# Arguments for and against DIY corpus tools creation:

A debate about programming

**Laurence Anthony**

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

Waseda University, Tokyo, Japan

anthony@waseda.jp

http://www.laurenceanthony.net/

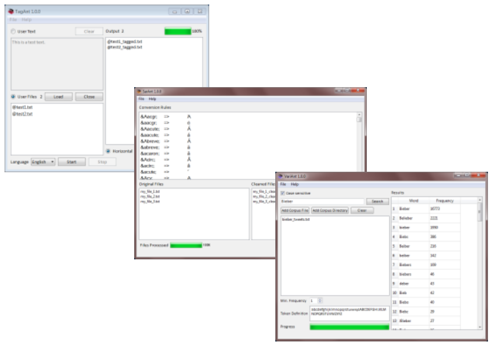
Corpus Statistics Group Launch Event, University of Birmingham, February 11, 2016

# Motivations

- Anthony, L. (2009). **Issues in the design and development of software tools for corpus studies: The case for collaboration**. In P. Baker (Ed.), Contemporary corpus linguistics (pp. 87-104). London, UK: Continuum Press.
- Anthony, L. (2013). **A critical look at software tools in corpus linguistics**. Linguistic Research, 2013, 30(2), 141-161.
- Anthony, L. (2014, September). **Corpus Tools Brainstorming Session**. Workshop given at the American Association for Corpus Linguistics (AACL 2014), September 25-28, 2014, Flagstaff, Arizona, US.
- Anthony, L., Wattam, S., Coole, M., Mariani, J., Rayson, P., and Vidler J. (2015, July). **Brainstorming the next generation of corpus software**. Workshop given at the Corpus Linguistics Conference (CL 2015), July 20-24, 2015. Lancaster University, UK.

## Overview

- **The current state of corpus linguistics tools**
  - four generations of corpus tools
  - a need for something new
- **DIY Corpus Tools – The Debate**
  - Arguments for learning to program
  - Arguments against learning to program
- **Thoughts on the future of programming and tools in corpus linguistics research**
  - Programming, tools, and statistics
  - Collaboration in project teams

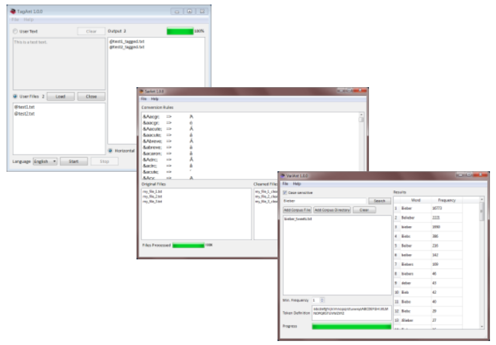


http://d.ibtimes.co.uk/en/full/339496/rock-john-cena.jpg?w=500

http://cdn.phys.org/newman/gfx/news/hires/2012/scientistsce.jp

# The current state of corpus linguistics tools

four generations of tools



# Current state of corpus linguistics tools:

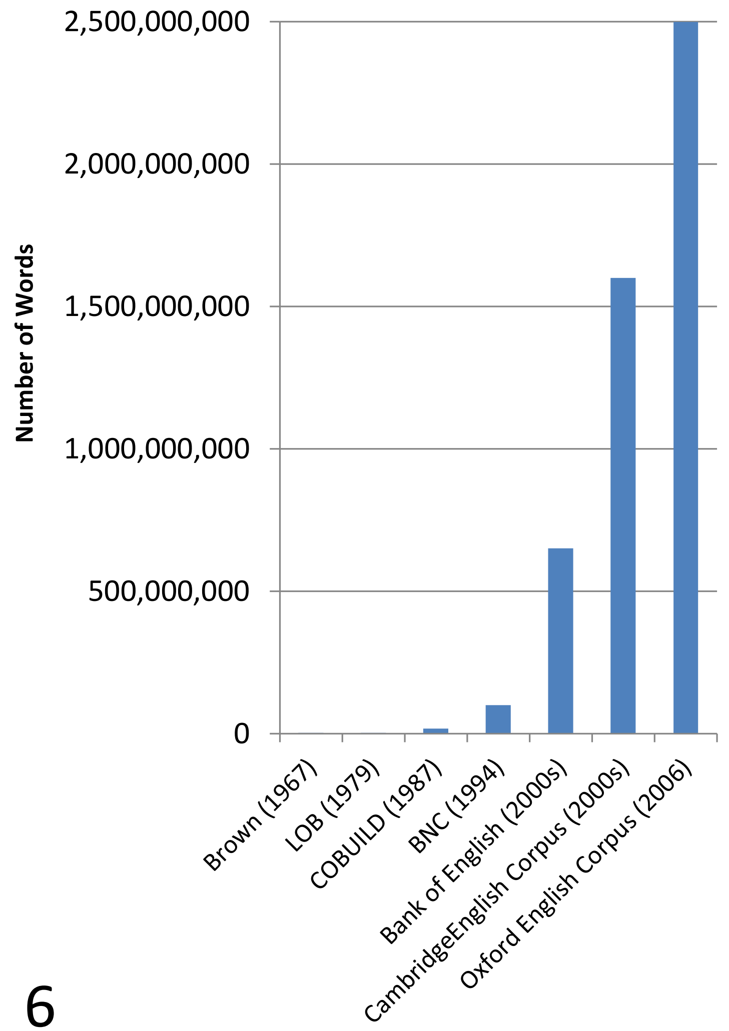
## A definition of corpus linguistics

- 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)

# Current state of corpus linguistics tools:

## From principled corpora to opportunistic corpora



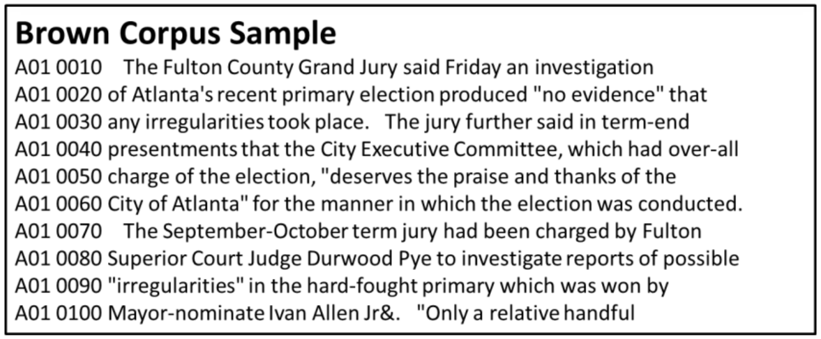
Google

CORPUS OF GLOBAL WEB-BASED ENGLISH (GloWbE)

CAMBRIDGE ENGLISH CORPUS · CAMBRIDGE QUALITY GUARANTEE

The Oxford English Corpus

BRITISH NATIONAL CORPUS



**Brown Corpus Sample**

```
A01 0010   The Fulton County Grand Jury said Friday an investigation
A01 0020 of Atlanta's recent primary election produced "no evidence" that
A01 0030 any irregularities took place.   The jury further said in term-end
A01 0040 presentments that the City Executive Committee, which had over-all
A01 0050 charge of the election, "deserves the praise and thanks of the
A01 0060 City of Atlanta" for the manner in which the election was conducted.
A01 0070   The September-October term jury had been charged by Fulton
A01 0080 Superior Court Judge Durwood Pye to investigate reports of possible
A01 0090 "irregularities" in the hard-fought primary which was won by
A01 0100 Mayor-nominate Ivan Allen Jr&.   "Only a relative handful
```

## Current state of corpus linguistics tools:
### Four generations of corpus tools (McEnery & Hardie, 2012)

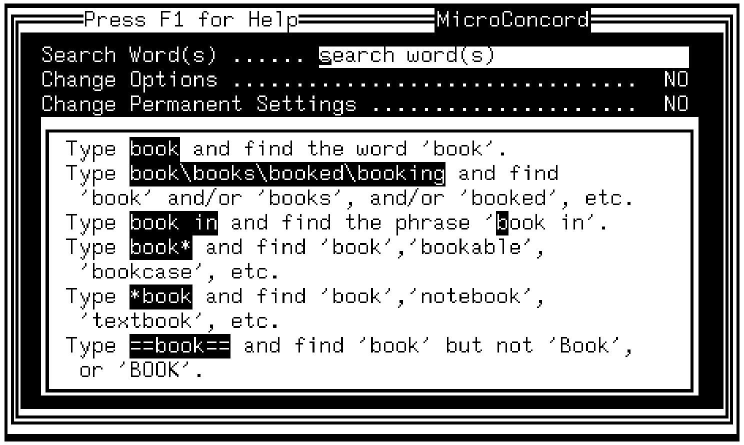
- 1st-generation (1960s-1970s)
  - run on mainframes, single function tools, ‘monolingual’ (ASCII-based), designed for tiny corpora (of the time)
    - e.g., A Concordance Generator (Smith, 1966)
    - e.g., Discon (Clark, 1966)
    - e.g., Drexel Concordance Program (Price, 1966)
    - e.g., Concordance (Dearing, 1966)
    - e.g., CLOC (Reed, 1978)

IBM 7090 Mainframe computer
http://thisdayintechhistory.com/11/30/ibm-7090-delivered/

# Current state of corpus linguistics tools:

## Four generations of corpus tools (McEnery & Hardie, 2012)

- 2nd-generation (1980s-1990s)
  - run on PCs, Roman-script language support, limited functions, designed for ‘small’ corpora
    - e.g., *Oxford Concordance Program (OCP)* (Hockey, 1988)
    - e.g., *Longman Mini-Concordancer* (Chandler, 1989)
    - e.g., *Kaye concordancer* (Kaye, 1990)
    - *e.g., MicroConcord* (Scott & Johns, 1993)

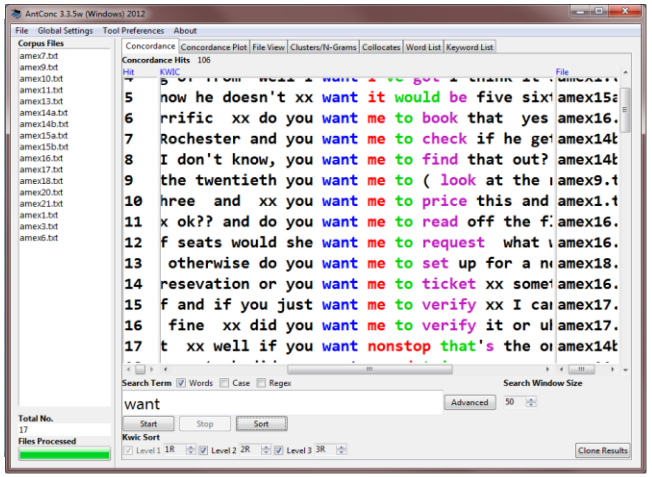


```
Press F1 for Help                    MicroConcord
 Search Word(s) ...... search word(s)
 Change Options .............................. NO
 Change Permanent Settings ................... NO

  Type book and find the word 'book'.
  Type book\books\booked\booking and find
   'book' and/or 'books', and/or 'booked', etc.
  Type book in and find the phrase 'book in'.
  Type book* and find 'book','bookable',
   'bookcase', etc.
  Type *book and find 'book','notebook',
   'textbook', etc.
  Type ==book== and find 'book' but not 'Book',
   or 'BOOK'.
```

MicroConcord (Scott & Johns, 1993)

## Current state of corpus linguistics tools:
Four generations of corpus tools (McEnery & Hardie, 2012)

- 3rd-generation (2000s-present)
  - run on PCs, partial (or full) Unicode support, more functions, designed for ‘bigger’ corpora, more statistical measures, easy-to-use
    - e.g., *WordSmith Tools* (Scott, 1996-2014)
    - e.g., *MonoConc Pro* (Barlow, 2000)
    - e.g., *AntConc* (Anthony, 2004-2014)



