Using a linguistic computer software to explore the N-grams in ESL college compositions

Rey John Castro Villanueva
Mariano Marcos State University-Batac, Ilocos Norte, The Philippines
arjaycastrovillanueva25@gmail.com

Abstract

This research made an initial exploration on the English N-grams commonly used by college students in their ESL writing classes. Previous studies found out that there are several types of N-grams; however, this research zeroed in on three- and four-word N-grams only as these are the most researched types and have been used in many related studies. Moreover, the functional classification of each of the identified three- and four-word N-grams was identified through the use of the taxonomy forwarded by Biber et al. (2004). The Antconc 3.2.4w retrieved a total of 31 English N-grams from the academic computer corpus that contains 100,000 running words. The findings reveal that referential N-grams are widespread, whereas both *stance* and *referential* N-grams record a minimal rate of occurrence. Several reasons in respect of the underuse of both *stance* and *discourse* N-grams were discussed; however, all of them require further analysis before any conclusion can be made.

Keywords: College compositions, functional classifications, linguistic computer software, N-grams

1. N-grams in College Compositions

Writing college compositions is one of the tasks given to students in ESL writing classes. One may note that such a classroom task is described as a very challenging activity as it compels learners' awareness and determination to produce an excellent academic writing output. In one of her papers, Zamel (1998) describes academic proses as those that have peculiar characteristics "... because it appears [sic] to require a kind of language with its own vocabulary, norms, sets of conventions, and modes of inquiry ... " (p. 187).

It should be noted that the N-gram is one of the distinctive parts of an academic prose that was initially described in Biber, Johansson, Leech, Conrad, and Finegan's (1999) *Longman Grammar of Spoken and Written English*, a massive scholarly work completely based on the 100 million-word *British National Corpus.* Biber and his colleagues described N-grams as sequences of frequently co-occurring words in particular registers. They are present in written registers, and they are considered as "basic building blocks for constructing… written discourse" (Biber & Conrad, 1999, p. 188).

In addition, Allen (2009) notes that these N-grams have shown that language is "register specific and perform a variety of discourse functions" (p. 367). Consequently, the application of these fixed expressions or the so-called N-grams reveals the competency level

and the success of language learners in that specific register (Haswel, 1999). Moreover, learning how to use the more frequent N-grams of a discipline can contribute to gaining communicative competence in a field of study, and one must note that there are advantages in identifying these N-grams to better help learners acquire the specific rhetorical practices of their communities (Hyland, 2008).

Allen (2009), however, claims that learners rarely have competent use of the N-grams when they begin to learn the ESL academic discourse, even if they have experience of participation in such communities in their native language. Several linguists and scholars found that learners of English from a specific language community or group produce language aspects in their writing production that do not conform with L1 speaker norms (Alternberg & Granger, 2001; Hyland & Milton, 1997). Therefore, language learners tend to have either insufficient or excessive use, or misemployment of certain language features such as N-grams. It should be noted that it is often a failure to use native-like N-grams that identify learners as outsiders, and there is a consensus that second language learners are habitually problematic in acquiring N-grams (Yorio, 1989).

Furthermore, Salazar (2011) opines in her trailblazing Philippine English research on N-grams that the frequent and appropriate use of N-grams is an important constituent of eloquent linguistic production in academic communities.

1.2 Previous Studies on English N-grams

The emergence of different computer corpora made the analysis and investigation of English N-grams in various registers straightforward. Moreover, the development of computer-operated text analysis tools such as the Antconc 3.4.4w made the statistical exploration of N-grams in discourse easier and more convenient because researchers could focus more their attention on what is frequent instead of looking into what is highly noticeable.

The first study that should be given attention is by Biber et al. (1999). This research used a monumental corpus that consists of both American and British English conversation and academic prose. In this particular study, Biber et al. introduced the concept of N-grams. They noted that "both conversation and academic prose use a large stock of different N-grams" (p. 993). Subsequently, such an assertion became a springboard in conducting other related studies on N-grams in different registers.

