Developing Corpora for Research:
From ready-built to custom-built corpora

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

Harnessing the latest corpus-based approaches for research,
The Open University of Hong Kong, Hong Kong, China, June 29-30, 2017

Overview

- **Understanding corpus linguistics research**
  - useful definitions
  - two paths to corpus research
- **Utilizing ready-built corpora**
  - availability of ready-built corpora
  - successful projects utilizing ready-built corpora
- **Designing and building custom corpora**
  - designing custom corpora
  - tools for collecting, cleaning, and processing custom corpora
- **Analyzing corpora**
  - an introduction








Understanding corpus linguistics research:
useful definitions


Understanding corpus linguistics research:
What is corpus linguistics?

- **It is an empirical (experimental) approach**
  - An analysis of actual patterns of use in target texts
- **It uses a corpus of natural texts as the basis for analysis**
  - Corpus = a representative sample of target language stored as an electronic database (plural = "corpora")
- **It relies on computer software for analysis**
  - Results are generated using automatic and interactive techniques
- **It depends on both quantitative and qualitative analytical techniques**
  - Observations are counted and results are interpreted

Biber, Conrad, and Reppen (1998)

Understanding corpus linguistics research:
What is corpus linguistics?

"A corpus is a collection of machine readable, authentic texts, which is sampled to be representative of a particular language or language variety." (McEnery et al., 2006: 5)

Understanding corpus linguistics research:
What are the main limitations?

- **If a word or phrase does not appear in a corpus, we cannot obtain any information about it**
  - Corpus studies are based on what we observe
- **The larger the corpus, the more reliable it will be about revealing information on language features**
  - Bigger corpora are (usually) better
- **But... however large a corpus is, it can never represent all the variation in a language (except in special cases)**
  - A corpus provides only an approximation to reality
  - It can suggest trends but not "facts" about language use
  - We need to use statistics to determine significant patterns

Understanding corpus linguistics research:
What are the main limitations?

- **Take the sentence…**
  - The cat sat on the mat.
- **How many "words" does it contain?**
  - Tokens = 6 (the, cat, sat, on, the, mat)
  - Types = 5 (the, cat, sat, on, mat)
- **Which word types are 'special'?**

	the	on	sat	cat	mat
target sentence	2	1	1	1	1
AmE06	60056	6932	148	49	10

Understanding corpus linguistics research:
What are the main limitations?

- **Take the sentence…**
  - The cat sat on the mat.
- **How many "words" does it contain?**
  - Tokens = 6 (the, cat, sat, on, the, mat)
  - Types = 5 (the, cat, sat, on, mat)
- **Which word types are 'special'?**

	the	on	sat	cat	mat
target sentence	2	1	1	1	1
AmE06	60056	6932	148	49	10
LL keyness	4.17	4.64	12.26*	14.46*	17.56*

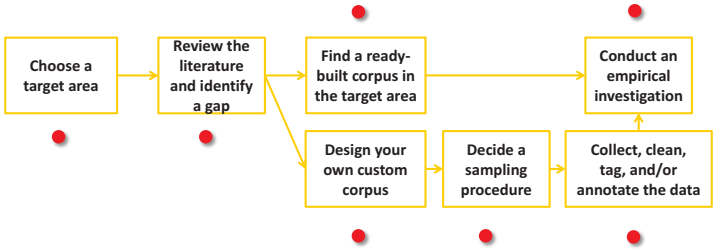
*p < 0.05 + Bonferroni correction

$$LL = 2\sum_i O_i \ln\left(\frac{O_i}{E_i}\right) \quad \text{where} \quad E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

Understanding corpus linguistics research:
two paths to corpus research

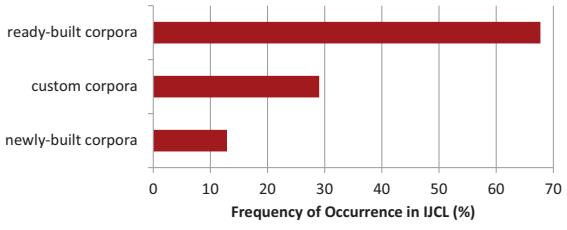
Understanding corpus linguistics research:
Two paths to corpus research

The "Corpus Approach"



Understanding corpus linguistics research:
Two paths to corpus research

- **Published research following the two paths**
  - **Journal**: International Journal of Corpus Linguistics (IJCL)
  - **Years**: 2013-2016 (3 years)
  - **Number of articles**: 64



