Statistics for Scientists: Incorporating Data-Driven Decision Making in the Publishing Process

Xavier Blake
Japan Advanced Institute of Science and Technology
John Blake
Japan Advanced Institute of Science and Technology

Introduction

A number of scientific researchers making use of a university proofreading service ignored the recommendations suggested. To encourage researchers to act on the suggestions, the use of statistics was adopted, based on the rationale that scientists are used to drawing conclusions from numerical data and tend to place faith in their own hypotheses.

The aim of this project is to enable scientific researchers to harness statistical data to evaluate their own writing. The data provided is based on corpus-based analyses of articles submitted for proofreading and specially-created corpora of the respective target journals.

Generic integrity is investigated using textual profiling tools to generate statistical data. Grammatical accuracy and appropriacy are assessed manually by an experienced proofreader. The authors are given a feedback sheet showing the data on lexicogrammatical accuracy and appropriacy, and generic integrity (Bhatia, 1993).

Corpus creation and selection

One target journal corpus was created for each article submitted for proofreading. The target corpora were compiled by collecting every research article published in English in the target journals for a period of between 1 year and 4 years. The target corpora ranged in size from 40,000 words to 2.1 million words.

Reference corpora were created, downloaded or assessed online. A 2.5-million word balanced corpus of Information Science (IS) research articles was created and tagged with parts of speech. The freely-available Brown corpus was downloaded. Two web-based corpora, namely, the 450-million word Corpus of Contemporary American English and the 100-million word British National Corpus, were also accessed. Finally, when dealing with very low frequency items, the web was used as a corpus.

Grammatical accuracy and appropriacy

A proofreader identified lexicogrammatical errors at two levels, namely those that were grammatically incorrect and those that were inappropriate. Each draft manuscript was read to identify language that grammarians would consider to be incorrect. The corrected forms were provided along with brief explanations or references on the feedback sheet. Each manuscript was re-read to identify style and usage errors. Style errors frequently included the use of informal language and incorrect citation mechanics. Usage errors often resulted in language that was marked (i.e. unusual or complex). When found, more appropriate forms of usage were suggested. To show that the marked form was less frequent and therefore less appropriate, the frequency of occurrence of both the marked and suggested phrases was found using the Keyword in Context function of AntConc (Anthony, 2011) and an appropriate corpus. The less frequently the item occurred, the larger the corpus needed.

Generic integrity

To identify any discrepancies in the generic integrity between the article submitted and the target journal, textual profiling tools were used to investigate the readability, vocabulary fit and vocabulary profile.

Readability statistics show the reading difficulty of the article versus that of the corpus of the target journal. The Gunning fog index, the Flesch-Kincaid grade level, and the mean sentence length were calculated for the draft article and the target journal. The greater the difference, the less likely the article would match the style of the journal.

Vocabulary fit refers to the lexical similarities between the article and the target journal. Scott and Tribble state that "keyness [is what a text] boils down to" (2006, p.56) and so the core semantic properties of a text are likely to be ascertained through an investigation of those key words. Keyword lists were made for each article and the related target journal corpus using the keyword list function of AntConc with the Brown corpus as a reference. The semantic similarity of the results was compared to assess the probability of a paper-journal fit (Hyland, 2011).

The vocabulary profile, or word type balance, shows the relative percentage of particular categories of words. The online vocab profiler Web VP Classic v.4 (Cobb, 2013) was used to identify words on the General Service List (the most frequent 2000 words), words on the Academic Word List (Coxhead, 2000) and those not on either list, i.e. off-list words.

The ratios of different word types in the submitted article and target journal were calculated and compared.

Feedback sheet

For each submission, a feedback sheet was created containing data categorized using the headings: readability, vocabulary fit, word type balance, style & usage, and grammatical errors. In order to provide an easy-to-understand overview, a rating is given on a five-point scale for each of the five aspects investigated. Items rated as "1" need substantial assistance while those rated as "5" are of publishable quality. An explanatory guide was also provided.

Further research

Although we have incorporated data-driven decision making in the publishing process for scientific writing, its effectiveness has not yet been established.

This can be addressed by comparing drafts submitted for proofreading to the published articles to ascertain the effect of the feedback on the final version. This could be complemented by conducting questionnaire and interview surveys with researchers participating in this study.

References

Anthony, L. (2011). *AntConc (Version 3.2.4) [Computer Software]*. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/

Bhatia, V. K. (1993). *Analysing genre: Language use in professional settings.* London, England: Longman.

Cobb, T. (2013). *Web Vocabprofile.* Available from http://www.lextutor.ca/vp/

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238.

Hyland, K. (2011). Welcome to the machine: Thoughts on writing for scholarly publication. *Journal of Second Language Teaching and Research, 1*(1), 58-68.

Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education.* Philadelphia, PA: John Benjamins.

Authors' bios

John Blake *has taught English at universities and schools for over 20 years in Thailand, Hong Kong, Japan and the UK. He is a research lecturer at the Japan Advanced Institute of Science and Technology. His research interests include ESP, EAP and corpus linguistics. johnb@jaist.ac.jp*

Xavier Blake *spent his gap year studying corpus linguistics, mathematics and science at the Japan Advanced Institute of Science and Technology. He has since enrolled in an engineering degree at Assumption University of Thailand.*

Received: October 1, 2013
Accepted: February 18, 2014