Logical, Single) and XADJPR (Lexico-Grammar, Adjectives, Dependent Preposition).

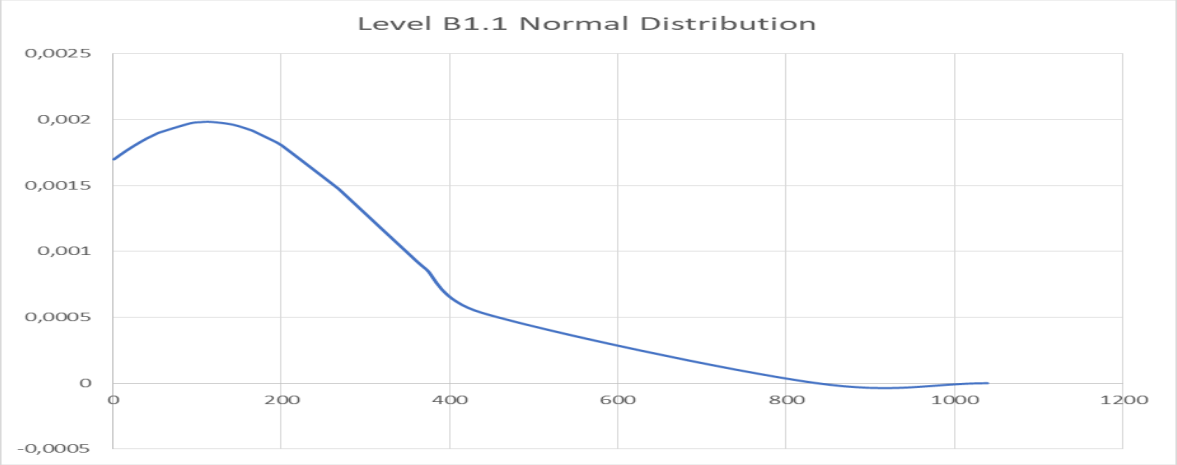



Figure 35. Distribution of errors in level B1.1.

The number of errors in this level accounted for 38% of the total in the corpus. The errors with most incidence in this level, as well as the percentages and means are presented in Table 59.

Table 59

*Errors with most incidence in level B1.1*

| **Error type** | **Total** | **Percentage** | ***Mean*** |
|---|---|---|---|
| FS | 1039 | 18.39% | *4.3* |
| GA | 833 | 14.74% | *3.47* |
| LS | 439 | 7.77% | *1.8* |
| GNN | 372 | 6.58% | *1.55* |
| LP | 345 | 6.11% | *1.4* |

**Analysis of results from B1.1**

The analysis of the results from level B1.1 will be based on the five errors with most incidence within the level.

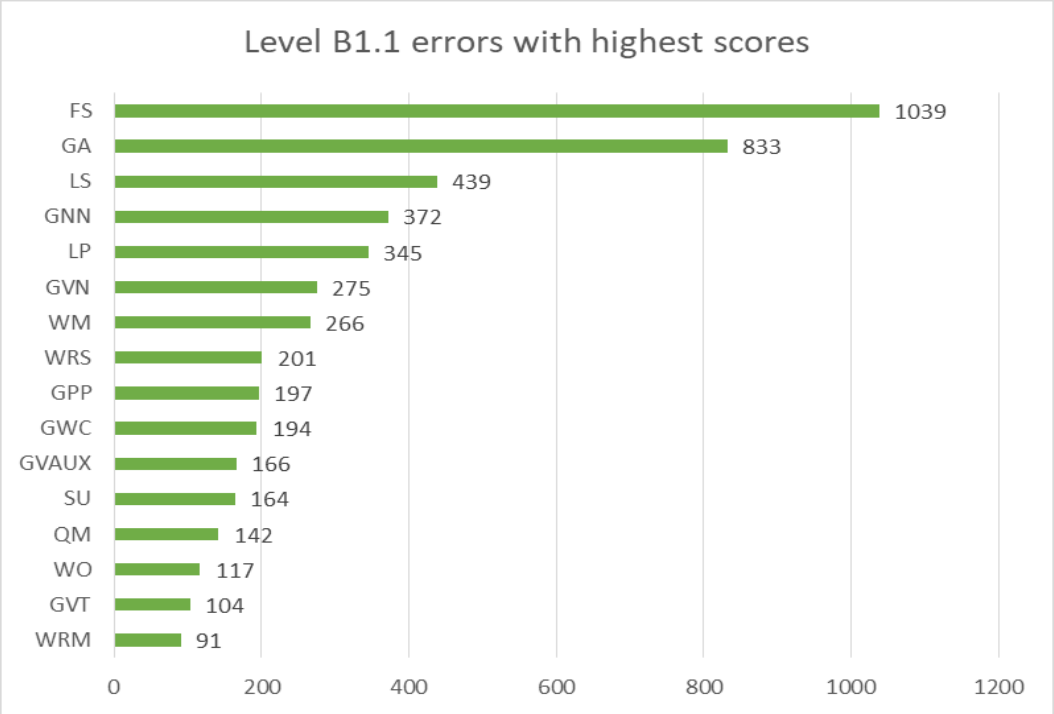



Figure 36. Most relevant errors in level B1.1.

As the data show, the error with the highest score was FS (Form, Spelling). This type of error refers not only to problems in the use of a correct spelling, but also to omission or misuse of capital letters, the borrowing of words from the mother tongue or the use of homophonous words. The high incidence of this error in this level is an indicator that students have not achieved proficiency in the use of spelling rules, or in other cases, it could be due to a lack of focus.

Regarding the second error with most incidence in this level was GA (Grammar, Article). This type of error refers to problems in the use of definite, indefinite and zero article. This error has a high incidence in this level because all students, Spanish speakers, tend to use articles before nouns imitating their mother tongue.

The third error with most incidence in this level was LS (Lexical, Single); this type of error refers to the wrong conceptual, collocational or connotative use of single words. In this case, the incidence might be related to the fact that, in many cases, students are still in the process of identifying and learning collocations and connotative meaning from words and expressions.

The fourth error with most incidence in this level was GNN (Grammar, Noun Number). This type of error refers to the addition or omission of the plural morpheme. Incidence of this error in this case is probably the result of a lack of attention because in their mother tongue the plural morpheme is used to differentiate singular nouns.

The fifth error with most incidence in this level was LP (Lexical Phrase). This type of error refers to the incorrect use of (semi) fixed multi-word expressions and idioms or when usual expressions have been paraphrased instead of using the corresponding lexical phrase in English. The incidence of this error in this level could be an indicator that students resort to translation of their mother language instead of searching for the appropriate expression in English.

#### 4.3.3.2. Error dispersion in Intermediate level (B1.2)

The total of errors in this category was 3,315 with a mean of *M=22.70* errors per student. Figure 37 shows the normal distribution of errors based on 48 error types. Additional to the two error types not found in level B1.1, there were other four error types not found in this level:

GADJN (Grammar, Adjectives, Number), LCLC (Lexis, Connectors, Logical, Complex), XNCO (Lexico, Grammar, Nouns, Complementation), XNPR (Lexico-Grammar, Nouns, Dependent Preposition).

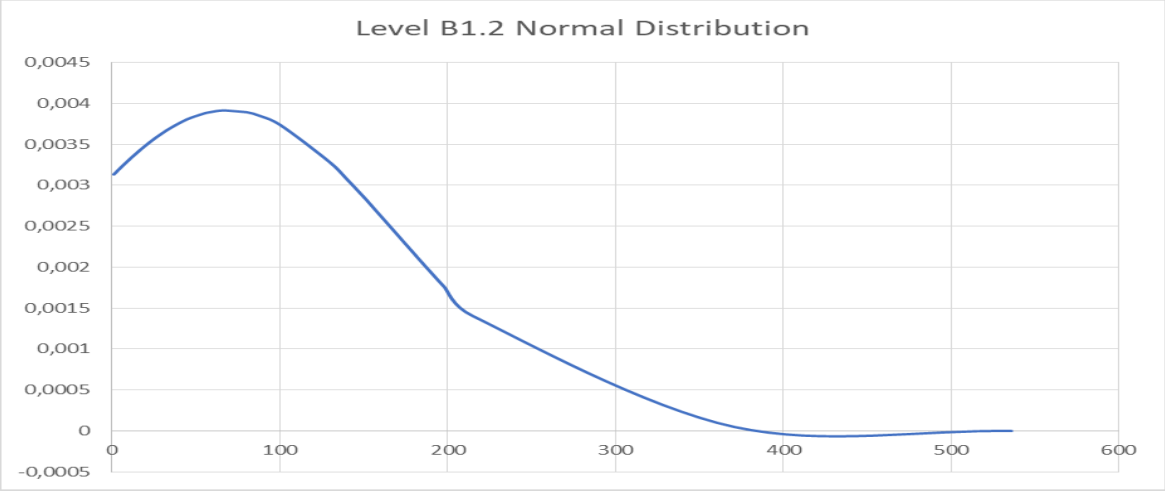



Figure 37. Distribution of errors in level B1.2.

The number of errors in this level accounted for 22% of the total in the corpus.

The results in Table 60 show the first five most relevant errors with the percentages and means within the level. The errors with most incidence in this level as well as the percentages and means are presented in Table 60.

Table 60

*Errors with most incidence in level B1.2*

| **Error type** | **Total** | **Percentage** | ***Mean*** |
|---|---|---|---|

| | | | |
|---|---|---|---|
| FS | 536 | 16.17% | *3.6* |
| GA | 371 | 11.19% | *2.5* |
| QM | 215 | 6.49% | *1.4* |
| LS | 198 | 5.97% | *1.3* |
| LP | 182 | 5.49% | *1.2* |

**Analysis of results from B1.2**

The analysis of the results from level B1.2 will be based on the five errors with most incidence within the level and compared with the previous analysed levels.

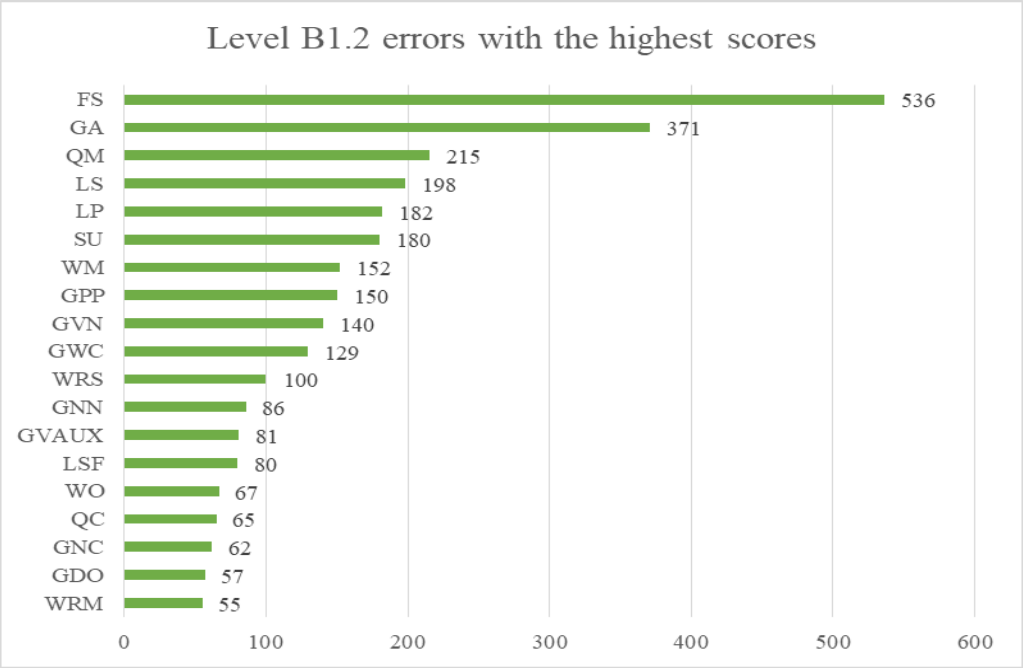


Figure 38. Most relevant errors in level B1.2.

As the data show and like the findings in level B1.1, the error with most incidence was FS (Form, Spelling). In this case, the incidence is lower than in the previous level. This is an indicator that the interlanguage from students is changing.

The second error with most incidence in this level was GA (Grammar, Article). In this case, the incidence is lower than in the previous level which indicates that again students are having progress in the use of the definite or indefinite article.

The third error with most relevance in this level was QM (Punctuation, Missing). This error is not in the list of the five main errors from level B1.1. Its incidence in this level could be an indicator that students are trying to use more complex structures and they are not aware of punctuation rules.

The fourth main error in this level was LS (Lexical Single). This error was also found in the list of main errors from level B1.1. In this case, it has a lower incidence, which again proves that students are in the transition to a higher level.

The fifth error with most incidence in this level was LP (Lexical Phrase). In this level, LP error has a lower influence than in B1.1. Again, its lower influence in this case might be an indicator that students are learning idioms and expressions in English that did not have in B1.1.

#### 4.3.3.3. Error dispersion in Intermediate II level (B1.3-B2.1)

The total of errors in this category was 581 with a mean of *M=34,17* errors per student. Figure 39 shows the normal distribution of errors based on 42 error types. There were four more error types not found in this level: GADVO (Grammar, Adverbs, Order), GDI (Grammar, Determiner, Indefinite), GPF (Grammar, Pronoun, Reflexive/Reciprocal) and GPU (Grammar, Pronoun, Unclear reference).

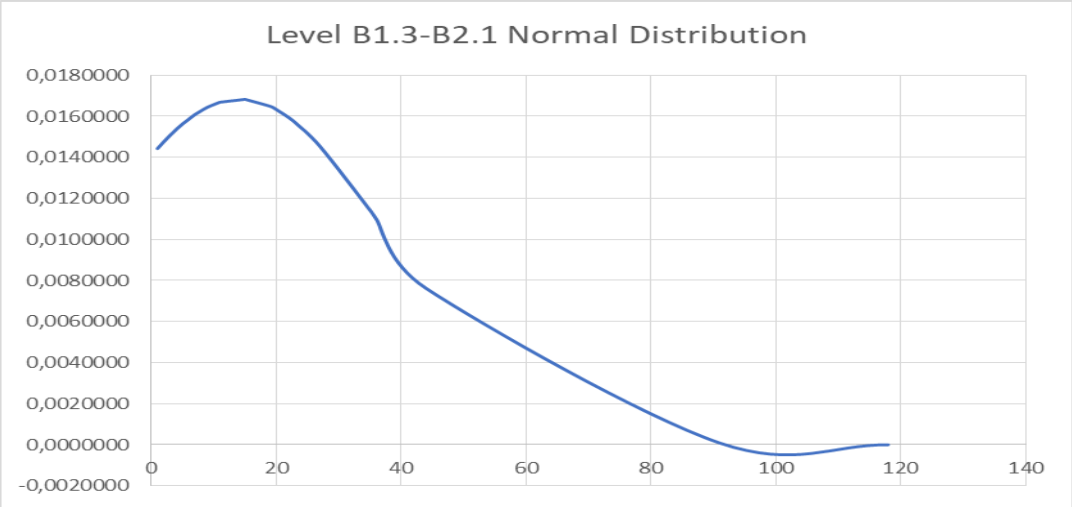



Figure 39. Distribution of errors in level B1.3-B2.1.

The number of errors in this level accounted for 0.039% of the total in the corpus. The errors with most incidence in this level as well as the percentages and means are presented in Table 61.

Table 61

*Errors with most incidence in level B1.3-B2.1*

| **Error type** | **Total** | **Percentage** | ***Mean*** |
|---|---|---|---|
| FS | 118 | 20.31% | *6.94* |
| GA | 91 | 15.66% | *5.35* |
| GNN | 44 | 7.57% | *2.58* |
| LS | 36 | 6.20% | *2.1* |
| SU | 35 | 6.02% | *2* |

**Analysis of results from B1.3-B2.1**

The analysis of the results from level B1.3-B2.1 will be based on the five errors with most incidence within the level. These levels are compared with the results from previous levels, in this case, levels B1.1 and B1.2.

Table 62

*Comparative of the first 5 errors in three different levels*

| Level B1.1 | | | Level B1.2 | | | Level B1.3-B2.1 | | |
|---|---|---|---|---|---|---|---|---|
| **E. Type** | **%** | ***Mean*** | **E. Type** | **%** | ***Mean*** | **E. Type** | **%** | ***Mean*** |
| FS | 18.39 | 4.3 | FS | 16.17 | 3.6 | FS | 20.31 | 6.94 |
| GA | 14.74 | 3.47 | GA | 11.19 | 2.5 | GA | 15.66 | 5.35 |
| LS | 7.77 | 1.8 | QM | 6.49 | 1.4 | GNN | 7.57 | 2.58 |
| GNN | 6.58 | 1.55 | LS | 5.97 | 1.3 | LS | 6.20 | 2.1 |
| LP | 6.11 | 1.4 | LP | 5.49 | 1.2 | SU | 6.02 | 2 |

Again, the data shows similar findings regarding the type of errors with most incidence.

Figure 40 shows the errors with most incidence in this level.

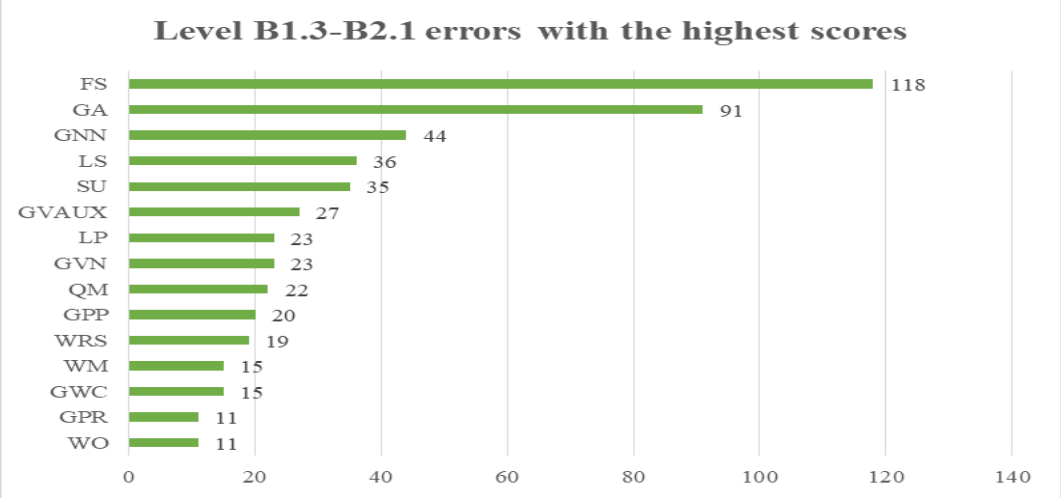



Figure 40. Most relevant errors in level B1.3-B2.1.

The error leading the account is again FS (Form, Spelling), but in this case, it has a higher incidence compared with the previous levels. Its incidence and means of errors per student increased. Comparing FS in the three levels from Table 62 there is an increase of at least 2 points in the present level. This result is in some ways unexpected because it shows how students have

not mastered spelling rules after three levels. It also shows that probably, due to the increase of new vocabulary learners are adapting to a new interlanguage level.

The second error with most incidence in this level was again GA (Grammar, Article). Like the former error in the two previous levels, its incidence is higher. It is an indicator that students still use the definite and indefinite article as they do in their mother tongue, Spanish.

The third error with most incidence in this level was GNN (Grammar, Nouns, Number). This error did not appear in level B1.2 and was in the fourth place in level B1.1 where it had a lower score. The incidence of this error in the present level is an indicator that learners have not mastered the use of plurals yet. This result is surprising taking into consideration that students have similar uses of plurals in their mother tongue.

The fourth error with most incidence in the present level was LS (Lexical Single). This error also had incidence in the two previous levels. Its incidence was higher in level B1.1, then decreased in level B1.2 and increased again in the present level. However, its incidence in the present level was the lowest from the three. It is an indicator that students are using vocabulary with accurate conceptual and connotative meaning and that they are more aware of the collocational options for each lexical item.

The fifth error with most incidence in this level was SU (Sentence Unclear). This error was not present in any of the previous levels. It is a great indicator of a different level of interlanguage. In this case, students are trying to use complex structures to express their

meanings. There is a search for sentences and compound clauses, but they do not master them yet.

#### 4.3.3.4. Error dispersion in Upper-Intermediate level (B2.2-B2.3)

The total of errors in this category was 5084 with a mean of *M= 45.39* errors per student. Figure 41 shows the normal distribution of errors based on 52 error types. There were two error types that were not found in this level: FSR (Form, Spelling, Regional) and XNCO (Lexico-Grammar, Nouns, Complementation).

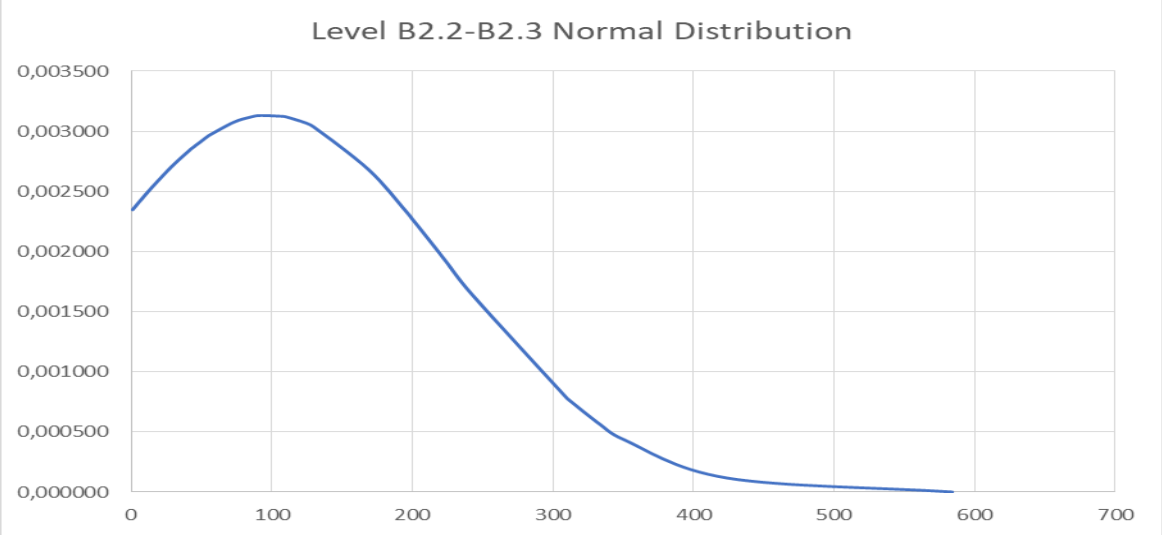



Figure 41. Distribution of errors in level B1.1.

The number of errors in this level accounted for 34% of the total in the corpus. The errors with most incidence in this level as well as the percentages and means are presented in Table 63.

Table 63

*Errors with most incidence in level B2.2-B2.3*

| Error type | Total | Percentage | *Mean* |
|---|---|---|---|
| LS | 584 | 11.00% | *5.21* |
| GA | 424 | 0.08% | *3.7* |
| WRS | 350 | 0.07% | *3.12* |
| GWC | 333 | 0.07% | *2.97* |
| GNN | 311 | 0.06% | *2.77* |

**Analysis of results from B2.2-B2.3**

The analysis of the results from level B2.2-B2.3 will be based on the five error with the most incidence within the level and compared with the 3 previously analysed levels, in this case, levels B1.1, B1.2, and B1.3-B2.1.

Table 64

*Comparative of the first 5 errors in four different levels*

| **Level B1.1** | | | **Level B1.2** | | | **Level B1.3-B2.1** | | | **Level B2.2-B2.3** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **E. Type** | **%** | ***M*** | **E. Type** | **%** | ***M*** | **E. Type** | **%** | ***M*** | **E. Type** | **%** | ***M*** |
| FS | 18.39 | *4.3* | FS | 16.17 | *3.6* | FS | 20.31 | *6.94* | LS | 11 | *5.2* |
| GA | 14.74 | *3.47* | GA | 11.19 | *2.5* | GA | 15.66 | *5.35* | GA | 0.08 | *3.7* |
| LS | 7.77 | *1.8* | QM | 6.49 | *1.4* | GNN | 7.57 | *2.58* | WRS | 0.07 | *3.1* |
| GNN | 6.58 | *1.55* | LS | 5.97 | *1.3* | LS | 6.20 | *2.1* | GWC | 0.07 | *2.9* |
| LP | 6.11 | *1.4* | LP | 5.49 | *1.2* | SU | 6.02 | *2* | GNN | 0.06 | *2.7* |

The data shows similar findings in the first three levels, but results change in the last level (B2.2-B2.3).

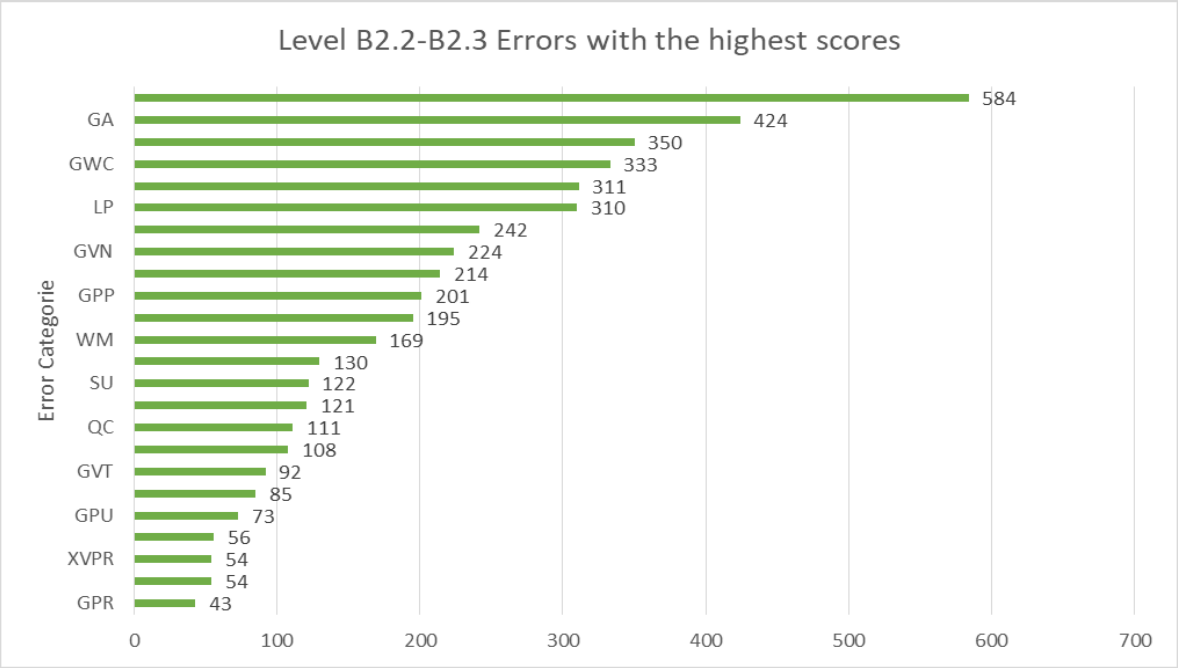



Figure 42. Most relevant errors in level B2.2-B2.3.