In 2002, Cortes explored the N-grams used by first-year tertiary students in writing their academic essays. After collecting 311 college compositions and using a computer-operated text analysis tool, she found a total of 93 different N-grams. Further analysis of the identified N-grams revealed that in terms of structure, these N-grams looked like the N-grams found in academic prose; while functionally, these frequently occurring expressions served as temporal or locative markers that created redundancy in freshman college writings.

Using a total of 160 monologic lectures from the British Academic Spoken English (BASE) and Michigan Corpus of Academic Spoken English Corpus (MICASE), Nesi and Basturkmen (2006) focused on the function of N-grams in academic lectures. The results of the analysis showed that N-grams can play a discourse-signaling role in lectures, and it is important to raise students' awareness on the use of N-grams.

In 2006, Cortes conducted another research that focused on a more pedagogical feature of N-grams in the classroom. Cortes explicitly taught N-grams to students in a writing-intensive history class. After investigating the effectiveness of the tasks she prepared

for teaching N-grams by comparing students' writings, she concluded that the students' use of target N-grams was rare and uneven and that having a few lessons that demonstrate some examples of N-grams in professional written production might not necessarily result in students using more N-grams in a more appropriate way.

Using the TOEFL 2000 Spoken and Written Academic Language or the T2K-SWAL corpus, Biber and Barbieri (2007) examined how N-grams are used in both non-academic university registers and core instructional registers. The results of this study showed that the use of N-grams is very common in instructional written course texts such as course syllabi. Therefore, this research hardly confirmed the findings shown by previous studies that N-grams were more common in speech than in writing.

Lastly, in the Philippines, Salazar (2010) made a frequency-driven investigation of the occurrence, including the grammatical and functional categories, of N-grams with verbs in Philippine and British scientific English. The findings of the study revealed that the British computer corpus generated a higher number of verbal N-grams compared to the other one, i.e., the Philippine corpus.

The review of the foregoing literature would reveal that only a few studies on L2 written data have identified the functional classifications of N-grams. Moreover, there seems to be only one study on N-grams conducted by a local researcher (i.e., Salazar, 2010). Although Salazar used other corpora, which included British English and Philippine English, she hardly considered a perspective from second-language learning. Alternatively, she treated and desribed the Philippine English corpus as "highly proficient," on the ground that all the components of her study corpus had been awarded high passes.

Furthermore, most of the foreign researchers who studied N-grams used computer corpora produced by L1 users of English. Therefore, one may note that little is known about the N-grams used by nonnative speakers who use English as a second or a foreign language, particularly Filipinos who often write English texts in various domains. The studies presented earlier reveal helpful information as regards the significance of English N-grams and how they differ functionally in various academic registers and in different conditions. Moreover, they provide opportunities in exploring N-grams in further studies, which was the impetus for the present research.

It should be noted that there are other classifications (e.g., two-, five-, and six-) of N-grams; however, this study zeroed in on three- and four-word N-grams only, for these are the most researched types of N-grams. Salazar (2011) notes that the longer the N-gram, the lower is its frequency; Cortes (2004), on the other hand, argues that three- and four-word N-grams provide researchers with more obvious varieties of discourse functions to investigate.

Since N-grams shape the meanings of written texts, the findings of this research may help students in understanding the various discourse functions of English N-grams in academic registers.

1.3 Research Questions

This research investigated the discourse functions of three- and four-word English N-grams in compositions written by Filipino university students. In particular, this study sought to answer the following questions:

a. What are the English N-grams found in the compositions written by university students?
b. What are the functional classifications of these N-grams?

