Anuario de Letras Modernas

VOLUMEN 25 | NÚMERO 1 | mayo – octubre 2022
issn: 2683-3352 | pp. 91-105
doi: 10.22201/ffyl.01860526p.2022.25.1.1708
Recibido: 29-10-2021 | Aceptado: 23-02-2022

Lexical Analysis Using Corpus Linguistics: Comprehending How EFL Students Are Writing

Análisis léxico a través de la Lingüística de Corpus: la comprensión de cómo escriben los estudiantes de ILE

Natin Guzmán Arce
Jimmy Ramírez Acosta
Sonia Rodríguez Salazar

Escuela de Literatura y Ciencias del Lenguaje
Universidad Nacional | Heredia, Costa Rica

Contacto: natin.guzman.arce@una.cr, jimmy.ramirez.acosta@una.cr, sonia.rodriguez.salazar@una.cr

Abstract

This article shows one specific object of study using Corpus Linguistics: adjectives from a corpus of academic texts from efl students collected between 2017 and 2018 revealing a dissimilar variety compared to native speakers' corpus in similar types of texts. This nearly half a million corpus comprises students' academic writings from two English Teaching majors and one English BA from the School of Literature and Language Sciences in Universidad Nacional in Heredia, Costa Rica. What the lexical unit of the study displays and what it reveals is surprisingly interesting because, as efl teachers, we expected a more diverse and profound use of adjectives as they are an important device in academic prose. Written texts seem to reflect a fault in the teaching and learning of this skill, calling for immediate attention to the matter. The use of the concordancer AntConc© as the linguistic analysis software selected to manage the lexical categories that the researchers picked as a starting point was truly significant in the methodological process. A corpus-driven approach was followed in this initial attempt to illustrate word choice in academic writings developed in the composition courses during the time period: English Integrated I, English Integrated II, Composition,

Resumen

El artículo muestra un objeto de estudio en específico utilizando la Lingüística de Corpus: los adjetivos de un corpus de textos académicos de aprendientes de inglés recolectado entre los años 2017 y 2018 y expone que la variedad es diferente al cotejarlos con un corpus de textos comparables de nativo hablantes de la misma lengua. Este corpus de casi medio millón de palabras contiene las composiciones académicas de estudiantes de dos licenciaturas en la Enseñanza del Inglés y de la Licenciatura en Inglés de la Escuela de Literatura y Ciencias del Lenguaje de la Universidad Nacional en Heredia, Costa Rica. Lo que el uso de esa unidad léxica refleja y revela es bastante interesante ya que, como profesores de lengua inglesa, esperábamos un uso más diverso y profundo de los adjetivos como elementos importantes en la narrativa académica. Estos textos escritos parecen reflejar una falla en la enseñanza y aprendizaje de esta destreza y llama a tomar acción inmediata al respecto. Fue muy importante en la metodología del estudio el uso del software o recurso de concordancia AntConc© que facilitó la administración de las categorías léxicas que se escogieron como etapa de inicio. Se siguió un enfoque inductivo (corpus-driven approach) para esta primera parte, como forma and Essay. This study was longitudinal as it involved the same students' writings during the two years of the corpus collection. General findings show that lexical complexity is deficient and suggest that more didactic efforts should be made to encourage the learning and acquisition of vocabulary, being the use of native speakers' corpus one example of such strategies.

Keywords: corpora (linguistics); English language teaching; academic writing; lexicology; foreign speakers

de mostrar cuál fue el léxico utilizado en los escritos académicos en los cursos de composición durante esos dos años, a saber: Inglés Integrado I, Inglés Integrado II, Composición y Ensayo. Este también es un estudio longitudinal porque tomó en cuenta a los mismos estudiantes durante el periodo de la recolección del corpus. Los resultados generales muestran que la complejidad léxica es deficiente y sugieren realizar mejores esfuerzos didácticos para el aprendizaje y la adquisición de vocabulario como, por ejemplo, el uso de corpus de nativo hablantes del inglés como estrategia de enseñanza.

