ICDEL Journal,
Vol.3, No. 1 (2018)
Design and Analysis of an Annotated Gender-Based English

Design and Analysis of an Annotated Gender-Based English Learner Corpus

María G. Ballesteros Chica & Javier Fernández-Cruz

Pontificia Universidad Católica del Ecuador Sede Esmeraldas.

Esmeraldas. Ecuador

Email for correspondence: javier.fernandez@pucese.edu.ec

Receipt date: March 7th, 2018

Approval date: May 6th, 2018

How to cite this article (APA Norms)

Ballesteros, M., & Fernández-Cruz, J. (2018). Design and Analysis of an Annotated Gender-Based English Learner Corpus. International Congress on the Didactics of the English Language Journal, Vol. 3, No.1. ISSN 2550-7036. Retrieved from http://revistas.pucese.edu.ec/ICDEL/index

ICDEL Journal,
Vol.3, No. 1 (2018)
Design and Analysis of an Annotated Gender-Based English

Abstract

This research focuses on the study of gender as a factor of influence of the performance from a corpus linguistics perspective. We compiled a university level English learner corpus from written sample texts produced in class, a linguistic analysis was carried out by means of the application of quantitative and qualitative methods. As a result, we obtained a systematic summary of errors in the learner corpus while taking several parameters into account, such as age, level and, especially, gender. Results revealed that errors were more prevalent in male students than on female at the time of composing general domain texts in English. In addition, we present a summary of the most frequent errors and a discussion of the factors that may have influenced these results.

Keywords: Corpus Linguistics, Learner corpus, gender studies, English as a Second Language, errors

Resumen

Esta investigación se centra en el estudio del género como factor de influencia del desempeño desde una perspectiva lingüística del corpus. Recopilamos un corpus de estudiantes de inglés de nivel universitario a partir de textos de muestra escritos producidos en clase, se realizó un análisis lingüístico mediante la aplicación de métodos cuantitativos y cualitativos. Como resultado, obtuvimos un resumen sistemático de los errores en el corpus de estudiantes, teniendo en cuenta varios parámetros, como la edad, el nivel y, especialmente, el género. Los resultados revelaron que los errores eran más frecuentes en los estudiantes varones que en las mujeres al momento de componer textos de dominio general en inglés. Además, presentamos un resumen de los errores más frecuentes y una discusión de los factores que pueden haber influido en estos resultados.

Palabas clave: Lingüística de corpus, Corpus de aprendices, estudios de género, inglés como segundo lenguaje, errores

Introduction

Does gender influence the type of errors committed during the English learning process? Many studies at the PUCE Esmeraldas observed the factors that have influenced the process of learning English as a Second Other Language (ESOL) such as interference (Olivo Tello, 2017; Rúa Castillo & Fernández Cruz, in this volume), motivation (Saltos Intriago, 2017; Estupiñán Boboy, 2017). The main goal of this chapter is to make a diagnosis of the English writing errors in students from 6th level of General English at PUCE Esmeraldas through the use of computational tools, but also as an introduction to the use of corpus-related research in Esmeraldas.

In the past, such an observation would be labor-intensive. Before the widespread use of computers, analyses of this nature were done by observation sheets and annotations. With the advent of computational linguistics, we can obtain accurate results through automatic processes. Specifically, corpus-based methodologies permit the empirical observation of linguistic evidence from large compilation of electronic texts containing real samples as produced by its speakers.

This research is focused on the grammatical analysis of texts written by students in 6th level of general English at PUCE Esmeraldas while considering the gender differences of the students. Learner corpora can be defined as electronic collections of language data produced by foreign-language learners.

These resources provide new horizons and opportunities for teaching and learning English as a foreign language. To sum up, this chapter aims to find answers to the questions asked previously and, from the preliminary results seeks to make several proposals in order to improve the observed problems.

