

# Addressing the Challenges of Data-Driven Learning Through Corpus Tool Design

## An Introduction to AntConc 4

### Laurence Anthony

Center for English Language Education in Science and Engineering (CELESE)

Faculty of Science and Engineering, Waseda University

<http://www.laurenceanthony.net/>

[anthony@waseda.jp](mailto:anthony@waseda.jp)

 [@antlabjp](https://twitter.com/antlabjp)



April 29 (Fri), 2022 (Univ. of Queensland - Online)

International Perspectives on Corpus Technology for Language Learning - Seminar Series

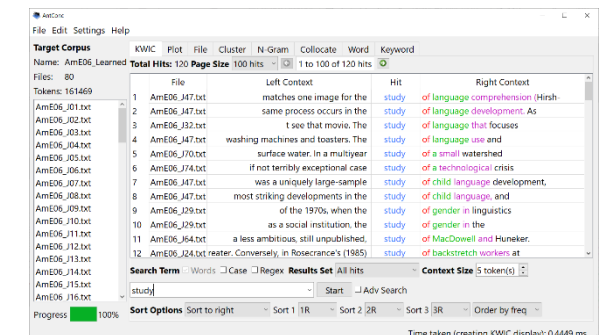


# Overview

- Understanding DDL
  - definitions and strengths
- Challenges in DDL
  - Challenges for the 'direct approach'
  - Challenges resulting from technical developments
- Introduction to AntConc 4
  - evolution over 20 years
  - design concept, features and functions
  - addressing the challenges of DDL



<https://simplelivingaustralia.com.au/wp-content/uploads/2016/02/financial-challenges-for-seniors-750x375.jpg>



---

# Understanding Data-Driven Learning (DDL)

definitions and strengths

---



# Understanding Data-Driven Learning (DDL)

DDL as an example of "Data Science"?

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to **extract knowledge and insights from noisy, structured and unstructured data**, [1][2] and **apply knowledge and actionable insights** from data across a broad range of application domains.

[https://en.wikipedia.org/wiki/Data\\_science](https://en.wikipedia.org/wiki/Data_science)



**WIKIPEDIA**  
The Free Encyclopedia

# Understanding Data-Driven Learning (DDL)

DDL as an example of "Active learning"?

"Active learning is a process whereby students engage in activities, such as reading, writing, discussion, or problem solving that promote analysis, synthesis, and evaluation of class content." (p. 223)

J. Engr. Education, 93(3), 223-231 (2004)

Michael J. Prince  
Professor of chemical engineering  
Bucknell University, US



# Understanding Data-Driven Learning (DDL)

DDL as an example of "Active learning"?

- Active learning overlaps with...
  - collaborative learning
    - "students work together in small groups toward a common goal"
  - cooperative learning
    - "students pursue common goals while being assessed individually"
  - problem-based learning
    - "relevant problems are introduced at the beginning of the instruction cycle and used to provide the context and motivation for the learning that follows"

Michael J. Prince  
Professor of chemical engineering  
Bucknell University, US



# Understanding Data-Driven Learning (DDL)

DDL as an example of "Active learning"?

Active learning works!

Reference	Learning Outcome	Effect Size
Johnson, Johnson and Smith [12]	Improved academic achievement	0.64
	Improved quality of interpersonal interactions	0.60
	Improved self-esteem	0.44
	Improved perceptions of greater social support	0.70
Johnson, Johnson and Smith [13]	Improved academic achievement	0.53
	Improved liking among students	0.55
	Improved self-esteem	0.29
	Improved perceptions of greater social support	0.51
Springer et al. [43]	Improved academic achievement	0.51
	Improved student attitudes	0.55
	Improved retention in academic programs	0.46

Table 1: Collaborative vs. individualistic learning

J. Engr. Education, 93(3), 223-231 (2004)

# Understanding Data-Driven Learning (DDL)

A simple definition of DDL

Language data science + Active Learning → Data-Driven Learning (DDL)

"learning how to **create, search, analyze, and interpret general and specialized language databases** (corpora)"

Anthony, L. (2016)



Learners as researchers or 'detectives' (John, 1991:30)  
(teachers as 'facilitators' of that research)



Meta analysis of 205 DDL empirical studies (Boulton & Cobb, 2017)  
Effect size = **0.95** (control/experimental design)  
Effect size = **1.50** (pre/post design)

# Understanding Data-Driven Learning (DDL)

## Strengths of DDL

- analyzing corpora provide learners with...
  - a rich exposure to "real language"
  - an alternative to the observations of 'native-speaker' informants
  - a wide range of cognitive skills  
[e.g., predicting, observing, noticing, thinking, reasoning, analyzing, interpreting, reflecting]
- interpreting corpora can...
  - raise the learners' contextual and linguistic awareness
  - reveal patterns and forms that are not obvious or visible in other language sources  
(e.g., dictionaries)
- working collaboratively with others helps learners to...
  - "learn how to learn" (Gee et al., 1996: 165)
  - improve their self-esteem and ability to interact with others (Prince, 2004)



# Challenges in Data-Driven Learning (DDL)

Challenges for the 'direct approach'

Challenges resulting from technical developments



# Challenges in Data-Driven Learning (DDL)

## Challenges for the 'direct approach'



Adel, 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.

# Challenges in Data-Driven Learning (DDL)

## Challenges for the 'direct approach'

- finding suitable target/discipline-specific corpora  
[online corpora are limited]
- creating custom corpora  
[personal corpora can be challenging and time-consuming to build]
- knowing...
  - what to search for in a corpus
  - how to manage 'data-overload' when analyzing results
  - how to incorporate results from corpora in learner language
- avoiding...
  - a focus on surface-level word/phrase patterns
  - a focus on technology over language

Adel, 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.



# Challenges in Data-Driven Learning (DDL)

## Challenges for the 'direct approach'



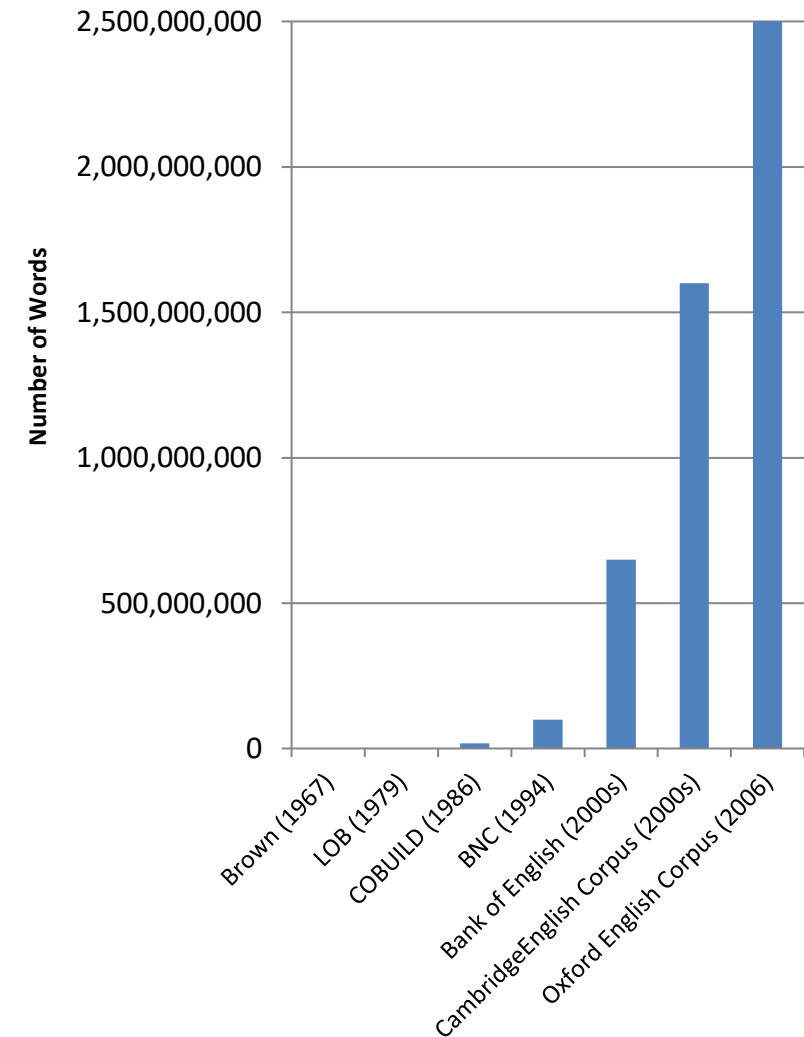
Jablonkai, R. J. (2022). **Transforming language teaching with corpora - Bridging research and practice**. International Perspectives on Corpus Technology for Language Learning - Seminar Series. Univ. of Queensland: Online.

- Learners might find it **technically challenging**
- **Time-consuming**
- Learners might feel **overwhelmed by the amount of data**
- Better suited for learners with **proficiency from an intermediate level** (BUT lower level learners benefit as well) (Boulton, 2009)
- Technology in language learning - **computer-anxiety** (Ortega, 1997; Sullivan & Pratt, 1996)

# Challenges in Data-Driven Learning (DDL)

Challenges resulting from technical developments - corpus sizes

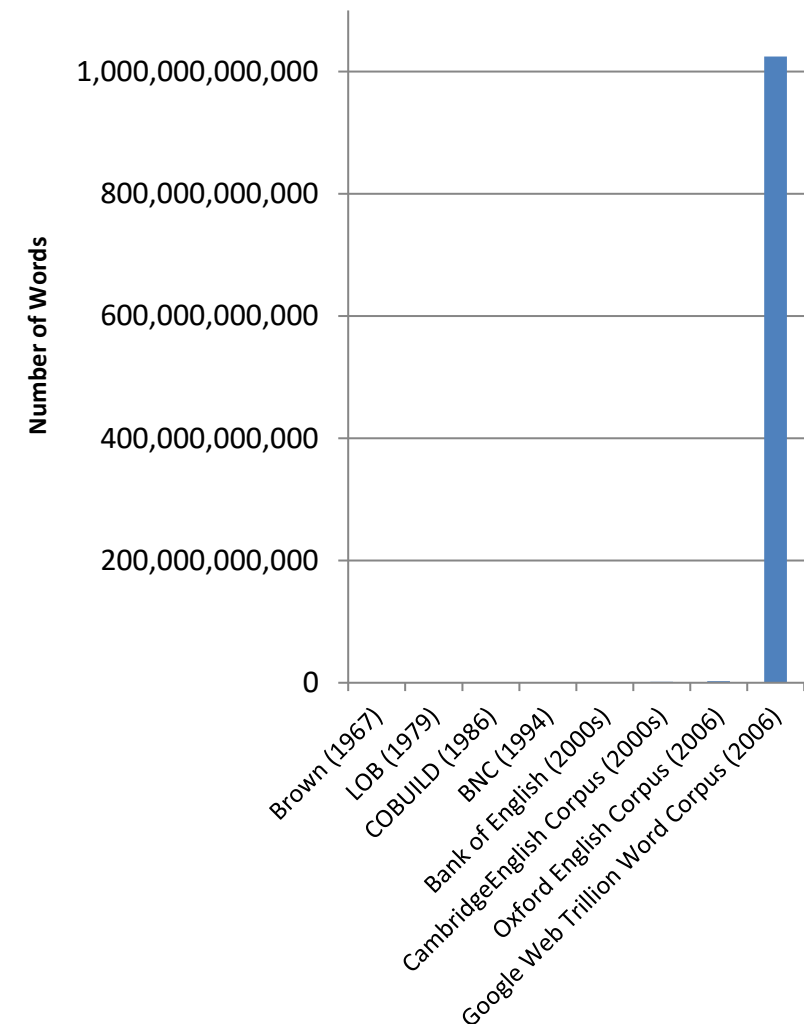
- 1960s-1970s
  - Brown Corpus (1964) - 1 m
  - LOB Corpus (1978) - 1 m
- 1980s-1990s
  - COBUILD Corpus (1986) - 18 m
  - British National Corpus (1994) - 100 m
- 2000s
  - Bank of English (2008) - 650 m
  - Cambridge English Corpus (2013) - 1.6 b
  - Oxford English Corpus (2013) - 2.5 b
  - Google Web Trillion Word Corpus (2006) - 1 t



