Naturalistae 17: 59-62 (Feb. 2013)

Original paper

An instant electronic dictionary of English collocations in natural history realized using a concordancer and open resources[1]

Hiroyuki TAKASAKI[2]

コンコーダンサーとコーパスで作る即席「自然誌英語活用電子辞典」[1]

高崎浩幸[2]

Abstract: In a short amount of time, the author actualized on his PC a personal electronic dictionary of English word collocations in natural history—the nursery of natural science. It was made of a corpus database of plain text files to be analyzed with a concordancer. To make the corpus, free open documents related to natural history in various fields were collected using the Web. A free concordancer was downloaded via the Internet. This article explains how to make a usable and redistributable "starter" corpus for beginners, who are not familiar with usage (word collocations in particular) of the English language in natural history. The prepared corpus is modifiable by adding data files if desired, and by removing inappropriate ones. The dictionary thus made will enhance the user's command of the trade language in natural history. A similar electronic dictionary could be likewise created in any field of science.

Keywords: natural history, collocation, concordancer, corpus, dictionary

No commercial dictionary of the English language satisfies a student of natural history who is not a native speaker of English. Unfortunately he (sensu "he/she" as well as anywere else throughout this article where applicable) is in a world where English has become the de facto lingua franca in the global community of scientists. When he writes a paper in English using technical terms in his field, it happens too often that his dictionaries hardly give him any good advice on the usage of the terms in detail. For example, none of the dictionaries in the author's bookshelf answers the following questions. Which prepositions follow the anatomical term "proximal" and "distal"? When they are used attributively, what are the most common nouns that follow them in morphological descriptions? Which prepositions follow "medial" and "lateral"? Not only the editorial board members and reviewers of *Naturalistae* but also the majority of colleagues for whom English is not their mother tongue are likely troubled by similar questions.

Even when papers are allowed to be written in the contributors' mother tongue, most scientific journals, including *Naturalistae*, require their titles and abstracts to be written also in English. This often handicaps both the contributors and the editorial staff with poor skills in the language. Because of this, sometimes, journals may suffer from shortages of submitted papers. In the worst case, they are forced due to limited time schedules for publication, to end up with nonsensical lines of English word-like character strings. Thus they may fail in providing titles and abstracts in English serviceable for scientific communication.

Such situations must be terminated. This article aims to provide a step forward to solve the problem, by presenting a method to build a custom electronic dictionary of English collocations tuned for natural history in particular. The dictionary consists of two components, (1) corpus data files ("corpus" for short), and (2) concordance software ("concordancer") to analyze the corpus for extracting collocation information, etc. on a

1．The author previously prepared a prototype corpus in a limited range of natural history, and has reported its usability elsewhere in Japanese (Takasaki, H. 2012. An instant private-edition electronic dictionary of English collocations in butterfly studies. *Butterflies* (Teinopalpus), 61: 48-51).

2．Department of Zoology, Faculty of Science, Okayama University of Science 1-1 Ridai-cho, Okayama-shi, Okayama-ken 700-0005, Japan.
〒700−0005 岡山市北区理大町1−1 岡山理科大学理学部動物学科．E-mail to: takasaki@zool.ous.ac.jp

given search term. Both components are obtained via the Internet. Therefore, any reader connected to the Web can make one. Component 1, corpus, is the harder part to make, and will be described in some detail later. Component 2, concordancer, is downloaded via the Internet.

As to the concordancer, the author's recommendation is AntConc (http://www.antlab.sci.waseda.ac.jp/software.html), although any other concordancer is probably equipped with similar functions. The reasons for his choice of AntConc are multiple. This well-made concordancer is free, and provides versions not only for Microsoft Windows but also for Macintosh OS X and Linux. Besides the official manual contained in the downloaded package, some guides are readily available in the Web (search with "AntConc use guide") including videos in YouTube. Consult them for how to use AntConc in detail. In short, the concordance menus mostly suffice for ordinary checks of word collocations (Fig. 1). Put in "Search Term," click "start" button, and wait until the analysis gets "FINISHED!" At this stage, click "sort" to align the word collocation lines as specified by the sort options. Then, collocations sorted according to the options of the search term will be given in the window, along with the source file names and some statistics. By scrolling, all the collocations of the search term contained in the corpus can be checked.[3]

To build the corpus, which is simply a set of texts stored in a directory (often further divided into subdirectories for categorization), it is only required to collect text files (to be in UTF-8 coding for use with AntConc) related to natural history. However, to make a corpus redistributable to colleagues and friends, the source text files must be in the public domain. To save labor and time, let us pygmies stand on the shoulders of giants.

The first site chosen for collecting text data for the corpus is Project Gutenberg (http://www.gutenberg.org/wiki/Main_Page [URLs are as of January 2013 throughout this article]). Appropriate titles in English in Plain Text UTF-8 are collected from the following categories and subcategories in the site: Animal, Astronomy, Ecology, Forestry, Geology, Botany, Cytology, Horticulture, Microbiology, Microscopy, Mycology, Natural History, Physiology, Scientific American, and Zoology.

