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

An Introduction to AntConc 4

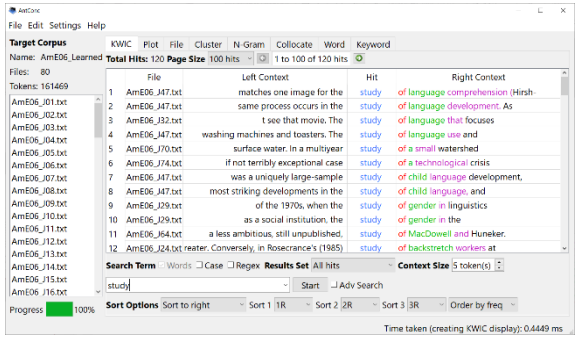
Laurence Anthony
Center for English Language Education in Science and Engineering (CELESE)
Faculty of Science and Engineering, Waseda University
http://www.laurenceanthony.net/
anthony@waseda.jp
@antlabjp

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

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

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

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'

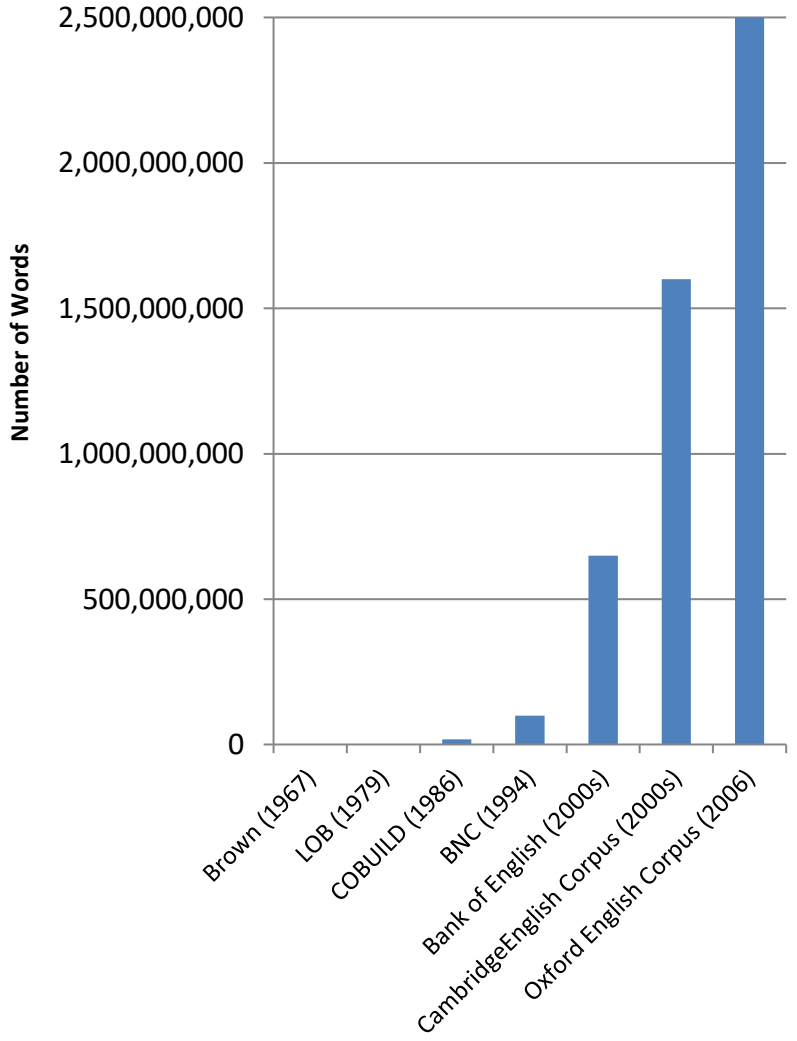
Jablonkai, R. J. (2022). **Transforming language teaching with corpora - Bridging research and practice**. International Perspectives on Corpus Technology for Language Learning - Seminar Series. Univ. of Queensland: Online.

- 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

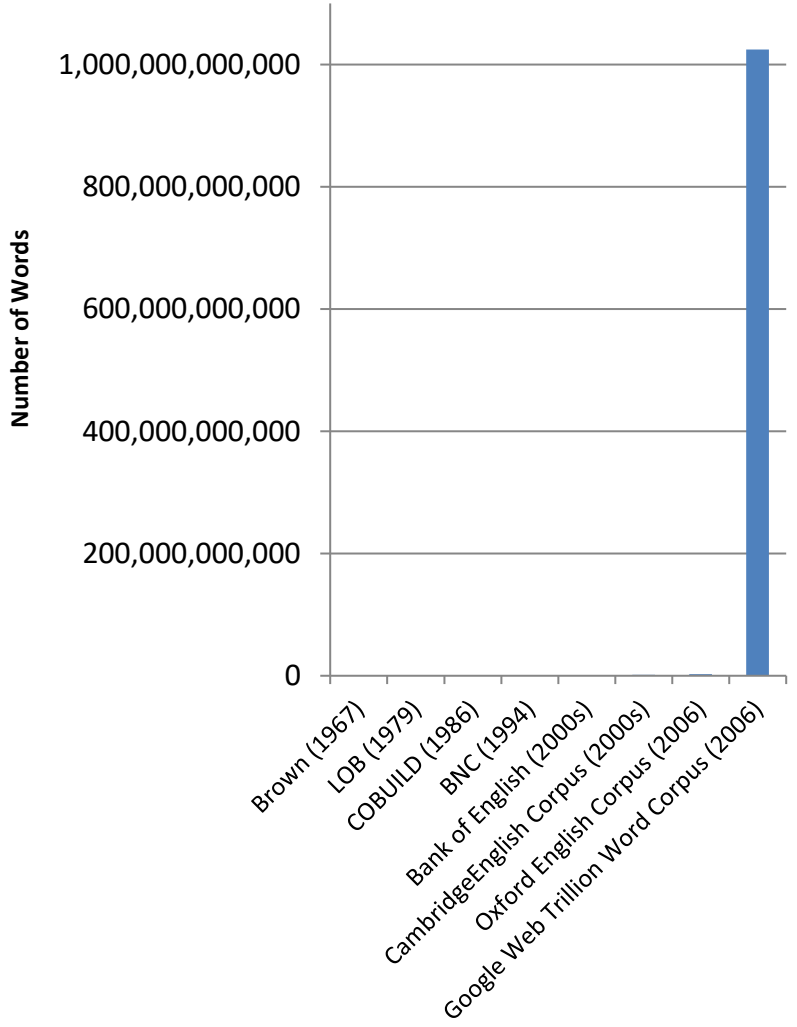
- 1960s-1970s
  - Brown Corpus (1964) - 1 m
  - LOB Corpus (1978) - 1 m
- 1980s-1990s
  - COBUILD Corpus (1986) - 18 m
  - British National Corpus (1994) - 100 m
- 2000s
  - Bank of English (2008) - 650 m
  - Cambridge English Corpus (2013) - 1.6 b
  - Oxford English Corpus (2013) - 2.5 b
  - Google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



Challenges in Data-Driven Learning (DDL)

Challenges resulting from technical developments - corpus sizes

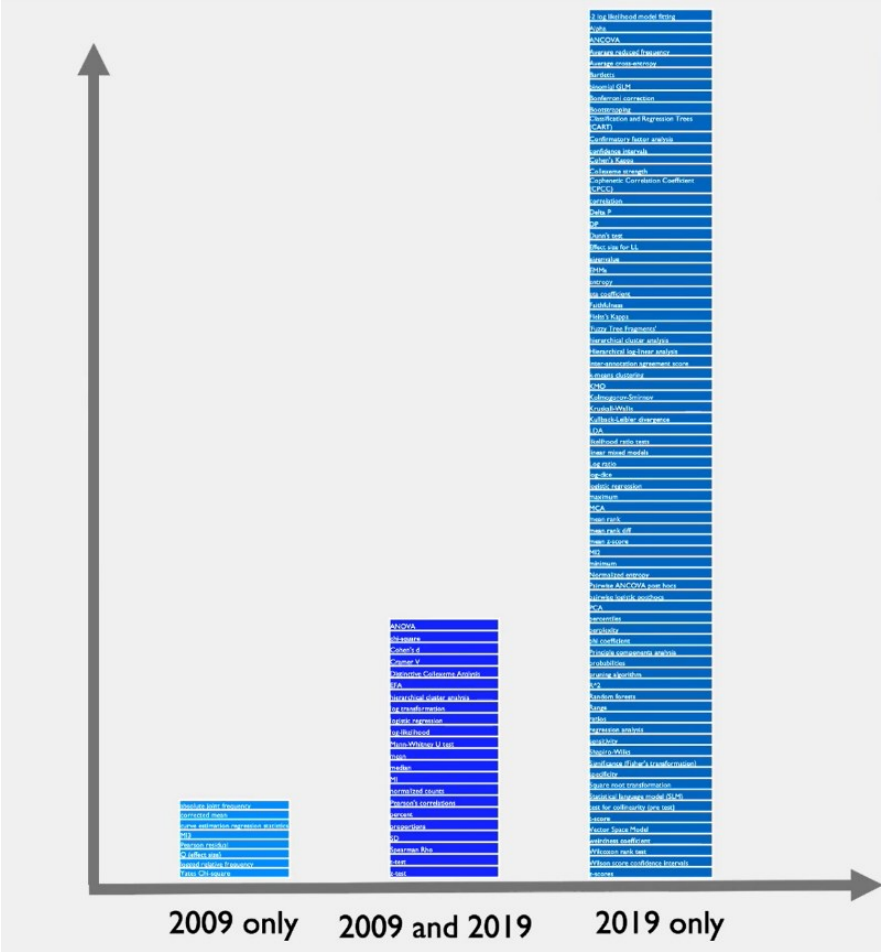
- 1960s-1970s
  - Brown Corpus (1964) - 1 m
  - LOB Corpus (1978) - 1 m
- 1980s-1990s
  - COBUILD Corpus (1986) - 18 m
  - British National Corpus (1994) - 100 m
- 2000s
  - Bank of English (2008) - 650 m
  - Cambridge English Corpus (2013) - 1.6 b
  - Oxford English Corpus (2013) - 2.5 b
  - Google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

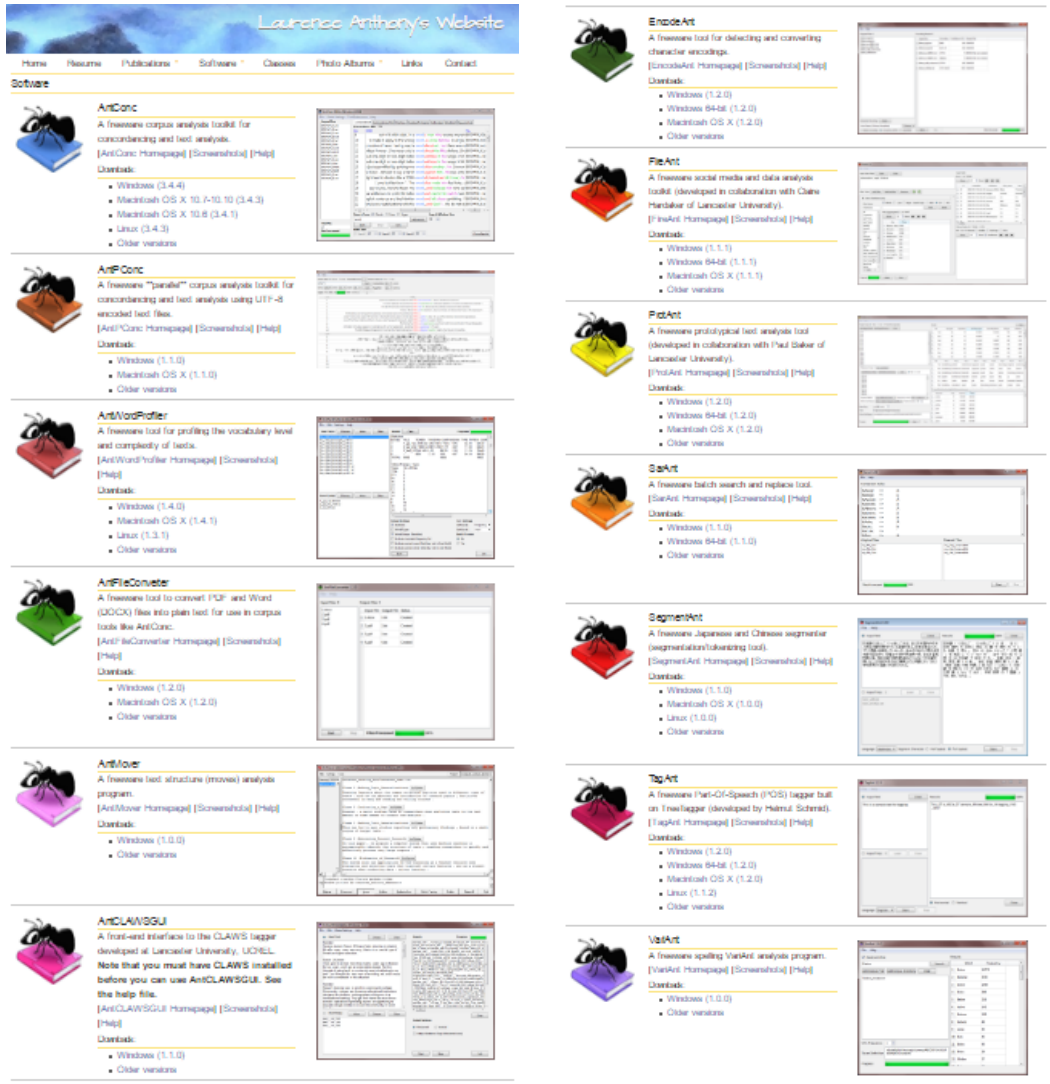
- Toward increased accuracy and transparency
- Toward linguistic interpretability
  - “Perhaps the most serious risk to researchers using available tools is that many of the quantitative measures provided by corpus analysis software do not have transparent linguistic interpretations."

