Corpus Tools Brainstorming Session

What tools do we need for the future?

Laurence ANTHONY
Professor, Center for English Language Education,
Waseda University, Japan
Honorary Research Fellow (2013-2015), Lancaster University

anthony@waseda.jp
http://www.antlab.sci.waseda.ac.jp/software/

AACL 2014, Northern Arizona University, Flagstaff, US
September 25-28, 2014

Overview of this brainstorming session

- Background
  - Reasons for organizing this session
  - Possible outputs from this session
  - The current state of corpus linguistics tools
  - The case for programming your own corpus tools
- Brainstorming:
  - Part 1: What do you currently do with corpora
  - Part 2: What do you want to do with corpora (that you cannot already)?
  - Part 3: What prevents you collaborating with a tools developer to create new tools… or programming your own?
  - Part 4: How can we increase corpus linguists' attention to corpus tools?

Background:

Reasons for organizing this session

Background:

Reasons for organizing this session

- To highlight the importance of tools in our field
  - Many corpus linguists forget that they look through the lens of a corpus tool when doing their research
- To introduce and discuss the current solutions to corpus linguists' needs
  - Many corpus linguists are not aware of currently available tools
- To generate ideas for new (and useful) corpus tools
  - What functions of a corpus tool are important to you?
  - What would you like corpus tools to do?
- But…
  - This is not a session to voice criticisms of current tools
    - (Unless the criticism is of *AntConc*!)

Background:

Possible outputs from this session

- The development of new corpus tools
  - e.g. realization of ideas generated in this session by current developers
  - e.g. collaborative tools development between developers and session participants
  - e.g. creation of new tools developers with an interest in programming
- A regular corpus tools section in corpus linguistics conferences
- More research presentations/publications on corpus tool innovations

Background:

The current state of corpus linguistics tools

Background:
The 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
    - a representative sample of target language stored as an electronic database
  - 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

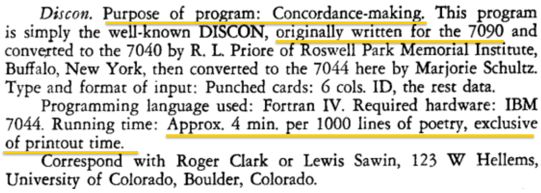
(Biber et al., 1998)

Background:
The current state of corpus linguistics tools

- Four Generations of Corpus Tools
  (see McEnery & Hardie, 2012)
  - 1st-generation (1960s-1970s)
    - run on mainframes, ASCII-based, very limited functions
      - 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)

Background:
The current state of corpus linguistics tools

- *Discon (Clark, 1966)*



Discon. Purpose of program: Concordance-making. This program is simply the well-known DISCON, originally written for the 7090 and converted to the 7040 by R. L. Priore of Roswell Park Memorial Institute, Buffalo, New York, then converted to the 7044 here by Marjorie Schultz. Type and format of input: Punched cards: 6 cols. ID, the rest data.

Programming language used: Fortran IV. Required hardware: IBM 7044. Running time: Approx. 4 min. per 1000 lines of poetry, exclusive of printout time.

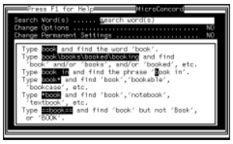
Correspond with Roger Clark or Lewis Sawin, 123 W Hellems, University of Colorado, Boulder, Colorado.

Computers and the Humanities (1966, Vol. 1, Issue 2, p. 39)

Background:
The current state of corpus linguistics tools

Background:
The current state of corpus linguistics tools

- Four Generations of Corpus Tools
  - 2nd-generation (1980s-1990s)
    - run on PCs, ASCII-based, limited functions, scalability problems
      - 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)

Background:
The current state of corpus linguistics tools

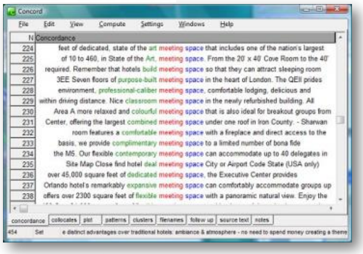
Tim Johns and Randolph Quirk (1982/1983)
Photo by John Higgins.