Palabras clave: corpus lingüístico; enseñanza del inglés; escritos académicos; lexicología; hablantes extranjeros

Theoretical framework

Corpus Linguistics is not a new methodology, yet it has greatly supported language studies and applied linguistics since its conception. Sara Laviosa (2002: 8) considers Corpus Linguistics a unique methodology for the study of language as it is strongly supported by four interdependent but equally important elements: data, description, theory, and methodology.

Corpus Linguistics studies language based on linguistic examples from real life (McEnery & Wilson, 2001: 1), which is possible because corpora have accurate data and provide empirical information. Corpus design reveals a model of the reality one desires to study, and it aids researchers and teachers comprehending how human language works (Torruela & Llisterri, 1999: 4). Corpus Linguistics is supported by specialized software that analyzes enormous quantities of language data. This software is used to observe, examine, and process a significant corpus more efficiently and faster than the human eye. Nonetheless, an organized and planned corpus design can benefit the organization, the presentation, and the validity of the texts to congruently systematize such data. A computerized corpus, as Torruela & Llisterri (1999: 7) call it, is a compilation of selected texts under some linguistic criteria, which are coded in a standard and homogeneous manner, with the purpose of being computer analyzed to present how a language or languages behave. It can be said that it is the Big Data of linguistic studies as the whole universe of texts is the evidence for researchers to explore, as opposed to a sample of texts, providing bold descriptions and conclusions provided that the methodology has been carried out rigorously.

A systemic construction and research of several monolingual and multilingual corpora have stepped into other study fields such as descriptive and applied linguistics. This is because Corpus Linguistics is solid and innovative given the evidence it can provide. These fields include translation, translation studies, and English as a second or foreign language (ESL/EFL). For instance, some researchers have launched descriptive studies for different language areas using computerized corpora, namely prosodic, lexis, morphology, syntax, history, and the like (Torruela & Llisterri, 1999: 3).

How reliable corpora are will depend on the type of research questions the researcher poses. McEnery & Wilson (2001: 27) suggest a clear idea: *what can a linguist ask without a corpus to reason the impact of the use of corpora?* There are many approaches to different areas of the language; one is grammatical description, the area of study this paper will focus on. Supported by concordancers, the number of queries researchers can develop in a corpus are endless, fast, and accurate as opposed to doing the same analysis without that data. Not that the hand-and-eye does not work, but how long will this procedure take? What is the expectancy of human error? (McEnery & Wilson, 2001: 27)

Compiling students' texts from EFL classes makes it possible to build a data source that shows the language interference or interlanguage between the first and second languages. It also establishes an empirical basis for error analysis and communicative strategies (Torruela & Llisterri, 1999: 5). Such bases are much closer to reality to language studies than intuitive methods (Torruela & Llisterri, 1999: 9). This new knowledge is very significant for language instructors to pinpoint how their students are writing and very likely identify why and what is occurring in the acquisition process. This study has aimed at showing how this process works.

The new theory that has been included for the current study is that of lexical complexity and richness, defined by Ai & Lu (2010) in the following subcategories:

1. Lexical Density: percentage of lexical words
2. Lexical Sophistication: coverage of advanced vocabulary

3. Lexical Diversity: the number of different words (word types) compared to the total running words (tokens)

All categories are self-explained clearly in their own right, yet lexical diversity will be illustrated with all the of examples that follow in the Results Discussion section.

Methodology

Corpus investigation has widened many fields of studies such as linguistics, sociolinguists, language evolution, culture-social dialectal, or epidemiological phenomena among others. Having this in mind, this corpus project carried out in the School of Literature and Language Sciences (ELCL, Spanish initials) yielded a product of approximately half a million word-tokens (the total of running words) from the gathered data based on the collection of written papers developed by the students in similar majors offered in the ELCL. Since the area of writing entails many aspects of language, the researchers decided to break down the information in different language aspects—for example, the analysis of modal verbs, nouns, adverbs, and adjectives usage, the latter being the only one illustrated here.

