Paper presented at the 20th KOTESOL International Conference, 20-21 October, 2012.
Sookmyung Women`s University, Seoul, Korea.

Harnessing technology to help researchers avoid plagiarism

John Blake

Japan Advanced Institute of Science and Technology, Nomi, Japan

Abstract

There is increasing pressure on academics for whom English is a second language to publish in international journals and conference proceedings. In fact, job prospects, tenure and salary may be directly related to this. Researchers not only have to master their content knowledge, but also come to grips with English. This presentation reports on the progress of the development of an online resource that inexperienced writers of English can use to help scaffold their attempts to draft research abstracts for scientific articles by selecting an appropriate template and functional exponents to use as a skeleton around which they can build the body of their abstract. Abstract writing is a complex field with an infinite amount of choices available to writers; yet there are well-defined shared expectations of the community of practice. The move structure of abstracts in one scientific journal was examined using genre analysis and systemic functional grammar. Each move within the abstract was manually labeled and a database was created. The data contained some phrases of language that students could draw on, such as 'in this study' while others could leave them open to accusations of plagiarism. Drawing on the concept of keyness, each commonly used word or phrase in that genre was tagged as 'key', that is, as having a high statistical probability of occurrence in this genre. Learners can, therefore, make an informed research-based choice of whether to harness a 'key' phrase or whether to select their own.

I. Introduction

Researchers need to be able to draft abstracts in English even when publishing in non-English journals targeted at non-English audiences. For those aiming to publish in prestigious journals and submit papers to international conferences, English abstracts are a must. The pressure to publish or perish has now become: the pressure to publish *in English* or perish.

Abstract writing is a complex craft with an infinite amount of choices available to writers; yet there are well-defined shared expectations of the community of practice. Novice writers often fail at the first step by omitting moves, or particular functions, that reviewers expect to be present. The problem may be exacerbated by the use of unexpected or odd-sounding terms that do not commonly occur in this genre.

In this paper the progress of the development of an online resource is described. This resource aims to assist inexperienced writers of English by scaffolding their attempts to draft research abstracts for scientific articles. Writers select an appropriate template and functional exponents to use as a skeleton onto which they can hang the body of their abstract.

II. Corpus analysis

The move structure of research abstracts in one scientific journal was examined using genre analysis and systemic functional grammar. The 3A perspective of annotation, abstraction and analysis (Wallis and Nelson, 2001) was adopted for this study. A corpus of scientific research abstracts published in the journal *AI and Society* (H index = 9) between January 2011 and August 2012 was compiled and double-checked by a research assistant. The corpus consisted of 54 abstracts, with a mean length of approximately 157 words, and a total word count of 8,454 words.

.

A. Corpus annotation

Annotation with parts of speech (POS) was conducted using GoTagger version 0.7, which employs rules from the Brill POS tagger. Drawing upon Hallidayan linguistics, we focus on the experiential metafunction which was later subsumed within the ideational metafunction (Halliday and Matthiessen, 2004). This was achieved by manually labeling each Process, Participant and Circumstance. Each of which was classified further according to its exact type using the standard systemic functional grammar (SFG) terminology.

B. Abstraction

Using the five-move framework disseminated by Swales and Feak (2009) and reproduced in Table 1, each move within the research abstracts was manually labeled and a database was created. This resulted in the identification of over 250 moves in the corpus.

Move	Common labels	Implied questions
1	Background / Introduction / Situation	What do we know about the topic? Why is the topic important?
2	Present research / Purpose	What is this study about?
3	Method / Materials/ Subjects / Procedures	How was it done?
4	Results / Findings	What was discovered?
5	Discussion / Conclusion / Implications / Recommendations	What do these findings mean?

Table 1: Move structure of research abstracts (Swales & Feak, 2009, p.5)

C. Analysis

The corpus contained a number of phrases that writers could draw on, such as ‘in this study’, but also contained a much larger set of expressions that if ‘borrowed’, could leave them open to accusations of plagiarism. In order to help writers decide which phrases were recyclable and which could be avoided the probability of each phrase was calculated. The initial idea was to harness the document similarity metric of statistically improbable phrases, but on further investigation, the concept of keyness proved easier to use. Statistical probability of words or phrases in a particular genre can be evaluated by measuring their frequency in that genre and comparing that with their probability in general usage such as in a larger reference corpus, such as the Brown Corpus and the Lancaster-Oslo-Bergen Corpus. Keyness is the linguistic term frequently used to describe whether a word or phrase is central or ‘key’ within a genre. In order to identify words with a high degree of keyness, namely 10 times higher than normal; the most frequently used unigrams, bigrams, trigrams, 4 n-grams and 5 n-grams were analysed using the AntConc3.2.4w concordancer (Anthony, 2012). Phrases that occurred proportionally much more frequently, such as ‘in this paper’, were classed as highly probable or ‘key’.

III. Template and functional exponent database

An integrated website was created which housed three separate tools, namely: skeleton templates, functional exponents and a frequency probability or ‘keyness’ checker.

A. Templates

Ten templates were created to provide writers with some degree of choice of skeleton frameworks onto which would act as a starting platform to help them start the drafting process. The skeleton templates are divided into five moves to guide the writers through each function in order to raise the writer’s awareness of the expected moves.

B. Functional exponents

Functional exponents and useful phrases that were used in the corpus were extracted and classified so that writers could browse through them while drafting their abstract. Functional exponents were ordered according to relative frequency so that writers who were not sure which particular exponent to select could simply opt for the first one, knowing that it is frequently used and therefore safe for them to use.

C. Frequency probability checker

When the first draft is completed, it can be submitted to the frequency probability checker. The submitted draft is searched using regular expressions. Any ‘key’ words and phrases that have been added to the MySQL database are highlighted to show the writer that those phrases are commonly used and therefore will not be considered as plagiarism, but simply as text that is recycled.

IV. Recommendations

This is the second reincarnation of an online tool to aid non-native English speakers to draft initial versions of research abstracts using skeleton templates to scaffold the creation of their own abstract. Recommendations stemming from this project and due to be incorporated into the next version are: (1) create a larger corpus to test the validity of generalisations made on the initial corpus, (2) incorporate filter questions to ensure writers start with the most appropriate template, (3) create a database of individual skeleton moves that can be combined to create a much larger selection of templates, and (4) utilise the SFG tagging more effectively.

References

Anthony, L. (2012). AntConc (Version 3.2.4) [Computer Software]. Tokyo, Japan: Waseda University.

Halliday, M.A.K. & Matthiessen (2004). *An introduction to functional grammar (3rd ed).* London: Hodder Education.

Swales, J.M., & Feak, C.B. (2009). *Abstracts and the writing of abstracts*. Michigan: University of Michigan Press.

Wallis, S. and Nelson G. (2001). Knowledge discovery in grammatically analysed corpora. *Data Mining and Knowledge Discovery, 5,* 307-340.

The Author

John Blake *has taught English and trained teachers for over 20 years in various universities and colleges around the world. He holds master’s degrees in applied linguistics and business administration, as well as postgraduate diplomas in education, management and language teaching. He has worked as a translator and is able to converse in Japanese, Thai and Cantonese. His main research interest is discourse analysis. Email: johnb@jaist.ac.jp*