Noun phrase complexity analyzer (NPCA)

Soyeon Sim
Department of Applied Linguistics & ESL
Georgia State University
ssim6@gsu.edu

Abstract

A growing body of research is showing that fine-grained complexity indices, such as noun phrase complexity (NPC), are reliable descriptors for L2 writing quality, successfully distinguishing the advanced and less advanced L2 writers (Biber, Gray, & Poonpon, 2011; Kyle & Crossley, 2018; Lan, Lucas, & Sun, 2019). Much of the existing body of research on L2 writers' NPC, however, required manual data coding to identify the noun structures. Realizing this limitation and the growing interest in NPC, an NLP-based tool that can automatically calculate the frequency of the 10 noun phrase structures in the texts in the corpus with reference to Biber et al.'s (2011) hypothesized developmental stages was developed. The tool, *noun phrase complexity analyzer* (NPCA), was mainly based on NLP *spaCy* to identify all the noun phrase structures in a text and report the raw and normalized frequencies of each structure. The precision and recall rates indicated that all 10 features were of satisfactory levels (i.e., 94.66% precision, 92.99% recall).

1 Introduction

The current study introduces the *noun phrase complexity analyzer* (NPCA), an NLP-based tool that aims to measure the phrasal complexity of noun phrase structures. Syntactic complexity has been defined as "the range of forms that surface in language production and the degree of sophistication of such form" (Ortega, 2003, p. 492). However, the definition and operationalization of complexity still vary widely (Bulté & Housen, 2012), and a myriad of measures have been used to explore this construct, ranging from global complexity indices, such as mean length of clause, to fine-grained indices of clausal and phrasal complexity indices, such as non-clausal features embedded in noun phrases.

Global complexity indices, such as T-unit-based measures, which are widely adopted measures of grammatical complexity, has been increasingly receiving criticism. For instance, research that uses 'the length of T-unit' as a measure of complexity assumes that the longer the T-units are, the more complex the text is. It has been claimed that analyses that rely solely on such measures on the clausal level, may not paint an accurate picture of syntactic complexity. Instead, a growing body of research aims to adopt fine-grained complexity measures, such as noun and verb phrase

Stage	Noun phrase structure	Example
2	Attribute adjective as premodifier	a nice **flavor**
3	Relative clause with animate head noun	the **man** that was nice to me
	Noun as a premodifier	cable **channel**
	Possessive noun as premodifier	Mary's **voice**
	Of phrase as postmodifier	**chair** of the committee
	Simple PP as postmodifier (prepositions other than *of*)	**house** in the country
4	Nonfinite relative clause	**studies** adopting this method
	More phrasal embedding in the NP (attributive adjectives, nouns as premodifiers)	positive propagule size **effects**
5	Complement clause controlled by a noun	the **hypothesis** that female body was more variable
	Extensive phrasal embedding in the NP (multiple prepositional phrases as postmodifiers, with levels of embedding)	the **presence** of layered structures at the borderline of cell territories

Table 1: Biber et al.'s (2011) hypothesized developmental stages of noun phrase complexity

complexity adding explanatory power to linguistic research on complexity (Biber, Gray, & Poonpon, 2011; Kyle & Crossley, 2018; Lan, Lucas, & Sun, 2019).

While an increasing amount of attention has been paid to second language (L2) learners' noun phrase complexity, most of the previous research adopted a manual data coding approach, which often requires an immense amount of time and effort to identify and code all nouns. The noun phrase complexity analyzer (NPCA) was developed to automatically measure noun phrase complexity.

2 Design and use of NPCA

Noun phrase complexity analyzer (NPCA) is an NLP-based tool that identifies all the noun modifiers in a text and reports the raw and normalized frequencies per 1,000 words of each structure. The structures that the tool identifies are the 10 types of noun phrase structures proposed by Biber et al. (2011), which are presented in Table 1. Biber et al. (2011) hypothesized that the structures in the lower stages are acquired before the structures in the higher stages. Therefore, those structures could be assumed to be adopted more frequently by L2 learners than the higher-stage structures. The complete code of the tool is publicly available at `https://github.com/soyeonsim1/npca/blob/main/NPCA_SoyeonSim.py`.

Using *spaCy's part-of-speech tagger*, the NPCA automatically calculates the raw and normalized frequency of the 10 noun phrase structures and present them in an output csv file. For instance, the structure '*complement clause controlled by noun*' in stage 5 has the following structure: [Noun + *that* + (independent clause)]. If a token in a text is identified as a noun or a pronoun, it then moves on to the following token. If the following token meets all of the following criteria, it is identified as the target structure (i.e.,):

(1) '*that*' as a lowercase string
(2) The part-of-speech is a subordinating conjunction. ('`SCONJ`')
(3) The syntactic dependency tag is a marker that introduces a clause subordinate to another clause. ('`mark`').