The learner corpus is not very large compared to some that surpass the million word-tokens, yet sufficient to achieve the main objective of the study. Biber, Conrad & Reppen (2014: 123) demonstrated that even 1000 words of data can give results that are reliable, and this learner corpus is not the exception as it followed strict criteria in the process of collection and analysis. In its initial phase, the study encompassed two major areas: building a specific learner corpus and an analysis via a concordance tool or concordancer. The examination is done through the lens of the quantitative method (word counting and organization) and also qualitative analysis from the researchers' academic formation in the field.

This longitudinal study collected a corpus of 487 304 word-tokens from a group of English language learners from three different majors of the ELCL. During the years 2017 and 2018, these students wrote several academic texts at four different levels and courses: Integrated English I, Integrated English II, Composition, and Essay. The two former courses were taught in 2017 and encompass the beginning and high beginning level of the major. The latter courses, taught in 2018, correspond to intermediate and high intermediate levels. These students had a total of 208 hours of in-class instruction during the four courses in the two years or four semesters of the corpus collection. All texts submitted were the last version students handed in to their teachers along the class period and were collected before they were given a grade. Ninety six percent of these students signed a consent form allowing their texts to be used for this study; such permission was compiled at the beginning of each semester and before the compilation of the texts. The academic compositions in the corpus varied in length and language level and consequently were divided in subcorpora as follows:

Integrated English I	17 955 words or word-tokens
Integrated English II	26 649 word-tokens
Composition	108 058 word-tokens
Essay	334 646 word-tokens

No names were recorded because the object of study was specifically the texts and not the authors of those texts in an indirect way. For the analysis, the concordancer program AntConc© supported this study to organize and count the tokens of the corpus and subcorpora, therefore searching through them.[1]

Results Discussion. Lexical Richness

The present study first included an overall evaluation of the lexical units from the entire corpus (487 304 word-tokens). The quality of the vocabulary students used can be easily determined by applying the *Type Token Ratio (TTR)* which is obtained by dividing the total number of unique words (word types) by the total number of words in the corpus (word-tokens). The result obtained ranges from 0.0 to 1.0; the closer the number to 1, the richer and more varied the vocabulary the students use. When the whole corpus is analyzed, the TTR reached is 0.036, which unfortunately is low, as Table 1 illustrates. If this information is analyzed per segments of students, the numbers are even more startling because it would be expected that, as students move forward in their learning process, so should the lexical richness; however, the numbers reveal a different scenario, shown in Table 2. This information can be easily seen in Graph 1 as it visually impacts the figures presented previously.

1 This tool can be downloaded at http://www.laurenceanthony.net/software.html and the creator, Laurence Anthony, has posted a series of video tutorials to use the concordance as well: https://www.youtube.com/user/AntlabJPN/featured