Corpus Linguistics and Learner Corpora

Sinclair (1996) defined corpus as a "a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research". Likewise, Hunston (2002) states that Linguistics has used the term corpus to describe a collection of naturally occurring examples of language, consisting of anything form a few sentences to a set of written texts or tape recordings, which have been collected for linguistic research. After the skyrocketing development of computing, developments of corpus collections have passed from containing a few thousand of words to the capacity of collecting multimillion corpora from the web or mass text processing at the speed of a click. Besides, thanks to the advantages of technology, corpora collect vast texts from many different sources suitable to be published online and ready to be analyzed by linguists around the world.

After three decades of epistemological and methodological development, depending on the needs of linguistic analysis, the main types of corpora are the following: specialized or not, monolingual/multilingual, parallel, un/annotated, synchronic/diachronic, etc. Particularly, according to O'Keeffe, McCarthy & Carter (2007), if texts are classified, corpora can provide insights on specific characteristics of language use. A properly tagged learner corpus may provide valuable information e.g. age, gender, location, school, level, teacher, class, etc. In order to provide a concise McEnery, Xiao & Tono (2007) described five different kinds of corpus. First, specialized corpora are collection of texts of a certain genre designed to investigate a specific type of language, such as newspaper editorials, textbooks, medicine academic articles, casual conversations, etc. Second, general corpora contain texts just in one language. This type of corpus is usually tagged for parts of speech and is used for a wide range of uses from lexicography to Natural Language Processing. Third, parallel corpora generally contain two monolingual corpora, being one corpus the translation of the other. They are generally aligned with software. For example: the European Central Bank press conferences and their translations. Fourth, a learner corpus is characterized as a group of texts produced by language learners. This type of corpus is used to study errors and problems students may have during while learning a second language. Fifth, diachronic corpora contain texts from extended time periods and labeled e.g. by year. They are used to study the development or language change. In addition, some corpus processing tools provide a

specialized diachronic feature referred as 'trends', which identifies words whose usage changes significantly during the selected period.

O'Keeffe et al. (2007) emphasize on the fact that corpus-based approaches provide the opportunity to apply quantitative and qualitative analysis that may provide many advantages in comparison with traditional linguistic approaches. First, a corpus can be more comprehensive and reliable. A corpus analysis offers linguistic evidence from a range of language speakers. This permits to observe the small picture and observing the use of language from individual speakers, but also it may offer several examples in a real communicative context. Second, a corpus can show us what is common and typical. For instance, learners of a foreign language can search specific terms to know whether the use a certain language pattern is common or not. Moreover, we can find differences that cannot be perceived by mere intuition. For instance, while coarsely considered synonyms, there are differences in the use of 'totally' and 'absolutely', 'utterly', 'completely' or 'entirely'. Third, from corpora we can obtain statistic data in order to quantify frequencies of single words, clusters or collocations. This allow researchers to obtain insights that are not observable to the naked eye, for example the percent of errors among men and women as in this research. After that, corpus data is more natural because it is used in real communications instead of being invented specifically for linguistics analysis. Besides, a constantly updated corpus can reflect even recent changes in the language, so it is important because we can learn about the new language tendencies to be applied in our speaking (Xiao & Tao, 2009).

Learner corpus research is already a mature branch of corpus (first studies date from late 1980s) linguistics but still developing with dynamism. Learner corpora collect texts produced by language learners. According to Granger (2008), they fulfil two distinct but intrinsically related purposes: they can contribute to further knowledge of SLA by offering a more precise account of interlanguage and a better understanding of the factors that influence it; and, as it is our case, they are ideal to the development of pedagogical implements and methods in order to aim for the needs of language learners in a more accurate way.