2. Method

2.1 Study Corpus

This research covered the compositions written by 130 senior college students of five Higher Education Institutions in the Philippines. These institutions include four Catholic schools and a state university. Also, these Philippine schools are all listed by a British company that specializes in education and study abroad as the best universities in Asia, and most of the time, the graduates of these top Philippine universities have higher chance of getting hired, according to the largest online employment company in Southeast Asia. All the students were enrolled in two ESL writing classes, i.e., Business English and Technical English. In choosing the student participants, the following inclusion criteria were considered: (a) must be a senior student in college, (b) must have a good scholastic standing, and (c) must have any of the Philippine Languages (e.g., Filipino) as L1 and English as L2. Thus, all the student informants are homogeneous in terms of their linguistic, educational, and socioeconomic background. They speak a Philippine language (e.g., Filipino) at home. In addition, all of them did not receive any English language instruction in English-speaking countries nor they had been to any English-speaking countries to have any kind of English exposure. They obtained their elementary and high school diplomas in the Philippines. The senior college students who met the above-mentioned criteria were tasked to write 1000-word college compositions on one occasion only during their vacant periods in order to avoid class interruptions. The essay prompt was about *any social issue in the Philippines*. Furthermore, all the papers were argumentative academic essays. It should be noted that in the collection of academic essays, when the desired number of words was not met, the researcher put in additional 'qualified' students to make up 1000 words. Table 1 elaborates some basic information about the corpus.

Table 1
Detailed content of the corpus

Corpus Component	Total Number of Texts	Total Number of Running Words
College Compositions	130	100,000

Table 1 reveals that the corpus has 130 English essays with a total of 100,000 running words. The size of the corpus may be considered relatively small, i.e., under one million words. However, a smaller size does not impede the usefulness of a corpus, as Kyto (2012) argued. Therefore, as what Bowker and Pearson (2002) and Pierini (2009) maintain, small corpora can be reliable and representative, especially when dealing with domains-specific languages. Furthermore, Fuster-Marquez (2014) seems to back up the argument of Kyto (2012), Bowker and Pearson (2002), and Pierini (2009) claiming that "size is not necessarily the most relevant criterion in corpus building" (p. 91)

2.2 N-gram Types, Frequency Threshold, and Occurrence Rates

In their research, Beng and Keong (2015) explained that "the normalized frequency threshold for large written corpora generally ranges from 20 to 40 per million words" (p. 81). However, this research took a sort of conservative approach by setting up the raw cut-off frequency at occurring 20 times per hundred thousand words because of the small size of the corpus.

Moreover, in order to avoid the students' idiosyncratic use, an N-gram (both three- and four-word) must be used in at least three college compositions written by different students.

2.3 The Linguistic Computer Software: Antconc 3.2.4w

The current research utilized the AntConc 3.2.4w, a computer text analysis tool invented by Laurence Anthony in 2007, and the researcher followed different steps in retrieving English N-grams from the academic corpus that composed mainly of college compositions:

a. All the collected college compositions were digitized, and each of them was saved in Plain Text format because the Antconc 3.2.4w cannot process a corpus saved in other document formats.

b. After the digitization process, the English N-grams were generated using the 'N-grams' feature of the linguistic concordancing software.

c. After the operation of Antconc 3.2.4w based on the abovementioned settings, an inventory of three- or four-word English N-grams were retrieved and the minimum frequency cut-off point for the range was computed manually.

2.4 Functional Taxonomies of English N-grams

The current research utilized the classification forwarded by Biber, Conrad, and Cortes (2004). This functional classification has three major categories – 'stance expressions,' 'discourse organizers,' and 'referential expressions.' The first category, 'stance expressions,' consists of phrases or groups of words that show the attitude, judgment, and perspective of the writer in terms of certainty or uncertainty, and proposition or ability. In their study, Biber et al. (2004) note:

> Stance expressions provide a frame for the interpretation of the following proposition, conveying two major kinds of meaning: epistemic and attitude/ modality. Epistemic stance expressions comment on the knowledge status of the information in the following proposition: certain, uncertain, or probable/possible (e.g., *I don't know if*, *I don't think so*). Attitudinal / Modality stance expressions express speaker attitudes towards the actions or events described in the following proposition (e.g. *I want you to*, *I'm not going to*). Stance expressions can be personal or impersonal. Personal stance expressions are overtly attributed to the speaker/writer, as in the examples given above. Impersonal stance expressions express similar

meanings without being attributed directly to the speaker/writer (e.g. *it is possible to*, *can be used to*). (p. 389)