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



Data Collection Tools

AntCorGen

AntFileConverter

AntFileSplitter

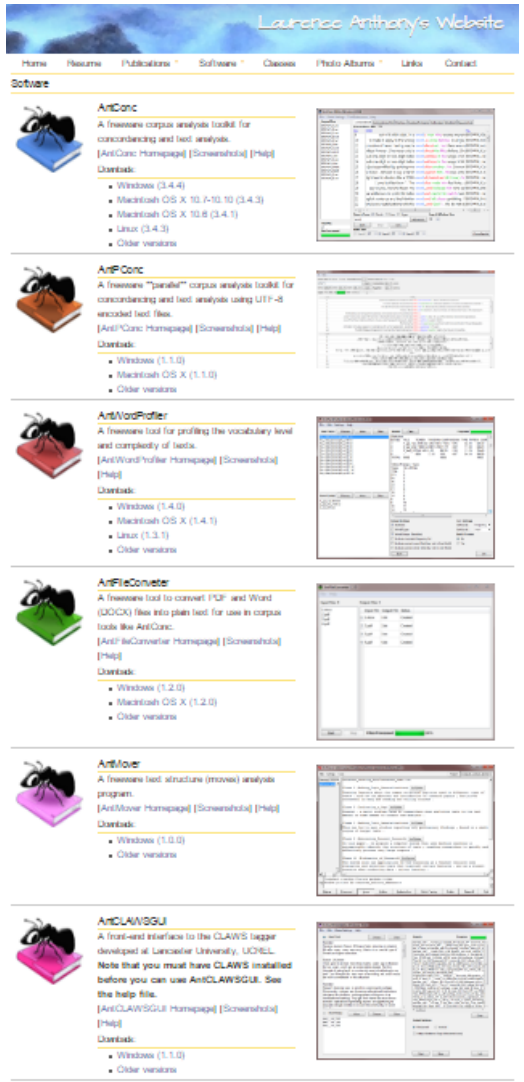
FireAnt

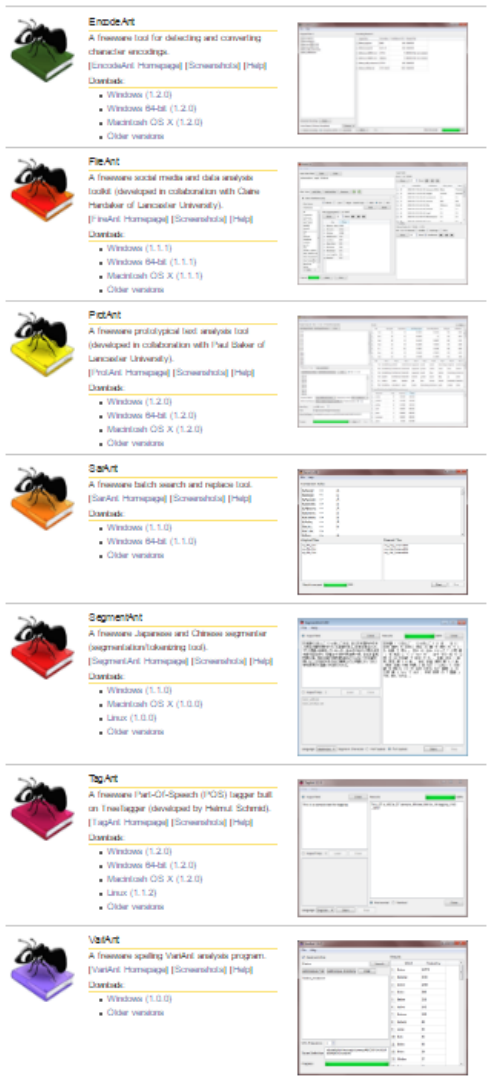
AntLab Tools

www.laurenceanthony.net/software

Introduction to AntConc 4:

A member of the AntLab tools family





Data Cleaning Tools

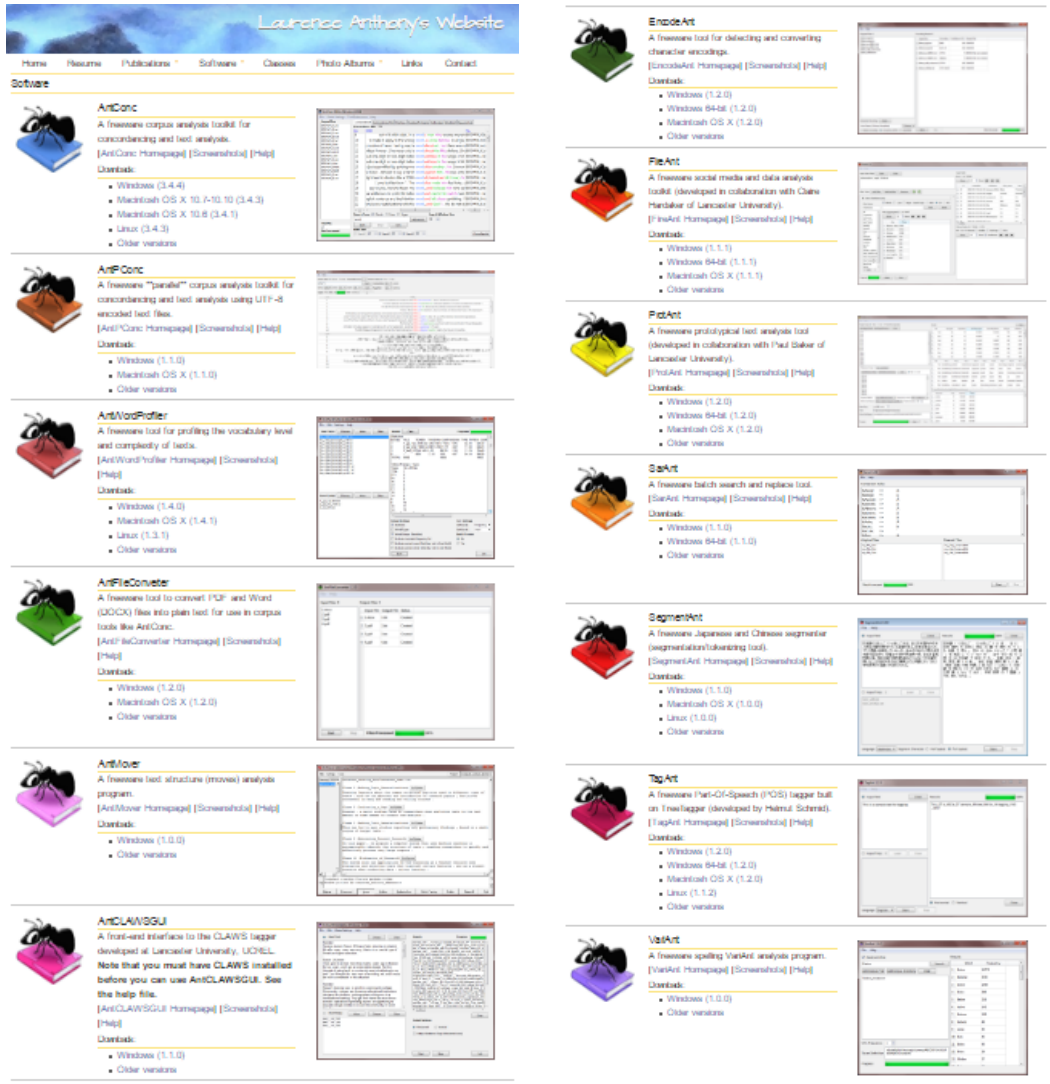
- EncodeAnt
- SarAnt
- SegmentAnt

AntLab Tools

www.laurenceanthony.net/software

Introduction to AntConc 4:

A member of the AntLab tools family



Data Annotation Tools

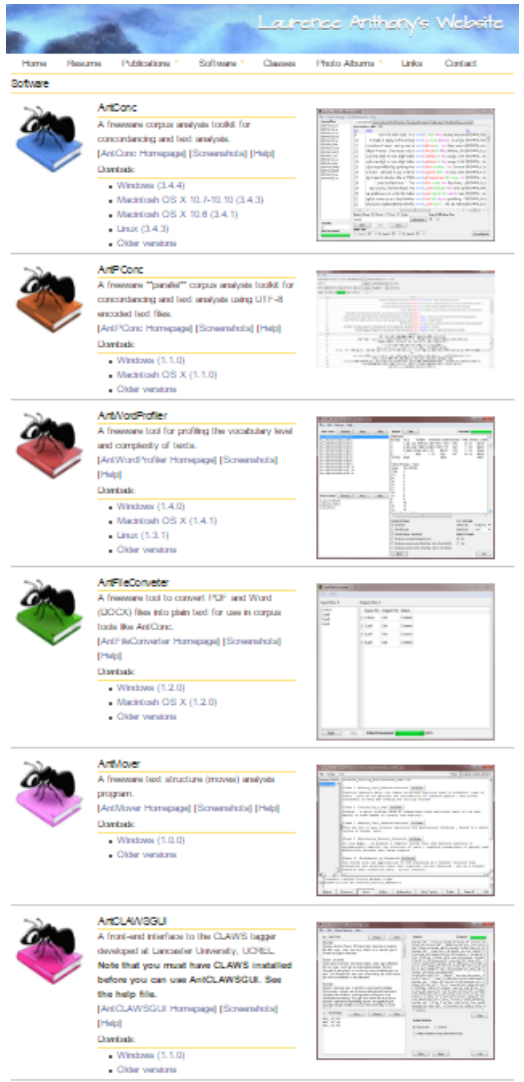
- AntMover
- CLAWSAnt
- SegmentAnt
- TagAnt

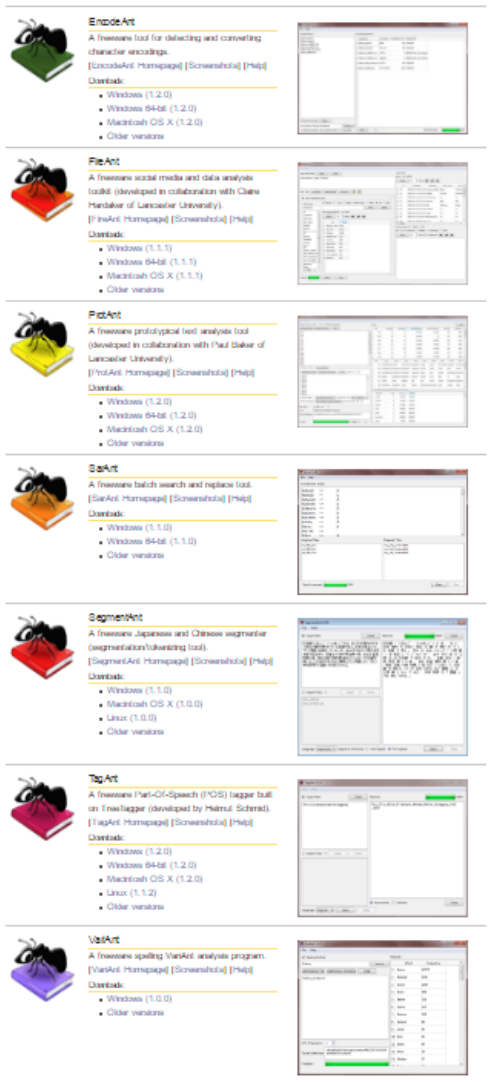
AntLab Tools

www.laurenceanthony.net/software

Introduction to AntConc 4:

A member of the AntLab tools family





Data Analysis Tools

- AntConc
- AntPConc
- AntWordProfiler
- FireAnt
- ProtAnt
- VariAnt

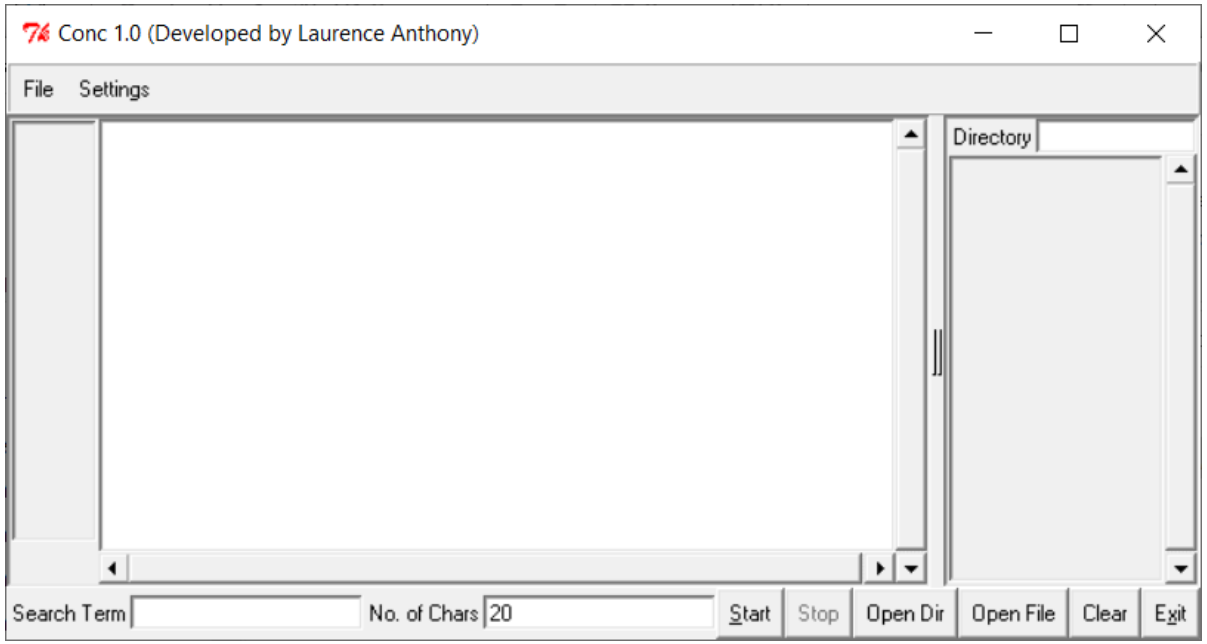
AntLab Tools

www.laurenceanthony.net/software

Introduction to AntConc 4:

Evolution over 20 years

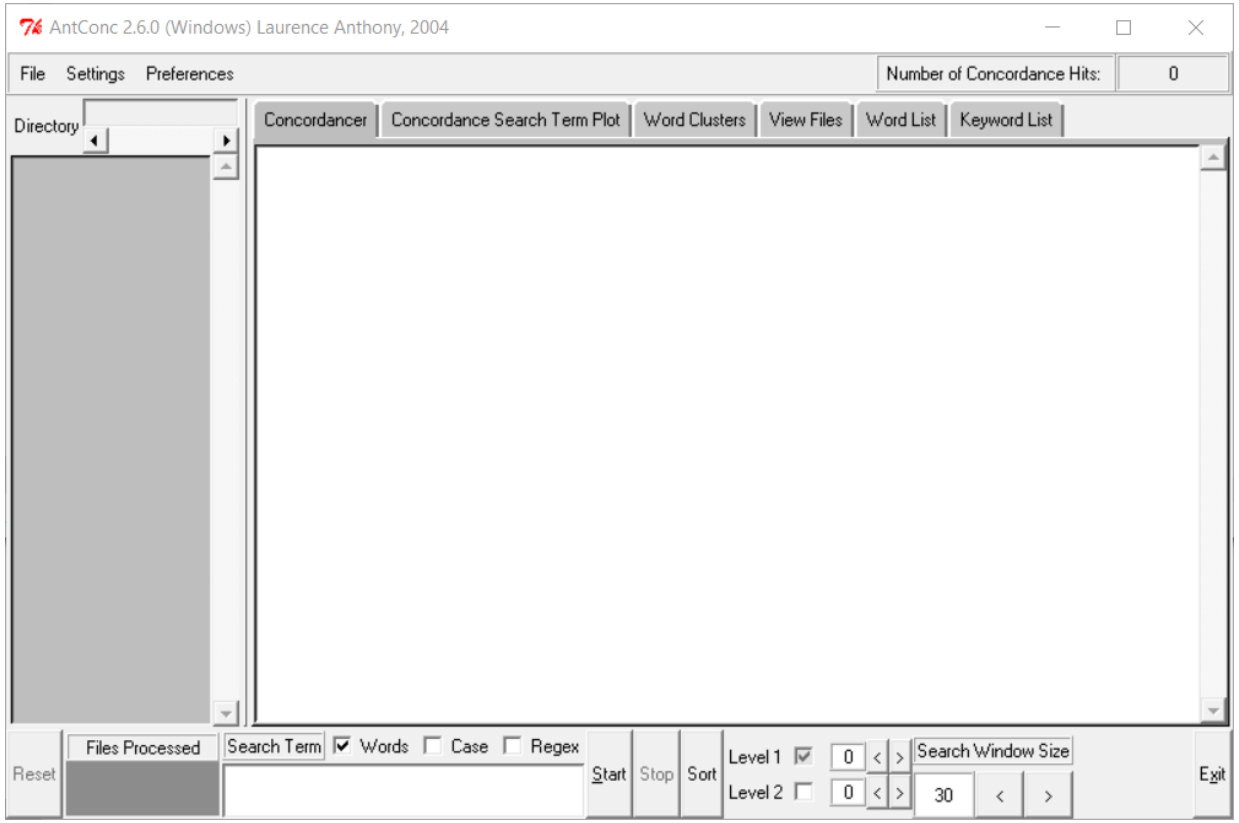
- Major updates
  - Version 1.0-1.1: Simple concordancer for Windows and Linux
    - tools: concordance
    - language: Perl + Tk



Introduction to AntConc 4:

Evolution over 20 years

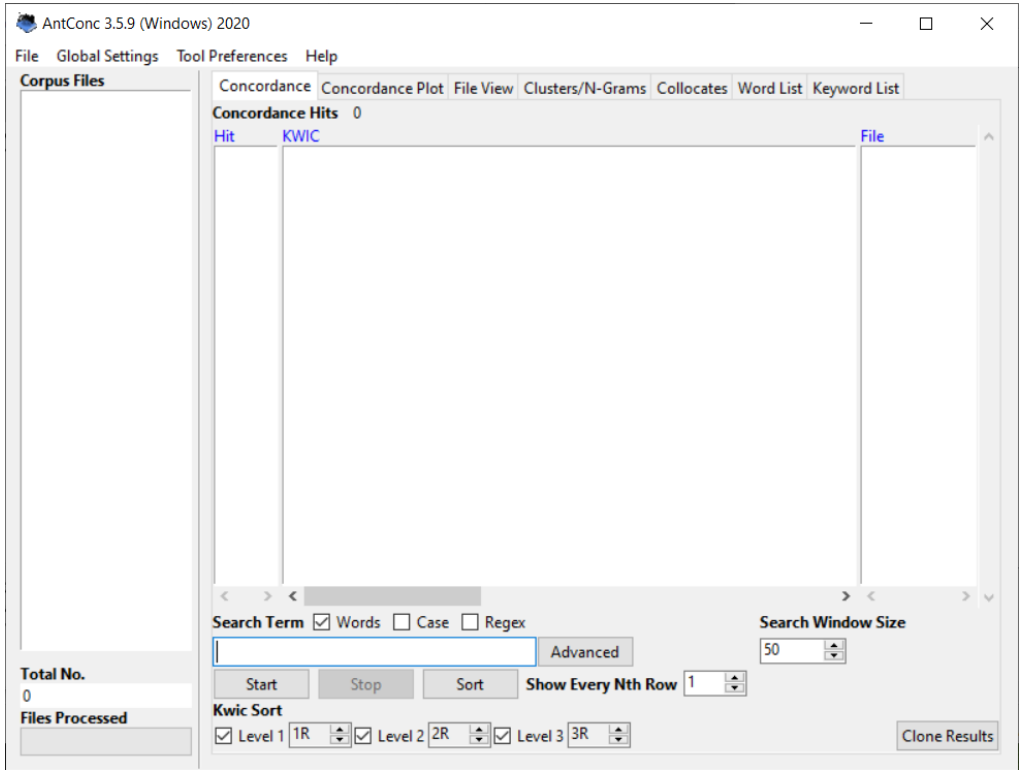
- Major updates
  - Version 2.0-2.6: Corpus toolkit for Windows and Linux
    - tools: concordance, plot, cluster/n-grams, file, word, keyword
    - language: Perl + Tk



Introduction to AntConc 4:

Evolution over 20 years

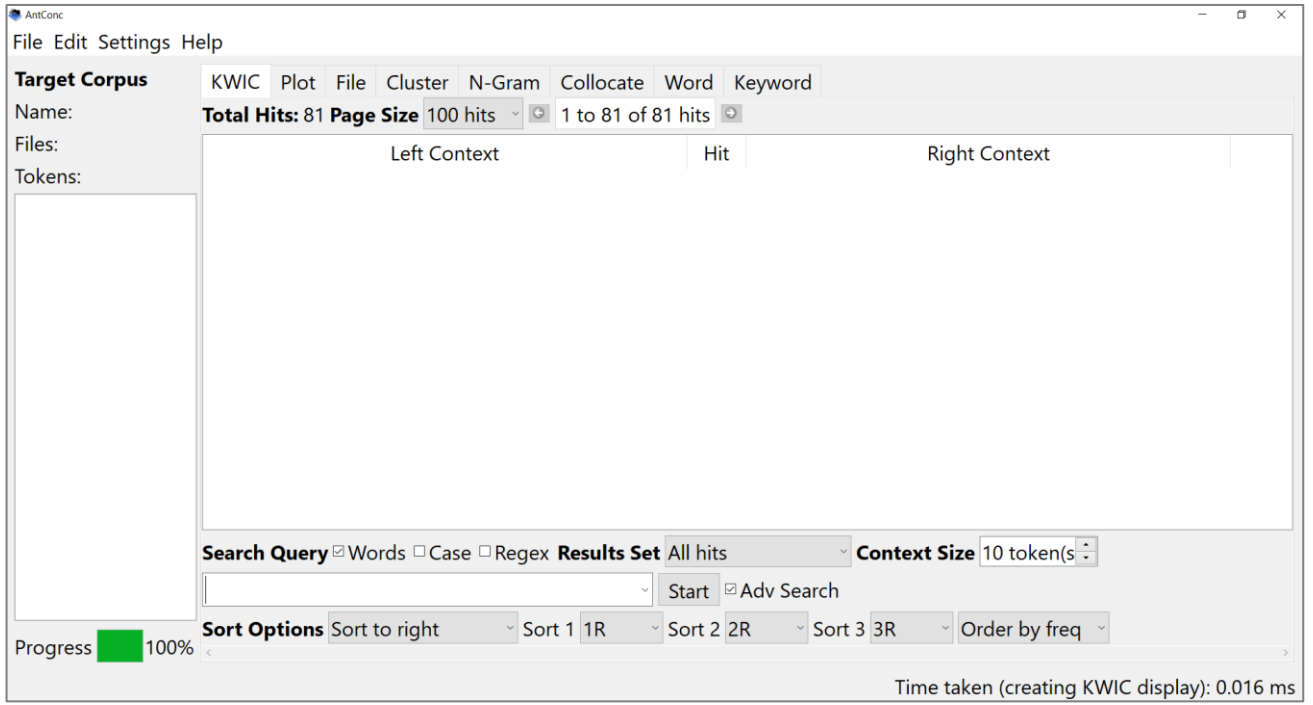
- Major updates
  - Version 3.0-3.5: Corpus toolkit for Windows, Linux, and MacOS (native)
    - tools: concordance, plot, file, cluster/n-grams, collocates, word, keyword
    - language: Perl + Tkx



Introduction to AntConc 4:

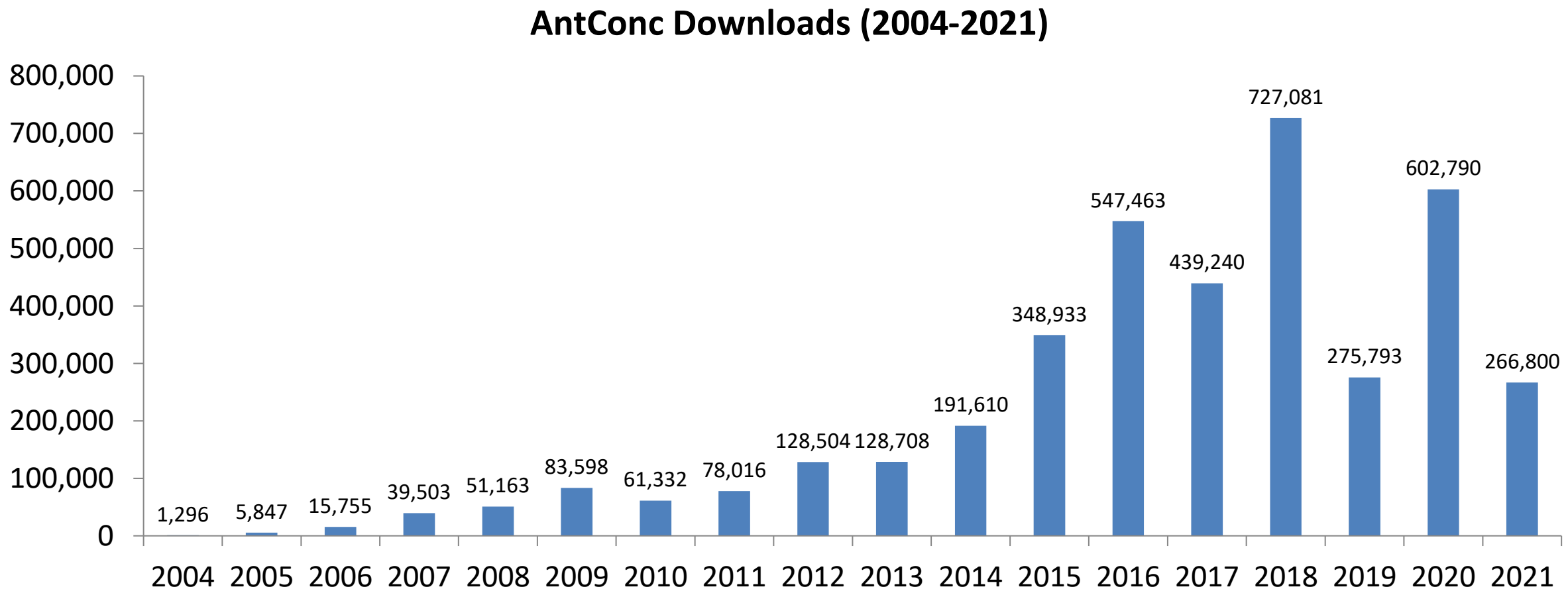
Evolution over 20 years

- 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:

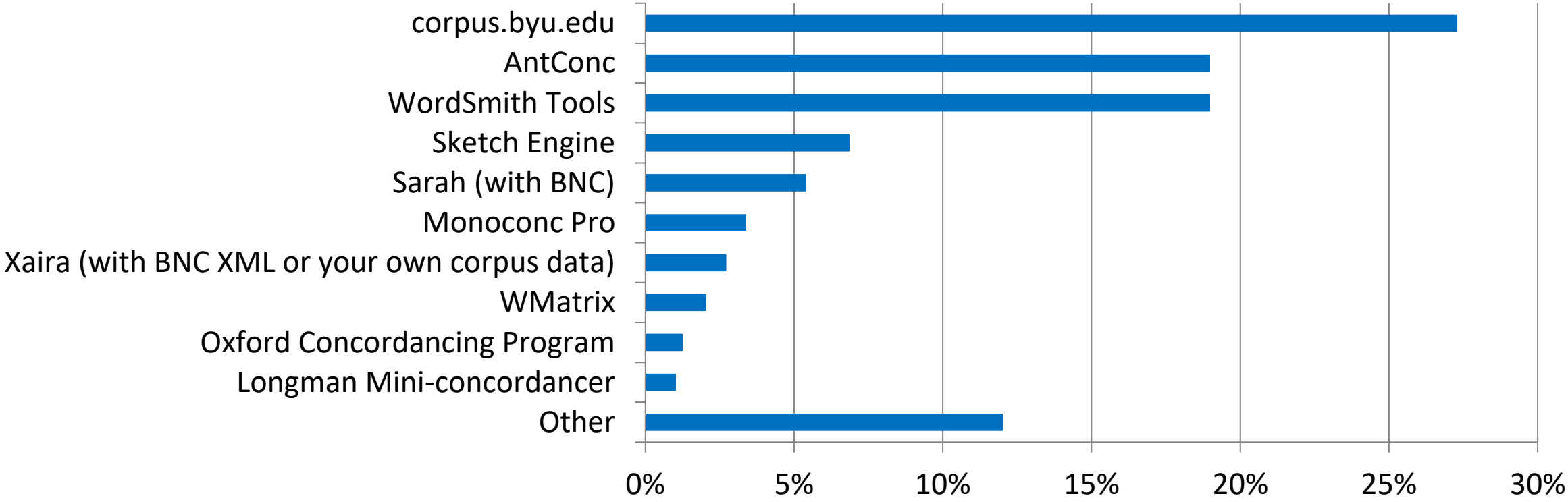
Evolution over 20 years



used in over 140 countries; tutorial videos viewed over 500,000 times

Introduction to AntConc 4:

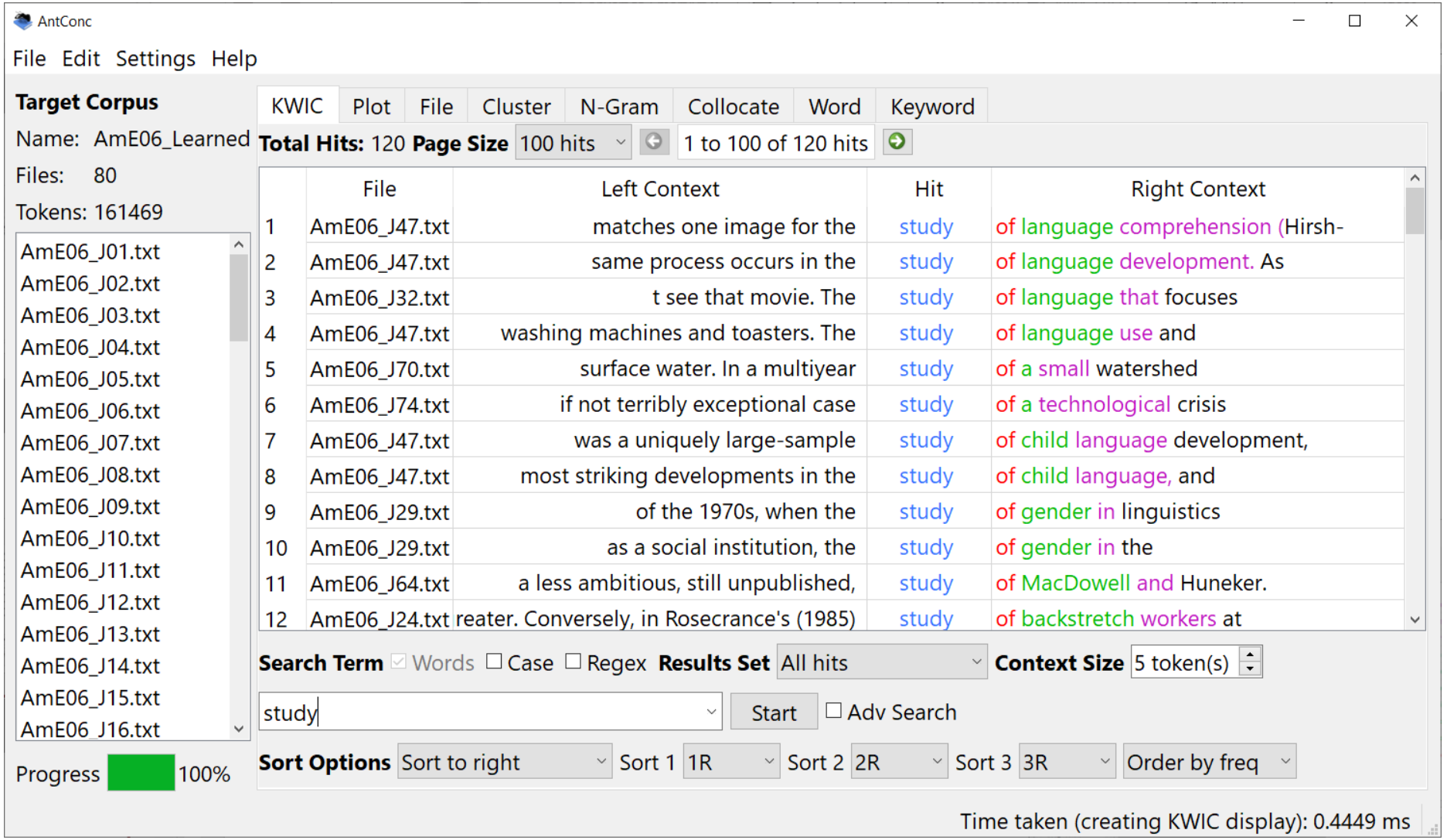
Evolution over 20 years



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

Introduction to AntConc 4:

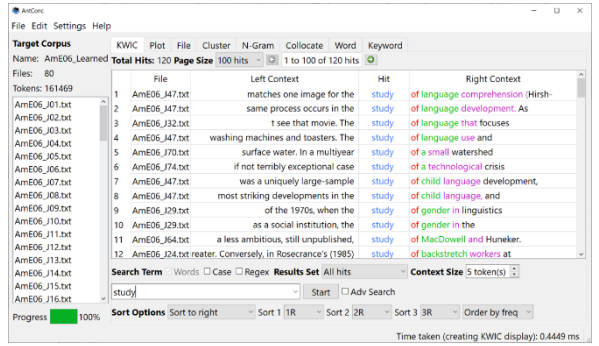
Evolution over 20 years



Introduction to AntConc 4:

Design concept - strengths

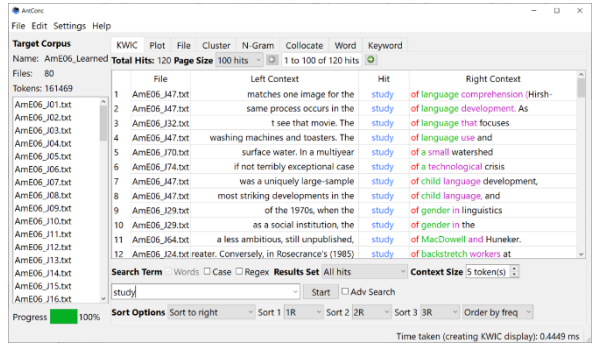
- easy to find, set up, and use
  - e.g., installer or portable version (no installation required)
  - e.g., help guides, YouTube tutorial videos, discussion forums, …
- safe and reliable
  - e.g., no viruses, community error reports, bug fixes, updates, …
- evolving
  - e.g., new tools, new features, 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:

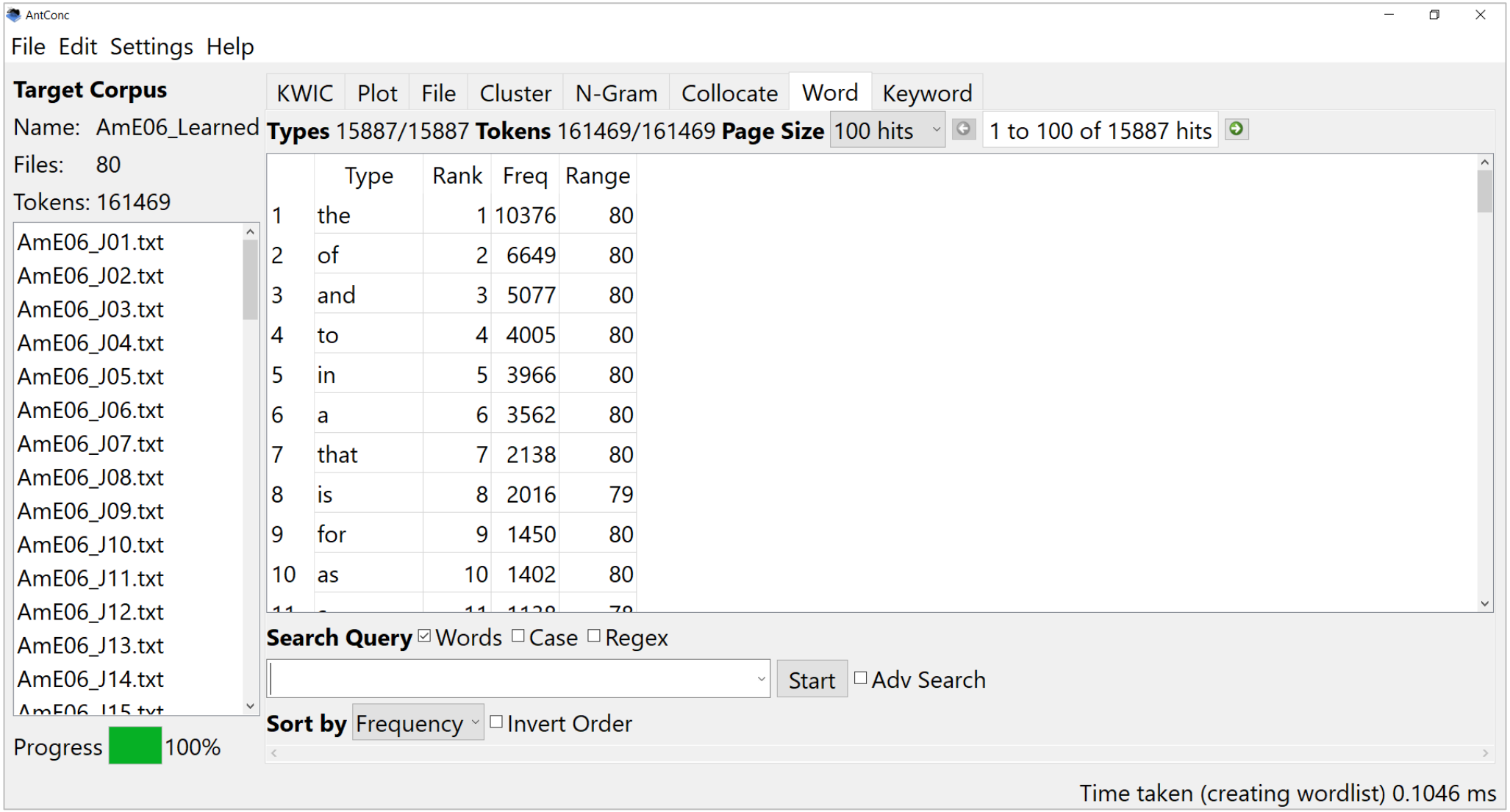
Design concept - weaknesses

- 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



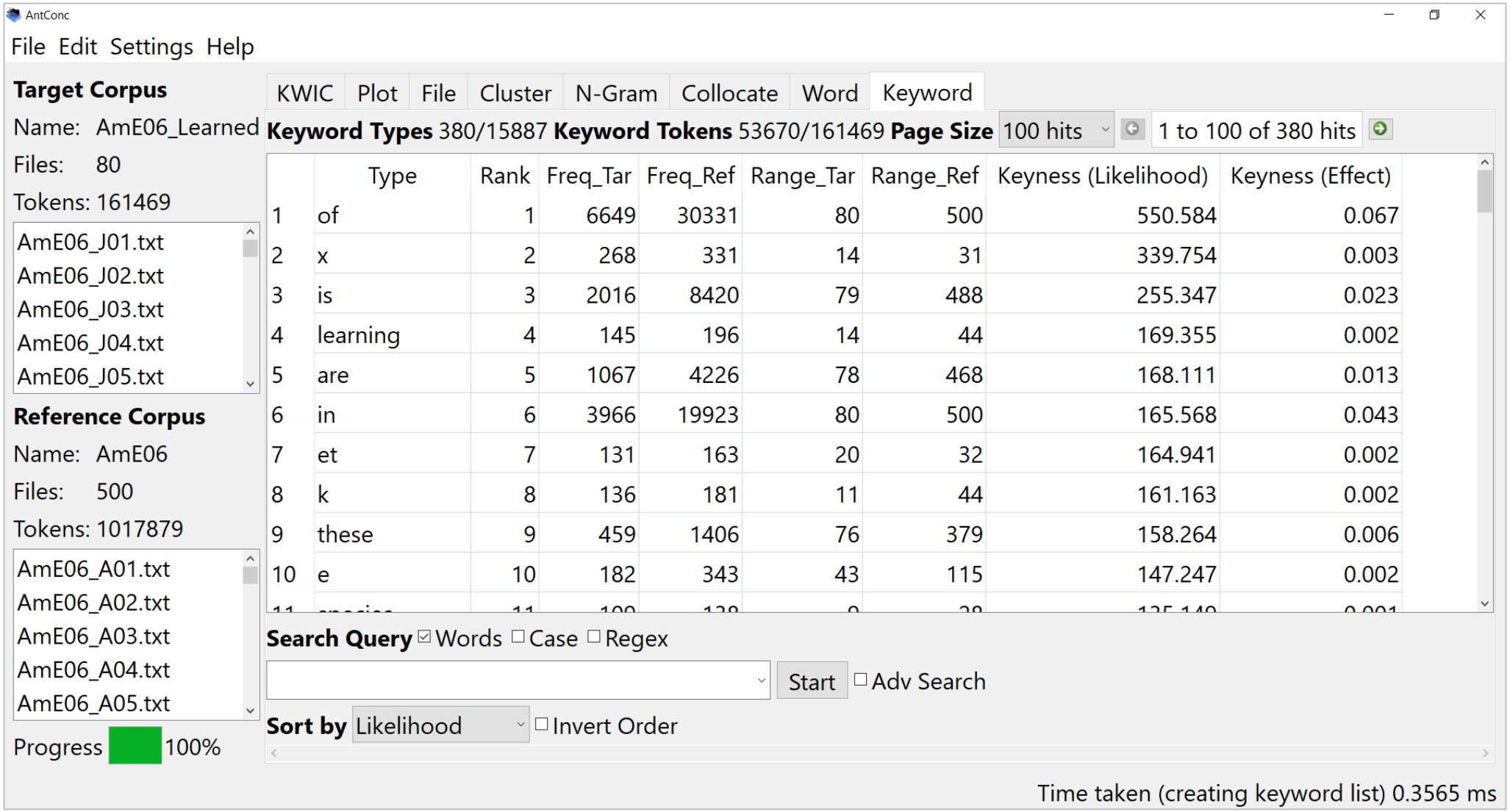
Introduction to AntConc 4:

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



Introduction to AntConc 4:

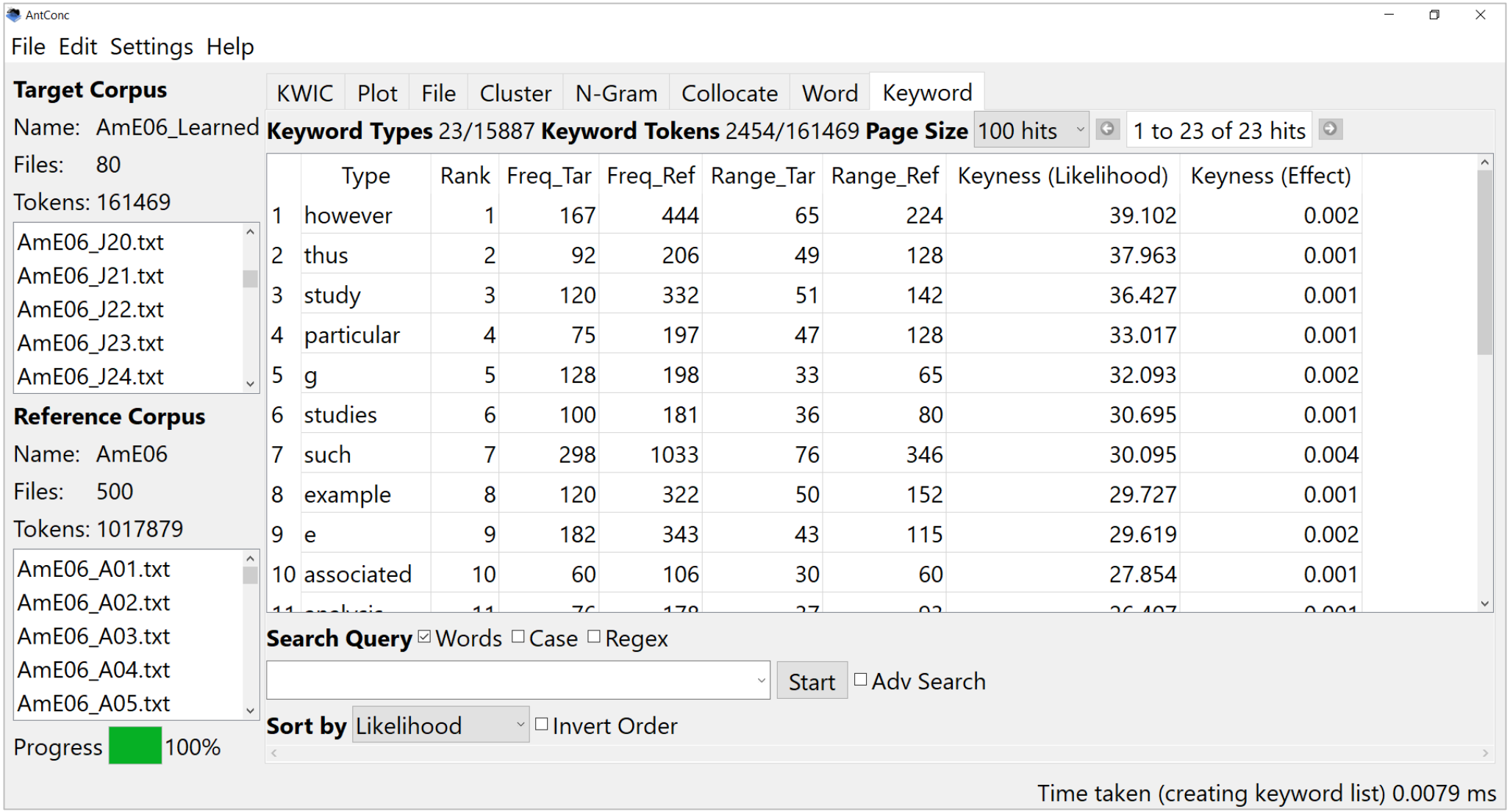
Features and functions - KWIC, Plot, File, Cluster, N-Gram, Collocate, Word, Keyword