The error with most incidence in this level is LS (Lexical Single), an error that was present in previous levels, but with less occurrence. Different from the previous levels FS is no more the error with the most incidence. The frequency of LS in this level is higher than it was in the three previous levels. It is an indicator, not only that learners are on the search of new vocabulary, but also an evidence that students have not mastered the use of hyphenated compounds and the conceptual, collocational and connotative use of words, that at this stage are more specialised.

The second error with most incidence in this level was GA (Grammar, Article). Again, GA, which concerns errors related to the misuse of definite and indefinite articles, is placed on the second position, like previous levels. With less incidence from the 3 previous levels, this is a persistent error that has been in all the learning process. The recurrence of this error is an indicator that students have not mastered the use of articles in English. It is surprising that after

all the stages in different levels students still tend to use articles the same way they do in Spanish which is their mother tongue.

The third error with most prevalence in this level was WRS (Word Redundant Single). A new error that was not in previous levels. This error refers to the unnecessary repetition of single words. In many cases, this error is due to their lack of attention. The incidence of this error could be related to typo problems since students from this level did their final assignment in digital version.

The fourth error with most frequency in this level was GWC (Grammar, Word Class). This is also a new error that was not present in the list of main errors from previous levels. With a high score in the present level, this error refers to the inappropriate use of a word class. The incidence of this error in the present level shows the search of students for a more refined vocabulary, but still learners confuse word class and could use an adjective instead of a noun, for instance.

The fifth error with most incidence was GNN (Grammar, Nouns, Number). This error, which regards the addition or omission of plural morphemes was present in levels B1.1 and B1.3-B2.1 with more frequency. This error is an indicator that students have not mastered the use of plural morphemes.

### 4.3.4. Overview of errors by gender in levels B1 and B2

The 515 participants in the present research were distributed into 238 males and 274 females. There were three files that did not report gender for they were anonymous; therefore, the data was kept for informative reasons, but did count in the final results. Figure 43 shows the quantity of students by gender in each level, the number of tokens and the number of errors in each case. The analysis of error by gender was performed according to CEFR levels.

| | Code | Level | Qty | Tokens | % Gender | % Token | # Of Errors | % Errors/Tokens |
|---|---|---|---|---|---|---|---|---|
| **Male** | 1 | **B1.1** | 104 | 19,301 | 43% | 41% | 2,402 | 12.4% |
| **Female** | 2 | **B1.1** | 135 | 27,592 | 56% | 59% | 3,226 | 11.7% |
| **No gender** | 0 | **B1.1** | 1 | 176 | 0.4% | 0.4% | 23 | 13.1% |
| **Male** | 1 | **B1.2** | 70 | 13,400 | 48% | 46% | 1,517 | 11.3% |
| **Female** | 2 | **B1.2** | 74 | 15,605 | 51% | 53% | 1,754 | 11.2% |
| **No gender** | 0 | **B1.2** | 2 | 437 | 1% | 1% | 44 | 10.1% |
| **Male** | 1 | **B1.3-B2.1** | 8 | 1,784 | 47% | 47% | 279 | 15.6% |
| **Female** | 2 | **B1.3-B2.1** | 9 | 1,977 | 53% | 53% | 302 | 15.3% |
| **Male** | 1 | **B2.2-B2.3** | 56 | 33,725 | 50% | 49% | 2,526 | 7.5% |
| **Female** | 2 | **B2.2-B2.3** | 56 | 35,328 | 50% | 51% | 2,558 | 7.2% |

Figure 43. Overview of errors by gender in each level.

The column errors/tokens describes the percentage of errors for every 100 tokens. It can be observed how males have a higher percentage of errors from females even though they are less productive than females.

#### 4.3.4.1. Errors in level B1 and B2

To do the analysis based on genre, it is necessary to clarify the following aspects: level B1 is formed by B1.1, B1.2, B1.3-B2.1 and level B2 is formed by B2.2-B2.3. The analysis in this case

will be based on the number of errors per every 100 words produced by learners to have a better understanding of the results.

Figure 44 shows the distribution of errors in level B1. The small box within it shows a comparative of production by gender in tokens and the percentage of errors per every 100 tokens from this level.

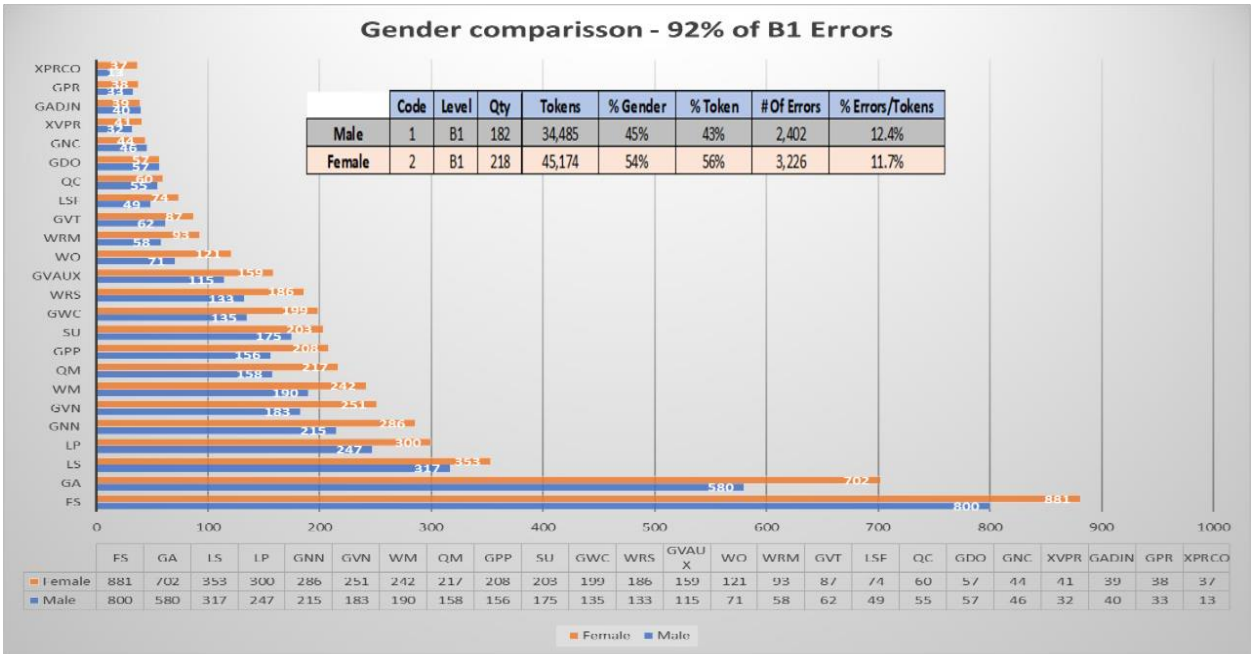


Figure 44. Overview of errors by gender in level B1.

There are several facts that affect the sample; first, there are 16 more females than males in B1. Second, females produce more written work than males in level B1, an average of 207 tokens per every female and 189 tokens per every male. Observing all errors from level B1 as a total, females make more written errors than their male counterparts do, but those errors are proportional to the number of females. The frequency of errors for males had a mean of *M= 13.19* and for females *M= 14.79*, but the percentage of errors for every 100 tokens was 12% for males and 11% for females.

In level B1 the first five errors are the same for males and females: FS (Form, Spelling), GA (Grammar, Articles), LS (Lexical Single), LP (Lexical Phrase) and GNN (Grammar, Nouns, Number).

Figure 45 shows the distribution of errors in level B2. The small box within it shows a comparative of production by gender in tokens and the percentage of errors per every 100 tokens from this level.

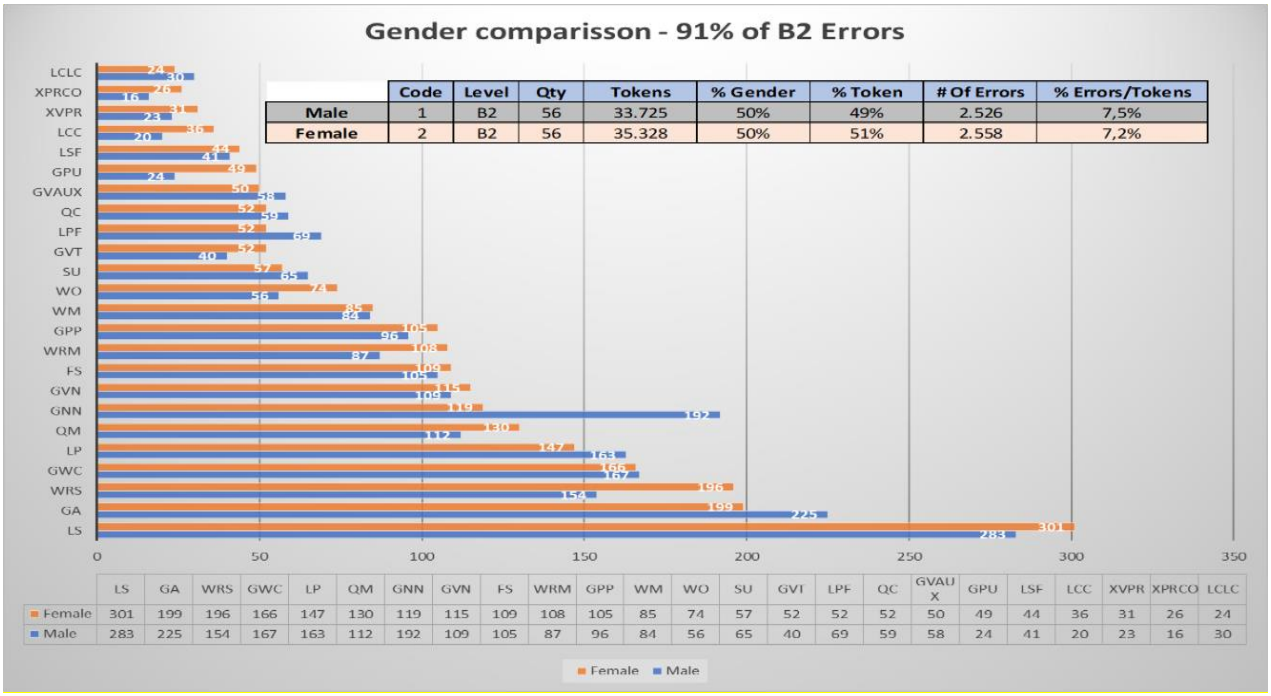
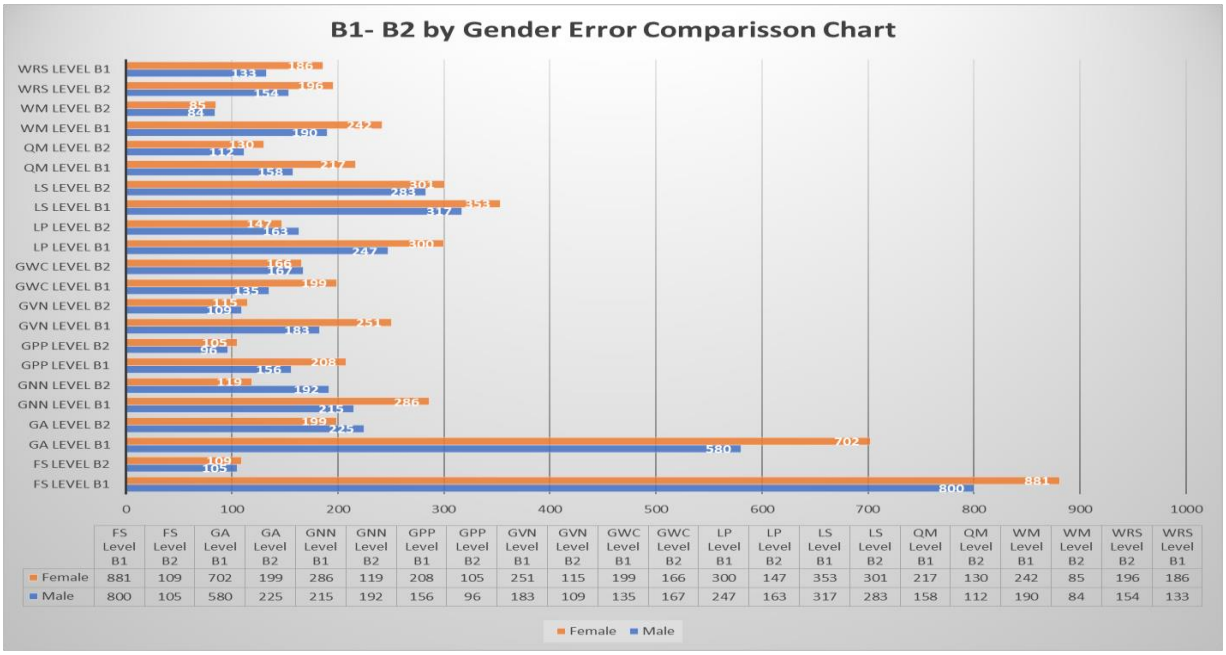


Figure 45. Overview of errors by gender in level B2.

Level B2 could be considered the most even from all levels because the number of students is the same for males and females. There were 56 males and 56 females in this level. The average tokens was 630 per text for females and 602 per text for males with a mean of 45 errors per male student and 45.6 errors per female student. The frequency of errors for males had a mean of *M= 45.00* and for females *M= 45.06*. The percentage of errors for every 100 tokens was 7.5% for males and 7.2% for females.

In this level, the errors with most prevalence changed from level B1. The first error with most incidence was LS (Lexical Single) and both males and females made almost the same number of hits from this error (268 and 269, respectively). The second error with most incidence was the same as in previous levels: GA (Grammar, Articles), but in this case the incidence is higher for males. The third error with most incidence for males was GNN (Grammar, Nouns, Number), but it is WRS (Word Redundant Single) for females.

A comparative graph of errors by gender in levels B1 and B2 shows the following:




| | FS Level B1 | FS Level B2 | GA Level B1 | GA Level B2 | GNN Level B1 | GNN Level B2 | GPP Level B1 | GPP Level B2 | GVN Level B1 | GVN Level B2 | GWC Level B1 | GWC Level B2 | LP Level B1 | LP Level B2 | LS Level B1 | LS Level B2 | QM Level B1 | QM Level B2 | WM Level B1 | WM Level B2 | WRS Level B2 | WRS Level B1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Female | 881 | 109 | 702 | 199 | 286 | 119 | 208 | 105 | 251 | 115 | 199 | 166 | 300 | 147 | 353 | 301 | 217 | 130 | 242 | 85 | 196 | 186 |
| Male | 800 | 105 | 580 | 225 | 215 | 192 | 156 | 96 | 183 | 109 | 135 | 167 | 247 | 163 | 317 | 283 | 158 | 112 | 190 | 84 | 154 | 133 |

Figure 46. Comparative of errors by gender in levels B1 & B2.

The interlanguage from males and females have a similar pattern of change. Similar type and number of errors can be found for males and females in B1 and B2. When the amount of production increases in level B2, both males and females make a great improvement in their grammar and in the amount of written work. The level of interlanguage with most changes was

B2. The first five type of errors decreased in great number when learners passed from B1 to B2. Let us see the first five errors for both levels.

Table 65

*Incidence and percentage in the corpus of the first five types of errors*

| **Type of error** | **Errors males B1** | **% in corpus** | **Errors females B1** | **% in corpus** | **Errors males B2** | **% in corpus** | **Errors females B2** | **% in corpus** |
|---|---|---|---|---|---|---|---|---|
| **FS** (Form, Spelling) | 800 | *0.05* | 881 | *0.06* | 105 | *0.007* | 109 | *0.007* |
| **GA** (Grammar, Articles) | 580 | *0.03* | 702 | *0.04* | 225 | *0.015* | 199 | *0.013* |
| **GNN** (Grammar, Nouns, Number) | 215 | *0.014* | 286 | *0.019* | 192 | *0.013* | 119 | *0.008* |
| **GPP** (Grammar, Pronoun, Personal | 156 | *0.010* | 208 | *0.014* | 96 | *0.006* | 105 | *0.007* |
| **GVN** (Grammar, Verbs, Number) | 183 | *0.012* | 251 | *0.017* | 109 | *0.007* | 115 | *0.0078* |

In all five cases the same errors from B1 had less incidence in level B2. On the contrary, observing all figure 46 can be seen that FS, the main error from B1, changes in B2 for LS.

Now follows the analysis for the sublevels from level B1: B1.1, B1.2, and B1.3-B2.1. Figure 47 shows the distribution of errors in level B1.1. The small box within it shows a comparative of

production by gender in tokens and the percentage of errors per every 100 words from level B1.1.

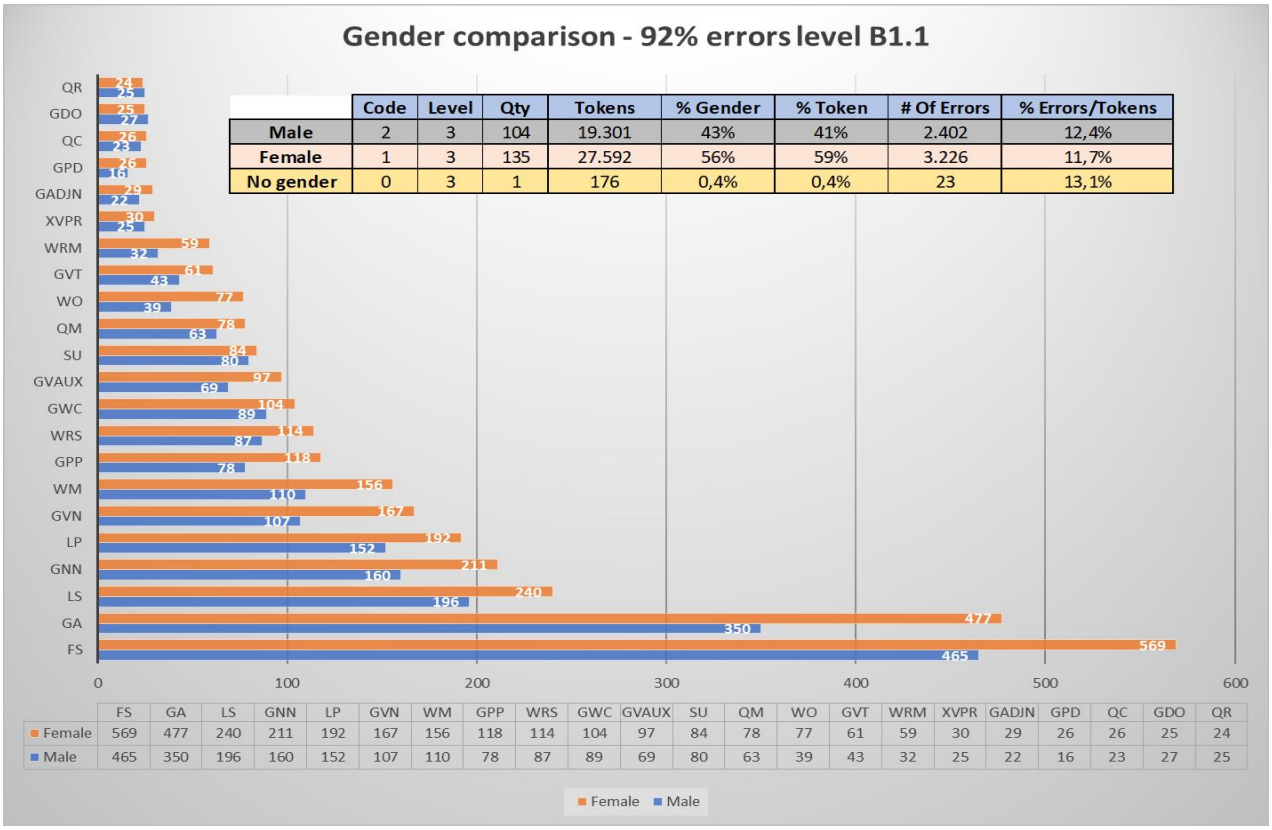


Figure 47. Distribution of errors by gender in level B1.1.

Level B1.1 had 104 males, 135 females and one case in which the student did not report his/her gender and the file did not have his/her name. In the general results, it can be observed that in level B1.1 the written production from females is more than from males, due in part to the greater number of females. However, looking at the results individually, males had an average of 185 tokens per text and females an average of 204 tokens per text. The frequency of errors for males had a mean of *M= 23.09* and for females *M= 23.89*. The percentage of errors for every 100 tokens was 12.4% for males and 11.7% for females.

The most prevalent error types for male and female are exactly the same with the difference that females have more frequency in all cases. Similar to the general corpus, the first three most prevalent errors are FS (Form, Spelling), GA (Grammar, Articles), and LS (Lexical Single).

Figure 48 shows the distribution of errors in level B1.2. The small box shows a comparative of production by gender in tokens and the percentage of errors per every 100 words from level B1.2.

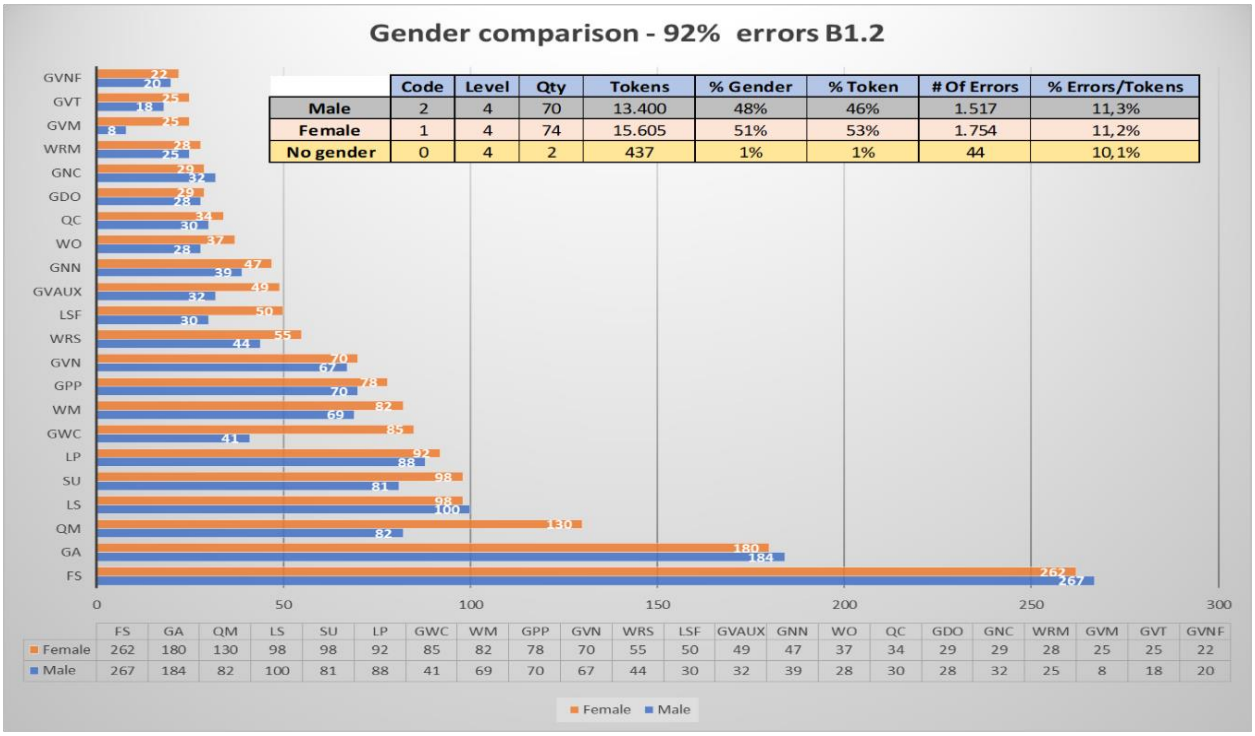


Figure 48. Distribution of errors by gender in level B1.2.

With 70 males and 74 females, level B1.2 had 4 more females than males, plus two cases in which the students did not identify their genders. Again, in this level the writen production from females is more abundant than from males. The average tokens per text for females was 210 and 185 for males. In this case, the first two errors with the most prevalence were the same from

B1.1 (FS and GA), but the third error with most incidence for males and females was QM (Punctuation, Missing). In the present level the same as the previous one, females had more frequency of errors with a mean of *M= 23.70* in front of a mean of *M= 21.67* from males, but the percentage of errors for every 100 tokens was of 11.3% for males and 11.2% for females.

Figure 49 shows the distribution of errors in level B1.3-B2.1. The small box shows a comparative of production by gender in tokens and the percentage of errors per every 100 words from level B1.3-B2.1.

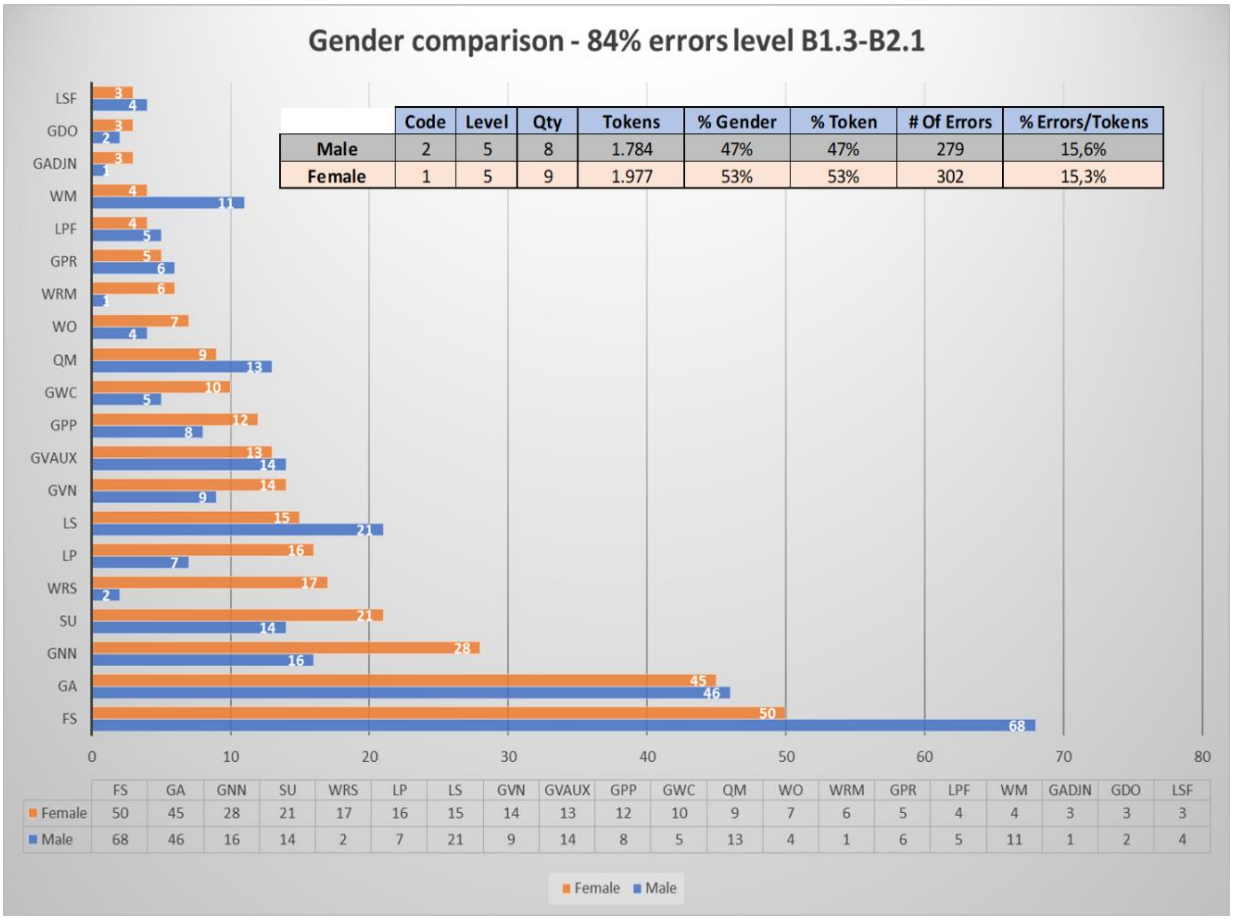


Figure 49. Distribution of errors by gender in level B1.3-B2.1.