The 'discourse organizers,' on the other hand, help in composing and in structuring the text itself by means of introducing a topic and clarifying or elaborating on the topic. *A little bit about* and *as well as the* are examples of N-grams that can be classified as such. Lastly, 'referential expressions which are used frequently in academic registers refer to a given attribute or condition, or pertain to number, amount, size, or quantity. It should also be noted that expressions, which bring out details about time and place, are also admitted in this classification. Furthermore, multi-functional referential expressions cover the N-grams that convey various referential functions in different contexts. N-grams such as *at the end of* can relate to place (*at the end* of this paragraph) or time (*at the end* of the 18th century).

Table 2
Functional types of N-grams

Classification	Example
1. Stance Expressions	
A. Epistemic Stance	
A.1. Personal	*I don't know if*
A.2. Impersonal	*are more likely to*
B. Attitudinal/Modality Stance	
B.1. Desire	*if you want to*
B.2. Obligation/Directive	
Personal	*you look at the*
Impersonal	*it is necessary to*
B.3. Intention/Prediction	
Personal	*what we are going to*
Impersonal	*is going to be*
B.4. Ability	
Personal	*to be able to*
Impersonal	*it is possible to*
2. Discourse Organizers	
A. Topic Introduction/Focus	*in this chapter we*
B. Topic Elaboration/Clarification	*on the other hand*
3. Referential Expressions	
A. Identification/Focus	*one of the most*
B. Imprecision	*and things like that*
C. Specification of Attributes	
C.1. Quantity Specification	*a lot of people*
C.2. Tangible Framing Attribute	*in the form of*
C.3. Intangible Framing Attribute	*in the case of*

Table 2 continued …

Classification	Example
D. Time/Place/Text Reference	
D.1. Place Reference	*in the United States*
D.2. Time Reference	*at the same time*
D.3. Text Deixis	*as shown in Figure N*
D.4. Multi-functional Reference	*at the end of*

3. Results and Discussion

3.1 Identifying the Target N-grams Present in College Compositions

The computer-operated text analysis tool used in this study retrieved a total of 59 N-gram types (the total number of N-gram tokens is 2,044); however, such a number was trimmed down to 31 because this research adapted the following rules forwarded by Salazar (2011), Chen and Baker (2010), and Biber and Conrad (1999) in the recognition of N-grams:

1. N-grams containing content words that are present in the essay questions/topics should be automatically deleted.

2. N-grams that incorporate proper names that have an indirect or direct connection to the informants or participants should also be excluded from the extracted N-gram list.

3. 'Overlapping' N-grams could inflate the results of quantitative analysis. Thus, overlapping word sequences should be combined into one longer unit so as to guard against inflated results.

4. Other N-grams that seem to be 'meaningless' should also be removed.

Table 3 reveals the inventory of English N-grams generated by the linguistic concordancing software. As presented in the said table, most of them are three-word N-grams as the computer-operated text analysis tool generated only one four-word N-gram, *is one of the*. This type of N-gram, *is one of the*, was used 33 times by the college students and occurred in 13 college compositions. The very low frequency of occurrence of this type of English N-gram may be explained by the complexity of their production. It takes the writer more effort and time to produce a four-word N-gram than three-word N-grams.

Table 3
An inventory of the target English N-grams

Three-word	No. of Occurrence	Four-word	No. of Occurrence
one of the	80	is one of the	33
a lot of	78		
because of the	49		
the fact that	42		
of the country	39		
of the Philippines	39		
that it is	37		
the Philippines is	37		
to have a	36		
is one of	35		
because it is	34		
be able to	32		
in the country	31		
it is a	30		
the use of	29		
of the people	28		
that they are	26		
it is not	25		
should not be	25		
in order to	23		
that there are	23		
according to the	22		
in the world	22		
it comes to	22		
they do not	22		
due to the	21		
of the most	21		
there is no	21		
in our country	20		
the people who	20		

As can be seen in Table 3, a majority of the English N-grams is three-word type of word sequences. Five of them, i.e., *one of the*, *a lot of*, *because of the*, *the fact that*, and *of the country*, appeared more than 38 times in the computer corpus. *One of the* was used 80 times by the college students and appeared in 25 different college compositions. While the three-word N-gram, *a lot of*, occurred 78 times in nine compositions, *because of the*, on the other hand, was used 49 times in 20 academic texts. The other three-word N-grams commonly used by the student respondents are *the fact that* and *of the country,* which occurred 42 and 39 times in 20 and 14 English college compositions, respectively.

