Data in Brief 16 (2018) 94–100

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

A corpus and a concordancer of academic journal articles

Deny A. Kwary

English Department, Faculty of Humanities, Universitas Airlangga, Jl. Dharmawangsa Dalam, Surabaya 60286, Indonesia

ARTICLE INFO

Article history:
Received 29 August 2017
Received in revised form
3 November 2017
Accepted 6 November 2017

ABSTRACT

This data article presents a corpus (i.e. a selection of a big number of words in an electronic form) and a concordancer (i.e. a tool to show the word in its context of use) of academic journal articles. As the title suggests, the data were collected from research articles published in academic journals. The corpus contains 5,686,428 words selected from 895 journal articles published by Elsevier in 2011–2015. The corpus is classified into four subject areas: Health sciences, Life sciences, Physical Sciences, and Social Sciences, following the classifications of Scopus, which is the largest abstract and citation database of peer-reviewed scientific journals, books and conference proceedings. To ease the access and utilization of the corpus, a program to produce the key word in context (KWIC) and word frequency was created and placed on the website: corpus.kwary.net. The corpus is a valuable resource for researchers, teachers, and translators working on academic English.

Specifications Table

Subject area	*Linguistics, Education, English Language*
More specific subject area	*Corpus Linguistics, Adult Education, English Language Teaching, English Language Learning, Academic Writing*
Type of data	*Texts and tables*

E-mail address: d.a.kwary@fib.unair.ac.id

https://doi.org/10.1016/j.dib.2017.11.023

How data was acquired	*Select the articles that meet the pre-determined criteria, download the articles, convert the texts into a.txt format, clean the corpus, and create the concordancer using PHP and MySQL.*
Data format	*Raw and processed*
Experimental factors	*The journals were classified into four subject areas: Health sciences, Life sciences, Physical Sciences, and Social Sciences, following the classifications of Scopus. The pre-determined criteria for the selection are: (1) The journals do not appear in more than one subject area, (2) the journals have a 5-year impact factor, (3) the articles are published in 2011 and 2015, (4) the articles are written in English, and (5) the articles are open access.*
Experimental features	*The corpus comprises well-selected and recent journal articles. The concordancer enables the search of particular words to determine its context of use.*
Data source location	*The data were gathered from the website:* sciencedirect.com
Data accessibility	*Free online access at corpus.kwary.net*

Value of the data

- Corpus linguists and English lecturers will be able to compare the results derived from these corpus and concordancer with the data found in other studies or textbooks.
- Researchers from the four subject areas (Health sciences, Life sciences, Physical Sciences, and Social Sciences) will be able to know the particular words which are frequently used in the recent journal articles in their own subject areas.
- Researchers from the four subject areas will be able to know how extensive a particular term or topic has been written in the recent journal articles in their own subject areas.
- Teachers of academic English will be able to know the common collocates of a particular word when it is used in journal articles.
- Translators will be able to select the right words and collocates when translating a particular word from a source language to English.

1. Data

The corpus comprises 5,686,428 words, classified into four subject areas: Health sciences, Life sciences, Physical Sciences, and Social Sciences, following the classifications of Scopus. The words were compiled from 895 journal articles published by Elsevier in 2011–2015. To ease the access and utilization of the corpus, a concordancer to produce the key word in context (KWIC) and a word frequency tabulation were created and placed on the website: corpus.kwary.net. When we open that website, we will be able to search for a word or several words to see the KWIC which is the position of the word together with its collocates (up to five word tokens to the left and five tokens to the right). There is also a link on the left-hand top of the page to access the frequency list, i.e. corpus.kwary.net/freq. This is useful to make a word list of the words found in each discipline.

2. Experimental design, materials and methods

A corpus as a collection of pieces of language text in electronic form, selected to represent a language or language variety [1]. A corpus enables the elaboration of better quality learner input and provides researchers and teachers with a wider, finer perspective into language in use [2]. The corpus available here is expected to be a resource for determining the behaviour of the words used in research articles published in international journals. Consequently, the data were selected from international journal articles. However, it is necessary to note that the behaviour or the usage of a