39th Annual Conference of the Japan Association for English Corpus Studies, 2013. Sendai, Japan.

John Blake

Personalised statistical writing analysis

English is the *de facto* language for scientific journals with the highest impact factors. This means that non-native English speaking (NNES) researchers have to not only master their specialism but also English, which for speakers of dissimilar languages, such as Japanese, is a significant hurdle. The aim of this pedagogically-driven project is provide objective statistical evidence that NNES researchers can harness to make decisions on how to improve the generic integrity and lexicogrammatical accuracy of their draft manuscripts.

Selected features of each draft research article (approx. 2,500-10,000 words) are compared with a specially-created corpus of the target publication (approx. 250,000 words) and a tagged corpus of the appropriate subject domain, namely information science, materials science or knowledge science (circa. 3 million words each). The subject domain corpora were annotated with parts of speech (POS) using GoTagger version 0.7, drawing upon rules from the Brill POS tagger. The generic integrity and lexical repertoire of drafts were then analyzed with respect to the target corpus and a subject-specific corpus using various textual analysis tools to generate data on vocabulary fit, readability, lexical profile, marked usage and grammatical accuracy.

The concept of keyness was used to identify the vocabulary fit with the respective target publication. The most frequently used unigrams, bigrams, trigrams, 4-grams and 5-grams were identified. Readability statistics, such as mean sentence length, lexical density, Gunning Fog index and Flesch reading ease, were also calculated and compared to the target corpus. The lexical profile was created using Tom Cobb's online vocab profiler (Web VP Classic v.4) to identify words listed in the general service and academic word lists. In addition, marked usage and lexicogrammatical errors were identified manually in the submitted draft. The keyword in context function was used to derive the statistical probability of the occurrence of marked forms. This was compared to an appropriate reference corpus, such as the untagged 100-million word British National Corpus or the tagged subject domain corpora. Grammatically incorrect usage was identified, corrected and commented on, and where relevant appropriate reference material was suggested.

Each researcher was presented with a personalized academic writing analysis showing a summary of the relevant statistics, followed by a more detailed analysis accompanied with an explanatory guide to help them interpret the statistics. Follow-up interviews were held with participating NNES researchers to identify the efficacy of the personalised statistical writing analysis. The future development and direction of this project will be discussed in light of this feedback.