Noun Phrase Complexity Analyzer (NPCA)

Soyeon Sim
Georgia State University, Atlanta, GA

Background

- Syntactic complexity: "the range of forms that surface in language production and the degree of sophistication of such form" (Ortega, 2003, p. 492)
- Measures of syntactic complexity
  - Global complexity measures
    - e.g., *mean length of clause*, *the number of dependent clauses per T-unit*
  - Fine-grained complexity measures
    - e.g., *non-clausal features embedded in noun phrases*

Q. Which sentence is more syntactically *complex*?

(a) Well since he got so upset, I just didn't think we would want to wait for Tina to come back. (the # of dependent clauses per T-unit: 4)

(b) This may be part of the reason for the statistical link between schizophrenia and membership in the lower socioeconomic classes. (the # of dependent clause per T-unit: 0)

➔ Using a global complexity measure, **(a)** is more complex than (b).

- 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).
- Most studies that explored noun phrase complexity, however, used manual coding of all noun phrases in the corpus making the analysis time-consuming and labor-intensive.

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 10 noun phrase structure types are based on Biber et al.'s (2011) hypothesized developmental stages. (**Table 1**)
- Using *spaCy*'s part-of-speech tagger and dependency parser, all 10 noun phrase structures were identified and put into a list.
- As an output, the tool generates a CSV file that includes all raw and normalized frequencies per 1,000 words for each of the noun phrase structures.
- The NPCA is designed to automatically measure noun phrase complexity and enable a more efficient analysis in NPC.

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

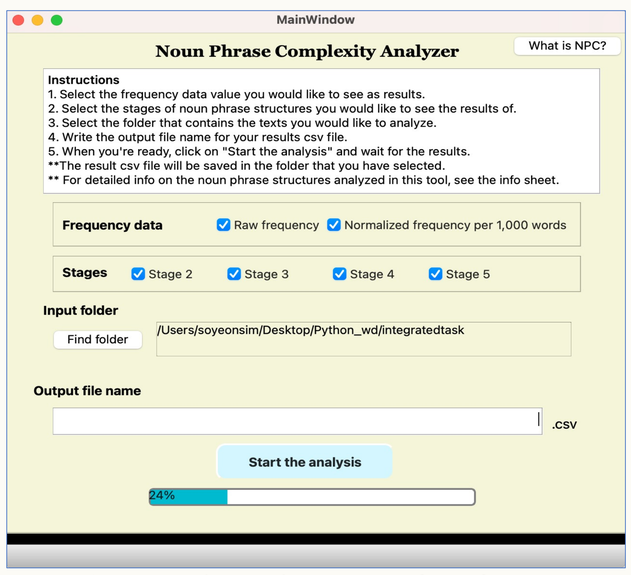

Figure 1. The graphic user-interface of NPCA

- A snippet of the code for the structure "Extensive phrasal embedding in the NP" (e.g. *the negative **effect** on the social life of an individual*)

```
if token.pos_ in ['NOUN', 'PRON']:
    # Initialize a variable to store the current phrase
    current_phrase = token.text
    next_right_token_found = False
    prep_count = 0
    prep_tokens = []
    for tok in token.rights:
        if tok.dep_ == 'prep':
            prep_tokens.append(tok)
            prep_count += 1
    while len(prep_tokens) != 0:
        prep_token = prep_tokens.pop()
        current_phrase = f"{current_phrase} {prep_token.text}"
        for right in prep_token.rights:
            if right.dep_ == 'pobj':
                current_phrase = f"{current_phrase} {right.text}"
                if token.i < right.i:
                    token = right
                for r in right.rights:
                    if r.dep_ == 'prep':
                        prep_tokens.append(r)
                        prep_count += 1
```

If the token is a noun or pronoun,

And the token next to it is a preposition,

Iterate over the right dependent tokens of the prepositional object ('pobj') to see whether there is an additional prepositional phrase

Accuracy test results

- Corpus: 120 source-based writing tasks written by Korean learners of English
  - Proficiency level: Intermediate-high to advanced
  - Topic: smart cars & Amtrak
- Precision and recall rates were calculated by manually coding 3 texts from the corpus for all 10 noun phrase structures
- 8 out of 10 structures had higher than 90% precision and recall rates
- The lowest accuracy rate was '*simple PP as postmodifier*' which had 89.3% of precision rate and 81.5% of recall rates.

