The Thirty-Third International
FLAIRS Conference (FLAIRS-33)

A Computational Approach to Assessing Nominalizations in Academic Writing

Yanisa Haley Scherber
Department of English, University of Alabama, Tuscaloosa, AL 35487
ykhaley@crimson.ua.edu

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 & History, and the fields with the lowest frequency of nominalization were Biological & Biomedical Sciences and Visual & 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.

Introduction

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)*.

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 & Matthiessen 2004).

Recent Work

The majority of current research on nominalization in writing investigates the number of nominalizations by searching for particular word endings associated with nominalizations (e.g. Biber et al. 1998; Biber et al. 1999; To et al. 2016; To & Mahboob 2018). A list of these common endings is included as Table 1 (Thomson & Droga 2012).

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

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

While this approach certainly captures many nominalizations, it is limited in its ability. For example, the ending *-th* will capture many nominalizations, such as *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*.

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 relatively small amount of research which exists on this topic, there is still much research needed to develop a larger body of knowledge around computational approaches to assessing nominalizations.

Purpose of the Study, Research Questions, and Hypotheses

This exploratory study attempts to investigate the aforementioned gap in research and provide information on how nominalization-use differs amongst various 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. The fields are (in order of descending popularity): Business-Management, Health Professions & Related Programs, Social Sciences & History, Psychology, Biological & Biomedical Sciences, Engineering, and Visual & Performing Arts.

The research questions investigated in this paper are as:

(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 & 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 & 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.

Methods

Data Collection

Data for this study was collected by extracting 1.8 million tokens of text from recent academic journals within the aforementioned subfields. Within the fields of Social Sciences & History, and Visual & Performing Arts, a sample of subfields was identified, given the large variety of subjects within each field. For Social Sciences & History, Economics, History, and Linguistics were selected as sampled subfields. For Visual & Performing Arts, Music and Theatre were selected as sampled subfields.

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.
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. Identified nouns not included in NOMLEX-PLUS were parsed through WordNet using NLTK. 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.
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.

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 & 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 & Related Programs	253,517	36,791	14.51%
Biological & Biomedical Sciences	242,202	30,543	12.61%
Visual & Performing Arts	206,967	25,769	12.45%
Music	*129,274*	*16,186*	*12.52%*
Theatre	*77,693*	*9,583*	*12.33%*

Table 2: Academic Fields and Corresponding Nominalization Frequencies

Results

Frequency of Nominalizations by Academic Field

Table 2 provides a summary of the frequency of nominalizations in each field. Business-Management, Psychology, and Social Sciences & History have the highest frequency of nominalization, which mostly conflicted with the hypothesis that Business-Management, Visual & Performing Arts, and Social Sciences & History would contain the lowest frequency of nominalization. In fact, Business-Management has the highest frequency of nominalization, and Biological & Biomedical Sciences is only .16 percentage points above the field with the lowest frequency, Visual & Performing Arts.

Most Frequent Nominalized Words

Table 3 shows bar graphs of the top three nominalizations in each academic field. As can be seen, 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.

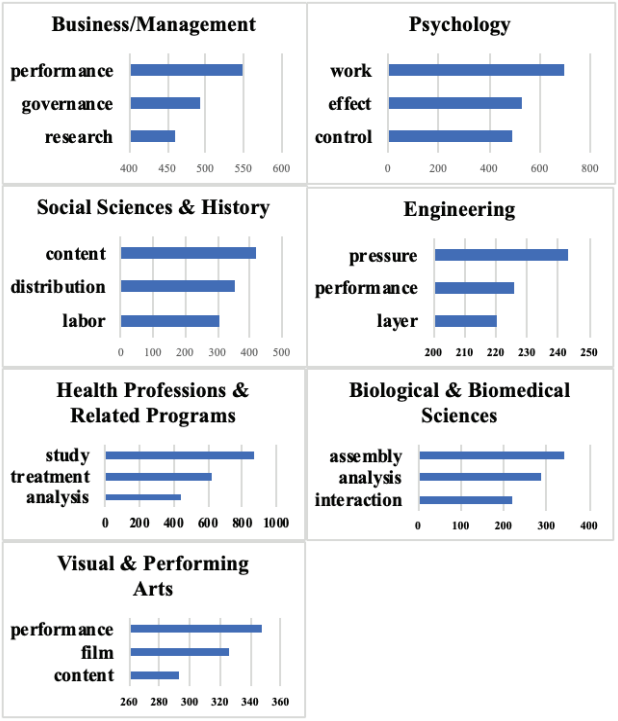

Table 3: Top Three Nominalized Words by Academic Field

Discussion

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 other aspects (e.g. lexical density). Additionally, the NSF is only one organization, and it could be very reasonably argued that Business-Management journals are also 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 & 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 indicate this method of searching for nominalizations re-

quires some revision. 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 within the field (e.g. *control* in *Group A was used as the control.*).

Additionally, in the process of identifying nouns not found in NOMLEX-PLUS, the criteria for identifying nominalizations was that the derivationally related form must have the same first three letters and 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 not *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*). While this second check did block many adjectivals from being labeled as nominalizations, 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 *drink* (verb), but erroneously identifies *water* as a nominalization for *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 undergeneralizes. In the future, a study comparing the accuracy of the two methods should be conducted.

Additionally, given the results of this computational approach, it is recommended that future study consider the addition of domain-specific filtering, and developing an established approach for counting common collocations.

Conclusion

This study aimed to explore the use of computational methods in assessing nominalizations in academic writing. Seven academic fields were assessed to determine the frequency of nominalization use and which words were being nominalized. It was found that Business-Management, Psychology, and Social Sciences & History contained the highest frequency of nominalizations, and Biological & Biomedical Sciences and Visual & 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., 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.

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

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.

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

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

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

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.