# Challenges in Data-Driven Learning (DDL)

Challenges resulting from technical developments - corpus sizes

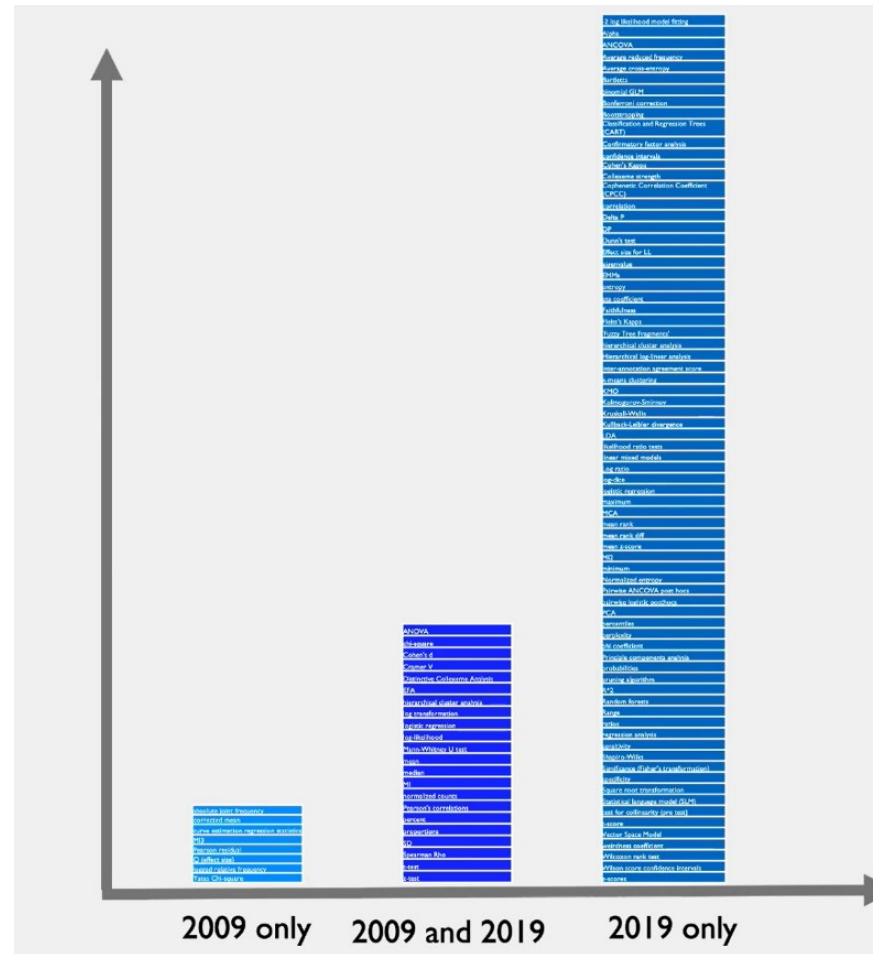
- 1960s-1970s
  - Brown Corpus (1964) - 1 m
  - LOB Corpus (1978) - 1 m
- 1980s-1990s
  - COBUILD Corpus (1986) - 18 m
  - British National Corpus (1994) - 100 m
- 2000s
  - Bank of English (2008) - 650 m
  - Cambridge English Corpus (2013) - 1.6 b
  - Oxford English Corpus (2013) - 2.5 b
  - Google Web Trillion Word Corpus (2006) - 1 t



# Challenges in Data-Driven Learning (DDL)

Challenges resulting from technical developments - statistical complexity

No. of statistical methods used in linguistic analyses



Larsson, T., Egbert, J. & Biber D. (2020) Do Corpus Linguists Focus on Statistics at the Expense of Linguistic Analysis? A Ten-year Perspective. ICAME 41.

# Challenges in Data-Driven Learning (DDL)

Challenges resulting from technical developments - statistical complexity



Larsson, T. (2022). **On the perils of opaque measures and methods: Toward increased transparency.** International Perspectives on Corpus Technology for Language Learning - Seminar Series. Univ. of Queensland: Online.

- Toward increased accuracy and transparency
- Toward linguistic interpretability
  - “Perhaps the most serious risk to researchers using available tools is that **many of the quantitative measures provided by corpus analysis software do not have transparent linguistic interpretations.**”

---

# Introduction to AntConc 4

evolution over 20 years; design concept, features and functions;  
addressing the challenges of DDL

---



# Introduction to AntConc 4:

## A member of the AntLab tools family

Laurence Anthony's Website

Home Resume Publications Software Classes Photo Albums Links Contact

**Software**

**AntConc**  
A freeware corpus analysis toolkit for concordancing and text analysis.  
[AntConc Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (3.4.4)
- Macintosh OS X 10.7-10.10 (3.4.3)
- Macintosh OS X 10.6 (3.4.1)
- Linux (3.4.3)
- Other versions

**AntPConc**  
A freeware "parallel" corpus analysis toolkit for concordancing and text analysis using UTF-8 encoded text files.  
[AntPConc Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (1.1.0)
- Macintosh OS X (1.1.0)
- Other versions

**AntWordProfiler**  
A freeware tool for profiling the vocabulary level and complexity of texts.  
[AntWordProfiler Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (1.4.0)
- Macintosh OS X (1.4.1)
- Linux (1.3.1)
- Other versions

**AntFileConverter**  
A freeware tool to convert PDF and Word (DOCX) files into plain text for use in corpus tools like AntConc.  
[AntFileConverter Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (1.2.0)
- Macintosh OS X (1.2.0)
- Other versions

**AntMover**  
A freeware tool structure (moves) analysis program.  
[AntMover Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (1.0.0)
- Other versions

**AntCLAWSGUI**  
A front-end interface to the CLAWS tagger developed at Lancaster University, UK. Note that you must have CLAWS installed before you can use AntCLAWSGUI. See the help file.  
[AntCLAWSGUI Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (1.1.0)
- Other versions

**EmbedAnt**  
A freeware tool for detecting and converting character encodings.  
[EmbedAnt Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (1.2.0)
- Windows 64-bit (1.2.0)
- Macintosh OS X (1.2.0)
- Other versions

**FileAnt**  
A freeware social media and data analysis toolkit (developed in collaboration with Claire Harber of Lancaster University).  
[FileAnt Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (1.1.1)
- Windows 64-bit (1.1.1)
- Macintosh OS X (1.1.1)
- Other versions

**FindAnt**  
A freeware prototypical text analysis tool (developed in collaboration with Paul Baker of Lancaster University).  
[FindAnt Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (1.2.0)
- Windows 64-bit (1.2.0)
- Macintosh OS X (1.2.0)
- Other versions

**SearchAnt**  
A freeware batch search and replace tool.  
[SearchAnt Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (1.1.0)
- Windows 64-bit (1.1.0)
- Other versions

**SegmentAnt**  
A freeware Japanese and Chinese segmenter (segmentation/tagging tool).  
[SegmentAnt Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (1.1.0)
- Macintosh OS X (1.0.0)
- Linux (1.0.0)
- Other versions

**TagAnt**  
A freeware Part-Of-Speech (POS) tagger built on TreeTagger (developed by Martin Schmalz).  
[TagAnt Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (1.2.0)
- Windows 64-bit (1.2.0)
- Macintosh OS X (1.2.0)
- Linux (1.1.2)
- Other versions

**ValidateAnt**  
A freeware spelling checker analysis program.  
[ValidateAnt Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (1.0.0)
- Other versions

## Data Collection Tools



AntCorGen



AntFileConverter



AntFileSplitter



FireAnt

# Introduction to AntConc 4:

## A member of the AntLab tools family

Laurence Anthony's Website

Home Resume Publications Software **Classes** Photo Albums Links Contact

Software

**AntConc**  
A freeware corpus analysis toolkit for concordancing and text analysis.  
[AntConc Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (3.4.4)
- Macintosh OS X 10.7-10.10 (3.4.3)
- Macintosh OS X 10.6 (3.4.1)
- Linux (3.4.3)
- Other versions

**AntPConc**  
A freeware "parallel" corpus analysis toolkit for concordancing and text analysis using UTF-8 encoded text files.  
[AntPConc Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (1.1.0)
- Macintosh OS X (1.1.0)
- Other versions

**AntWordProfiler**  
A freeware tool for profiling the vocabulary level and complexity of texts.  
[AntWordProfiler Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (1.4.0)
- Macintosh OS X (1.4.1)
- Linux (1.3.1)
- Other versions

**AntFileConverter**  
A freeware tool to convert PDF and Word (DOCX) files into plain text for use in corpus tools like AntConc.  
[AntFileConverter Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (1.2.0)
- Macintosh OS X (1.2.0)
- Other versions

**AntMover**  
A freeware text structure (moves) analysis program.  
[AntMover Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (1.0.0)
- Other versions

**AntCLAWSGUI**  
A front-end interface to the CLAWS tagger developed at Lancaster University, UK. **Note that you must have CLAWS installed before you can use AntCLAWSGUI. See the help file.**  
[AntCLAWSGUI Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (1.1.0)
- Other versions

**EncodeAnt**  
A freeware tool for detecting and converting character encodings.  
[EncodeAnt Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (1.2.0)
- Windows 64-bit (1.2.0)
- Macintosh OS X (1.2.0)
- Other versions

**FileAnt**  
A freeware social media and data analysis toolkit (developed in collaboration with Claire Harber of Lancaster University).  
[FileAnt Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (1.1.1)
- Windows 64-bit (1.1.1)
- Macintosh OS X (1.1.1)
- Other versions

**FindAnt**  
A freeware prototypical text analysis tool (developed in collaboration with Paul Baker of Lancaster University).  
[FindAnt Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (1.2.0)
- Windows 64-bit (1.2.0)
- Macintosh OS X (1.2.0)
- Other versions

**SearchAnt**  
A freeware batch search and replace tool.  
[SearchAnt Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (1.1.0)
- Windows 64-bit (1.1.0)
- Other versions

**SegmentAnt**  
A freeware Japanese and Chinese segmenter (segmentation/tagging tool).  
[SegmentAnt Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (1.1.0)
- Macintosh OS X (1.0.0)
- Linux (1.0.0)
- Other versions

**TagAnt**  
A freeware Part-Of-Speech (POS) tagger built on TreeTagger (developed by Martin Schmalz).  
[TagAnt Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (1.2.0)
- Windows 64-bit (1.2.0)
- Macintosh OS X (1.2.0)
- Linux (1.1.2)
- Other versions

**UnAnt**  
A freeware spelling UnAnt analysis program.  
[UnAnt Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (1.0.0)
- Other versions

## Data Cleaning Tools



EncodeAnt



SarAnt



SegmentAnt

# Introduction to AntConc 4:

## A member of the AntLab tools family

Laurence Anthony's Website

Home Resume Publications Software Classes Photo Albums Links Contact

**Software**

**AntConc**  
A freeware corpus analysis toolkit for concordancing and text analysis.  
[AntConc Homepage] [Screenshots] [Help]  
Download:  

- Windows (3.4.4)
- Macintosh OS X 10.7-10.10 (3.4.3)
- Macintosh OS X 10.6 (3.4.1)
- Linux (3.4.3)
- Other versions

**AntPConc**  
A freeware "parallel" corpus analysis toolkit for concordancing and text analysis using UTF-8 encoded text files.  
[AntPConc Homepage] [Screenshots] [Help]  
Download:  

- Windows (1.1.0)
- Macintosh OS X (1.1.0)
- Other versions

**AntWordProfiler**  
A freeware tool for profiling the vocabulary level and complexity of texts.  
[AntWordProfiler Homepage] [Screenshots] [Help]  
Download:  

- Windows (1.4.0)
- Macintosh OS X (1.4.1)
- Linux (1.3.1)
- Other versions

**AntFileConverter**  
A freeware tool to convert PDF and Word (DOCX) files into plain text for use in corpus tools like AntConc.  
[AntFileConverter Homepage] [Screenshots] [Help]  
Download:  

- Windows (1.2.0)
- Macintosh OS X (1.2.0)
- Other versions

**AntMover**  
A freeware tool structure (moves) analysis program.  
[AntMover Homepage] [Screenshots] [Help]  
Download:  

- Windows (1.0.0)
- Other versions

**AntCLAWSGUI**  
A front-end interface to the CLAWS tagger developed at Lancaster University, UK. Note that you must have CLAWS installed before you can use AntCLAWSGUI. See the help file.  
[AntCLAWSGUI Homepage] [Screenshots] [Help]  
Download:  

- Windows (1.1.0)
- Other versions

**EmbedAnt**  
A freeware tool for detecting and converting character encodings.  
[EmbedAnt Homepage] [Screenshots] [Help]  
Download:  

- Windows (1.2.0)
- Windows 64-bit (1.2.0)
- Macintosh OS X (1.2.0)
- Other versions

**FileAnt**  
A freeware social media and data analysis toolkit (developed in collaboration with Claire Harber of Lancaster University).  
[FileAnt Homepage] [Screenshots] [Help]  
Download:  

- Windows (1.1.1)
- Windows 64-bit (1.1.1)
- Macintosh OS X (1.1.1)
- Other versions