Level B1.3-B2.1 only had 17 students, but it is relevant the analysis in order to see the progress of the interlanguage from level B1. In any case, conclusions will be drawn based on level B1 as a whole compound.

Level B1.3-B2.1 had 8 females and 9 males. The average tokens per text was 223 per text for females and 219 per text for males. The mean of errors for males was *M= 37.75* and for females *M= 31.00*. The percentage of errors for every 100 tokens was 14,1% for males and 16,9% for females.

Again, in this case, the two first most prevalent error types for male and female were exactly the same (FS) and (GA). The incidence of FS in this level is higher for males and GA has a quite even incidence for both genders. There is a new error in the third place; GNN (Grammar, Nouns, Number) had a higher incidence in females than males. As level B2 compiles all errors from level B2.2-B2.3 it is not necessary to analyse this level again.

#### 4.3.4.2. General analysis of errors by gender

In section 4.3.4. errors were presented as a whole in the corpus but divided according to gender. From 515 participants, three files were not identified with gender because students did not report it and the files did not have their names, for that reason those files were analysed apart and did not count for the final results. The following are some of the main points analysed in the present section:

1. In level B1 the three most prevalent errors from males and females were the same, FS, GA, and LS.

2. In all sub-levels from B1, females accounted for more quantity of errors than males in part because females did more written production than males and because females were 1.15% more than males. Nevertheless, the mean of errors per every 100 tokens revealed that females had a lower percentage of errors than males in all levels.

3. In level B2, females decreased the amount of errors, but still overpassed males by 32 errors. Again, in this level females had a lower percentage of errors from males for every 100 tokens.

4. The most prevalent error in all level B1 was FS (Form, Spelling), but that tendency changes in level B2 where the most prevalent error becomes LS (Lexical Single).

It can be concluded that in the present research the results matched with the statement from Babayigid (2015, p.33) affirming, “girls outperformed on all dimensions of written expression, except for organization.” In every sub-level from B1 to B2 even though females accounted for more errors, the means show that the average of errors for females is always less than for males.

The results from the present research matched with the stated by Ghani et al (2011) claiming, “The female students committed less errors of L2 writing compared to male students.” In addition “females can be said better language learners than males.” (Saeed, Ghani, and Ramzan 2011, p.1).

### 4.3.5. Errors and Common European Framework of Reference (CEFR)

It is necessary to understand the incidence of errors related to the CEFR level in which students were classified to verify if the interlanguage of students is developing according to what is expected. For the present research the criteria from CEFR using the descriptions from English Profile (UCLES/CUP, 2011) were taken into account in order to have a clearer view from the possible error types found in the different levels.

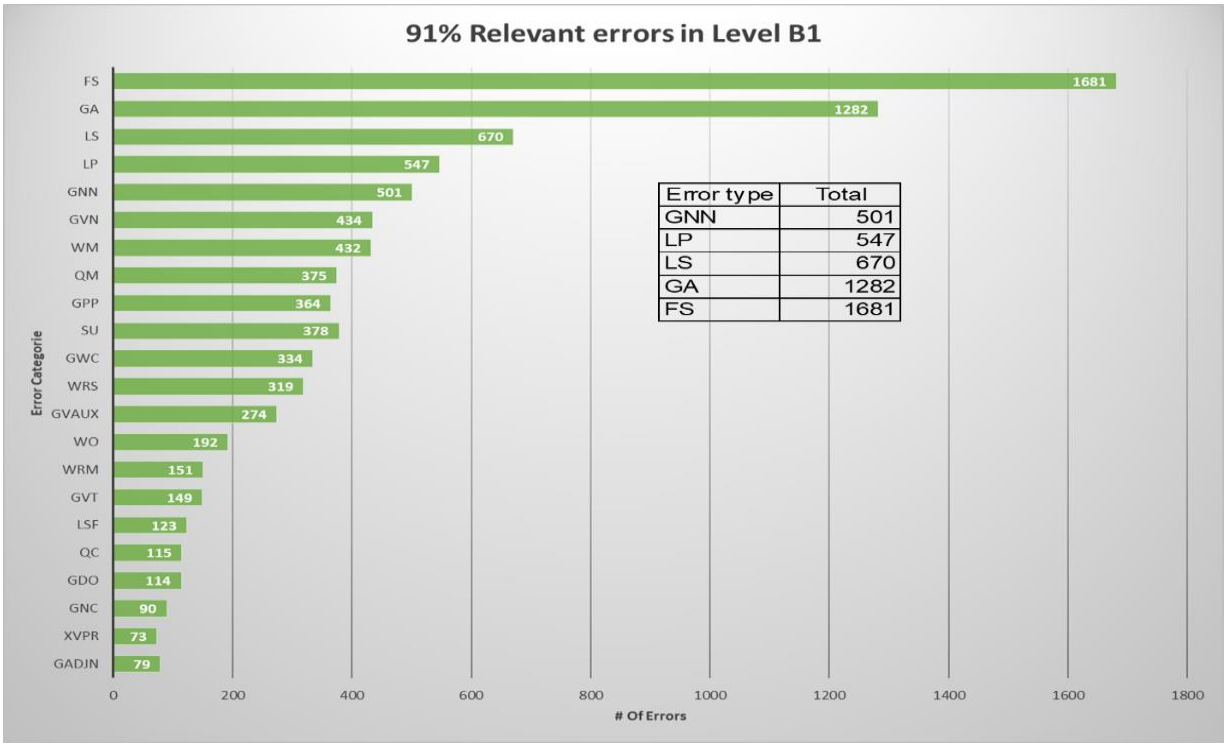



Figure 50. Most frequent written errors found in level B1.

Looking at errors from B1 and B2 a difference in the type and quantity can be seen. For instance, the most recurrent error in level B1 was Form Spelling while in level B2 was Lexical Single.

Figure 51 shows the most frequent errors in level B2.

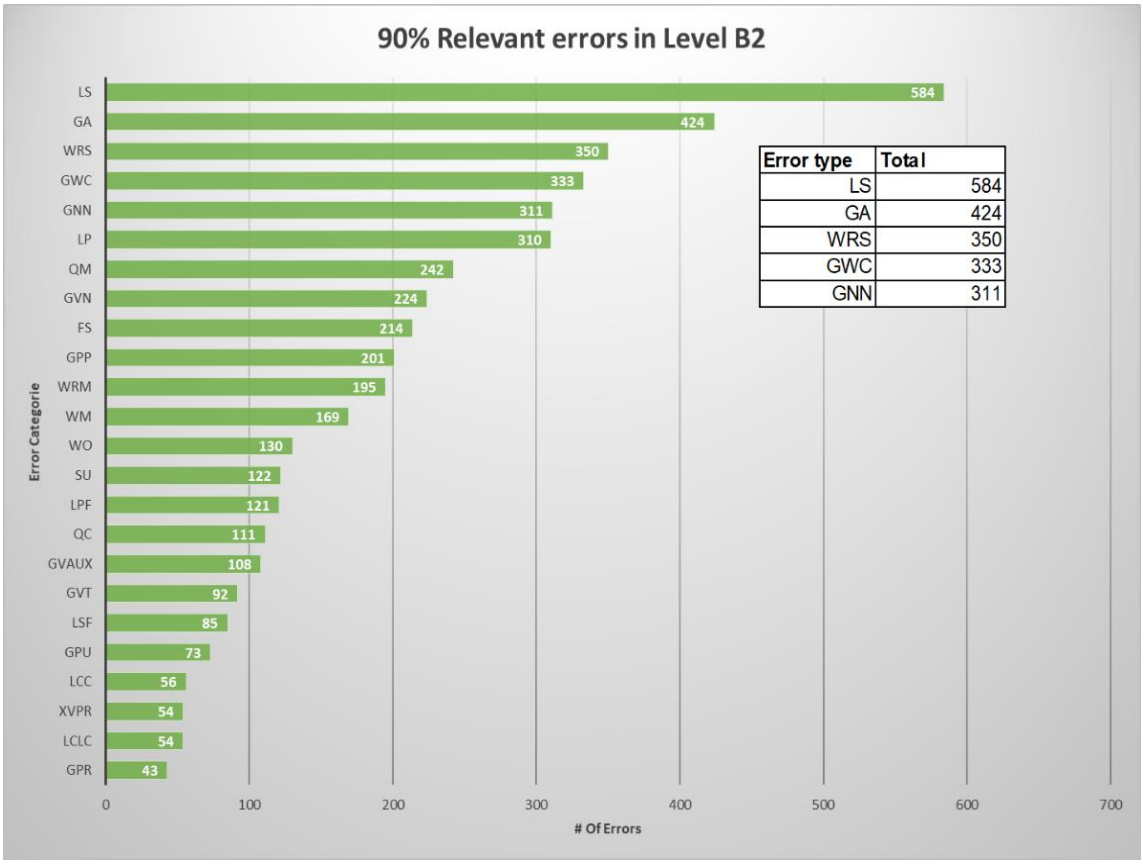



Figure 51. Most frequent written errors found in level B2.

Table 66 shows the comparative of errors in levels B1 and B2.

Table 66

*Comparative chart of errors in B1 to B2*

| **Level B1.1** | | **Level B1.2** | | **Level B1.3-B2.1** | | **Level B2.2-B2.3** | |
|---|---|---|---|---|---|---|---|
| **E. Type** | **%** | **E. Type** | **%** | **E. Type** | **%** | **E. Type** | **%** |
| FS | 18.39 | FS | 16.17 | FS | 20.31 | LS | 11 |
| GA | 14.74 | GA | 11.19 | GA | 15.66 | GA | 0.08 |
| LS | 7.77 | QM | 6.49 | GNN | 7.57 | WRS | 0.07 |
| GNN | 6.58 | LS | 5.97 | LS | 6.20 | GWC | 0.07 |
| LP | 6.11 | LP | 5.49 | SU | 6.02 | GNN | 0.06 |

Comparing the total results from B1 and B2, it can be observed that there are errors with a prevalence in both levels, which is the case of GA (Grammar, Article), LS (Lexical Single) and GNN (Grammar, Nouns, Number). LS is a recurrent error in all the stages of learning that becomes the prominent error in level B2. The change between B1 and B2 is given by the new errors that become visible in B2: WRS (Word redundant Single) and GWC (Grammar, Word Class), all the previous ones appeared at least once.

Table 67 shows the type of errors that according to English Profile (UCLES/CUP, 2011) improve significantly between B1 to B2.

Table 67

*Errors that improve from B1 to B2*

| | **Error type** |
|---|---|
| **1.** | **Derivation of Conjunction (Link Word)**<br>Where a conjunction / link word resembles, or includes the stem of, a valid word but has been incorrectly derived, usually because it has been given an incorrect affix, it is a Derivation of Conjunction error. |
| **2.** | **Derivation of Determiner**<br>Where a determiner resembles, or includes the stem of, a valid determiner but has been incorrectly derived, usually because it has been given an incorrect affix, it is a Derivation of Determiner error. |
| **3.** | **Form of Determiner**<br>When the articles ‘a’ and ‘an’ are confused. |
| **4.** | **Inflection of Determiner**<br>When the learner has created a feasible but non-valid inflected form of the determiner, usually because of a mistaken belief that the determiner must agree in number with the noun which it precedes. |
| **5.** | **Inflection of Quantifier**<br>When the learner has created a feasible but non- valid inflected form of the quantifier. |
| **6.** | **Inflection of Verb**<br>When the learner has made a false assumption about whether a verb is regular or irregular and inflected it accordingly. Most commonly, the error is caused by putting regular inflections on irregular verbs. |

Source: table and data from UCLES/CUP, (2011)

The errors from the present corpus in levels B1- B2 do not really match the main, typically expected errors from CEFR.

### 4.3.6. Overview of errors by strata

One of the objectives from this research was to determine if involvement in leisure events such as travel abroad or access to private schools was related to the written production of English as a foreign language. Since involvement in such leisure events is determined by the variable of socio-economic stratum, a hypothesis was set up and students answer a survey that included sociocultural facts, described in section 4.2. The hypothesis was: "There are statistically significant differences in the median of errors from written production of EFL university students in relation to the stratum classification given in Colombia."

The variable of stratum was unfortunately affected by the fact that the students from this research in most cases did not confirm the strata in their writings as it was requested. From 515 participant students in this research only 191 mentioned their strata. Table 68 shows the distribution of students who confirmed their strata.

Table 68

*Distribution of students that confirmed their strata*

| | **Total Students** | **Males who confirmed** | **Females who confirmed** | **Total** | **% per level** |
|---|---|---|---|---|---|
| B1 | 403 | 55 | 50 | 105 | 26% |
| B2 | 112 | 43 | 43 | 86 | 77% |
| | | | | Total: 191 | |

Having in mind that from level B1 only 26% of students confirmed their strata the analysis from this level will not be conclusive because the sample is not enough. Nevertheless, for the case of level B2 as 77% of students answer their strata, it is possible to draw some conclusions.

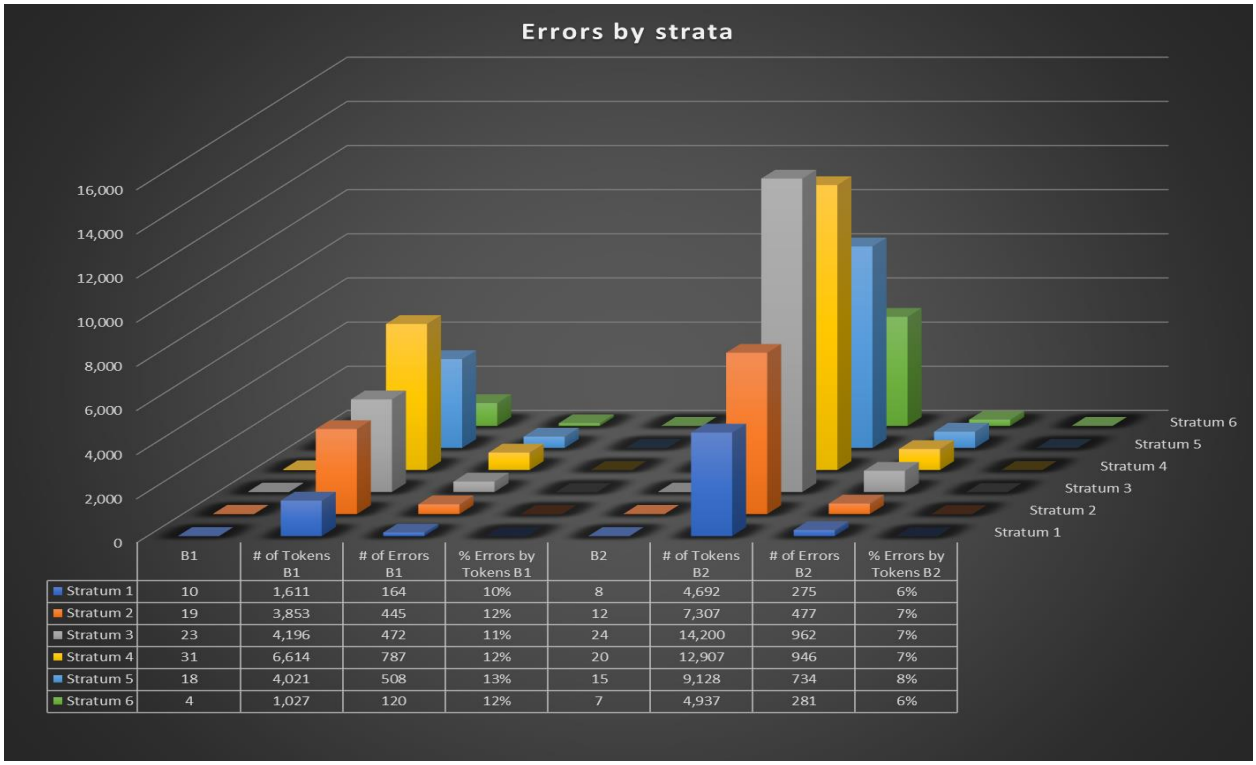
Students from this research belonged to six different strata as follows:

Table 69

*Classification of students by strata*

| **Stratum** | **Males L. B1** | **Females L. B1** | **Males L. B2** | **Females L. B2** |
|---|---|---|---|---|
| 1 | 7 | 3 | 7 | 1 |
| 2 | 9 | 10 | 4 | 8 |
| 3 | 14 | 9 | 15 | 9 |
| 4 | 19 | 12 | 7 | 13 |
| 5 | 3 | 15 | 7 | 8 |
| 6 | 6 | 1 | 3 | 4 |
| Total | 55 | 50 | 43 | 43 |

The distribution of errors according to the strata is as follows:

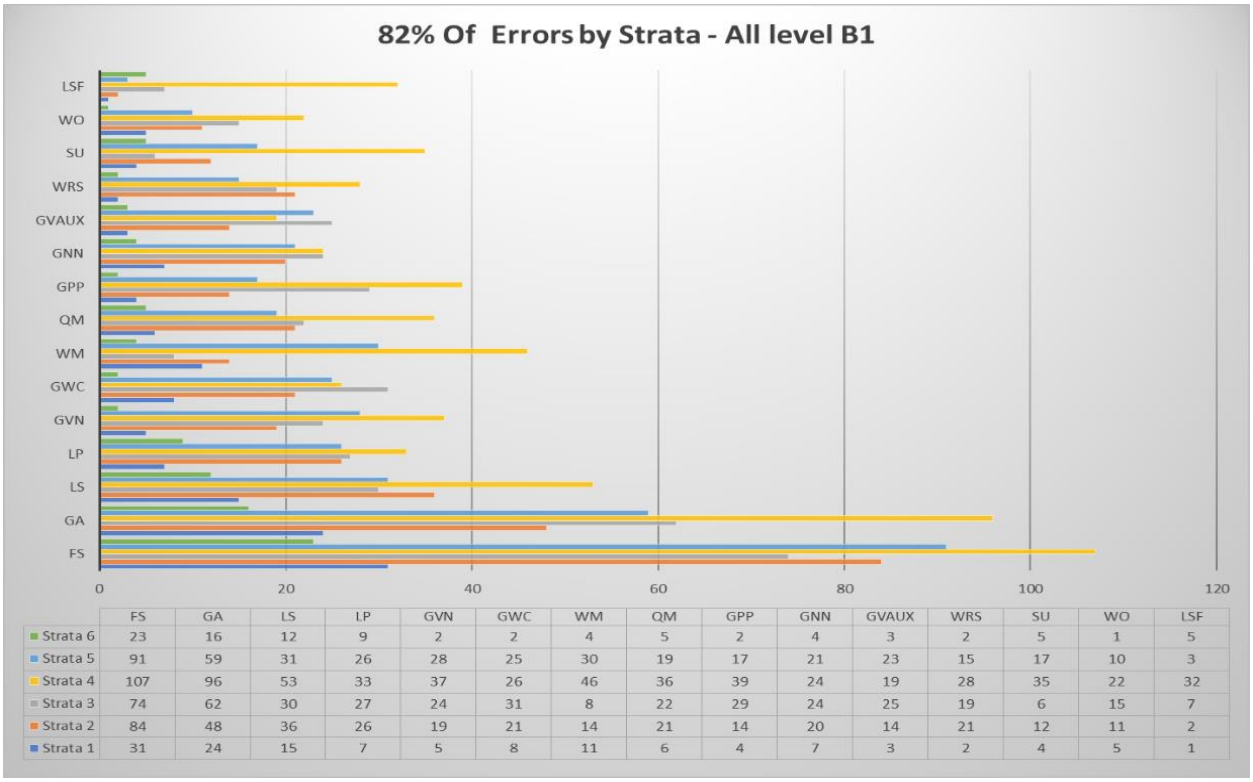



| | B1 | # of Tokens B1 | # of Errors B1 | % Errors by Tokens B1 | B2 | # of Tokens B2 | # of Errors B2 | % Errors by Tokens B2 |
|---|---|---|---|---|---|---|---|---|
| Stratum 1 | 10 | 1,611 | 164 | 10% | 8 | 4,692 | 275 | 6% |
| Stratum 2 | 19 | 3,853 | 445 | 12% | 12 | 7,307 | 477 | 7% |
| Stratum 3 | 23 | 4,196 | 472 | 11% | 24 | 14,200 | 962 | 7% |
| Stratum 4 | 31 | 6,614 | 787 | 12% | 20 | 12,907 | 946 | 7% |
| Stratum 5 | 18 | 4,021 | 508 | 13% | 15 | 9,128 | 734 | 8% |
| Stratum 6 | 4 | 1,027 | 120 | 12% | 7 | 4,937 | 281 | 6% |

Figure 52. Distribution of errors by strata in levels B1 and B2.

It can be observed that errors with most incidence in level B1 and B2 are distributed in strata 3 and 4. The stratum with the least incidence of errors per 100 tokens is stratum 1 in both levels.

### 4.3.6.1. Analysis of errors by strata in level B1




| | FS | GA | LS | LP | GVN | GWC | WM | QM | GPP | GNN | GVAUX | WRS | SU | WO | LSF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Strata 6 | 23 | 16 | 12 | 9 | 2 | 2 | 4 | 5 | 2 | 4 | 3 | 2 | 5 | 1 | 5 |
| Strata 5 | 91 | 59 | 31 | 26 | 28 | 25 | 30 | 19 | 17 | 21 | 23 | 15 | 17 | 10 | 3 |
| Strata 4 | 107 | 96 | 53 | 33 | 37 | 26 | 46 | 36 | 39 | 24 | 19 | 28 | 35 | 22 | 32 |
| Strata 3 | 74 | 62 | 30 | 27 | 24 | 31 | 8 | 22 | 29 | 24 | 25 | 19 | 6 | 15 | 7 |
| Strata 2 | 84 | 48 | 36 | 26 | 19 | 21 | 14 | 21 | 14 | 20 | 14 | 21 | 12 | 11 | 2 |
| Strata 1 | 31 | 24 | 15 | 7 | 5 | 8 | 11 | 6 | 4 | 7 | 3 | 2 | 4 | 5 | 1 |

Figure 53. Distribution of errors by strata in level B1.

The following table shows the number of tokens and errors according to the strata for males in level B1.

Table 70

*Errors and social strata from male students in B1*

| Strata | Male B1 | Tokens | # OfErrors | % Errors |
|---|---|---|---|---|
| 1 | 7 | 1,076 | 122 | 11.3% |
| 2 | 9 | 1,753 | 186 | 10.6% |
| 3 | 14 | 2,612 | 281 | 10.8% |
| 4 | 19 | 3,747 | 422 | 11.3% |
| 5 | 3 | 587 | 80 | 13.6% |
| 6 | 3 | 823 | 115 | 14.0% |
| | 55 | 10,598 | 1,206 | |
| 0 | 127 | 23,887 | 2,992 | |

The percentage of error for every 100 tokens increases with the stratum. While in stratum 1 errors were 11% for every 100 tokens in stratum 5 and 6 the percentage increases to 13 and 14 respectively. The last file with zero refers to the males from B1 who did not report their stratum. As mentioned before, some participants did not report their strata when they did their final work.

The following table shows the number of tokens and errors according to the strata for females in level B1.

Table 71

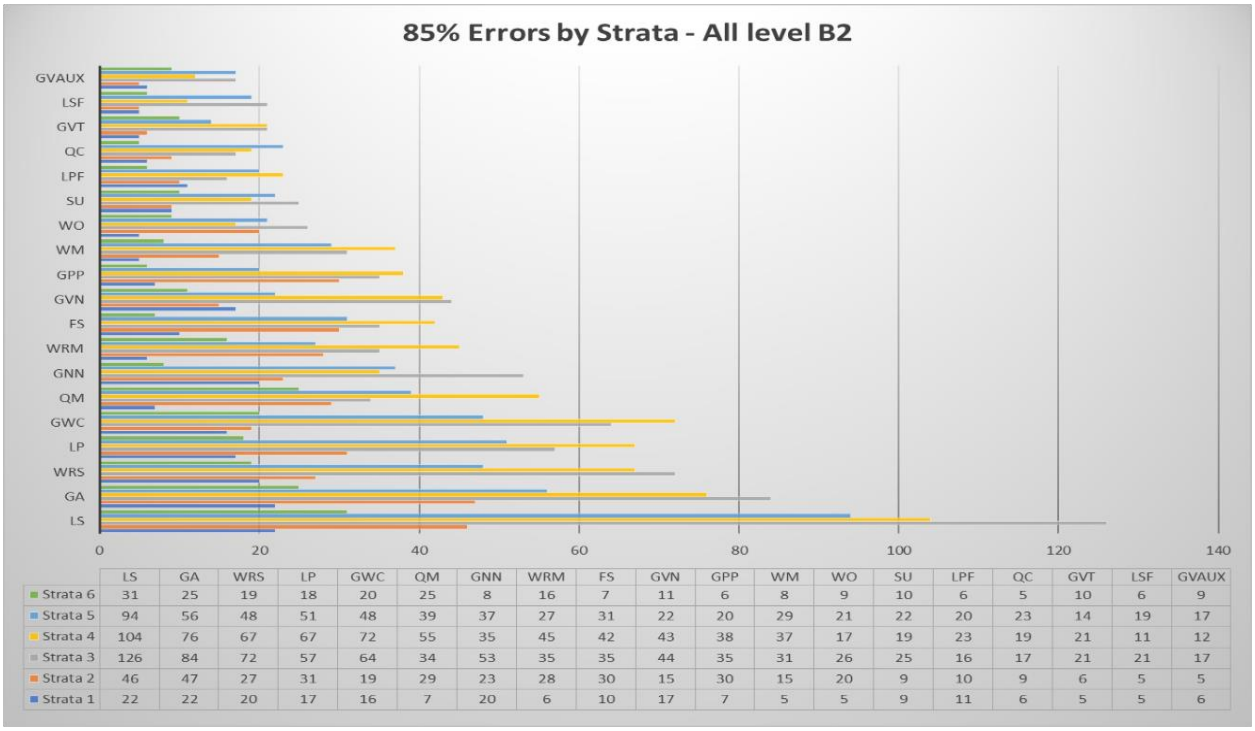
*Errors and strata from female students in B1*

| Strata | Female B1 | Tokens | # OfErrors | % Errors |
|---|---|---|---|---|
| 1 | 3 | 535 | 42 | 7.9% |
| 2 | 10 | 2,100 | 259 | 12.3% |
| 3 | 9 | 1,584 | 191 | 12.1% |
| 4 | 12 | 2,867 | 365 | 12.7% |
| 5 | 15 | 3,434 | 428 | 12.5% |
| 6 | 1 | 204 | 5 | 2.5% |
| | 50 | 10,724 | 1,290 | |
| 0 | 169 | 34,450 | 3,992 | |

Again, as for the male analysis, the percentage of error for every 100 tokens increased with the stratum. In this case, the percentage of errors increased from stratum 2 and kept steady until level 5. In level 6, there was only one female reported. For that reason, the percentage decreased. The last file with zero refers to the females from B1 who did not confirm their stratum.