Concordance Hits 106

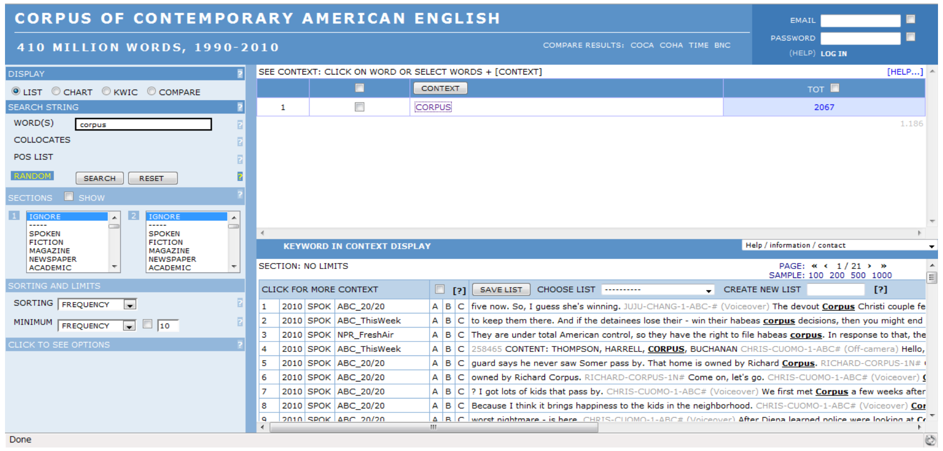
```
Hit KWIC                                                  File
5   now he doesn't xx want it would be five six amex15a
6   rrific  xx do you want me to book that  yes amex16.
7   Rochester and you want me to check if he ge amex14b
8   I don't know, you want me to find that out? amex14b
9   the twentieth you want me to ( look at the  amex9.t
10  hree  and  xx you want me to price this and amex1.t
11  x ok?? and do you want me to read off the f amex16.
12  f seats would she want me to request  what  amex16.
13   otherwise do you want me to set up for a n amex18.
14  resevation or you want me to ticket xx some amex16.
15  f and if you just want me to verify xx I ca amex17.
16   fine  xx did you want me to verify it or u amex17.
17  t  xx well if you want nonstop that's the o amex14b
```

Search Term: want
Search Window Size: 50
Kwic Sort: Level 1 1R, Level 2 2R, Level 3 3R
Total No. 17

WordSmith Tools (Scott, M., 2014)
AntConc (Anthony, L., 2014)

## Current state of corpus linguistics tools:
### Four generations of corpus tools (McEnery & Hardie, 2012)

- 4th-generation (late 2000s-present)
  - run on a server (accessed via a browser), partial (or full) Unicode support, simple (or advanced) functions, designed for pre-installed (copyrighted) corpora, simple to advanced statistical measures, easy-to-use
    - e.g., corpus.byu.edu (Davies, 2011), *CQPweb* (Hardie, 2011), *SketchEngine* (Kilgariff, 2011), *Wmatrix* (Rayson, 2011)



WORD(S): corpus

SORTING: FREQUENCY

MINIMUM: FREQUENCY 10

SEE CONTEXT: CLICK ON WORD OR SELECT WORDS + [CONTEXT]

| | | CONTEXT | TOT |
|---|---|---|---|
| 1 | | CORPUS | 2067 |

1.186

SECTION: NO LIMITS

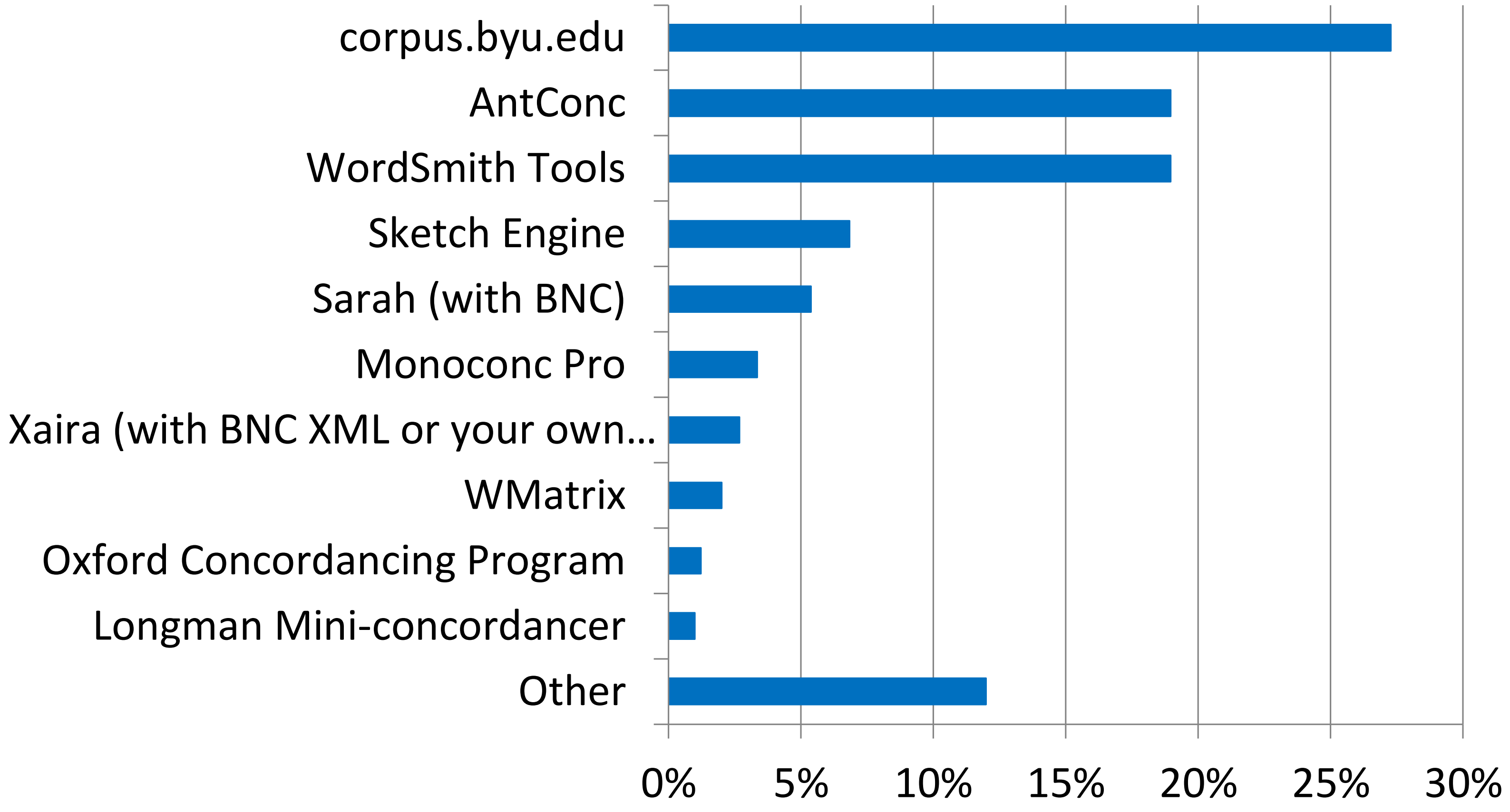
PAGE: « ‹ 1 / 21 › »
SAMPLE: 100 200 500 1000

| CLICK FOR MORE CONTEXT | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | 2010 | SPOK | ABC_20/20 | A | B | C | five now. So, I guess she's winning. JUJU-CHANG-1-ABC-# (Voiceover) The devout **Corpus** Christi couple fe |
| 2 | 2010 | SPOK | ABC_ThisWeek | A | B | C | to keep them there. And if the detainees lose their - win their habeas **corpus** decisions, then you might end |
| 3 | 2010 | SPOK | NPR_FreshAir | A | B | C | They are under total American control, so they have the right to file habeas **corpus**. In response to that, the |
| 4 | 2010 | SPOK | ABC_ThisWeek | A | B | C | 258465 CONTENT: THOMPSON, HARRELL, **CORPUS**, BUCHANAN CHRIS-CUOMO-1-ABC# (Off-camera) Hello, |
| 5 | 2010 | SPOK | ABC_20/20 | A | B | C | guard says he never saw Somer pass by. That home is owned by Richard **Corpus**. RICHARD-CORPUS-1N# |
| 6 | 2010 | SPOK | ABC_20/20 | A | B | C | owned by Richard Corpus. RICHARD-CORPUS-1N# Come on, let's go. CHRIS-CUOMO-1-ABC# (Voiceover) |
| 7 | 2010 | SPOK | ABC_20/20 | A | B | C | ? I got lots of kids that pass by. CHRIS-CUOMO-1-ABC# (Voiceover) We first met **Corpus** a few weeks after |
| 8 | 2010 | SPOK | ABC_20/20 | A | B | C | Because I think it brings happiness to the kids in the neighborhood. CHRIS-CUOMO-1-ABC# (Voiceover) |
| 9 | 2010 | SPOK | ABC_20/20 | A | B | C | worst nightmare - is here. CHRIS-CUOMO-1-ABC# (Voiceover) After Diena learned police were looking at |

COCA (Davies, M., 2016)

# Current state of corpus linguistics tools:

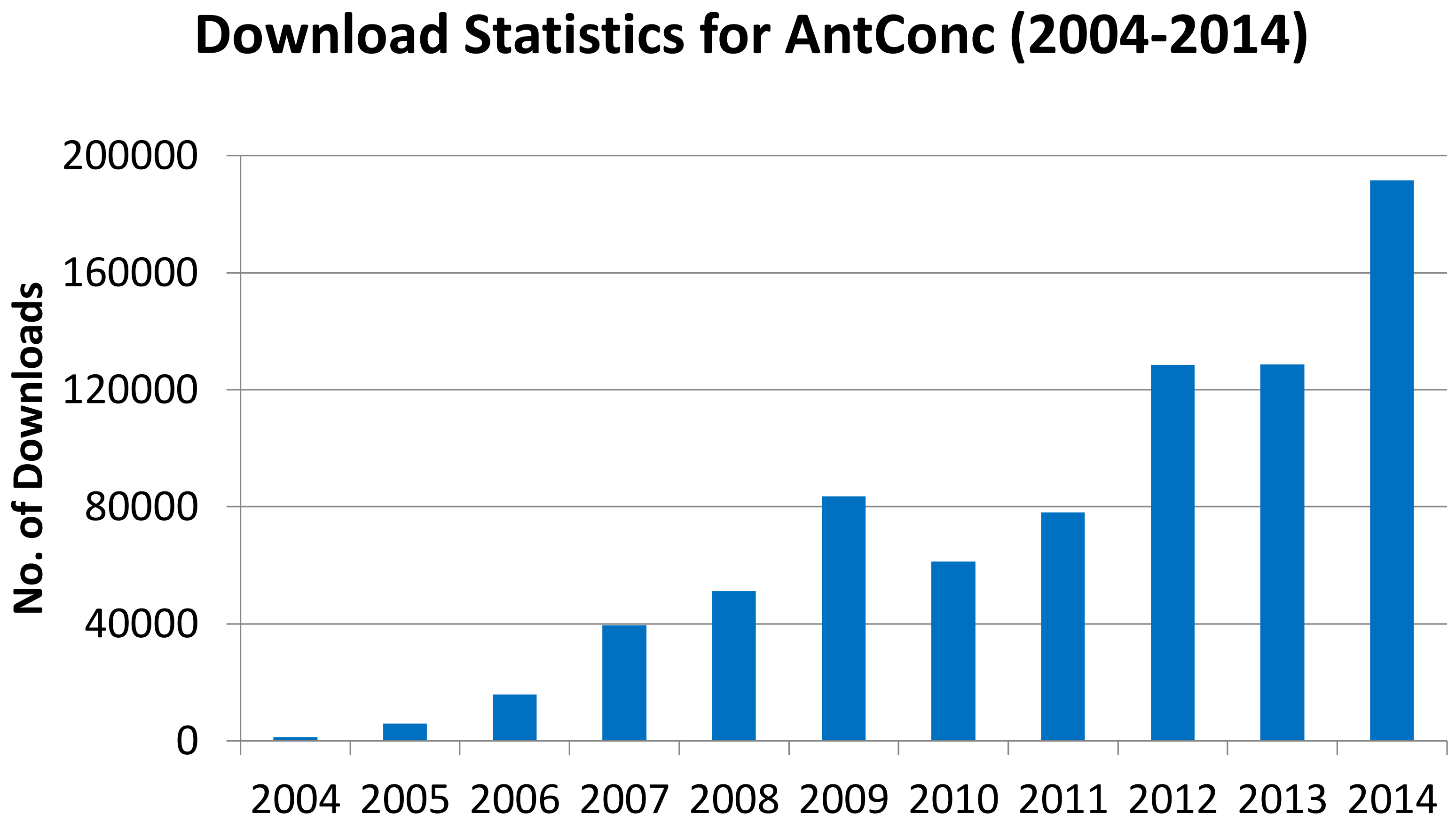
## Most popular tools for analyzing corpora