**FileAnt**  
A freeware prototypical text analysis tool (developed in collaboration with Paul Baker of Lancaster University).  
[FileAnt Homepage] [Screenshots] [Help]  
Download:  

- Windows (1.2.0)
- Windows 64-bit (1.2.0)
- Macintosh OS X (1.2.0)
- Other versions

**SearchAnt**  
A freeware batch search and replace tool.  
[SearchAnt Homepage] [Screenshots] [Help]  
Download:  

- Windows (1.1.0)
- Windows 64-bit (1.1.0)
- Other versions

**SegmentAnt**  
A freeware Japanese and Chinese segmenter (segmentation/tagging tool).  
[SegmentAnt Homepage] [Screenshots] [Help]  
Download:  

- Windows (1.1.0)
- Macintosh OS X (1.0.0)
- Linux (1.0.0)
- Other versions

**TagAnt**  
A freeware Part-Of-Speech (POS) tagger built on TreeTagger (developed by Martin Schmalz).  
[TagAnt Homepage] [Screenshots] [Help]  
Download:  

- Windows (1.2.0)
- Windows 64-bit (1.2.0)
- Macintosh OS X (1.2.0)
- Linux (1.1.2)
- Other versions

**UnAnt**  
A freeware spelling UnAnt analysis program.  
[UnAnt Homepage] [Screenshots] [Help]  
Download:  

- Windows (1.0.0)
- Other versions

## Data Annotation Tools



AntMover



CLAWSAnt



SegmentAnt



TagAnt

# Introduction to AntConc 4:

## A member of the AntLab tools family

Laurence Anthony's Website

Home Resume Publications Software Classes Photo Albums Links Contact

**Software**

**AntConc**  
A freeware corpus analysis toolkit for concordancing and text analysis.  
[AntConc Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (3.4.4)
- Macintosh OS X 10.7-10.10 (3.4.3)
- Macintosh OS X 10.6 (3.4.1)
- Linux (3.4.3)
- Other versions

**AntPConc**  
A freeware "parallel" corpus analysis toolkit for concordancing and text analysis using UTF-8 encoded text files.  
[AntPConc Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (1.1.0)
- Macintosh OS X (1.1.0)
- Other versions

**AntWordProfiler**  
A freeware tool for profiling the vocabulary level and complexity of texts.  
[AntWordProfiler Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (1.4.0)
- Macintosh OS X (1.4.1)
- Linux (1.3.1)
- Other versions

**AntFileConverter**  
A freeware tool to convert PDF and Word (DOCX) files into plain text for use in corpus tools like AntConc.  
[AntFileConverter Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (1.2.0)
- Macintosh OS X (1.2.0)
- Other versions

**AntMover**  
A freeware tool structure (moves) analysis program.  
[AntMover Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (1.0.0)
- Other versions

**AntCLAWSGUI**  
A front-end interface to the CLAWS tagger developed at Lancaster University, UK. Note that you must have CLAWS installed before you can use AntCLAWSGUI. See the help file.  
[AntCLAWSGUI Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (1.1.0)
- Other versions

**EmbedAnt**  
A freeware tool for detecting and converting character encodings.  
[EmbedAnt Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (1.2.0)
- Windows 64-bit (1.2.0)
- Macintosh OS X (1.2.0)
- Other versions

**FireAnt**  
A freeware social media and data analysis toolkit (developed in collaboration with Claire Harber of Lancaster University).  
[FireAnt Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (1.1.1)
- Windows 64-bit (1.1.1)
- Macintosh OS X (1.1.1)
- Other versions

**ProtAnt**  
A freeware prototypical text analysis tool (developed in collaboration with Paul Baker of Lancaster University).  
[ProtAnt Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (1.2.0)
- Windows 64-bit (1.2.0)
- Macintosh OS X (1.2.0)
- Other versions

**SeAnt**  
A freeware batch search and replace tool.  
[SeAnt Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (1.1.0)
- Windows 64-bit (1.1.0)
- Other versions

**SegmentAnt**  
A freeware Japanese and Chinese segmenter (segmentation/tagging tool).  
[SegmentAnt Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (1.1.0)
- Macintosh OS X (1.0.0)
- Linux (1.0.0)
- Other versions

**TagAnt**  
A freeware Part-Of-Speech (POS) tagger built on TreeTagger (developed by Helmut Schmid).  
[TagAnt Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (1.2.0)
- Windows 64-bit (1.2.0)
- Macintosh OS X (1.2.0)
- Linux (1.1.2)
- Other versions

**VariAnt**  
A freeware spelling VariAnt analysis program.  
[VariAnt Homepage] [Screenshots] [Help]  
Downloads:  

- Windows (1.0.0)
- Other versions

## Data Analysis Tools



AntConc



AntPConc



AntWordProfiler



FireAnt



ProtAnt



VariAnt

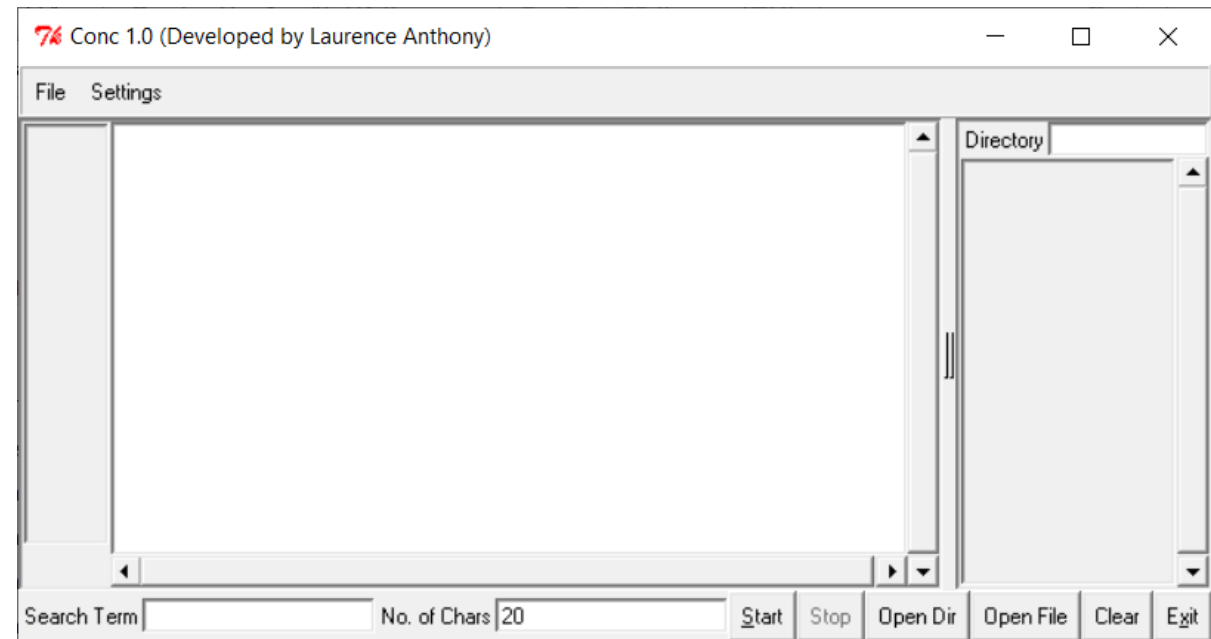
AntLab Tools

www.laurenceanthony.net/software

# Introduction to AntConc 4:

Evolution over 20 years

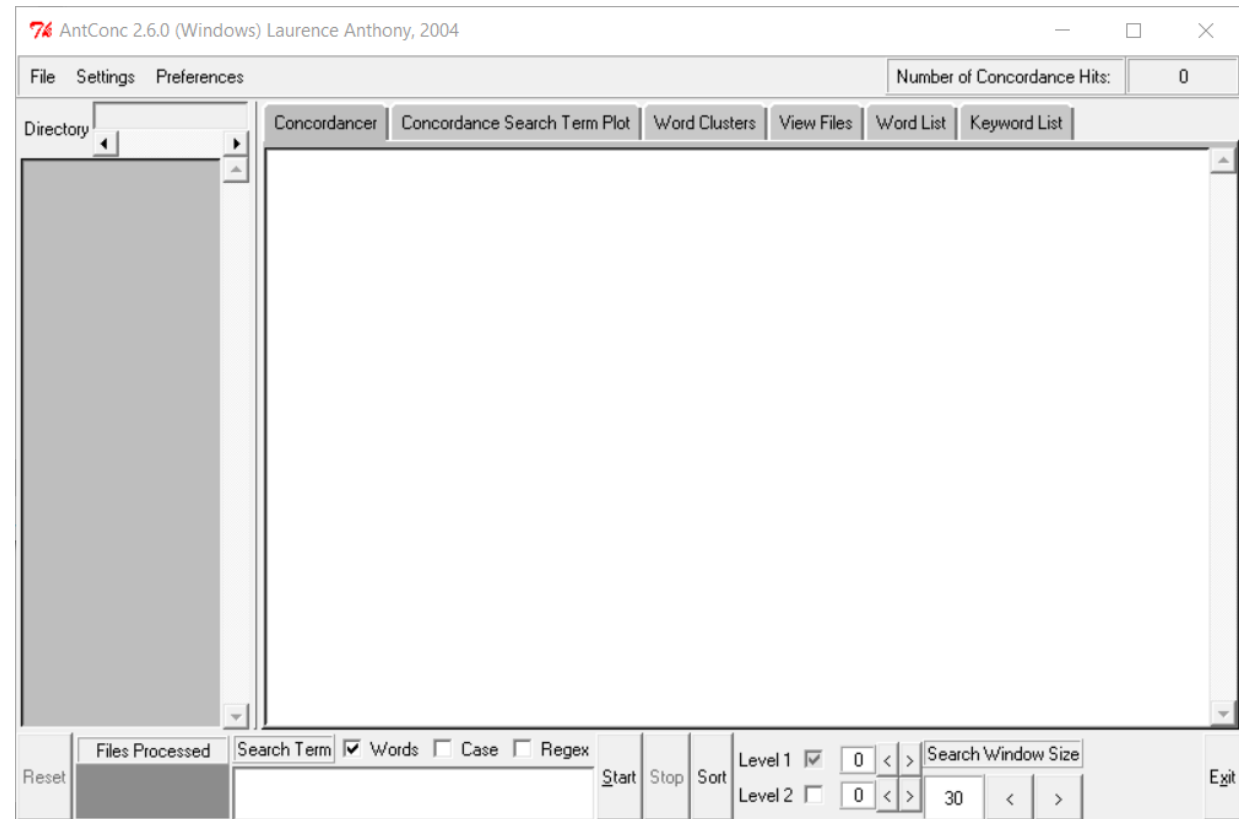
- Major updates
  - Version 1.0-1.1: Simple concordancer for Windows and Linux
    - tools: concordance
    - language: Perl + Tk



# Introduction to AntConc 4:

## Evolution over 20 years

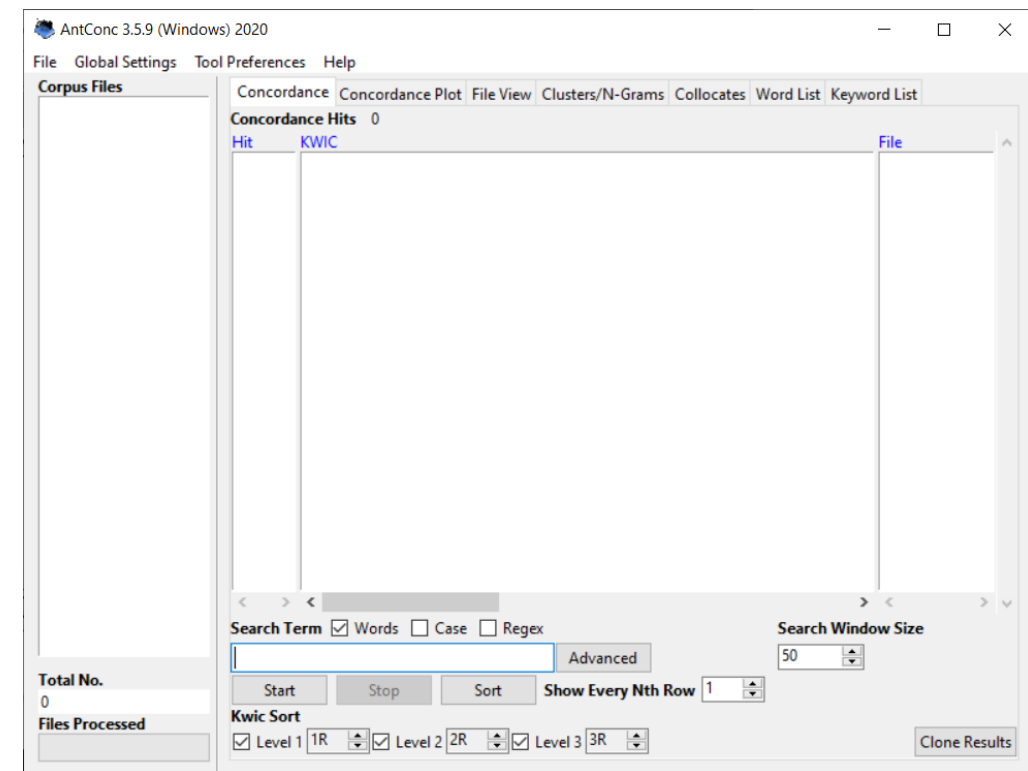
- Major updates
  - Version 2.0-2.6: Corpus toolkit for Windows and Linux
    - tools: concordance, plot, cluster/n-grams, file, word, keyword
    - language: Perl + Tk