Utilizing ready-built corpora:
availability of ready-built corpora

Understanding corpus linguistics research:
Selection of ready-built (general) corpora

- **Australian Corpus of English (ACE)**
- **British Academic Written English (BAWE) Corpus**
- **British National Corpus (BNC)**
- **Brown Corpus (+FROWN)**
- **Corpus of Contemporary American English (COCA)**
- **Lancaster Olsen Bergen (LOB) Corpus (+ FLOB, +BE06, AmE06)**
- **Open American National Corpus (OANC)**
- **Wellington Corpus of Written NZ English (WWC)**
- **…**

Understanding corpus linguistics research:
Selection of ready-built (specialized) corpora

- **Business Letter Corpus**
  - http://www.someya-net.com/concordancer/
- **Enron Email Dataset**
  - http://www.cs.cmu.edu/~./enron/
- **Michigan Corpus of Academic Spoken English (MICASE)**
  - https://quod.lib.umich.edu/cgi/c/corpus/corpus?c=micase;page=simple
- **Michigan Corpus of Upper-Level Student Papers (MICUSP)**
  - http://micusp.elicorpora.info/
- **PERC Corpus ("Corpus of Professional English")**
  - http://scn.jkn21.com/~perc04/
- **PolyU Business Corpora**
  - http://langbank.engl.polyu.edu.hk/corpus/polyu_business.html
- **SRI American Express travel agent dialogue corpus**
  - http://www.ai.sri.com/~communic/amex/amex.html
- **The Twitter Political Corpus**
  - http://www.usna.edu/Users/cs/nchamber/data/twitter/

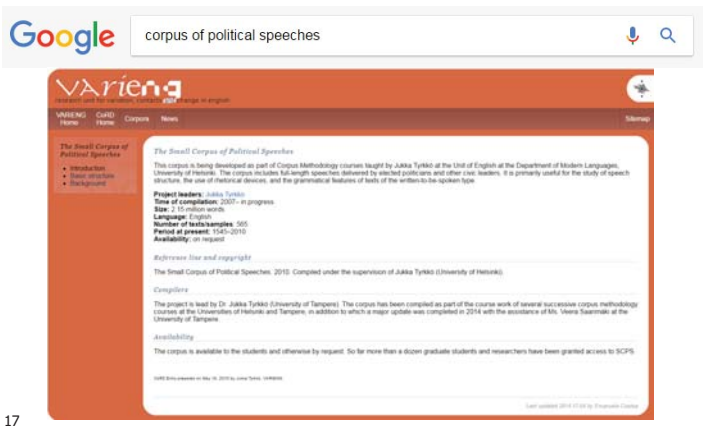
Understanding corpus linguistics research:
Selection of lists of ready-built corpora

- **"Corpus-based linguistics links"**
  - http://martinweisser.org/corpora_site/CBLLinks.html
- **"List of corpora"**
  - http://www.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/corpora/list/index2.html
- **"Texts & corpora"**
  - http://linguistlist.org/sp/GetWRListings.cfm?WRAbbrev=Texts
- **Wikipedia**
  - https://en.wikipedia.org/wiki/List_of_text_corpora
- **Web searches**
  - http://www.google.com/

Google Is Your Friend, GIYF is a term sometimes used in chat and forums to let the person asking the question know the answer could have been found by using the Google.
https://www.computerhope.com/jargon/g/giyf.htm

Understanding corpus linguistics research:
Selection of lists of ready-built corpora

Understanding corpus linguistics research:
Selection of lists of ready-built corpora



Utilizing ready-built corpora:
successful projects utilizing ready-built corpora

Utilizing ready-built corpora:
IJCL Research articles utilizing ready-built corpora

BNC, A Widow for One Year	Čermáková, A. (2015). Repetition in John Irving's novel A Widow for One Year
BAWE	Park, K., & Lu, X. (2015). Automatic analysis of thematic structure in written English.
Bergen Corpus of London Teenage English	Larrivee, P., & Duffley, P. (2014). The emergence of implicit meaning: scalar implicatures with some.
BNC, COCA	Dichtel, F. (2016). A quantifier used on many occasions.
CEPhiT and CELiST	Monaco, L. M. (2016). Was late Modern English scientific writing impersonal?
Corpus of Spoken Dutch	Rysii, J., & De Cuypereii, L. (2014). Variable satellite placement in spoken Dutch.
Fisher corpus	Koops, C., & Lohmann, A. (2015). A quantitative approach to the grammaticalization of discourse markers
ICCI	Lenko-Szymanska, A. (2014). The acquisition of formulaic language by EFL learners
New York Times Annotated Corpus	De Smet, H. (2016). The root of ruthless.
Society for the Reformation of Manners Corpus	Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context