There are three notable merits in collecting corpus data from Project Gutenberg. (1) Almost all texts are redistributable, for academic and non-commercial use, in countries where the copyright expires after the same number of years as in the US or earlier. (2) The large number of texts are classified by category and subcategory for easy location of files by topics of interest. (3) The texts contain only small numbers of typos thanks to repeated proofreading before their release. The only shortcoming is that the texts are rather old. Note, however, old-fashioned usage of the language is infinitely better to learn than nonsensical jumble of words.

The second collection site is Wikipedia in English (http://en.wikipedia.org/wiki/Main_Page). Open appropriate articles, and save them (copy and paste from the browser window to any text editor [e. g. Notepad]) as text files in UTF-8 coding with new names. Then, place them in the corpus directory or appropriate subdirectories. The corpus made up to this point is legally redistributable.

The third collection site is the Biodiversity Heritage Library (http://www.biodiversitylibrary.org/). Browse the subjects (http://www.biodiversitylibrary.org/browse/subject) for appropriate titles. Many entries of this site are in the public domain as in the Project Gutenberg. Unfortunately, however, clean text files are not available, but only OCR output files with names ending with _djvu.txt. They have to be proofread and corrected before being added to the corpus. Although proofreading is time-consuming and tiring, it also trains the proofreader's command of the language. In some titles published in the 21st century, clean text files may be obtained by conversion from .pdf files. However, addition of those whose copyright status is unclear may make redistribution of the corpus illegal. Such files are separately saved for later addition to the corpus limited to private use. Only those judged legally safe for redistribution are added to the redistributable "starter" corpus.

Finally, GNU Free Documentation License (http://www.gnu.org/licenses/fdl-1.3.txt), as a reminder for redistribution as well as a corpus data

3. Even without any concordancer available for the OS on some PC, the corpus is usable for checking word collocations if the OS has a Unix-based terminal mode. First open the terminal, move to the directory which contains the corpus, and run the following command:

```
grep $1 --color=always -R . | less -R
```

where $1 denotes the search term. Then the result somehow like in Fig. 1 but before sorting is returned. Try also:

```
grep "$1 .* $2" --color=always -R . | less -R
```

where $1 and $2 denotes the search term one and two which are guessed to collocate with each other.

4. "Free" here does not mean to permit free redistribution as their copyrights have not perished yet.