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

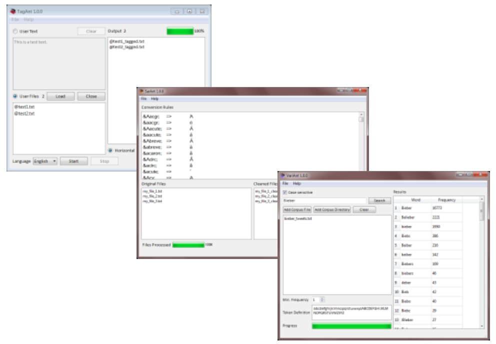
# Current state of corpus linguistics tools:

## Most popular tools for analyzing corpora



## The current state of corpus linguistics tools

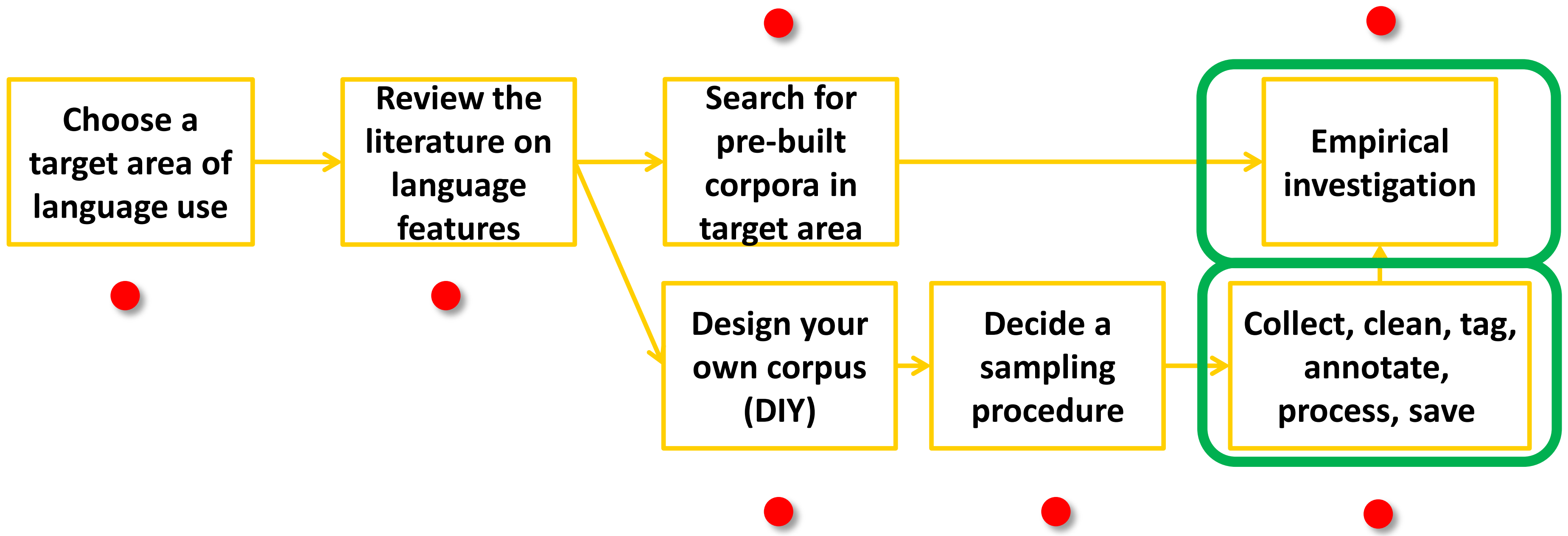
a need for something new



## Current state of corpus linguistics tools:

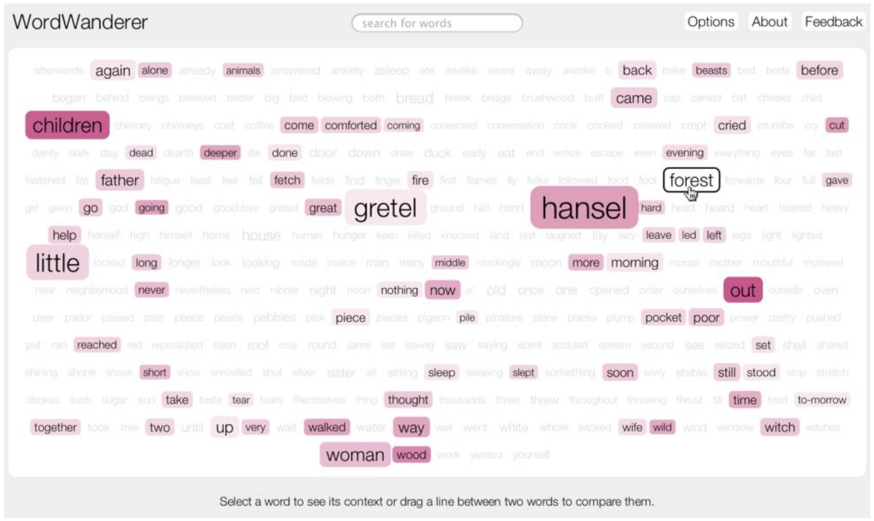
A need for something new

### Corpus Development Process


Choose a target area of language use
Review the literature on language features
Search for pre-built corpora in target area
Empirical investigation
Design your own corpus (DIY)
Decide a sampling procedure
Collect, clean, tag, annotate, process, save

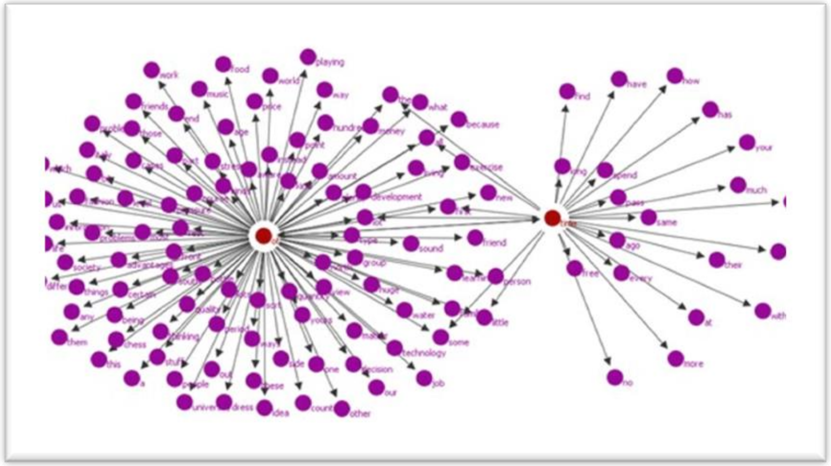
# Current state of corpus linguistics tools:

## A need for something new (exploratory tools)



Select a word to see its context or drag a line between two words to compare them.

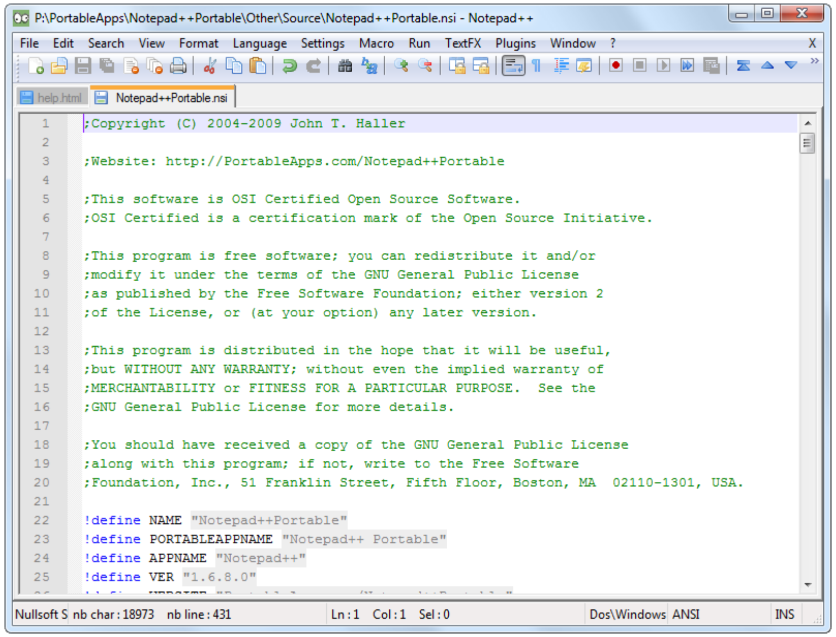
WordWanderer
http://wordwanderer.org/



GraphColl
http://www.extremetomato.com/projects/graphcoll/

# Current state of corpus linguistics tools:

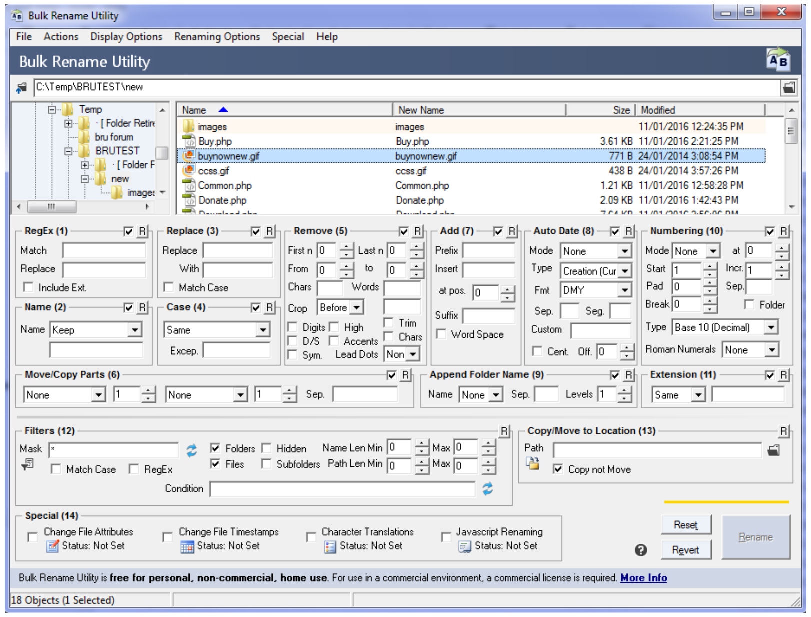
## A need for something new (collecting, cleaning, tagging tools)


P:\PortableApps\Notepad++Portable\Other\Source\Notepad++Portable.nsi - Notepad++
File Edit Search View Format Language Settings Macro Run TextFX Plugins Window ?
help.html
Notepad++Portable.nsi

```
;Copyright (C) 2004-2009 John T. Haller

;Website: http://PortableApps.com/Notepad++Portable

;This software is OSI Certified Open Source Software.
;OSI Certified is a certification mark of the Open Source Initiative.

;This program is free software; you can redistribute it and/or
;modify it under the terms of the GNU General Public License
;as published by the Free Software Foundation; either version 2
;of the License, or (at your option) any later version.

;This program is distributed in the hope that it will be useful,
;but WITHOUT ANY WARRANTY; without even the implied warranty of
;MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
;GNU General Public License for more details.

;You should have received a copy of the GNU General Public License
;along with this program; if not, write to the Free Software
;Foundation, Inc., 51 Franklin Street, Fifth Floor, Boston, MA  02110-1301, USA.

!define NAME "Notepad++Portable"
!define PORTABLEAPPNAME "Notepad++ Portable"
!define APPNAME "Notepad++"
!define VER "1.6.8.0"
```

```
Nullsoft S nb char : 18973   nb line : 431   Ln : 1   Col : 1   Sel : 0   Dos\Windows ANSI   INS
```

NotePad++
https://notepad-plus-plus.org/



C:\Temp\BRUTEST\new

| Name | New Name | Size | Modified |
|---|---|---|---|
| images | images | | 11/01/2016 12:24:35 PM |
| Buy.php | Buy.php | 3.61 KB | 11/01/2016 2:21:25 PM |
| buynownew.gif | buynownew.gif | 771 B | 24/01/2014 3:08:54 PM |
| ccss.gif | ccss.gif | 438 B | 24/01/2014 3:57:26 PM |
| Common.php | Common.php | 1.21 KB | 11/01/2016 12:58:28 PM |
| Donate.php | Donate.php | 2.09 KB | 11/01/2016 1:42:43 PM |

Name: Keep; Case: Same; Auto Date Mode: None; Type: Creation (Cur; Fmt: DMY; Numbering Mode: None; Start: 1; Incr.: 1; Pad: 0; Break: 0; Type: Base 10 (Decimal); Roman Numerals: None; Extension: Same; Mask: *; Copy not Move

Status: Not Set

Bulk Rename Utility is **free for personal, non-commercial, home use**. For use in a commercial environment, a commercial license is required. More Info