Table 1
Type Token Ratio

Word Tokens	Word Types	TTR
487 308	17 686	0.036

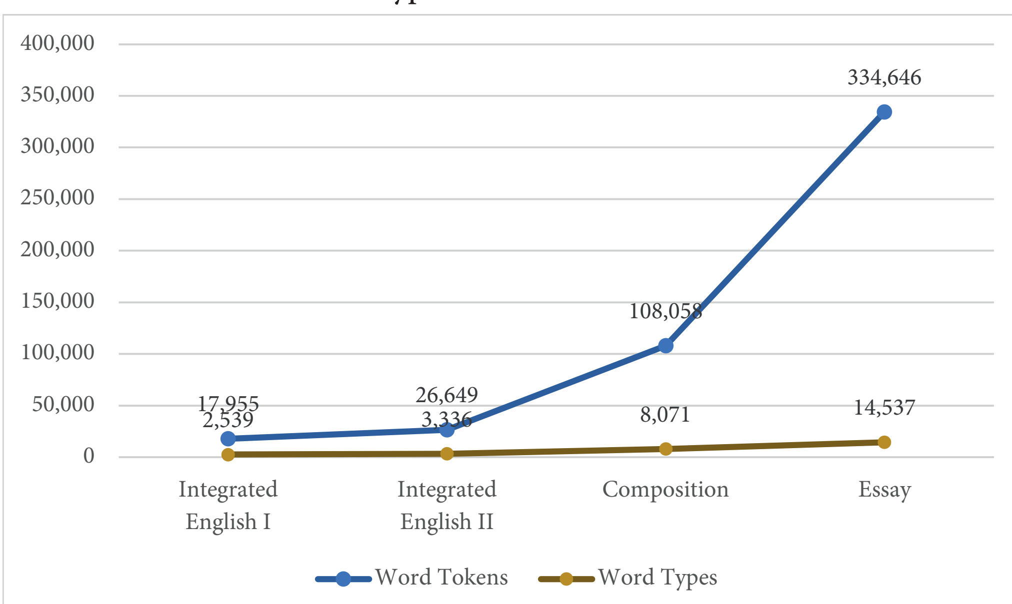
Source: 2017-2019 ELCL Corpus

Table 2
Type Token Ratio by Segment

Segment =	Integrated English I	Integrated English II	Composition	Essay
Word Tokens	17 955	26 649	108 058	334 646
Word Types	2539	3336	8071	14 537
TTR	0.14	0.13	0.075	0.043

Source: 2017-2019 ELCL Corpus

Graph 1
Type Token Ratio



Source: 2017-2019 ELCL Corpus

Considering the graph, it cannot be said that the students are not improving their lexicon. What can be concluded is that the students are using the same words repeatedly. Table 2 clearly shows the increase in the number of word tokens from Integrated English II to Composition—an outcome that is expected as students move from the beginning to the intermediate level. In the Composition course, many more writing tasks were assigned, and participants became sophomore students, which is believed to support the maturity of their writings. Table 3 establishes the percentage of new vocabulary that the students are using from one segment to the other.

The impact in the vocabulary-usage increase from the high beginning level to the intermediate (from Integrated English II to Composition) is positive and crucial to point out, as these texts were improving considerably. However, the intermediate to high level usage (from Composition to Essay) froze and held a quite slightly higher level, yet not the one that it needs to be reached as the students should be mastering advanced writing levels. As Dewi (2017) asserts "Due to the fact, the existence of lexical complexity in students' academic texts sets forth the students' writing proficiency. Therefore, lexical complexity proficiency in writing academic texts such as research articles is undoubtedly required" (161).

These general remarks have set the context in which the next section (adjectives) develops providing that word usage remains low and so do the different parts of speech within the corpus. Nonetheless, displaying the word types of the study will provide a good vision we intend to illustrate, seeing the actual frequency of adjectives supports the concept of simplicity that the study has found in the corpus.

Table 3
New Vocabulary

Segment=	**Integrated English I**	**Integrated English II**	**Composition**	**Essay**
Word types	2539	3336	8071	14 537
New words used	—	797	4.735	6466
%	—	31.30	141.93	80.11