Background:
The current state of corpus linguistics tools

- Four Generations of Corpus Tools
  - 3rd-generation (2000s-present)
    - more functions, better statistics, improved scalability, multi-language support , more user-friendly, simple, flexible
      - e.g., *WordSmith Tools* (Scott, 1996-2014)
      - e.g., *MonoConc Pro* (Barlow, 2000)
      - e.g., *AntConc* (Anthony, 2004-2014)
    - Problems:
      - limited functionality (still)
      - limited access to corpora containing copyrighted data
      - limited ability to scale to massive (100 million+) corpora

Background:
The current state of corpus linguistics tools

WordSmith Tools

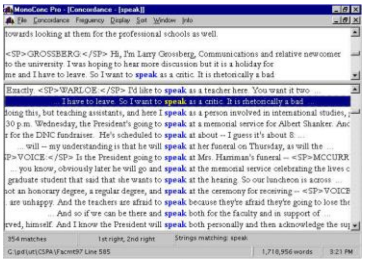

Concord
File Edit View Compute Settings Windows Help
concordance collocates plot patterns clusters filenames follow up source text notes

```
  N Concordance
224       feet of dedicated, state of the art meeting space that includes one of the nation's largest
225        of 10 to 460, in State of the Art, meeting space. From the 20' x 40' Cove Room to the 40'
226    required. Remember that hotels build meeting space so that they can attract sleeping room
227        3EE Seven floors of purpose-built meeting space in the heart of London. The QEII prides
228       environment, professional-caliber meeting space, comfortable lodging, delicious and
229  within driving distance. Nice classroom meeting space in the newly refurbished building. All
230      Area A more relaxed and colourful meeting space that is also ideal for breakout groups from
231   Center, offering the largest combined meeting space under one roof in Iron County. - Sharwan
232            room features a comfortable meeting space with a fireplace and direct access to the
233       basis, we provide complimentary meeting space to a limited number of bona fide
234      the M5. Our flexible contemporary meeting space can accommodate up to 40 delegates in
235          Site Map Close find hotel deal meeting space City or Airport Code State (USA only)
236     over 45,000 square feet of dedicated meeting space, the Executive Center provides
237   Orlando hotel's remarkably expansive meeting space can comfortably accommodate groups up
238    offers over 2300 square feet of flexible meeting space with a panoramic natural view. Enjoy the
454   Set    e distinct advantages over traditional hotels: ambiance & atmosphere - no need to spend money creating a theme
```

WordSmith Tools (Scott, M. , 2014)

Background:
The current state of corpus linguistics tools

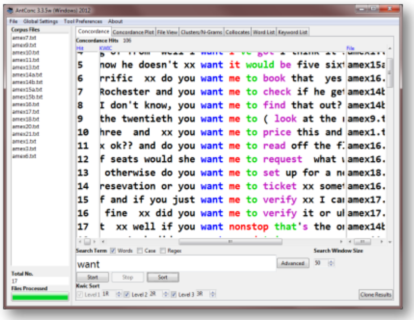
athelstan


MonoConc Pro - [Concordance - [speak]]
File Concordance Frequency Display Sort Window Info

```
towards looking at them for the professional schools as well.

<SP>GROSSBERG:</SP> Hi, I'm Larry Grossberg, Communications and relative newcomer
to the university. I was hoping to hear more discussion but it is a holiday for
me and I have to leave. So I want to speak as a critic. It is rhetorically a bad

Exactly. <SP>WARLOE:</SP> I'd like to speak as a teacher here. You want it two ...
        ... I have to leave. So I want to speak as a critic. It is rhetorically a bad ...
doing this, but teaching assistants, and here I speak as a person involved in international studies, ;
30 p.m. Wednesday, the President's going to speak at a memorial service for Albert Shanker. Anc
r for the DNC fundraiser.  He's scheduled to speak at about -- I guess it's about 8 ...
   ... will -- my understanding is that he will speak at her funeral on Thursday, as will the ...
P>VOICE:</SP> Is the President going to speak at Mrs. Harriman's funeral -- <SP>MCCURR
 ... you know, obviously later he will go and speak at the memorial service celebrating the lives c
 graduate student that said that she wants to speak at the hearing. So our luncheon is across ...
hot an honorary degree, a regular degree, and speak at the ceremony for receiving -- <SP>VOICE
. are unhappy. And the teachers are afraid to speak because they're afraid they're going to lose the
        ... And so if we can be there and speak both for the faculty and in support of ...
rved, himself. And I know the President will speak both personally and then acknowledge the su
354 matches        1st right, 2nd right        Strings matching: speak
G:\pd\ut\CSPA\Facmt97 Line 585                           1,718,956 words    3:21 PM
```

