Research Abstract Templates: The Skeletons are Out of the Virtual Cupboard

John Blake

Japan Advanced Institute of Science and Technology (Japan)
johnb@jaist.ac.jp

Abstract

Web-based resources are an ideal way to support individualised learning as they provide an eminently customizable environment in which the level of scaffolding can be tailored to the user. Templates for writing tasks in language learning textbooks are often rather simplistic, adopting a one-size fits all approach; however, web-based templates can be customised rather easily by the user and provide a much more specific template that is better able to scaffold complex writing processes. One particularly complex task is the drafting of scientific research abstracts. Their high lexical density makes research abstracts difficult to write [1], which is why mastering the process of abstract writing may be considered as the "rite de passage" of entry into the scientific community [2]. This study can be divided into three phases. First, we designed an abstract creation tool (ACT) for postgraduate students in the School of Information Science at the Japan Advanced Institute of Science and Technology. Second, we created an online prototype. There are three notable features to the ACT prototype, namely: (1) the selection of the most relevant template from the collection based on filter questions, (2) the provision of dropdown menus to enable users to vary the functional exponents within the template and add an extra layer of tailoring, and (3) colour coding to show the statistical improbability of phrases, which in turn enables users to decide whether their usage could be considered as plagiarism. Third, we evaluated ACT by collating feedback on the perceived usefulness of the tool from the users themselves via observation, questionnaires and interview surveys. A number of abstracts created with help from ACT were also evaluated for content and language by experienced reviewers of journal articles and conference proceedings. Based on the results of these evaluations, recommendations for further development are given.

1. Introduction

1.1 Background

Non-native English speakers have to not only master their specialism but also overcome the difficulties [1] inherent in the "*rite de passage*" of mastering the genre of abstracts and joining the discourse community [2]. English is now the *de facto* language for scientific journal articles [3] and research abstracts are one of the *de facto* gateways to the research literature [4]. The oft-quoted phrase 'publish or perish' has now come to mean 'publish in English or perish' for the scientific research community. Even for papers written in Japanese for publication in Japanese scientific journals and read primarily by Japanese researchers, abstracts in English are required.

In order to "give the reader an exact and concise knowledge of the full article" [5], a full-length research paper needs to be stripped down to its kernel. This significantly increases the lexical density and linguistic complexity of abstracts. These factors combine together to increase comprehension difficulty.

Given the word constraints and the necessity for precision, scientific research abstracts frequently utilize technical and obscure terms that act as a form of shorthand to those familiar with the terminology, but simultaneously create an intimidating and impenetrable barrier for lay readers unfamiliar with the jargon.

Research abstracts, like other texts, contain moves, which put simply, could be described as words or sentences that carry out a particular function. Moves themselves can be subdivided into steps, which also have functions that are subsumed under the function assigned to the move. Numerous studies have analysed the move structure, or functional organisation, of research abstracts of journal articles. The consensus appears to be that they are either composed or four moves [6] [7] or five moves [8] [9] with the key difference being the presence or absence of the first move. Swales and Feak [10] in their seminal guide to writing research abstracts include five moves in a table, which is reproduced below:

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)

Research abstracts have been classified in the research literature for many years as indicative [11], informative or indicative-informative depending on whether particular moves are described in general, in detail or simply omitted [12]. This project focuses on informative scientific research abstracts for journal articles, which contain all five moves.

1.2 Overview
The overarching aim of this project is to create a prototype Abstract Creation Tool that Japanese postgraduate students and researchers in the School of Information Science at the Japan Advanced Institute of Science and Technology (JAIST) can use to create initial drafts of research abstracts. This prototype will serve as a quick-and-dirty testing ground for a more comprehensive integrated tool.
The preparation phase consisted of the design, creation and annotation of a corpus of research abstracts. Various analyses were conducted at genre, move, functional exponent, phrase and word levels to gather the necessary data.
The creation phase began with the extraction of skeleton sentences based on the corpus analyses. A website was created and data stored in a MySQL database. Filter questions were created to enable the relevant skeleton sentences to be accessed and combined together to create a tailor-made template based on user selections. A vertical dropdown menu was created to display functional exponents that users could draw upon while drafting their abstract. A frequency probability checker was also incorporated into ACT. This checker harnesses the concept of regular expressions (regexp) by searching the submitted abstracts for items occurring in the database of expressions that were found to be frequently used in the corpus. Any common expressions found in the submitted abstract are highlighted. Users can then focus more carefully on the less statistically probable aspects of their draft.
In the evaluation phase both the usability of ACT and the generic integrity of the abstracts produced were evaluated.