In addition, it is interesting to note that in spite of the fact that the college compositions constituting the entire academic corpus used in this research were produced by second language users of English, eight (i.e., *one of the*, *the fact that*, *the use of*, *in order to*, *according to the*, *due to the*, *there is a*, and *is one of the*) of the 31 English N-grams retrieved by the Antconc 3.2.4w were also retrieved by previous researchers such as Hyland (2008) who used academic texts written by L1 users of English as his corpus. This means that even if this research used a relatively small size of corpus, one may note that L2 speakers of English also use some of the N-grams that are often present in native English writing productions.

Hyland (2008) identified a total of 50 three- and four-word N-grams in the 3.5 million word academic corpus. The most frequently used three-word N-grams was *in order to*, and such was used 1,629 times in the corpus. For the four-word N-grams, *on the other hand*, which appeared 726 times in the corpus, is viewed as the one that is oftentimes used by native speakers of English.

Further, Table 3 shows that the most frequent N-grams found were *one of the*, *a lot of, because of the*, *the fact that*, and *is one of the*. All these N-grams keyed out as infrequent in previous studies. While Hyland (2008) described *in order to* and *on the other hand* as the most common N-grams in academic corpus, Biber et al. (1999), on the other hand, considered *in the case of* and *on the other hand* as the ones that are oftentimes used in the *Longman Grammar of Spoken and Written English*. The seemingly obvious reason why all these N-grams were not observed as frequent in the corpus employed in this study is because of the conception that one of the attributes of N-grams is the naturalness of language production (Rafiee, Travakoli, & Amirian, 2011). Consequently, one may describe the written texts produced by native users of English as more 'bundleized' than the English writing production of L2 users of English. The Filipino students' persistent use of some of the N-grams (e.g., *one of the*, *the fact that*, *the use of*, *in order to*, *according to the*, *due to the*, *there is a*, *is one of the*) may be due to the fact that they have already been exposed to such N-grams several times in their prior readings of various kinds of English literature (Biber, et al., 1999; Biber & Barbieri, 2007). In other words, L2 writers tend to produce something analogous to L1 writers' production they were previously exposed to. Therefore, they tried to utilize N-grams that were used by L1 writers, but sometimes, they may overuse them to show that they are proficient enough to be deemed as writers in a particular discipline.

3.2 Discourse Functions of the Target N-grams in College Compositions

Table 4 reveals the results of the functional analysis. The English N-grams generated by the computer-operated text analysis tool, Antconc 3.2.4w, were categorized based on the descriptions provided by Biber et al. (2004). Explanations for each of the major categories as well as sample sentences to show how the N-grams that fall under these major categories are used are also presented. However, it should be noted that some of the sentences are ungrammatical because the corpus used in this study consists of 'unedited' college compositions.

Table 4
Functional classifications of the target English N-grams

Classification	Lexical Bundle
1. Stance Expressions	
A. Epistemic Stance	
A.1. Personal	
A.2. Impersonal	*the fact that*
B . Attitudinal/Modality Stance	
B.1. Desire	
B.2. Obligation/Directive	
Personal	
Impersonal	*that it is, should not be, it comes to*
B.3. Intention/Prediction	
Personal	
Impersonal	*to have a, they do not*
B.4. Ability	
Personal	*be able to, in order to*
Impersonal	
2. Discourse Organizers	
A. Topic Introduction/Focus	*according to the*
B. Topic Elaboration/Clarification	
3. Referential Expressions	
A. Identification/Focus	*one of the* (other variation of: *is one of*, *is one of the*, *of the most*) *that there are*, *there is no, because of the, because it is, it is a, due to the, that they are*, *it is not*
B. Imprecision	
C. Specification of Attributes	
C.1. Quantity Specification	*a lot of*
C.2. Tangible Framing Attribute	
C.3. Intangible Framing Attribute	*the use of*
D. Time/Place/Text Reference	
D.1. Place Reference	*of the country*, *of the Philippines*, *in the country*, *in the world*, *in our country, the Philippine is*
D.2. Time Reference	
D.3. Text Deixis	
D.4. Multi-functional Ref.	*the people who*, *of the people*

