A Computational Approach to Assessing Nominalizations in Academic Writing

Yanisa Haley Scherber

Fall 2019

Abstract

Previous research suggests that scientific writing contains a higher frequency of nominalization than other fields. It is the intent of this paper to investigate this topic further, and explore the use of computational methods to assess nominalizations in academic writing. In this study, 1.8 million tokens were extracted from journal articles across seven academic fields to determine the frequency of nominalization by each field. The results indicate that the fields with the highest frequency of nominalization were Business/Management, Psychology, and Social Sciences and History, and the fields with the lowest frequency of nominalization were Biological and Biomedical Sciences and Visual and Performing Arts, which partially opposes the previous research. The words with the highest frequencies of nominalizations were also domain-specific. In the discussion of the findings, this paper suggests further study on this topic, and provides several recommended approaches to improving the computational method used.

1. Introduction

1.1 Background

Nominalization in language refers to a type of process in which a noun is derived from another word class (Eggins, 2004; Martin, 2008). This formation is particularly common in academic writing for its ability to convey information efficiently; however, high frequency of nominalization also increases grammatical density and complexity, which contributes to academic writing being perceived as more difficult to read (Halliday, 1993a). The comparison of a nominalization to its verb form can be observed in the following two sentences:

(a) *The evaporation of water occurs in hot weather.*

(b) *Water evaporates in hot weather.*

In the above, sentence (a) represents a nominalized version of the verb *evaporate (evaporation)*, and sentence (b) represents a non-nominalized version *(evaporates)*.

Following the grammatical description of systemic functional linguistics (Halliday, 1985; further extended in Halliday and Matthiessen, 2004 and Halliday and Matthiessen, 2014), language allows for simultaneous meanings to be expressed in clause structures, and individuals use language in a way which represents the choices they make within these structures. The concept of "grammatical metaphor" was created by Halliday (1993b) to categorize the substitution of one grammatical class for another, such as the process of nominalization, in which a verb, adjective, or adverb is metaphorically realized as a noun. Through the use of grammatical metaphor, individuals are able to adjust their language through grammar, which allows for nuanced variations in meaning that are difficult to achieve via manipulation of lexical items.

For instance, take the example below:

(a) Jacob quickly analyzed the findings yesterday.

(b) Jacob's quick analysis of the findings was done yesterday.

To express grammatical metaphor, nominalization is used to transform *analyzed* (verb) to *analysis* (noun). It is through this nominalization that a nuanced differentiation in meaning can be conveyed. In sentence (a), the grammatical subject of the sentence is *Jacob*, also the actor/agent whom is performing the action, but in sentence (b), the grammatical subject becomes *Jacob's quick analysis*. From sentence (a) to (b), the grammatical subject is switched from *Jacob* to *Jacob's quick analysis,* which communicates a slight alteration in meaning, as the mental "focus" of the sentence switches from *Jacob* to the *analysis*. This type of nominalized construction frequently results in more lexically dense writing; in other words, there are more lexical items included in each clause (Halliday and Matthiessen, 2004).

1.2 Recent Work

Little computational research has been conducted on the use of nominalizations in writing, and the majority of the research which does exist investigates the number of nominalizations by searching for particular word endings associated with nominalizations (e.g. Biber, 1986; Biber et al. 1998; Biber et al. 1999; To et al., 2016; To and Mahboob, 2018). A list of these common endings is included as Table 1 (Thomson and Droga, 2012):

Table 1: Nominal Endings for Verbs and Adjectives (Thomson and Droga, 2012)

Nominal Endings for Verbs	Nominal Endings for Adjectives
ion: cohesion, coercion *ment:* treatment, resentment *ation:* animation, sterilization *ing:* thinking, blocking *ance:* assistance*,* avoidance	*ity:* authority, equality *ery:* bribery, debauchery *ance:* abundance*,* balance *ness:* forgiveness, witness *th:* growth, worth *gy:* apology, strategy

While this approach certainly captures many nominalizations, it is limited in its ability. For example, the ending *-ity* will capture many nominalizations, such as *complexity, elasticity,* and *authenticity*; however, it also captures words which are not nominalizations, such as *velocity, fruity,* and *city*. Additionally, the ending *-th* captures the nominalizations *growth, worth,* and *birth*, but it also erroneously captures *tooth, sloth,* and *cloth*. This approach also makes it difficult to capture nominalizations which do not fit into these common endings, such as *change* in *Her change of address*, or *loss* in *Her loss of a friend*.