Utilizing ready-built corpora:
IJCL Research articles utilizing ready-built corpora

- **Interesting trends**
  - studying language at the word, grammar, and discourse level
  - studying langauge in quantitative and qualitative ways
  - comparing results from custom-built target corpora with ready-built reference corpora
  - building custom corpora which are utilized in future research as ready-built corpora
  - sampling texts from ready-built corpus to build a new, custom-built corpus
    - e.g. Kreyer, R. (2015). "Funky fresh dressed to impress": A corpus-linguistic view on gender roles in pop songs. *International Journal of Corpus Linguistics*, *20*(2), 174-204.
      - From the Giessen Bonn corpus of Popular music...
        - lyrics by females ($Corpus_f$) + lyrics by males ($Corpus_m$), no albums

Designing and building custom corpora:
designing custom corpora

Designing custom corpora:
Step 1: Understand the different types of source texts

- **Plain text (.txt)**
  - can be used in corpus software without changes
- **XML (.xml) / HTML (.html/.htm)**
  - specially formatted plain text file (use as is or with tags deleted)
- **Microsoft DOC/DOCX (.doc / .docx)**
  - DOC: special binary file (choose "save as text")
  - DOCX: specially formatted .xml file that is zipped (choose "save as text")
- **Adobe PDF (.pdf)**
  - Text-based PDF:
    - requires a conversion tool to work with corpus software ("save as text")
    - usually introduces noise into the plain text file (e.g. headers/footers)
  - Graphic-based PDF:
    - requires OCR software to recreate text inside of the graphics (try "save as text"

Designing custom corpora:
Step 2: Understand (.txt) character encodings

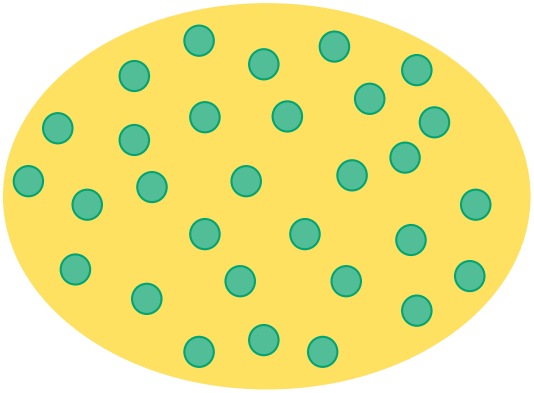
Character Encoding	Included characters	Windows (File Name)	Windows (File Content)	Linux/Mac (File Name)	Linux/Mac (File Content)
ASCII (the first)	A-Z + α				
ANSI (ASCII + local characters)	A-Z + α + Chinese (cp950 - Big5)	✓	✓ (Save option "ANSI")		
UTF-16LE (Windows internal)	A-Z + α + all languages	✓	✓ (Save option "Unicode")		
UTF-8 (International standard + Linux/Mac internal)	A-Z + α + all languages		✓ (Save option "UTF-8")	✓	✓

Designing custom corpora:
Step 3: Decide a suitable sampling frame

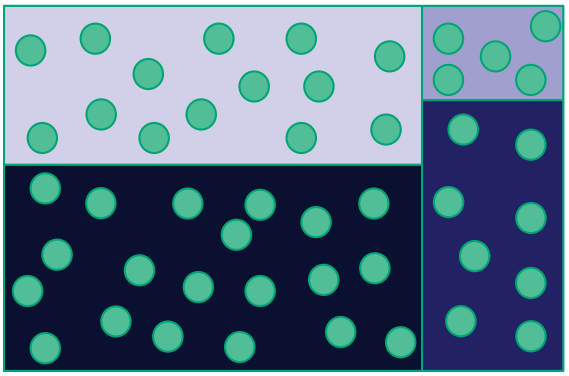
"... a well conducted poll of 1,000 people can, most of the time, give us an idea of what the country as a whole is thinking"