# Introduction to AntConc 4:

Evolution over 20 years

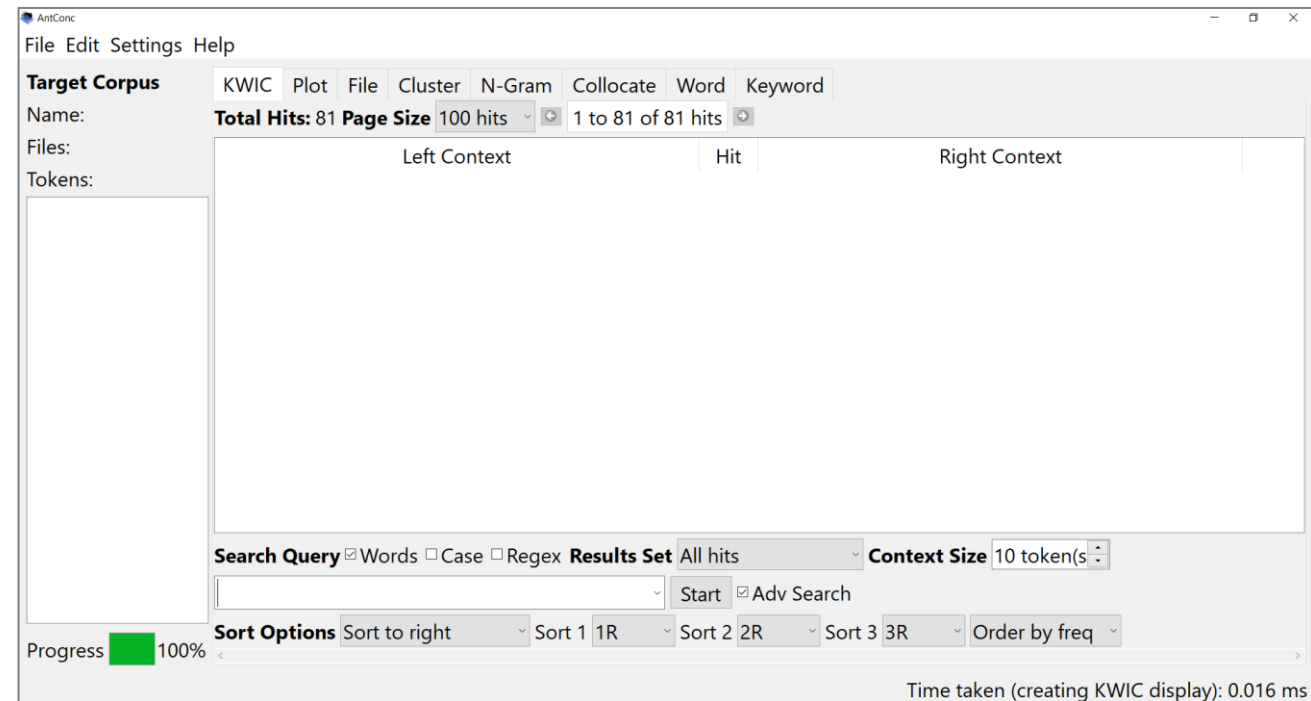
- Major updates
  - Version 3.0-3.5: Corpus toolkit for Windows, Linux, and MacOS (native)
    - tools: concordance, plot, file, cluster/n-grams, collocates, word, keyword
    - language: Perl + Tkl



# Introduction to AntConc 4:

## Evolution over 20 years

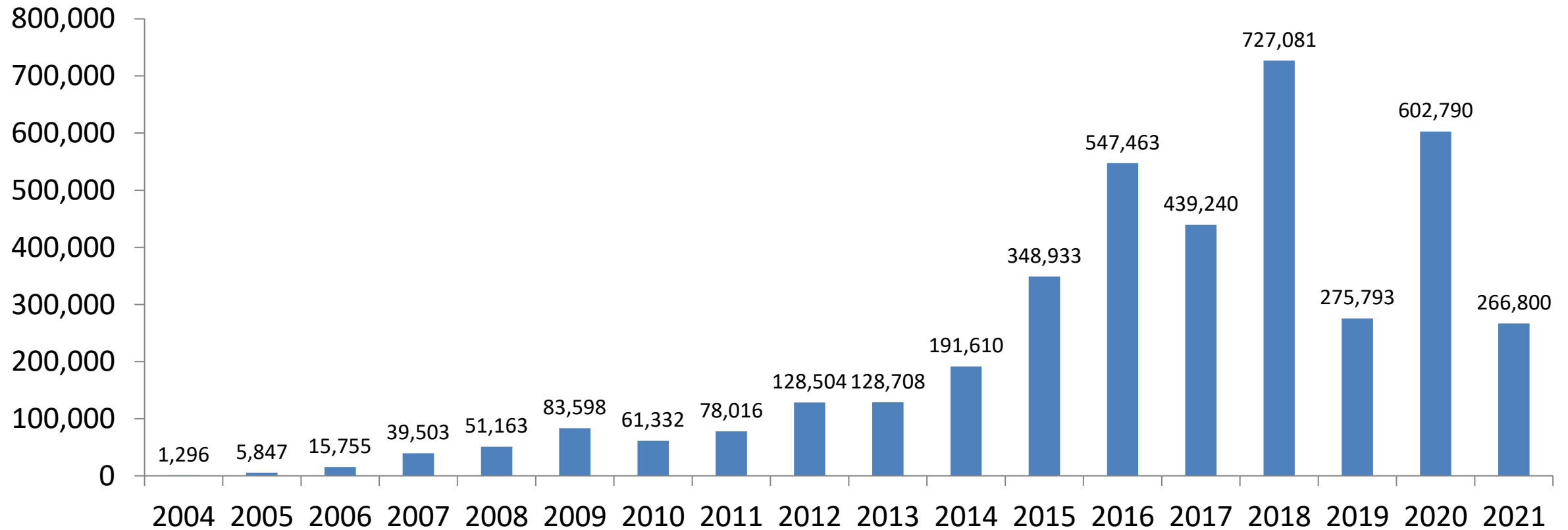
- Major updates
  - Version 4.0: Advanced corpus toolkit for Windows, Linux, and MacOS (native)
    - tools: concordance, plot, file, cluster, n-grams, collocates, word, keyword
    - language: Python + Qt



# Introduction to AntConc 4:

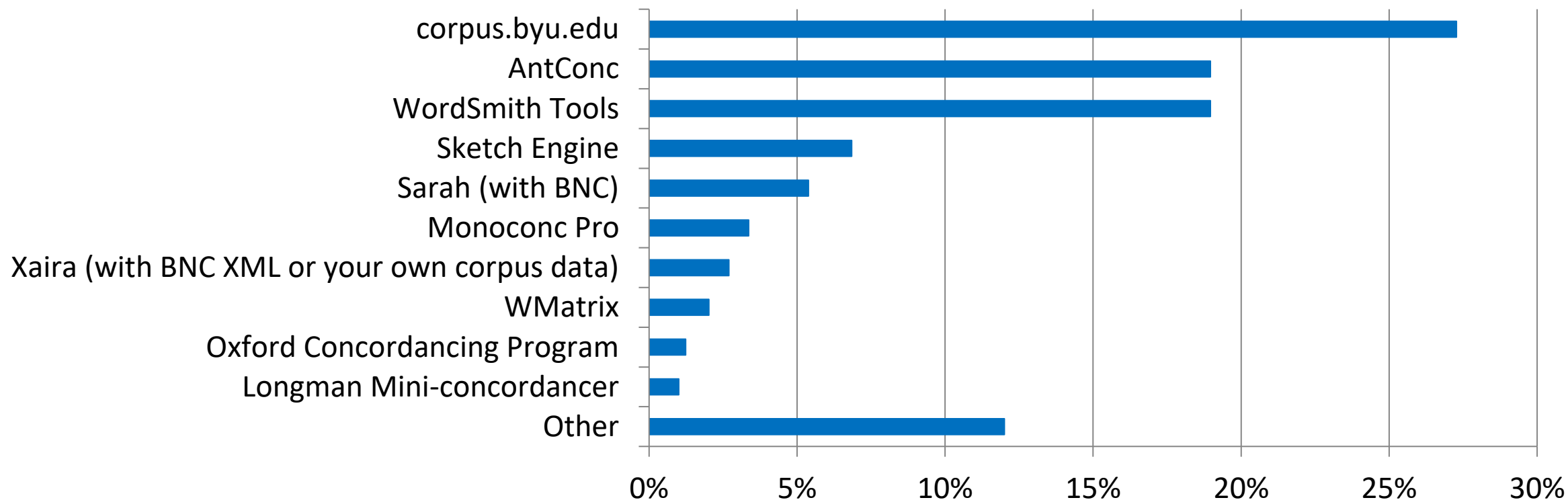
Evolution over 20 years

## AntConc Downloads (2004-2021)



# Introduction to AntConc 4:

Evolution over 20 years



"Which computer programs do you use for analysing corpora?"  
International survey of corpus linguists. Responses: 891. (Tribble, 2012)



# Introduction to AntConc 4:

## Evolution over 20 years

AntConc

File Edit Settings Help

**Target Corpus**

Name: AmE06\_Learned  
Files: 80  
Tokens: 161469

AmE06\_J01.txt  
AmE06\_J02.txt  
AmE06\_J03.txt  
AmE06\_J04.txt  
AmE06\_J05.txt  
AmE06\_J06.txt  
AmE06\_J07.txt  
AmE06\_J08.txt  
AmE06\_J09.txt  
AmE06\_J10.txt  
AmE06\_J11.txt  
AmE06\_J12.txt  
AmE06\_J13.txt  
AmE06\_J14.txt  
AmE06\_J15.txt  
AmE06\_J16.txt

Progress  100%

KWIC Plot File Cluster N-Gram Collocate Word Keyword

**Total Hits:** 120 **Page Size** 100 hits 1 to 100 of 120 hits

	File	Left Context	Hit	Right Context
1	AmE06_J47.txt	matches one image for the	study	of language comprehension (Hirsh-
2	AmE06_J47.txt	same process occurs in the	study	of language development. As
3	AmE06_J32.txt	t see that movie. The	study	of language that focuses
4	AmE06_J47.txt	washing machines and toasters. The	study	of language use and
5	AmE06_J70.txt	surface water. In a multiyear	study	of a small watershed
6	AmE06_J74.txt	if not terribly exceptional case	study	of a technological crisis
7	AmE06_J47.txt	was a uniquely large-sample	study	of child language development,
8	AmE06_J47.txt	most striking developments in the	study	of child language, and
9	AmE06_J29.txt	of the 1970s, when the	study	of gender in linguistics
10	AmE06_J29.txt	as a social institution, the	study	of gender in the
11	AmE06_J64.txt	a less ambitious, still unpublished,	study	of MacDowell and Huneker.
12	AmE06_J24.txt	reater. Conversely, in Rosecrance's (1985)	study	of backstretch workers at

**Search Term**  Words  Case  Regex **Results Set** All hits **Context Size** 5 token(s)

study Start  Adv Search

**Sort Options** Sort to right Sort 1 1R Sort 2 2R Sort 3 3R Order by freq

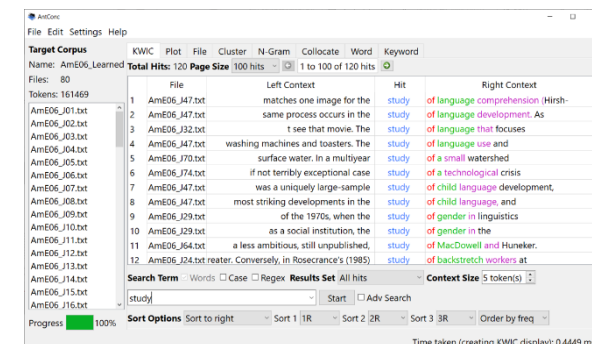
Time taken (creating KWIC display): 0.4449 ms



