A list of **tools and resources that facilitate Data-Driven Learning (DDL)**, to be used as a complement to the Cambridge Element:

Pérez-Paredes, P., & Boulton, A. (2025). *Data-driven Learning in and out of the language classroom.* Cambridge University Press.

Sections

Collocations

Collocates Data: list of collocates of English, about 13.5 million pairs.
https://www.collocates.info

Hyper Collocation: finds example sentences from almost 1 million English papers in arXiv.
https://hypcol.marutank.net

Netspeak: provides easy-to-use complex searches of word use in English and German, providing relevant frequency information and examples from Google Books.
https://netspeak.org

Oxford Online Collocation Dictionary: uses the British National Corpus to generate common word combinations in English.
https://www.freecollocation.com

Sketch Engine for Language Learning (SKELL): provides collocations and syntactic patterning of common word combinations in English, German, Italian and other languages.
https://skell.sketchengine.eu

StringNet: searches for collocations and patterns in English.
http://nav.stringnet.org

Webcorp: searches the web as a corpus.
https://www.webcorp.org.uk/live

Corpus-assisted writing

ColloCaid: supports writers with collocations that are typical of written academic English across a wide range of disciplines.
https://collocaid.uk

Corpus software

AntConc: a freeware corpus analysis toolkit for concordancing and text analysis.
https://www.laurenceanthony.net/software/antconc

CasualConc: generates KWIC concordance lines, word clusters, collocation analysis, and word count.
https://sites.google.com/site/casualconc

LancBox X: a powerful tool for analysing millions or billions of words.
https://lancsbox.lancs.ac.uk

Sketch Engine: very fast with large corpora, with many built-in for different languages, and the innovative word sketch function.
https://www.sketchengine.eu

TextSTAT 3: works with different formats, including online texts.
https://neon.niederlandistik.fu-berlin.de/en/textstat

N-gram tools

Google Books Ngram Viewer: shows the frequency and evolution of words and phrases from millions of books over time.
https://books.google.com/ngrams

Online NGram Analyzer: input text and output from single items to 5-grams.
http://guidetodatamining.com/ngramAnalyzer

Online corpora

BNC: see English-Corpora.

COCA: see English-Corpora.

Corpus del Español del Siglo XXI (CORPES): 21st-century corpus of Spanish, from the Royal Spanish Academy.
https://apps2.rae.es/CORPES

CEDEL 2: written Spanish learner language corpus (version 2).
http://cedel2.learnercorpora.com

Corpus de Aprendices de Español (CAES): corpus of Spanish learner language.
https://galvan.usc.es/caesv20/search

Corpus del Español: a collection of Spanish language corpora.
https://www.corpusdelespanol.org

CorpusMate: simplified language data analysis experience for younger learners of English.
https://corpusmate.com

Corpus of Study for Contemporary French (CEFC): a collection of spoken and written corpora for French.
https://repository.ortolang.fr/api/content/cefc-orfeo/11/documentation/site-orfeo/index.html

English-Corpora: probably the easiest-to-access collection of English language corpora including the British National Corpus (BNC), the Corpus of Contemporary American English (COCA), the News on the Web Corpus (NOW), the TV corpus, the Movie Corpus, the Corpus of American Soap Operas, the Time magazine Corpus or, among others, the Coronavirus Corpus.
https://www.english-corpora.org

MICASE: the Michigan Corpus of Academic Spoken English, a collection of transcripts of academic speech events recorded at the University of Michigan.
https://quod.lib.umich.edu/cgi/c/corpus/corpus?c=micase;page=simple

MICUSP: the Michigan Corpus of Upper-Level Student Papers, a collection of 'good' student writing.
https://micusp.elicorpora.info/main

TED Corpus Search Engine: offers searchable transcripts of some 5,000 TED Talks on a variety of topics.
http://yohasebe.com/tcse

Pedagogic corpora

SACODEYL: interviews with British, French, German, Lithuanian, Rumanian and Spanish teenagers between 13 and 18 years of age, using annotated XML corpora.
https://www.perezparedes.es/sacodeyl-xml-corpora

Phrase banks and grammar patterns

Collins Dictionary Grammar Patterns: lists grammar patterns used in English, and all the words regularly used with a given pattern.
https://grammar.collinsdictionary.com/grammar-pattern

The Manchester Academic Phrasebank: provides the phrases used for various functions in academic writing.
https://www.phrasebank.manchester.ac.uk

Resources for language learning and teaching

British Academic Written English Quicklinks: a database of concordance lines chosen for learners of Academic English.
https://bawequicklinks.coventry.domains

Compleat Lexical Tutor: dozens of online tools for learning English and French language learning and develop DDL activities.
https://www.lextutor.ca

English Grammar Profile: allows language teachers and researches to see how language learners develop competence in grammatical form and meaning, as well as pragmatic appropriateness, as they move up the Common European Framework of Reference for languages (CEFR) levels.
https://www.englishprofile.org/english-grammar-profile/egp-online

English Vocabulary Profile: offers corpus-based information about which words, word senses and phrases are known and used by learners at each level of the Common European Framework of Reference for languages (CEFR).
https://www.englishprofile.org/wordlists/evp

Playphrase: search for a word or phrase from films and TV shows, to compare pronunciation in context.
https://www.playphrase.me

WebCorp Learn: context-based inductive English language learning through exploratory experimentation.
https://www.webcorp.org.uk/wcx/learn and
https://bawequicklinks.coventry.domains/encyclopedia

Youglish: provides examples of words and phrases from YouTube videos – a kind of multimedia concordancer for spoken English and other languages.
https://youglish.com

Text exploration tools

Online Text Comparator: compare the words and word frequency in two texts.
http://guidetodatamining.com/ngramAnalyzer/comparator.php

Corkpit: a multipurpose corpus analysis toolkit.
https://interrogator.github.io/corpkit

Corpus Tools: a collection of easy-to-use tools for processing text.
https://corpus.tools

JustText: remove boilerplate from any URL and just keep the text.
https://nlp.fi.muni.cz/projects/justext

Pérez-Paredes, P. and Boulton, A. (2025). *Data-driven Learning in and out of the language classroom.* Cambridge University Press.

Laurence Anthony's website: a range of tools designed to facilitate all stages of text analysis, from compiling and editing.
https://www.laurenceanthony.net/software.html

Text Analyzer: identifies the level of a text according to the Common European Framework or Reference for language (CEFR).
http://www.roadtogrammar.com/textanalysis

Versatile: word cloud generator, concordancer and vocabulary profiler.
https://www.versatile.pub/versatext-info.html

Word lists

Academic Word List (AWL): a list of 570 headwords and 3000 words in total, from Averil Coxhead's AWL.
https://www.uefap.com/vocab/select/awl.htm

AWL highlighter: highlight words in the Academic Word List (AWL) in your own papers:
https://www.nottingham.ac.uk/alzsh3/acvocab/awlhighlighter.htm

AWL Gapmaker: produces gap-fill exercises for academic texts.
https://www.nottingham.ac.uk/alzsh3/acvocab/awlgapmaker.htm

Webcorp wordlist tool: input a text or URL and generate a wordlist.
https://www.webcorp.org.uk/live/wdlist.jsp

Word lists in Oxford Learner's Dictionaries: various lists from OUP, including he 3000 most important words to learn in English, AWL, etc.
https://www.oxfordlearnersdictionaries.com/wordlist