A typical learner corpus is characterized by the design of an annotated corpus or *corpus mark-up* which provides advantages if compared with traditional research, as the information noted to the text (i.e. error tags) permits to quantify and observe data e.g. the type of errors. Besides, tagging can be automated through tagging software (e.g. Penn TreeBank) or manual (this task is made by humans who follow tagging conventions). As Hunston (2002) points out, the level of accuracy achieved depends on the automatic or manual tagging. Even though in the first case a corpus can be tagged entirely, the level of accuracy is lower (however in our decade the standards are quite high!) than those resulting from manual tagging, however this task is time consuming and, due to economic reasons, it is only feasible when the corpus is small.

Corpus-based studies are scarce in Ecuador. In the case of Esmeraldas, Castillo (2016) used corpora as a teaching-learning material. She concluded that the use of corpora is not the definitive solution for improving vocabulary level yet, it proved to be an effective optional tool available for teachers at the time of designing materials for their classes. In addition, the results of the survey suggested that working with corpora can provide some benefits such as exposure to realistic language, and interactivity which may challenge students and increase their interest.

ICDEL Journal,
Vol.3, No. 1 (2018)
Design and Analysis of an Annotated Gender-Based English

Gender, writing and errors

For learners, writing in a foreign language is not a simple task. It is muddled to write in other language and sometimes most of the understudies try to decipher or translate words, expressions, and sentences from the first language to English obtaining unpleasant outcomes (Benson, 2002). Agreeing with Richards & Renandya (2002), writing is the most problematic skill for EFL students to master. The problem here is creating and assembling thoughts while having an accurate production, lexicogrammatically accurate, as well as in making an interpretation of these ideas into understandable content.

For us, the main challenge that teaching presents here is the discovery of strategies that may allow to activate definitively written skills. At that point, it is basic here to show a difference between errors and mistakes. For instance, Corder (1967) reveals a criterion that helps us to do so: mistakes can be self-corrected, but an error cannot. Errors are systematic i.e. liable to occur habitually and unrecognized by the learner and thus, only a third person is suitable to point them out (Gass & Selinker, 1994).

The following concern is related to gender differences during their SL performance. Gender has a leading role among the factors in the use of foreign language as males and females visibly differ in various perceptual, motor and cognitive domains (Halpern & La May, 2000; Kimura, 1999; Ali, 2016).

A considerable number of investigations (Noguchi,1991, Green & Oxford, 1995; Boroomand & Rostami Abusaeedi, 2013) evidenced that gender had, up to a certain extent, an effect on how students learn a language. Many studies that examined gender as a variable in the use of language learning strategies reported that there were significant gender differences, i.e. there is an agreement in of a better performance of language learning strategies in female students. As for illustration purposes, Zoghi et al. (2013) studied the performance of EFL learners from a group of 50 boys and 50 girls aged 12-14. Accomplishment tests were used as instrument of observation including four sections: vocabulary, language structure, sentence capacity, and perusing appreciation, which evaluated general capacity in student's production. Results revealed a considerable impact of gender in the students' accomplishment test.

Method

A corpus-based methodology was designed in order to detect features in order to detect writing errors according to gender. To do this, a five-step methodology was followed: (1) compilation of writing output through an online form, (2) manual detection and annotation of errors, (3) XML labeling and final corpus compilation, (4) automatic calculation of error statistics, (5) a final observation of examples in order to detect exceptions and relevant features.

The population of this study was a class of 6th level General English at PUCE Esmeraldas (like ECFL level B1) which include students from different programs, being most of them undergraduate students of Environmental Management. Our intentional sample included all the 24 students that regularly attend the class. There were 11 male and 13 female students.

The resulting corpus, codenamed *Esmeraldas Learner Corpus*, includes 3732 tokens, and was produced after an in-class writing exercise. To do this, an activity was designed in Google Forms. The activity consisted of writing a paragraph about travelling, and demanded metadata such as age, gender and degree. Results were converted to UTF8 and its metatags were labeled in XML. The errors were annotated manually

using Dahlmeier, Ng & Wu's (2013) error tags. The resulting corpus was analyzed with the AntConc concordancer (Anthony, 2014). A concise list of error tags is found in Table 1 and the learner corpus is available under demand.