BBC News Politics, "How poll tracker works", 2015
http://www.bbc.co.uk/news/uk-politics-13248622

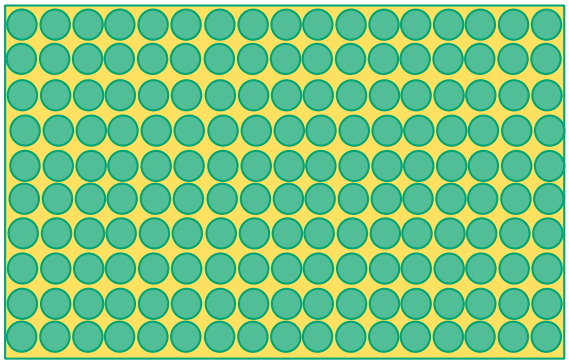
Designing custom corpora:
Step 3: Decide a suitable sampling frame (random sampling)



Designing custom corpora:
Step 3: Decide a suitable sampling frame (stratified sampling)



Designing custom corpora:
Step 3: Decide a suitable sampling frame (whole population)



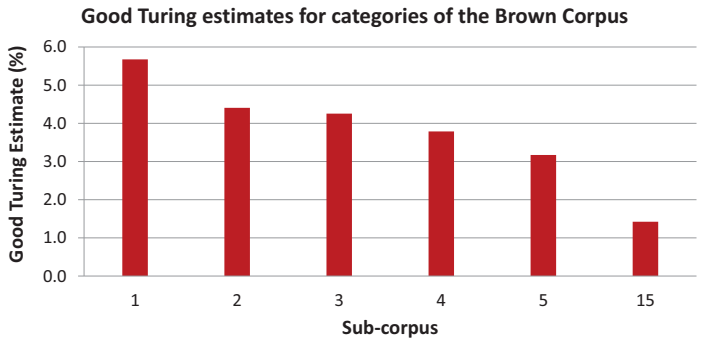
Designing custom corpora:
Step 4: Estimate a good corpus size (improving representativeness)

- A Good-Turing Estimate for Finding New Types:
  - I. J. Good (1953), Biometrika 40(3-4), pp. 237-264.

$$\text{Probability of new type} = \frac{\text{Number of Types with Freq.of 1}}{\text{Total Number of Tokens}}$$

  - Example:
    - In the Brown Corpus (Section A), there are 97389 tokens and 5528 types that occur only once.
      - Good-Turing estimate = 5528/97389 = **5.7%**
    - In the Brown Corpus (Section A+B), there are 157517 tokens and 6940 types that occur only once.
      - Good-Turing estimate = 6940/157517 = **4.4%**
    - the Brown Corpus (Section A+B+C), there are 196612 tokens and 8365 types that occur only once.
      - Good-Turing estimate = 8365/196612 = **4.3%**

Designing custom corpora:
Step 4: Estimate a good corpus size (improving representativeness)



Designing custom corpora:
Step 5: Other suggestions

- **To improve your corpus design, ...**
  - understand sampling theory and apply it to corpus building
  - create better operational definitions of the target population
  - focus on narrower target populations
    - e.g. academic English written by students in Asia
    - e.g. textbook English in UK university science courses
- **To improve the impact of your corpus research, ...**
  - focus less on descriptive research
  - focus more on predictive research

Designing custom corpora:
Step 5: Other suggestions

"you [can] always make a model to explain your data. That's not the hard thing. Now give me some predictions..."

Eric Lander, 7.012 Introduction to Biology, MIT OCW, 2004
http://ocw.mit.edu/courses/biology/7-012-introduction-to-biology-fall-2004/

Designing and building custom corpora:
tools for collecting, cleaning, and processing custom corpora

Building and utilizing custom corpora:
Tools for collecting corpus data

- **BootCat** (Freeware tool to build corpora from the web)
  - http://bootcat.sslmit.unibo.it/
- **CorpusCreator** (Freeware tool to build corpora from the web)
  - http://www.staff.uni-mainz.de/fantinuo/info_corpuscreator.html
- **WebBootCat** (Commercial interface to BootCat)
  - https://www.sketchengine.co.uk/documentation/wiki/SkE/Help/WebBootCat
- **DownThemAll** (Firefox file download manager)
  - https://addons.mozilla.org/en-US/firefox/addon/downthemall/
- **Chrono Download Manager** (Chrome file download manager)
  - http://www.chronodownloader.net/
- **NotePad++** (Win text editor) or **TextWrangler** (Mac text editor)