### 4.3.6.2. Analysis of errors by social strata in level B2

Since 77% of students from level B2 confirmed their strata, the following section will present the findings in graphics that show an overview of errors by strata and then the type of errors and quantities in every stratum.




| | LS | GA | WRS | LP | GWC | QM | GNN | WRM | FS | GVN | GPP | WM | WO | SU | LPF | QC | GVT | LSF | GVAUX |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Strata 6 | 31 | 25 | 19 | 18 | 20 | 25 | 8 | 16 | 7 | 11 | 6 | 8 | 9 | 10 | 6 | 5 | 10 | 6 | 9 |
| Strata 5 | 94 | 56 | 48 | 51 | 48 | 39 | 37 | 27 | 31 | 22 | 20 | 29 | 21 | 22 | 20 | 23 | 14 | 19 | 17 |
| Strata 4 | 104 | 76 | 67 | 67 | 72 | 55 | 35 | 45 | 42 | 43 | 38 | 37 | 17 | 19 | 23 | 19 | 21 | 11 | 12 |
| Strata 3 | 126 | 84 | 72 | 57 | 64 | 34 | 53 | 35 | 35 | 44 | 35 | 31 | 26 | 25 | 16 | 17 | 21 | 21 | 17 |
| Strata 2 | 46 | 47 | 27 | 31 | 19 | 29 | 23 | 28 | 30 | 15 | 30 | 15 | 20 | 9 | 10 | 9 | 6 | 5 | 5 |
| Strata 1 | 22 | 22 | 20 | 17 | 16 | 7 | 20 | 6 | 10 | 17 | 7 | 5 | 5 | 9 | 11 | 6 | 5 | 5 | 6 |

Figure 54. Distribution of errors by strata in level B2.

The following figures show the distribution of errors by every stratum in level B2.

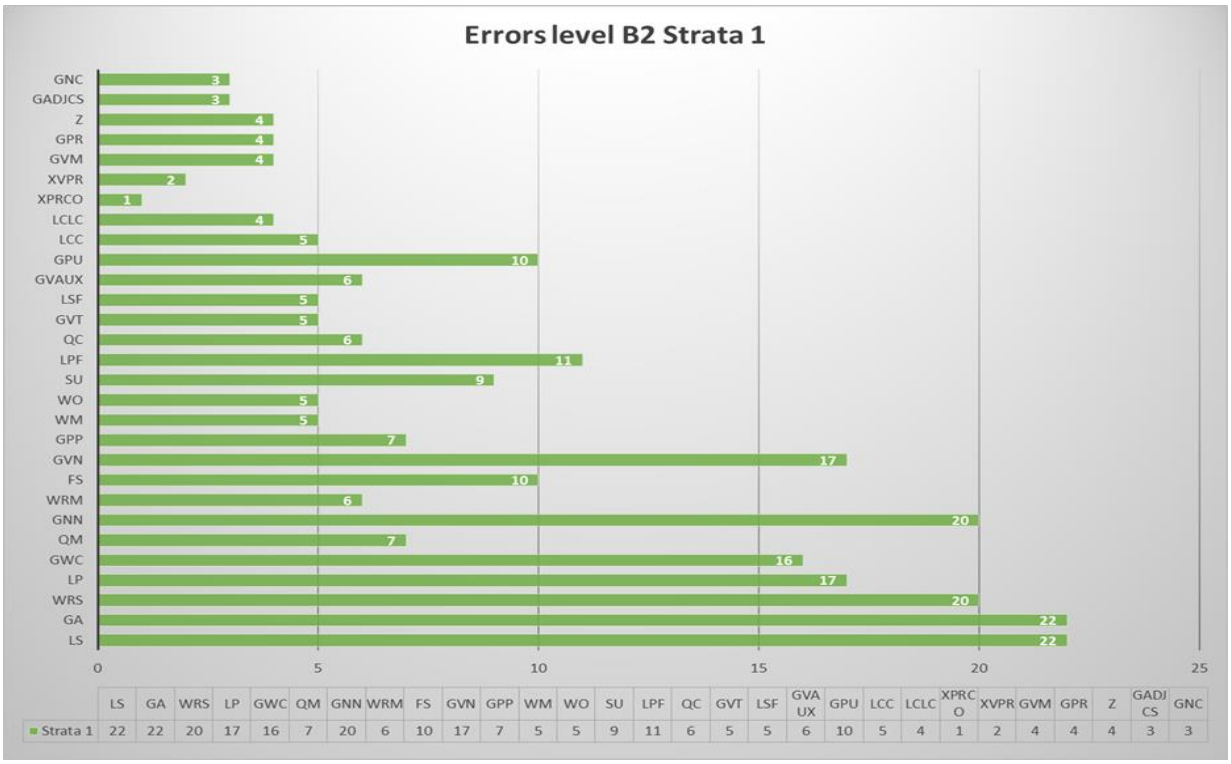



Figure 55. Distribution of errors level B2 stratum 1.

It can be observed that the three errors with most incidence in stratum 1 from level B2 are Lexical Single, Grammar Article and Word Redundant Single.

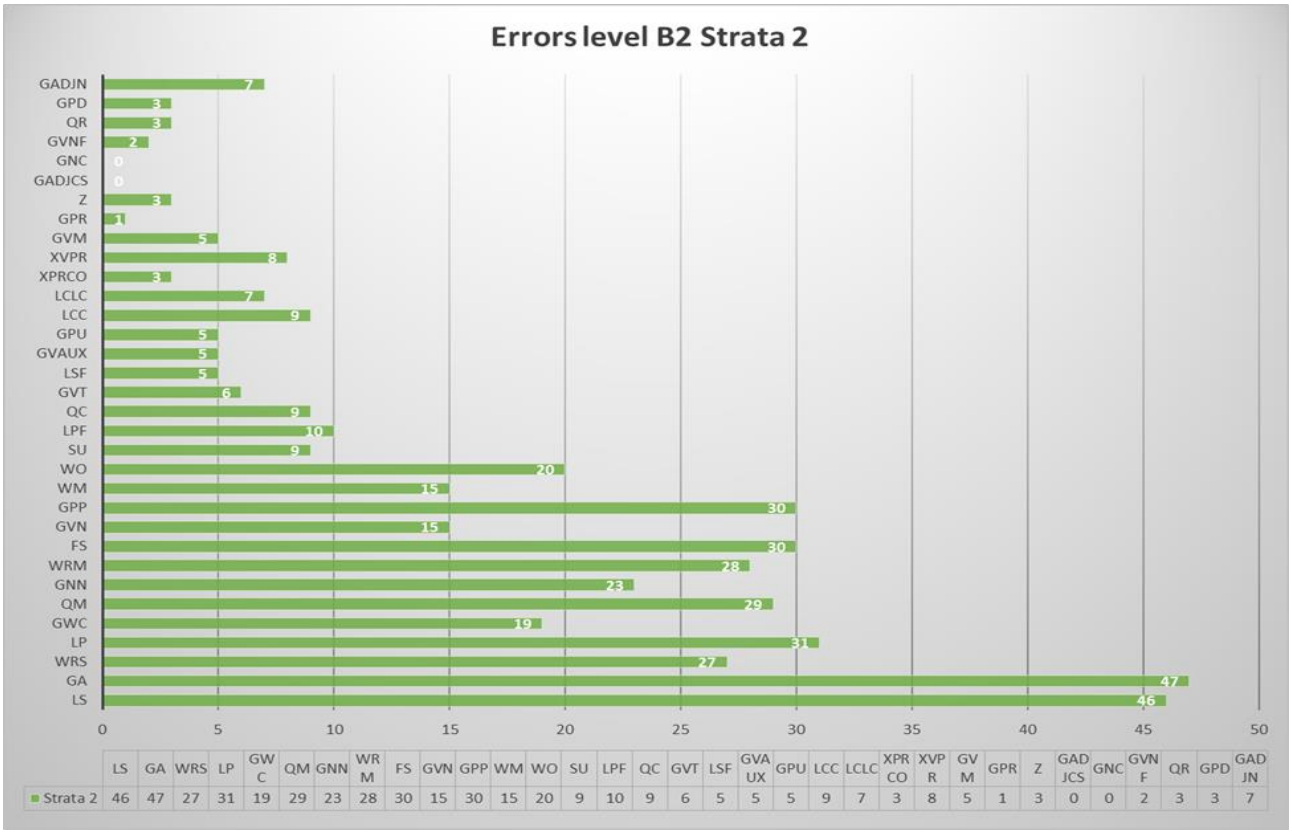



Figure 56. Distribution of errors level B2 stratum 2.

In the case of stratum 2 for the same level B2, the first two errors with most incidence remain the same as in stratum 1, Lexical Single, Grammar Article, but the third error with most incidence is Word Redundant Single.

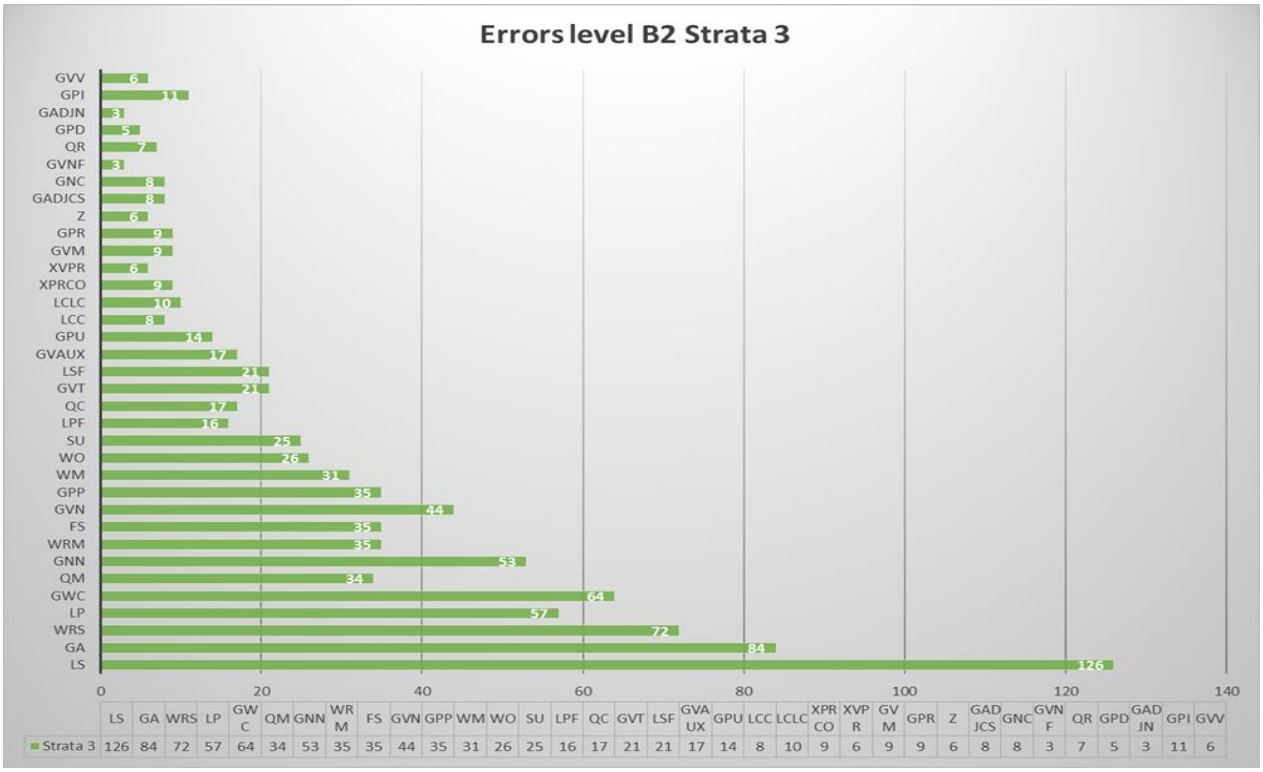


Figure 57. Distribution of errors level B2 stratum 3.

Stratum 3 from level B2 presents exactly the same first three errors as stratum 2 and stratum 4. The only difference in these strata are presented in the quantity of errors, not in the type.

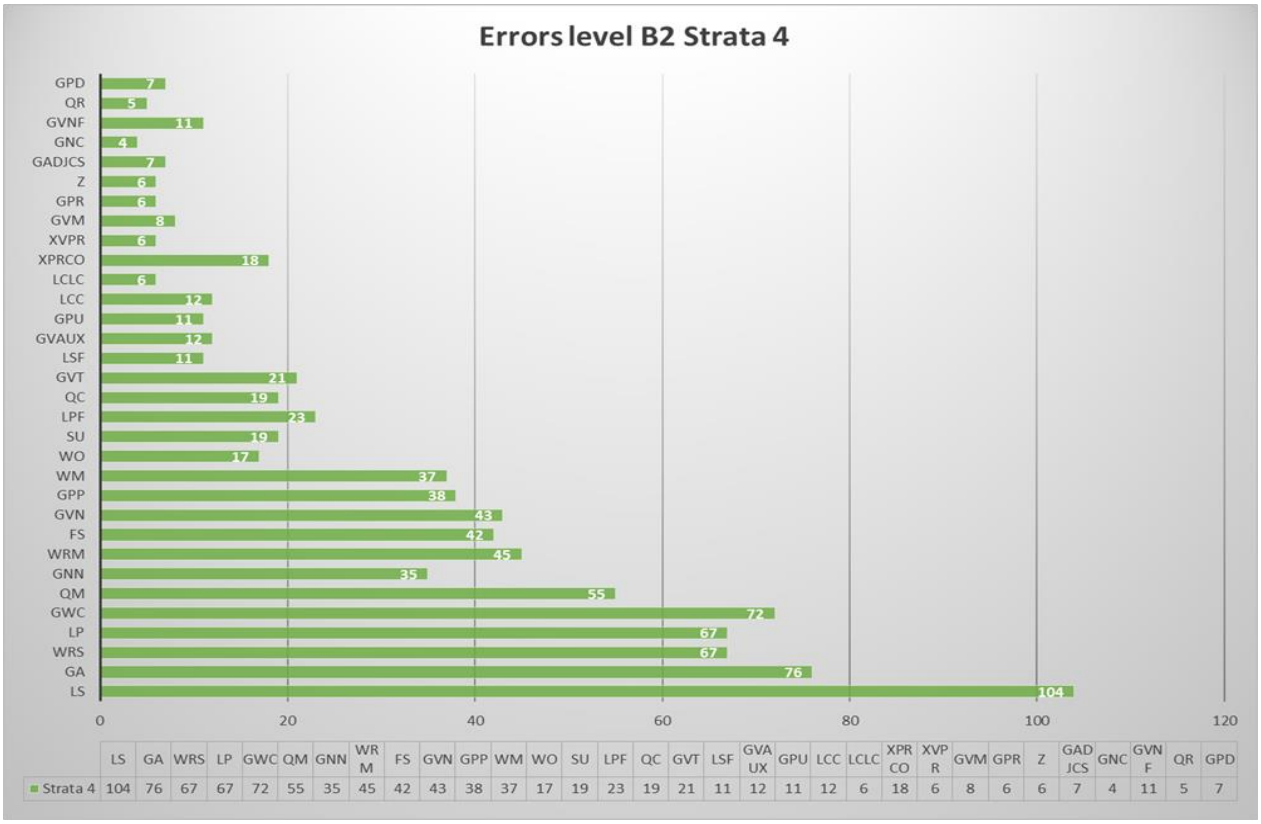



Figure 58. Distribution of errors level B2 stratum 4.

Error Lexical Single decreases from 126 in stratum 3 to 104 in stratum 4.

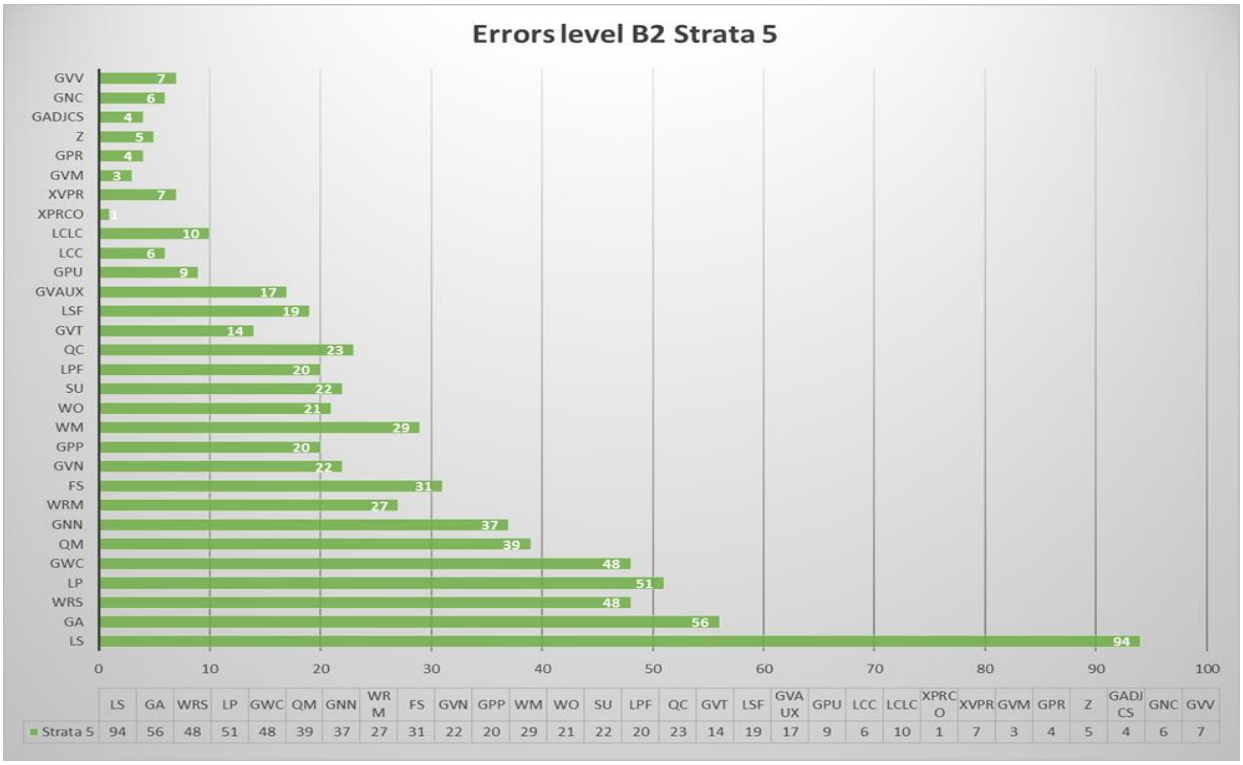



Figure 59. Distribution of errors level B2 stratum 5.

In stratum 5 the same first 3 errors persist, and the percentages decrease as an indicator that students have improved their interlanguage.

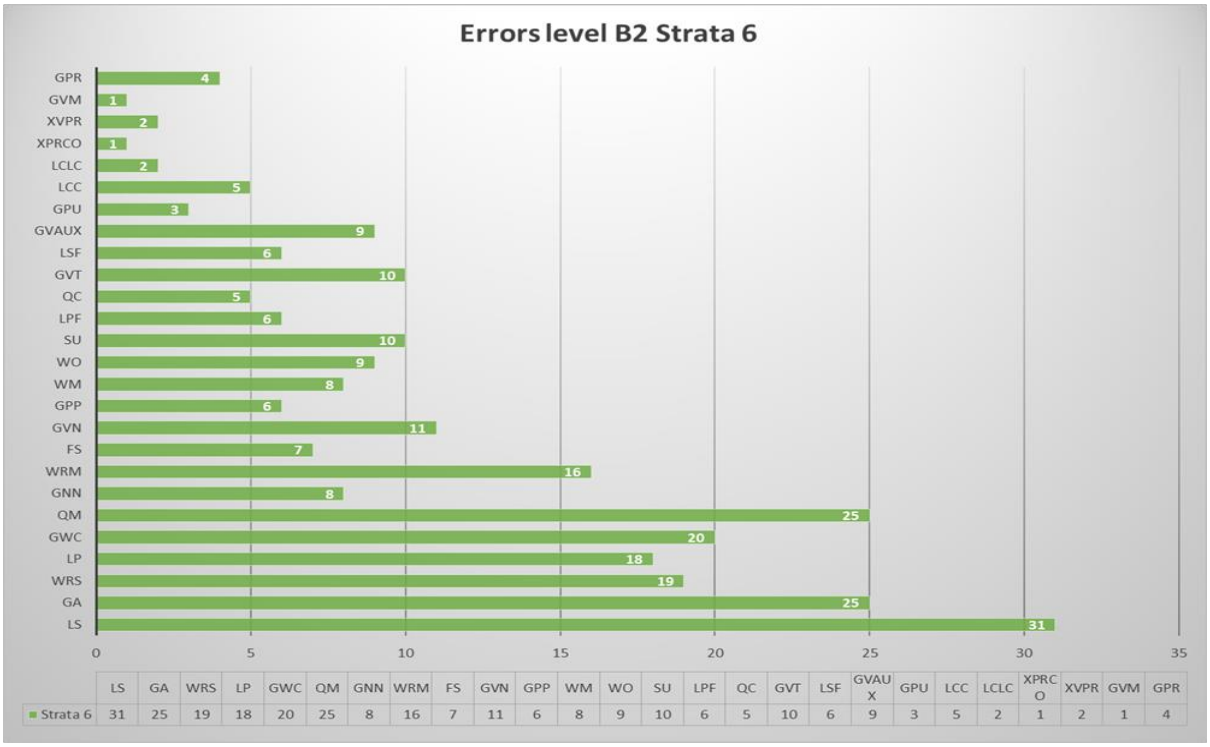



Figure 60. Distribution of errors level B2 stratum 6.

The final figure from stratum 6 shows exactly the same three types of errors with most incidence, but in this case, the incidence is the lowest from all strata.

The following table shows the number of tokens and errors according to the strata for males in level B2.

Table 72

*Errors and strata from male students in B2*

| Strata | Male B2 | Tokens | # Of Errors | % Errors |
|---|---|---|---|---|
| 1 | 7 | 3,871 | 224 | 5.8% |
| 2 | 4 | 2,522 | 158 | 6.3% |
| 3 | 15 | 8,311 | 600 | 7.2% |
| 4 | 7 | 4,998 | 352 | 7.0% |
| 5 | 7 | 4,339 | 360 | 8.3% |
| 6 | 3 | 2,095 | 151 | 7.2% |
| | 43 | 26,136 | 1,845 | |
| 0 | 13 | 7,589 | 681 | |

The percentage of error for every 100 tokens increases along with the strata. While in social stratum 1 errors were 5.8% for every 100 tokens in stratum 5 the percentage reached 8.3% and in social stratum 6 decreased approximately one point. Therefore, the general tendency and the tendency for males in level B2 is that errors increase when social stratum is higher. It might mean that students from higher strata are overconfident in their abilities with the foreign language for the possibility to travel abroad. Students from lower strata are less confident and probably this lack of confidence make them study more. Regarding the average of written production in the different strata, it can be observed that the number of tokens increased along with the stratum. The following list shows how the mean of tokens vary per male student in level B2.

- Stratum 1 *M=553*
- Stratum 2 *M=630*
- Stratum 3 *M= 554*

- Stratum 4 *M= 714*
- Stratum 5 *M= 619*
- Stratum 6 *M= 698*

It can be said that, in higher strata, males increased their written production and the most productive strata were 4 and 6.

The last file with zero refers to the males from B1 who did not confirm their strata.

Table 73

*Errors and strata from female students in B2*

| Strata | Female B2 | Tokens | # Of Errors | % Errors |
|---|---|---|---|---|
| 1 | 1 | 821 | 51 | 6.2% |
| 2 | 8 | 4,785 | 319 | 6.7% |
| 3 | 9 | 5,889 | 362 | 6.1% |
| 4 | 13 | 7,909 | 594 | 7.5% |
| 5 | 8 | 4,789 | 374 | 7.8% |
| 6 | 4 | 2,842 | 130 | 4.6% |
| | 43 | 27,035 | 1,830 | |
| 0 | 13 | 8,293 | 728 | |

In this case, errors also gradually increased with the stratum. Females from level B2 started with a percentage of 6.2% errors per 100 tokens in stratum 1 and increased until reaching 7.8% errors in stratum 5. In stratum six, the percentage of errors per 100 tokens decreased to 4.6% which is an indicator that in stratum 6, females have a better performance in written output than the lower strata. The average of tokens was varied in the different strata. The following list shows how the mean of tokens vary per female student in level B2.

- Stratum 1 *M=821*
- Stratum 2 *M=598*
- Stratum 3 *M= 654*
- Stratum 4 *M= 608*
- Stratum 5 *M= 598*
- Stratum 6 *M= 710*

It can be said that the tendency is to have more production in higher strata, but in this case, one student from stratum 1 overpassed the mean from all strata which is an exception to the tendency in the sample. The other students kept increasing their production along with their strata. The most productive strata in this case are stratum 1 and 6. The tendency is quite different from the one from male counterparts.

Table 74

*Main five errors and quantities in B2*

| **Strata** | **Error** | **Qty** | **Error** | **Qty** | **Error** | **Qty** | **Error** | **Qty** | **Error** | **Qty** |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | LS | 22 | GA | 22 | WRS | 20 | LP | 17 | GWC | 16 |
| 2 | LS | 46 | GA | 47 | WRS | 27 | LP | 31 | GWC | 19 |
| 3 | LS | 126 | GA | 84 | WRS | 72 | LP | 57 | GWC | 64 |
| 4 | LS | 104 | GA | 76 | WRS | 67 | LP | 67 | GWC | 72 |
| 5 | LS | 94 | GA | 56 | WRS | 48 | LP | 51 | GWC | 48 |
| 6 | LS | 31 | GA | 25 | WRS | 19 | LP | 18 | GWC | 20 |

When observing the main five errors by stratum in level B2, the tendency is to have fewer errors in strata 1 and 2, then in strata 3 to 5 errors increase and again in strata 6 errors decrease.

