Automatic Creation and Discourse-Level Annotation of Individualized Discipline-Specific Corpora for the Data-Driven Learning (DDL) Classroom

Laurence Anthony (Waseda University, Japan) and Paul Thompson (University of Birmingham, UK)

Data-Driven Learning (DDL) has been increasingly used in the technical writing classroom as an effective method for introducing general- and discipline-specific language features to learners (Boulton 2012). One of the many strengths of DDL over traditional prescriptive approaches to writing instruction is that it allows learners to utilize or construct discipline-specific corpora that meet their unique target language needs. As a result, even within a heterogeneous class, each learner can discover the characteristic features of their own discipline, and through discussion and collaboration with others working with different corpora, learn about discipline specificity (Anthony, 2012), interdisciplinary features of writing (Bhatia, 2010), and general patterns of language use. However, DDL also suffers from a number of weaknesses. In particular, there is a general lack of discipline-specific corpora (Ädel, 2010). This means that the instructor or learner must construct an individualized corpus, but this requires locating, downloading, cleaning, and tagging relevant language data that is often beyond their technical skill level. In addition, searching within an individualized corpus for discourse-level language features is extremely limited due to the lack of any annotation.

In this research, the lack of availability of suitable discipline-specific corpora is addressed through the development of *AntCorGen*, a freeware corpus generation software tool that can automatically search, collect, clean, tag, and annotate very large discipline-specific corpora. The *AntCorGen* tool runs on all major operating systems. It is standalone and portable, and requires no installation or security permissions. Also, the tool is designed to be simple and easy to use without the need for instruction guides or tutorials. These features make it ideally suited for use by instructors in the creation of DDL class materials and by learners as part of in-class DDL learning tasks and activities.

AntCorGen utilizes the PLOS API [1], which provides access to the PLOS ONE multidisciplinary Open Access journal [2]. Through a user-friendly graphical interface to the API, users can search for relevant articles in the journal database using broad or very narrowly-defined parameters including subject category, keywords, date of publication, popularity, and type of article. The tool then automatically downloads relevant articles that match the search criteria and stores the data on the user's local file system in plain text and PDF formats. This allows the data to be immediately viewed in its published form or analyzed with traditional desktop corpus tools, such as *AntConc* (Anthony 2017). Using the metadata included with the articles returned by the API, *AntCorGen* can automatically divide the texts into sections and annotate this data accordingly, saving the data as separate files and placing them in folders that are named according to their headings. This function makes it simple for a user to search for language patterns within, for example, the title, abstract, introduction, materials/methods, results, or discussion sections of a research article.

To evaluate the utility of *AntCorGen*, 236 research articles in the area of "Human Mobility" and 126 articles in the area of "Temperate Forests" (both subject categories within PLOS ONE) were collected using the tool. Word lists and p-frame lists were then generated for each rhetorical section category in both subject areas using *AntConc* (Anthony, 2016) for the word list generation and *AntGram* (Anthony, 2017) for the p-frame generation. Using these lists, learners are able to investigate the linguistic features of each section of a research article through corpus analysis. For example, in the "Human Mobility" data, learners can find that the frame *<that the # of>* is relatively frequent in both the introduction and discussion sections, and the frame *<we # that the>* is relatively frequent in the "Temperate Forests" discussion sections. Such observations raise questions about why certain words and p-frames are more frequent in one section over another, which can lead learners to a deeper analysis of the relevant concordance data. In the case of *<we # that the>*, for example, learners can find that half of the instances are reports on what the researchers *found*, *demonstrated*, or *showed*, while the other half utilize present simple verbs and congregate around acts of proposing, hypothesizing and suggesting (e.g., *<we suggest that the>*). Having seen that *we* is used in the discussion section in this way, the learners can then investigate whether or not *we* is used in the same way in other sections of a research article.

In conclusion, the *AntCorGen* corpus generation tool presented here makes it possible for learners to create individualized discipline-specific corpora for use in DDL, which have the added facility of subdivision of the research articles into principal rhetorical sections.

[1] PLOS API. Information available at: http://api.plos.org/

[2] PLOS ONE. Information available at: htttp://journals.plos.org/plosone/s/journal-information

References

Ädel, A. (2010). Using corpora to teach academic writing: Challenges for the direct approach. In M. C. Campoy-Cubillo, B. Belles-Fortuño & M. L. Gea-Valor (Eds.), *Corpus-based approaches to English language teaching* (pp. 18-35). London: Continuum.

Anthony, L. (2012). Products, processes, and practitioners: A critical look at the importance of specificity in ESP. *Taiwan International ESP Journal*, 3(2), 1-18.

Anthony, L. (2016). *AntConc* [Computer Software]. Tokyo, Japan: Waseda University. Available online at http://www.laurenceanthony.net/software.

Anthony, L. (2017). *AntCorGen* [Computer Software]. Tokyo, Japan: Waseda University. Available online at http://www.laurenceanthony.net/software.

Anthony, L. (2017). *AntGram* [Computer Software]. Tokyo, Japan: Waseda University. Available online at http://www.laurenceanthony.net/software.

Bhatia, V. K. (2010). Interdiscursivity in Professional Communication. *Discourse and Communication*, 21(1), 32-50.

Boulton, A. (2012). Corpus consultation for ESP: A review of empirical research. In A. Boulton, S. Carter-Thomas, and E. Rowley-Jolivet (Eds.), *Corpus-informed research and learning in ESP: Issues and Applications* (pp. 261-291). Amsterdam, Netherlands: John Benjamins.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge, UK: Cambridge University Press.