Meeting all three criteria, the noun token, along with the following token and its right children tokens are appended to the total list of the noun phrase structures, after which the frequency of this

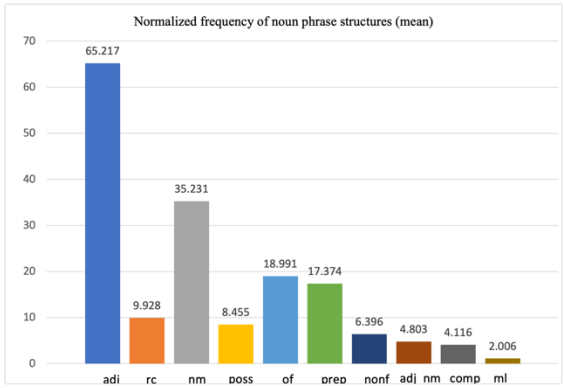

Figure 1: A sample visualization of the mean normalized frequency for the 10 noun phrase structures

structure is calculated. The rest of the nine structures followed a similar approach, in which the noun is identified and the preceding or the following tokens are analyzed in terms of part-of-speech and/or syntactic dependency. Using *Qt Designer*, the graphic user-interface (GUI) of the tool was designed for users' convenience.

An illustration of the application of the NPCA results is shown in Figure 1, which shows the mean normalized frequency rates of the 10 structures. The corpus used in this example consists of 120 source-based writing tasks written by Korean L2 college English learners. Such visualizations could be a useful starting point for tracking the developmental trajectories of L2 learners especially if there are multiple corpora of different proficiency levels. Researchers can also zoom in on the specific language structures that do not seem to fit the hypothesis (as in '*rc*' and '*poss*' in Figure 1) and explore why that is the case by qualitatively analyzing the texts.

3 Accuracy test results

Using the same corpus in the previous section, accuracy tests were conducted. Three text files were randomly selected from the corpus and all 10 noun phrase structures were manually coded for precision and recall rates. It was found that nine out of 10 structures had higher than 90% precision and recall rates, with the lowest rate being 92% for the recall rate of '*noun as a premodifier*'. One structure, '*simple PP as postmodifier*' had the precision rate of 89.3% and recall rate of 81.5%, which is still fairly high rate for reliable use of the tool.

Limitations

Although the NPCA can provide useful help to researchers and teachers, it is not without its limitations. First, it falls short of accounting for the language errors within the noun phrases (e.g., 'the *man* (main) problem). Considering that L2 learners are especially more prone to making simple mechanical or grammatical errors, further fine-tuning of the program should be conducted for higher accuracy. Second, the 10 noun phrase structures could be further specified for more nuanced research. For instance, some researchers have been attempting to include additional indices or separating one index into two specific indices, such as '*-ed participles as post-modifiers*' and '*-ing participles as post-modifiers*' (Sarte & Gnevsheva, 2022). Adding these additional indices may help the tool to paint a more comprehensive picture regarding syntactic complexity.

Ethics Statement

The development process of the tool complies with the ACL Code of Ethics. The corpus that was used in this study for accuracy tests has been approved by the Institutional Review Board at Georgia State University of the data collection and informed consent forms had been collected from the participants prior to the collection of the data to protect the rights and welfare of the human participants.

References

Douglas Biber, Bethany Gray, and Kornwipa Poonpon. 2011. Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, 45(2): 5-35.

Bram Bulté and Alex Housen. 2012. Defining and operationalizing L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency – Investigating complexity, accuracy and fluency in SLA* (pp. 21-46). Amsterdam: John Benjamins.

Kristopher Kyle and Scott. A. Crossley. 2018. Measuring syntactic complexity in L2 writing using fine-grained and phrasal indices. *The Modern Language Journal*, 102(2): 333-349.

Ge Lan, Kyle Lucas, and Yachao Sun. 2019. Does L2 writing proficiency influence noun phrase complexity? A case analysis of argumentative essays written by Chinese students in a first-year composition course. *System*, 85:1-13.

Lourdes Ortega. 2003. Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, 24(4):492-518.

Kayla M. Sarte & Ksenia Gnevsheva. 2022. Noun phrasal complexity in ESL written essays under a constructed-response task: Examining proficiency and topic effects. *Assessing Writing*, 51.