# Introduction to AntConc 4:

## Design concept - strengths

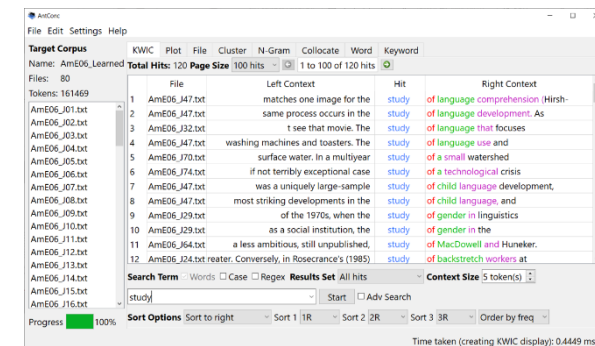
- easy to find, set up, and use
  - e.g., installer or portable version (no installation required)
  - e.g., help guides, YouTube tutorial videos, discussion forums, ...
- safe and reliable
  - e.g., no viruses, community error reports, bug fixes, updates, ...
- evolving
  - e.g., new tools, new features, new documentation, ...
- fast, comprehensive set of tools
  - e.g., KWIC, Plot, File, Cluster, N-gram, Collocate, Word (list), Keyword (list)
  - e.g., fully indexed SQLite database, transparent data interoperability
- designed for a broad user base
  - e.g., multiplatform support (Win, Mac, Linux)
  - e.g., multilingual support (UTF-8 compliant)
  - e.g., supports LTR (e.g., English, Japanese) and RTL languages (e.g., Arabic)



# Introduction to AntConc 4:

## Design concept - weaknesses

- designed for a broad user base
  - e.g., only standard corpus tools and statistics are included [although best practices are followed]
- 'black box' technology
  - e.g., the software is 'freeware' but not 'open-source'
- inflexible
  - e.g., users cannot easily add new features or statistical methods



# Introduction to AntConc 4:

Features and functions - KWIC, Plot, File, Cluster, N-Gram, Collocate, **Word**, Keyword

The screenshot shows the AntConc 4.0.0.0 interface. The 'Word' tab is active, displaying the following information:


- Target Corpus:** AmE06\_Learned
- Files:** 80
- Tokens:** 161469
- Types:** 15887/15887
- Tokens:** 161469/161469
- Page Size:** 100 hits
- Results:** 1 to 100 of 15887 hits

	Type	Rank	Freq	Range
1	the	1	10376	80
2	of	2	6649	80
3	and	3	5077	80
4	to	4	4005	80
5	in	5	3966	80
6	a	6	3562	80
7	that	7	2138	80
8	is	8	2016	79
9	for	9	1450	80
10	as	10	1402	80
11	-	11	1128	78

Search Query:  Words  Case  Regex

Start  Adv Search

Sort by: Frequency  Invert Order

Progress:  100%

Time taken (creating wordlist) 0.1046 ms

# Introduction to AntConc 4:


Features and functions - KWIC, Plot, File, Cluster, N-Gram, Collocate, Word, **Keyword**

AntConc

File Edit Settings Help

**Target Corpus**  
Name: AmE06\_Learned  
Files: 80  
Tokens: 161469

**Reference Corpus**  
Name: AmE06  
Files: 500  
Tokens: 1017879

Progress  100%

KWIC Plot File Cluster N-Gram Collocate Word **Keyword**

**Keyword Types** 380/15887 **Keyword Tokens** 53670/161469 **Page Size** 100 hits 1 to 100 of 380 hits

	Type	Rank	Freq_Tar	Freq_Ref	Range_Tar	Range_Ref	Keyness (Likelihood)	Keyness (Effect)
1	of	1	6649	30331	80	500	550.584	0.067
2	x	2	268	331	14	31	339.754	0.003
3	is	3	2016	8420	79	488	255.347	0.023
4	learning	4	145	196	14	44	169.355	0.002
5	are	5	1067	4226	78	468	168.111	0.013
6	in	6	3966	19923	80	500	165.568	0.043
7	et	7	131	163	20	32	164.941	0.002
8	k	8	136	181	11	44	161.163	0.002
9	these	9	459	1406	76	379	158.264	0.006
10	e	10	182	343	43	115	147.247	0.002
11	...	11	100	130	0	20	135.140	0.001

**Search Query**  Words  Case  Regex

Adv Search

**Sort by** Likelihood  Invert Order

Time taken (creating keyword list) 0.3565 ms

# Introduction to AntConc 4:


Features and functions - KWIC, Plot, File, Cluster, N-Gram, Collocate, Word, **Keyword**

AntConc 4.0.0.0

File Edit Settings Help

**Target Corpus**  
Name: AmE06\_Learned  
Files: 80  
Tokens: 161469

**Reference Corpus**  
Name: AmE06  
Files: 500  
Tokens: 1017879

Progress  100%

KWIC Plot File Cluster N-Gram Collocate Word **Keyword**

**Keyword Types** 23/15887 **Keyword Tokens** 2454/161469 **Page Size** 100 hits 1 to 23 of 23 hits

	Type	Rank	Freq_Tar	Freq_Ref	Range_Tar	Range_Ref	Keyness (Likelihood)	Keyness (Effect)
1	however	1	167	444	65	224	39.102	0.002
2	thus	2	92	206	49	128	37.963	0.001
3	study	3	120	332	51	142	36.427	0.001
4	particular	4	75	197	47	128	33.017	0.001
5	g	5	128	198	33	65	32.093	0.002
6	studies	6	100	181	36	80	30.695	0.001
7	such	7	298	1033	76	346	30.095	0.004
8	example	8	120	322	50	152	29.727	0.001
9	e	9	182	343	43	115	29.619	0.002
10	associated	10	60	106	30	60	27.854	0.001
11	analysis	11	76	178	27	92	26.407	0.001

**Search Query**  Words  Case  Regex

Adv Search

**Sort by** Likelihood  Invert Order

Time taken (creating keyword list) 0.0079 ms

# Introduction to AntConc 4:

Features and functions - KWIC, Plot, File, Cluster, N-Gram, Collocate, Word, Keyword

The screenshot shows the AntConc 4 application window. The 'Target Corpus' is 'AmE06\_Learned' with 80 files and 161469 tokens. The search query is 'process', and the results are sorted by frequency. The interface includes a menu bar (File, Edit, Settings, Help), a toolbar with search options (KWIC, Plot, File, Cluster, N-Gram, Collocate, Word, Keyword), and a main results table. A progress bar at the bottom left shows 100% completion, and a status bar at the bottom right indicates the time taken for creating plot results is 0.3018 ms.

	File	Left Context	Hit	Right Context
1	AmE06_J07.txt	wever, prompted by the need to place the	process	of taking moments in context. Moments of
2	AmE06_J33.txt	education. Successful online teaching is a	process	of taking our very best practices in the clas
3	AmE06_J25.txt	homes. The findings demonstrate that the	process	of assimilation was not uniform for all grou
4	AmE06_J43.txt	nunist Party of Indonesia, which was in the	process	of being eliminated by Soeharto's New Orc
5	AmE06_J65.txt	canon of modern children's literature. The	process	of creating or augmenting professional ide
6	AmE06_J13.txt	lack that protein. Now scientists are in the	process	of figuring out which proteins are coded fo
7	AmE06_J80.txt	and Martin (2004), and Pysek et al. (2004).	Process	of invasion At one level, the issue of invasiv
8	AmE06_J33.txt	participants is formed, through which the	process	of knowledge acquisition is collaboratively
9	AmE06_J51.txt	ference. An inference, in turn, is a mental	process	of linking propositions by offering support
10	AmE06_J25.txt	appropriate for explaining the adaptation	process	of newcomers who arrived in America in th

# Introduction to AntConc 4:

Features and functions - KWIC, Plot, File, Cluster, N-Gram, Collocate, Word, Keyword

AntConc 4.0.0.0

File Edit Settings Help

**Target Corpus**

Name: AmE06\_Learned **Total Hits:** 106 **Total Files With Hits:** 34

Files: 80  
Tokens: 161469

- AmE06\_J20.txt
- AmE06\_J21.txt
- AmE06\_J22.txt
- AmE06\_J23.txt
- AmE06\_J24.txt
- AmE06\_J25.txt
- AmE06\_J26.txt
- AmE06\_J27.txt
- AmE06\_J28.txt
- AmE06\_J29.txt
- AmE06\_J30.txt
- AmE06\_J31.txt
- AmE06\_J32.txt
- AmE06\_J33.txt
- AmE06\_J34.txt

Progress  100%

KWIC Plot File Cluster N-Gram Collocate Word Keyword

DocID	DocPath	DocTokens	Freq	NormFreq	Dispersion	Plot
1	AmE06_J28.txt	2021	4	1979.218	0.592	
2	AmE06_J80.txt	1966	5	2543.235	0.553	
3	AmE06_J02.txt	2036	4	1964.637	0.447	
4	AmE06_J51.txt	2000	4	2000.000	0.447	
5	AmE06_J55.txt	2071	4	1931.434	0.447	
6	AmE06_J77.txt	1995	4	2005.013	0.447	
7	AmE06_J05.txt	2025	9	4444.444	0.440	
8	AmE06_J25.txt	1968	7	3556.911	0.396	
9	AmE06_J33.txt	2029	12	5914.243	0.345	
10	AmE06_J13.txt	2051	4	1950.268	0.333	

**Search Query**  Words  Case  Regex **Results Set** All hits **Plot Zoom** 1.00 x **Overlay**  Color ■

process of Start  Adv Search

**Sort by** Dispersion  Invert Order

Time taken (creating plot results): 0.2925 ms

# Introduction to AntConc 4:

Features and functions - KWIC, Plot, File, Cluster, N-Gram, Collocate, Word, Keyword

AntConc

File Edit Settings Help


**Target Corpus** KWIC Plot File Cluster N-Gram Collocate Word Keyword

Name: AmE06\_Learned **File Hits** 10 **File Types** 696 **File Tokens** 2029 **File Name** AmE06\_J33.txt

Files: 80

Tokens: 161469

AmE06\_J20.txt  
AmE06\_J21.txt  
AmE06\_J22.txt  
AmE06\_J23.txt  
AmE06\_J24.txt  
AmE06\_J25.txt  
AmE06\_J26.txt  
AmE06\_J27.txt  
AmE06\_J28.txt  
AmE06\_J29.txt  
AmE06\_J30.txt  
AmE06\_J31.txt  
AmE06\_J32.txt  
AmE06\_J33.txt  
AmE06\_J34.txt

Progress  100%

environment cannot be passive. If students do not enter into the online classroom - do not post a contribution to the discussion - the instructor has almost no way of knowing whether they have been there. So students are not only responsible for logging on but they must also contribute to the learning **process** by posting their thoughts and ideas to the online discussion. Learning is an active **process** in which both the instructor and the learners must participate if it is to be successful. In the **process**, a web of learning is created. In other words, a network of interactions between the instructor and the other participants is formed, through which the **process** of knowledge acquisition is collaboratively created. (See Chapters Eight and Nine for a discussion of collaborative learning and the transformative nature of the learning **process**.) Outcomes of this **process**, then, should not be measured by the number of facts memorized and the amount of subject matter regurgitated but by the depth of knowledge and the number of skills gained. Evidence of critical thinking and of knowledge acquired are the desired learning outcomes. Consequently, cheating on exams should not be a major concern in an effective online environment because knowledge is acquired collaboratively through the development of a learning community. (The assessment of student performance in this environment is discussed in Chapter Ten.) Institutions entering the distance learning arena must be prepared to tackle these issues and to develop new approaches and new skills in order to create an empowering learning **process**, for the creation of empowered learners is yet another desired outcome of online distance education. Successful online teaching is a **process** of taking our very best practices in the classroom and

**Search Query**  Words  Case  Regex **Hit Location** 1

process Start  Adv Search

Time taken (creating file view) 0.0579 ms

# Introduction to AntConc 4:

Features and functions - KWIC, Plot, File, **Cluster**, N-Gram, Collocate, Word, Keyword

AntConc 4.0.0.0

File Edit Settings Help

**Target Corpus**  
Name: AmE06\_Learned  
Files: 80  
Tokens: 161469

KWIC Plot File **Cluster** N-Gram Collocate Word Keyword

Cluster Types 43 Cluster Tokens 87 Page Size 100 hits 1 to 43 of 43 hits

	Cluster	Rank	Freq	Range
1	process of	1	19	15
2	process that	2	9	9
3	process and	3	5	3
4	process the	3	5	3
5	process to	5	4	4
6	process a	6	3	3
7	process was	6	3	2
8	process for	8	2	2
9	process in	8	2	2
10	process known	8	2	2
11	process residents	11	1	1