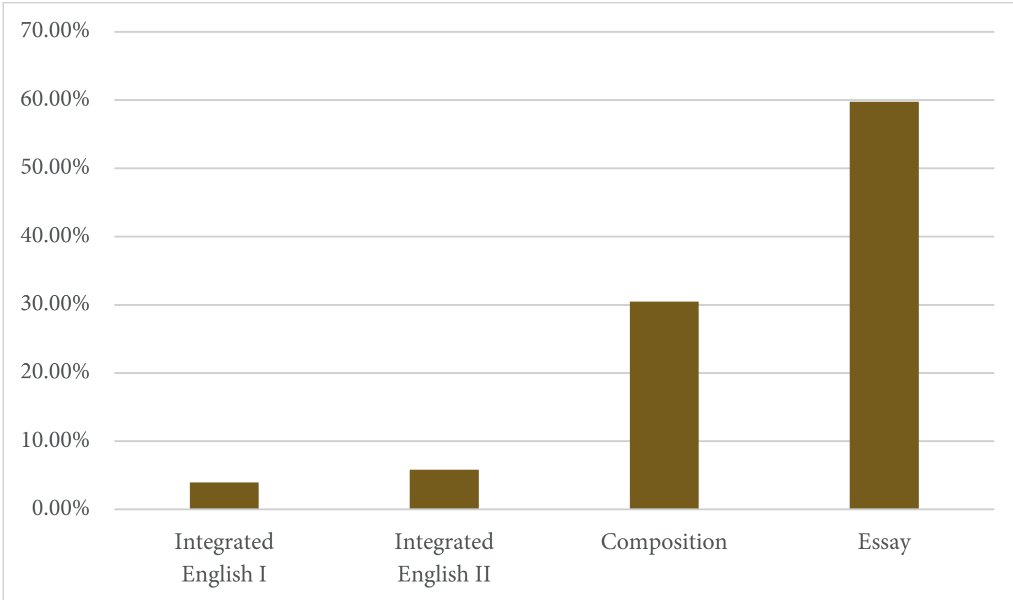
Source: 2017-2019 ELCL Corpus

Adjectives

This section explains the findings related to adjectives used by students starting with Integrated English I, then Integrated English II, Composition, and the Essay course. This part of speech was selected from the rest of categories for the present study as descriptors and classifiers (functions of adjectives) are relevant to the type of texts in the corpus. In the study by Biber, Conrad & Leech (2002: 190) there is clear evidence that adjectives are a common part of speech and add up to the concrete meaning of the nouns in academic prose.

From the 487 308 tokens, 4088 of them were adjectives counted only one time. Graph 2 indicates how students progressed from Integrated English I to Essay, and it shows that there is a gradual trend in the use of adjectives from one level to the other. In Integrated English I, there is a use of adjectives of 3.5 percent. For example, students used very personal descriptive adjectives such as *amazing*, *beautiful*, or *nice* due to the topics that were probably covered in the syllabus or topics the instructors chose to develop. Unlike course I, Integrated English II shows a small increment in adjective usage: 5.85 percent represents the use of adjectives, which means that there is an increment in the development of adjectives usage of 1.90 percent from one course to the other. In this second level, adjectives were still on the personal descriptive level, but more intended to describe situations or pleasures. Since the courses Integrated English I and II are taught in the first year of the three majors and the students are getting to know the language and learning it, the percentages drawn from the data go according to the initial stages of acquiring a second language; nonetheless, there is a shift on the other levels in regards to adjective use progression. The third column of Graph 2 draws an interesting statistical point: 30.45 percent involves the use of adjectives in the Composition course, projecting an increase of 24.60 percent from the previous level. The increase of adjective usage encloses more sophisticated and academic adjectival structures, and there are hyphenated adjectives not used before. The last column on the graphic corresponds to the use of adjectives in the Essay course, where 59.75 percent signifies the adjective usage. Remarkably, there is a rising in adjective usage from Composition to Essay of 29.30 percent. In short, there is a tendency of positive progression in adjective use that goes along with the development of the major curriculum if one only looks at the numbers (repetition) and not the lexical richness.

Graph 2
Adjective Progression per Level



Source: 2017-2019 ELCL Corpus

According to the Graph 2, adjective progression seems to be aligned with the main objective of the English Teaching Bachelor which is, in general terms, to help students master properly the L2. However, there is revealing information from the data related to the number of times adjectives were utilized. It was noticeable that 48 percent of the adjectives were used only one time, followed by 17.32 percent two times, 3.8 percent three times, and, lastly, 1.52 percent four times. The main purpose of underlying these percentages is to perceive if participants broaden their lexicon in terms of semantical sentence descriptions. Whereas there is an obvious growth, this one, in fact, not even surpasses the 50 percent of the adjectives being used in the corpus.

Linked to this issue is the complexity and intricacy of the adjectives used in the texts. In language learning, linguists know that at the beginner level (Integrated English I and II in this particular case), the lexicon is not only simpler, but also generally descriptive having in mind that tasks are mostly based on personal or anecdotical information. Nonetheless, when it comes to see the types of adjectives applied in Composition or Essay courses (high intermediate levels), the overall usage of adjectives continues a tendency which does not fit the academic prose except for a little increase of participial adjectives, hyphenated adjectives, attributive nouns as well as more sophisticated lexicon-adjective related to the themes. This increase represents 18.28 percent of the total adjectives analyzed within the corpus.

