Generic integrity: Visualizing lexicogrammatical features in scientific articles

Blake, John
University of Aizu, Aizu-Wakatsu, Japan

Problem

Publish in English or perish (Lillis and Curry, 2010, p.1) is used to describe the pressure faced by non-Anglophone university faculty. However, this academic pressure is no longer limited to faculty. At a public university in Japan, all computer science majors are required to submit a short research article written in English to fulfil their graduation requirement. The article must be factually accurate and suitably academic, for example, in terms of clarity and formality. However, the standards required in terms of originality, substance and rigour are less stringent than typically required for academic publications. Scientific research articles *per se* are a particularly challenging genre for novice writers due to their complexity and sophistication. The entry barrier for users of English as an additional language is especially high (Flowerdew, 2019) given the necessity to master both the research field and conform to the implicit generic expectations of the community of practice (Englander, 2014, p.81). Writing a research article is a huge hurdle for students who rarely function in English and have had little or no exposure to the target genre. To help fourth-year undergraduates rise to the challenge, a research writing course is offered. A key problem for the course teachers is providing examples and advice suitable for all registered students. This is because students write different types of research articles including mathematical proofs, software builds and controlled experimental studies.

Proposed solution

To address both problems, an online language feature visualizer was created to provide students and teachers access to exemplar research articles categorized by subgenre. As students use toggle buttons to reveal or hide lexicogrammatical features, awareness of the form and function of prototypical features should increase. Multimodal and bilingual explanations were added to allow students to select their preferred mode and medium of instruction. Students can, therefore, choose their own learning path, exploring and discovering the language features in research articles. Students focus on individual language features before seeing how all the features combine to create a coherent text (Nation, 2018). The noticing hypothesis proposed by Schmidt (2010) states that it is necessary to notice a language feature before learning it. Although there is debate about its empirical support, this theory explains the practice of many language teachers. Data-driven language learning (Flowerdew, 2015) encourages students to discover language features but is often plagued by logistic and technological difficulties. The user-friendly graphical user interface of visualization tool aims to solve this by making it simple for users to visualize the pre-selected features in the preloaded corpus of short research articles.

Development of visualizer

The articles were selected based on their subgenre, organizational structure, clarity and readability. At present, some prototypical language features need to be identified by human annotators and then rule-based parsing is used to visualize the features in the browser. This process of annotation, however, is time-consuming and impinges on the scalability of this tool. Where possible raw (non-annotated) text is searched to identify language features using either rule-based or probabilistic parsing. In line with the increasing trend to harnessing artificial intelligence in language analysis, machine and deep learning is used to reveal complex linguistic patterns (Manning, 2015). An online platform for texts, pattern-searching algorithms and explanatory materials was created (Blake, 2019). This version houses on short research articles in the domain of computer science, but the architecture is designed for multiple text types. Based on the results of an investigation of an in-house corpus of student-written research articles, four subgenres of research articles were identified, namely: practical, experimental, theoretical and empirical. Practical articles report the construction

of a product such as software or hardware, experimental articles describe controlled experiments, theoretical articles contain mathematical proofs while the empirical articles focus on software usability, user experience and accuracy. Four exemplar articles for each subgenre were selected for inclusion in the preloaded dataset. Articles were selected based on their accuracy, clarity and readability.

The lexicogrammatical features selected for inclusion in the prototype visualizer were chosen from those that are frequently referred to in the pedagogic literature and are included in the in-house research writing course. These features range from discoursal (e.g. sections and moves), to grammatical (e.g. tense and voice) and lexical (e.g. linking words). Some features, such as linking words and passive voice, can be identified by using language processing pipelines without the need for manual annotation. However, annotation is needed to identify features that cannot be discovered automatically with sufficient accuracy. To help students understand their form and function, colorization and labelling is used. Students can reveal or hide both language features and their associated explanations with a single click.

Discussion

The most striking difference among the four subgenres is the organisation. The section structure and rhetorical moves differ in each subgenre. Overall, however, there was little variation among the four types of research articles for the remaining language features. Although some functionalities rely on the texts to be pre-annotated before loaded into the database, where possible, scripts were written to automatically identify and visualize language features in raw text. The lack of necessity to annotate streamlines the procedure to expand the content and enhances the scalability of the visualizer. This pedagogic tool enables users to explore the form and function of lexicogrammatical features. Through exploring the visualizations and interacting with multimedia explanations, awareness of generic expectations is raised. The visualizer provides digital scaffolding to familiarize users with appropriate discoursal practices.

Future work

The initial idea was the exemplar research articles would need to be heavily annotated to enable the visualization of various features, such as voice and tense. However, the current aim is to minimize annotations and harness natural language processing pipelines to parse for language features more precisely. The existing content for the language feature visualizer is short computer science research articles and associated multimodal explanations. However, we plan to extend the system to include academic essays, specifically those written by science majors undertaking compulsory English language courses at universities in Hong Kong.

References

Blake, J. (2019). Annotated scientific text visualizer: Design, development and deployment. In Fanny Meunier, Julie Van de Vyver, Linda Bradley, and Sylvie Thouësny (Eds.), *CALL and complexity – short papers from EUROCALL 2019, (pp.45-50)*. UCLouvain, Belgium.
https://doi.org/10.14705/rpnet.2019.38.984

Englander, K. (2014). Writing and Publishing Scientific Research Papers in English: A Global Perspective. New York: Springer.

Flowerdew, L. (2015). Data-driven learning and language learning theories. In Leńko-Szymańska, A., & Boulton, A. (Eds.), *Multiple affordances of language corpora for data-driven learning, (pp.15-36).* Amsterdam: John Benjamins Publishing.

Flowerdew, J. (2019). The linguistic disadvantage of scholars who write in English as an additional language: Myth or reality. *Language Teaching, 52,* 249–260. https://doi.org/10.1017/S0261444819000041
Lillis, T., & Curry, M. J. (2010). *Academic writing in a global context: The politics and practices of publishing in English.* Abingdon: Routledge.

Manning, C. D. (2015). Last words: Computational linguistics and deep learning. *Computational Linguistics, 41*(4), 701-707. https://doi.org/10.1162/COLI_a_00239

Nation, P. (2018). Keeping it practical and keeping it simple. *Language Teaching, 51*(1), 138-146. https://doi.org/10.1017/S0261444817000349

Schmidt, R. (2010). Attention, awareness, and individual differences in language learning. In W. M. Chan, S. Chi, K. N. Cin, J. Istanto, M. Nagami, J. W. Sew, T. Suthiwan, & I. Walker, *Proceedings of CLaSIC 2010, Singapore, December 2-4 (pp. 721-737).* Singapore: National University of Singapore, Centre for Language Studies.