In addition to the aforementioned work, there have been few studies which use computational approaches (e.g. Lapata, 2002; Liu et al., 2017); however, these studies do not focus on scientific or academic writing, and, given the small amount of research which exists on this topic at all, there is still much research needed to develop a larger body of knowledge around computational approaches to assessing nominalizations.

1.3 Purpose of the Study, Research Questions, and Hypotheses

The presence of nominalization in language has been studied from a theoretical standpoint (e.g. Lees, 1960; Chomsky, 1970); however, the research from a computational standpoint is limited. This exploratory study attempts to investigate this gap in research and provide information on how nominalization-use differs amongst academic writing within varying fields of academia. Seven academic fields were investigated, and they were chosen based on the popularity of college majors, according to the United States' National Center for Educational Statistics data from the 2015-2016 academic year. The fields are (in order of descending popularity): Business/Management, Health Professions and Related Programs, Social Sciences and History, Psychology, Biological and Biomedical Sciences, Engineering, and Visual and Performing Arts.

The research questions investigated in this paper are as follows:

(1) Which subjects have the highest and lowest frequencies of nominalizations?

(2) What words are most frequently nominalized in each field?

According to Halliday (2004), high use of grammatical metaphor, which includes nominalization, is characteristic of scientific writing, so it is hypothesized that the academic fields with the lowest frequency of nominalizations will be Business/Management and Visual and Performing Arts, as those are the only fields in this study which are not approved research areas by the National Science Foundation (NSF, 2019). Additionally, it is hypothesized that Social Sciences and History will also have a lower frequency, since History is not a research area approved by the NSF. It is also hypothesized that the words with the highest frequency of nominalization will be domain-specific words within each academic field, since these words appear most frequently. These domain-specific words will not be parsed out of the raw data in order to determine if a correlation exists.

2. Methods

2.1 Data Collection

Data for this study was collected by extracting text from academic journals within the aforementioned subfields. In this study, tokens, rather than words, will be referenced. For the purpose of this study, these two concepts can be thought of similarly; however, there is some slight difference between the two. The most notable difference is that contractions (e.g. *can't* in *You can't have that.)* and words containing the genitive *–'s* (e.g. *David's* in *David's car is blue.*)

were separated by the apostrophe and counted as two individual tokens, rather than one token (word).

Around 1.8 million tokens were assessed, and this total was separated amongst the seven fields. Each field contained between 200,000 and 320,000 tokens. After collection, the raw data was parsed into tokens through spaCy (Honnibal and Montani, n.d.), which is trained on the OntoNotes 5 corpus (Weischedel et al., 2013). Symbols, numbers, punctuation, extended white spaces, and unknown characters were not counted as tokens. Journal articles from each of the seven academic fields were sampled, and all sampled texts were published in 2019. Within the academic fields of Social Sciences and History, and Visual and Performing Arts, a sample of subfields was identified, given the large variety of subjects within each field. For Social Sciences and History, Economics, History, and Linguistics were selected as sampled subfields. For Visual and Performing Arts, Music and Theatre were selected as sampled subfields. Text samples were parsed in their respective academic fields and subfields, and an aggregate of all text samples was not created or tested.

2.2 Nominalization Search

To search for nominalizations within the text, a tool was built to perform the following procedure:

1. Tokens were part-of-speech (POS) tagged using spaCy. SpaCy's POS tagger has a proven accuracy of around 97% (Honnibal and Montani, n.d.), which is at or above the expected reliability of human POS-tagging (Manning, 2011).

2. Identified nouns were extracted and compared against the list of nominalizations in NOMLEX-PLUS, which is a large lexical database of over 7,000 nominalizations created by New York University (Macleod et al., 1998).