MonoConc Pro (Barlow, M. , 2000)

Background:
The current state of corpus linguistics tools


AntConc 3.3.5w (Windows) 2012
File Global Settings Tool Preferences About
Corpus Files
Concordance Concordance Plot File View Clusters/N-Grams Collocates Word List Keyword List
Hit KWIC File
Start Stop Sort
Advanced
Clone Results

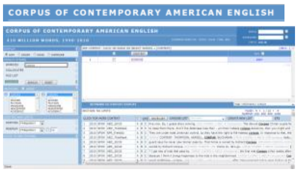
```
Corpus Files   Concordance Hits  106
amex7.txt      5   now he doesn't xx want it would be five sixt amex15a
amex9.txt      6   rrific  xx do you want me to book that  yes  amex16.
amex10.txt     7   Rochester and you want me to check if he get amex14b
amex11.txt     8   I don't know, you want me to find that out?  amex14b
amex13.txt     9   the twentieth you want me to ( look at the   amex9.t
amex14a.txt    10  hree  and  xx you want me to price this and  amex1.t
amex14b.txt    11  x ok?? and do you want me to read off the f  amex16.
amex15a.txt    12  f seats would she want me to request  what   amex16.
amex15b.txt    13   otherwise do you want me to set up for a n  amex18.
amex16.txt     14  resevation or you want me to ticket xx somet amex16.
amex17.txt     15  f and if you just want me to verify xx I ca  amex17.
amex18.txt     16   fine  xx did you want me to verify it or u  amex17.
amex20.txt     17  t  xx well if you want nonstop that's the o  amex14b
amex21.txt
amex1.txt
amex3.txt
amex6.txt

Total No.      Search Term ☑ Words ☐ Case ☐ Regex      Search Window Size
17             want                                    50
Files Processed
               Kwic Sort
               ☑ Level 1 1R  ☑ Level 2 2R  ☑ Level 3 3R
```

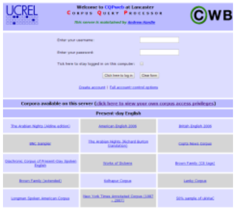
AntConc (Anthony, L., 2014)

Background:
The current state of corpus linguistics tools

- Four Generations of Corpus Tools
  - 4th-generation (late 2000s-present)
    - better scalability, access to copyrighted data, available anywhere or everywhere (via browsers)
      - e.g., corpus.byu.edu (Davies, 2011), *CQPweb* (Hardie, 2011), *SketchEngine* (Kilgariff, 2011), *Wmatrix* (Rayson, 2011)
    - Problems:
      - limited functionality (still)
      - overkill for many purposes (e.g., analyzing small corpora)
      - logistics: registration, setup, licenses, payment
      - limited access to public-domain corpora
      - no access to "personalized" corpora
        - e.g., institution-owned collections of copyrighted material
        - e.g., user-created collections of copyrighted material
      - explosion in one-off, web-based, single-corpus interfaces

Background:
The current state of corpus linguistics tools


CORPUS OF CONTEMPORARY AMERICAN ENGLISH
CORPUS OF CONTEMPORARY AMERICAN ENGLISH
410 MILLION WORDS, 1990-2010



Welcome to CQPweb at Lancaster
Corpus Query Processor

Enter your username:

Enter your password:

Tick here to stay logged in on this computer: ☐

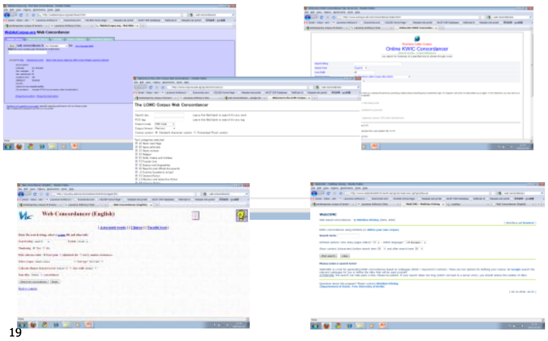
Create account | Full account control options

Corpora available on this server (click here to view your own corpus access privileges)

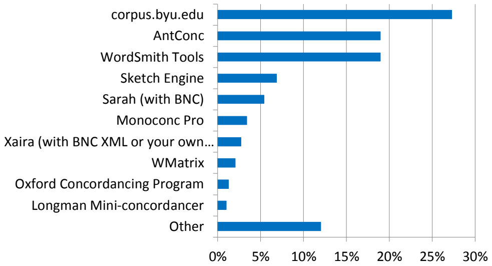
Present-day English

The Arabian Nights (Aldine edition)	American English 2006	British English 2006
BNC Sampler	The Arabian Nights (Richard Burton translation)	[illegible]
Diachronic Corpus of Present-Day Spoken English	Works of Dickens	Brown Family (C8 tags)
Brown Family (extended)	Kolhapur Corpus	Lanka Corpus
Longman Spoken American Corpus	New York Times Annotated Corpus (1987 - 2007)	50% sample of ukWaC

Background:
The current state of corpus linguistics tools



Background:
The current state of corpus linguistics tools

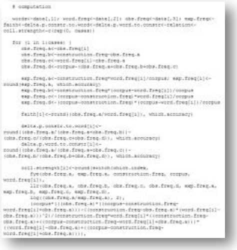


"Which computer programs do you use for analysing corpora?"
International survey of corpus linguists. Reponses: 891. (Tribble, 2012)

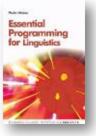
Background:
The case for programming your own corpus tools

Background:
The case for programming your own corpus tools

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



```
# 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)))),
```



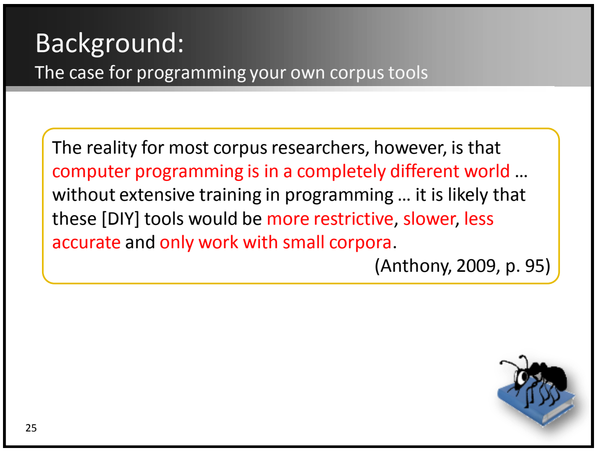
Background:
The case for programming your own corpus tools

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)

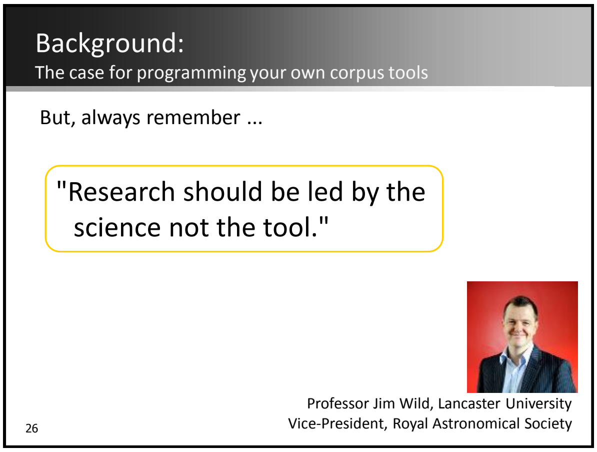
Background:
The case for programming your own corpus tools

> "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)



Background:
The case for programming your own corpus tools

> 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)