Results

Male students committed a total sum of 119 errors (averaging 10.82 per individual). Meanwhile, women, being more abundant, only summed only 109 errors (an average of 8.38 per individual). In Table 1 we can observe that males were especially outperformed in noun and article use and spelling (mechanics), while women tended to be more redundant and performed especially worse at the use of transitional connectors. A summary of errors can be found bellow for illustration purposes.

Verb Errors

In this section 11 of 24 students used erroneously the verb selection and tense of whom 60% of were produced by males. For example:

(1) Within (Wcip) Colombia I know (vt) (vform) Cali, (trans) Ipiales.
(2) My best vacation was go to yasuni itt where we were 1 weekend and visit (vt) many places interesting (WOadv).
(3) To spend the day at the beach with friends or family make (vt) games and eat (vt).

Article / Determiner Errors

67% of the errors corresponded to males while the 33% corresponded to females. For all, this figure shows to 6 of 24 students had difficulties in this case. For instance:

(4) (mec) have (ArtOrDet) good experience.
(5) Sunday at noon we usually go to (ArtOrDet) temple with my family

Nouns

Generally, students have evidenced good management in the use of nouns in sentences. Out of the 5 students that committed errors in this category. In this case, only 40% of the errors were committed by men. For example:

(6) This countries (nn) is (vform) impresionated.
(7) Sports and health life are my other hobbie (nn) .

At the time of using genitive, but a low number f students had problems with this rule. In fact, 4 students (2M, 2F) committed errors:

(8) I need a capacitation about English by Language student's. (Npos)

Word Order

Word order is a common problem in students mainly because of L1 interference, especially adjective use. This figure shows to 4 students who wrote incorrectly the adjective position. 86% of errors were committed by men and the other 14% were committed by women.

(9) I love them very much and have I (WOinc) a lot of fun with them.

Table 1

Summary of results

ICDEL Journal,
Vol.3, No. 1 (2018)
Design and Analysis of an Annotated Gender-Based English

Error Tag	Error Categories	M	F	Total
Verbs				
VT	Verb tense	12	8	20
VFORM	Verb form	4	1	5
Subject-Verb Agreement				
SVA	Subject-Verb-Agreement	7	1	2
Articles/Determiners				
ArtOrDet	Article or determiner	4	2	6
Nouns				
Nn	Noun number	2	3	5
Npos	Noun possessive	2	2	4
Pronouns				
Pform	Pronoun form	6	1	7
Pref	Pronoun reference	0	5	5
Wcip	Wrong collocations, idioms and prepositions	2	3	5
Sentence structure				
Srun	Runnons, comma splice	31	32	63
Word order				
Woinc	Incorrect sentence form	9	11	20
Woadv	Adverb/adjective position	6	1	7
Transitions				
Trans	Link words/phrases	9	20	29
Mechanics				
Mec	Punctuation, capitalization, spelling and typos	30	15	45
Redundancy				
Rloc	Local redundancy	1	4	5
TOTAL		119	109	228
AVG PER				
ST.		10.81	8.38	
%		52.19	47,81	

Discussion and conclusions

Corpus provides an interesting and easy way of observing the English learning process as it provides relevant information, while data are considered fast to obtain, realistic, aseptic, and confidential. It simplified the analysis of a small 3172 token corpus in a short time span, which allowed to spend more time observing language use than obtaining data.

This is one of the first approaches to corpus studies on learner produced text in Esmeraldas. Even if it has been a very preliminary analysis, the experience and results are encouraging and this I a first step towards a future large-scale corpus-based analysis by expanding the sample and population.

Results evidence that female language learners commit less writing errors (avg. 10.81 males and 8.38 females). In conclusion, women perform better than man when they are writing in EFL. All in all, students presented more problems in writing errors referring to punctuation, followed by mechanic errors, such as capitalization and spelling. Finally, the third highest range of errors was related to transition words.