**Conclusions on the findings of errors by strata**

This section presented the data to test the hypothesis about the relationship between the median of errors from written production of EFL university students in relation to the strata classification given in Colombia. On the one hand, the sample is divided in levels B1 and B2 and the researcher established that few students (26%) from level B1 confirmed their strata; therefore, results in this level are not conclusive. On the other hand, 77% of students from level B2 confirmed their strata; therefore, results in this level are statistically significant. The following are the most relevant aspects in the findings from level B2.

1. In level B2, male and female students increased the mean of errors per 100 tokens along with the strata
2. The written production in B2 was more abundant in higher than in lower strata except for one female student that overpassed the production from the group
3. The strata with the highest number of errors in B2 level for males was 3 and for females was 4
4. In level B2, errors increased for males and females in stratum 3 and 4
5. In level B2, errors decreased in stratum 6 for females but increased for males.

Therefore, it can be said that in level B2 there are some differences in the median of errors from written production of EFL university students in relation to the stratum classification given in Colombia(DANE, n.d.). The higher the stratum, the higher the incidence of errors. These results might show how students who come from lower strata schools, with a low English level,

study more to keep up with their peers and to achieve the EFL required proficiency. Another possible cause is that students from higher strata who have travelled abroad believe they have a better level of English and are overconfident; therefore, they rely on informal English empirically learned. This findings contradict the theories from Lin (1999, p.407) claiming "the middle-class students brought with them the right kind of habitus (i.e. cultural capital). They had the correct attitudes and interest and the correct linguistic skills." And the studies from Vandrick, (1995); Arikan, (2010); Morales, (2017) claiming that privileged students from higher social strata might have better EFL proficiency for the opportunity to practice English language learned in class. On the contrary, the most successful students were in lowest strata.

### 4.4. Incidence of errors by category: Overview of eight error categories

From the eight error categories analysed in the written production of university students from *Universidad del Norte* in Barranquilla, Colombia. The following are the results found first as a great total and then in all categories. Further analysis will be done in every category individually.

The most frequent errors belong to different categories. Even though grammatical category accounted for 42% of the total of errors in the corpus, only one type of error from this class GA (Grammar, Article) accounted for 13% and the others did not account for more than 0.05% which is the case, for instance of GNN (Grammar, Noun Number). Let us see the graphic.

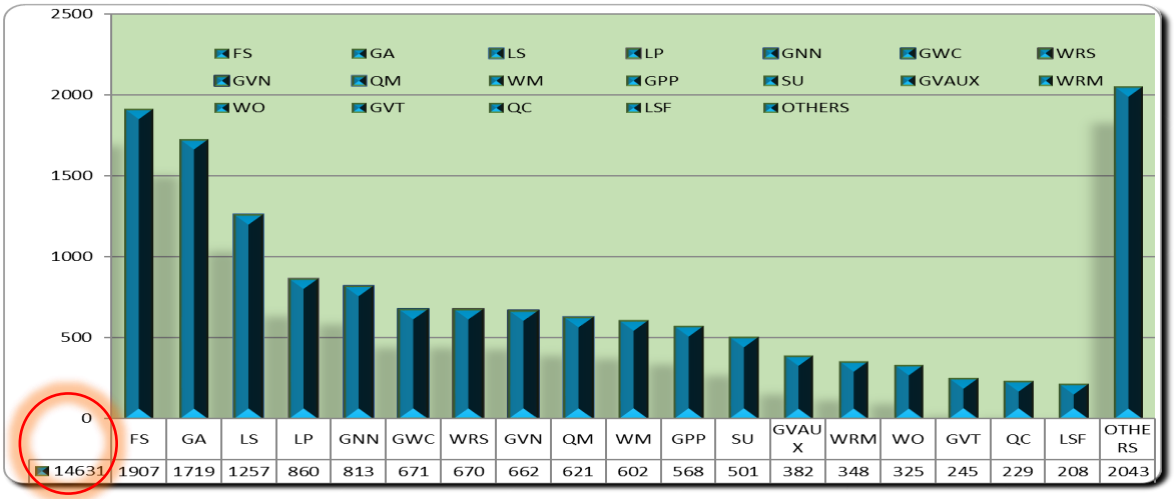



Figure 61. Most relevant errors by type.

Graphic 61 shows 18 types of the most relevant errors found in the written corpus. The column entitled as "others" refer to other errors with less incidence that cannot be viewed in this screenshot. Errors in this graphic belong to different categories.

Error categories analysed in the present research are the ones proposed by Louvain University, introduced in section 3.1.1.

Table 75

*Categories of errors*

| **Categories of errors** | **Code (letter that stands for error category)** |
|---|---|
| Form errors | F |
| Grammatical errors | G |
| Lexico-Grammar errors | X |
| Lexis errors | L |
| Word errors | W |
| Punctuation errors | Q |
| Style errors | S |
| Infelicities | Z |

Source: data obtained from (Dagneaux et al., 2005).

The great total of errors from eight categories proposed in the tagger added up 14,631 in a corpus of 515 files, 149,325 tokens, 12,337 lemmas and 12,164 types. Figure 62 shows the total of errors by category and Table 76 presents the percentages and means in each category.

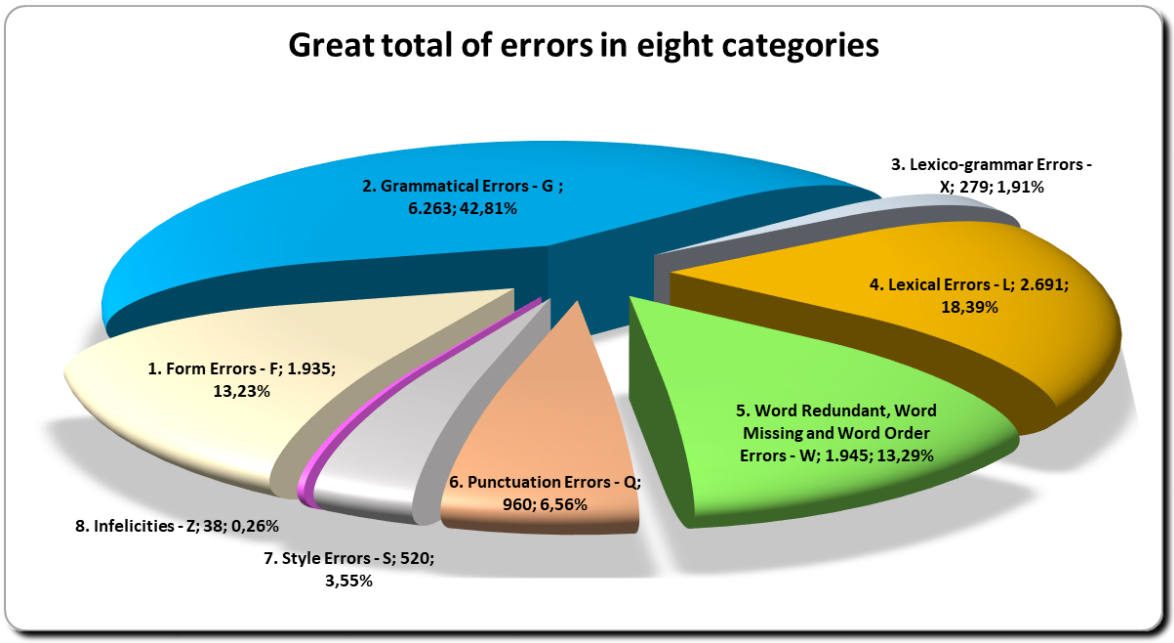



Figure 62. Great total of errors in eight categories.

Table 76

*Total of errors by categories with percentages and means in corpus*

| Error category | | Number of errors | Percentage % | Mean in corpus |
|---|---|---|---|---|
| Grammatical | (G) | 6,263 | 42.81 | *M*=12.16 |
| Lexis | (L) | 2,691 | 18.39 | *M*=5.22 |
| Words | (W) | 1,945 | 13.29 | *M*=3.77 |
| Form | (F) | 1,935 | 13.23 | *M*=3.75 |
| Punctuation | (Q) | 960 | 6.56 | *M*=1.86 |
| Style | (S) | 520 | 3.55 | *M*=1.00 |
| Lexico-grammar | (X) | 279 | 1.91 | *M*=0.54 |
| Infelicities | (Z) | 38 | 0.26 | *M*=0.073 |
| Total of errors: | | 14,631 | 100.0 | |

Figure 62 is a summary of the whole corpus that shows, from the highest to the lowest, the errors with the most and least incidences. In this case, the category of Grammar Errors has the highest incidence.

Table 77 shows the number of errors in each category to see their incidence in the corpus.

Table 77

*Numbers of errors tags*

| Categories of errors | Code | Number of type of errors |
|---|---|---|
| Form errors | F | 3 |
| Grammatical errors | G | 25 |
| Lexico-Grammar errors | X | 9 |
| Lexis errors | L | 8 |
| Word errors | W | 4 |
| Punctuation errors | Q | 4 |
| Style errors | S | 2 |
| Infelicities | Z | 1 |

Source: data retrieved from (Dagneaux et al., 2005).

Grammar errors account for 42% of errors in the whole corpus. This result is also related to the fact that the category of Grammar Errors has 25 type of errors from the list of 56 errors from all categories. Therefore, this category has a high incidence in the total results. In some way, the proportion of errors from each category vary according to the number of error types from each category.

Every category of errors will be analysed separately to have detailed information of the results. The description of errors is based on the categories and subcategories proposed by the Error Tagger from Louvain University, version 2. (E. Dagneaux et al., 2005)

### 4.4.1. Analysis of Grammatical errors (G)

The first category with most incidence within the corpus was Grammar (G). The category of grammatical errors refers to errors that break the rules of English grammar. This category is divided into 25 error types.

Errors in this category accounted for 6,263 hits that make 42% of the corpus. Grammar errors were dispersed within the 25 error types from this category.

Table 78

*Five errors with most incidence in the grammar category*

| Error type | Number of errors | Percentage in corpus | Mean per student |
|---|---|---|---|
| GA | 1719 | 11 | *3.33* |
| GNN | 813 | 0.05 | *1.57* |
| GWC | 671 | 0.04 | *1.30* |
| GVN | 662 | 0.045 | *1.28* |
| GPP | 568 | 0.038 | *1.10* |

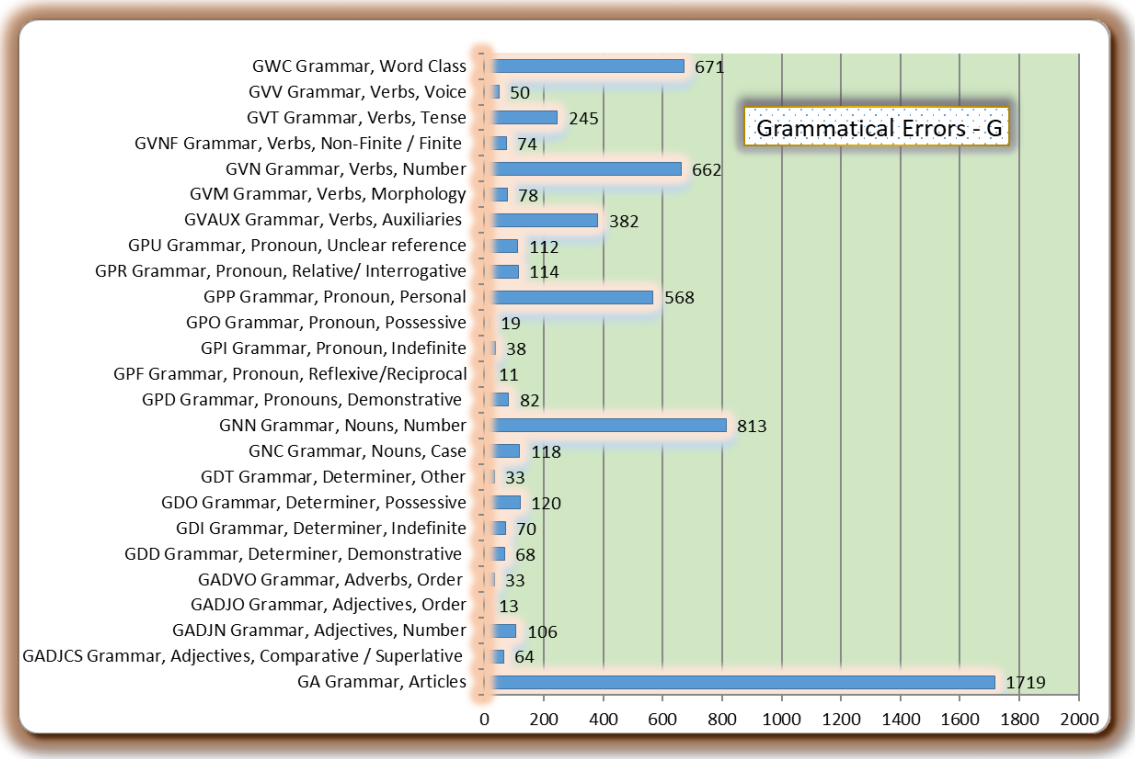



Figure 63. Account of grammatical errors.

The analysis from this category will be based on the first five error types that had the highest scores. The error with most incidence in this category is GA (Grammar, Articles) with 1,719 errors and 11% in the whole corpus. This error type accounts for all problems in definite, indefinite or zero article. The following are some examples from the corpus.[4]

1. 7496 before, because now you have **(GA)** 0 $a$ compromise with you
2. 7526 , can be terrible at the end. **(GA)** The $0$ people that get m
3. 3808 more different ways to enjoy **(GA)** the $0$ life, the$0
4. 7562 they think that **(GA)** the $0$ capital punishment
5. 4005 sionate people. They know that (**GA)** the$0$ emotions are in the
6. 9186 children leave their nests. **(GA)** The $0$ tourists always l
7. 9717 g and it is dangerous because **(GA)** a $an$ animal can try to e

[4] http://grupotnt.udea.edu.co/django/clec/corpus/

As can be seen in examples 1-6, this type of error has different ways to appear. In example 1, there is an omission of the indefinite article. In examples 2-6, there is no need to use an article, therefore, those are examples of addition. This kind of error is probably related to the use of articles in Spanish because examples 2-6 would use an article in typical sentences from Spanish. Example 7 is an incorrect use of the indefinite article followed by a vowel sound. From examples 2 to 6, Spanish speakers would say in Spanish:

Table 79

*Examples of GA errors and possible cause*

| English version with correction between $ $ symbols | Possible Spanish version |
|---|---|
| 2. **(GA)** The $0$ people | La gente |
| 3. **(GA)** the $0$ life | La vida |
| 4. **(GA)** the $0$ capital punishment | La pena de muerte |
| 5. **(GA)** the $0$ emotions | Las emosiones |
| 6. **(GA)** The $0$ tourists always l | Los turistas siempre |

Example number 6 is a misuse of the article according to the vowel sound of the word that follows.

The second error with most incidence in this category is GNN (Grammar, Noun, Number) with a total of 813 and an incidence of 0.05% in the corpus. This error type involves errors due to addition or omission of the plural morpheme. The following are some examples from the corpus.

7. 9265 (GA)The$0$ **(GNN**) tourist $tourists$ are ba
8. 9283 r our life$?$. There are many **(GNN)** reason $reasons$ that af
9. 9309 companies use (GA)the $0$ **(GNN)**commercial $commercials$

10. 9354 me I think that (WRS)be $0$ **(GNN)**tourist $tourists$ (GVN$
11. 9822 say in the tv $?$. The final **(GNN)** reasons$reason$ is that
12. 10176 affect this environment. A (**GNN)** consequences $consequence$

Examples 7 to 10 refer to the use of a singular instead of a plural and examples 11 and 12 refer to the use of a plural instead of singular word.

The third error with most incidence in this category is GWC (Grammar Word Class) with 671 and an incidence of 0.04% in the corpus. This error type refers to the inappropriate use of a word class e.g. the use of a noun instead of an adjective, or an adverb instead of an adjective. Let us see some examples from the corpus.[5]

13. 49 be associated as the most **(GWC)** danger $dangerous$ race.
14. 114 acts, abuses and other racial **(GWC)** critics $criticism$. Tha
15. 116 ay$ that (GA) the$0$ equality **(GWC)** beginning $begins$ from
16. 824 0$ generate loss of appetite, **(GWC)** nauseous $nausea$, const
17. 1163 tion (LS) of $from$ the daily **(GWC)** live $life$, in so many
18. 2266 $ feelings like irritation or **(GWC**) angry $anger$ appear. In

Example 13 refers to the use of a noun instead of an adjective in example 14 even though the author uses a noun it is the incorrect one –a critic is the person who criticises, and criticism is what the critics say about something they analyse. Example 15 refers to the use of an adjective instead of a verb. Example 16 ,17, 18 and 19 refer to the use of an adjective instead of a noun.

The fourth error with most incidence in the grammar category is GVN (Grammar, Verbs, Number) with 662 and an incidence of 0.045% in the corpus. This error type accounts for all errors of concord between a subject and its verb e.g., a plural verb with singular subject, a

[5] http://grupotnt.udea.edu.co/django/clec/corpus/

singular verb with plural subject, a singular verb with collective subject. Let us see some examples from the corpus.[6]

| | | |
|---|---|---|
| 19. | 3147 | solve $solved$. Barranquilla **(GVN)** estimate $estimates$ tha |
| 20. | 7019 | Although (QM)0 $,$ everybody **(GVN)** aren't$isn't$ (WRS) of $ |
| 21. | 883 | cancer effects. More evidence **(GVN)** show $shows$ the synergy |
| 22. | 8769 | ries$ (GA) the $0$ traditions **(GVN)** is $are$ very important |
| 23. | 8599 | he (GWC)true $truth$ everyone **(GVN)** have $has$ to pass away |
| 24. | 9058 | things for the planet. People **(GVN)** wants $want$ to help the |

Examples 19 to 21 refer to the use of a plural verb with singular subject. Examples 22 and 23 refer to the use of singular verb with plural subjects and example 24 refers to the use of a singular verb with collective subject.

The fifth error with most incidence in the grammar category is GPP (Grammar Personal Pronoun) with 568 and an incidence of 0.038% in the corpus. This error type accounts for errors that affect personal pronouns and the generic pronoun “one” that means people in general. Let us see some examples from the corpus.

| | | |
|---|---|---|
| 25. | 10187 | oblem is the gasoline because **(GPP)** 0 $it$ is a toxic (FS) s |
| 26. | 10599 | the$ floor. For example, when **(GPP)**0 $they$ go walking and ( |
| 27. | 12413 | , we are expected (WM) 0 $to$ **(GPP)** we $0$ believe that it i |
| 28. | 9892 | $ you could see a doctor that **(GPP)** he$0$ is saying some thi |
| 29. | 8752 | FS) conclutions $conclusion$, **(GPP)** 0 $it$ is important for |
| 30. | 12842 | mercials on TV just show what **(GPP)** it $they$ (XVPR) want 0 |

Examples 25, 26 and 29 refer to omissions of a personal pronoun. Examples 27 and 28 refer to an addition of a personal pronoun. Example 30 refers to the interchange of a pronoun by another one. Example 30 refers to the use of a singular pronoun instead of a plural pronoun.

**Summary of Grammatical Errors category**

[6] http://grupotnt.udea.edu.co/django/clec/corpus/

In this category, the five most relevant errors were presented and their incidence in the whole corpus was analysed along with the percentages and means. GA errors with an incidence of 11% in the corpus are the type of errors with the most relevance in this category. It was analysed how this error type could be caused by an interference in the use of articles from the mother tongue, in this case, Spanish. More errors in this category with less incidence include the wrong use of adjectives, nouns, or determiners, among others. Grammar errors add up 1830 divided into 20 grammar error types.

### 4.4.2. Analysis of Lexis Errors (L)

The second category with most incidence within the corpus was Lexis (L). This category refers to errors involving the semantic properties of words being collocational, of concept or connotative. According to the manual (Estelle et al., 2005), it is divided into three subcategories: Lexical Single, Lexical Phrase, and Connectors. Lexical errors accounted for 2691 that make 18% of the corpus. Errors in this category were dispersed into 6 error types.

Table 80

*First three errors with most incidence in the lexis category*

| Error type | Number of errors | Percentage in corpus | Mean per student |
|---|---|---|---|
| LS | 1257 | 0.085 | *2.44* |
| LP | 860 | 0.058 | *1.66* |
| LSF | 208 | 0.014 | *0.40* |

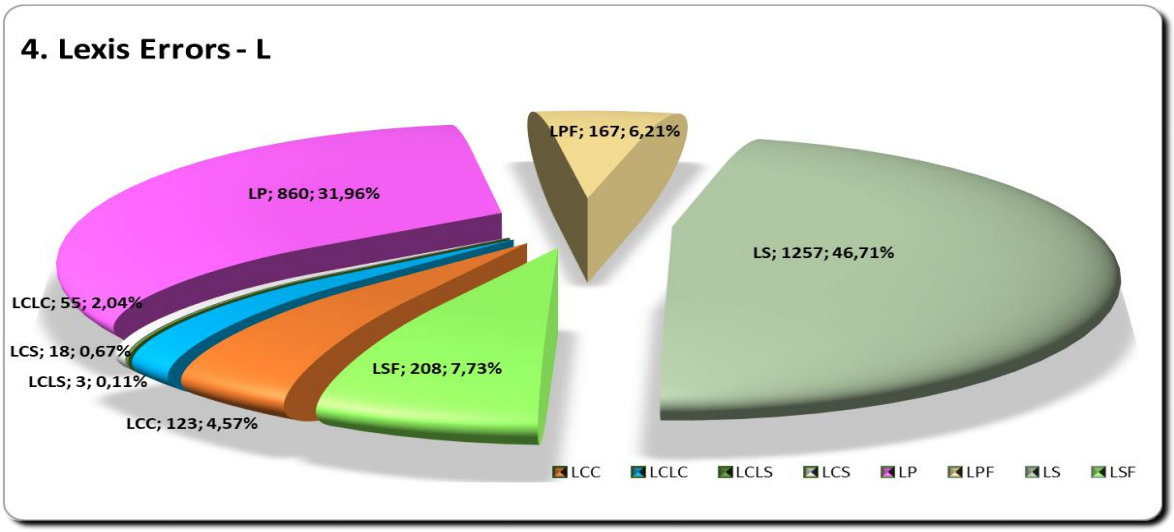



Figure 64. Account of Lexis errors.

The analysis from the present category will be based on the first three error types with most incidence in this category. The error with most incidence in this category is LS (Lexical Single) with 1,257 errors and a percentage of 0.085%. In the corpus. This error type accounts for problems in conceptual, collocational or connotative lexical errors in single words only. This sub-category includes solid and hyphenated compounds. The following are some examples from the corpus.[7]

| | | |
|---|---|---|
| 31. | 1976 | th education, and in having a **(LS)** work $job$ inside your ow |
| 32. | 3372 | tream, (GPP) 0 $they$ help to **(LS)** safe $save$ a car, (LCC) |
| 33. | 6231 | In conclusion (QM)0 $,$ think **(LS)** good $well$ when (GPP) 0 |
| 34. | 5019 | en noted, the risks of being **(LS)** damaged $hurt$ by an injury |
| 35. | 9722 | it and die. For the birds the **(LS)** candies $chewing$ gum tha |
| 36. | 12496 | of Africa. (GPP) 0 $It$ Is a **(LS)** country $continent$ with more |

Examples 31, and 32 refer to conceptual LS problems. Examples 33, 34 and 35 refer to collocational problems. Example 36 refers to a connotative problem. This error shows how the

[7] http://grupotnt.udea.edu.co/django/clec/corpus/

learners have not achieved proficient use of single words due to errors in conceptual, collocational and connotative matters.

The second error with most incidence in this category is LP (Lexical Phrase) with 860 errors and a percentage of 0.058% in the corpus. This error type accounts for three types of problems. 1. Problems in (semi-) fixed multi-word expressions and idioms. 2. When the learner paraphrases instead of using the corresponding English LP. 3. Problems with phrasal verbs.

The following are some examples from the corpus.[8]

37. 1809 lly the patients have to take (LP) medical excuse $sick days$
38. 4376 ) in $0$ (GWC) working $work$ (LP) half time $part time$. (L
39. 4583 (Z) don't $don't$ know how to (LP) carry out $carry on$ thei
40. 4832 e (GWC) addicted $addictive$. (LP) In another hand $On the other hand$
41. 5081 our kids (GWC) save $safe$ . (LP) Today the $Nowadays$ kids
42. 5661 that all (FS) People $people$ (LP) to make a fraud $who commit fraud$

Example 37 and 38 refer to errors in (semi-) fixed multi-word expressions. Example 39 refers to an error in a phrasal verb. Examples 40 to 42 refer to a paraphrase instead of using the corresponding English LP.

The third error with most incidence in this category is Lexical Single False Friends (LSF) with 208 and a percentage of 0.014% in the corpus. This error type refers to errors result from the influence of a formally similar word in the learner's modern tongue e.g. false friends. The following are some examples from the corpus.

43. 1015 hat$ (GVN) need $needs$ a big **(LSF)** inversion $investment$.

[8] http://grupotnt.udea.edu.co/django/clec/corpus/

44. 2331 lls (QC) . $;$ But (QM) 0 $,$ **(LSF)** lamentably $unfortunately$
45. 2854 $. (WRS) Consequently $0$ the **(LSF)** actual $current$ prison
46. 3958 nd other (LS) works $jobs$ to **(LSF)** maintain $provide for$ y
47. 4334 ne who knows about the topic, **(LSF)** localize $track down$ th
48. 7835 ty$ (QR) . $0$ (WRM) Is a big **(LSF)** responsability $responsibility$

All the previous examples show how the learners are influenced by their mother tongue when they use false friends or expressions that do not belong to the English language. LSF is an error worth to analyse for the relationship with the mother tongue.

LSF errors result from the influence of a similar word in the learner's mother tongue. In the present corpus, several students used words in their original language (Spanish) to complete a sentence in English and in some cases, they added graphic stress as is done in Spanish. Even though the manual from Louvain University says, those cases should be marked as FS, for the present research they were left as LFS because are all related to the influence of the mother tongue. The possible incidence of this type of errors is the lack of competence in the foreign language. Let us see some examples.

49. 6950 ge$ (QM) 0 $,$(GA) 0 $a$ life **(LSF)** decisión $decision$ I th
50. 4650 of$ perception and it is very **(LSF)** perjudicial $harmful$ fo
51. 2806 annual flooding. There are ' **(LSF)** Arroyos $flashfloods$' i
52. 6948 fe$. If you aren't mature and **(LSF)** responsable $responsible
53. 8947 $ (GWC) married $marriage$ is **(LSF)** esencial $essential$ for

All examples from 49 to 52 show the use of Spanish words instead of the English ones. Learners go to their mother tongue to borrow the vocabulary they do not have yet. They consistently show a lack of lexical competence.

**Summary of Lexis Errors (LS) category**

In this category the three most relevant errors are analysed accounting also for their incidence in the whole corpus. The list with percentages and means in the whole corpus was presented in Table 80, page 203.

LS category, with an incidence of 0.085% in the corpus, is the most relevant error in the present category. LP and LSF are less relevant for the number in the corpus, but LSF was analysed for showing the mother tongue as a possible source of errors.