2. Preparation phase
In this phase two corpora were compiled and various statistical corpus analyses conducted.

2.1 Corpus creation
Artificial intelligence, one of the five research domains of the School of Information Science at JAIST, was chosen as the first domain to be targeted based on the number of postgraduate students within JAIST aiming to publish in English in that field.
The Springerlink journal AI and Society, which is indexed in Scopus and has an H index of 9, was chosen on the recommendation of a focus group. This group consisted of researchers and postgraduate students who felt that it was an appropriate target journal. The variables for an ideal corpus were established [13]. A whole population corpus sample method was chosen. Abstracts for all research articles published in AI and Society from January 2002 (Issue 1-2, Volume 16) to August 2012 (Issue 3, Volume 27) were collected and logged by a research assistant who followed a standardized operating procedure. The corpus was double-checked using a similarity assessment tool to ensure that no abstracts had accidently been entered twice. The reference corpus contained over ten years of abstracts with exactly 257 abstracts, consisting of 36,306 words, including author names and article titles. A sub-corpus of recent abstracts (without names or titles) published from January 2011 was also compiled, comprising 54 abstracts, taken from the most recent 7 issues, resulting in a total of 8,454 words. The mean abstract length was around 157 words. Both the

reference and the recent corpora were annotated with parts of speech (POS) using GoTagger version 0.7, which draws upon rules from the Brill POS tagger.

2.2 Corpus analysis

All the analyses were conducted on the corpus of recent abstracts, and the results compared with the reference corpus. Following the practice in other studies [14] [15] [16], each abstract was manually divided into moves and each move classified according to the five-step model. This was conducted twice by the same researcher with an intra-coder reliability of .98. The moves were subdivided into steps based on those described in the Alto University Academic English Writing guide. The AntConc3.2.4w concordancer [17] was used to identify the most frequently used unigrams, bigrams, trigrams, 4 n-grams and 5 n-grams. The results of which were stored in a database for use with the regexp search function in the frequency probability checker. Commonly occurring functional exponents were entered into a separate database. The functional exponents were categorised according to their relative frequency within the reference corpus and the move in which they most frequently occurred. Figure 1 displays a visual representation of the most frequent words using Wordle, whose algorithm increases the font size in proportion to the relative frequency.

Figure 1: Most frequent words in the reference corpora

3. Creation phase

In this phase skeleton sentences were created which combine together to create templates, and a web interface was developed.

3.1 Template creation

The templates created are not for the abstract as a whole, but are in fact many sentence-level mini-templates that combine together in numerous ways to create a wide range of possible templates. In short, scaffolding is provided at the level of step rather than complete abstract. Skeleton sentences and sentence stems were created by extracting generative structures from the recent corpus. The skeleton sentences were classified according function and tagged according to move and step. The aim was that users would fill in the specific details of their research into the template and add descriptions, explanations and examples as necessary. Five skeleton sentences were identified for each of the four steps in moves 1 and 5, and ten sentences for each of the moves 2, 3 and 4. Each move is therefore realised in one to four sentences. The skeleton sentences were designed so that there was a high probability that they could follow on from any of the skeleton sentences occurring in the previous and following steps or moves. Although there are only 65 skeleton sentences in the database, the simple theoretical combination of these sentences used in the order given (5^4 x 10^3 x 5^4) results in a total of 7,812,500 different permutations. This should go some way to alleviate potential criticism of some of the limitations that a template may impose. Filter questions were devised at this stage to enable users to select the most appropriate skeletons to build their template. An example of a series of filter questions is shown in Figure 2.

Do you want to:
1. emphasize the relevance?   Yes / No
2. emphasize the importance?   Yes / No
3. emphasize the significance ?   Yes / No

4. describe the current common practice?	Yes / No
5. describe something that is already known	Yes / No

Figure 2: Extract of filter questions on Abstract Creation Tool

It is also hypothesised that simply by answering all the questions posed in ACT, users will have a clearer idea of the format and content of their abstracts regardless of whether or not they make use of the tailor-made abstract.