18 Objects (1 Selected)

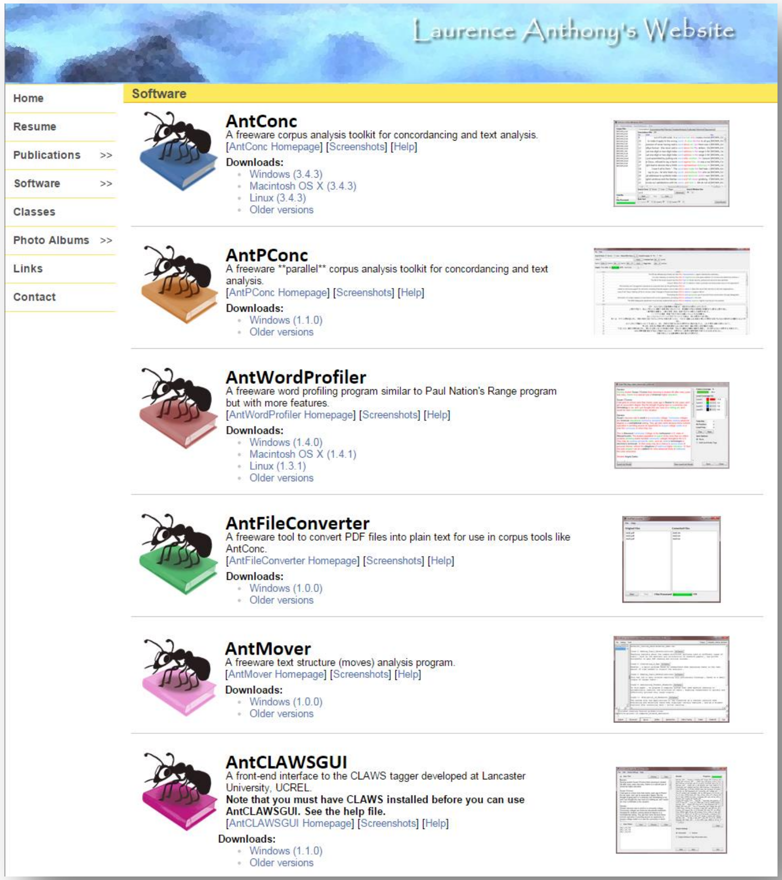
BulkFileRenamer
http://www.bulkrenameutility.co.uk/Main_Intro.php

# Current state of corpus linguistics tools:

## A need for something new

### AntLab Tools

www.laurenceanthony.net/software


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 (3.4.3)
- Macintosh OS X (3.4.3)
- Linux (3.4.3)
- Older versions

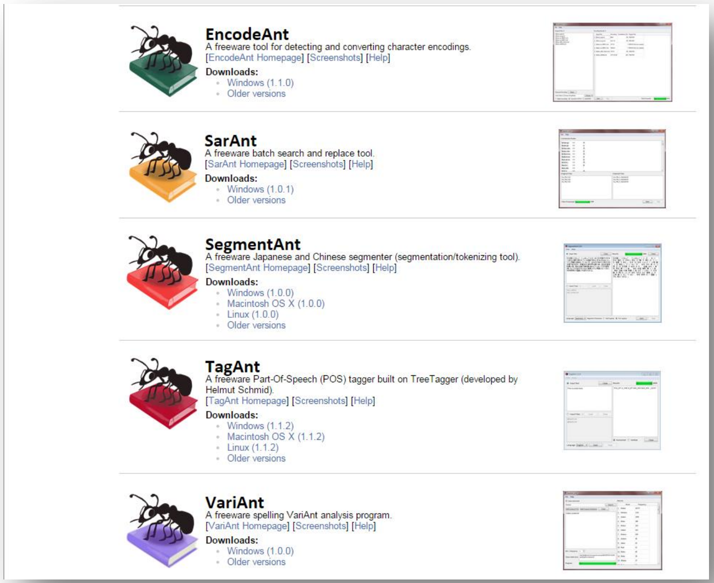
**AntPConc**
A freeware **parallel** corpus analysis toolkit for concordancing and text analysis.
[AntPConc Homepage] [Screenshots] [Help]
Downloads:
- Windows (1.1.0)
- Older versions

**AntWordProfiler**
A freeware word profiling program similar to Paul Nation's Range program but with more features.
[AntWordProfiler Homepage] [Screenshots] [Help]
Downloads:
- Windows (1.4.0)
- Macintosh OS X (1.4.1)
- Linux (1.3.1)
- Older versions

**AntFileConverter**
A freeware tool to convert PDF files into plain text for use in corpus tools like AntConc.
[AntFileConverter Homepage] [Screenshots] [Help]
Downloads:
- Windows (1.0.0)
- Older versions

**AntMover**
A freeware text structure (moves) analysis program.
[AntMover Homepage] [Screenshots] [Help]
Downloads:
- Windows (1.0.0)
- Older versions

**AntCLAWSGUI**
A front-end interface to the CLAWS tagger developed at Lancaster University, UCREL.
**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)
- Older versions

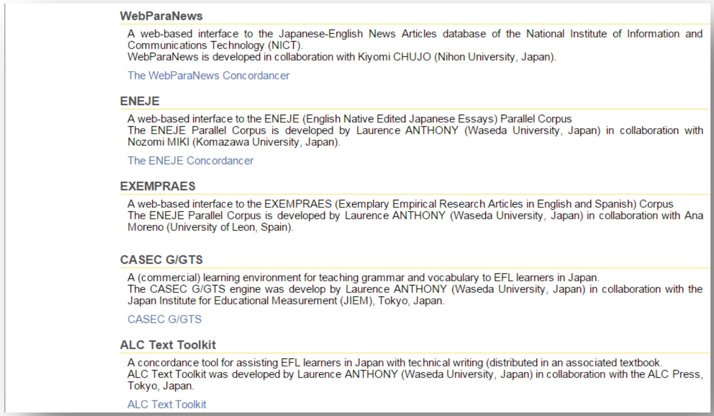

**EncodeAnt**
A freeware tool for detecting and converting character encodings.
[EncodeAnt Homepage] [Screenshots] [Help]
Downloads:
- Windows (1.1.0)
- Older versions

**SarAnt**
A freeware batch search and replace tool.
[SarAnt Homepage] [Screenshots] [Help]
Downloads:
- Windows (1.0.1)
- Older versions

**SegmentAnt**
A freeware Japanese and Chinese segmenter (segmentation/tokenizing tool).
[SegmentAnt Homepage] [Screenshots] [Help]
Downloads:
- Windows (1.0.0)
- Macintosh OS X (1.0.0)
- Linux (1.0.0)
- Older versions

**TagAnt**
A freeware Part-Of-Speech (POS) tagger built on TreeTagger (developed by Helmut Schmid).
[TagAnt Homepage] [Screenshots] [Help]
Downloads:
- Windows (1.1.2)
- Macintosh OS X (1.1.2)
- Linux (1.1.2)
- Older versions

**VariAnt**
A freeware spelling VariAnt analysis program.
[VariAnt Homepage] [Screenshots] [Help]
Downloads:
- Windows (1.0.0)
- Older versions


**WebParaNews**
A web-based interface to the Japanese-English News Articles database of the National Institute of Information and Communications Technology (NICT).
WebParaNews is developed in collaboration with Kiyomi CHUJO (Nihon University, Japan).
The WebParaNews Concordancer

**ENEJE**
A web-based interface to the ENEJE (English Native Edited Japanese Essays) Parallel Corpus
The ENEJE Parallel Corpus is developed by Laurence ANTHONY (Waseda University, Japan) in collaboration with Nozomi MIKI (Komazawa University, Japan).
The ENEJE Concordancer

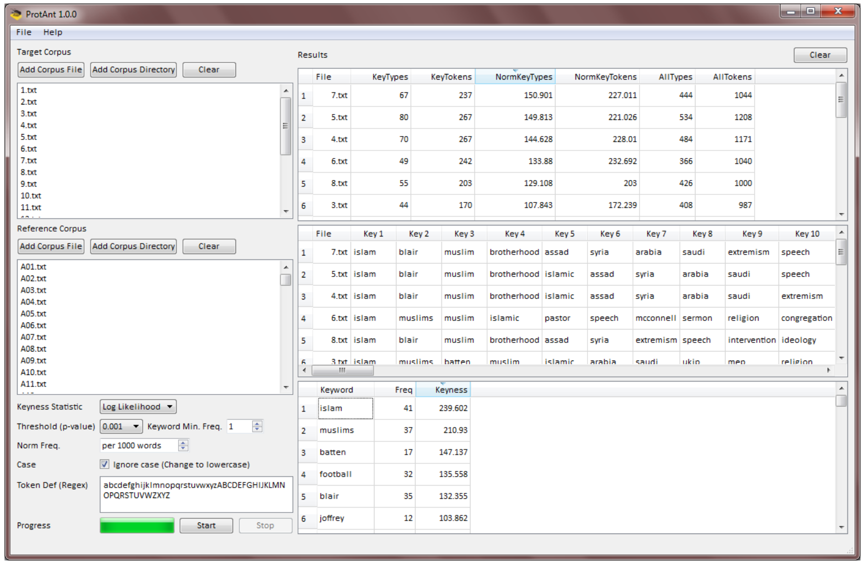
**EXEMPRAES**
A web-based interface to the EXEMPRAES (Exemplary Empirical Research Articles in English and Spanish) Corpus
The ENEJE Parallel Corpus is developed by Laurence ANTHONY (Waseda University, Japan) in collaboration with Ana Moreno (University of Leon, Spain).

**CASEC G/GTS**
A (commercial) learning environment for teaching grammar and vocabulary to EFL learners in Japan.
The CASEC G/GTS engine was develop by Laurence ANTHONY (Waseda University, Japan) in collaboration with the Japan Institute for Educational Measurement (JIEM), Tokyo, Japan.
CASEC G/GTS

**ALC Text Toolkit**
A concordance tool for assisting EFL learners in Japan with technical writing (distributed in an associated textbook.
ALC Text Toolkit was developed by Laurence ANTHONY (Waseda University, Japan) in collaboration with the ALC Press, Tokyo, Japan.
ALC Text Toolkit

# Current state of corpus linguistics tools:

## A need for something new (exploratory tools)



ProtAnt 1.0.0

File Help

Target Corpus

1.txt
2.txt
3.txt
4.txt
5.txt
6.txt
7.txt
8.txt
9.txt
10.txt
11.txt

Reference Corpus

A01.txt
A02.txt
A03.txt
A04.txt
A05.txt
A06.txt
A07.txt
A08.txt
A09.txt
A10.txt
A11.txt

Keyness Statistic: Log Likelihood
Threshold (p-value): 0.001 Keyword Min. Freq.: 1
Norm Freq.: per 1000 words
Case: [✓ Ignore case (Change to lowercase)]
Token Def (Regex): abcdefghijklmnopqrstuvwxyzABCDEFGHIJKLMN OPQRSTUVWZXYZ
Progress

