Corpus methods for sociolinguistics

Emily M. Bender
bender@csli.stanford.edu
NWAV 31 - October 10, 2002

Introduction

Overview

- Introduction
- Corpora of interest
- Software for accessing and analyzing corpora (demo)
- Basic programming tools
- Creating & publishing corpora

Sociolinguistics IS *corpus linguistics*

- Study naturally occurring data
- . . . in context
- . . . including frequency of (co)-occurrence

Goals

- What kinds of resources are out there
- How to learn more about those resources
- How to find more resources
- Encourage you to create & publish corpora

Rules of thumb

If it's tedious, a computer could probably do it for you.

If you'll be doing much more of it, or doing it again later, it's probably worth figuring out how to get a computer to do it for you.

The only URL you need to know

`http://www-csli.stanford.edu/~bender/corpora_sociolx.shtml`

Corpora of Interest

BNC

- 1994
- 100,000,000+ words (90% written, 10% spoken)
- Some coding for age, gender, region, social class, audience, genre
- Available for purchase (£250/network license, £50 single user license) or online subscription (price depending on number of machines it will be used on)
- Some (limited) access is available for free online

ANC

- In progress
- Modeled on the BNC
- Core corpus: 100,000,000 words, similar genre distribution to BNC
- Plus potentially several hundreds of millions more words

BNC

- Supported by bnc-discuss, a mailing list on the use of the BNC
- Marked up with SGML
- Comes with SARA software for easy access

ANC

- First installment (10 million words) this fall
  - preliminary search tools
  - spoken data: LDC Switchboard & CallHome (2 million words)
  - written data: NYT (1.5 million words), ephemera, novels
- Completion in 2004

ICE

- Parallel corpora from 20 sites around the world
- Spoken and written, 1990–1996
- Spoken genres include conversations, classroom lessons, broadcast interviews, legal cross-examination, parliamentary debate
- 1,000,000 words in each corpus

A few others

- LIPPS (TalkBank): Language Interaction in Plurilingual and Plurilectal Speakers (code-switching data)
- CHILDES (TalkBank): Language acquisition data (child and adult, first and second language)

A few others

- Switchboard (LDC): strangers speaking to each other over the telephone on randomly selected topics (speech files & transcripts) (American English)
- CallHome (LDC): telephone conversations between close friends & family members. (speech files & transcripts)
  (American English, Egyptian Arabic, German, Japanese, Mandarin, Spanish)
- CallFriend (LDC): like CallHome, more languages, not (yet?) transcribed

Where to find corpora

- TalkBank
- LDC: Linguistic Data Consortium
- ELRA: European Language Resources Association
- ICAME: International Computer Archive of Modern and Medieval English
- Indices maintained by individuals
- The corpora mailing list

Software

Kinds of useful software

- Preprocessing: taggers, tokenizers, parsers
- Searching
- Coding
- Transcribing

Taggers/Tokenizers

- AMALGAM: pos tagger for English, available over the internet
- ChaSen: tagger, morphological analyzer and tokenizer for Japanese (free download)
- …

Searching: BNCweb

- A beautiful search interface for the BNC (World Edition)
- Links up to SARA
- In principle could be used with other corpora, provided they were formatted & marked up properly
- Available for 30 Euros
- → demo

Searching: TIGERSearch

- A search engine for searching treebanks
- Query language is akin to TFS formalisms
- Available for free

Transcribing: TalkBank tools

- CLAN: An editor for files in CHAT (like CHILDES) or CA (Conversation Analysis) format (free)
- Transana: A tool designed to facilitate transcription and analysis of video data (free)
- Transcriber: A tool for segmenting, labeling, and transcribing speech (free)

Coding: Goldsearch

- Software for creating input file for VARBRUL
- Input:
  - Text file annotated with independent variable values
  - Speaker file indicating variable values for each speaker
- Output:
  - File suitable for VARBRUL input
  - Speaker variables recorded for each token
  - Any other annotations recorded for each token
- Available for free
- $\rightarrow$ demo

Basic Programming Tools

Grep (& other unix commands)

- Generalized regular expression printer
- Useful for pulling examples out of text files
- Regular expression syntax similar to that of emacs, perl
- $\rightarrow$ demo
- $\rightarrow$ web-based tutorial

Perl

- General purpose programming language
- . . . tuned to be useful for manipulating text files
- Interpreted (rather than compiled) language
- Not that hard to learn
- Recommended reading:

  Schwartz, Randal L. 1993. *Learning Perl*. Sebastopol, CA: O'Reilly & Associates.

Creating & Publishing Corpora

Why

- More value for effort
- Comparative studies
- Speech data paired with published ethnographic work particularly interesting
- Video data also interesting

How

- Independently
- Through the LDC
- Through TalkBank (corpora created with TalkBank tools are expected to be contributed to TalkBank)
- → Human subjects considerations

Human Subjects Considerations

- Obtain consent (plan ahead!)
- Preserve anonymity in both speech files and transcripts
- Consult committee for the protection of human subjects at your institution

Conclusion

Goals

- What kinds of resources are out there
- How to learn more about those resources
- How to find more resources
- Encourage you to create & publish corpora