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

Samuel Johnson

Nelson Francis and Henry Kucera

Mitch Marcus
Empiricists

Samuel Johnson

Nelson Francis and Henry Kucera

Mitch Marcus
What are corpora?

corpus |ˈkərpəs|
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.
• 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)
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.
“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\textsuperscript{th} 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]
1. no genuine semantic structure from which you can determine its meaning, for example *kick the bucket* means die and you don't get that in the meaning of *kick the bucket*. but notice *kick the bucket* appears as a verb phrase and eat humble pie get your knickers in a twist.

2. means die and you don't get that in the meaning of *kick the bucket*. but notice *kick the bucket* appears as a verb phrase and eat humble pie, get your knickers in a twist.

3. is the meaning of the whole phrase. So like I said *kick the bucket*, the meaning of that idiomatically is just die, sorry die. It's

4. just you know. So, although in all these three, *kick the bucket*, eat humble pie, get your knickers in a twist or all look like

5. leg, to have a bee in one's bonnet, to *kick the bucket*, to cook someone's goose, to be off one's rocker, round

6. Thus, The aspidistra kicked the bucket exemplifies inappropriateness because replacing *kick the bucket* with its cognitive synonym die removes the dissonance.
<table>
<thead>
<tr>
<th>PART OF SPEECH</th>
<th>NO LIMITS (HELP)</th>
<th>SECTION</th>
<th>NO LIMITS</th>
<th>SAMPLE: 100 ENTRIES</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 C85 W_fict_prose</td>
<td>work in a place like this. A small flame of hope lit. She <strong>snuffed it</strong> out. But it was to burn again in the candlelit drawing-room where the</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 CT W_fict_prose</td>
<td>Holly, it's the most extraordinary thing but the chap's dead, <strong>snuffed it</strong>. He had the best medical treatment --; well, you'll not be</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 EA5 W_fict_prose</td>
<td>'er. I ain't seen' er about fer ages. P'praps she's <strong>snuffed it</strong>. 'Praps she'as,' he replied quietly. Bessie was not</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 EA5 W_fict_prose</td>
<td>to shake the chattering Bessie Chandler by the scruff of her neck until she <strong>snuffed it</strong>. 'Praps she'as,' he replied quietly. Bessie was not</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 FB9 W_fict_prose</td>
<td>of a fellow for crystallized fruits and waiters. Well, when Paddy Pottleton <strong>snuffed it</strong>, I still got the rag, entirely for your articles. I ca</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 HA0 W_fict_prose</td>
<td>Lord Jesus had a better idea. He knew nothing dies. Even when he <strong>snuffed it</strong> on Mount Cavalry, he knew he would live again.'</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 HR9 W_fict_prose</td>
<td>, I'd only been in the hospital for about ten minutes before I <strong>snuffed it</strong>. 'It must have been weird lying there. Presumably trying to move</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 F9W W_ac_soc_science</td>
<td>has taken her away from us. B: You mean the old girl's <strong>snuffed it</strong>. (2.4) Referring expressions These are words whose meaning can only be</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9 EUU W_commerce</td>
<td>meet his maker. He's propping up the daisies. He's fucking <strong>snuffed it</strong>. 'In actuality, in this particular case, the widow of the deceased</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**BYU-BNC: BRITISH NATIONAL CORPUS**

(100 MILLION WORDS, 1980s-1993)

Mark Davies / Brigham Young University
### BYU-BNC: BRITISH NATIONAL CORPUS
Mark Davies / Brigham Young University

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

#### KEYWORD IN CONTEXT (KWIC)

<table>
<thead>
<tr>
<th>ID</th>
<th>CC</th>
<th>SE</th>
<th>KEYWORD</th>
<th>CONTEXT</th>
<th>TOT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>a stiff</td>
<td>A STIFF</td>
<td>6</td>
</tr>
</tbody>
</table>

**MORE INFORMATION...**

**SAMPLE:** 100 ENTRIES

1. **CLICK ON TITLE FOR MORE CONTEXT**

   1. CKU  W_fict_prose  , afore any of your lot got here. But for her he'd be a **stiff** 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 a **stiff**. "Yeah, that's right," growled the Battler. I
   3. FAP  W_fict_prose  Regal Arms' deal. What have they got? Two guys carting a **stiff**. 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 a **stiff** 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 a **stiff** 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 a **stiff**, round thin stem. When growing emersed, these leaves are blue-green or emerald
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.
• 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?
• 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 zoop zorrillas zounds zu zur \{0,T\} ]
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
>>> [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