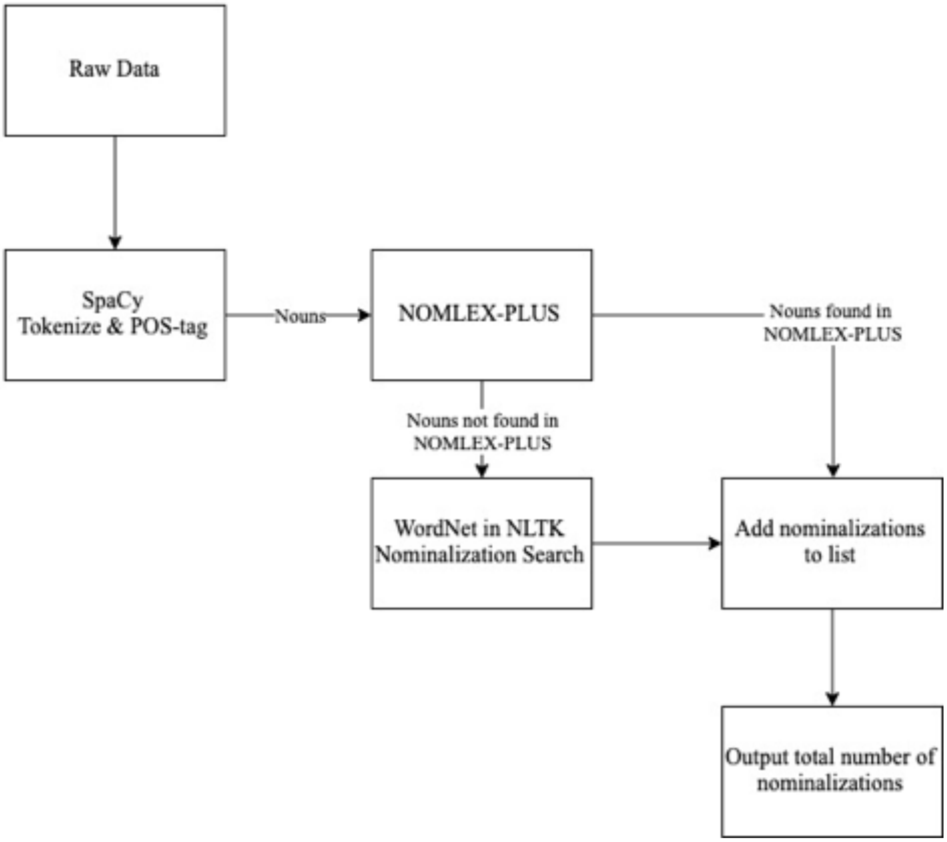
3. Any nouns which were identified by the POS tagger, but not included in NOMLEX-PLUS, were parsed through WordNet (Princeton University, 2010), using NLTK (Bird et al., 2009). First, sets of synonyms (synsets) were retrieved for each noun, lemmas were retrieved for these synsets, and derivationally related forms were retrieved for each lemma in the synset. The original words were considered nominalizations if they contained any derivationally related forms which were verbs, adjectives, or adverbs, began with the same first three letters as the original word, and were not longer in length than the original word. The length comparison component was included to avoid erroneously counting nouns as nominalizations for certain adjectival forms which share the first three letters, such as *alcohol* as a nominalization for *alcoholic* in the sentence *We drink alcohol.*

4. The nominalizations found in NOMLEX-PLUS and WordNet were added together to determine the total number of nominalizations in the text sample.

5. The frequency of nominalized tokens in comparison to the total number of tokens was calculated.

Figure 1 shows a diagram of this procedure.

Figure 1: Data Pipeline


Raw Data
SpaCy Tokenize & POS-tag
Nouns
NOMLEX-PLUS
Nouns found in NOMLEX-PLUS
Nouns not found in NOMLEX-PLUS
WordNet in NLTK Nominalization Search
Add nominalizations to list
Output total number of nominalizations

3. Results

3.1 Frequency of Nominalizations by Academic Field

Table 2 provides a summary of the frequency of nominalizations in each text sample, in descending order from the highest frequency.

Table 2: Academic Fields and Corresponding Nominalization Frequencies

Academic Fields	Total Number of Tokens	Number of Nominalizations	Percentage of Nominalizations
Business/Management	279,875	52,590	18.79%
Psychology	240,567	43,712	18.17%
Social Sciences and History	313,978	50,098	15.96%
Linguistics	*79,604*	*14,967*	*18.80%*
Economics	*174,632*	*27,609*	*15.81%*
History	*59,742*	*7,522*	*12.59%*
Engineering	253,167	39,586	15.64%
Health Professions and Related Programs	253,517	36,791	14.51%
Biological and Biomedical Sciences	242,202	30,543	12.61%
Visual and Performing Arts	206,967	25,769	12.45%
Music	*129,274*	*16,186*	*12.52%*
Theatre	*77,693*	*9,583*	*12.33%*