Search Query  Words  Case  Regex Cluster Size 2 Min. Freq 1 Min. Range 1

process of Start  Adv Search

Sort by Frequency  Invert Order Search Term Position  On Left  On Right  On Left/Right

Progress 100%

Time taken (creating file view) 0.0579 ms

# Introduction to AntConc 4:

Features and functions - KWIC, Plot, File, Cluster, N-Gram, Collocate, Word, Keyword


AntConc

File Edit Settings Help

**Target Corpus**

Name: AmE06\_Learned  
Files: 80  
Tokens: 161469

AmE06\_J20.txt  
AmE06\_J21.txt  
AmE06\_J22.txt  
AmE06\_J23.txt  
AmE06\_J24.txt  
AmE06\_J25.txt  
AmE06\_J26.txt  
AmE06\_J27.txt  
AmE06\_J28.txt  
AmE06\_J29.txt  
AmE06\_J30.txt  
AmE06\_J31.txt  
AmE06\_J32.txt  
AmE06\_J33.txt  
AmE06\_J34.txt

Progress  100%

KWIC Plot File Cluster **N-Gram** Collocate Word Keyword

**N-Gram Types** 94736 **N-Gram Tokens** 161389 **Page Size** 100 hits 1 to 100 of 94736 hits

	Type	Rank	Freq	Range
1	of the	1	1460	80
2	in the	2	1008	80
3	to the	3	507	80
4	and the	4	379	76
5	on the	5	284	75
6	to be	6	272	74
7	it is	7	260	62
8	for the	8	250	70
9	that the	9	248	70
10	as a	10	233	71
11	of	11	222	67

**Search Query**  Words  Case  Regex **N-Gram Size** 2 **Open Slots** 0 **Min. Freq** 1 **Min. Range** 1

Start  Adv Search

**Sort by** Frequency  Invert Order

Time taken (creating ngram list) 0.8965 ms

# Introduction to AntConc 4:

Features and functions - KWIC, Plot, File, Cluster, N-Gram, Collocate, Word, Keyword

AntConc

File Edit Settings Help

**Target Corpus**

Name: AmE06\_Learned

Files: 80

Tokens: 161469

AmE06\_J20.txt

AmE06\_J21.txt

AmE06\_J22.txt

AmE06\_J23.txt

AmE06\_J24.txt

AmE06\_J25.txt

AmE06\_J26.txt

AmE06\_J27.txt

AmE06\_J28.txt

AmE06\_J29.txt


AmE06\_J30.txt

AmE06\_J31.txt

AmE06\_J32.txt

AmE06\_J33.txt

AmE06\_J34.txt

Progress  100%

KWIC Plot File Cluster **N-Gram** Collocate Word Keyword

**N-Gram Types** 155809 **N-Gram Tokens** 161387 **Page Size** 100 hits 1 to 100 of 155809 hits

	Type	Rank	Freq	Range
1	in the united states	1	25	12
2	is the case that	2	20	3
3	it is the case	2	20	3
4	the ac tc ratio	4	19	1
5	in the context of	5	18	12
6	at the same time	6	16	11
7	in the form of	7	15	14
8	commissioners aranoff and hillman	8	14	1
9	ω x k s	9	13	1
10	the end of the	10	12	9
11	in the case of	11	11	0

**Search Query**  Words  Case  Regex **N-Gram Size** 4 **Open Slots** 0 **Min. Freq** 1 **Min. Range** 1

Adv Search

**Sort by** Frequency  Invert Order

Time taken (creating ngram list) 2.4093 ms

# Introduction to AntConc 4:

Features and functions - KWIC, Plot, File, Cluster, N-Gram, Collocate, Word, Keyword

The screenshot shows the AntConc 4 interface with the N-Gram tab selected. The main window displays a table of N-Gram results for the search query "in the \* that". The table has columns for Rank, Freq, Range, S1\_TT, S1\_Ent, S2\_TT, and S2\_Ent. The top result is "in the + that" with a rank of 1 and a frequency of 15. An "Open Slot Viewer" dialog is open, showing the S1 and S2 slots for the selected entry. The S1 slot is empty, and the S2 slot contains a list of words with their frequencies: case (2), sense (1), chapters (1), fact (1), symptom (1), brain (1), sections (1), and destabilization (1).

**Target Corpus**  
Name: AmE06\_Learned  
Files: 80  
Tokens: 161469

**N-Gram Types** 15 **N-Gram Tokens** 30 **Page Size** 100 hits 1 to 15 of 15 hits

	Type	Rank	Freq	Range	S1_TT	S1_Ent	S2_TT	S2_Ent
1	in the + that	1	15	13			0.933	0.991
2	in + case that	2	2	1	0.5	0.0		
3	in + analysis that	3	1	1	1.0	0.0		
4	in + brain that	3	1	1	1.0	0.0		
5	in + chapters that	3	1	1	1.0	0.0		
6	in + destabilization that	3	1	1	1.0	0.0		
7	in + fact that	3	1	1	1.0	0.0		
8	in + fields that	3	1	1	1.0	0.0		
9	in + grooves that	3	1	1	1.0	0.0		
10	in + s that	3	1	1	1.0	0.0		
11	in + sections that	3	1	1	1.0	0.0		

**Search Query**  Words  Case  Regex **N-Gram Size** 4 **Open Slots** 1 **Min. Freq** 1 **Min. Range** 1

in the \* that Start  Adv Search

**Sort by** Frequency  Invert Order

Progress 100%

Time taken (creating ngram list) 0.7917 ms

# Introduction to AntConc 4:

Features and functions - KWIC, Plot, File, Cluster, N-Gram, Collocate, Word, Keyword

AntConc

File Edit Settings Help

**Target Corpus**

Name: AmE06\_Learned

Files: 80

Tokens: 161469

AmE06\_J20.txt

AmE06\_J21.txt

AmE06\_J22.txt

AmE06\_J23.txt

AmE06\_J24.txt

AmE06\_J25.txt

AmE06\_J26.txt

AmE06\_J27.txt

AmE06\_J28.txt

AmE06\_J29.txt


AmE06\_J30.txt

AmE06\_J31.txt

AmE06\_J32.txt

AmE06\_J33.txt

AmE06\_J34.txt

Progress  100%

KWIC Plot File Cluster N-Gram Collocate Word Keyword

Collocate Types 9 Collocate Tokens 120 Page Size 100 hits 1 to 9 of 9 hits

	Collocate	Rank	FreqLR	FreqL	FreqR	Range	Likelihood	Effect
1	learning	1	9	8	1	3	27.680	3.526
2	chemical	2	6	5	1	2	24.376	4.288
3	large	3	6	1	5	3	22.705	4.077
4	globalization	4	4	3	1	1	20.126	5.012
5	attends	5	2	1	1	1	20.110	8.536
6	postulate	6	2	1	1	1	17.125	7.536
7	assimilation	7	3	1	2	1	15.533	5.121
8	gaseous	8	2	1	1	1	15.444	6.951
9	the	9	86	56	30	31	15.024	0.621

Search Query  Words  Case  Regex Window Span From 5L To 5R Min. Freq 1 Min. Range 1

process Start  Adv Search

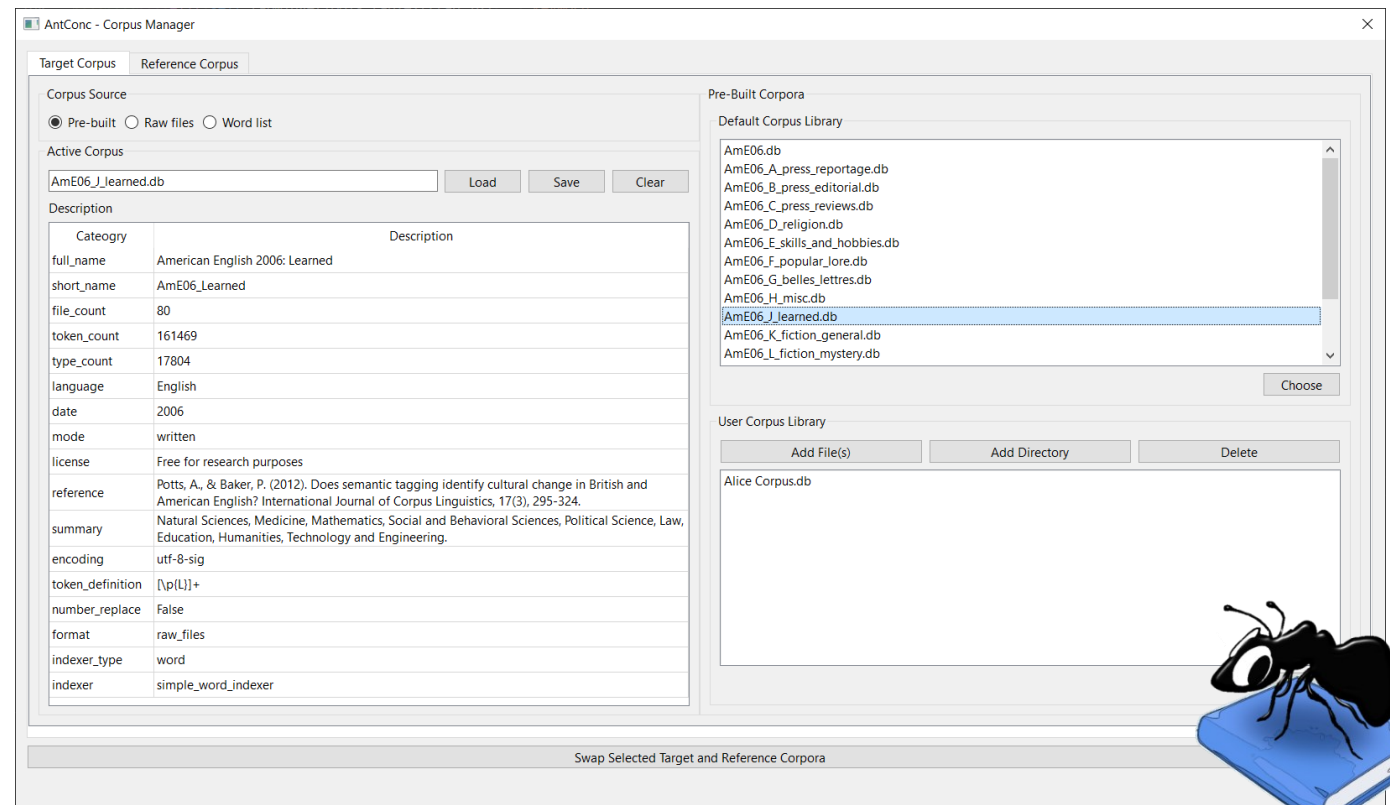
Sort by Likelihood  Invert Order

Time taken (creating collocate list) 0.001 ms

# Introduction to AntConc 4:

Addressing the challenges of DDL - 1) finding suitable target/discipline-specific corpora

- AntConc "Corpus Manager" - **default/user corpus library**
  - a repository of pre-built corpora
    - AmE06
      - AmE06 press
      - AmE06 learned (academic)
      - ...
    - BE06
      - BE06 press
      - BE06 learned (academic)
      - ...
    - BASE/BAWE (coming soon)
    - ...
    - your contributions??