Results

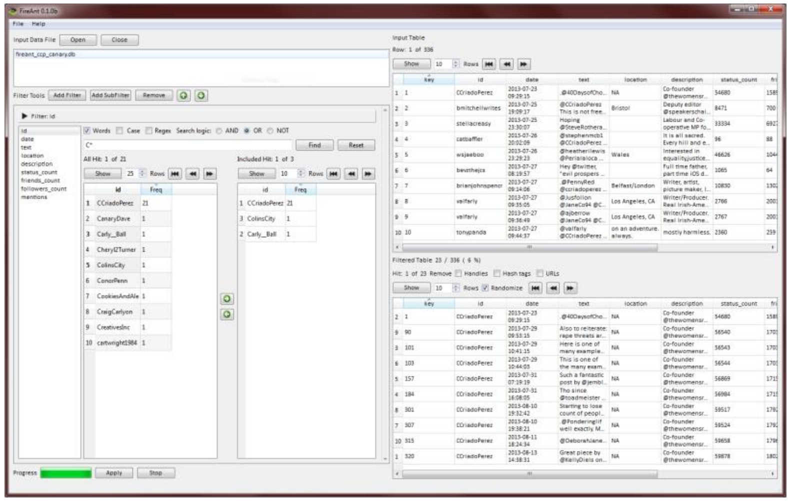
| | File | KeyTypes | KeyTokens | NormKeyTypes | NormKeyTokens | AllTypes | AllTokens |
|---|---|---|---|---|---|---|---|
| 1 | 7.txt | 67 | 237 | 150.901 | 227.011 | 444 | 1044 |
| 2 | 5.txt | 80 | 267 | 149.813 | 221.026 | 534 | 1208 |
| 3 | 4.txt | 70 | 267 | 144.628 | 228.01 | 484 | 1171 |
| 4 | 6.txt | 49 | 242 | 133.88 | 232.692 | 366 | 1040 |
| 5 | 8.txt | 55 | 203 | 129.108 | 203 | 426 | 1000 |
| 6 | 3.txt | 44 | 170 | 107.843 | 172.239 | 408 | 987 |

| | File | Key 1 | Key 2 | Key 3 | Key 4 | Key 5 | Key 6 | Key 7 | Key 8 | Key 9 | Key 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 7.txt | islam | blair | muslim | brotherhood | assad | syria | arabia | saudi | extremism | speech |
| 2 | 5.txt | islam | blair | muslim | brotherhood | islamic | assad | syria | arabia | saudi | speech |
| 3 | 4.txt | islam | blair | muslim | brotherhood | islamic | assad | syria | arabia | saudi | extremism |
| 4 | 6.txt | islam | muslims | muslim | islamic | pastor | speech | mcconnell | sermon | religion | congregation |
| 5 | 8.txt | islam | blair | muslim | brotherhood | assad | syria | extremism | speech | intervention | ideology |
| 6 | 3.txt | islam | muslims | batten | muslim | islamic | arabia | saudi | ukip | men | religion |

| | Keyword | Freq | Keyness |
|---|---|---|---|
| 1 | islam | 41 | 239.602 |
| 2 | muslims | 37 | 210.93 |
| 3 | batten | 17 | 147.137 |
| 4 | football | 32 | 135.558 |
| 5 | blair | 35 | 132.355 |
| 6 | joffrey | 12 | 103.862 |

ProtAnt

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



FireAnt 0.1.0b

File Help

Input Data File: fireant_ccp_canary.db

[✓ Words] Case Regex Search logic: AND [✓ OR] NOT

C*

All Hit: 1 of 21 — Show 25 Rows

| | id | Freq |
|---|---|---|
| 1 | CCriadoPerez | 21 |
| 2 | CanaryDave | 1 |
| 3 | Carly__Ball | 1 |
| 4 | Cheryl2Turner | 1 |
| 5 | ColinsCity | 1 |
| 6 | ConorPenn | 1 |
| 7 | CookiesAndAle | 1 |
| 8 | CraigCarlyon | 1 |
| 9 | CreativesInc | 1 |
| 10 | cartwright1984 | 1 |

Included Hit: 1 of 3 — Show 10 Rows

| | id | Freq |
|---|---|---|
| 1 | CCriadoPerez | 21 |
| 3 | ColinsCity | 1 |
| 2 | Carly__Ball | 1 |

Progress

Input Table

Row: 1 of 336 — Show 10 Rows

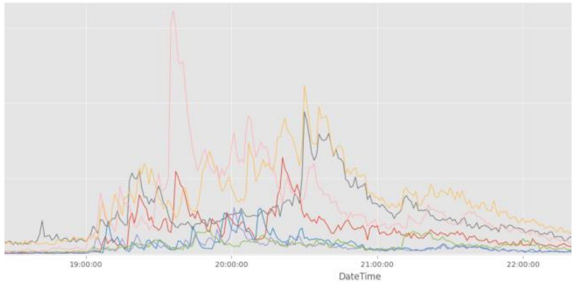
| | key | id | date | text | location | description | status_count |
|---|---|---|---|---|---|---|---|
| 1 | 1 | CCriadoPerez | 2013-07-23 09:29:15 | .@40DaysofCho... | NA | Co-founder @thewomensr... | 54680 |
| 2 | 2 | bmitchellwrites | 2013-07-25 19:09:17 | @CCriadoPerez This is not free... | Bristol | Deputy editor @speakerschai... | 8471 |
| 3 | 3 | stellacreasy | 2013-07-25 23:30:07 | Hoping @SteveRothera... | | Labour and Co-operative MP fo... | 33334 |
| 4 | 4 | catbaffler | 2013-07-26 20:02:09 | @stephenmcb1 @CCriadoPerez ... | | It is all sacred. Every hill and e... | 96 |
| 5 | 5 | wsjaeboo | 2013-07-26 23:29:23 | @heatherllewis @Perlaialoca ... | Wales | Interested in equality,justice... | 46626 |
| 6 | 6 | bevthejcs | 2013-07-27 08:19:57 | Hey @twitter, "evil prospers ... | | Full time father, part time iOS d... | 1065 |
| 7 | 7 | brianjohnspencr | 2013-07-27 09:14:06 | .@PennyRed @ccriadoperez ... | Belfast/London | Writer, artist, picture maker, I... | 10890 |
| 8 | 8 | valfarly | 2013-07-27 09:35:05 | @Jusfollon @JaneCo94 @C... | Los Angeles, CA | Writer/Producer. Real Irish-Ame... | 2766 |
| 9 | 9 | valfarly | 2013-07-27 09:36:49 | @ajberrow @JaneCo94 @C... | Los Angeles, CA | Writer/Producer. Real Irish-Ame... | 2767 |
| 10 | 10 | tonypanda | 2013-07-27 09:44:37 | @valfarly @CCriadoPerez ... | on an adventure. always. | mostly harmless. | 2360 |

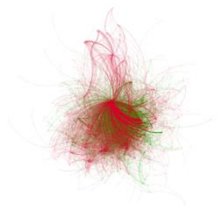
Filtered Table 23 / 336 ( 6 %)

Hit: 1 of 23 Remove Handles Hash tags URLs

Show 10 Rows [✓ Randomize]

| | key | id | date | text | location | description | status_count |
|---|---|---|---|---|---|---|---|
| 2 | 1 | CCriadoPerez | 2013-07-23 09:29:15 | .@40DaysofCho... | NA | Co-founder @thewomensr... | 54680 |
| 9 | 90 | CCriadoPerez | 2013-07-29 09:53:15 | Also to reiterate: rape threats ar... | NA | Co-founder @thewomensr... | 56540 |
| 3 | 101 | CCriadoPerez | 2013-07-29 10:41:15 | Here is one of many example... | NA | Co-founder @thewomensr... | 56543 |
| 6 | 103 | CCriadoPerez | 2013-07-29 10:44:03 | This is one of the many exam... | NA | Co-founder @thewomensr... | 56544 |
| 5 | 157 | CCriadoPerez | 2013-07-31 07:19:19 | Such a fantastic post by @jembl... | NA | Co-founder @thewomensr... | 56869 |
| 4 | 184 | CCriadoPerez | 2013-07-31 16:08:05 | Tho since @toadmeister ... | NA | Co-founder @thewomensr... | 56984 |
| 8 | 301 | CCriadoPerez | 2013-08-10 19:32:42 | Starting to lose count of peopl... | NA | Co-founder @thewomensr... | 59517 |
| 7 | 307 | CCriadoPerez | 2013-08-10 19:38:21 | .@Ponderinglif well exactly M... | NA | Co-founder @thewomensr... | 59524 |
| 10 | 315 | CCriadoPerez | 2013-08-11 18:24:34 | @DeborahJane... | NA | Co-founder @thewomensr... | 59658 |
| 1 | 320 | CCriadoPerez | 2013-08-13 14:38:31 | Great piece by @KellyDiels on... | NA | Co-founder @thewomensr... | 59878 |



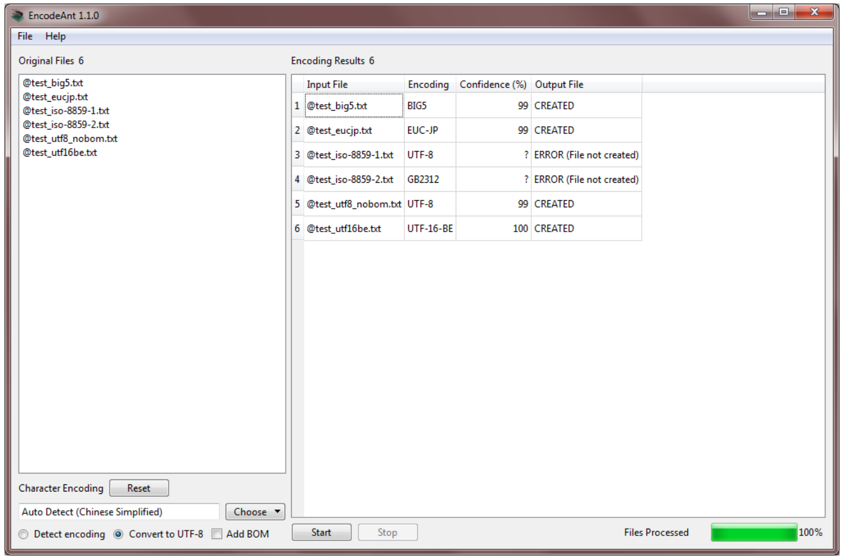


FireAnt

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

# Current state of corpus linguistics tools:

## A need for something new (collecting, cleaning, tagging tools)



Original Files 6

@test_big5.txt
@test_eucjp.txt
@test_iso-8859-1.txt
@test_iso-8859-2.txt
@test_utf8_nobom.txt
@test_utf16be.txt

Encoding Results 6

| | Input File | Encoding | Confidence (%) | Output File |
|---|---|---|---|---|
| 1 | @test_big5.txt | BIG5 | 99 | CREATED |
| 2 | @test_eucjp.txt | EUC-JP | 99 | CREATED |
| 3 | @test_iso-8859-1.txt | UTF-8 | ? | ERROR (File not created) |
| 4 | @test_iso-8859-2.txt | GB2312 | ? | ERROR (File not created) |
| 5 | @test_utf8_nobom.txt | UTF-8 | 99 | CREATED |
| 6 | @test_utf16be.txt | UTF-16-BE | 100 | CREATED |

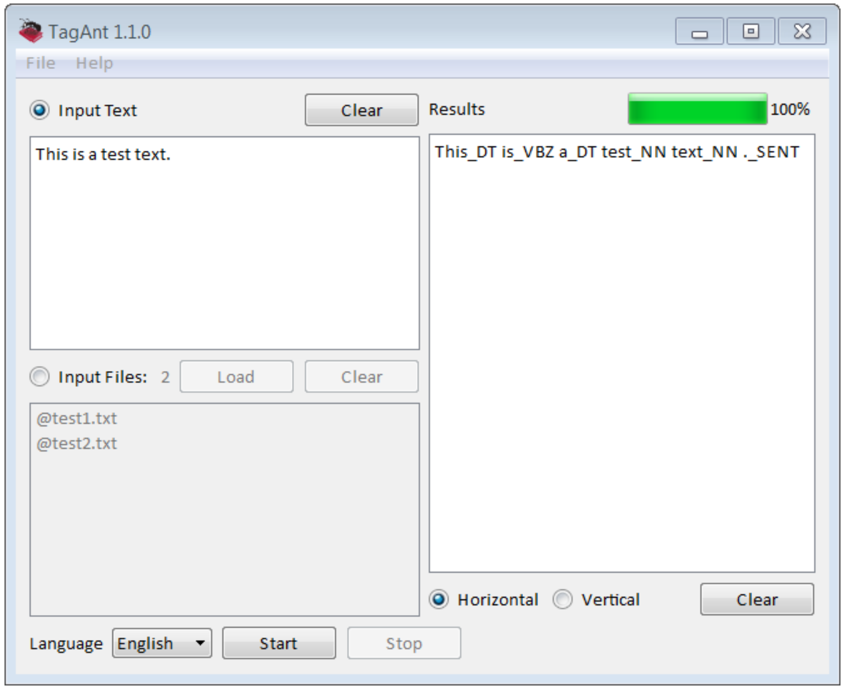
Character Encoding: Auto Detect (Chinese Simplified)

Detect encoding [✓ Convert to UTF-8] Add BOM

Files Processed 100%

EncodeAnt

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



[✓ Input Text]

This is a test text.