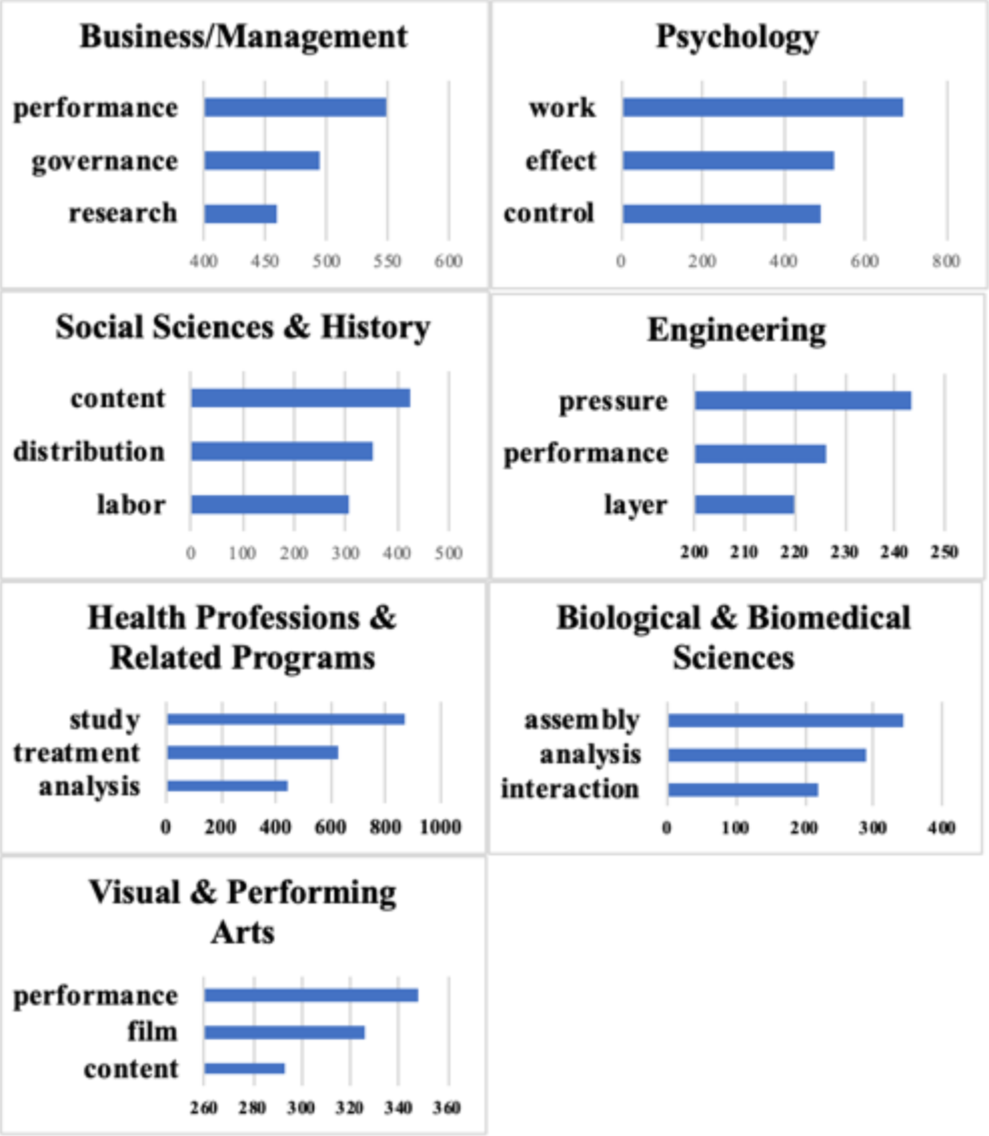
Fields are ranked in order of popularity out of all bachelor's degrees conferred by postsecondary institutions in 2015-2016 (National Center for Education Statistics)

As can be seen, Business/Management, Psychology, and Social Sciences and History have the highest frequency of nominalization, which mostly conflicted with the hypothesis that Business/Management, Visual and Performing Arts, and Social Sciences and History would contain the lowest frequency of nominalization. In fact, Business/Management has the highest frequency of nominalization across all sampled fields, and Biological and Biomedical Sciences, which is an approved research area according to the NSF (2019), is only .16 percentage points above Visual and Performing Arts, which has the lowest frequency of nominalization (as hypothesized).