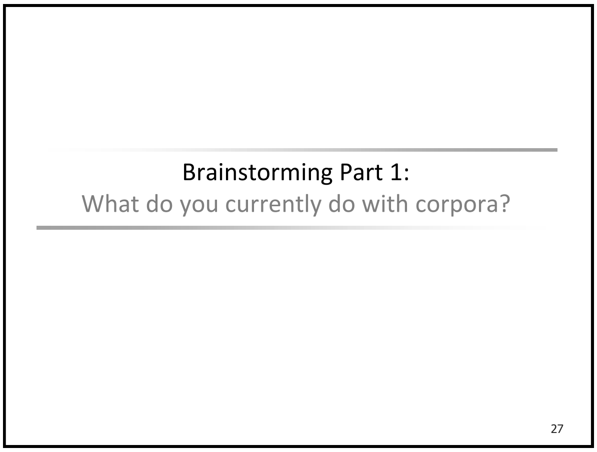


Background:
The case for programming your own corpus tools

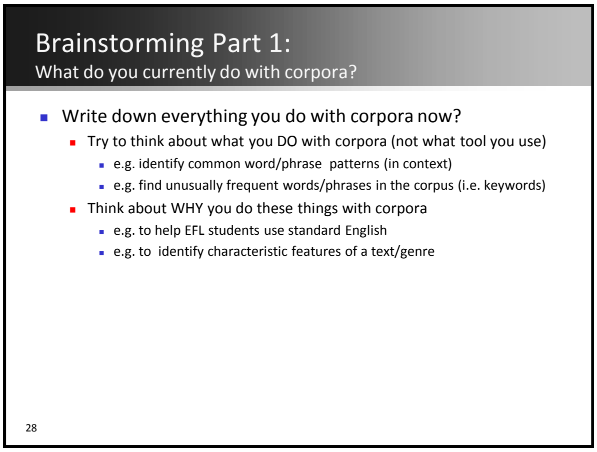
But, always remember …

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

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

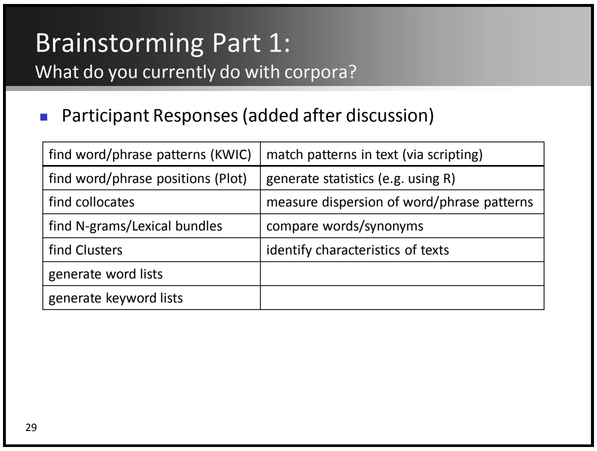


Brainstorming Part 1:
What do you currently do with corpora?



Brainstorming Part 1:
What do you currently do with corpora?

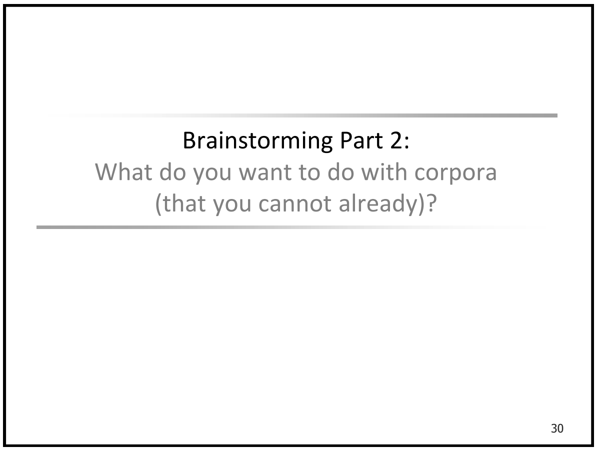
- Write down everything you do with corpora now?
  - Try to think about what you DO with corpora (not what tool you use)
    - e.g. identify common word/phrase patterns (in context)
    - e.g. find unusually frequent words/phrases in the corpus (i.e. keywords)
  - Think about WHY you do these things with corpora
    - e.g. to help EFL students use standard English
    - e.g. to identify characteristic features of a text/genre