Efforts at the Lexicography and Corpus research team at PUCE Esmeraldas are focused on providing training in corpus-based analysis to both teachers and students. Corpus-based approaches will be applied as a tool in EFL classes, which would provide a powerful empirical tool for teacher's research, in order to obtain large amounts of language samples from real situations and to provide situated solutions in Esmeraldas in order to determine the sociocultural causes of the higher prevalence of errors.

References

Ali, H. O. (2016). Gender differences in using language in the EFL classes: From teachers‟ views. International Journal of Humanities and Cultural Studies (IJHCS) ISSN 2356-5926, 2(4), 73-91.

Anthony, L. (2014). AntConc [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Boroomand, Faezeh & Abusaeedi, Ali. (2013). A gender-based analysis of Iranian EFL learners' types of written errors. International Journal of Research Studies in Language Learning. 2. 10.5861/ijrsll.2013.287.

Castillo Jaen, S. J. (2017). Do Corpora Benefit the Level of Vocabulary of 10th Year of Basic Education Students? Case Study. *International Congress on the Didactics of the English Language*, *2*(1).

Corder, S. P. (1967). The significance of learner's errors. IRAL-International Review of Applied Linguistics in Language Teaching, 5(1-4), 161-170.

Dahlmeier, D., Ng, H. T., & Wu, S. M. (2013, June). Building a Large Annotated Corpus of Learner English: The NUS Corpus of Learner English. In BEA@ NAACL-HLT, 22-31.

Estupiñán Boboy, E. M. (2016). Motivation in the Classroom: Factors That Motivate Students to Learn English at PUCESE. International Congress on the Didactics of the English Language, 1(1).

Gass, S. M., & Selinker, L. (1994). Topics in applied psycholinguistics. Second language acquisition: An introductory course. Hillsdale, US: Lawrence.

Granger S. (2008). Learner corpora. In Lüdeling, A. & Kytö, M. (eds.) Corpus Linguistics. An International Handbook. Volume 1. Berlin & New York: Walter de 5 Gruyter, 259-275.

Green J. & Oxford RL.1995: A Closer Look at Learning Strategies, L2 Proficiency, and Gender. TESOL Quarterly, 29, 261-297.

Halpern, D. F., & La May, M. L. (2000). The smarter sex: A critical review of sex differences in intelligence. Educational Psychology Review, 12(2), 229-246.

Hunston, S. (2002). Corpora in applied linguistics. Cambridge: CUP.

Kimura, D., & Clarke, P. G. (2002). Women's advantage on verbal memory is not restricted to concrete words. Psychological reports, 91(3_suppl), 1137-1142.

Noguchi T. 1991. Review of language learning strategy research and its implications. Unpublished bachelor's thesis, Tottori University, Tottori, Japan.

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). From corpus to classroom: Language use and language teaching. Cambridge: CUP.

Olivo Tello, K. M. (2017). The Influence of Spanish on the Pronunciation of the English Phonemes /t/ and /d/ in Students of the Eighth Level of the International Commerce Career at PUCESE. *International Congress on the Didactics of the English Language*, *2*(1).

Richards, J. C., & Renandya, W. A. (Eds.). (2002). Methodology in language teaching: An anthology of current practice. Cambridge: CUP.

Saltos Intriago, M. F. (2017). Intrinsic and Extrinsic Motivation in Senior High School Students from Margarita Cortés Educational Institution in Esmeraldas, Ecuador. International Congress on the Didactics of the English Language, 2(1).

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

Xiao, R., & Tao, H. (2007). A corpus-based sociolinguistic study of amplifiers in British English. Sociolinguistic studies, 1(2), 241-273.

Zoghi, M., Kazemi, S. A., & Kalani, A. (2013). The effect of gender on language learning. *Journal of Novel Applied Sciences*, 1124-1128.