3.2.1 Stance Expressions

Biber (2006) defines 'stance expressions' as those that convey feelings, attitudes, perspectives, certainties, uncertainties, and the like. As Table 4 shows, this functional classification consists of two subcategories—the 'epistemic stance' N-gram and the 'attitudinal/modality stance' N-gram.

As what Biber et al. (2004) explain, the first subcategory, 'epistemic stance,' refers to an expression that provides information about certainty (impersonal) and uncertainty (personal). The Antconc 3.2.4w detected only one 'impersonal epistemic stance' N-gram, *the fact that*, and its use is shown in the sentence below:

> You just have to accept ***the fact that*** they should live.

Next to 'epistemic stance' N-gram is the 'attitudinal/ modality' stance. This particular subgroup of the stance expressions is further divided into four subcategories—'desire,' 'obligation/directive,' 'intention/prediction,' and 'ability.' Biber et al. (2004) note that the English N-grams that fall under these subcategories convey personal attitudes. Each of the sentences below contains N-grams that fall into this 'attitudinal/modality stance' category:

(1) Obligation/Directive – Impersonal
> They believe ***that it is*** God's gift.

(2) Intention/Prediction – Impersonal
> Many Filipinos are suffering which ***they do not*** deserve.

(3) Ability – Personal
> One of the main things that seems to make Ebola viruses especially deadly is that they seem to ***be able to*** evade much of the human immune system.

3.2.2 Discourse Organizers

In their study, Biber et al. (2004) explicated that the functions of the English N-grams categorized as 'discourse organizers' are to introduce and to elaborate/clarify a topic. The computer software, Antconc 3.2.4w, generated only one N-gram (i.e., *according to the*), and it was classified as a 'topic introduction' N-gram. The following sentences show how *according to the* is used in introducing a topic.

> ***According to the*** bible, God worked for six days to create everything we see including us, human beings—from the plants to the smallest organisms we've known which ever existed.

3.2.3 Referential Expressions

This particular major functional classification of N-grams consists of four subcategories: 'identification/focus,' 'imprecision,' 'specification of attributes,' and 'time/place/text

reference.' Therefore, the N-grams in this category "generally identify an entity or single out some particular attribute of an entity as especially important" (Biber et al. 2004, p. 393). Moreover, these subcategories are further divided into subgroups, and as what Table 4 reveals, a majority of the N-grams retrieved by the Antconc 3.2.4w for the 'referential N-gram' is sorted out as 'identification focus' (*one of the*, *is one of*, *is one of the, that there are*, *there is a, of the most, because of the, because it is, it is a, due to the, that they are*, *it is not*); 'specification of attributes,' especially its subgroups 'quantity specification' (*a lot of*, *the use of*) and 'intangible framing attribute' (*the use of*); and 'time/place/text reference,' particularly its more narrow categories, 'place reference' (*of the country*, *of the Philippines*, *in the country*, *in the world*, *in our country, the Philippine is*) and 'multi-functional reference' (*the people who, of the people*).

In addition, the following sentences explicate how the abovementioned 'referential N-grams' are used by the student participants in writing their college compositions:

(1) Identification/Focus
Also, they have never failed to constantly remind us ***that there are*** some we are not allowed to.

(2) Specification of Attributes – Quantity Specification
There are ***a lot of*** news around the world and we hear the news, read newspapers, and watch news everyday life.