NP	Precision	Recall
Attribute adjective as premodifier	98%	97.96%
Relative clause with animate head noun	91.2%	89.3%
Noun as a premodifier	100%	92.01%
Possessive noun as premodifier	94.2%	96.5%
Of phrase as postmodifier	100%	100%
Simple PP as postmodifier	89.3%	81.5%
Nonfinite relative clause	93.4%	99.1%
More phrasal embedding in the NP	92.3%	100%
Complement clause controlled by a noun	100%	100%
Extensive phrasal embedding in the NP	100%	100%

Table 2. Precision and recall rates for the 10 noun phrase structures

Applications

Research Implications

- Quantitative analyses of noun phrase complexity

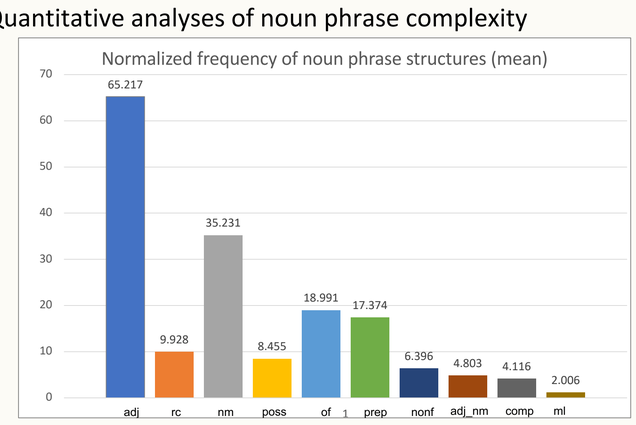

Figure 2. A sample visualization of the mean normalized frequency for the 10 noun phrase structures

- **Figure 2** shows a visualization of the mean normalized frequency for 10 structures in the corpus.
- 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 could also examine the specific language structures that do not seem to fit the hypotheses ('rc' and 'poss' in Figure 2)
- Comparing the English proficiency levels in the corpus (**Figure 3**), the intermediate learners tended to use the lower-stage noun phrase structures more frequently, and the advanced learners adopted the higher-stage NP structures more frequently.

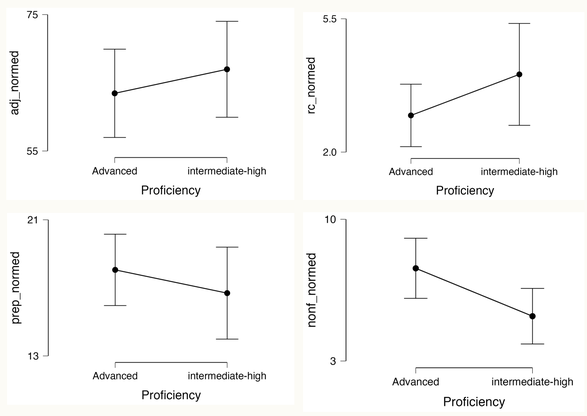

Figure 3. Descriptive plots of the four NP structures across the proficiency levels

Pedagogical Implications

- Based on the results, learner writing could be qualitatively analyzed, which could be useful reference in L2 teaching.

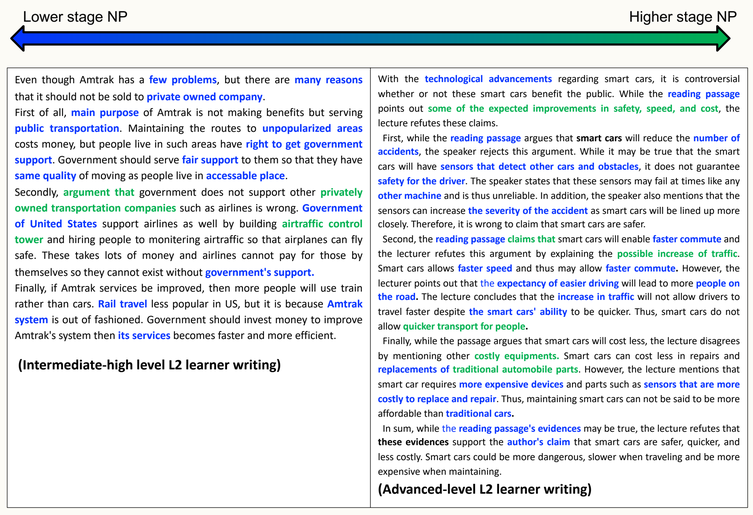

Figure 3. Qualitative analyses of the complex NP structures in INT and ADV writing

Limitations & Future directions

- The tool generates the raw and normalized frequencies of NP structures, which may be useful for quantitative analysis, but does not provide the list of identified noun phrases, which could be used for qualitative analysis.
- Future development of the tool could show each of the structures in full context or concordance format.
- The tool currently does not take into account spelling or typing errors, which are common in timed essays or learner writing. The tool could be trained using machine learning methods.
- More fine-tuned semantic criteria could be added → animate nouns & inanimate nouns

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.