Some examples of participial adjectives seen are *addicted person*, *botched work*, *taught student*, *wooden house*, *managing time*, *flowering moment*, for instance. Regarding the use of hyphenated adjectives, the illustrations perceived in the compositions of the study are either related to the themes where one can see adjectives such as *chemotherapy-related treatment*, *cell-based use*, *drunk-driving accident*, *eco-technological procedure*, or fixed expressions like *well-known*, *self-confident*, *high-risk*, *well-paid*, *well-done*, *so-called*, or *potty-mouthed* among others. Aligned with this, attribute nouns working as adjectives found in the corpus followed a similar pattern; some examples are *videogame*, *secondhand*, *nationwide*, *northwest*, *mainstream*, *lifelong*. Last of all, there are the less common adjectives: some words extracted from the corpus are *strenuous*, *slanderous*, *psychoactive*, *outstanding*, *nefarious*, *noxious*, *nitrous*, *mediocre*, or *lucrative*. In short, this increment represents a benefit in the students' writing learning process, but there is still a huge gap in terms of helping students expand their adjectival lexicon increment. For instance, the data suggests that there is a repetition of adjectives from Composition to Essay that students reused in their written works meaning 26.73 percent between the last two levels. This percentage is very significant because the lower it is, the better it indicates that students are hardly mastering the adjectival structural usage of the target language.

Another important point is to see the general frequency of the tokens found in the analysis of this corpus in relation to the adjectives used. In this case, the word-token *many* appears in the corpus 1399 times used as an adjective that indicates plural nouns (large number), as in *many families* or *many people*. In contrast, synonyms to this word, such as *several*, was used 241 times, *numerous* 22, *various* 26, and *countless* 7 times. Aside from that, there are other synonyms that can be taught to expand the students' vocabulary such as *immeasurable*, *innumerable*, *incalculable*, *uncountable*, and many more. Likewise, *many*, *good*, *different*, and *important* give the impression to reveal similar situations. *Good* was used 1292 times while *different* and *important*, 1167 and 879 times respectively. This analysis per frequency is extremely vital because it discloses how the students have not mastered a varied use of adjectives and it sheds light on this important issue while providing enough input to start carrying out activities to help learners advance in the acquisition of the L2.

As opposed to the data presented in the previous paragraph, adjectives of the study that were used in all levels represent 1.52 percent of all the adjectival tokens. To mention the highest frequency adjectives, we have *social* used 727 times, *psychological* used 126 times, *national* 58 times, *professional* 54 times, *historical* 26 times, and *economical* 29 times. A great deal of adjectives used in all levels classified as general ones that can be substituted for much more evocative or eloquent synonyms that would enhance writings. For instance, the adjective *different* was employed by the students 1167 times, representing 28.54 percent of the adjective usage in the corpus. The adjective *important* was used 879 times, representing 21.50 percent while the adjective *big*, 403 times, represents 9.85 percent. The data reveals that adjectives learned in Integrated English I were applied by learners constantly and their utilization incremented in more advanced courses, yet with no great variety or richness. Other basic adjectives regularly observed in the corpus with a minor percentage, but not insignificant, were *young* (224 times), *easy* (204), *difficult* (198), or *happy* (153 times), to pinpoint a few. The see-through information states a noteworthy pattern among these writers of the L2 language which is the recurrence of plain adjectives, providing abundant evidence to do material development or creating teaching material to widen the lexis spectrum in terms of synonyms or alternative words. This finding leads the study to show the Type Token Ratio (TTR) in terms of lexical richness in adjective usage within the corpus and subcorpora, which reached 0.07, a result that, as explained in the first section of this article, shows how low the variety in the use of adjectives is.