### 4.4.3. Analysis of Word errors (W)

The third category with most incidence within the corpus was W (Word). This category is divided into three subcategories: WR, Word Redundant, WM, Word Missing, WO, and Word Order. WR is also divided into two sub-categories: WRS: Word Redundant Single, if one word is redundant, and WRM: Word Redundant Multiple, if several words are redundant. WM involves the omission of words. WO involves problems of word order except for categories of adverb order and adjective order.

Word errors accounted for 1945 errors that make 13% of the corpus. Errors in this category were dispersed in 4 error types.

Table 81

*Word errors with percentages and means*

| Error type | number of errors | percentage in corpus | Mean in corpus *M* |
|---|---|---|---|

| | | | |
|---|---|---|---|
| WRS | 670 | 0.045 | *1.3* |
| WM | 602 | 0.041 | *1.16* |
| WRM | 348 | 0.023 | *0.67* |
| WO | 325 | 0.022 | *0.63* |
| Total | 1945 | 0.131 | *3.78* |

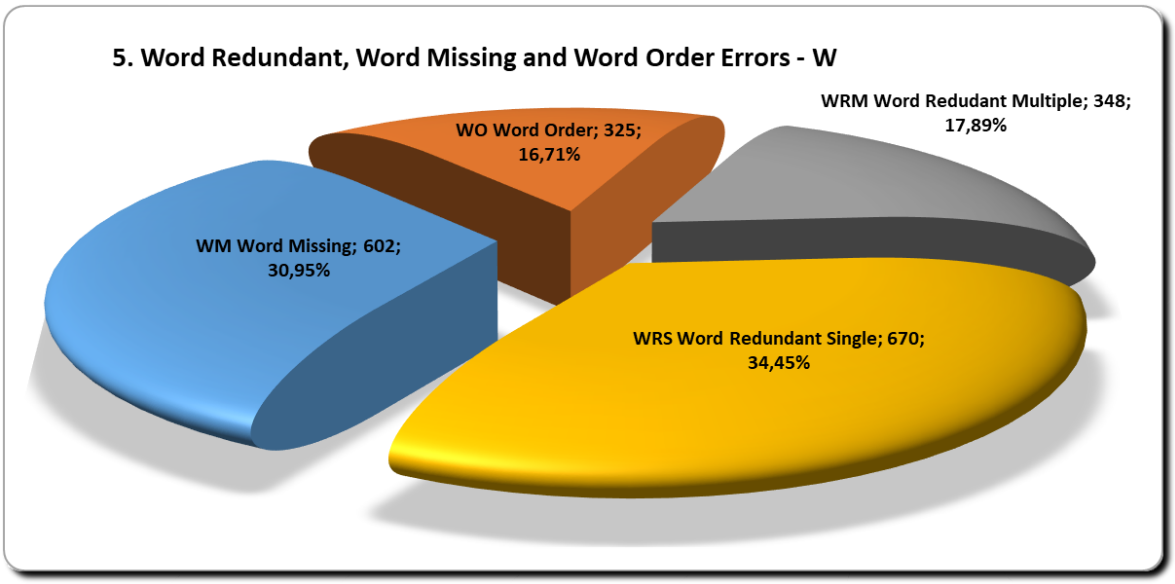



Figure 65. Account of Word errors.

The analysis from the present category will include its 4 types of errors, since they account for an important part of errors in the corpus. The error with most incidence in this category is WRS (Word Redundant Single) with 670 errors and a percentage of 0.045% in the whole corpus. This error type accounts for unnecessary repetitions of single words. The following are some examples from the corpus.

54. 818 n drugs and nutrition, can be **(WRS)** feel $0$ fatigue and (LP
55. 922 ugh chemotherapy is also hard **(WRS)** too $0$, because people
56. 1390 r (WRS) and $0$ to extend and **(WRS**) to $0$ have a better life

57. 1927 the risks of fixing o not **(WRS)** fixing $0$ Barranquilla's flashfloods
58. 2155 k$. Some treatments that help **(WRS)** to$0$ people to reduce t
59. 2296 In fact, one of the teachers **(WRS)** he $0$ asked about what

As it can be seen, examples 54-59 refer to the same situation of words that are redundant and the correction is $0$.

The second error with most influence in this category is WM (Word Missing) with 602 errors and a percentage of 0.041% in the whole corpus. This error type accounts for the omission of words except when they are pronouns, dependent preposition, articles, auxiliaries and connectors. The following are some examples from the corpus.

60. 1293 This problem since many years **(WM)** 0 $has$ affected things
61. 1459 if you go to jail you do not **(WM)**0 $know$ (WO) who is your
62. 1622 lties and the noise pollution **(WM)**0 $caused$ by construction
63. 1728 cognize$ that chemotherapy is **(WM)** 0 $used$ especially for
64. 2441 nts who (GVAUX) does$ do$ not **(WM)**0$ need$ special attention
65. 4768 eaten$ everyday and they must **(WM**) $be$ careful with the information

All examples from 60 to 65 refer to missing words in sentences that affect the meaning or leave the sentence incomplete.

The third error with most influence in this category is WRM (Word Redundant Multiple) with 348 errors and a percentage of 0.023% in the whole corpus. This error type accounts for unnecessary repetitions of multiple words. The following are some examples from the corpus.

66. 661 also has the advantage of **(WRM)** that can $0$ treating al
67. 920 are a lot of types of cancer **(WRM)** that exists $0$, and the
68. 1611 ideas, we need measures **(WRM)** that make possible $0$ t
69. 1790 and an appropriate curricula **(WRM)** in which is $0$ designed to
70. 1892 n (QM)0 $,$ cancer treatments **(WRM)** in the body $0$ generate
71. 2245 (QM) 0 $,$ you cannot prevent **(WRM)** you of $0$ future cancer

All examples from 66 to 71 show the use of unnecessary groups of words.

The final error with least incidence in the category of Word is WR (Word Redundant) with 325 and a percentage of 0.022% in the whole corpus. This error type accounts for problems of word order that do not fall into adverbs or adjective order. The following are some examples from the corpus.

72. 2599 different learning style$ but **(WO)** this not is $this is not$
73. 2791 rrect this major problem **(WO)** that has Barranquilla $that Barranquilla has$
74. 2793 all that goes directly to the **(WO)** river 'Magdalena'.$Magdalena river$
75. 3000 urrent mobility problems. Why **(WO)** we should $should we$ (LS
76. 3116 e any kind of discrimination. **(WO)** To not $Not to$ separate
77. 3236 because these students **(WO)** to want learn fast $want to learn$

All examples from 72 to 77 clearly show problems in word order.

**Summary of Word (W) category**

In this category, four types of errors that account for 13% from the whole corpus have been analysed. The error with most incidence in this category was WRS (Word Redundant Single). Errors in this category are probably the result of typing mistakes or in the case of WM (Word Missing) could be ignorance of a missing word.

### 4.4.4. Analysis of Form errors (F)

The fourth category with most incidence within the corpus was Form (F). This category refers to errors that affect the form of words. It is divided into 3 subcategories: FM (Form,

Morphology) that refers to errors in inflections or derivate words. FS (Form, Spelling) that refers to errors in spelling, omission of capital letters, borrowings, homophones, among others. The last sub-category is FSR (Form, Regional Spelling) that refers to non-consistent spelling being American or British within a text.

Form errors accounted for 1935 hits that make 13.22% of the whole corpus. Errors in this category were dispersed into three error types: Form, Spelling (FS), Form, Morphology (FM) and Form, Spelling, Regional (FSR).

Table 82

*Form errors with percentages and means*

| Error type | Number of errors | Percentage in corpus % | Mean in corpus *M* |
|---|---|---|---|
| FS | 1907 | 0.13 | *3.70* |
| FM | 23 | 0.0015 | *0.04* |
| FSR | 5 | 0.0003 | *0.009* |

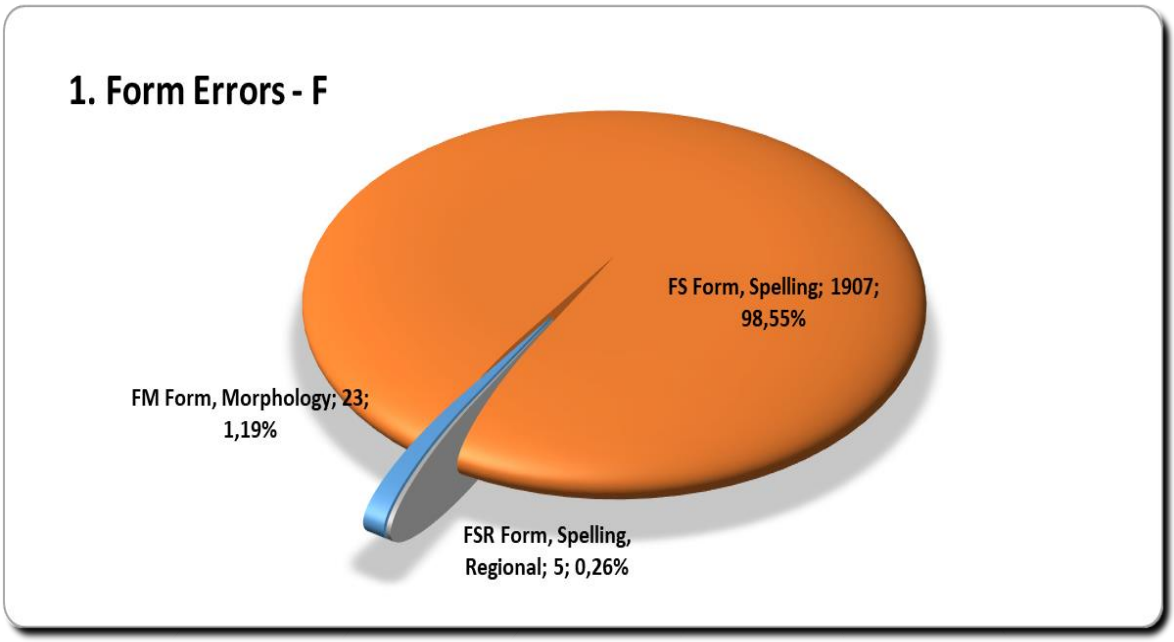



Figure 66. Account of Form errors.

The analysis from the present category will be based only on FS that accounts for 1907 errors and make 13.03% of the total errors in the corpus. FS accounts for the highest number of errors in the whole corpus. Even though this category is on the fourth place of incidence from the total, this type of error is on the first place of incidence in the whole corpus. The following are some examples from the corpus.

| | |
|---|---|
| 78. 1423 | lement a new security system. **(FS)** Thrid $third$, (FS)cr'ate |
| 79. 1505 | to be a place in isolation$. **(FS)** Its $ It's because the pr |
| 80. 1547 | k (GPP) they $their$ kids are **(FS)** saver $safer$ in separate |
| 81. 2217 | eating $treating$ cancer with **(FS)** Diets $diets$ or an alter |
| 82. 2507 | mplemented all over the city. **(FS)** Now a day $Nowadays$ (SU) |

Error 78 and 80 are spelling errors. Error 79 is a homophone. Error 81 is the misuse of a capital letter. Error 82 is a misspelling of a long word.

**Summary of Form (F) category**

In this category, FS errors were analysed. Since Form Morphology and Form Spelling Regional do not have a high incidence in the corpus. FS errors account for 13.03% of the total of errors in the corpus and is the type of errors with most incidence in the corpus. This type of error appears in all levels of learning from the corpus.

### 4.4.5. Analysis of Punctuation errors (Q)

The fifth category with most incidence within the corpus was Punctuation (Q). This category refers to four types of errors: QM (missing punctuation), QR (redundant punctuation), QC (confusion of punctuation), and QL (punctuation mark instead of a lexical item).

Punctuation category accounted for 960 errors that make 6.56% of the whole corpus. Errors in this category were dispersed into four error types: Punctuation Confusion (QC), Punctuation Lexical (QL), Punctuation Missing (QM) and Punctuation Redundant (QR).

Table 83

*Punctuation errors with percentages and means*

| Error type | Number of errors | Percentage in corpus | Mean in corpus *M* |
|---|---|---|---|
| QM | 621 | 0.042 | *1.2* |
| QC | 229 | 0.015 | *0.44* |
| QR | 93 | 0.006 | *0.18* |
| QL | 17 | 0.0011 | *0.03* |

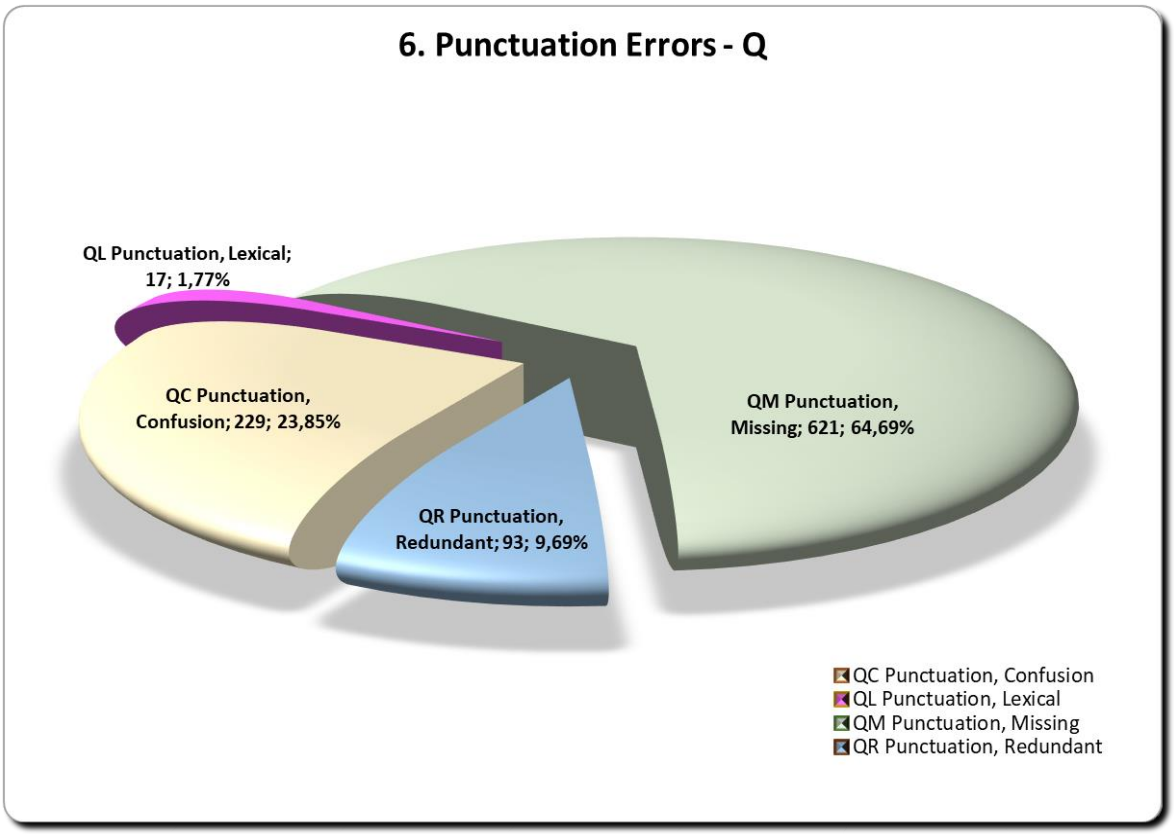



Figure 67. Account of Punctuation errors.

The analysis from the present category will be based on the error with most incidence within this category: QM with 621 errors that make 0.042% of the whole corpus. The following are some examples from the corpus.

| | |
|---|---|
| 83. 178 | do anything. For that reason (QM)0$,$ people (LP) is disgra |
| 84. 400 | oth genders (QC) ,$;$ however (QM) 0 $,$ (GPU) 0 $they$ acqu |
| 85. 1298 | vic culture$?$. In conclusion (QM)0 $,$ (GPP)0 $it$ is impor |
| 86. 4920 | 'safety to control your risk' (QM) On the other hand$,$ (GPP |
| 87. 5278 | of fraud be punished severely (QM) 0 $?$ (GA) The $0$ con me |
| 88. 6085 | munion, it's part of our life (QM)0 $.$ (SU) and if we don't |

Examples 83 84, 85, and 86 refer to missing commas. In the case of example 87, there is a missing question mark. Example 88 refers to a missing period.

**Summary of punctuation category**

In this category, QM errors were analysed. This error type has most incidence within the category. Nevertheless, its incidence in the corpus is minimum. Other errors in this category hardly account for 0.015% in the whole corpus.

### 4.4.6. Analysis of Style errors (S)

The sixth category with most incidence within the corpus was Style (S). This category refers to two cases: incomplete sentences (Sentence, Incomplete. SI) and unclear sentences (Sentence, Unclear. SU). SI includes sentence fragments, for example without a verb. SU is used when there is an unclear phrase or sentence.

Style errors accounted for 520 errors that make 0.03% of the whole corpus. Errors in this category were dispersed into two error types: Sentence Unclear (SU) and Sentence Incomplete (SU).

Table 84

*Style errors with percentages and means*

| Error type | Number of errors | Percentage in corpus | Mean in corpus *M* |
|---|---|---|---|
| SU | 501 | 0.034 | *0.97* |
| SI | 519 | 0.001 | *0.36* |

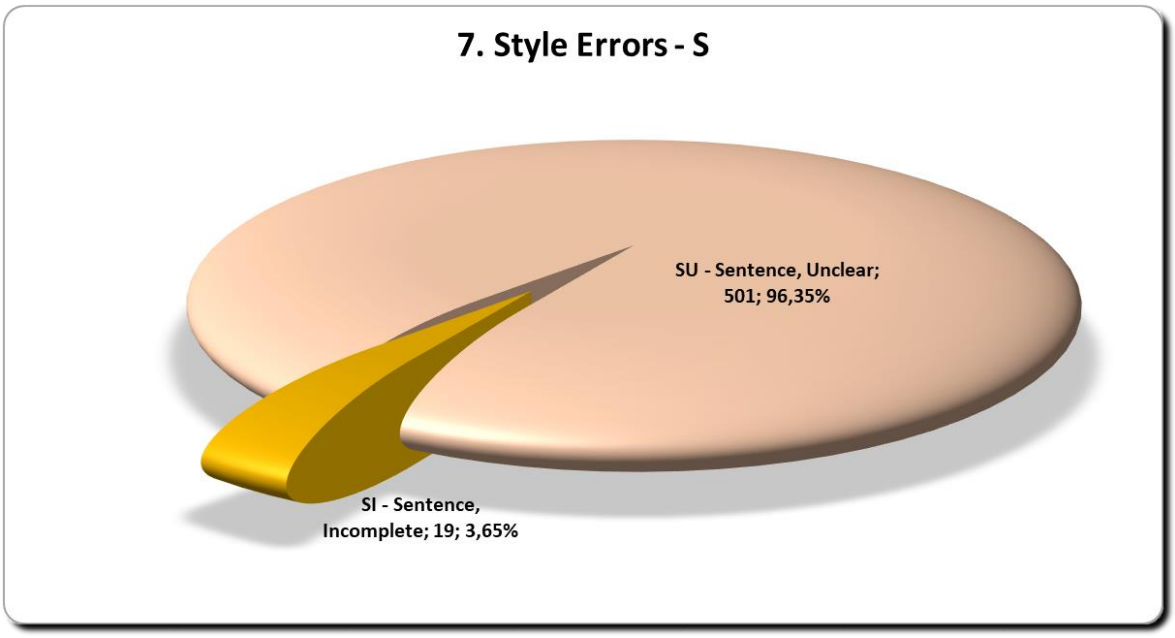



Figure 68. Account of Style errors.

The analysis from the present category will be based on the error with most incidence within this category: SU that accounted for 501 errors and make 0.034% from the corpus. The following are some examples from the corpus.

**89. 6176 (SU)** Bu the best idea, Its only a dream, Its a tradition say for our grandmothers.$?$
**90. 6179 (SU)** Now, I a woman and man lo love each one, they can live together, prove how is their feeling and after decide if they marriage or not$?$.
**91. 6086 (SU)** and if we don't wan't to marry and lose this tradition think about if you want to lose the others, I think no$?$.
**92. 6033 (SU)** The ideal situation would be where the move is by mutual Agreement, with the parents and the young person feeling that the time is right$?$.

Examples 89 to 92 show sentences that are unclear and do not make sense.

**Summary of Style category**

In this category, SU errors were analysed since it is the error type with 96% of the errors in the Style category. The incidence of this category is only 0.03%, for that reason no more analysis was done.

### 4.4.7. Analysis of Lexico-Grammar errors (X)

The seventh category with most incidence within the corpus was Lexico-Grammar (X). This category refers to errors "where the morpho-syntactic properties of a word have been violated." (E. Dagneaux et al., 2005, p.27). This category is divided into 3 subcategories: Complementation of adjectives, conjunctions etc. Dependent prepositions of adjectives, verbs etc. and Nouns uncountable/countable. This category is divided into 9 error types.

Lexico-Grammar errors accounted for 279 errors that make 0.019% of the corpus. Errors in this category were dispersed into 7 error types.

Table 85

*Lexico-Grammar errors with percentages and means*

| Error type | Number of errors | Percentage in corpus | Mean in corpus *M* |
|---|---|---|---|
| XVPR | 127 | 0.008 | *0.024* |
| XPRCO | 92 | 0.006 | *0.17* |
| XNUC | 43 | 0.0029 | *0.08* |
| XADJPR | 7 | 0.00042 | *0.013* |
| XVCO | 6 | 0.00041 | *0.011* |
| XNPR | 3 | 0.0002 | *0.005* |
| XNCO | 1 | 0.00006 | *0.001* |

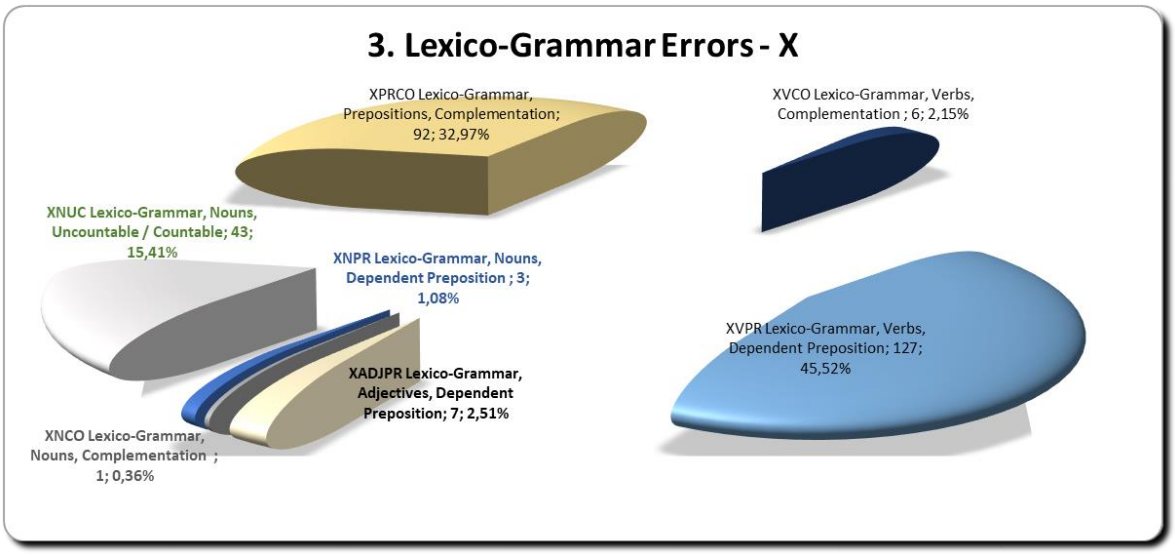



Figure 69. Account of Lexico-Grammar errors.

The analysis from the present category will be based on the first two errors that accounted for more than 70% within this category: XVPR and XPRCO.

XVPR refers to the use of verbs with the wrong dependent preposition. This error accounted for 127 hits in total. The following are some examples from the corpus.

93. 5574 $ people (FS) depent $depend$ (**XVPR)**depend of $depend on$ th
93. 6576 t$ (FS)peoples $people$ don't **(XVPR)** think in $think about$
94. 2341 e deteriorated. Diet consists **(XVPR)** consists to $consists of$
95. 4643 at more when your mind is not (**XVPR)** concerned in $concerned about$
96. 13053 py, (FS) acording $according$ **(XVPR)** according with $according to$
97. 6394 0$ the real life where no one **(XVPR)** depends by $depends on$

All examples from 93 to 97 refer to the misuse of verb-preposition.

XPRCO refers to the erroneous complementation of prepositions. This error accounted for 92 hits in total in level B2. The following are some examples from the corpus.

98. 5091 rcials have false advertising (XPRCO) for sell $ for selling$
99. 5475 e information and description (XPRCO) before buy $before buying$
100. 5791 ng adult (GVN) think $thinks$ (XPRCO) about leave $about leaving$
101. 5949 r killing$ (GA) a $0$ people, (XPRCO)for let $for letting$ a
102. 5950 killer can pay for his crime, (XPRCO) for kill $for killing$
103. 6008 $ in Colombia have kids (XPRCO) without marriage $without marrying$

All examples from 98 to 103 correspond to the wrong preposition complementation.

**Summary of Lexico-Grammar category**

In this category, XVPR and XPRCO were analysed since both errors add up more than 70% of the total in this category. The incidence from these errors in the corpus is minimum, therefore this category does not have much incidence in the corpus.

### 4.4.8. Analysis of Infelicities (Z)

The eighth category with most incidence in the corpus was Infelicities (Z). This category does not refer to real errors but it refers to register problems, questions of political correctness and stylistic problems. The only tag for this category is Z.

Infelicities accounted for 38 errors in the corpus that make 0.0025 from the corpus.

Table 86

*Infelicities with percentages and means*

| Error type | Number of errors | Percentage in corpus % | Mean in corpus *M* |
|---|---|---|---|
| Z | 38 | 0.00025 | *0.07* |

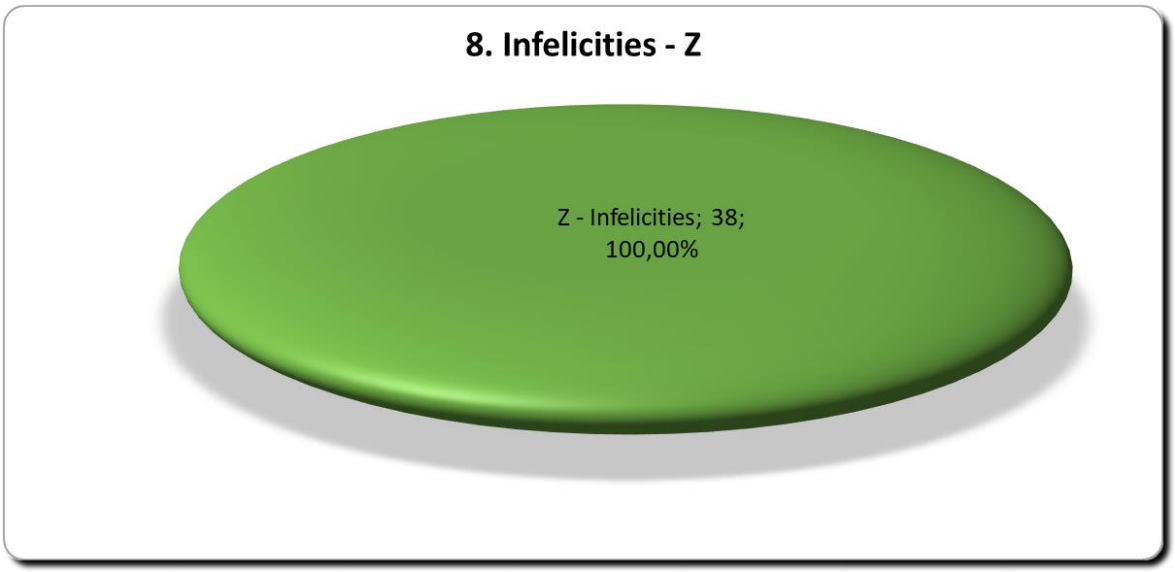



Figure 70. Account of Infelicities.