(3) Specification of Attributes – Intangible Framing Attribute
Thousands upon thousands of Filipino people are benefiting from this rapid innovation that even our everyday undertakings and endeavors are reflected upon ***the use of*** technology and social media.

(4) Time/Place/Text Reference – Place Reference
The social issues ***of the Philippines*** nowadays is how calamity budget occur, the people of some regions are experiencing neglected.

(5) Time/Place/Text Reference – Multi-functional Reference
Most of the political leaders use the money ***of the people*** to cater their own personal needs.

The results presented above disclose that a majority of the English N-grams retrieved by the Antconc 3.2.4w from the academic corpus used in this research conforms well to the functional classifications put forth by Biber et al. (2004).

Moreover, in accordance with the results of the functional analysis shown above, one may note the too little number of English N-grams categorized as *stance expressions* and *discourse organizers*. It seems inconclusive, however, to assert that these insufficient pieces of evidence for these two discourse-function classifications in students' use of N-grams put forward the lack of any knowledge of *stance expressions* and *discourse organizers* or the inability to put the knowledge or a part thereof into good use (e.g., in their academic writing).

A number of reasons could have played a part in elucidating the insufficient use of both *stance expressions* and *discourse organizers*. One reason is that the student participants in the current research might not have felt confident enough to take a stance or that they might not have seen themselves fit to utilize discourse organizers in the writing task they were asked to perform. Another important reason for the underuse of *stance expressions* and *discourse organizers* in the computer corpus used in this study may be attributed to the size of the corpus (Tomasello & Stahl, 2004). One may need a rather large corpus (i.e., more than 500,000 words) for the *stance expressions* or the *discourse organizers* to crop up in the ESL writing production of the senior tertiary students. As per the review of previous studies, the present study is the first to examine the discourse functions of N-grams in L2 compositions of Filipino tertiary students; thus, it would be immature to completely explicate the underuse of both *stance expressions* and *discourse organizers* in L2 writing production. Therefore, this writing deficiency needs additional exploration.

Moreover, the frequent use of *referential expressions*, especially *identification/focus* type of English N-grams, may be attributed to the fact that this particular group of N-grams might have been embedded and could be called upon when it is necessary. Another elucidation is that the writing activity calls for the utilization of *referential expressions*. For instance, the student participants were asked to write about *any social issue in the Philippines*. A majority of the college students framed their writing outputs in such a way that *referential expressions* (e.g., *one of the*, *of the most*) were necessary for the task completion.

4. Conclusion

This study aimed to have an initial exploration on the English N-grams that commonly appear in the college compositions of students in Philippine Higher Education Institutions. Although the review of previous studies revealed that there are numerous types of N-grams, the current research focused on three- and four-word N-grams only as these are the most researched types and have been used in several related studies. In addition, the functional classification of each of the target N-grams was identified using the taxonomy put forth by Biber et al. (2004). This particular taxonomy was chosen because its application is quite broad, and it can be used to analyze discourse functions realized in any discourse. A majority of the retrieved English N-grams did fit neatly into the three major categories: *stance expressions*, *discourse organizers*, and *referential expressions*.

Moreover, the results of the investigation reveal that the tertiary ESL learners heavily relied on referential English N-grams in their writing production. Nevertheless, as far as the research findings from the corpus exploration are concerned, only a few student participants use *stance expressions* and *discourse organizers*, the rate of occurrence being so little made these two discourse function categories nearly unnoticed in the investigation. Such an argument may raise an interesting question of whether or not the targeted student participants possess the knowledge of both *stance* and *discourse* N-grams, the point of which can be further investigated in future studies. In order to further explicate the whys and wherefores of the underuse of both *stance* and *discourse* N-grams in L2 speakers' writing outputs, the current research calls for further exploration on the same line.

References

Allen, D. (2009). Lexical bundles in learner writing: An analysis of formulaic language in the ALESS learner corpus. *Komaba Journal of English Education, 1*, 105-127.

Altenberg, B., & Granger, S. (2001). *Lexis in contrast: Corpus-based approaches*. Amsterdam, Philadelphia: John Benjamins Publishing Company.

