

# Corpus-based computational linguistics

Chris Brew and Michael White  
14-19 July 2008  
OSU Mini-Institute





# Outline

- What are corpora?
- What are corpora for?
- How are they used?
- How could they be used?

# Empiricists

---

Nelson Francis and Henry Kucera

Mitch  
Marcus

Samuel Johnson

---

# Empiricists

Nelson Francis and Henry Kucera



Samuel Johnson

Mitch  
Marcus

# Empiricists

Nelson Francis and Henry Kucera



Samuel Johnson



Mitch  
Marcus

# Empiricists

Nelson Francis and Henry Kucera



Samuel Johnson



Mitch  
Marcus



# What are corpora?

corpus |'kôrpəs|

noun ( pl. **-pora** | -pərə | or **-puses** )

**1** a collection of written texts, esp. the entire works of a particular author or a body of writing on a particular subject : *the Darwinian corpus*.

- a collection of written or spoken material in machine-readable form, assembled for the purpose of studying linguistic structures, frequencies, etc.

**2** Anatomy the main body or mass of a structure.

- the central part of the stomach, between the fundus and the antrum.

**ORIGIN** late Middle English (denoting a human or animal body): from Latin, literally 'body.' Sense 1 dates from the early 18th cent.



# Pre-history

- Johnson's dictionary established norms of lexicography that still prevail.
  - Collected illustrative sentences on slips of paper.
  - Made a numbered list of meanings, starting with the one he called the "natural and primitive signification", going on to others that are less central.
  - Tried to use good modern writers. Writing in 1755, he aimed for later than 1558 (start of Elizabeth I's reign)



# das Kädingsche Ur- Korpus

- Kading J. (1879) *Häufigkeitswörterbuch der deutschen Sprache*, and subsequent consultancy work
  - Motivation: court stenography. Work out which keys should lie under which fingers
  - Method: systematic count of letter-letter pairs by 5000 helpers.
  - eventually got to 100 million words.

# Survey of English Usage: Quirk



<http://www.ucl.ac.uk/english-usage/about/history.htm>

“The million-word Survey Corpus, now complete, samples written and spoken British English produced between c.1955 and 1985. It comprises 200 texts, each of 5,000 words. The spoken texts include both dialogue and monologue, while the written texts include not only printed and manuscript material but also examples of English read aloud, as in broadcast news and scripted speeches.”



# What are corpora?

- A waste of time.

In 1962, when I was in the early stages of collecting the Brown Standard Corpus of American English, I met Professor Robert Lees at a linguistic conference. In response to his query about my current interests, I said that I had a grant from the U.S. Office of Education to compile a million-word corpus of present-day American English for computer use. He looked at me in amazement and asked, “Why in the world are you doing that?” I said something about finding out the true facts about English grammar. I have never forgotten his reply: “That is a complete waste of your time and the government’s money. You are a native speaker of English; in ten minutes you can produce more illustrations of any point in English grammar than you will find in many millions of words of random text.”

**(Francis, 1979:110)**



# What are corpora?

- Not necessarily a waste of time.
  - No native speakers of 14<sup>th</sup> C English available. So no option but to work from the data.
  - Standardized corpora allow reproducible quantitative experiments, and comparisons between different approaches.
  - Native speakers lack usable intuitions about many things which turn out to be useful.  
[Example: die: <http://corpus.byu.edu/bnc/x.asp>]

**BYU-BNC: BRITISH NATIONAL CORPUS** (100 MILLION WORDS, 1980s-1993)  
 Mark Davies / Brigham Young University

D SEE CONTEXT: CLICK ON WORD OR SELECT WORDS + [CONTEXT] [\[HELP...\]](#)

	<input type="checkbox"/>	<b>CONTEXT</b>	TOT <input type="checkbox"/>
1	<input type="checkbox"/>	KICK THE BUCKET	7

SE 0.578

- SE
- Wt >
- CC >
- PC >
- US >
- Sl

1. **KEYWORD IN CONTEXT (KWIC)** More information...