Building and utilizing custom corpora:
Tools for cleaning and processing corpus data

Cleaning Tools

- AntFileConverter
- EncodeAnt
- SegmentAnt
- SarAnt

www.laurenceanthony.net/software

Building and utilizing custom corpora:
Tools for cleaning and processing corpus data

Tagging Tools

- TagAnt
- CLAWSAnt
- AntMover

www.laurenceanthony.net/software

Building and utilizing custom corpora:
Tools for cleaning and processing corpus data

Analysis Tools

- AntConc
- AntPConc
- AntWordProfiler
- ProtAnt
- VariAnt

www.laurenceanthony.net/software

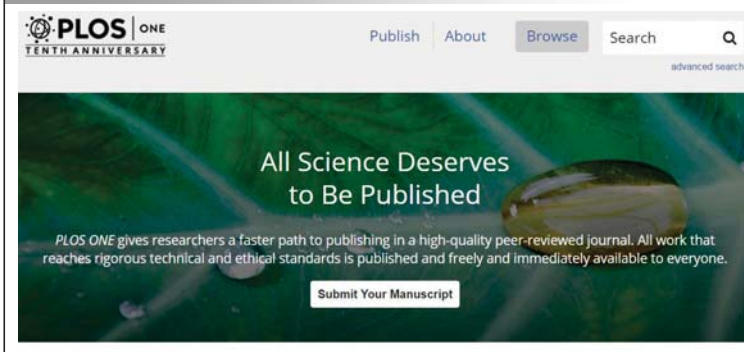
Introduction to *AntCorGen*

an automated corpus generation tool

http://www.laurenceanthony.net/software/antcorgen/

AntCorGen

Version 1.0.0

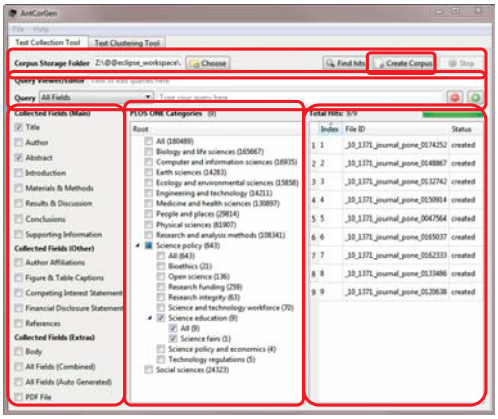


Join the worldwide community of researchers - from Nobel laureates to early career researchers - who choose *PLOS ONE*.

38

AntCorGen

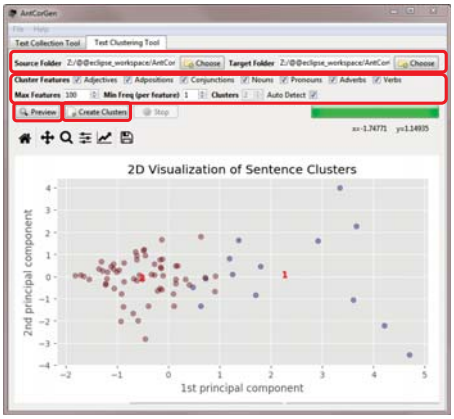
Version 1.0.0



39

AntCorGen

Version 1.0.0



40

AntCorGen

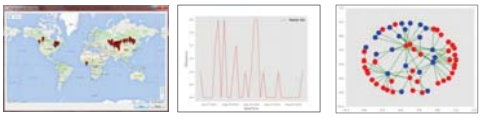
Version 1.0.0

- **Freeware**
  - Download from: http://www.laurenceanthony.net/software/antcorgen/
- **Portable**
  - Requires no installation. Runs directly from a USB stick
- **Multiplatform**
  - Windows, Mac OS X, Linux
- **Development environment**
  - Python 3.5.3