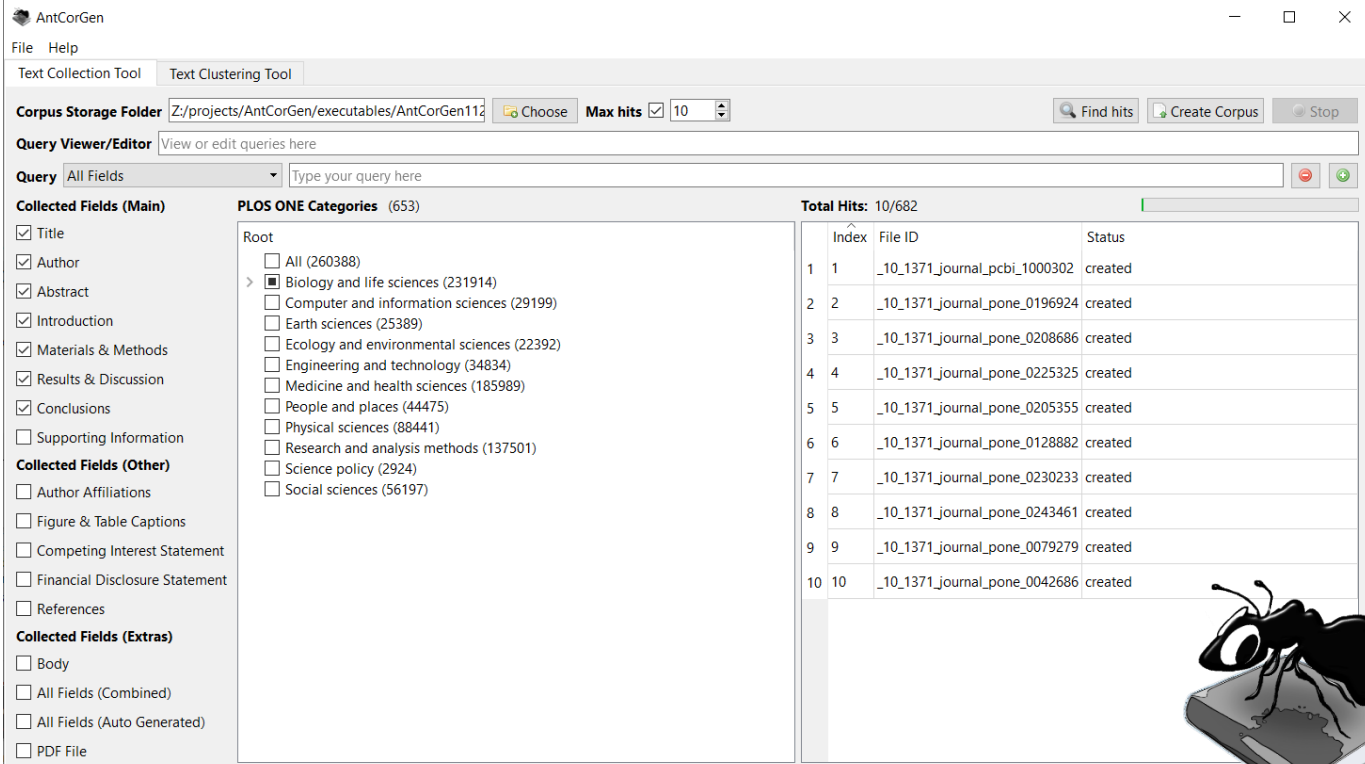


# Introduction to AntConc 4:

Addressing the challenges of DDL - 1) finding suitable target/discipline-specific corpora

## ■ AntCorGen - A PLOS ONE multi-disciplinary research article collection tool

- Anthony, L. (2022). AntCorGen (Version 1.2.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from <https://www.laurenceanthony.net/software/antcorgen>



The screenshot displays the AntCorGen software interface. The window title is "AntCorGen". The menu bar includes "File" and "Help". There are two tabs: "Text Collection Tool" and "Text Clustering Tool". The "Text Collection Tool" tab is active, showing a "Corpus Storage Folder" field with the path "Z:/projects/AntCorGen/executables/AntCorGen112", a "Choose" button, and a "Max hits" dropdown set to "10". There are "Find hits", "Create Corpus", and "Stop" buttons. Below this is a "Query Viewer/Editor" with a text input field and "View or edit queries here" text. The "Query" dropdown is set to "All Fields" and there is a "Type your query here" input field. The main area is divided into three sections: "Collected Fields (Main)", "PLOS ONE Categories (653)", and "Total Hits: 10/682".

**Collected Fields (Main)**

- Title
- Author
- Abstract
- Introduction
- Materials & Methods
- Results & Discussion
- Conclusions
- Supporting Information

**Collected Fields (Other)**

- Author Affiliations
- Figure & Table Captions
- Competing Interest Statement
- Financial Disclosure Statement
- References

**Collected Fields (Extras)**

- Body
- All Fields (Combined)
- All Fields (Auto Generated)
- PDF File

**PLOS ONE Categories (653)**

- Root
  - All (260388)
  - Biology and life sciences (231914)
    - Computer and information sciences (29199)
    - Earth sciences (25389)
    - Ecology and environmental sciences (22392)
    - Engineering and technology (34834)
    - Medicine and health sciences (185989)
    - People and places (44475)
    - Physical sciences (88441)
    - Research and analysis methods (137501)
    - Science policy (2924)
    - Social sciences (56197)

**Total Hits: 10/682**

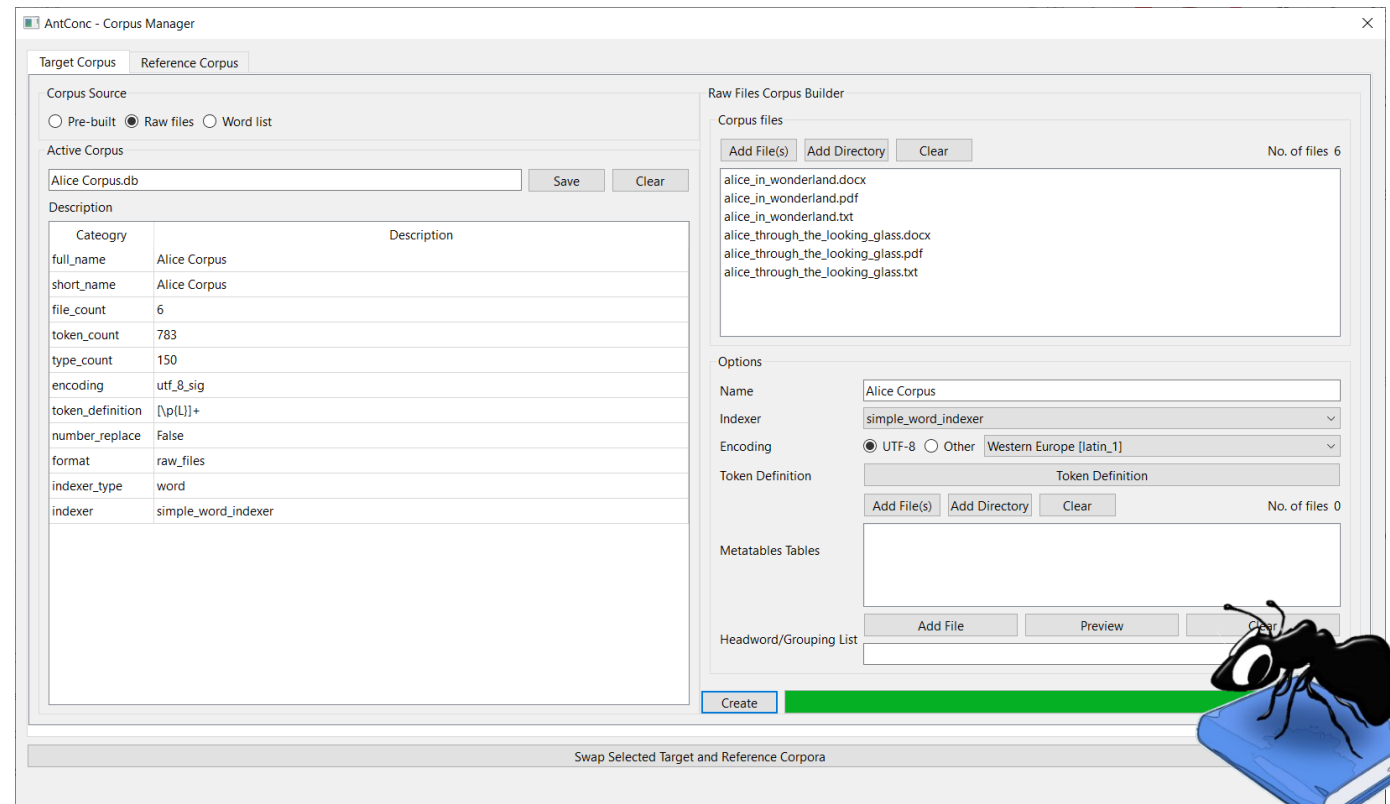
	Index	File ID	Status
1	1	_10_1371_journal_pcbi_1000302	created
2	2	_10_1371_journal_pone_0196924	created
3	3	_10_1371_journal_pone_0208686	created
4	4	_10_1371_journal_pone_0225325	created
5	5	_10_1371_journal_pone_0205355	created
6	6	_10_1371_journal_pone_0128882	created
7	7	_10_1371_journal_pone_0230233	created
8	8	_10_1371_journal_pone_0243461	created
9	9	_10_1371_journal_pone_0079279	created
10	10	_10_1371_journal_pone_0042686	created



# Introduction to AntConc 4:

## Addressing the challenges of DDL - 2) creating custom corpora

- AntConc "Corpus Manager" - **corpus builder**
  - Word (.docx), PDF (.pdf), TEXT (.txt) corpus creation via drag n' drop
    - student papers
    - research papers
    - e-books
    - copy/paste articles
    - ...



# Introduction to AntConc 4:

Addressing the challenges of DDL - 3) knowing what to search for in a corpus

- AntConc tools for language discovery - **Word, Keyword, (open slot) N-gram**
  - discover 'important' (frequent) words using the Word (list) tool
  - discover 'important' (unusually frequent) words using the Keyword (list) tool
  - discover frequent word patterns (lexical bundles & positional frames) using the N-gram tool

The screenshot shows the AntConc 4.0.0.0 interface. The 'Target Corpus' is 'AmE06\_Learned' with 80 files and 161469 tokens. The 'N-Gram' tool is active, showing a search query of 'in the \* that'. The results table is as follows:

Type	Rank	Freq	Range	S1_TT	S1_Ent	S2_TT	S2_Ent
1 in the + that	1	15	13			0.933	0.991
2 in + case that	2	2	10.5	0.0			
3 in + analysis that	3	1	11.0	0.0			
4 in + brain that	3	1	11.0	0.0			
5 in + chapters that	3	1	11.0	0.0			
6 in + destabilization that	3	1	11.0	0.0			
7 in + fact that	3	1	11.0	0.0			
8 in + fields that	3	1	11.0	0.0			
9 in + grooves that	3	1	11.0	0.0			
10 in + s that	3	1	11.0	0.0			
11 in + ... that	3	1	11.0	0.0			

An 'Open Slot Viewer' window is open for the top result, showing the following data:

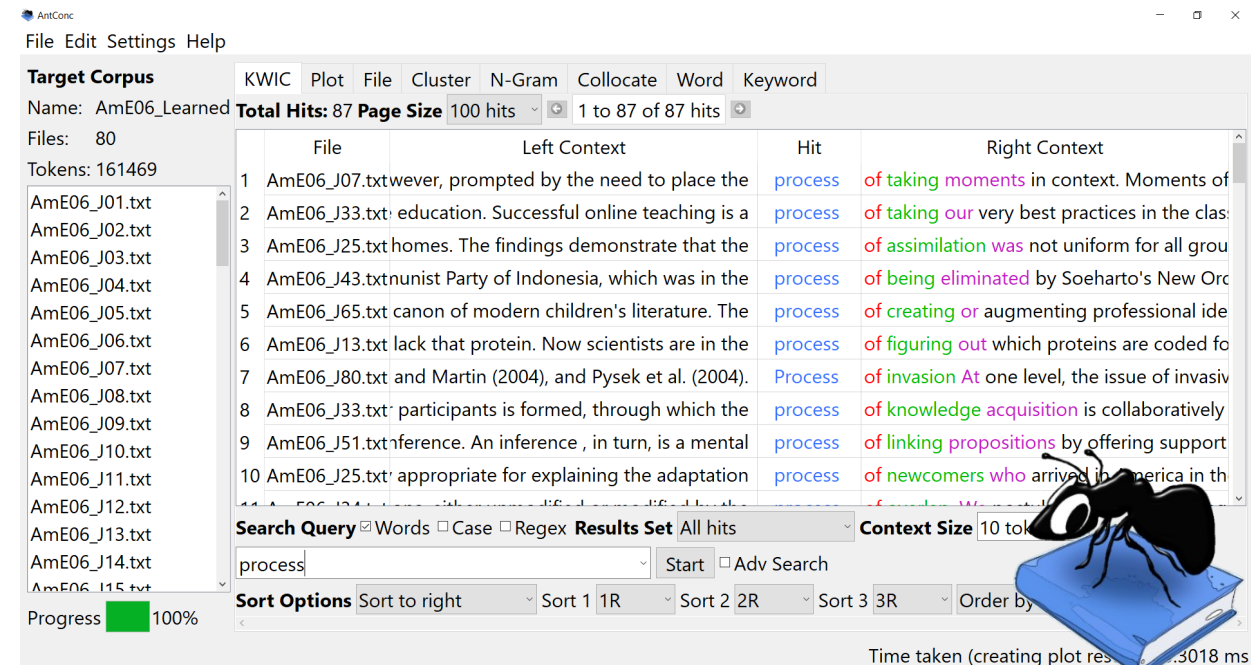
S1	S1 freq	S2	S2 freq
		case	2
		sense	1
		chapters	1
		fact	1
		symptom	1
		brain	1
		sections	1
		destabilization	1

The interface also shows search options like 'Words', 'Case', 'Regex', 'N-Gram Size 4', and 'Open Slots'. A progress bar at the bottom indicates 100% completion. A small illustration of an ant on a book is visible in the bottom right corner.