3.2 Website creation

The first prototype for the ACT website was built using frames, but was soon abandoned because of negative comments regarding the limited size of the usable screen space. This paper, however, describes the development of the second prototype. ACT essentially comprises a website linked to a MySQL database. The webpage interface was designed so that users initially answer multiple choice questions about their abstract. An individualised template using the most appropriate combination of the 65 skeleton sentences is generated from their answers. This template is displayed alongside detailed prescriptive advice on what to include in each move of the abstract. This advice contains information on the audience, purpose, organisation, flow, style and presentation of an abstract [18]. Links are also provided to relevant learning resources, such as exploratory tasks and data-driven learning tools. On completion of their abstract, users submit the first draft of their abstract into the frequency probability checker. The submitted draft is searched for the commonly used phrases extracted from the corpus of most frequently used phrases using regexp. The phrases that are found are highlighted so that users can focus their error checking on the statistically less probable aspects of their abstract. At present, there is no means to incorporate statistically improbable phrases, but this is planned for inclusion into a later version once technological hurdles have been overcome.

4. Evaluation phase

Discount usability testing [19] was used to make ACT more user-friendly. Users were allocated specific tasks and feedback was received through talk-aloud protocol and subsequent small-scale questionnaire and interview surveys. Suggestions were made for (1) the inclusion of complete templates which could be browsed, (2) the incorporation of an automatic error detector and corrector, and (3) a video explanation of how to use ACT. Abstracts produced by ACT were evaluated by experienced reviewers from within the same discipline. Despite the shortcomings in the abstracts evaluated, the reviewers thought that the overall structure of the abstracts was appropriate, but in some abstracts there was insufficient thematic development. It was also suggested that more skeleton sentences could be added to the database to extend the range of choices available for moves 2 to 4 (purpose, method and results).

5. Further development

The users` requests for a browsable library of templates and a video explanation will be implemented. Automatic error correction is technologically much more challenging and is beyond the scope of this project. Thematic development could be enhanced by asking users to input alternative terms to refer to their research, so these terms can automatically be incorporated into the templates, starting with the longest term for the first mention. Additional skeleton sentences will be added to the database and the selection algorithm altered accordingly.

References

[1] Halliday, M.A.K., & Martin, J.R. (1993). *Writing Science: Literacy and Discursive Power.* London: Falmer Press.
[2] Swales, J. M. (1990). *Genre Analysis.* Cambridge: Cambridge University Press.
[3] Simionescu, M., & Simion, E. (2004). *ALLEA Biennial Yearbook Critical Topics in Science and Scholarship* (pp.129-133). Amsterdam: ALLEA.
[4] Hartley, J., & Benjamin, M. (1998). An evaluation of structured abstracts in journals published by the British psychological society. *British Journal of Educational Psychology, 68*, pp.443-456.
[5] Bhatia, V. K. (1993). *Analysing genre: Language use in professional settings.* London: Longman.

[6] Salager-Meyer, F. (1994). Hedges and textual communicative function in medical English written discourse. *English for Specific Purposes, 13* (2), 149-171.

[7] Salager-Meyer, F.M., Ariza, A.A., & Zambrano, N. (2003).The scimitar, the dagger and the glove: Intercultural differences in the rhetoric of criticism in Spanish, French and English medical discourse (1930–1995). *English for Specific Purposes, 22* (3) 223–247.

[8] Hyland, K. (2001). *Disciplinary discourses: Social interactions in academic writing.* London: Longman.

[9] Santos, M.B.D. (1996). The textual organization of research paper abstracts in applied linguistics. *Text, 16,* 481-499.

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

[11] Rowley, J.E. (1988). *Abstracting and indexing, 2nd ed.* London: Bingley.

[12] Cross, C., & Oppenheim, C. (2005). A genre analysis of scientific abstracts. *Journal of Documentation, 62* (49), 428-446.

[13] Atkins, S., Clear, J., and Ostler, N. (1992). Corpus design criteria. *Literary and Linguistic Computing, 7*, 116.

[14] Tseng, F. (2011). Analyses of move structure and verb tense of research article abstracts in applied linguistics journals. *International Journal of English Linguistics, 1* (2), 27-39.

[15] Samraj, B. (2002). Introductions in research articles: Variations across disciplines. *English for Specific Purposes, 21,* 1-17.

[16] Lorés, R. (2004). On RA abstracts: from rhetorical structure to thematic organisation. *English for Specific Purposes, 23,* 280-302.

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

[18] Swales, J.M. (1990). *Genre Analysis.* Cambridge: Cambridge University Press.

[19] Nielsen, J. (1994). Guerrilla HCI: Using discount usability engineering to penetrate the intimidation barrier. In R. G. Bias & Mayhew, D. (Eds.) *Cost-justifying usability,* (pp.45-272). Boston: Academic Press.