41

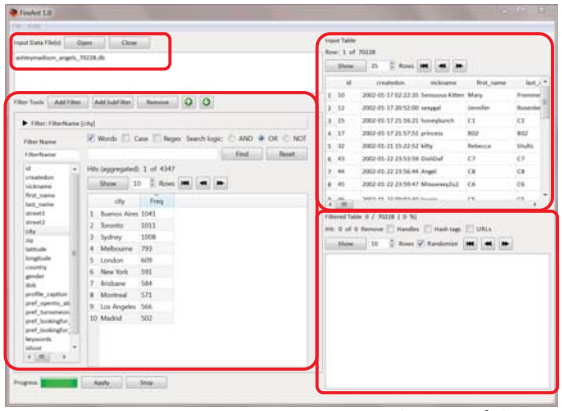
Introduction to *FireAnt*

a social media (Twitter) data collection tool and general corpus visualization tool



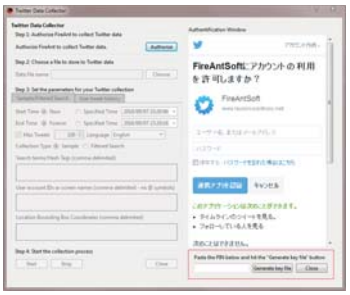
http://www.laurenceanthony.net/software/fireant/

FireAnt:
Filter, Identify, Report, and Export Analysis Tool

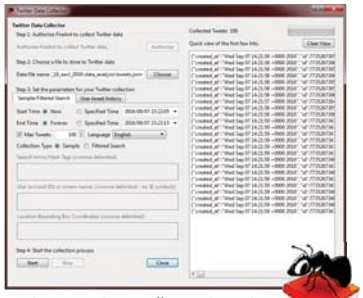


Main screen of *FireAnt*

FireAnt:
Filter, Identify, Report, and Export Analysis Tool

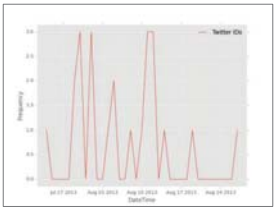


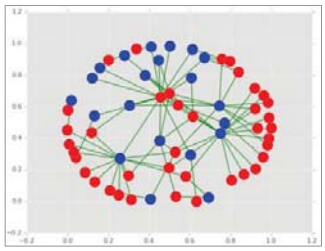
FireAnt Twitter authentication window



FireAnt Twitter collector in action

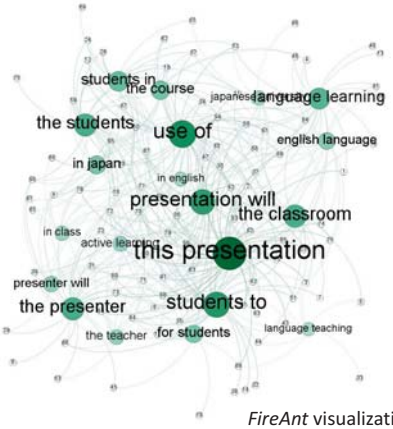
FireAnt:
Filter, Identify, Report, and Export Analysis Tool





FireAnt visualization tools

FireAnt:
Filter, Identify, Report, and Export Analysis Tool

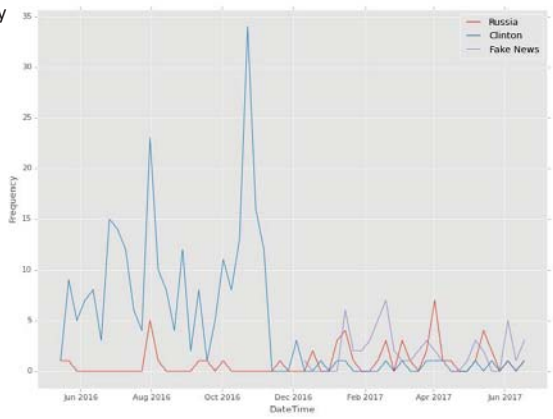


FireAnt visualization tools

FireAnt:
Filter, Identify, Report, and Export Analysis Tool

Trump Tweet history
(2017/06/18)

Total tweets:
3016 (no retweets)

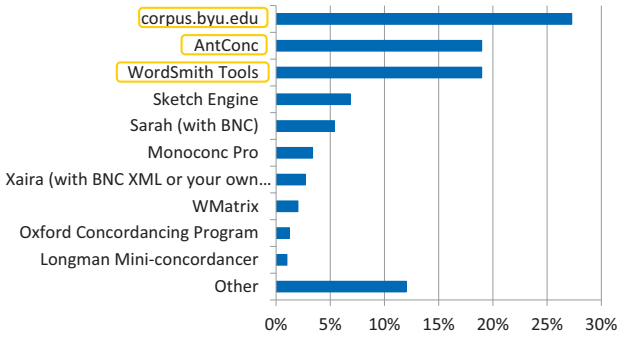


FireAnt:
Filter, Identify, Report, and Export Analysis Tool

- **Freeware**
  - Download from: http://www.laurenceanthony.net/software/fireant/
- **Portable**
  - Requires no installation. Runs directly from a USB stick
- **Multiplatform**
  - Windows, Mac OS X, Linux
- **Development environment**
  - Python 2.7