The following are some examples from the corpus.

104. 3009 $floods$ Barranquilla is the **(Z)** 4th $fourth$ business city
105. 3344 loods$ some people lose their **(Z)** stuff $belongings$, (GNN)
106. 4605 althy$?$. Following this line **(Z)** you`ll $you will$ gain wei
107. 7190 $ (FS) goverment $government$ **(Z)** can not $cannot$ establish
108. 11583 anies around the world really **(Z)** don't $do not$ (XVPR) care

All examples from 104 to 108 refer to the use of English in an oral context but are not really erroneous output from students.

**Summary of Infelicities category**

In this category, the incidences were analysed in the only tag it has. It was established that this category refers not to real errors, but to problems in register, style or matters of political correctness.

It can be seen that the first seven errors with the most incidence in the corpus and that make 53% of the total in the corpus are: FS (Form Spelling), GA ( Grammar, Articles), LS (Lexical Single), LP (Lexical Phrase), GNN (Grammar, Nouns, Number), GWC (Grammar, Word Class), and WRS (Word Redundant Single).

**Recapitulation Chapter 4**

This chapter presented the verification or proof of the established hypotheses. It presented a description of the main findings making in some cases an analysis related to the sociocultural variables. This chapter described the research outcomes according to the research objectives with an exaustive analysis of errors by category giving examples from the corpus and analysing possible sources of errors.

# 5. CONCLUSIONS AND RECOMMENDATIONS

The main objective from the present research was to investigate the relationship between the main errors found in written compositions from students at university level and the socio-demographic factors that could have incidence in the development of the written skills at *Universidad del Norte* in Barranquilla, Colombia. To accomplish this goal, there were two propositions tested:

a. Male and female students greatly differ in the type and quantity of written errors
b. There are statistically significant differences in the median of errors from written production of EFL university students in relation to the stratum classification given in Colombia.

To achieve the main goal, an instrument was designed to search for the students' profile, the academic profile and the socio-demographic factors that could have incidence in the learning of a foreign language. Furthermore, this thesis accounted for the most important aspects of Error Analysis (EA) analysing its criticism and limitations. It presented the state of the art regarding Error Analysis and Computerised Learner Corpus (CLC) finding that most studies on error analysis using CLC have been done in Europe, and that few works have been done in Latin America and that in Colombia this kind of work is almost non-existent. It presented a framework for the diagnostic of errors, and an overview of corpus linguistics as a methodological approach to do scientific linguistic analysis. Learner Corpus (LC) features were explained head-to-head with the acquisition of a second or a foreign language.

In order to test the propositions regarding the relationship of gender and strata in the type and quantity of errors, it introduced an elaborated Error Analysis framework as the methodology

followed in order to do Error Analysis. It showed how the learner corpus was collected and how the data was analysed. It described the type of software used for the extraction of data and the annotation system employed. Furthermore, it described the interface design, which is an application that will be useful for future research. Additionally, it presented the results from the instrument designed to find out the sociolinguistic aspects that could affect acquisition in a foreign language.

Finally, this thesis accounted for the verification or proof of the established hypotheses. It presented a description of the main findings matching the analysis to the socio-demographic variables analysed in this thesis: gender and strata. It described the research outcomes according to the research objectives with an exhaustive analysis of errors by category giving examples from the corpus and analysing possible sources of errors.

### 5.1. Synthesis of results

1. **Regarding the main question from the present research:** The most prevalent errors from EFL written compositions at university level are: FS (Form Spelling), GA (Grammar, Articles), LS (Lexical Single), LP (Lexical Phrase), GNN (Grammar, Nouns, Number), GWC (Grammar, Word Class), and WRS (Word Redundant Single). These error types showed to be the most recurrent in the different instances in which they were examined, not only at category level, but also at general level.

3. **Validation or proof of the hypotheses:** This thesis focused on two situations: **Situation 1:** H0: Male and female, university students, do not present statistically significant differences in the median of errors from written production of English as a Foreign Language (EFL). Some tests were necessary to determine if females make less errors than males in the production of EFL written texts at university level as differences in EFL learning from male and female have been previously established by the literature. To do this test, the present research was constituted by a dichotomous variable (gender, 1-> male; 2->female) and one quantitative variable: written errors. **Situation 2**: H0: There are no statistically significant differences in the median of errors from written production of EFL university students in relation to the strata classification given in Colombia. Some tests were necessary to determine if access to private schools or involvement in sociocultural factors such as travel abroad were related to the written production of English as a foreign language. Involvement in such leisure events is determined by the variable of socio-economic stratum. To do this test, the present research was constituted by a qualitative variable, a component of six levels (strata from 1 to 6) and a quantitative variable that is written errors. It was necessary to analyse if the quantitative variable of errors followed a normal probability distribution to determine what type of test should be used. Two normality tests: Kolmogorov-Smirnonov and Shapiro-Wilk were performed. Since the significance was less than 0.05 (p-value < 0.05), in both normality tests (Kolmogorov-Smirnov and Shapiro-Wilks), it was determined that the variable errors did not follow a pattern of normal distribution. For the previous reason, it was necessary to use a non-parametric test to accept or reject the null hypothesis in situations 1 and 2.

Nevertheless, taking into account that parametric tests are more reliable than the non-parametric tests, it was decided to do a logarithmic transformation of the variable errors. After performing the logarithmic transformation, it was not possible to achieve normality. Given that the stated situations refer to independent samples, the U de Mann Whitney test was used for situation 1 and the K Kruskal-Wallis test for the situation 2. **In situation 1**, since the significance was 0.541 (p-value > 0.05) H0 was accepted. For the previous reason, it was stated that written errors from male and female EFL university students do not present statistically meaningful differences. **In situation 2**, it was mentioned that 191 out of 515 students reported their strata to match their written work with the results from the survey on socio-demographic aspects. For the previous reason, those assignments without strata were given a number zero for stratum. This situation hinders the analysis process and, in some ways, it was established that it might contribute to the misinterpretation of the data and the testing of this hypothesis, not for the test itself, but because stratum zero does not exist, and therefore it does not correspond to the reality. For the previous reason, the same test was performed in the sample that confirmed their strata (191 in total). After applying the Kolmogorov-Smirnov normality test it was established that the variable errors did not achieve normality in this sample; therefore, the non-parametric K-Kruskal-Wallis test for polytomous variables was performed in the sample that confirmed their strata. The significance was Sig =.092 therefore, the H0 was accepted. It was established, according to the results, that there are no statistically meaningful differences in the median of errors according to strata. After analysing the previous facts, the researcher found out that students who established their strata were distributed into 86 from level B2 (77%) and 105 from level B1 (26%). The preceding is reason enough to be convincing because 77% of the population from level B2 is sufficient to draw conclusions. Consequently, the researcher only analysed the percentage of students who

stated their strata and belonged to level B2. The previous fact allowed the researcher to establish that the results from B2 are totally consistent and conclusive. It can be said that in level B2 there are some differences in the median of errors from written production of EFL university students in relation to the stratum classification given in Colombia. The higher the stratum, the higher the incidence of errors. The findings contradict some theories regarding the claims from some authors about the privileges of some social classes that have access to travel abroad to improve their level of the language.

3. **The incidence of errors in each category**: From the eight error categories, some error types had a recurrent tendency that was reflected not only in the category, but also in the general results. Errors with the most predominance in the corpus were, in some cases, leading the account from their own category. The category with most incidences was the grammatical (G) category, in part because it has 25 error types from the 56 found in the error tagger. From this category, only three error types were in the list of main errors from the general corpus. The second category with most incidence was F (Form). The reason was the leader error from the entire corpus: FS. This error overpassed all errors from any category and was steady along level B1. The third category with most incidence was (L) Lexical. This category with the error LS becomes the most important in level B2. The category with the least incidence was infelicities (Z).

4. **The incidence of errors according to gender**: In the general results, the first three main errors for males and females were Form Spelling (FS), Grammar Article (GA) and Lexical

Single (LS). Females have more written production than males and therefore more errors in general, but when obtaining the means of errors per student, males have more incidence of errors. The most prevalent error in all level B1 was Form Spelling, but in level B2, it changes for Lexical Single. It can be said from levels B1 and B2, that the results from this research matched the ones from previous research that stated that females are better language learners.

5. **The incidence of errors according to the CEFR**: The CEFR has a high impact on language testing, but it is difficult to establish a way to follow up learners' writing progress because there is no real clarity about possible written errors in each level. It could be said that there are some error types that are steady from B1 to B2. They have a "growing cycle", they decrease, and after, they appear again with more strength. The case of LS is an example of an error that was in all the interlanguage from level B1, and kept "alive" with different scores until becomes the first error in level B2. Another similar case is error GA that was on the second place in all levels from B1 to B2. Comparing the typical errors that according to English Profile (UCLES/CUP, 2011) improve significantly from B1 to B2 the errors from the present corpus do not match the ones from English Profile.

6. **The incidence of errors by strata**: Since only 26% of students from level B1 confirmed their social strata the results from that level were presented without drawing any conclusions. From level B2, 77% of students confirmed their strata, therefore the analysis was based on that subgroup. Errors in B2 increased along with the strata, especially in strata 3 to 5. The five most frequent errors in all strata were Lexical Single, Grammar Article, Word Redundant Single, Lexical Phrase and Grammar Word Class. The average of written production increased in strata

3 and 6 for females and in strata 4 and 5 from males. In some strata males are more productive than females, but the mean of errors is higher in higher strata for males and females. These results might show how students who come from lower strata schools, with a low English level, study more to keep up with their peers and achieve the EFL required proficiency. Another possible cause is that students from higher strata who have travelled abroad believe they have a better level of English and are overconfident; therefore, they rely on informal English empirically learned. This findings contradict the theories from Lin (1999) about the advantages of some privileged social classes for having access to travel abroad and improve their level of the language. On the contrary, the most successful students were in the lowest strata.

7. **Similar work in the same topic:** Comparing the results from the present thesis with previous work by MacDonald (2017) explained in section 2.2, it can be established that similar results were found regarding the incidence of grammar error as the most frequent type in a similar corpus. Another research that has comparable results is from Diez-Bedmar (2011) who did error analysis using the error tagger manual 1.1 from Louvain University, (1996). In any of those cases, the comparison of results could not be conclusive for the difference in the error taggers.

8. **The incidence of Error Analysis in the teaching of languages**: It can be said that the results from Error Analysis are a big source of data that can be used to do better teaching practices and more accurate materials according to the student's needs. Error Analysis using a Computerised Learner Corpus shows the current state of the interlanguage and its evolution in order to correct the course of unfocused teaching practices or materials and it is also a source of empirical data.

## 5.2. Scope and limitations

The significance of this research can be established in the gap left for many years in the literature concerning a scientific analysis of the EFL learners' output, as a source that shows the evolution of learners' interlanguage, in Latin America, and more specifically in Colombia, where this type of work is almost inexistent. Even though the present work has been put into parallel with work from authors that analysed Spanish-speakers learning English, including authors from Latin America, the results from the present thesis only have internal validity because there is not a comparable work yet. It is necessary to do more research using the same error tagger and similar conditions with population from other universities in order to validate the results. Therefore, the results from the present thesis could in the future, be comparable to the results of research that use the same type of error tagger with a similar population of Spanish-speaker learners.

As mentioned before, not all students answered all variables, nevertheless, the statistic tests were done to stabilise the sample and the results show some important tendencies regarding gender and strata.

As explained in section 3.3, one of the results from this research was the design of an application. CLEC, The Colombian-Learner English Corpus, the annotated error corpus of all learners was processed and put online under the IMS Corpus Workbench technology, developed by the *Institut für Maschinelle Sprachverarbeitung* at the University of Stuttgart. CLEC can be found at the following **URL <http://grupotnt.udea.edu.co/django/clec/>** as well as help on how to use it for teaching or research purposes. CLEC is fully available to the academic

community under registration; guest users may have limited access to corpus results. This application can be used by teachers and researchers and is worth as a source of material to improve materials design or to have data from familiar context. This kind of data is also useful to find out the level of interlanguage from students and what to know what to expect from them.

This research could be a starting point for new research on Error Analysis using a computerised learner corpus, especially in Latin America and more specifically in Colombia. The findings are comparable with future research using the same error tagger.

### 5.3. Problems found in the process

Logistic inconveniences: Even though all teachers were informed beforehand about the research project, some of them never informed the students and that could presumably be the reason for having a low amount of surveys filled out. In some other cases, the researcher had to contact some of the teachers to obtain the students' work and that resulted in a waste of time because in some cases the work was never sent.

The error tagger: The error tagger has a great description of errors, but there should be more clear directions about its use or a document or page with FAQ in case the user needs it and because waiting for a response from the provider takes too long, especially for the difference in the standard time. The researcher in many cases ended up learning by try and error, and that resulted in a waste of time. Designers of the software assume all researchers know how to use it, but that is not always the case. For example, when there is a long name of a document included in another document, the tagger will not open it. You must create a short way to access it through the tagger. On the other hand, spelling errors should be automatically tagged. As a final reasoning, the researcher

insists on the need to have a collective judgement for an error tagger that benefits the whole research community, which is available for research purposes and not for selling.

Error classification: according to the researchers' experience with the error tagger used in this research, error tagging should be done by categories one by one to focus just in one category and not get dispersed by an overwhelming quantity of errors in several categories. Nevertheless, there are some error tags that are difficult to decide in which type they match better, for that reason I know in several cases some errors could have been classified differently, but it was a matter of consistency.

### 6.4. Recommendations

Since there are some errors that keep steady in all interlanguage stages, it would be advisable to use additional activities and materials that make students aware of those kinds of errors and how to avoid them. More errors in quantity and type could be avoided if students worked exercises more adapted to the needs of Spanish speakers, members of the Colombian culture.

The errors found in this research are a screen shot from the learners' interlanguage in *Universidad del Norte* that can serve as a reference and can be compared to similar research in other universities at national and international levels.

Errors show learners' level of interlanguage and are unavoidable in the learning process, nevertheless, the general mean of written errors per student could be lowered adapting materials to the real needs of students.

There should be more similar work at large scale in order to compare results in public and private universities.

It is advisable to do research using intervention models in order to know if Error Analysis combined with an intervention model improves the results of written production.

In part, the results presented in this research are because of the type of methodology used to do the analysis. Nevertheless, the arguments of authors that did not trust Error Analysis and Corpus Linguistics as methodologies to do scientific-linguistic research are not valid anymore since the use of new technologies allow reliable scientific-proven results.

The present thesis is the result of an association of Error Analysis and Corpus Linguistics, two methodologies, one for the analysis of errors and the other for the scientific process of corpus compilation that are blended to produce a reliable scientific work. This thesis is a pioneering work in its magnitude in Latin America, but specially in Colombia where few works have been done and with few participants.

Having analysed the main written errors from EFL students at university level and the incidence of strata and gender in the production of errors, I consider that the main scope from this research has been accomplished. Having valid results supported by the data, the present thesis is consistent in the classification of errors, in the theories by which it is supported and in the methodology used to obtain the findings.

## REFERENCES

Abercrombie, D. (1965). *Studies in phonetics and linguistics*. Oxford University Press.

Aliakbari, M., & Mahjub, E. (2010). Analytical/intuitive EFL learners and gender effect. *International Journal of Pedagogies & Learning*, *6*(1), 41–48. http://search.proquest.com.ezp.lib.unimelb.edu.au/docview/791529295?accountid=12372%5Cnhttp://sfx.unimelb.hosted.exlibrisgroup.com/sfxlcl41?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:journal&genre=article&sid=ProQ:ProQ%3Aeducation&atitle=Analyti

Aries, E. (1976). Interaction patterns and themes of male, female and mixed groups. *Small Group Behavior*, *7*(1), 7–18.

Arikan, A. (2010). Prospective English language teachers' perceptions of the target language and culture in relation to their socioeconomic status. *English Language Teaching*, *4*(3), 232–242.

Babayiğit, S. (2015). The dimensions of written expression: Language group and gender differences. *Learning and Instruction*, *35*, 33–41. https://doi.org/10.1016/j.learninstruc.2014.08.006

Bell, K. (2011). *How ESL and EFL classrooms differ*. Oxford University Press. https://oupeltglobalblog.com/2011/07/12/how-esl-and-efl-classrooms-differ/

Bell, R. (1974). Error analysis: A recent pseudo-procedure in Applied Linguistics. *ITL Review of Applied Linguistics*, *25*(6), 35–39.

Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing*, *8*(4), 243–257. https://www.mendeley.com/catalogue/representativeness-corpus-design/

Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on

collocation networks. *International Journal of Corpus Linguistics*, *20*(2), 139–173. https://doi.org/10.1075/ijcl.20.2.01bre

Burt, M. (1975). *Error Analysis in the adult EFL Classroom*. TESOL Quarterly. https://www.jstor.org/stable/3586012?seq=1#page_scan_tab_contents

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge M.I.T. Press.

Collins, J. (2006). Where is class in second language learning? *Working Papers. Urban and Language Literacies.*, *41*, 1–9. https://www.academia.edu/6464773/WP41_Collins_2006._Wheres_class_in_second_language_learning

Corder, S. (1967). The significance of learner's errors. *IRAL - International Review of Applied Linguistics in Language Teaching*, *5*(1–4), 161–170. https://doi.org/10.1515/iral.1967.5.1-4.161

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

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR).* Cambridge University Press.

Crystal, D. (1997). *English as a global language, Second edition*. Cambridge University Press.

Crystal, D. (2008). *Dictionary of Linguistics and Phonetics* (6th.). Blackwell Publishing.

Dagneaux, E., Denness, S., Granger, S., Meunier, F., Neff, J., & Thewissen, J. (2005). *Error Tagging Manual Version 1.2*. Centre for English Corpus Linguistics, Université Catholique de Louvain.

Dagneaux, Estelle, Denness, S., & Granger, S. (1998). Computer-aided error analysis. *System*, *26*(2), 163–174. https://doi.org/10.1016/S0346-251X(98)00001-3

DANE. (n.d.). *Estratificación socioeconómica*. Socio-Economic Estratification for Household

Public Services. Retrieved October 24, 2018, from https://www.dane.gov.co/index.php/servicios-al-ciudadano/servicios-de-informacion/estratificacion-socioeconomica

Díez-Bedmar, M. B. (2011). Spanish pre-university students' use of English: CEA results from the University Entrance Exam. *International Journal of English Studies*, *11*(2), 141–158. http://revistas.um.es/ijes/article/view/149681

Du, X. (2012). *A Brief Introduction of Skopos Theory*. https://doi.org/10.4304/tpls.2.10.2189-2193

Dulay, H; Burt, M; Krashen, S. (1982). *Language Two*. Oxford University Press.

Durán, N. (2011). Exploring gender differences in the EFL classroom. *Colombian Applied Linguistics Journal*, *01*(8), 123–136.

Eckert, P., & McConnell, S. (2003). *Language and gender*. Cambridge University Press.

Eckman, F. R. (1977). Some Theoretical a N D Pedagogical Implications of the Markedness Differential Hypothesis. *SSLA*, 289–307. https://doi.org/10.1017/S0272263100005544

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

Estelle, D., Sharon, D., Sylviane, G., Fanny, M., Joanne, N., & UCL Louvain. (2005). *Error Tagging Manual*. 1–46.

Europe, C. of. (2001). 1 The Common European Framework of Reference for Languages: Learning, teaching, assessment. *Common European Framework*. https://doi.org/10.1093/elt/cci105

Gilquin, G. (2015). From design to collection of learner corpora. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 9–34). Cambridge M.I.T. Press.

González, L. (2017). *Valanglia*. https://www.google.com.co/search?q=González,+L.+(2017).+Valanglia+cefr&tbm=isch&source=iu&ictx=1&fir=hLaTuG_g1N-tMM%253A%252CAUvl4AtiWBm20M%252C_&usg=AI4_-kQ3bcTUhprjiTGNERRCEgGPPnwpug&sa=X&ved=2ahUKEwiymPvDpOPeAhWFrVMKHdZfBzUQ9QEwAHoECAAQBA#imgrc=hLaTuG_

Granger, S, Gilquin, G, Meunier, F. (Ed.). (2015). *The Cambridge Handbook of Learner Corpus Research* (Sylviane G). Cambridge University Press.

Granger, S. (2006). *Center for English Corpus Linguistics CECL*. International Corpus of Learner English.

Henao, N., Ortega, M., & Tamayo, A. (2018). *CLEC Colombian Learner English Corpus*.

Ishikawa, Y. (2015). Gender Differences in Vocabulary Use in Essay Writting by University Students. *Procedia Social and Behavioral Sciences*, 593–600.

James, C. (1998). *Errors in Language Learning and Use. Exploring Error Analysis.* Rouledge.

James, Carl. (1998). *Errors in Language Learning and Use. Exploring Error Analysis.* (S. Candlin Christopher (Macquarie University (Ed.)). Rouledge.

Krashen, S. D. (1982). Principles and Practice in Second Language Acquisition. In *The Modern Language Journal* (Vol. 73, Issue 2). https://doi.org/10.2307/328293

Lakoff, R. (2003). Language, Gender, and Politics: Putting “Women” and “Power” in the Same Sentence. In J. Holmes & M. Meyerhoff (Eds.), *The Handbook of Language and Gender* (p. 722). Blackwell Publishing Ltd. https://doi.org/10.1002/9780470756942

Lee, J., Yeung, C. Y., Zeldes, A., Reznicek, M., Lüdeling, A., & Webster, J. (2015). CityU corpus of essay drafts of English language learners: a corpus of textual revision in second

language writing. *Language Resources and Evaluation*, *49*(3), 659–683. https://doi.org/10.1007/s10579-015-9301-z

Leech, G. N. (2005). *Developing linguistic corpora a guide to good practice* (M. Wyne (Ed.)). Oxford:Oxbow Books.

Lennon, P. (1991). Error: Some problems of definition, identification, and distinction. *Applied Linguistics*, *12*(2), 180–196. https://doi.org/10.1093/applin/12.2.180

Lin, A. (1999). Doing-English Lessons in the Reproduction or Transformation of Social Worlds? *TESOL Quaterly*, *33*, 393–412.

Ludeling, A., & Hirschmann, H. (2015). Error annotation systems. In *The Cambridge handbook of learner corpus research* (pp. 135–157). Cambridge University Press.

MacDonald, P. (2017). "We All Make Mistakes!". Analysing an Error-coded Corpus of Spanish University Students' Written English. *Complutense Journal of English Studies*, *24*(0), 103–129. https://doi.org/10.5209/CJES.53273

Mark, K., & Engels, F. (1998). *Communist manifesto*. Merlin Press.

McDowell, L. (2016). *An Error Analysis of Japanese scientists' research articles*. https://www.researchonline.mq.edu.au/vital/access/services/Download/mq:57812/SOURCE1?view=true

McEnery, A, Xiao, R., & Tono, Y. (2006). *Corpus-Based Language Studies: An advanced Resource Book.* Routledge.

McEnery, Anthony, & Hardie, A. (2011). *Corpus Linguistcis: Method, theory and practice*. Cambridge University Press.

McEnery, T., & Wilson. (2001). *Corpus Linguistics An Introduction* (Second). Edinburgh University Press.

Morales-Reyes, A., & Soler, I. G. (2016). Transfer and semantic universals in the L2 acquisition of the English article system by child L2 learners. *Language Acquisition*, *23*(1), 57–74. https://doi.org/10.1080/10489223.2015.1067318

Morales, S. (2017). Relationship between Social Context and L2 Learning of EFL Students in Terciary Level. *English Language Teaching*, *10*(10), 87–91. https://doi.org/http://doi.org/10.5539/elt.v10n10p87

Nash, R., Burt, M. K., & Kiparsky, C. (2006). The Gooficon: A Repair Manual for English. *TESOL Quarterly*. https://doi.org/10.2307/3585665

Nemser, W. (1969). Approximative Systems of Foreign Language Learners. *NA*, *NA*(NA), 13. https://files.eric.ed.gov/fulltext/ED026639.pdf

Nord, C. (2001). Dealing with Purposes in Intercultural Communication : Some Methodological Considerations. *Revista Alicantina de Estudios Ingleses*, *14*, 151–166.

Nord, C. (2012). *Texto Base-Texto Meta Un modelo funcional de análisis pretraslativo.* Universitat Jaume I.

Oxford University Press. (2017). *Online Placement Tests*. Oxford University Press. https://www.oxfordenglishtesting.com/DefaultMR.aspx?id=3034&menuId=1#

Parodi, G. (2008). Linguistica de corpus: una introduccion al ambito. *RLA, Revista de Linguistica Teorica y Aplicada*, *46*(1), 93–119. https://doi.org/10.4067/S0718-48832008000100006

Pavlov, I. (1927). *Conditioned Reflexes: an investigation of the physiological activity of the cerebral cortex.* Oxford University Press.

Psacharopoulos, G. (1987). Public Versus Private Schools in Developing Countries: Evidence From Colombia and Tanzania. *International Journal of Educational Development*, *7*(1), 59–

67.

Reib, K., & Vemeer, H. (2013). *Towards a General Theory of Translational Action: Skopos Theory Explained*. Routledge.

Ribas, R., & D'Aquino, A. (n.d.). *La corrección de errores como instrumento didáctico*. Actas Del Programa de Formación Para El Profesorado de Español Como Lengua Extarnjera.

Richards, J., & Schmidt, R. (1985). *Dictionary Of Language Teaching And Applied Linguistics*.

Romaine, S. (2003). Variation in Language and Gender. In J. Holmes & M. Meyerhoff (Eds.), *The Handbook of Language and Gender* (pp. 98–118). Blackwell Publishing Ltd.