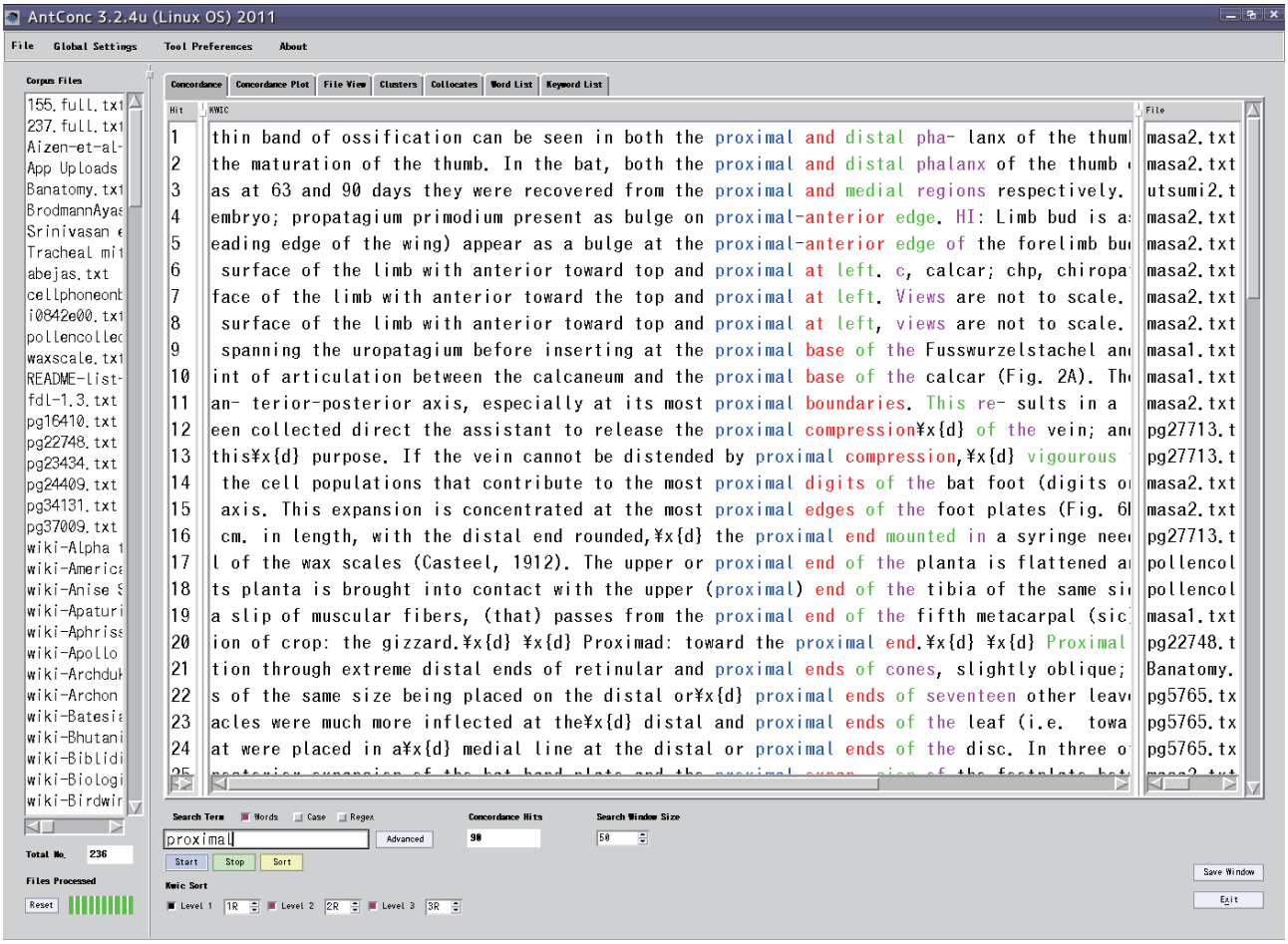

Fig. 1. Collocation examples of "proximal" extracted from the prepared natural history corpus using AntConc.

file, is added to the "starter" corpus. This corpus may be copied and shared among colleagues and friends.

In the actual use of the electronic collocation dictionary, the user may add any appropriate text files to the "starter" corpus. Also he may remove any inappropriate ones. Many of the contemporary papers in scientific journals distributed as pdf files are convertible to text files, and may be added to it as well to enhance its usability. Many papers that are in .pdf format, some free[4] while others charged, can be searched and downloaded via Google Scholar.

The realized electronic collocation dictionary gives the answers to the questions asked in the opening paragraph of this article. "Proximal" and "distal" are followed by "to." And they are often combined in the phrase containing the noun "end," which is further often followed by "of," viz. "proximal/distal end of...." "Medial" is hardly followed by any preposition. "Lateral" is sometimes followed by "along," and less often by "on." By using a modified version with additional corpus data files from reproductive physiology, the author, being not a physiologist, learned in 2012 the following facts about "androgen" and "testosterone." The singular form "androgen" is rarely used as the noun but often used adjectivally, while the plural "androgens" are often used as the noun. And "testosterone" is one of the "androgens." These are not readily found in any commercial dictionary.

This custom dictionary, a unique variant of which anyone can make, is far better than any online or web concordancer. It will help a novice student of natural history whose skill in English may be poor, and its process in the making as well will train him in the language. Eventually it will lighten the burdens of editorial staff of various journals including *Naturalistae*. A person who wishes to make a similar custom electronic dictionary in any field of science other than natural history may be advised to reread this article by substituting "natural science" for "natural history." It is only required to replace the corpus with one appropriate for the particular field.

Acknowledgements: The students who attended the author's socioecology seminar in 2012 helped him to conceive the idea that has been realized here. They also helped him in testing the actualized dictionary for its usability with additions of the text data files which they collected while they read for their own study. Project Gutenberg, Wikipedia, and the Biodiversity Heritage Library were indispensable giants whose shoulders he, only a pygmy of a scholar, was kindly allowed to stand on. Lukas Bonick kindly read the manuscript to eliminate linguistic errors. To these people and institutions, he is grateful.

高崎浩幸：コンコーダンサーとコーパスで作る即席「自然誌英語活用電子辞典」

要約

短時日で個人用の「自然誌英語活用電子辞典」をパソコン上に現実した．その構成は，コーパス・データベース（プレーン・テキスト・ファイルからなる）とそれを分析するコンコーダンサーである．コーパスには，自然科学を育んできた，さまざまな自然誌分野に関連したフリーの公開文書を，ウェブ上で集めた．さらにフリーのコンコーダンサーをインターネット経由でダウンロードした．本稿は，自然誌での英語の用法（とくに連語の用法）に不慣れな初学者が使えて，なおかつ再配布可能な「スターター」となるコーパスを製作する方法を説明する．必要に応じてデータ・ファイルを追加することによって，また不適当なものを取り除くことによって，このコーパスは修正・拡張可能である．このようにして自作する活用辞典を使えば，使用者の自然誌における英語力を増強することができる．類似の電子活用辞典は科学のどのような分野でも作成することができるだろう．

（Accepted 7 January 2013）