Anthony, L. (2007). Antconc 3.2.1w: Freeware corpus analysis toolkit. [on-line]. Retrieved from http://www.antlab.sci.waseda.ac.jp/

Beng, C., & Keong, Y. (2015). Functional types of lexical bundles in reading texts of Malaysian University English Test: A corpus study. *Journal of Language Studies, 15*(1), 77-90.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers.* Amsterdam, Philadelphia: John Benjamins Publishing Company.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes, 26,* 263-286.

Biber, D., & Conrad, S. (1999). Lexical bundles in conversations and academic prose. In H. Hasselgard, & S. Oksefjell (Eds.), *Out of corpora: Studies in honor of Stig Johansson* (pp. 181-190). Amsterdam: Rodopi.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at ...: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25,* 371-405.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *The Longman grammar of spoken and written English*. London: Longman.

Bowker, L., & Pearson, J. (2002). *Working with specialized language: A practical guide to using corpora.* London: Routledge.

Chen, Y., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning and Technology, 14*(2), 30-49.

Cortes, V. (2002). Lexical bundles in freshman composition. In R. Reppen, S. M. Fitzmaurice, & D. Biber (Eds.), *Using corpora to explore linguistic variation* (pp. 131-145). Amsterdam: John Benjamins Publishing Company.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes, 23,* 397-423.

Cortes, V. (2006). Teaching lexical bundles in the disciplines: An example from a writing intensive history class. *Linguistics and Education, 17,* 391-406.

Fuster-Marquez, M. (2014). Lexical bundles and phrase frames in the language of hotel websites. *English Text Construction, 7*(1), 84-121.

Haswel, R. (1999). *Gaining ground in college writing: Tales of development and interpretation.* Dallas: Southern Methodist University Press.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27,* 4-21.

Hyland, K., & Milton, J. (1997). Qualification and certainty in L1 and L2 students writing. *Journal of Second Language Writing, 6*(2), 183-205.

Kyto, M. (2012). Introduction. In M. Kyto, (Ed.), *English corpus linguistics: Crossing paths (Language and Computers – Studies in practical linguistics)* (pp. 1-6). Amsterdam, Netherlands: Rodopi.

Nesi, H., & Basturkmen, H. (2006). Lexical bundles and discourse signaling in academic lectures. *International Journal of Corpus Linguistics, 11,* 283-304.

Pierini, P. (2009). Adjectives in tourism English on the web: A corpus-based study. *Circulo de Linguistica Aplicada a la Comunicacion, 40,* 93-116.

Rafiee, M., Travakoli, M., & Amirian, Z. (2011). *Structural analysis of lexical bundles across two types of English newspapers edited by native and non-native speakers*. Retrieved from http://www.mjal.org/removedprofiles/2013/11.Structural%20Analysis%20of%20Lexical%20Bundles%20Across%20Two%20Types%20of%20English%20News%20Papers%20Edited%20by%20Native%20and%20Non-nat.pdf

Salazar, D.J.L. (2010). Lexical bundles in Philippine and British English. *Philippine Journal of Linguistics, 41,* 94-109.

Salazar, D.J.L. (2011). *Lexical bundles in scientific English: A corpus-based study of native and non-native writing* (Doctoral dissertation). Universitat de Barcelona. Retrieved from http://www.tdx.cat/bitstream/handle/10803/52083/DJLS_DISSERTATION.pdf

Tomasello, M., & Stahl, D. (2004). Sampling children's spontaneous speech: How much is enough? *Journal of Child Language*, *31*, 101-121.

Yorio, C. (1989). Idiomaticity as an indicator of second language proficiency. In K. Hyltenstam, & K. Obler (Eds.), *Bilingualism across the lifespan* (pp. 55-72). Cambridge: Cambridge University Press.

Zamel, V. (1998). Questioning academic discourse. In V. Zamel, & R. Spack (Eds.), *Negotiating academic literacies: Teaching and learning across languages and cultures* (pp. 187-197). Mahwah, NJ: Erlbaum.