In sum, as academic writings or texts are the object of the study, we want to highlight three factors from Biber, Conrad & Leech (2002) who indicate that "in news and academic prose, attributive adjectives are an important device used to add information to noun phrases" (189)—more than predicative adjectives (189), yet adjectives after all. Adjectives are much more common in academic writing than in the news, fiction, and conversation genres, as the *Longman Student Grammar of Spoken and Written English* (*LSWE*) shows a solid corpus of 40 million words (Biber *et al*., 2002: 190) used to compare the corpus of this study at the statistical level. The most common adjectives in academic prose, according to the LSWE corpus, are *long*, *small*, *great*, *high*, *low*, *large*, *new*, *old*, *young*, *good*, *best*, *right*, *important*, *simple*, *special*, *basic*, *common*, *following*, *higher*, *individual*, *lower*, *particular*, *similar*, *specific*, *total*, *various*, *whole*, *different*, *full*, *general*, *major*, *final*, *main*, *single*, *local*, *natural*, *oral*, *physical*, *public*, *sexual*, *political*, *social*, *human*, *international*,

national, and *economic* (Biber *et al*., 2002: 200). This short yet quite illustrative list frames the discussion about the lexical diversity that advanced EFL learners' texts need to display.

Conclusions

This paper illustrates the very first steps that can be carried out with a Corpus Linguistics methodology by language specialists and what they are able to deduce. Among the multiple tasks a language researcher can fulfill with a concordancer program, one of the simplest is to organize the data while the interpretation of the information lies on an analytical human eye. That is, the insights arise from the associated analysis of the quantitative and qualitative methods. A frequency list or word list, called like that in AntConc©, offers all the support that can suffice to begin an analysis and set the starting line.

As language professors, but mainly observers of the learner corpus, we have presented data that illustrates the stages of written language acquisition leaving intuition aside. With reliable facts we demonstrated that with a corpus made from every text collected in the time span selected, the results derived from frequency are conclusive. Also, significant trends or regularities can be observed in the language production of the L2 users. One can call this type of study an Ethnographic one to show how language (whether from native speakers or learners) behaves.

Though the steps for building a corpus were not a section of this paper, it is clearly stated that a good quality design and consistency collecting a corpus provides a decent amount of data for analysis that is reliable and pertinent. Being able to collect almost the entire collection of students' writings draws stronger conclusions undoubtedly.

No corpus is too small for a meaningful study. Though authors vary concerning different corpus size aptness, if teachers quest the habit of surveying their students' writings, for example, they will hold excellent basis of what their pupils are producing and if their teaching practices' outcome is positive and efficient or not: obtaining a snapshot of the *status quo* never fails and can certainly support many other further studies.

Undertaking longitudinal studies combined with Corpus Linguistics is worth the effort as they provide students' real progress concerning language development because it enables the observation of any changes in variables that occur over time. It is the ultimate evaluation of the process, as opposed to testing students at the end of the course with all the variables—human and physical variables that, in the end, may change or distort assessment objectives.

Even though adjectival usage was the object of the study for this paper, choosing to show the level of lexical richness of the entire corpus was significant to provide the context of the overall data in case one wonders if adjectives were the one part of speech that did not comply with the expectations of language development. As can be observed, the larger the subcorpus (Composition and Essay), the lower the TTR—that is, much is being written or produced, yet the quality of word usage is problematic. It is important to add that basic usage of the language in academic prose does not reveal any advance in the acquisition of the language. Students may be expressing their ideas, but their academic texts do not convey a complex usage according to the level. It is imperative to recall that the courses of the study where the texts were collected are college level ones, aimed to train English teachers so that they become masters of the language. Therefore, what are the standards that the majors are pursuing? A look at those standards proves that the ELCL defines clearly in the English majors' syllabi the objectives the students have to reach and lead to the conclusion that the teaching needs reinforcement. If the exit profile of the students is precise, a lack of lexicon variety occurs and more reinforcement in this matter is urgent. With the findings of this study, teachers should be aiming to address the teaching of lexical sophistication and diversity more precisely than the correct morphological and syntactical use of a word—that is, encouraging the acquisition of more words amidst the rest that need learning: using native speakers' corpora to learn and increase vocabulary is an example of a teaching strategy that leads learners to discover words independently.