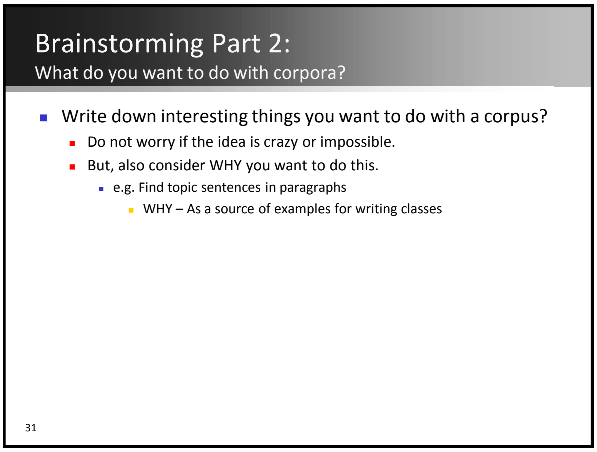
Brainstorming Part 1:
What do you currently do with corpora?

- Participant Responses (added after discussion)

find word/phrase patterns (KWIC)	match patterns in text (via scripting)
find word/phrase positions (Plot)	generate statistics (e.g. using R)
find collocates	measure dispersion of word/phrase patterns
find N-grams/Lexical bundles	compare words/synonyms
find Clusters	identify characteristics of texts
generate word lists	
generate keyword lists	



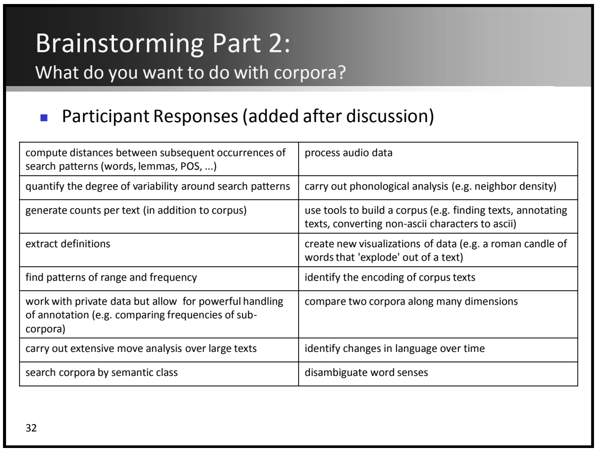
Brainstorming Part 2:
What do you want to do with corpora (that you cannot already)?



Brainstorming Part 2:
What do you want to do with corpora?

- Write down interesting things you want to do with a corpus?
  - Do not worry if the idea is crazy or impossible.
  - But, also consider WHY you want to do this.
    - e.g. Find topic sentences in paragraphs
      - WHY – As a source of examples for writing classes

31

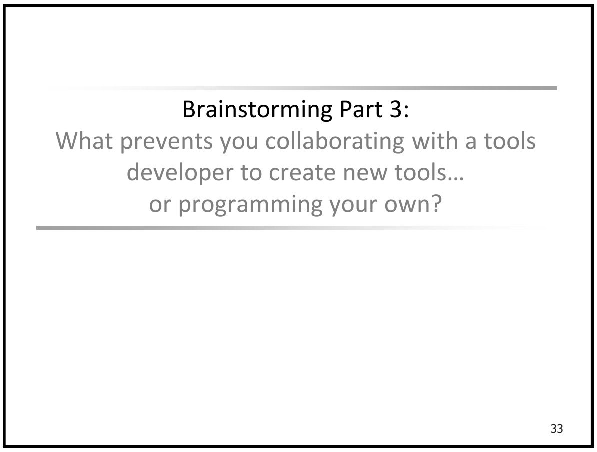


Brainstorming Part 2:
What do you want to do with corpora?