Analyzing corpora:
an introduction

Analyzing corpora:
Choosing a corpus analysis tool



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

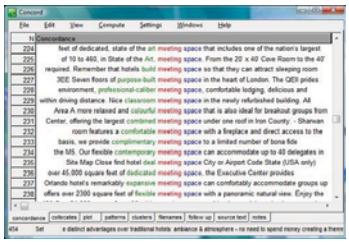
Analyzing corpora:
Choosing an online corpus analysis tool

Corpus of Contemporary American English (COCA)
http://corpus.byu.edu/coca/

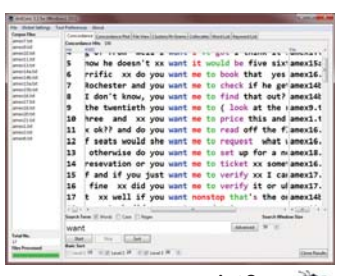
Analyzing corpora:
Choosing an online corpus analysis tool

SketchEngine
https://www.sketchengine.co.uk/

Analyzing corpora:
Choosing an offline corpus analysis tool

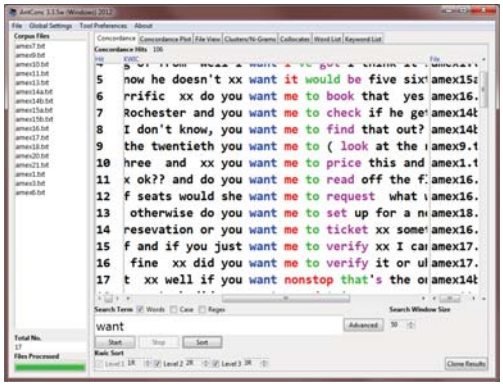


WordSmith Tools
Scott, M. (2015)



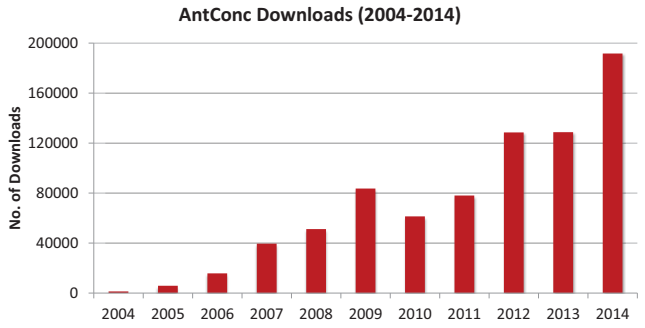
AntConc
Anthony, L. (2017)

Analyzing corpora:
Choosing an offline corpus analysis tool



AntConc
Anthony, L. (2017)

Getting started in corpus linguistics:
Stage 2: Choose a software tool



Server log analysis. (Anthony, 2014)

Analyzing corpora:
Overview of AntConc (www.laurenceanthony.net/software/)

- **Freeware**
- **Multiplatform**
  - Windows, Mac OSX, Linux
- **Portable**
  - no installation
  - runs from a USB
- **Unicode compliant**
- **HTML/XML tag handing**
- **Search Features**
  - words, strings (case)
  - wildcards
  - regular expressions
- **Tools**
  - KWIC Concordancer
  - Distribution Plot
  - File View
  - Clusters/N-grams
  - Collocates
  - Word Frequency
  - Keyword Frequency

Summary and Questions:
Where is your next step?

Summary and Questions

- **Understanding corpus linguistics research**
  - What do you want to know?
  - What is the best corpus that will help you find the answer?
- **Utilizing ready-built corpora**
  - What resources are already available?
  - Can these resources help you find the answer to your question?
- **Designing and building custom corpora**
  - What steps do you need to follow to create your own corpus?
  - What tools can you use to automate some of these steps?
- **Analyzing corpora**
  - What ready-made tools can you use to help you find the answer to your question?
  - Do you need to design your own custom tools?