This first attempt to dig into the students' word choice, specifically, adjective use, led the researchers into the following assumptions. First, it is vital to state that this part of speech was not randomly picked. As experienced teachers, time has demonstrated that nouns and verbs, or the use of names and actions, should be as specific as possible when communicating ideas. Descriptors and classifiers, on the other hand, are an excellent starting point to begin observation because it allows to discover the accuracy of the ideas expressed in writing. It is well-known that adjectives are a regular characteristic in literary texts; however, the compositions in these courses are mainly expository and descriptive texts. When moving to argumentative texts (Essay), images should be bold enough to be convincing and descriptive words ought to be mastered.

That being said, the enormous repetition of basic adjectives along the acquisition of the language leads to conclude that learners could be paying more attention to syntax than lexicon, and this latter situation is not being corrected or given the attention it needs. That is, usage in writing is correct, yet the entire texts have not reached the corresponding English lexicon level as shown in the *LSWE* corpus in Biber, Conrad & Leech (2002). It seems that the students pass composition courses due to syntax, not due to lexical diversity. The following questions should be posed, then: *What are the roles of the teachers? How can they approach this issue and improve lexical acquisition? How hard it is to teach lexical sophistication?* Even if the study collected academic writings, these writings are not technical nor disciplinary to justify the lack of adjective variety. One should read these courses program objectives and analyze the level of lexicon each program requests. The bottom line is all these questions lead to a prompt reaction in the English course syllabi of the ELCL.

The corpus collected is currently available to the academic staff in the English Department of the School of Literature in question. We seek to contribute to the empirical knowledge of Corpus Linguistics and try to collaborate with other teachers to either provide this unique corpus or help them build their own learner corpus for a variety of hypothesis based on a corpus-based or corpus-driven research. Ultimately, sharing this corpus with other foreign colleagues who have their own corpora that meets the same criteria can lead to more conclusions in the region and that is a plan in the horizon we need to accomplish.

Limitations of this research

The present study does not attempt to be exhaustive by any means. It presents the description of frequency and word count of a learner corpus collected in a two-year spam with the same group of individuals. With such corpus, any language researcher is able to dig desired areas of interest such as grammar, word use, lexicography, and the like. This corpus was not an annotated one either, yet the size was manageable enough to identify each part of speech manually with the human eye using the concordancer. Every example of word-types that was used one time has been selected with care, yet a little *cherry picking* influenced the choice. A bit of intuition biased the sort of illustrations in this paper due to our solid 10-year-plus experience as language professors.

References

Ai, Haiyang; Lu, Xiaofei. (2010). "A Web-based System for Automatic Measurement of Lexical Complexity". Paper presented at the 27th Annual Symposium of the Computer Assisted Language Instruction Consortium (CALICO-10). Amherst, MA. June 8-12. http://dx.doi.org/10.13140/RG.2.2.16499.07208

Biber, Douglas; Conrad, Susan; Leech, Geoffrey. (2002). *Longman Student Grammar of Spoken and Written English*. Longman.

Biber, Douglas; Conrad, Susan; Reppen, Randi. (2014). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press.

Dewi, Ratna. (2017). "Lexical Complexity in the Introduction of Undergraduate Students' Research Articles". *Jurnal Pendidikan Bahasa dan Sastra Inggris, 6*(2), 161-172. https://doi.org/10.26618/EXPOSURE.V6I2.1179

Laviosa, Sara. (2002). *Corpus-Based Translation Studies: Theory, Findings, Applications*. Rodopi.

McEnery, Tony; Wilson, Andrew. (2001) *Corpus Linguistics* (2nd Ed.). Edinburgh University Press.

Torruela, Joan; Llisterri, Joaquim. (1999). "Diseño de corpus textuales y orales". En José Manuel Blecua, Gloria Clavería, Carlos Sánchez, Joan Torruela (Eds.), *Filología e Informática. Nuevas tecnologías en los estudios lingüísticos* (pp. 45-77). Universidad Autónoma de Barcelona; Milenio.