PART OF SPEECH: NO LIMITS (HELP) SAMPLE: 100 ENTRIES  
 SECTION: NO LIMITS

- <
- <
- <
- <
- CL
- SEt
- OP

CLICK ON TITLE FOR MORE CONTEXT

1	HE0	S_lect_soc_science	no genuine semantic structure from which you can determine its meaning, for example <b><u>kick the bucket</u></b> means die and you don't get that in the meaning of kick t
2	HE0	S_lect_soc_science	means die and you don't get that in the meaning of <b><u>kick the bucket</u></b> . but notice kick the bucket appears as a verb phrase and eat humble pie
3	HE0	S_lect_soc_science	get that in the meaning of kick the bucket. but notice <b><u>kick the bucket</u></b> appears as a verb phrase and eat humble pie, get your knickers in a
4	HE0	S_lect_soc_science	is the meaning of the whole phrase. So like I said <b><u>kick the bucket</u></b> , the meaning of that idiomatically is just die, sorry die. It's
5	HE0	S_lect_soc_science	just you know. So, although in all these three, <b><u>kick the bucket</u></b> , eat humble pie, get your knickers in a twist er all look like
6	FAC	W_ac_soc_science	leg, to have a bee in one's bonnet, to <b><u>kick the bucket</u></b> , to cook someone's goose, to be off one's rocker, round
7	FAC	W_ac_soc_science	. Thus, The aspidistra kicked the bucket exemplifies inappropriateness because replacing <b><u>kick the bucket</u></b> with its cognitive synonym die removes the dissonance.



**BYU-BNC: BRITISH NATIONAL CORPUS** (100 MILLION WORDS, 1980s-1993)  
 Mark Davies / Brigham Young University

D SEE CONTEXT: CLICK ON WORD OR SELECT WORDS + [CONTEXT] [\[HELP...\]](#)

	<input type="checkbox"/>	CONTEXT	TOT <input type="checkbox"/>
1	<input type="checkbox"/>	SNUFFED IT	9
2	<input type="checkbox"/>	SNUFF IT	8
3	<input type="checkbox"/>	SNUFFING IT	3
4	<input type="checkbox"/>	SNUFFS IT	2
		TOTAL	22

CC > 0.375

PC >

US >

SI

1.

KEYWORD IN CONTEXT (KWIC) More information...

PART OF SPEECH: NO LIMITS (HELP)  
 SECTION: NO LIMITS

SAMPLE: 100 ENTRIES

CLICK ON TITLE FOR MORE CONTEXT

1	C85	W_fict_prose	work in a place like this. A small flame of hope lit. She <b>snuffed it</b> out. But it was to burn again in the candlelit drawing-room where the
2	CJT	W_fict_prose	Holly, it's the most extraordinary thing but the chap's dead, <b>snuffed it</b> . He had the best medical treatment --; well, you'll not be
3	EA5	W_fict_prose	'er. I ain't seen'er about fer ages. P'raps she's <b>snuffed it</b> ." Fred cut into the pieces of meat with a vengeance, fighting
4	EA5	W_fict_prose	to shake the chattering Bessie Chandler by the scruff of her neck until she <b>snuffed it</b> . "P'raps she'as," he replied quietly. Bessie was not
5	FB9	W_fict_prose	of a fellow for crystallized fruits and waiters. Well, when Paddy Pottleton <b>snuffed it</b> , I still got the rag, entirely for your articles. I ca
6	HA0	W_fict_prose	Lord Jesus had a better idea. He knew nothing dies. Even when he <b>snuffed it</b> on Mount Cavalry, he knew he would live again." "Easy
7	HR9	W_fict_prose	, I'd only been in the hospital for about ten minutes before I <b>snuffed it</b> !" It must have been weird lying there. Presumably trying to move
8	F9W	W_ac_soc_science	has taken her away from us. B: You mean the old girl's <b>snuffed it</b> . (2.4) Referring expressions These are words whose meaning can only be
9	EUU	W_commerce	meet his maker. He's propping up the daisies. He's fucking <b>snuffed it</b> ." In actuality, in this particular case, the widow of the deceased

CL  
SEI  
OP

**BYU-BNC: BRITISH NATIONAL CORPUS** (100 MILLION WORDS, 1980s-1993)  
 Mark Davies / Brigham Young University

SEE CONTEXT: CLICK ON WORD OR SELECT WORDS + [CONTEXT] [\[HELP...\]](#)

	<input type="checkbox"/>	<b>CONTEXT</b>	TOT <input type="checkbox"/>
1	<input type="checkbox"/>	A STIFF	6

0.250

- SE
- Wt >
- CC >
- PC >
- US >
- S

1. **KEYWORD IN CONTEXT (KWIC)** More information...