Input Files: 2

@test1.txt
@test2.txt

Results 100%

This_DT is_VBZ a_DT test_NN text_NN ._SENT

[✓ Horizontal] Vertical

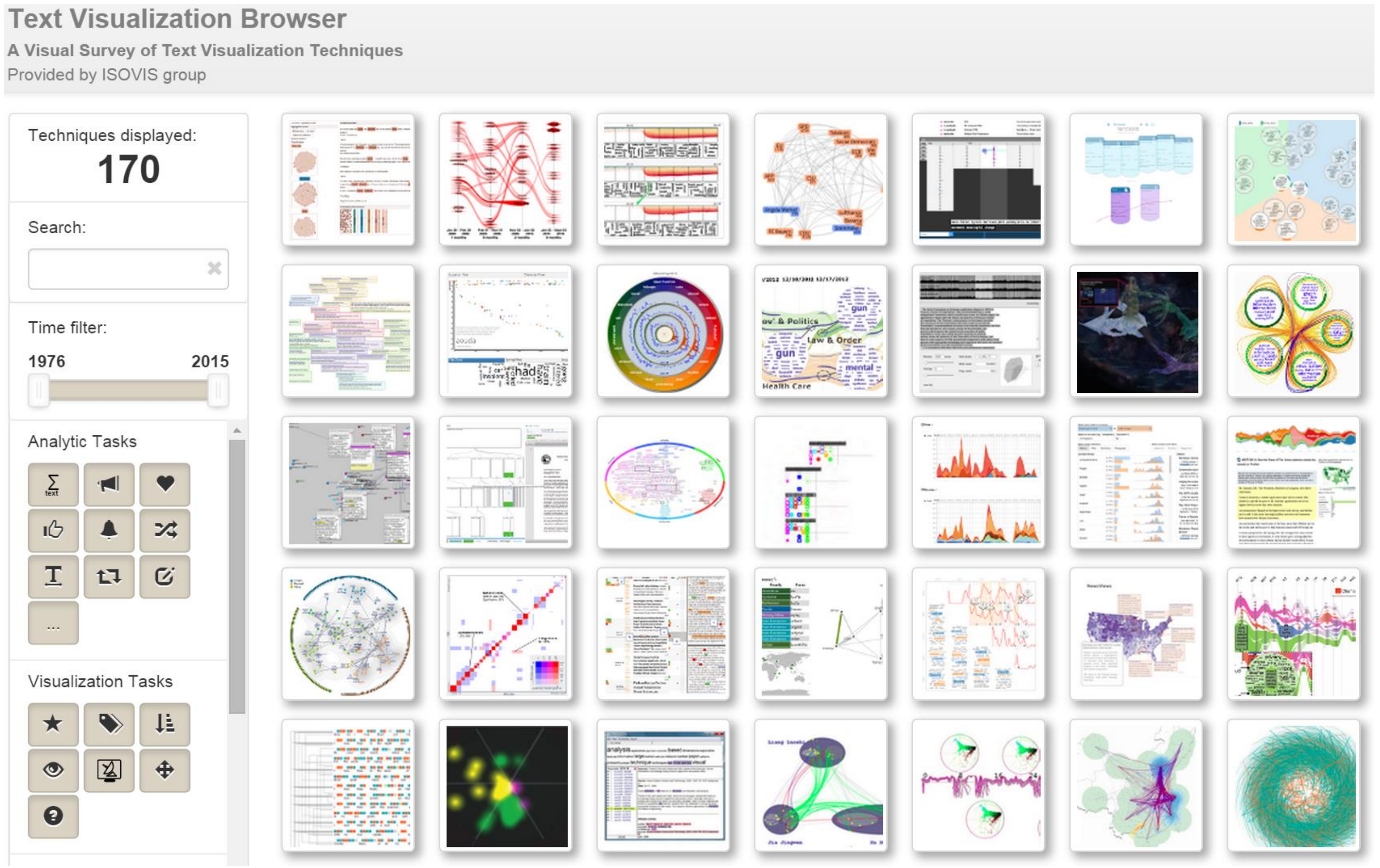
Language English

TagAnt

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

# Current state of corpus linguistics tools:

## The power of R (and Python, Java, …)


Text Visualization Browser

A Visual Survey of Text Visualization Techniques

Provided by ISOVIS group

Techniques displayed:

170

Search:

Time filter:

1976 2015

Text Visualization Browser
http://textvis.lnu.se/

## Current state of corpus linguistics tools:
The power of R (and Python, Java, …)

**In a search for ditransitive constructions…**

“As a heuristic, we used a script (less than 60 lines) that recovered all verb tokens tagged as used ditransitively in the ICE-GB, looked up the lemmas for these tokens in a lemma list, looked up all the forms for these lemmas in the lemma list… and then outputted a concordance of all matches of those forms in the learner corpus”

“This is not perfect, but it is easy to see that no ready-made program could ever do this (especially not quickly)”

(Gries, 2011: 93-94)

## DIY Corpus Tools – The Debate

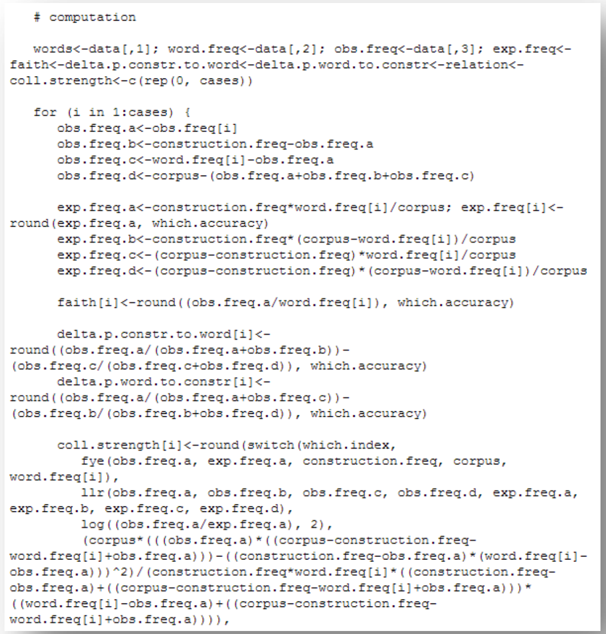
Arguments for learning to program

http://d.ibtimes.co.uk/en/full/339496/rock-john-cena.jpg?w=500

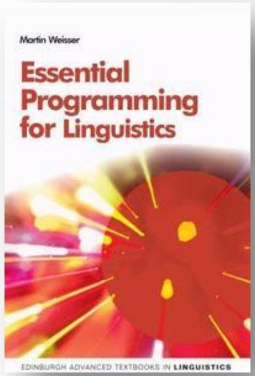
# DIY Corpus Tools – The Debate:

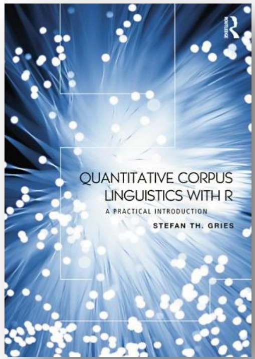
## Arguments for learning to program

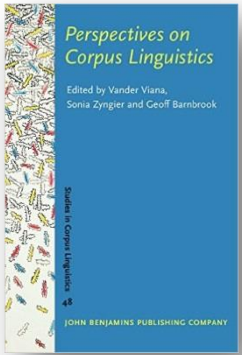
- Corpus linguists should learn to program…
  - (Biber, Weisser, Gries, Davies, ….)



```
   # computation

   words<-data[,1]; word.freq<-data[,2]; obs.freq<-data[,3]; exp.freq<-
faith<-delta.p.constr.to.word<-delta.p.word.to.constr<-relation<-
coll.strength<-c(rep(0, cases))

   for (i in 1:cases) {
      obs.freq.a<-obs.freq[i]
      obs.freq.b<-construction.freq-obs.freq.a
      obs.freq.c<-word.freq[i]-obs.freq.a
      obs.freq.d<-corpus-(obs.freq.a+obs.freq.b+obs.freq.c)

      exp.freq.a<-construction.freq*word.freq[i]/corpus; exp.freq[i]<-
round(exp.freq.a, which.accuracy)
      exp.freq.b<-construction.freq*(corpus-word.freq[i])/corpus
      exp.freq.c<-(corpus-construction.freq)*word.freq[i]/corpus
      exp.freq.d<-(corpus-construction.freq)*(corpus-word.freq[i])/corpus

      faith[i]<-round((obs.freq.a/word.freq[i]), which.accuracy)

      delta.p.constr.to.word[i]<-
round((obs.freq.a/(obs.freq.a+obs.freq.b))-
(obs.freq.c/(obs.freq.c+obs.freq.d)), which.accuracy)
      delta.p.word.to.constr[i]<-
round((obs.freq.a/(obs.freq.a+obs.freq.c))-
(obs.freq.b/(obs.freq.b+obs.freq.d)), which.accuracy)

      coll.strength[i]<-round(switch(which.index,
         fye(obs.freq.a, exp.freq.a, construction.freq, corpus,
word.freq[i]),
         llr(obs.freq.a, obs.freq.b, obs.freq.c, obs.freq.d, exp.freq.a,
exp.freq.b, exp.freq.c, exp.freq.d),
         log((obs.freq.a/exp.freq.a), 2),
         (corpus*(((obs.freq.a)*((corpus-construction.freq-
word.freq[i]+obs.freq.a)))-((construction.freq-obs.freq.a)*(word.freq[i]-
obs.freq.a)))^2)/(construction.freq*word.freq[i]*((construction.freq-
obs.freq.a)+((corpus-construction.freq-word.freq[i]+obs.freq.a)))*
((word.freq[i]-obs.freq.a)+((corpus-construction.freq-
word.freq[i]+obs.freq.a)))),
```


Martin Weisser
Essential Programming for Linguistics
EDINBURGH ADVANCED TEXTBOOKS IN LINGUISTICS




Perspectives on Corpus Linguistics
Edited by Vander Viana, Sonia Zyngier and Geoff Barnbrook
Studies in Corpus Linguistics 48
JOHN BENJAMINS PUBLISHING COMPANY

## DIY Corpus Tools – The Debate:

### Arguments for learning to program

If you program ...

"you can do analyses not possible with concordancers ...
you can do analyses more quickly and more accurately ...
you can tailor the output to fit your own research needs ...
you can analyze a corpus of any size"

(Biber et al., 1998, p. 256)

## DIY Corpus Tools – The Debate:

### Arguments for learning to program

"when you use pre-configured corpus programs, you're a little bit at the mercy of the company or individual selling them …

One final big advantage of programming languages, therefore, is that you are in the driver's seat."

(Gries, 2009, p. 11-12)

# DIY Corpus Tools – The Debate:

## Arguments for learning to program

"Every corpus-linguistic researcher should have some programming skills"

Reason 1: software many people use is severely limited in terms of

- availability (OS, cost)
- functionality (only what is hardwired)
- user-control (at the mercy of developers)

Reason 2: "inflexible software creates inflexible researchers"

(Gries, 2011: 92-94)

# DIY Corpus Tools – The Debate:

## Arguments for learning to program

"Here, I think, we must divide researchers into two camps – corpus users and corpus creators. Corpus users can often get by with stand-alone tools or web-based corpora…For corpus creators, however, I would say that some experience with programming is a necessity."

(Davies, 2011: 77)

## DIY Corpus Tools – The Debate:
### Arguments for learning to program

A caveat...

For corpus users...
"even a little knowledge of regular expressions will go a long way in helping with more complex queries."

For corpus creators (in more simple cases)...
"perhaps regular expressions and a simple knowledge of semi-automated file handling would be sufficient."