- Participant Responses (added after discussion)

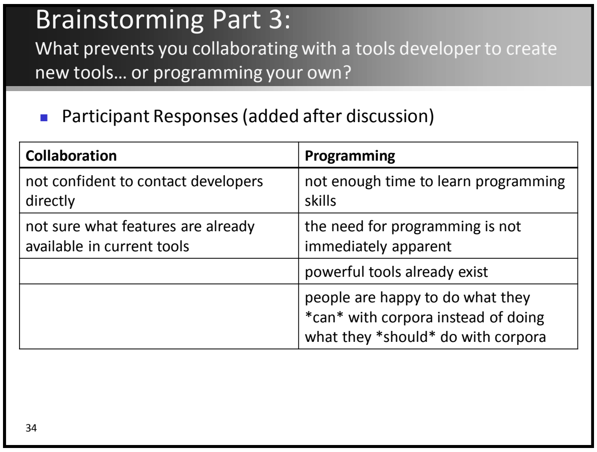
compute distances between subsequent occurrences of search patterns (words, lemmas, POS, ...)	process audio data
quantify the degree of variability around search patterns	carry out phonological analysis (e.g. neighbor density)
generate counts per text (in addition to corpus)	use tools to build a corpus (e.g. finding texts, annotating texts, converting non-ascii characters to ascii)
extract definitions	create new visualizations of data (e.g. a roman candle of words that 'explode' out of a text)
find patterns of range and frequency	identify the encoding of corpus texts
work with private data but allow for powerful handling of annotation (e.g. comparing frequencies of sub-corpora)	compare two corpora along many dimensions
carry out extensive move analysis over large texts	identify changes in language over time
search corpora by semantic class	disambiguate word senses

32



Brainstorming Part 3:
What prevents you collaborating with a tools developer to create new tools… or programming your own?

33

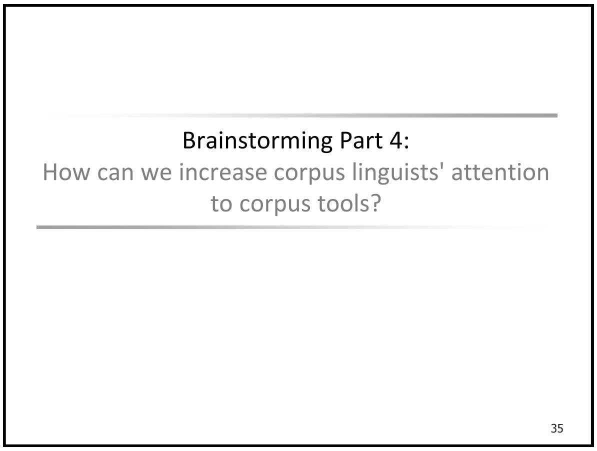


Brainstorming Part 3:
What prevents you collaborating with a tools developer to create new tools… or programming your own?

- Participant Responses (added after discussion)

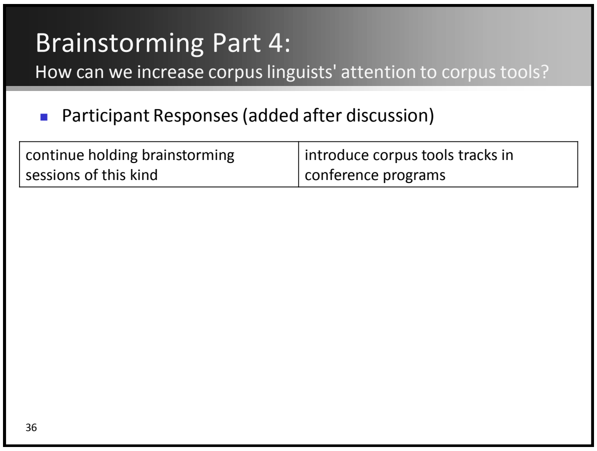
Collaboration	Programming
not confident to contact developers directly	not enough time to learn programming skills
not sure what features are already available in current tools	the need for programming is not immediately apparent
	powerful tools already exist
	people are happy to do what they *can* with corpora instead of doing what they *should* do with corpora

34



Brainstorming Part 4:
How can we increase corpus linguists' attention to corpus tools?

35



Brainstorming Part 4:
How can we increase corpus linguists' attention to corpus tools?

- Participant Responses (added after discussion)

continue holding brainstorming sessions of this kind	introduce corpus tools tracks in conference programs

36