# Introduction to AntConc 4:

Addressing the challenges of DDL - 4) managing 'data-overload'

- AntConc "KWIC Patterns"
  - KWIC concordance line ordering by pattern frequency (not alphabetical)
- Results "Paging" and "Filtering"
  - all tools show 'paged' (10, 50, 100, ..., all hits) results (cf. Web browser searching)
  - KWIC results can be 'trimmed' (e.g., 10, 20, ... random hits)
- Best practices" for ranking/filtering
  - all tools show frequency & range
  - all tools show normalized values
  - all statistics are 'meaningful'
    - keywords (keyness + effect size)
    - collocates (keyness + effect size)
    - KWIC results (ranked by frequency)



The screenshot displays the AntConc 4.0.0 application window. The 'Target Corpus' is 'AmE06\_Learned'. The search query is 'process', and the results are displayed in a table with columns for File, Left Context, Hit, and Right Context. The results are sorted by frequency, with 'process' being the most frequent hit. The interface includes a menu bar (File, Edit, Settings, Help), a toolbar with various search options (KWIC, Plot, File, Cluster, N-Gram, Collocate, Word, Keyword), and a progress bar at the bottom showing 100% completion. A small illustration of an ant on a book is visible in the bottom right corner.

File	Left Context	Hit	Right Context
AmE06_J07.txt	wever, prompted by the need to place the	process	of taking moments in context. Moments of
AmE06_J33.txt	education. Successful online teaching is a	process	of taking our very best practices in the clas
AmE06_J25.txt	homes. The findings demonstrate that the	process	of assimilation was not uniform for all grou
AmE06_J43.txt	nunist Party of Indonesia, which was in the	process	of being eliminated by Soeharto's New Orc
AmE06_J65.txt	canon of modern children's literature. The	process	of creating or augmenting professional ide
AmE06_J13.txt	lack that protein. Now scientists are in the	process	of figuring out which proteins are coded fo
AmE06_J80.txt	and Martin (2004), and Pysek et al. (2004).	Process	of invasion At one level, the issue of invasiv
AmE06_J33.txt	participants is formed, through which the	process	of knowledge acquisition is collaboratively
AmE06_J51.txt	ference. An inference, in turn, is a mental	process	of linking propositions by offering support
AmE06_J25.txt	appropriate for explaining the adaptation	process	of newcomers who arrived in America in th

# Introduction to AntConc 4:

Addressing the challenges of DDL - 5) incorporating results from corpora in learner language

- AntConc "**KWIC Patterns**" bring a revolutionary change to DDL (and corpus linguistics research in general)
  - ordering KWIC results by pattern frequency...
    - immediately shows salient patterns (without the need for scrolling)
    - (can) reveal extremely surprising results (e.g., sorting to the left vs. right)
    - hugely simplifies the DDL process

AntConc 4.0.0.0

File Edit Settings Help

**Target Corpus**  
Name: AmE06\_Learned  
Files: 80  
Tokens: 161469

AmE06\_J01.txt  
AmE06\_J02.txt  
AmE06\_J03.txt  
AmE06\_J04.txt  
AmE06\_J05.txt  
AmE06\_J06.txt  
AmE06\_J07.txt  
AmE06\_J08.txt  
AmE06\_J09.txt  
AmE06\_J10.txt  
AmE06\_J11.txt  
AmE06\_J12.txt  
AmE06\_J13.txt  
AmE06\_J14.txt  
AmE06\_J15.txt

Progress 100%

KWIC Plot File Cluster N-Gram Collocate Word Keyword

Total Hits: 87 Page Size 100 hits 1 to 87 of 87 hits

File	Left Context	Hit	Right Context
AmE06_J07.txt	wever, prompted by the need to place the	process	of taking moments in context. Moments of
AmE06_J33.txt	education. Successful online teaching is a	process	of taking our very best practices in the clas
AmE06_J25.txt	homes. The findings demonstrate that the	process	of assimilation was not uniform for all grou
AmE06_J43.txt	nunist Party of Indonesia, which was in the	process	of being eliminated by Soeharto's New Orc
AmE06_J65.txt	canon of modern children's literature. The	process	of creating or augmenting professional ide
AmE06_J13.txt	lack that protein. Now scientists are in the	process	of figuring out which proteins are coded fo
AmE06_J80.txt	and Martin (2004), and Pysek et al. (2004).	Process	of invasion At one level, the issue of invasiv
AmE06_J33.txt	participants is formed, through which the	process	of knowledge acquisition is collaboratively
AmE06_J51.txt	ference. An inference, in turn, is a mental	process	of linking propositions by offering support
AmE06_J25.txt	appropriate for explaining the adaptation	process	of newcomers who arrived in America in th

Search Query  Words  Case  Regex Results Set All hits Context Size 10 tokens

process Start  Adv Search

Sort Options Sort to right Sort 1 1R Sort 2 2R Sort 3 3R Order by

Time taken (creating plot res 3018 ms

# Introduction to AntConc 4:

Addressing the challenges of DDL - 5) incorporating results from corpora in learner language

- AntConc "Plot" and range values...
  - help identify texts of interest
    - e.g., frequent use of target words
    - e.g., texts with frequent uses of words/patterns (i.e., exemplar texts - cf. *ProtAnt*)
  - highlight individual variation
  - reveal outlier texts

AntConc  
File Edit Settings Help

**Target Corpus**  
Name: AmE06\_Learned **Total Hits: 106 Total Files With Hits: 34**

Files: 80  
Tokens: 161469

DocID	DocPath	DocTokens	Freq	NormFreq	Dispersion	Plot
1	AmE06_J28.txt	2021	4	1979.218	0.592	
2	AmE06_J80.txt	1966	5	2543.235	0.553	
3	AmE06_J02.txt	2036	4	1964.637	0.447	
4	AmE06_J51.txt	2000	4	2000.000	0.447	
5	AmE06_J55.txt	2071	4	1931.434	0.447	
6	AmE06_J77.txt	1995	4	2005.013	0.447	
7	AmE06_J05.txt	2025	9	4444.444	0.440	
8	AmE06_J25.txt	1968	7	3556.911	0.396	
9	AmE06_J33.txt	2029	12	5914.243	0.345	
10	AmE06_J13.txt	2051	4	1950.268	0.333	


Search Query  Words  Case  Regex **Results Set** All hits **Plot Zoom** 1.00 x

process of   Adv Search

Sort by Dispersion  Invert Order

Progress 100%

Time taken (creating plot res... 2925 ms

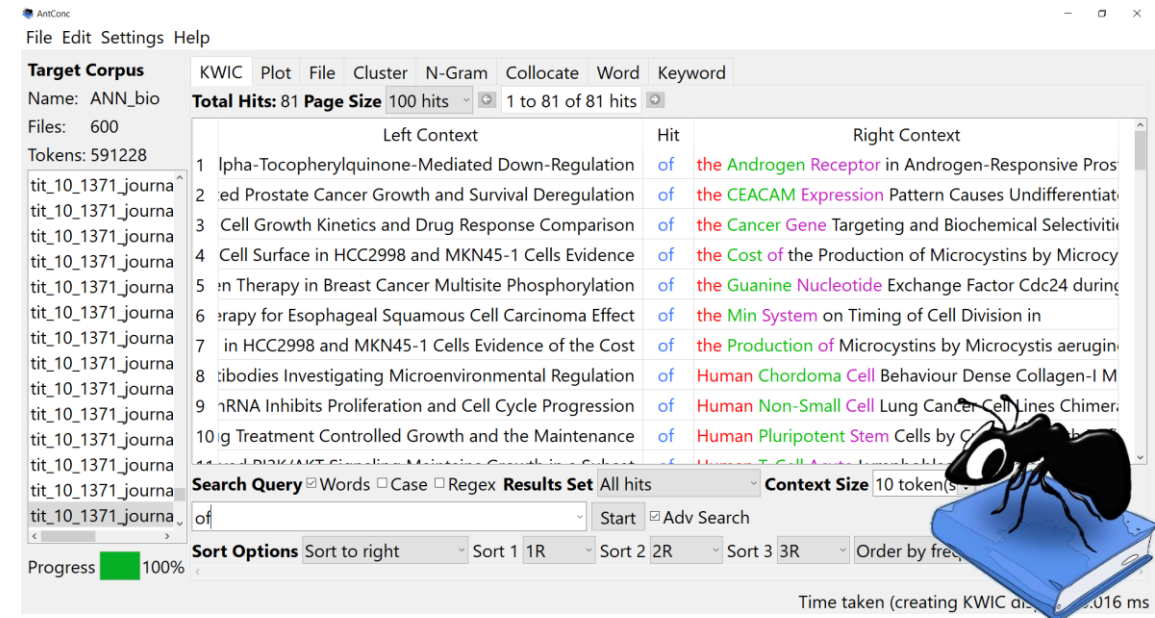


# Introduction to AntConc 4:

Addressing the challenges of DDL - 6) going beyond surface-level word/phrase patterns

## ■ AntConc "Text-, Sub-Text, Word-Level Processing"

- corpora can be saved with multi-level metadata [e.g., word, part-of-speech (POS), LEMMA, text-level annotations]
- part-of-speech (POS) can be searched and viewed in a transparent way
- all searches can be combined with metadata queries via the advanced search option
  - e.g., "adjective + noun" patterns in novels
  - e.g., "however" as used in abstracts
  - ...



The screenshot displays the AntConc application window. The 'Target Corpus' is 'ANN\_bio' with 600 files and 591,228 tokens. The search query is 'of', and the results are sorted by frequency. The interface shows a list of search results with columns for 'Left Context', 'Hit', and 'Right Context'. The search query is 'of', and the results are sorted by frequency. The interface also shows a progress bar at 100% and a time taken of 0.016 ms.

Left Context	Hit	Right Context
1 lpha-Tocopherylquinone-Mediated Down-Regulation	of	the Androgen Receptor in Androgen-Responsive Pros
2 ed Prostate Cancer Growth and Survival Deregulation	of	the CEACAM Expression Pattern Causes Undifferentiat
3 Cell Growth Kinetics and Drug Response Comparison	of	the Cancer Gene Targeting and Biochemical Selectivit
4 Cell Surface in HCC2998 and MKN45-1 Cells Evidence	of	the Cost of the Production of Microcystins by Microcy
5 n Therapy in Breast Cancer Multisite Phosphorylation	of	the Guanine Nucleotide Exchange Factor Cdc24 during
6 rapy for Esophageal Squamous Cell Carcinoma Effect	of	the Min System on Timing of Cell Division in
7 in HCC2998 and MKN45-1 Cells Evidence of the Cost	of	the Production of Microcystins by Microcystis aerugin
8 ibodies Investigating Microenvironmental Regulation	of	Human Chordoma Cell Behaviour Dense Collagen-I M
9 rRNA Inhibits Proliferation and Cell Cycle Progression	of	Human Non-Small Cell Lung Cancer Cell Lines Chimer
10 g Treatment Controlled Growth and the Maintenance	of	Human Pluripotent Stem Cells by C

# Introduction to AntConc 4:

Addressing the challenges of DDL - 7) focusing on language (not technology)

- AntConc "Classroom-proof" design
  - packaging and cost
    - multiplatform app (Windows, MacOS, Linux)
    - digitally signed and notarized by Windows and MacOS (no security warnings)
    - simple installation (or use as a portable app)
    - freeware
  - design
    - clean, familiar interface
    - scaling of fonts for in-class demonstrations
    - drag n' drop corpus creation and export of results
  - speed/scalability
    - very fast searching (millisecond response times)
    - works with corpora of **thousands** to **billions** of words (via a fully indexed database)
    - incorporates the latest 'best practice', 'meaningful' statistical methods



---

## Summary and Conclusions

---

# Summary and conclusions:

- Data-Driven Learning (DDL) is...
  - a combination of language data science and active learning
    - "learning how to create, search, analyze, and interpret general and specialized language databases (corpora)"
  - an effective approach in second and foreign language learning
    - "it works!"
- Data-Driven Learning (DDL) is often considered to be challenging and better suited to intermediate/advanced learners because...
  - teachers expose students to tools designed for other purposes/users
    - e.g., tools designed for corpus linguists, lexicographers, and/or materials developers
  - the traditional, core "KWIC" view is extremely poor at revealing patterns
- AntConc 4 attempts to...
  - address these core challenges
  - allow DDL to be used in a much broader range of contexts