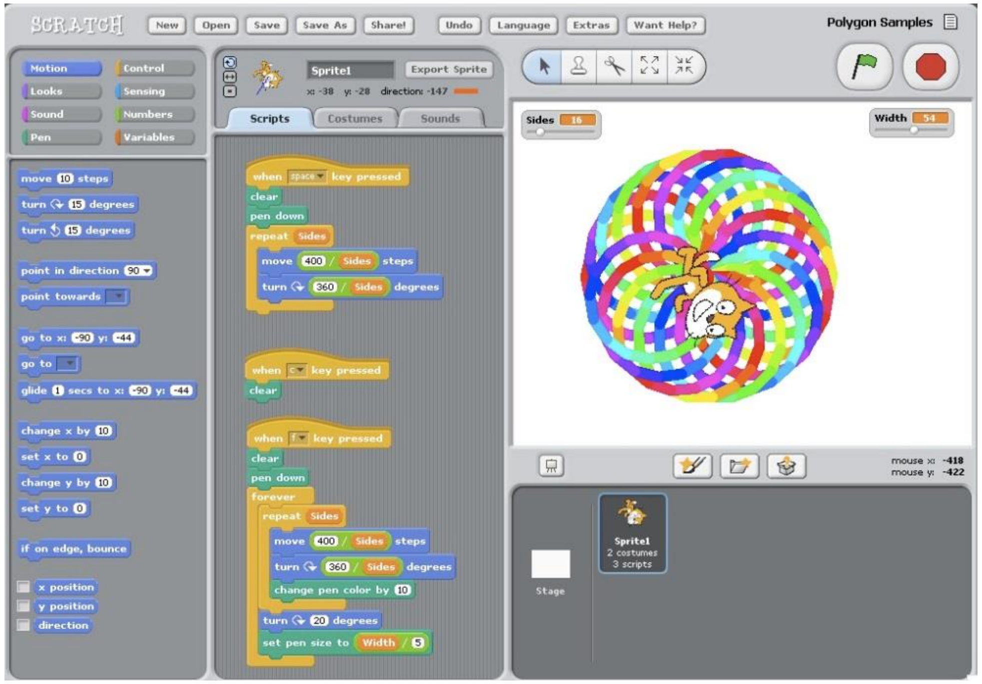
(Davies, 2011: 77-78)

# DIY Corpus Tools – The Debate:
## Arguments for learning to program

- Where to start with programming…
  - Pick a popular language
    - My languages of choice
      - Python (standalone tools), JavaScript/PHP (web programming)
    - Suggestions for **learning programming** (in order of preference)
      - Scratch (BYOB), Python, Java
    - Suggestions for **learning corpus linguistics programming** (in order of preference)
      - R, Python, JavaScript (PHP?, Java??, Perl???, Pascal????)
  - Read a programming book or online tutorial (or join a MOOC)
    - e.g. Teach Yourself Perl 5 in 21 Days
    - e.g. Learn Python The Hard Way (http://learnpythonthehardway.org/book/)
  - Join **the one truly amazing** programming forum
    - Stack Overflow (http://stackoverflow.com/)

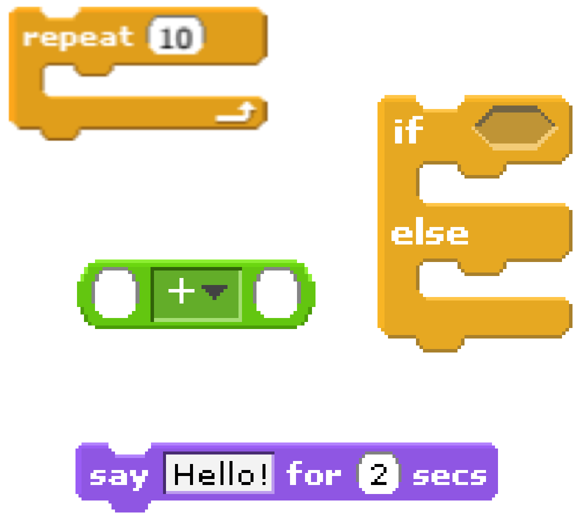
# DIY Corpus Tools – The Debate:
## Arguments for learning to program

- The Scratch programming interface



```
move 10 steps
turn ↻ 15 degrees
turn ↺ 15 degrees
point in direction 90
point towards
go to x: -90 y: -44
go to
glide 1 secs to x: -90 y: -44
change x by 10
set x to 0
change y by 10
set y to 0
if on edge, bounce
x position
y position
direction
```

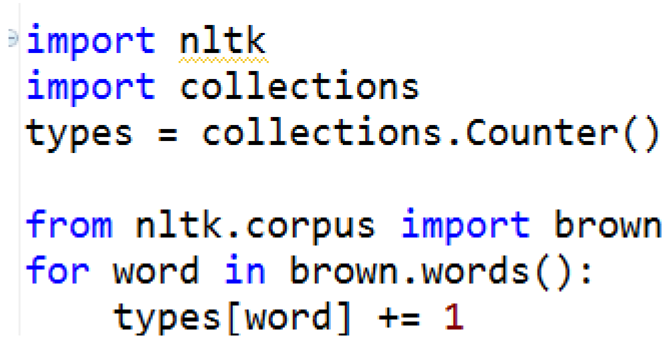
```
when space key pressed
clear
pen down
repeat Sides
    move 400 / Sides steps
    turn ↻ 360 / Sides degrees

when c key pressed
clear

when f key pressed
clear
pen down
forever
    repeat Sides
        move 400 / Sides steps
        turn ↻ 360 / Sides degrees
        change pen color by 10
    turn ↻ 20 degrees
    set pen size to Width / 5
```



# DIY Corpus Tools – The Debate:

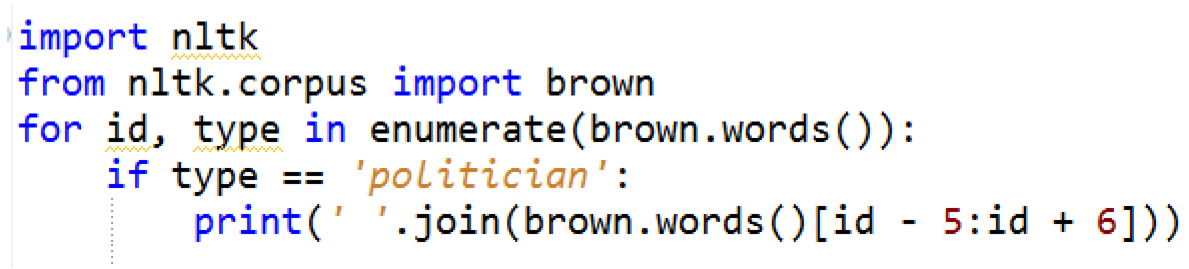
## Arguments for learning to program

- A (Brown Corpus) word list tool in Python…



```python
import nltk
import collections
types = collections.Counter()

from nltk.corpus import brown
for word in brown.words():
    types[word] += 1
```

- A (Brown Corpus) KWIC Concordancer in Python…



```python
import nltk
from nltk.corpus import brown
for id, type in enumerate(brown.words()):
    if type == 'politician':
        print(' '.join(brown.words()[id - 5:id + 6]))
```

## DIY Corpus Tools – The Debate

Arguments against learning to program

http://d.ibtimes.co.uk/en/full/339496/rock-john-cena.jpg?w=500

# DIY Corpus Tools – The Debate:

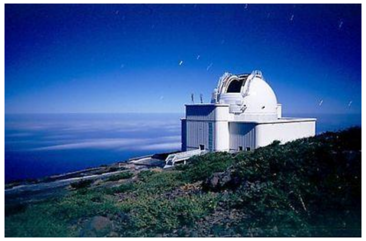
## Arguments against learning to program

- Argument 1:
  - Most corpus linguists are corpus users (not corpus creators)
    - [We can 'get by' with current corpus tools]
- Rebuttal
  - Using ready-built tools 'imprisons' the corpus linguist preventing them from developing new methods, analyzing interesting data sets, and deriving novel interpretations of that data
    - "AntConc's availability only in compiled form makes running it problematic" (AntConc userFeb. 10, 2016)
    - "Many 'stand-alone' programs to analyze corpora are not scalable enough to handle new, 'super-sized' corpora (Davies, 2011: 74)
    - "if the commercial software is not designed to produce the desired results, then the corpus linguist without programming experience either has to live with a potentially foul compromise or drop the project" (Gries, 2011: 94)

## DIY Corpus Tools – The Debate:
### Arguments against learning to program

- Argument 2:
  - Researchers in many fields do not develop their own tools
    - e.g. astronomers, biologists, (doctors), …

Home (Optical) Telescope


Issac Newton (Optical) Telescope

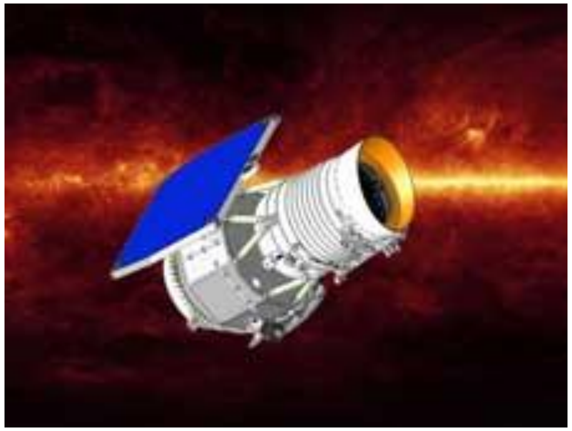
The Hubble (Optical) Telescope

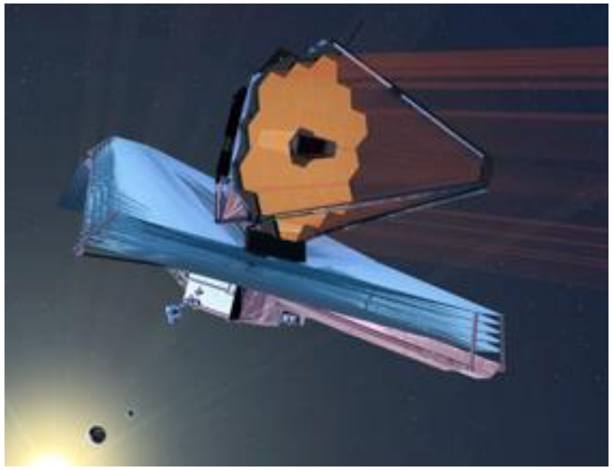
The Fermi Gamma-ray Space Telescope

Jodrell Bank Radio Telescope

## DIY Corpus Tools – The Debate:
### Arguments against learning to program

- "How are new tools developed for astronomy?


Wide-Field Infrared Explorer

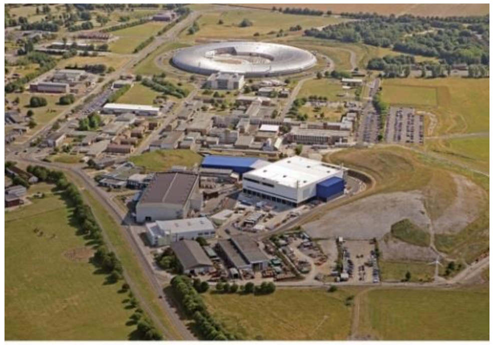

James Webb Space Telescope

Professor Jim Wild, Lancaster University
Vice-President, Royal Astronomical Society

# DIY Corpus Tools – The Debate:

## Arguments against learning to program

- "How are new tools developed for astronomy?"
  - Collaborations between astronomers and engineers
  - Massive funding for tool development
    - STFC (Science and Technology Facilities Council) national laboratories
      - ISIS, Diamond, Central Laser Facility
    - STFC technology centers
    - Specialist laboratories
      - Rutherford Appleton Laboratory
      - Space Magnetometer Laboratory (Imperial College)
      - Particle sensors laboratory (University College London)



http://www.clf.stfc.ac.uk/CLF/resources/image/jpg/ral_aerial_photo.jpg

## DIY Corpus Tools – The Debate:

### Arguments against learning to program

- Why should we be hermits trying to develop corpus tools on our own?

http://www.sai.msu.su/cjackson/dou/dou32.jpg

## DIY Corpus Tools – The Debate:

### Arguments against learning to program

- **Arts and Law**
  - English, Drama and American & Canadian Studies; History and Cultures; Languages, Cultures, Art History and Music; Birmingham Law School; Philosophy, Theology and Religion
- **Engineering and Physical Sciences**
  - Chemistry; Chemical Engineering; Civil Engineering; Computer Science; Electronic, Electrical and Computer Engineering; Mathematics; Mechanical Engineering; Metallurgy and Materials; Physics and Astronomy
- **Life and Environmental Sciences**
  - Biosciences; Geography, Earth and Environmental Sciences; Psychology; Sport and Exercise Sciences
- **Medical and Dental Sciences**
  - Cancer Sciences; Clinical and Experimental Medicine; Dentistry; Health and Population Sciences; Immunity and Infection
- **Social Sciences**
  - Birmingham Business School; Education; Government and Society; Social Policy
- **Liberal Arts and Sciences**

## DIY Corpus Tools – The Debate:
### Arguments against learning to program

- Argument 3:

The reality for most corpus researchers, however, is that computer programming is in a completely different world … without extensive training in programming … it is likely that these [DIY] tools would be more restrictive, slower, less accurate and only work with small corpora.

(Anthony, 2009, p. 95)

## DIY Corpus Tools – The Debate:

### Arguments against learning to program

- Where to start when using standard tools…
  - Decide your research question before selecting your tool/method

"Research should be led by the science not the tool."