	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

Introduction to AntConc 4:

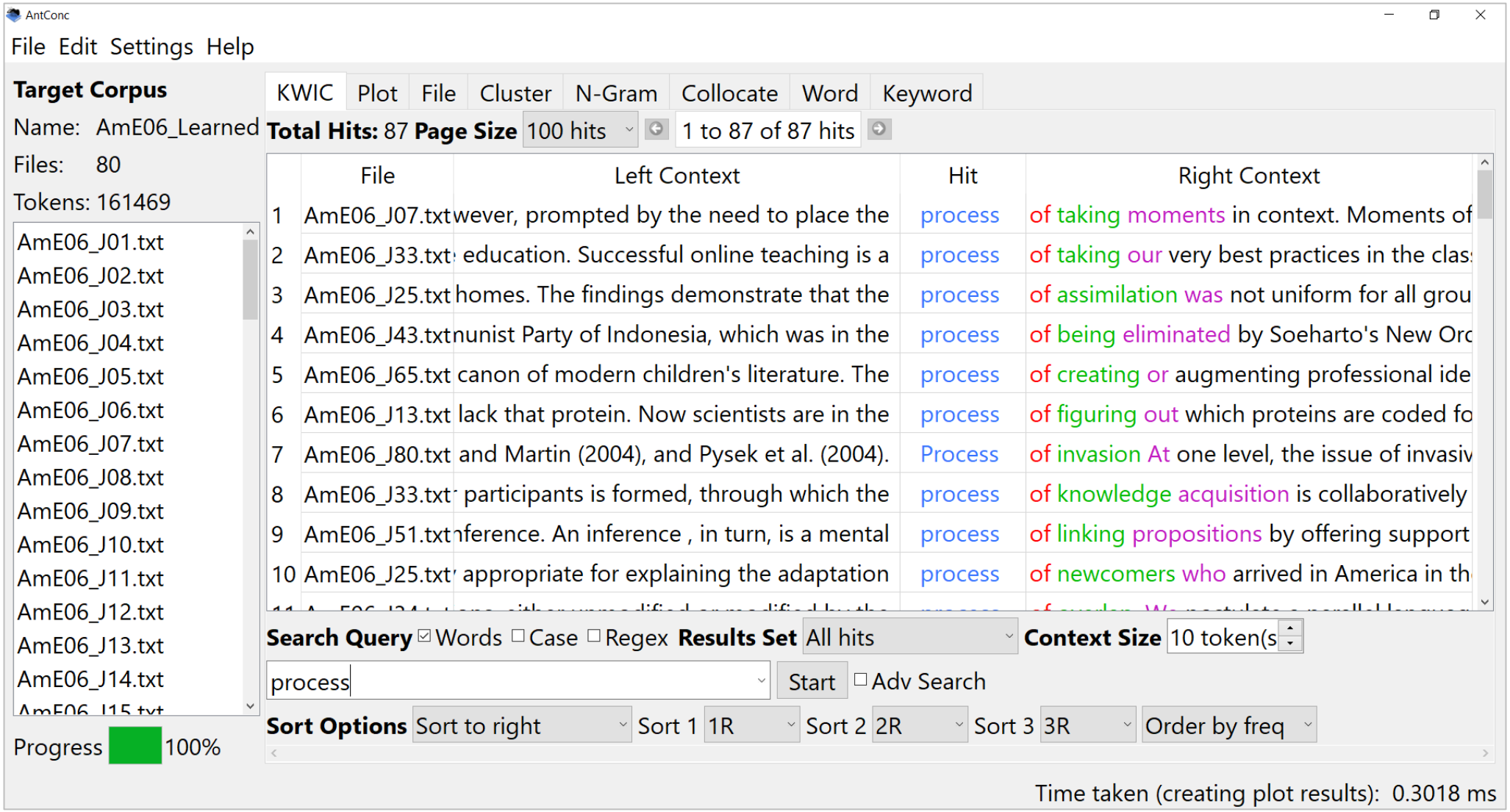
Features and functions - KWIC, Plot, File, Cluster, N-Gram, Collocate, Word, Keyword



	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

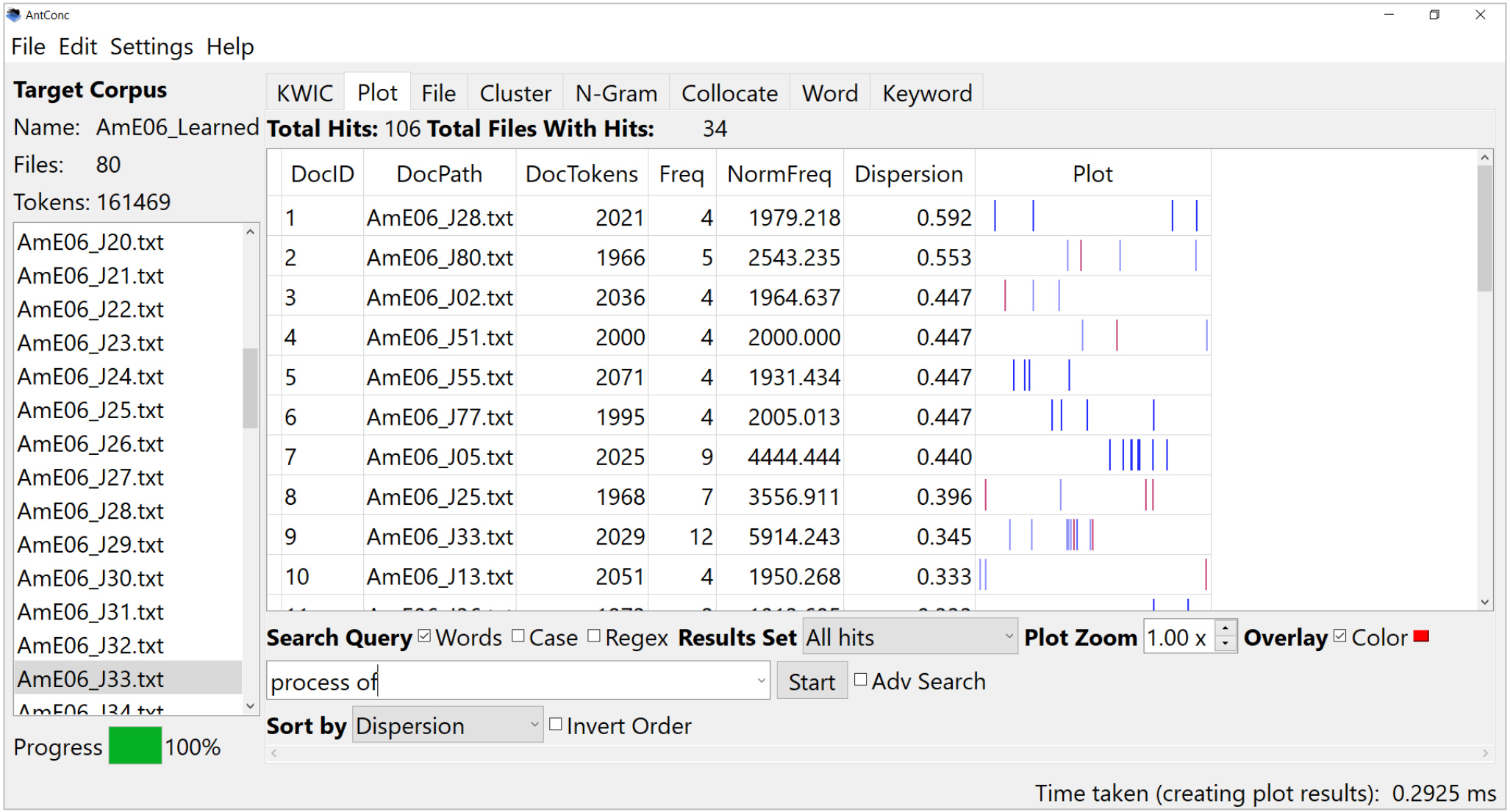
Introduction to AntConc 4:

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



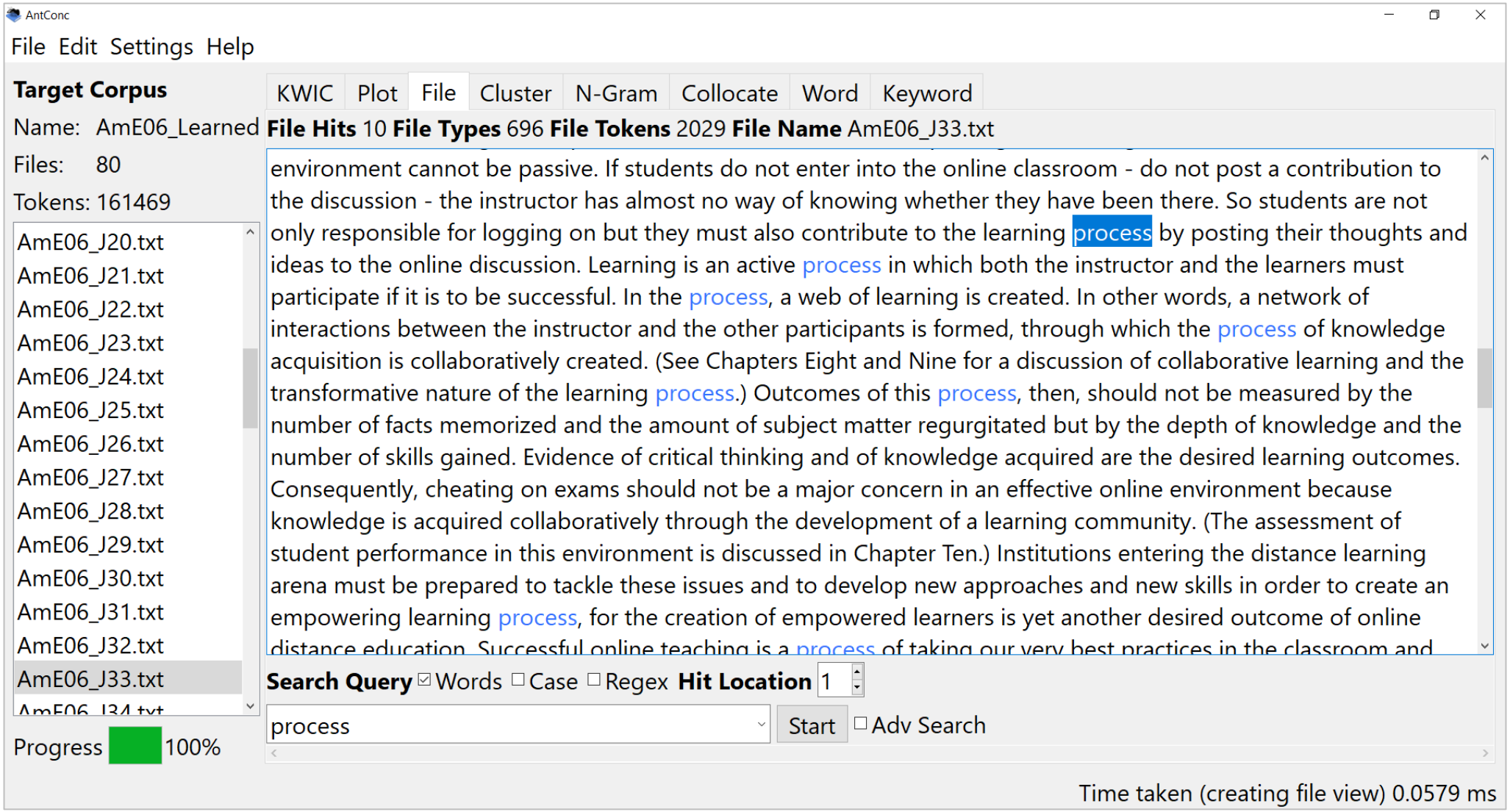
Introduction to AntConc 4:

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



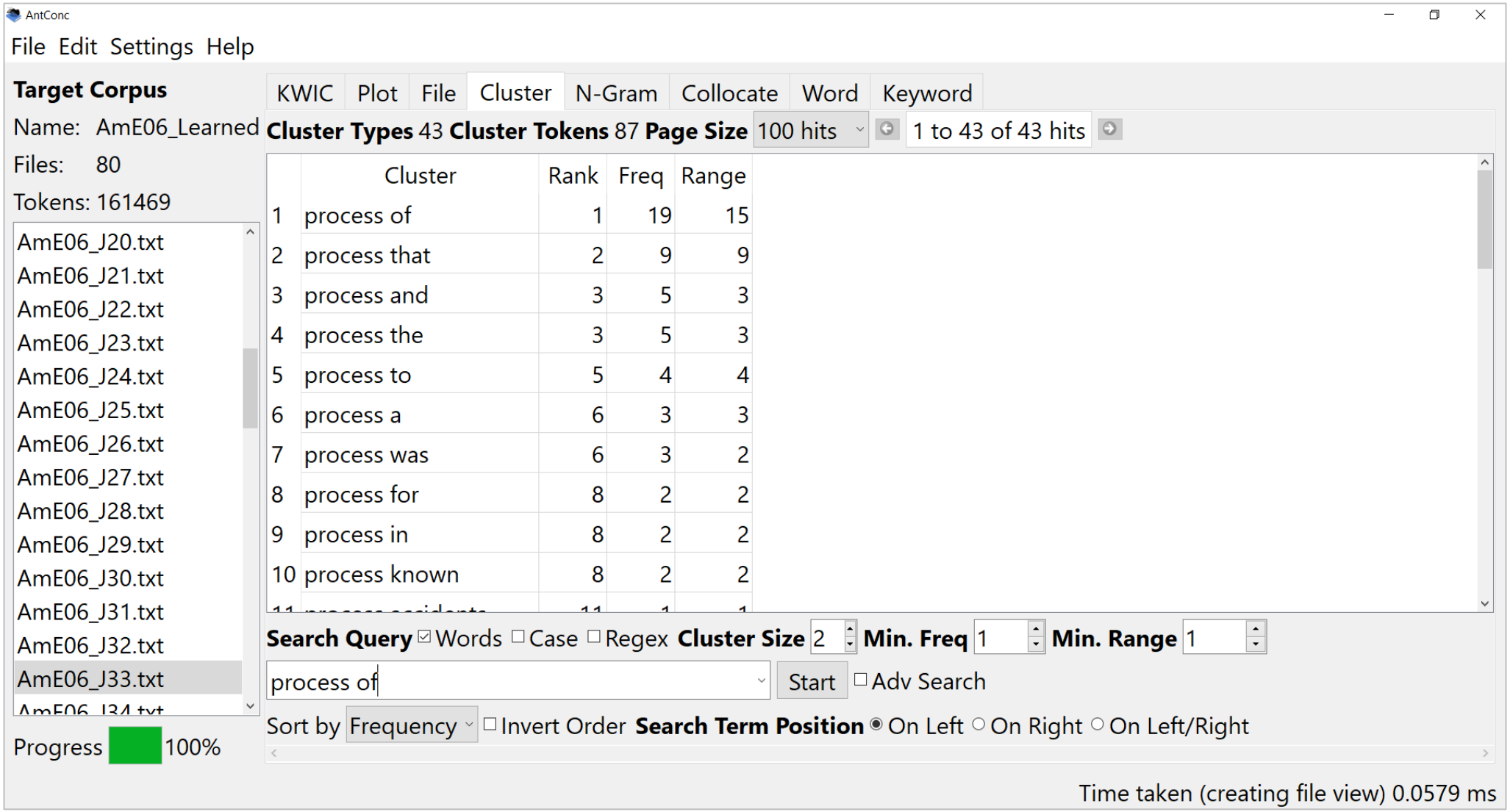
Introduction to AntConc 4:

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



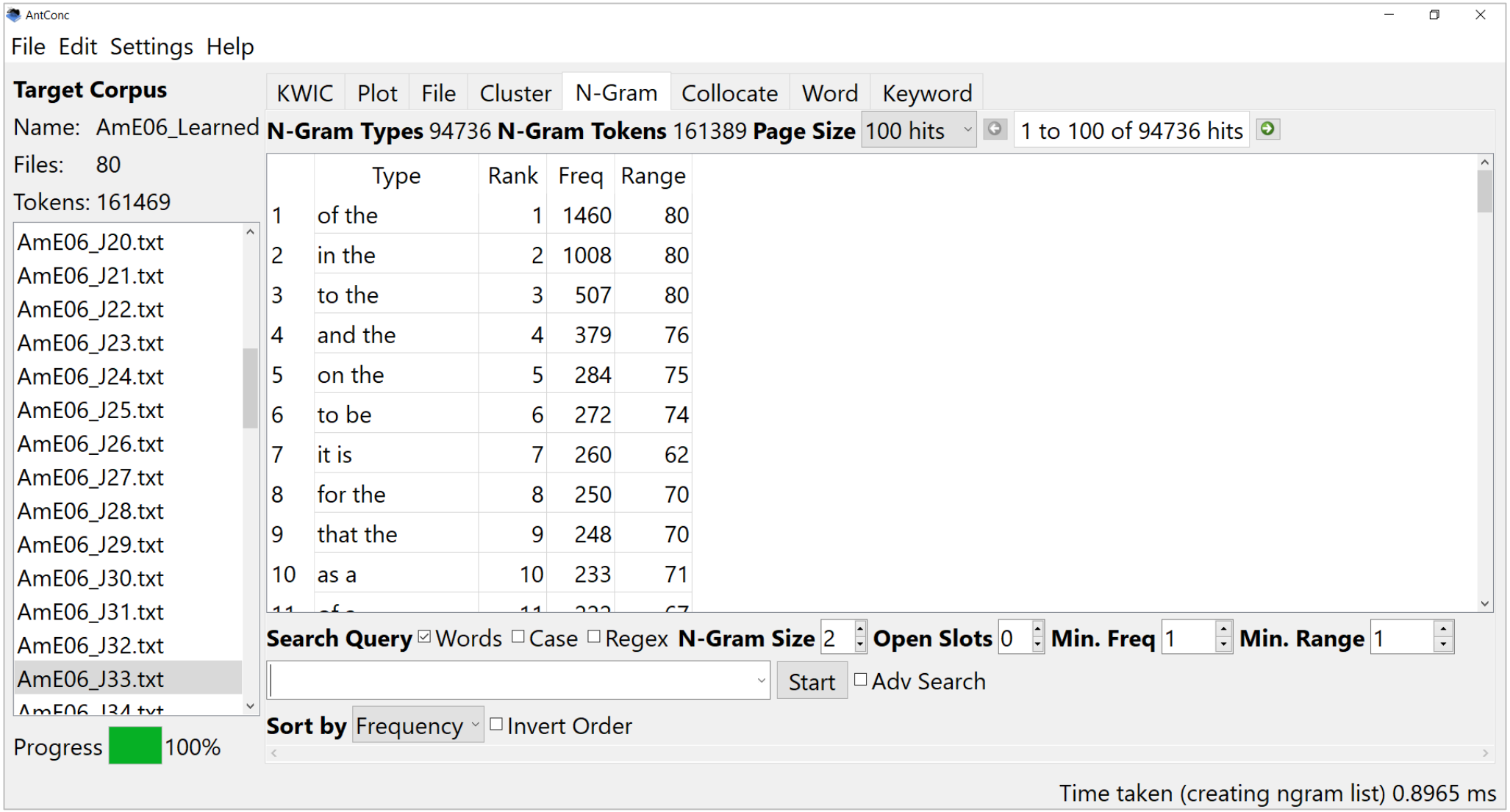
Introduction to AntConc 4:

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



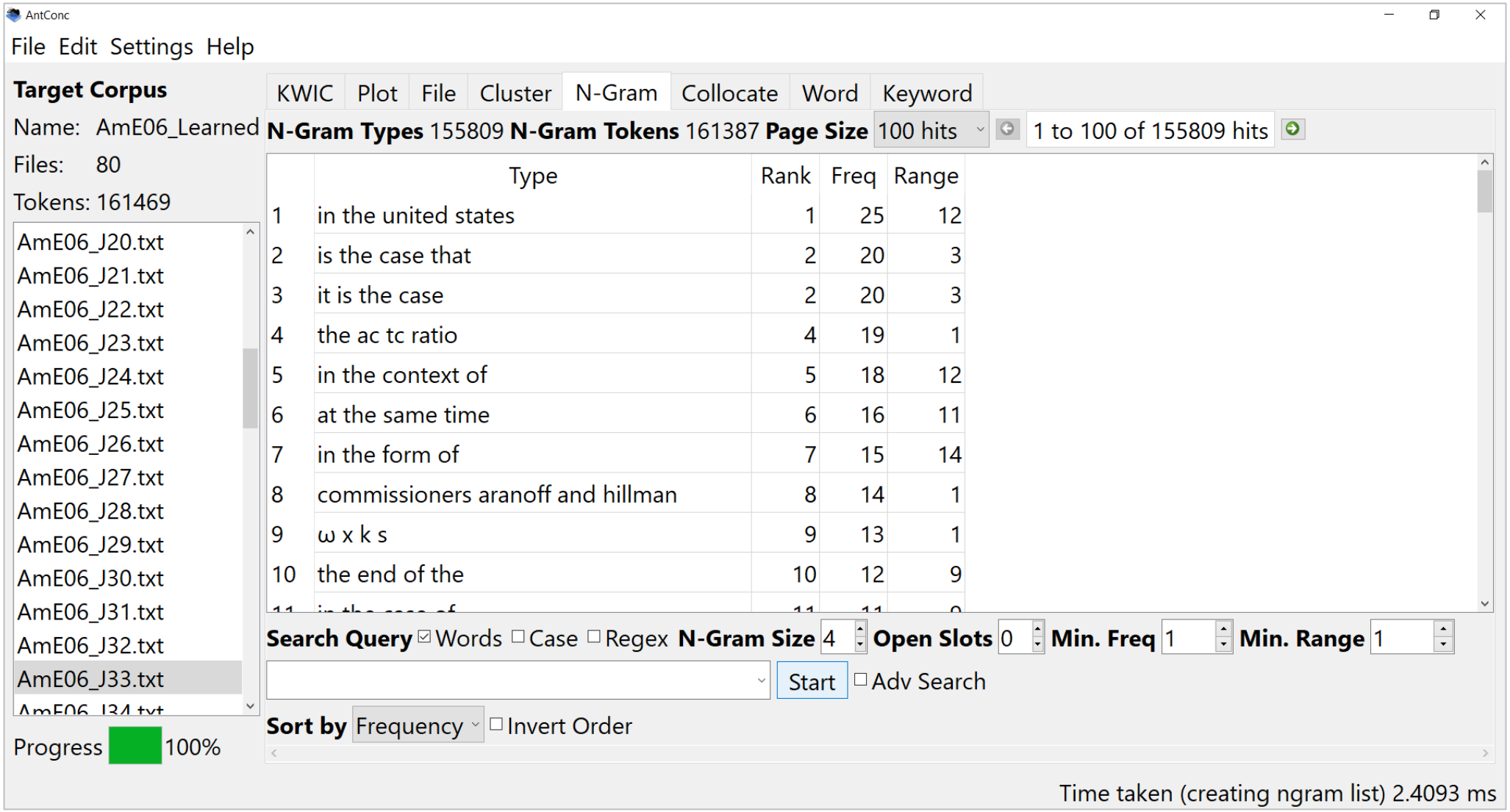
Introduction to AntConc 4:

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



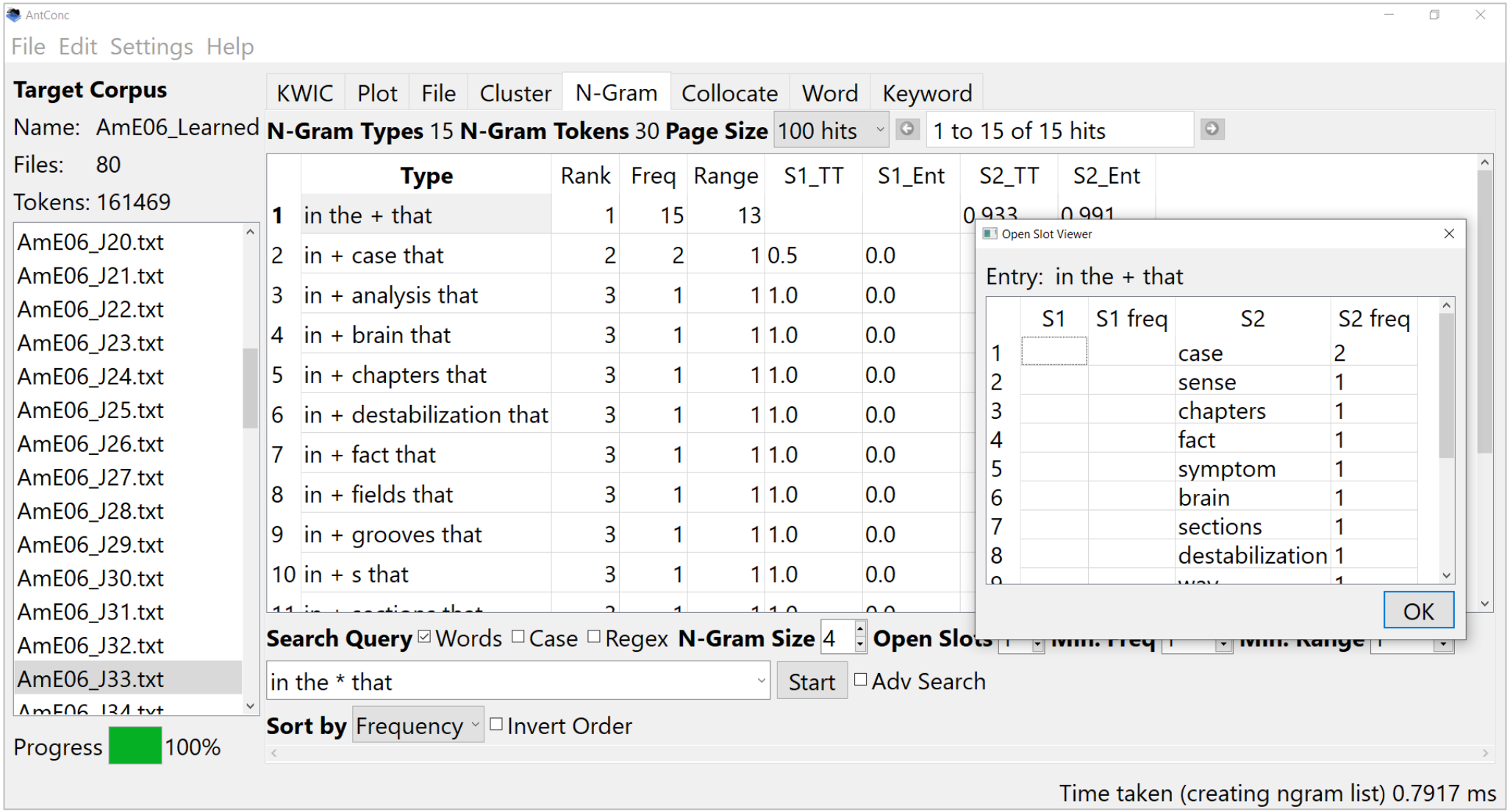
Introduction to AntConc 4:

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



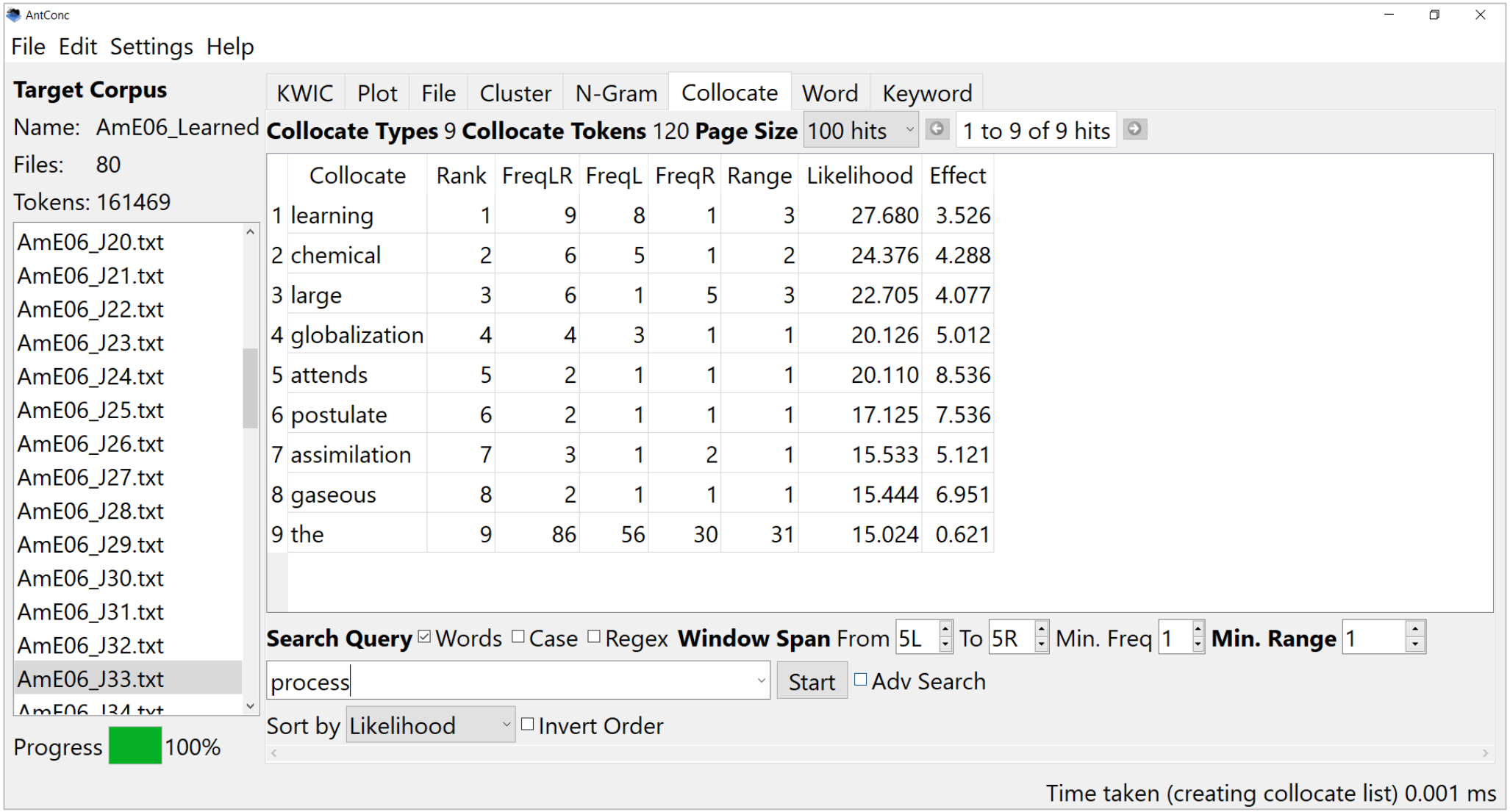
Introduction to AntConc 4:

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



Introduction to AntConc 4:

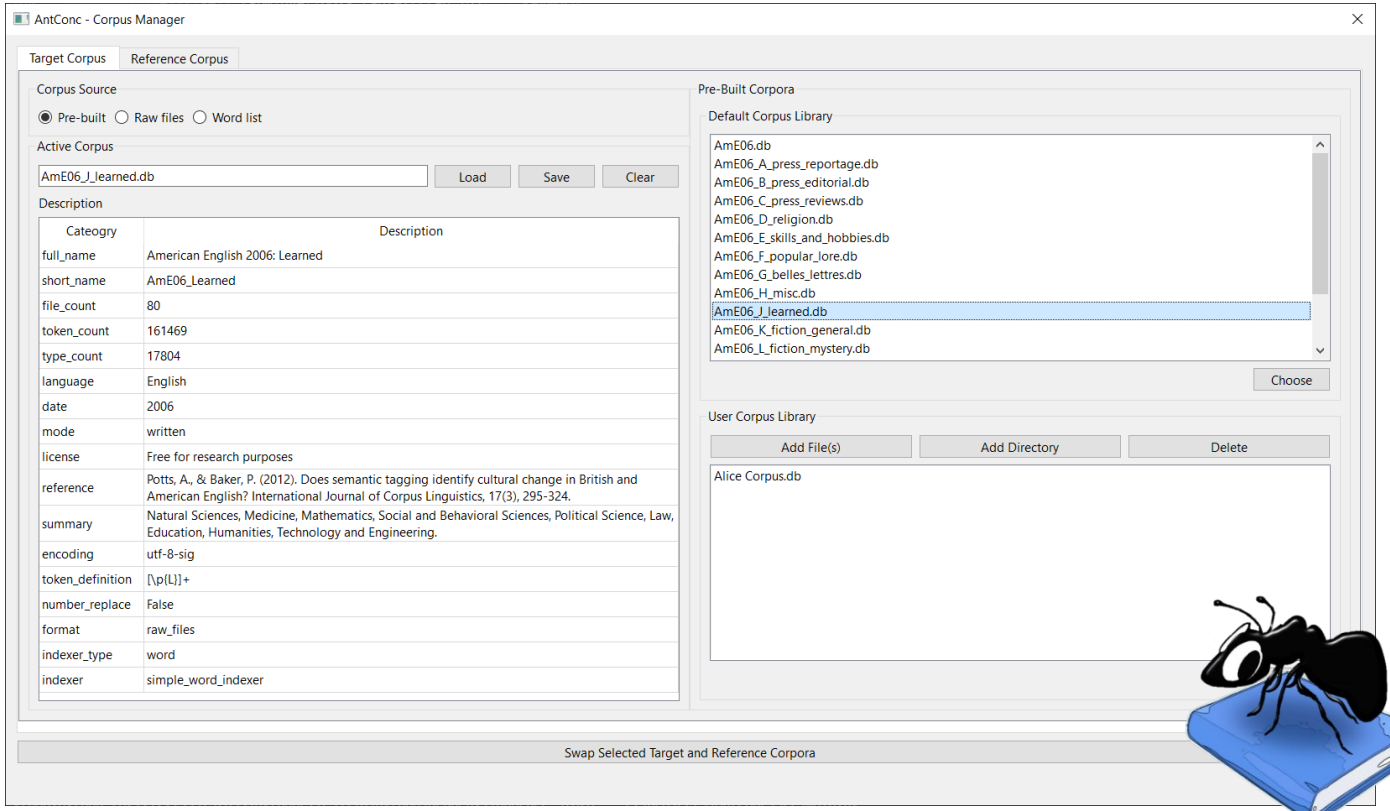
Features and functions - KWIC, Plot, File, Cluster, N-Gram, Collocate, Word, Keyword



Introduction to AntConc 4:

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

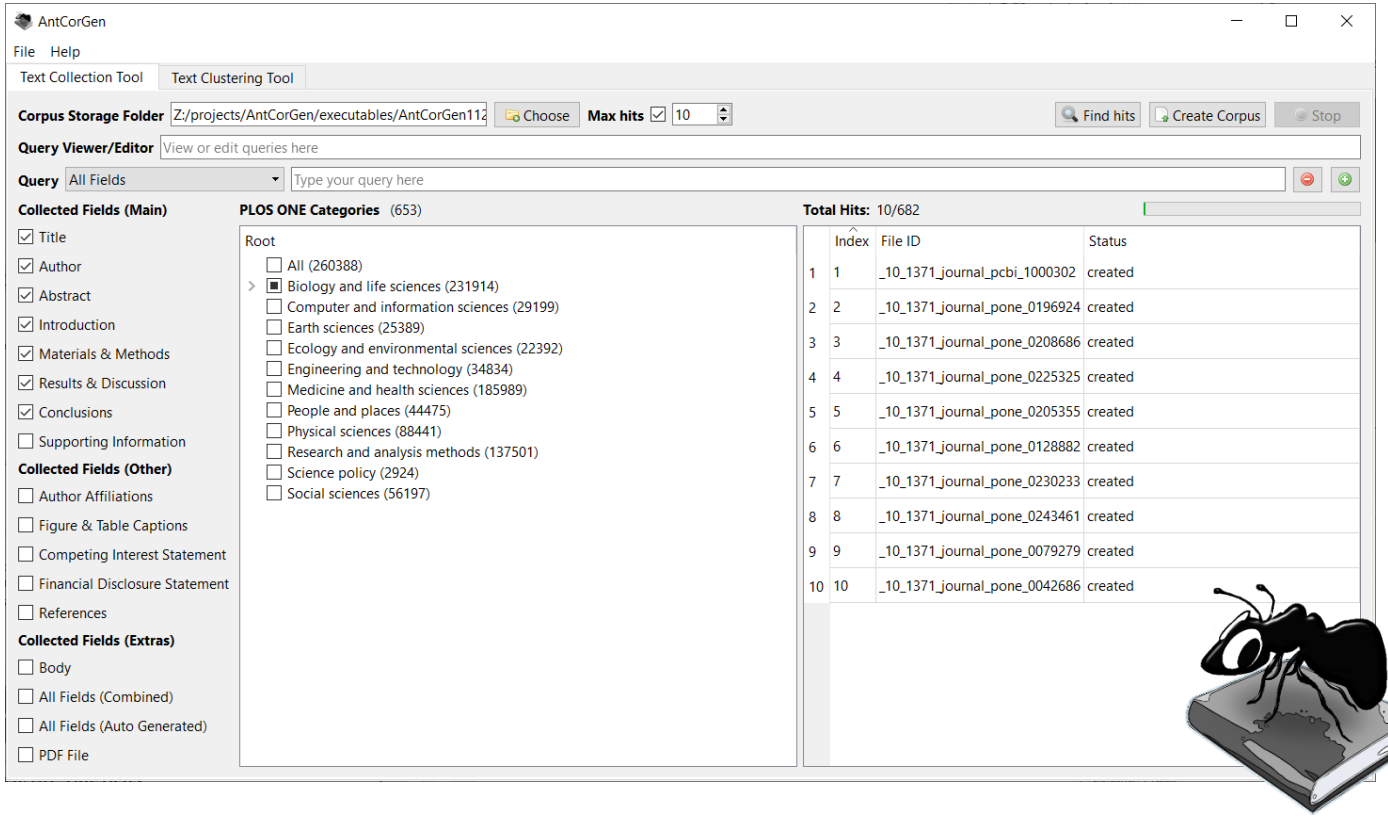
- AntConc "Corpus Manager" - default/user corpus library
  - a repository of pre-built corpora
    - AmE06
      - AmE06 press
      - AmE06 learned (academic)
      - …
    - BE06
      - BE06 press
      - BE06 learned (academic)
      - …
    - BASE/BAWE (coming soon)
    - …
    - your contributions??



Introduction to AntConc 4:

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

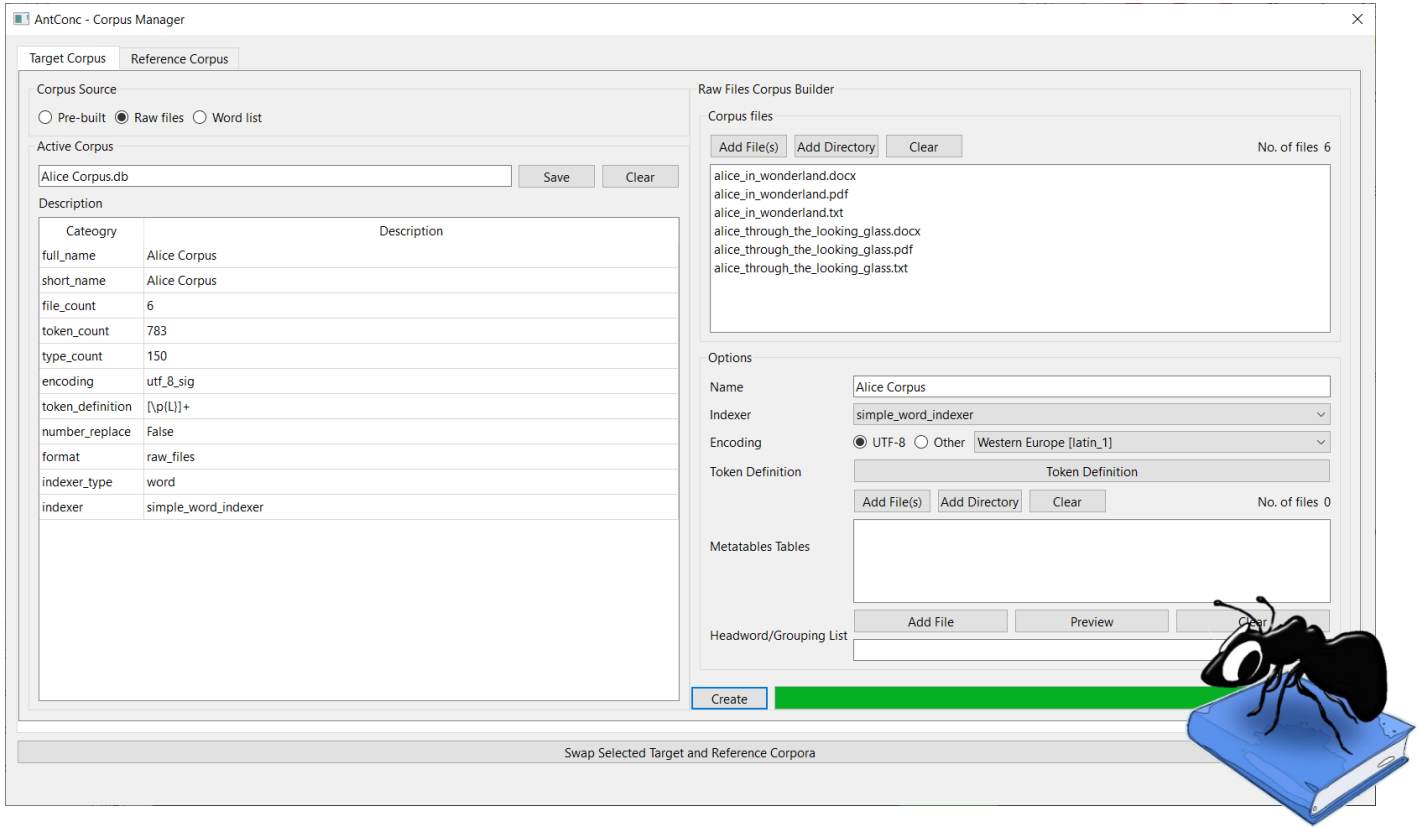
- **AntCorGen** - A PLOS ONE multi-disciplinary research article collection tool
  - Anthony, L. (2022). AntCorGen (Version 1.2.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software/antcorgen



Introduction to AntConc 4:

Addressing the challenges of DDL - 2) creating custom corpora

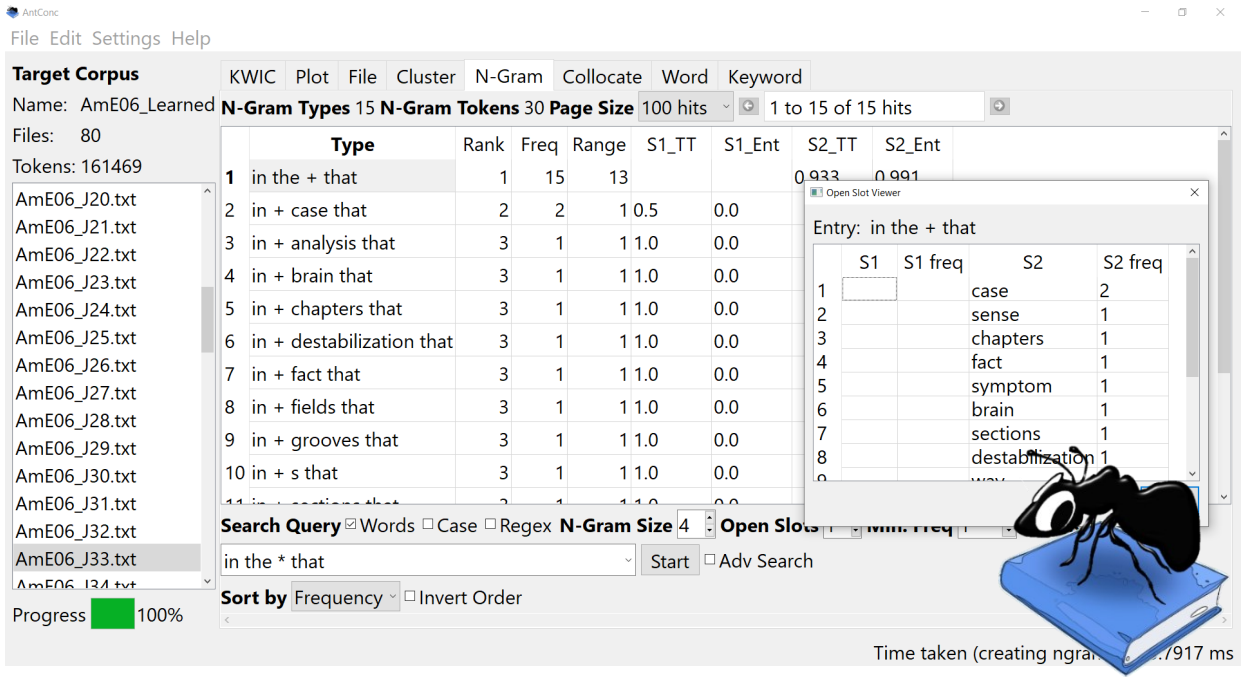
- AntConc "Corpus Manager" - corpus builder
  - Word (.docx), PDF (.pdf), TEXT (.txt) corpus creation via drag n' drop
    - student papers
    - research papers
    - e-books
    - copy/paste articles
    - …



Introduction to AntConc 4:

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

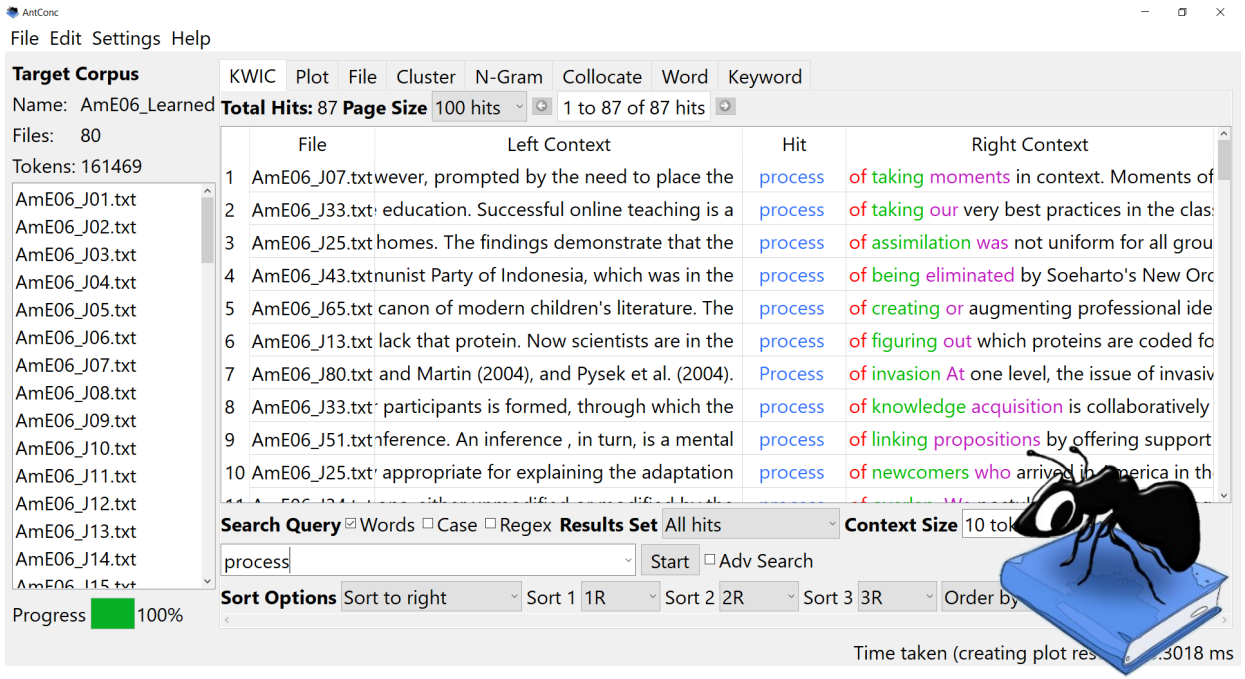
- AntConc tools for language discovery - Word, Keyword, (open slot) N-gram
  - discover 'important' (frequent) words using the Word (list) tool
  - discover 'important' (unusually frequent) words using the Keyword (list) tool
  - discover frequent word patterns (lexical bundles & positional frames) using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



Introduction to AntConc 4:

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

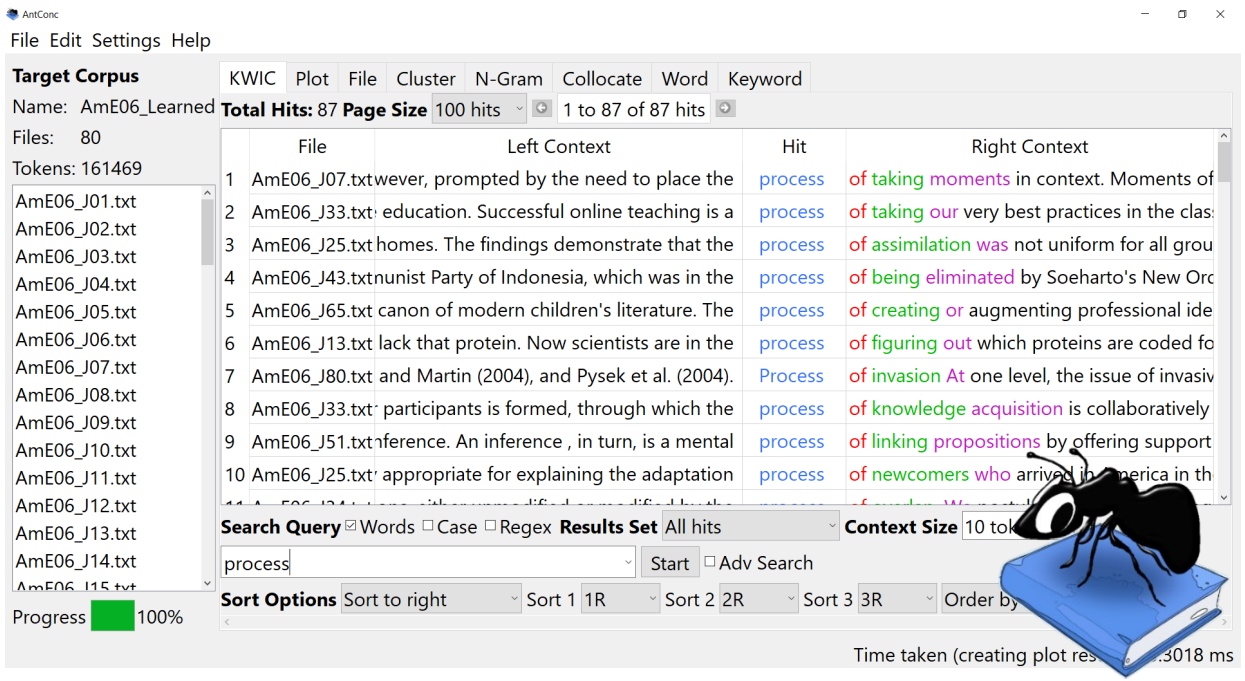
- AntConc "KWIC Patterns"
  - KWIC concordance line ordering by pattern frequency (not alphabetical)
- Results "Paging" and "Filtering"
  - all tools show 'paged' (10, 50, 100, …, all hits) results (cf. Web browser searching)
  - KWIC results can be 'trimmed' (e.g., 10, 20, … random hits)
- Best practices" 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)



Introduction to AntConc 4:

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

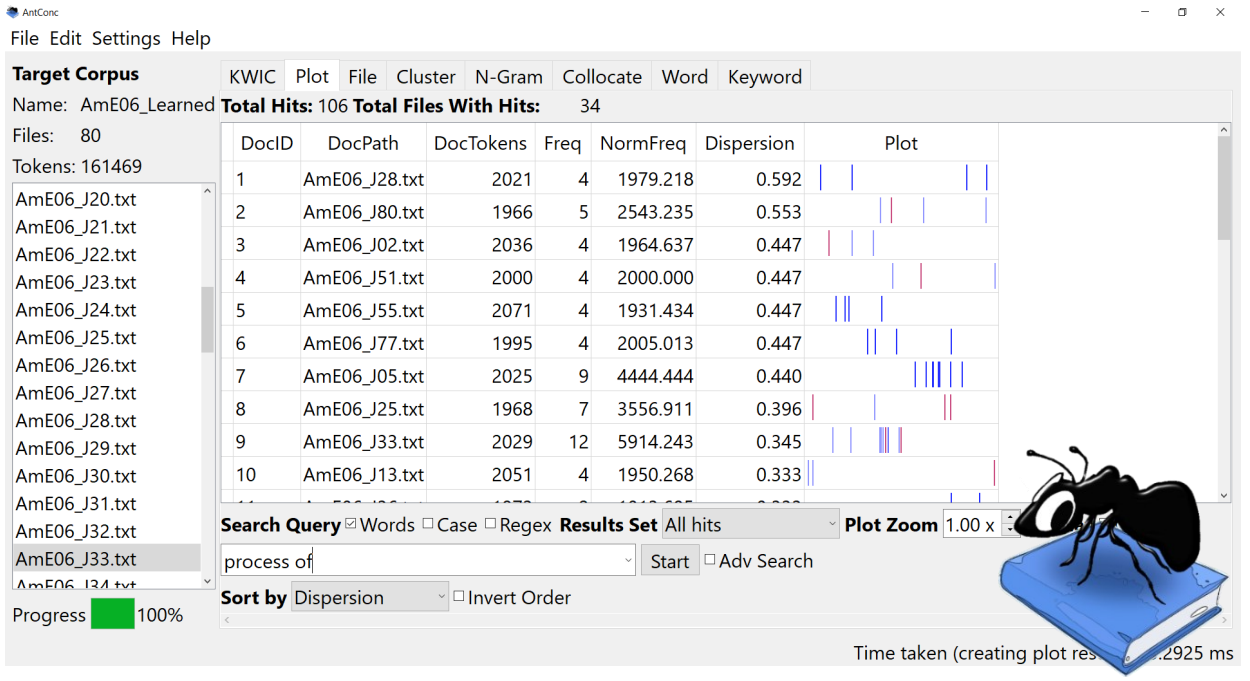
- AntConc "KWIC Patterns" bring a revolutionary change to DDL (and corpus linguistics research in general)
  - ordering KWIC results by pattern frequency…
    - immediately shows salient patterns (without the need for scrolling)
    - (can) reveal extremely surprising results (e.g., sorting to the left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



Introduction to AntConc 4:

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

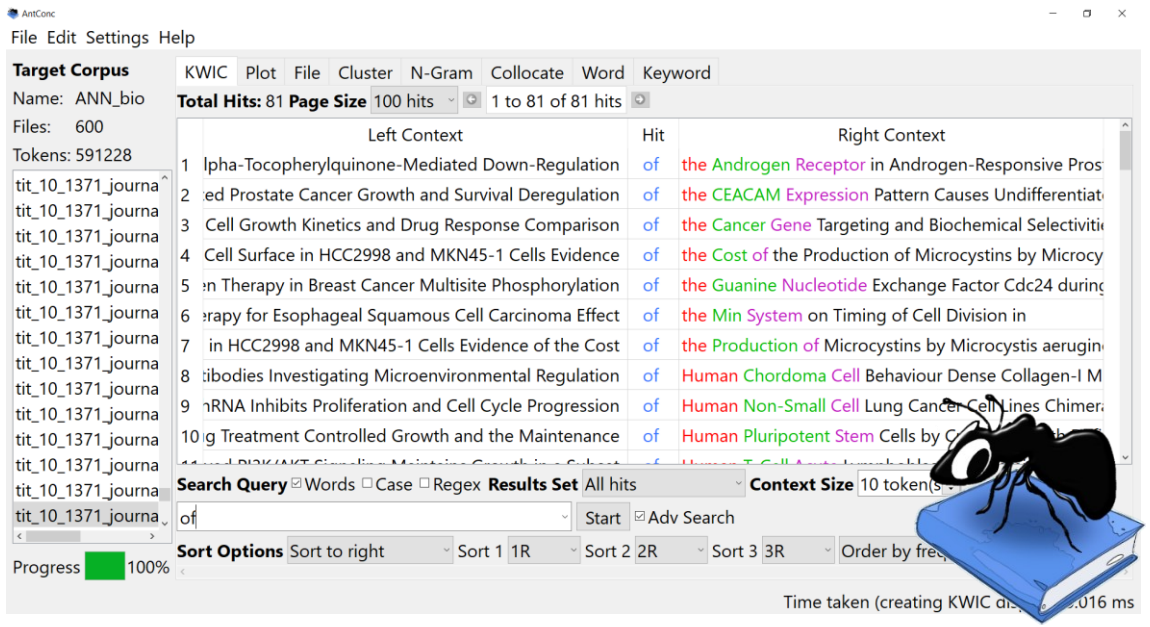
- AntConc "Plot" and range values…
  - help identify texts of interest
    - e.g., frequent use of target words
    - e.g., texts with frequent uses of words/patterns (i.e., exemplar texts - cf. *ProtAnt*)
  - highlight individual variation
  - reveal outlier texts



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
    - …



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