PART OF SPEECH: [N%]  
 SECTION: NO LIMITS

SAMPLE: 100 ENTRIES

CLICK ON TITLE FOR MORE CONTEXT

1	CK0	W_fict_prose	, afore any of your lot got here. But for her he'd be <b>a stiff</b> this very minute, and I'm not telling you no lies! Proper bleeding
2	FAP	W_fict_prose	there you was stretched out on the floor. We thought you was <b>a stiff</b> ." "Yeah, that's right," growled the Battler. I
3	FAP	W_fict_prose	Regal Arms" deal. What have they got? Two guys carting <b>a stiff</b> . They shoot at a prowl-car, and the prowl boys shoot straighter. Two
4	H9N	W_fict_prose	piratical chauffeur, while he shot back and forth under my feet like <b>a stiff</b> in a mortuary drawer. The whistling had a peculiarly aggravating quality. When I
5	J13	W_fict_prose	thousand, and I wouldn't tell the law that you'd left <b>a stiff</b> behind the Windsor." "Fuck all that. I'm talking about this
6	CBL	W_misc	are light green, feather-like, arranged in whorls of 4-6, on <b>a stiff</b> , round thin stem. When growing emersed, these leaves are blue-green or emerald

- CL
- SEI
- OP



# What is a corpus?



# What is a corpus?

- Any body of text is a corpus if we can think of a way of doing linguistic things with it.



# What is a corpus?

- Any body of text is a corpus if we can think of a way of doing linguistic things with it.
- OK, so what kind of things then?



# Machine readable corpora

- Francis and Kucera's Brown Corpus
  - 500 texts of American English, chosen from 15 text categories. Total of 1,161,192 "words", 57,340 "sentences".
  - 56,057 different "words" [alphabetically last 10 are: zoology zoomed zooming zooms zooop zorrillas zounds zu zur {0,T} ]

# Natural Language Toolkit

...software, data sets and tutorials for natural language processing...

search this site




GO

SEARCH

## navigation

- ▶ [Main Page](#)
- ▶ [Book](#)
- ▶ [Code](#)
- ▶ [Contribute](#)
- ▶ [Corpora](#)
- ▶ [Courses](#)
- ▶ [Development](#)
- ▶ [Documentation](#)
- ▶ [download](#)
- ▶ [FAQ](#)
- ▶ [Getting Started](#)
- ▶ [News Archive](#)
- ▶ [People](#)
- ▶ [Projects](#)
- ▶ [Quotes](#)
- ▶ [Screenshots](#)
- ▶ [Teaching](#)
- ▶ [Video](#)

## toolbox

- ▶ [What links here](#)
- ▶ [Related changes](#)
- ▶ [Upload file](#)
- ▶ [Special pages](#)
- ▶ [Printable version](#)
- ▶ [Permanent link](#)

[article](#)

[discussion](#)

[view source](#)

[history](#)

## Main Page

**NLTK — the Natural Language Toolkit — is a suite of open source Python modules, data and documentation for research and development in natural language processing. NLTK contains [Code](#) supporting dozens of NLP tasks, along with 40 popular [Corpora](#) and extensive [Documentation](#) including a 375-page online [Book](#). Distributions for Windows, Mac OSX and Linux are available.**

- ▶ [News](#) - NLTK presented at ACL conference [June 2008]; Version 0.9.3 Released [June 2008]; NLTK at LinuxFest Northwest [April 2008]; NLTK in Google Summer of Code [April 2008]; Python Software Foundation adopts NLTK for Google Summer of Code application [March 2008]; NLTK 0.9.2 Released [March 2008]; NLTK video posted on YouTube [January 2008]
- ▶ [Quotes](#) - what users have said about NLTK
- ▶ [Documentation](#) - book, articles, guides, reviews, API documentation
- ▶ [Donate](#) - support ongoing NLTK development

### Getting Started

- ▶ [Download](#) - instructions for downloading and installing Python and NLTK on all platforms
- ▶ [Getting Started](#) - simple things to try, including NLTK's demonstrations
- ▶ [Mailing List](#) - subscribe to important announcements about NLTK by email
- ▶ [Powerpoint](#) - overview and screenshots

### Getting Help

- ▶ [FAQ](#) - answers to frequently asked questions
- ▶ [Screenshots](#) - some graphical and textual demonstrations
- ▶ [User forum](#) - mailing list for discussion amongst NLTK users



# nlTK usage example

```
>>> [random.choice(lbrs) for i in  
range(10)]
```

```
['Acala', 'Grumble', 'Stardel',  
'trimmed', 'Ahm', 'Aerobacter',  
'pets', 'Archbishop', 'Iraqw',  
'pre-literate']
```

# Most frequent tags

NN	152470
IN	120557
AT	97959
JJ	64028
.	60638
,	58156
NNS	55110
CC	37718
RB	36464
NP	34476
VB	33693
VBN	29186
VBD	26167
CS	22143
PPS	18253
VBG	17893
PP\$	16872
TO	14918
PPSS	13802
CD	13510