Syntax and Discourse in Good and Poor Scientific Articles

Yuichiro KOBAYASHI and Shosaku TANAKA
College of Letters
Ritsumeikan University
Kyoto, Japan
kobayashi0721@gmail.com

The aim of this study is to profile the use of syntax and discourse in English scientific articles. Based on a corpus of five million words, we clarify the difference between good and poor scientific articles in terms of the frequencies and usage of part-of-speech (POS) n-grams and metadiscourse markers.

This study draws on a corpus of English scientific articles with information on the quality of their linguistic features. The procedure of building this corpus includes the following three steps. First, electronic articles put on the Web are collected automatically using a search engine and keywords of scientific categories. Second, the quality of linguistic features, not of the content, of each article is evaluated by professional reviewers. To put it more concretely, the quality assessment is based on the number of grammatical errors, unnatural collocations, over-complicated sentences. Finally, these articles are organized as a specialized corpus of English scientific articles. The corpus used in this study contains 384 "good" articles which have fewer than two grammatical errors per 250 words and 397 "poor" articles which have more than four grammatical errors per 250 words.

The statistical method for the comparison of good and poor scientific articles is random forests (Breiman, 2001). It can be defined as a technique for pattern recognition and machine learning, and it is known as a powerful method for feature extraction and text classification. In this study, the explanatory variables are the frequencies of POS n-grams (n=2~4) and metadiscourse markers. POS information for n-grams is automatically tagged by TreeTagger (Marcus, *et al.*, 1993), and the list of metadiscourse markers is based on Hyland (2005).

Using random forests with POS n-grams, we identify syntactic patterns in good and poor scientific articles. The most prominent feature of syntax in poor articles is the overuse of passive voice (e.g. *VBZ-VVN-IN*, *VBP-VVN-IN*). There is a common misconception that the passive voice will sound more impressive in an academic text, and the major cause of dullness in scientific writing is the overuse of passive voice (e.g. Gross & Sis, 1982). Another notable feature is coordinate conjunctions (e.g. *CC-DT-NN, NN-NN-CC*). On the other hand, good articles are characterized by sequences of adjectives (e.g. JJ-JJ) and those of adverbs (e.g. RB-RB).

We also identify discourse patterns in good and poor scientific articles using random forests with metadiscourse markers. While poor articles are characterized by transitions (e.g. *also, on the other hand, therefore*) and self-mentions (e.g. *we*), good articles are marked by a great variety of hedges (e.g. a*pparent, appear, could, largely, likely, may, perhaps, would*). Hedges represent a major "rhetorical gaps" (Hyland, 1995) that academic writers have to cross before they can gain membership of a scientific community.

These results are very informative for teaching English for specific purposes. Most of syntactic and discourse patterns in poor scientific articles are not *grammatical errors* but *stylistic deviations* which would hardly have been detected in traditional error analysis.

References

Breiman, L. (2001). 'Random forests.' *Machine Learning* 24: 123-140.

Gross, D. V., & Sis, R. F. (1982). 'Scientific writing: The good, the bad, and the ugly.' *Veterinary Radiology & Ultrasound* 23: 131-134.

Hyland, K. (1995). 'The author in the text: Hedging scientific writing.' *Hong Kong Papers in Linguistics and Language* 18: 33-42.

Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing*. New York: Continuum.

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). 'Building a large annotated corpus of English: The Penn Treebank.' *Computational Linguistics* 19: 313-330.