Rosen, A., Jirka, H., Stindlová, B., Feldman, A., & Svatava, S. (2012). Evaluating and automating the annotation of a learner corpus. *Language Resources and Evaluation*, *48*(1), 65–92. https://doi.org/10.1007/s10579-013-9226-3

Rúa, P. L., & López Rúa, P. (2006). The sex variable in foreign language learning: an integrative approach. *Porta Linguarum*, *6*(1994), 99–114. http://dialnet.unirioja.es/servlet/articulo?codigo=2371605

Saeed, A., Ghani, M., & Ramzan, M. (2011). Gender Difference and L2 Writing. *International Research Journal of Arts and Humanities*, *39*(39), 29–40. http://search.proquest.com.ezaccess.library.uitm.edu.my/docview/1354331813?accountid=42518

Sánchez, A. (2013). *Bilingüismo en Colombia* *.

Sánchez, M. E., Sevilla, Y., & Bachrach, A. (2014). Agreement processing in control and raising structures. Evidence from sentence production in Spanish. *Lingua*, *177*, 60–77. https://doi.org/10.1016/j.lingua.2015.12.014

Scott, M. (2008). *WordSmith*.

Scott, Mike. (2005). *WordSmith*. http://lexically.net/wordsmith/research/

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching*, *10*(1–4), 209–232.

Sinclair, J. (2004). *Developing Linguistic Corpora: a Guide to Good Practice*. AHDS Literature Languages and Linguistics.

Skinner, B. F. (1953a). Operant behavior. *Science and Human Behavior*, 59–90. https://doi.org/10.3390/ijerph8093528

Skinner, B. F. (1953b). Operant behavior. *Science and Human Behavior*, 59–90. https://doi.org/10.3390/ijerph8093528

Spada, N., & Lightbown, P. (2013). *How languages are learned*. Oxford University Press.

Startvik, Jan; Enkvist, Nils Erik; Nickel, Gerald; Hammarberg, Bjorn; Corder, Pit; Johanson, Faith; Johansson, Stig; Rossipul, Hans; Lindell, Ebbe; Stendahl, Christina; Edstrom, Esmari; Hyldgaard-Jenssen, Karl; Moller, Elizabeth; Gorosch, Marx;, J. (1973). *Errata: Papers in Error Analysis* (J. (Lund university) Startvik (Ed.)). Errrata Papers.

Stenson, N. (1983). *Second Language Learning: Contrastive Analysis, Error Analysis, and Related Aspects.* (J. Robinett, Betty; Schachter (Ed.)). The University of Michigan Press.

Stromquist, N. (2004). Inequality as a way of life. Education and social class in Latin America. *Pedagogy, Culture and Society.*, *12*(1), 95–120.

*Survey Gizmo*. (n.d.).

Tognini-Bonelli, E. (2001). *Corpus linguistics at work. Studies in Corpus Linguistics*. John Benjamins Publishing Company.

Tono, Y. (2003). Learner corpora : design , development and applications. *Proceedings of the 2003 Corpus Linguistics Conference*, 800–809.

https://doi.org/http://dx.doi.org/10.1016/j.ijheatmasstransfer.2005.07.046

UC Louvain. (2018). *Centre for English Corpus Linguistics*. Learner Corpora Around the World. https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

UCLES/CUP. (2011). *English Profile: Introducing the CEFR for English* (Issue August).

van Rooy, B. (2015). Annotating learner corpora. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 79–105). Cambridge University Press.

Vandrick, S. (1995). Privileged ESL University Students. *TESOL Quaterly*, *29*(2), 375–381.

Vásquez, D. A. L. (2008). Error analysis in a written composition. *Profile*, *10*, 135–146. http://www.revistas.unal.edu.co/index.php/profile/article/view/10619/11079

Weinberger, U. (Lancaster U. (2002). *Error analysis with computer learner corpora: a corpus-based study of errors in the written German of British University students*. Lancaster.

# Annexes

## 1. Annex 1. Error tagging manual 1.2

## 2. Annex 2. Writing assessment directions Level B1.1

Universidad del Norte – Instituto de Idiomas Level Three: **Writing Assessment Instructions & Writing Prompts**

**What am I going to do?**
You are going to ***write an opinion paragraph***. Your paragraph should include the following (see rubric):

a) include at least one compound sentence (a sentence that joins two independent clauses using a coordinating conjunction like *and, or, but, so*)

| |
|---|
| ***Example:*** Uninorte is the best university in the Caribbean, but it is much more expensive than Atlántico. |

b) include at least one complex sentence (a sentence that joins a dependent clause with an independent clause using a subordinating conjunction like *because, if, after, whenever, although, as long as*)

| |
|---|
| ***Example:*** Because there are many new free trade agreements, Colombia's economy will continue to grow |

OR

| |
|---|
| ***Example:*** Colombia's economy will continue to grow because there are many new free trade agreements |

c) include at least one compound-complex sentence. You can combine a *simple sentence* with a *complex sentence* to create a *compound-complex sentence*

| |
|---|
| ***Example:*** As soon as the movie started, my phone rang, but I did not answer it. |

OR

| |
|---|
| ***Example:*** I did not answer my phone when it rang because the movie had started. |

**Step One:** Begin by choosing one of the following topics:

| Option One | Option Two | Option Three |
|---|---|---|

| | | |
|---|---|---|
| Are commercials on TV honest? Provide several good examples to support your answer. | Should con men and women who are caught in the act of fraud be punished severely? Explain. | In your opinion, do tourists pose a danger to the environment? Support your argument with several good examples. |

**Step Two: Brainstorming** (5 Minutes)

a) Take a few minutes to think about which option you want to write about.

b) Then, think about what you are going to write. Make a list of your ideas in the space provided.

**Step Three: Write an OUTLINE** (10 minutes)

a) Begin by thinking about a strong **opinion** statement. This is your TOPIC sentence. Write it down in the space provided for the topic sentence. It should be brief and to the point.

b) Now write at least 2 or 3 good reasons that answer why your opinion makes sense. Write your reasons in the space provided. Your reasons should be brief and to the point. Include examples or explanation for each reason.

c) Finally, write down a conclusion that restates your topic sentence in a different way. Make sure to use different vocabulary to express your conclusion.

**Step Four: Write a First Draft** (15 minutes)

a) Rewrite your OUTLINE as a paragraph on the space provided. Remember to **double-space** so your work is clear.

b) When you finish, read your paragraph and go through the **checklist** (5 minutes). Once your checklist is complete, it is time to write your final draft.

**Step Five: Rewrite Final Draft** (10 minutes)

a) Rewrite your paragraph and hand ALL your work in before you leave. **Do not forget to include your name.**

**You have 10 minutes extra. Use that time to review your work one last time and fix any mistakes you find.**

| CHECKLIST (5 minutes) | YES | NO |
|---|---|---|
| I completed the writing process (brainstorming, outlining and drafting). | | |
| I wrote between 8 and 14 sentences. A sentence begins with a capital and ends with a period. | | |
| I wrote a strong opinion-based topic sentence. | | |
| I wrote at least 3 reasons answering why I think this way. | | |
| I wrote at least 3 good examples or explanations to support my reasons. | | |

| | | |
|---|---|---|
| I included at least one compound sentence. | | |
| I included at least one complex sentence. | | |
| I included at least one compound-complex sentence. | | |
| I included at least 3 of the vocabulary words from the option I selected. | | |
| I reviewed my writing for spelling and grammar errors and corrected any mistakes I found. | | |

Sign/Firmar:_____Date: _____

## 3. Annex 3. Writing assessment directions level B1.2

Universidad del Norte – Instituto de Idiomas Level Four: **Writing Assessment Instructions & Writing Prompts**

**What am I going to do?**

You are going to ***write an opinion paragraph***. Your paragraph should include the following (see the rubric):

a) be ***approximately 8-14 sentences***. Remember, a sentence begins with a capital letter and ends with a period.
b) include at least one *compound sentence* (a sentence that joins two independent clauses using a coordinating conjunction like *and, or, but, so*)

> ***Example:*** Uninorte is the best university in the Caribbean, but it is much more expensive than Atlantico.

c) include at least one *complex sentence* (a sentence that joins a dependent clause with an independent clause using a subordinating conjunction like *because, if, after, whenever, although, as long as*)

> ***Example:*** Because there are many new free trade agreements, Colombia's economy will continue to grow.

OR

> ***Example:*** Colombia's economy will continue to grow because there are many new free trade agreements.

d) include at least one compound-complex sentence. You can combine a *simple sentence* with a *complex sentence* to create a *compound-complex sentence*

> ***Example:*** As soon as the movie started, my phone rang, but I did not answer it.

OR

> ***Example:*** I did not answer my phone when it rang because the movie had started.

**Step One:** Begin by choosing one of the following topics:

| Option One | Option Two | Option Three |
|---|---|---|
| Is it important for young adults to leave their parent's home and live on their own? Explain. | Is getting married still an important part of our traditions? Give good reasons why or why not. | Should capital punishment be abolished throughout the world? Explain why or why not. |

**Step Two: Brainstorming** (5 Minutes)

a) Take a few minutes to think about which option you want to write about.
b) Then, think about what you are going to write. Make a list of your ideas in the space provided.

**Step Three: Write an OUTLINE** (10 minutes)

a) Begin by thinking about a strong **opinion** statement. This is your TOPIC sentence. Write it down in the space provided for the topic sentence. It should be brief and to the point.
b) Now write at least 2 or 3 good reasons that answer why your opinion makes sense. Write your reasons in the space provided. Your reasons, or supportive details, should be brief and to the point.
c) Finally, write down a conclusion that restates your topic sentence in a different way. Make sure to use different vocabulary to express your conclusion.

**Step Four: Write a First Draft** (10 minutes)

a) Rewrite your OUTLINE as a paragraph on the space provided. Remember to **double-space** so your work is clear.
b) When you finish, read your paragraph and go through the **checklist**. Once your checklist is complete, it is time to write your final draft.

**Step Five: Rewrite** (5 minutes)

a) Rewrite your paragraph and hand it in before you leave. **Do not forget to include your name, date and a title.**

| CHECKLIST | YES | NO |
|---|---|---|
| I completed the writing process (brainstorming, outlining and drafting). | | |
| I wrote between 8 and 14 sentences. | | |
| I wrote a strong opinion-based topic sentence. | | |
| I wrote at least 3 reasons answering why I think this way. | | |
| I wrote at least 3 good examples to support my reasons. | | |
| I included at least one compound sentence | | |
| I included at least one complex sentence | | |
| I included at least one compound-complex sentence | | |
| I included at least 3 of the vocabulary words from the option I selected | | |
| I reviewed my writing for spelling and grammar errors and corrected any mistakes I found. | | |

Sign/Firmar:_____Date: _____

## 4. Annex 4. Writing assessment directions level B1

Universidad del Norte – Instituto de Idiomas Level Five: **Writing Assessment Instructions & Writing Prompts**

**What am I going to do?**
You are going to ***write an opinion in two paragraphs***. Your paragraphs should include the following:

e) be ***approximately 14-20 sentences***. Remember, a sentence begins with a capital letter and ends with a period.
f) include at least one *compound sentence* (a sentence that joins two independent clauses using a coordinating conjunction like *and, or, but, so*)

| |
|---|
| ***Example:*** Uninorte is the best university in the Caribbean, but it is much more expensive than Atlántico. |

g) include at least one *complex sentence* (a sentence that joins a dependent clause with an independent clause using a subordinating conjunction like *because, if, after, whenever, although, as long as*

| |
|---|
| ***Example:*** Because there are many new free trade agreements, Colombia's economy will continue to grow. |

OR

| |
|---|
| ***Example:*** Colombia's economy will continue to grow because there are many new free trade agreements. |

h) include at least one compound-complex sentence. You can combine a *simple sentence* with a *complex sentence* to create a *compound-complex sentence*

| |
|---|
| ***Example:*** As soon as the movie started, my phone rang, but I did not answer it. |

OR

| |
|---|
| ***Example:*** I did not answer my phone when it rang because the movie had started. |

**Step One:** Begin by choosing one of the following topics:

| Option One | Option Two | Option Three |
|---|---|---|
| Are commercials on TV honest with what they offer? Give examples. | Do you think that tourists pose a danger to the environment? Support your answer. | Should people who commit fraud be severely punished? Support your answer. |

**Step Two: Brainstorming** (5 Minutes)

c) Take a few minutes to think about which option you want to write about.

d) Then, think about what you are going to write. Make a list of your ideas in the space provided.

**Step Three: Write an OUTLINE** (10 minutes)

d) Begin by thinking about a strong **opinion** statement. This is your TOPIC sentence. Write it down in the space provided for the topic sentence. It should be brief and to the point.
e) Now write at least 2 or 3 good reasons that answer why your opinion makes sense. Write your reasons in the space provided. Your reasons, or supportive details, should be brief and to the point.
f) Finally, write down a conclusion that restates your topic sentence in a different way. Make sure to use different vocabulary to express your conclusion.

**Step Four: Write a First Draft** (10 minutes)

c) Rewrite your OUTLINE as a paragraph on the space provided. Remember to **double-space** so your work is clear.
d) When you finish, read your paragraph and go through the **checklist**. Once your checklist is complete, it is time to write your final draft.

**Step Five: Rewrite** (5 minutes)

b) Rewrite your paragraph and hand it in before you leave. **Don't forget to include your name, date and a title.**

| CHECKLIST | YES | NO |
|---|---|---|
| I completed the writing process (brainstorming, outlining and drafting). | | |
| I wrote between 14 and 20 sentences. | | |
| I wrote a strong opinion-based topic sentence. | | |
| I wrote at least 3 reasons answering why I think this way. | | |
| I wrote at least 3 good examples to support my reasons. | | |
| I included at least one compound sentence | | |
| I included at least one complex sentence | | |
| I included at least one compound-complex sentence | | |
| I included at least 3 of the vocabulary words from the option I selected | | |
| I reviewed my writing for spelling and grammar errors and corrected any mistakes I found. | | |

Sign/Firmar:_____Date:
_____

## 5. Annex 5. Writing assessment directions level B2

Universidad del Norte – Instituto de Idiomas
**INGLES CONTENIDO 1 (Level 7)**

**Writing Assessment for Multifaceted Project: Compare & Contrast Essay**

**Name:**

***Warning: Did you give your teacher your essay from Chapter 5?*** He/she won't accept this assignment until he/she sees your comparison & contrast essay.

---

**What am I going to do?**
You are going to write a ***compare and contrast essay.*** Your essay must have **four or five paragraphs**. Remember, a paragraph usually has 5-7 sentences. Choose one of the following options for your essay:

| **Option 1** |
| --- |
| The risks of fixing Barranquilla´s flashflood problem<br><br>***VERSUS***<br><br>The risks of not fixing Barranquilla´s flashflood problem. |

| **Option 2** |
| --- |
| The risks of putting students with special needs, whether it be physical or mental, in separate classrooms<br><br>***VERSUS***<br><br>The risks of having classrooms that students with and without special needs both share |

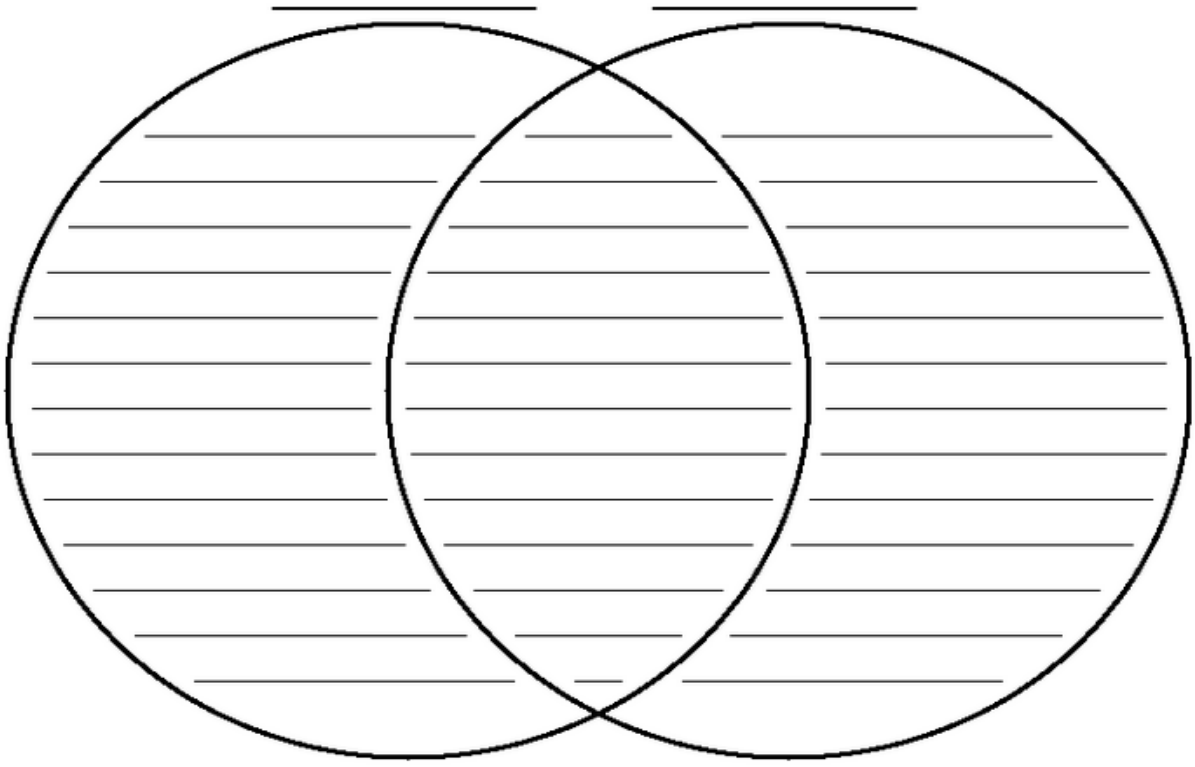
| **Option 3** |
| --- |
| The risks of violent and nonviolent criminals being sent to the same prisons<br><br>***VERSUS***<br><br>The risks of creating prisons for different types of offenders |

| **Option 4** |
| --- |
| The risks of treating cancer through diet only (or Alternative medicine)<br><br>***VERSUS***<br><br>The risks of treating cancer through chemotherapy |

| **Option 5** |
| --- |
| The risks of the artist producing work strictly using established formulas<br><br>***VERSUS***<br><br>The risks of the artist creating work exclusively incorporating original concepts |

**Step One: Brainstorming** (45 minutes)

e) Work with a partner or a group. Take turns talking about all four options. You don't have to write anything.
f) Independently, decide which option you want to write about.
g) Think about what you are going to write about. Put your ideas in the Venn Diagram provided below. Include **examples and details**. Your ideas should be rough and in point form. Don't worry about grammar, spelling, or using complete sentences when you are brainstorming.
h) After you finish, show your Venn Diagram to your teacher. He or she will tell you if you need more ideas.



i) Write comparison and contrast sentences based off your diagram in the space below.

Refer to the comparison and contrast conjunctions on p.99 (Reading & Writing Chapter 5).

_____
_____
_____

**Step Two: Research** (60 minutes)

a) One, out of three, of your sources must be from our textbook. You must include a direct or indirect quote from one of the three readings in Chapter 7: *How risky is it, really? Why our fears don´t always match the facts* (p. 139), *Risk: a practical guide for deciding what´s really safe and what´s really dangerous in the world around you* (p.145-147), or *Risk, democratic citizenship and public policy* (p.150-152).
b) Read the information below about how to use Universidad del Norte's library database.
c)

**How to Use the Library Database for Writing an Essay in English**

1. Log into your Uninorte account at http://www.uninorte.edu.co/

2. Click on "Mis Servicios."

3. Click on "Biblioteca Karl C. Parrish."

4. Click on "Multidisciplinarias." On the right-hand side, you´ll find a list of databases. Possible ones you can use are Encyclopedia Britannica Online, ProQuest, Jstor. If you use the Encyclopedia Britannica Online, it counts as only one source of information. You still need to find one other source. One, out of three, of your sources must be from the library data base.

5. Here`s an example of how to navigate ProQuest when searching for information: Click on "ProQuest."

6. Click on "Ciencias sociales."

7. Scroll down until you find "ProQuest Research Library: Social Sciences." Click on "Search."

8. Click on "Avanzada." In the spaces provided, enter terms to search, for example "construction risks." You might find that your original search doesn't bring up any material; conversely, you might find that your search is too general and you have too many options to look through. "Construction risks" brought up 127,366 results. So, I need to further refine my search. Go back to the "Avanzada" page and add more words if you need to. I added "coast" and "ecological." That resulted in 3,076 results. Don´t get frustrated as it can take time to find good sources. Try changing the words that you're using to search for material.

9. On the "Avanzada" page, there is a "Opciones de busqueda" section. There you can click on "Textocompleto" which will limit your search to complete articles. You need complete articles for your paper. A summary, review, or synopsis of an article will not work. **All of your sources must be in English.**

10.You can click on the "Citar" button in the top right-hand corner of the webpage to see the reference information for each source.You will want to copy and paste the citation for each of your sources as you will need these for the Reference List for your paper.

d) Using the above instructions, find **two** sources for your essay and save them. You can email the PDF file to your Uninorte email account or save the file on a flash drive (USB). Do not email your sources to your Hotmail or Facebook accounts as many lab computers block access to these accounts. **Remember, you will be writing your essay in class so you need to have easy access to your sources.**

e) While <u>you must write the paper in class</u>, you can complete your References ahead of time and email them to yourself. Remember that your reference list must be in alphabetical order.

f) You will receive a zero (0%) on this assignment if you do not write both first and final drafts of the paper, write the paper outside of class, or plagiarize, such as copy and paste from a website or another student's paper.

**Step Three: Prepare to Write** (60 minutes)

a) Using the information you wrote and gathered in Step One and Step Two, choose ONE of the outline templates below and create an outline for your essay. The outline may include, at most, 2 complete sentences per section. You should not write your first draft in your

outline. You need to include three citations from your sources. You need to read your articles before class so that you can spend your class time writing your outline, not reading.

**Block Style Outline: Compare & Contrast 2 Risks**

| **Paragraph 1: Introduction w/ Thesis** |
| --- |
| A. Hook:<br><br>B. Connecting information:<br><br>C. Thesis statement: |
| **Body Paragraph 2: Risk One** |
| A. Topic Sentence:<br><br>B. Supporting details & examples:<br><br>C. Supporting details & examples:<br><br>D. Supporting details & examples:<br><br>E. Supporting details & examples: |
| **BodyParagraph 3: Risk Two** |
| A. Topic Sentence:<br><br>B. Supporting details & examples:<br><br>C. Supporting details & examples:<br><br>D. Supporting details & examples: |

| |
|---|
| E. Supporting details & examples: |
| **Paragraph 4: Conclusion** |
| A. Restated Thesis:<br><br>B. Final thoughts/suggestions/ideas: |

**Point-by-Point Style Outline: Compare & Contrast 2 Risks**

| | |
|---|---|
| **Paragraph 1: Introduction w/ Thesis** | |
| A. Hook:<br><br>B. Connecting information:<br><br>C. Thesis statement: | |
| **Body Paragraph 2: 1st Comparison & Contrast** | |
| A. Topic Sentence: | |
| **Risk One** | **Risk Two** |
| B. Supporting details & example | C. Supporting details & examples: |
| **Body Paragraph 3: 2nd Comparison & Contrast** | |
| A. Topic Sentence: | |
| **Risk One** | **Risk Two** |
| B. Supporting details & examples: | C. Supporting details & examples: |
| **Body Paragraph 4: 3rd Comparison & Contrast** | |

| A. Topic Sentence: | |
|---|---|
| **Risk One** | **Risk Two** |
| B. Supporting details & examples: | C. Supporting details & examples: |
| **Paragraph 5: Conclusion** | |
| A. Restated Thesis:<br><br>B. Final thoughts/suggestions/ideas: | |

b) When you have completed your outline, show it to your teacher. He or she will tell you if you need to add more information.

**Step Four: Write a First Draft** (120 minutes)

a) Look at the rubric that your teacher will use to give you a grade. What is he or she looking for? Remember these points when you begin to write.

b) You must write your essay in the lab, on the computer. Make sure you type double-spaced in Times New Roman 12pt font.

c) Make sure you include ***an introduction paragraph, two or three body paragraphs (with topic sentences), and a conclusion paragraph***.

d) Remember to include ***a thesis statement*** and to use a variety of ***simple sentences***, ***compound sentences, complex, compound-complex sentences***, **and adjective clauses**, as well as proper capitalization in your title.
You can't work on your essay at home, but you should bring your outline to class. However, you can use your textbook, the Writing Supplement, Spanish-English Linkers, and your three sources, which have been approved by your teacher. You can also ask classmates for help! The following websites are useful tools:
- Online dictionary (www.wordreference.com or http://www.merriam-webster.com/)
- Thesaurus for synonyms and antonyms (http://thesaurus.com/)
- Purdue's OWL website for APA formatting: In Text Citations(https://owl.english.purdue.edu/owl/resource/560/02/) & Reference List (https://owl.english.purdue.edu/owl/resource/560/05/)
- Collocations, like common preposition and verb combinations (http://www.just-the-word.com/)

e) Integrate two direct quotes of no more than 30 words each and one indirect quote from your academic sources and your textbook. Keep in mind that you need to introduce the

quote, use correct citation, and explain the relevancy of the quote to your paper. **Do not simply stick a quote in your paper without connecting it to your own ideas.** Refer to pages 110-113 in *Chapter 5 Reading & Writing* on how to do correct in-text citations of direct and indirect quotes as well as referencing.

f) Change your keyboard to English by clicking on the little "ES" symbol at the bottom right-hand corner of the taskbar. Select "EN" for Ingles.

g) Keep in mind the following Microsoft Word tips:

- Language Selection - Highlight text. Click on "Revisar". Click on "Idioma". Select "Ingles".
- Double-space - Highlight text. Right click over highlighted text. Click on "párrafo". Select "doble".
- Spell Check – Click on "Revisar" and "ABC Ortografia y gramática". This will help you with spelling and grammar mistakes.

h) **Do not use Google Translate or Wikipedia while you are writing your essay!**

i) Writing your essay is an assessment. Therefore**, using your cell phone while writing is prohibited.**

j) Refer to *Model 6* on pages 187-190in of your textbook for help with compare and contrast essay structure, pages 147-148 in *Chapter 7 Reading & Writing* for help with summarizing, and pages 120-123 in *Chapter 6 Reading & Writing* for help with paraphrasing.

k) In order to submit your paper, go to our Blackboard page and click on Content. Then find the Safe Assign check mark icon called "Compare & Contrast Essay." Click on the ´´View/Complete ´´ link. Upload your essay from the computer. Click on the ´´Submit´´ button. Now, your teacher has your essay as well as a report on how much, if any, material was plagiarized.

l) Keep in mind, **your first draft is worth 5% out of your 15% essay grade so do the best that you can on your first draft.**

**Step Five: Write a Final Draft** (120 minutes)

e) Look at your teacher's feedback on your first draft. Make sure you understand all his or her comments. Use this information to help you improve your final draft. You must revise content as well as grammatical errors.

f) Rewrite your final draft ***in class***. Again, remember to type your essay double-spaced in Times New Roman 12pt font.

g) When you finish, read your essay one more time. Are there any mistakes?

h) On Blackboard, email your teacher your final draft. He or she will compare the content and grammatical changes you made between your first and final draft when giving you a grade. Keep in mind that **your final draft is worth 10% out of your 15% essay grade.**