3.2 Most Frequent Nominalized Words

Table 3 shows bar graphs of the top three nominalizations in each academic field.

Table 3: Top Three Nominalized Words by Academic Field



As can be seen in Table 3, most of the words are domain-specific, and many of the other words are related to research. This is congruent with the original hypothesis, that the most frequent words would be domain-specific.

4. Discussion

This exploratory study provides evidence that further study on this topic is necessary. The results from the ranking of academic fields in frequency of nominalizations was surprising, and it was unexpected that Business/Management would contain the highest frequency of nominalization, while Biological and Biomedical Sciences was near the bottom of the list. This warrants further investigation, as it not only does not match the hypothesis of this study, but it also shows diversion from Halliday (2004), who stated that grammatical metaphor, which includes nominalization, is characteristic of scientific writing.

While the findings were partially incongruent with Halliday (2004), it is important to note that increased grammatical metaphor only represents one aspect of Halliday’s claims on scientific writing, and this paper did not attempt to address the other aspects (e.g. lexical density). Additionally, the NSF is only one organization, and it could be very reasonably argued that Business/Management journals should also be considered scientific writing. Following this argument, the findings from this study do more closely match Halliday (2004); however, the relatively low frequency of nominalization in Biological and Biomedical Sciences must still be considered.

There were also some limitations in this study. For instance, the results from the second portion of this study, which aimed to identify the most frequent nominalizations in each field, indicates that this method of searching for nominalizations requires revision for additional sophistication. Many of the words captured in this nominalization search were domain-specific, and while some of these domain-specific words were true nominalizations (e.g. *performance* in *It was a result of the financial performance.*), some of the nominalizations captured are debatable, since they are widely-accepted terms (nouns) within the field (e.g. *control* in *Group A was used as the control.*).

The automatic POS-tagger and aforementioned process of identifying nominalizations which were not in NOMLEX-PLUS also showed some limitations. For example, the POS-tagger erroneously tags *rich* in *The rich get richer* as an adjectival, rather than a noun, since *rich* is generally used as an adjective/adjectival. This type of error would result in some words not being included in the nominalization search. Additionally, in the process of identifying nouns which were not in NOMLEX-PLUS, the criteria for identifying nominalizations was that the derivationally related form for the original word (i.e. possible nominalization) must have the same first three letters as the original word, and must not be longer in length than the original word. The first check was intended to filter for words containing the same root (e.g. so *recognition* would correctly be labeled as a nominalization for *recognize*, and would not be labeled as a nominalization of *acknowledge*), and the second check was intended to filter out adjectivals from being erroneously categorized as nominalizations (e.g. so *class* would not be considered a nominalization of *classic*). This second check did, justifiably, block many adjectivals from being labeled as nominalizations; however, it did not filter out all erroneous categorizations. An example can be found in the sentence *The drinking of water*, which correctly identifies *drinking* as a nominalization for its derivationally related form of *drink* (verb), but it erroneously identifies *water* as a nominalization for its derivationally related form of *water* (verb).

The computational approach to assessing nominalizations used in this study does appear to overgeneralize the number of nominalizations found in text; however, it can be argued that the current approach of looking for particular suffixes associated with nominalizations undergeneralizes. In the future, a study comparing the accuracy of the two methods should be conducted.

Additionally, in future studies using this computational approach, some additional considerations should be made. First, the addition of domain-specific filtering should be considered. This could involve parsing for the most common words in the raw data of each academic field, looking at their common usages, and determining if any should be omitted from the study. Second, there should be an established approach for counting common collocations. For example, in the approach used in this study, the common collocation *control variable* was counted as two nouns, and, subsequently, two nominalizations. This approach is inaccurate, since *control variable* should either be looked at as one noun/common collocation (i.e. a common "chunk" of language), or *control* should be considered an adjectival for the noun *variable*.

5. Conclusion

This study aimed to explore the use of computational methods in assessing nominalizations in academic writing. In this study, seven academic fields were assessed to determine the frequency of nominalization use overall, and what words were being nominalized. It was found that scientific writing may not contain a higher frequency of grammatical metaphor (e.g. nominalization), as was previously thought. Business/Management, Psychology, and Social Sciences and History contained the highest frequency of nominalizations, and Biological and Biomedical Sciences and Visual and Performing Arts contained the lowest frequency of nominalizations. Additionally, it was determined that nominalized words were typically domain-specific. While this study did explore the use of computational methods in assessing nominalizations in academic writing and uncovered some interesting findings, further research on this topic is needed in order to develop definite claims and determine how the frequency of nominalization impacts the academic writing of different fields.