Professor Jim Wild, Lancaster University
Vice-President, Royal Astronomical Society

# DIY Corpus Tools – The Debate:

## Arguments against learning to program

- Where to start when using standard tools…
  - Decide your research question before selecting your tool/method
  - Learn to use a "good" text editor:
    Notepad++ (Win), TextWrangler (Mac)
    - Unicode support (reading/converting text encodings)
    - Batch file handling
    - Regular Expressions (Regex) search/replace
  - Read the user guide of your chosen tool
    - Can it handle Unicode data?
    - Can it perform Regular Expression (Regex) searches?
    - Can it output results that you can feed into other tools (e.g. Excel/SPSS)?
  - Be proactive in contacting software developers
    - Explain clearly want you want to do (not how you think you should do it)
    - Provide motivation for them to get involved (what will they get out of it?)
    - Treat them as part of the team (not just a technical staff member)

## DIY Corpus Tools – The Debate:

### Arguments against learning to program

"Regardless of the project or the resources being used, researchers should attempt to understand (a) the limitations of the tools they are using and (b) what the alternatives are."

(Davies, 2011: 77)

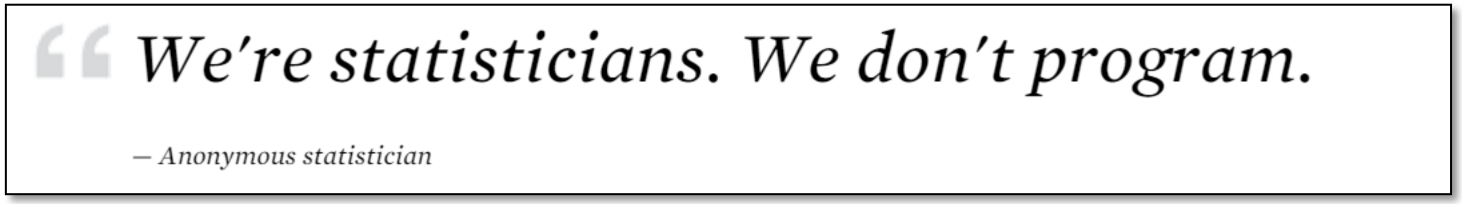
## Thoughts on the future of programming and tools in corpus linguistics research

programming, tools, and statistics, project teams

http://cdn.phys.org/newman/gfx/news/hires/2012/scientistsce.jp

# Thoughts on the future of programming and tools in corpus linguistics research:

- Programming, tools, and statistics

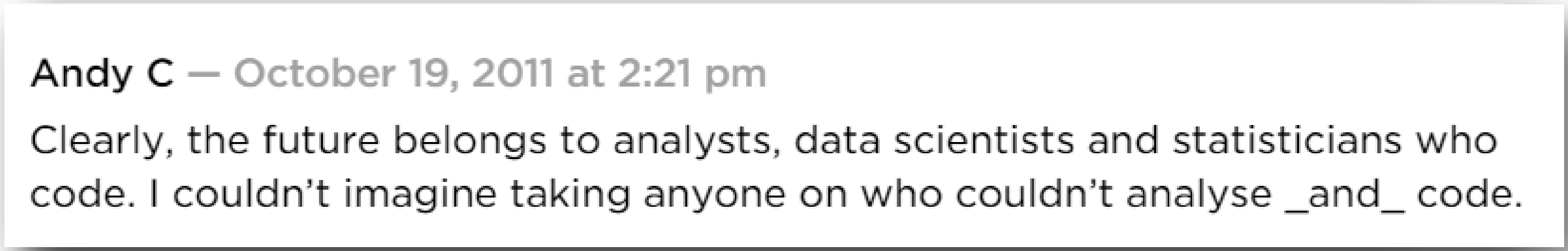


*We're statisticians. We don't program.*

*— Anonymous statistician*

https://flowingdata.com/2011/10/18/statisticians-dont-program/

- Statistics has evolved enormously through the development of software (and hardware) tools
- But… not all statisticians are programmers
- Should statisticians program?

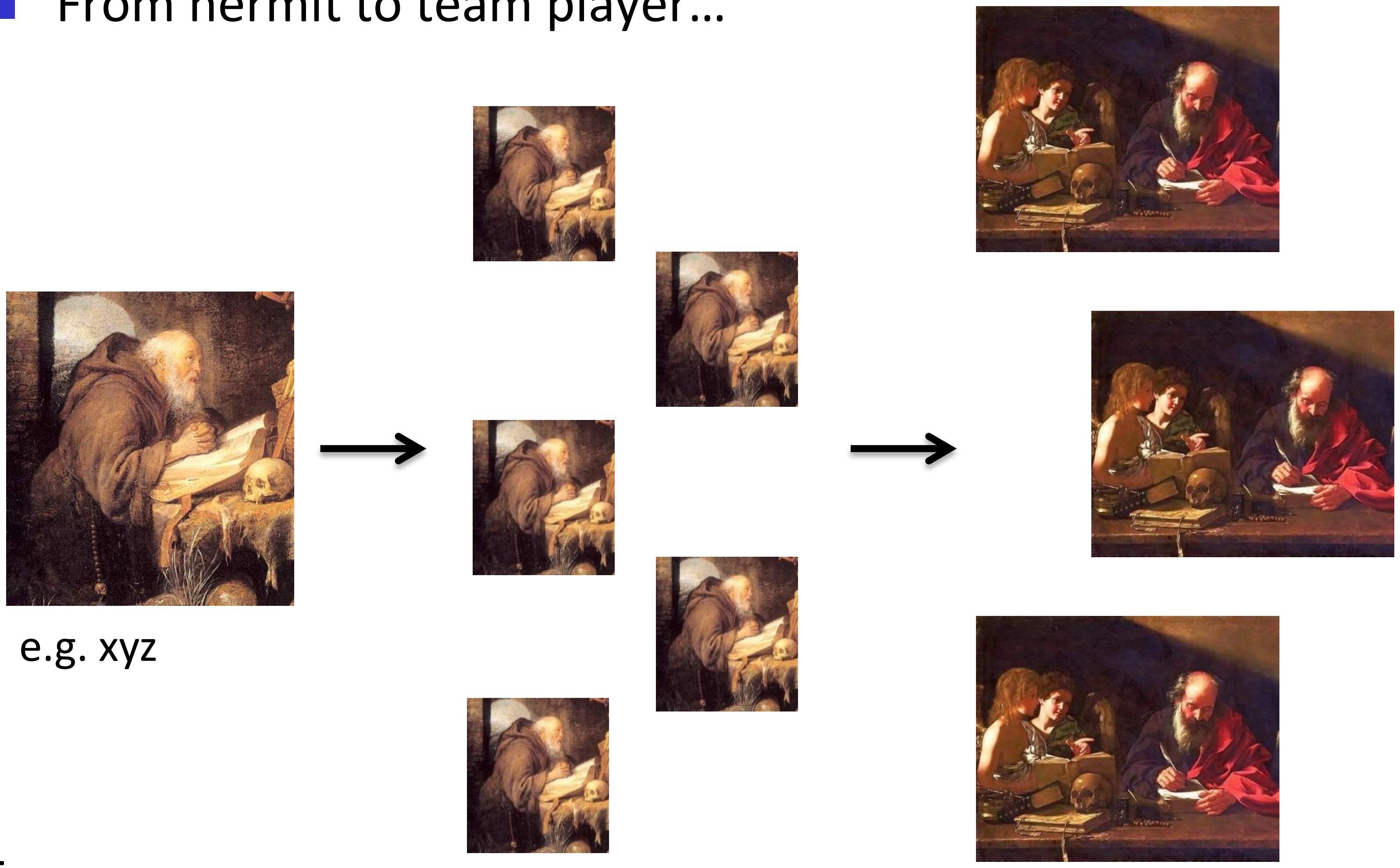


Andy C — October 19, 2011 at 2:21 pm

Clearly, the future belongs to analysts, data scientists and statisticians who code. I couldn't imagine taking anyone on who couldn't analyse _and_ code.

# Thoughts on the future of programming and tools in corpus linguistics research:

- From hermit to team player…



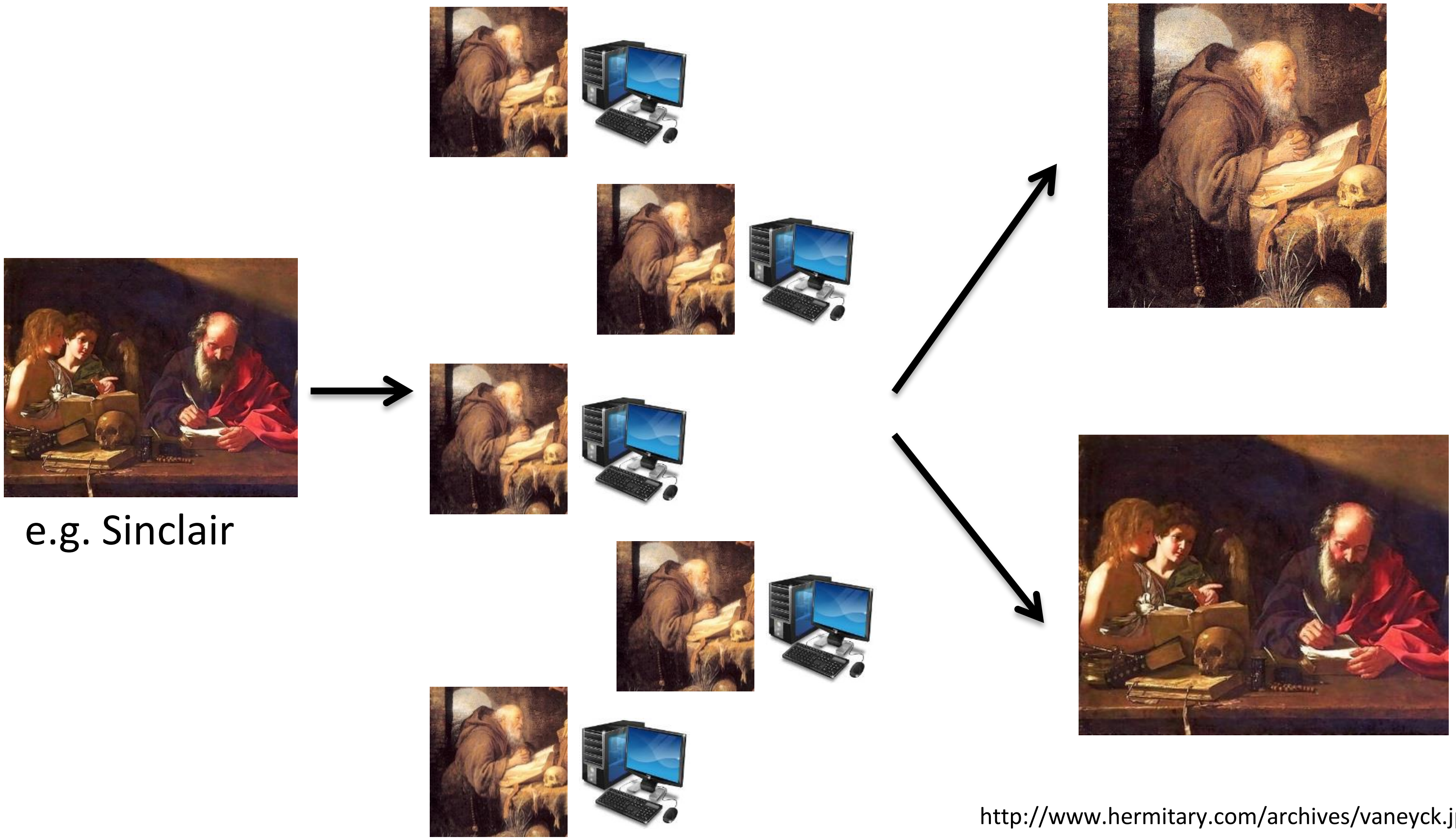
e.g. xyz

http://www.hermitary.com/archves/vaneyck.jpg

https://upload.wikimedia.org/wikipedia/commons/d/d6/St.-Jerome-In-His-Study.jpg

# Thoughts on the future of programming and tools in corpus linguistics research:

- From team player to …



http://www.hermitary.com/archives/vaneyck.jpg

https://upload.wikimedia.org/wikipedia/commons/d/d6/St.-Jerome-In-His-Study.jpg

## Conclusions

- Corpus linguistics research is rapidly changing in terms of corpus size, design, and applications
  - Many interesting corpus linguistics problems can only be solved with new and interesting tools
- Many corpus linguists struggle to collect, clean, tag, annotate and analyze their corpus in new and interesting ways
  - Developing a generation of corpus linguistics who understand basic text handling and processing is essential
- Future corpus tools development and research designs can be improved (most rapidly) through researcher interaction within and across disciplines
  - Creating successful project teams will need infrastructure and financial support by institutions, societies, and funding agencies