References

Biber, D. (1986). Spoken and Written Textual Dimensions in English: Resolving the Contradictory Findings. Language, 62(2): 384-414.

Biber, D., Conrad, S. and Reppen, R. (1998). Corpus Linguistics: Investigating Language Structure and Language Use. Cambridge: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. (1999). The Longman Grammar of Spoken and Written English. London: Longman.

Bird, S., Loper, E., and Klein, E. (2009). Natural Language Processing with Python. O'Reilly Media Inc.

Chomsky, N. (1970). Remarks on Nominalization. In R. A. Jacobs and P. S. Rosenbaum (Eds.). Readings in English Transformational Grammar. Waltham, Massachusetts: Ginn, pp. 184-221.

Eggins, S. (2004). An Introduction to Systemic Functional Linguistics (2nd edition). London: Continuum.

Halliday, M. A. K., (1985). *An Introduction to Functional Grammar*. London: Arnold.

Halliday, M. A. K. (1993a). The Construction of Knowledge and Value in the Grammar of Scientific Discourse: Charles Darwin's The Origin of the Species. In M. A. K. Halliday & J. K. Martin (Eds), *Writing Science: Literacy and Discourse Power*, 86–105. Washington/ London: Falmer.

Halliday, M. A. K. (1993b). Some Grammatical Problems in Scientific English. In M. A. K. Halliday & J. R. Martin (Eds), *Writing Science: Literacy and Discursive Power*, 69–85. Washington/London: Falmer.

Halliday, M.A.K. (2004). *The Language of Science*. London: Continuum.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2004). *An Introduction to Functional Grammar*. London: Hodder Education.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2014). *Halliday's Introduction to Functional Grammar*. London: Routledge.

Hewings, A., & Hewings, M. (2005). *Grammar and Context: An Advanced Resource Book.* London: Routledge.

Honnibal, M. & Montani, Ines. (n.d.). spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing. Location: https://spacy.io/

Lapata, M. (2002). The Disambiguation of Nominalizations. *Computational Linguistics*, *28*(3), 357–388. doi: 10.1162/089120102760276018

Lees, R. B. (1960). The Grammar of English Nominalizations. The Hague: Mouton de Gruyter.

Liu, Y., Fang, A. C., & Wei, N. (2017). A Corpus-Based Study of Syntactic Patterns of Nominalizations Across Chinese and British Media English. *Researching Chinese English: The State of the Art Multilingual Education*, 77–92. doi: 10.1007/978-3-319-53110-6_6

Macleod, C., Grishman, R., Meyers, A., Barrett, L., & Reeves, R. (1998). NOMLEX: A Lexicon of Nominalizations. *Proceedings of EURALEX '98.*

Manning, C. D. (2011). Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics? *Computational Linguistics and Intelligent Text Processing Lecture Notes in Computer Science*, 171–189. doi: 10.1007/978-3-642-19400-9_14

Martin, J. R. (2008). Incongruent and Proud: De-vilifying 'Nominalization'. *Discourse Society*, 19 (6): 801–810. https://doi.org/10.1177/0957926508095895

National Science Foundation. (2019). *Research Areas.* Available at: https://www.nsf.gov/about/research_areas.jsp

Princeton University (2010). "About WordNet." WordNet. Princeton University.

Thomson, E., & Droga, L. (2012). Effective Academic Writing: An Essay-Writing Workbook for School and University. Australia: Phoenix Education.

To, V., Fan, S., & Le, Q. (2016). Research Writing. *What Is Next in Educational Research?*, 341–352. doi: 10.1007/978-94-6300-524-1_29

To, V. T., & Mahboob, A. (2018). Complexity of English Textbook Language: A Systemic Functional Analysis. *Linguistics and the Human Sciences*, *13*(3), 264–293. doi: 10.1558/lhs.31905

U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics.

Weischedel, R., et al. (2013). OntoNotes Release 5.0 LDC2013T19. Web Download. Philadelphia: Linguistic Data Consortium.