# Introduction to Corpora at Stanford

Anubha Kothari

For Ling 395A Research Methods, 04/18/08

# OUTLINE

- Data
  - What kinds of corpora do we have?
- Access
  - How do we get to them?
- Search
  - How do we find what we need?

- (Thanks to Hal Tily whose presentation was partially copied/adapted for this one.)

# Why Corpora are Useful…

Though messy and requiring much time, care, and technical expertise, corpora are great to work with!

- Large amounts of searchable data consisting of real language use
- Saves you time from collecting data yourself
- Analyze the distribution of some linguistic variant and understand what factors might explain its variation
- Discover new linguistic and non-linguistic association patterns
- Investigate predictions of your theory

# What kinds of Corpora do we have?

- Text and speech in various languages
  - English, German, Spanish, Mandarin, Arabic, French, Japanese, Czech, Korean, Old & Middle English, Russian, Hindi, Tamil, etc.
- Complete inventory at http://www.stanford.edu/dept/linguistics/corpora/inventory.html
  - LDC and non-LDC corpora available
- Also scout around on the web for corpora (e.g. Native American languages), and if they aren't free just ask the corpus TA – it might be possible to acquire them!
- Many possible types of annotations:
  - Syntactic structure, part-of-speech, coreference chains, animacy, information status, pitch accents, word times, speaker gender, dialect, age, education level, etc.
  - You can always take a (partially annotated) corpus and add your own annotations!

# Some Multilingual Corpora

- Parallel texts (translations)
  - Arabic-Mandarin-English news (TDT corpus)
  - English-Arabic parsed newswire (English-Arabic treebank)
  - Cantonese-English literature (Hong Kong hansards)
  - French-English literature (Canadian hansards)
- Phone conversations (CALLHOME)
- Gigawords (English, Chinese, French, Arabic, Spanish)
- Recent acquisitions:
  - Emille/CIIL (lots of South Asian languages)
  - Spoken Dutch corpus

# STANDARD RAW TEXT CORPORA

- Brown corpus
  - 1 million words
  - Written American English (before 1960)
  - Genres balanced to reflect published quantities
- British National Corpus
  - 100 million words
  - Mostly British English (~85%)
  - Mostly written (90%); 10% transcribed naturalistic speech
  - Genres selected for breadth (not necessarily reflecting use)
- English Gigaword
  - 3 million words
  - Mostly American English, some international Englishes
  - Newswire from multiple sources
- Web 1T 5-gram
  - Google's corpus of uni-, bi-, tri-, 4-, and 5-grams and their frequency counts
  - Any English on the web
  - No part-of-speech information
- CELEX
  - English, Dutch, and German
  - Lexical database containing frequency counts, word class info, subcategorization/arg-structure info, orthography variations, phonetic transcriptions, variations in pronunciation, syllable structure, primary stress, derivational and compositional structure, inflectional paradigms)

# STANDARD PARSED CORPORA

- Penn Treebank
  - 1 million word Wall Street Journal (WSJ) section
  - 1 million word Brown corpus section
  - Spoken data from Switchboard

# Standard speech corpora

- Switchboard I
  - Phone conversations between strangers
  - Approx. 3 million words (but many versions exist)
  - Time-aligned and parsed portions exist
  - Switchboard LINK has lots of annotations including animacy, information status, coreference, kontrast, etc.
- CALLHOME/CALLFRIEND
  - Phone conversations between friends/family
  - Arabic, English, German, Canadian French, Japanese, Mandarin, Spanish, Hindi, Tamil, Farsi, Vietnamese
- Santa Barbara Corpus of Spoken American English
  - More language situations (including face-to-face)
  - More diverse speakers
- Fisher
- Boston University Radio Speech Corpus

# Getting to the Data

- To use LDC data or corpora with special access restrictions, first read the instructions at http://www.stanford.edu/dept/linguistics/corpora/access.html.
- To get the data:
  - Either borrow a physical disk from the Chair’s Office
    - Fill out the sign-out sheet when taking or returning a disk
  - Or, (as indicated on the corpus inventory page) use a copy that’s been uploaded to Stanford’s AFS system at /afs/ir/data/linguistic-data/
    - To understand what AFS is all about, see http://www.stanford.edu/services/afs/
    - Basic Unix commands are described at http://www.stanford.edu/services/unix/unixcomm.html

# Working with Corpora on AFS

- Use SSH to connect to a Stanford server (*cardinal, elaine, tree, vine, bramble, hedge*, etc.)
  - Use the terminal on MacOS or Linux (`ssh anubha@vine.stanford.edu`); use Putty on Windows
  - Note:  Our `tgrep2` works only on linux machines (*vine, bramble, hedge*)
  - Use `cd` to get to the right directory
    - `cd /afs/ir/data/linguistic-data/`
  - Use `ls` to see the contents
  - Explore the directory structure
  - Pay attention to the readme file(s)
  - Read documentation using `less` or `more`
  - Will have to save outputs of searches to your personal AFS space

# Searching Corpora for Data you need

- There are many data gathering tools out there!
- Most of the commonly-used tools are listed at http://www.stanford.edu/dept/linguistics/corpora/tools.html
- Some corpora come with search software of their own – just check your favorite corpus' documentation.
- Most commonly used
  - `grep` (for raw text)
  - `tgrep2` (for parsed text)
- Extracting sound samples (along with other annotations)
  - Jason Brenier's `ExtractUnitAcoustics` script
  - Gabe Recchia's STRATA tool
- Also learn some quick command-line tools like `sort`, `uniq`, `awk`, etc.  See "Unix for Poets" at http://people.sslmit.unibo.it/~baroni/compling04/UnixforPoets.pdf.

# GREP

- You can use it to search and count for occurrences of a (text) expression
- A good tutorial – http://www.panix.com/~elflord/unix/grep.html
- Let's search for "gourmet" in ICAME-Brown1
  - Change to the directory which holds your data
    - `cd /afs/ir/data/ling    uistic-data/Brown/ICAME-Brown1`
  - Search for all occurrences of a word in all (*) files in that directory:
    - `grep gourmet *`
  - For search expressions longer than one word, put quotes around the expression
  - Retrieve context if you like with flags
    - Example: `grep -A2 gourmet *` for 2 lines of following context
    - -A*n* provides n lines of following context
    - -B*n* provides n lines of preceding context
    - -C*n* provides n lines of surrounding context

# Retrieving Counts

- Use the `–c` flag to return counts for each file in the directory
- `grep -c gourmet *`
- Use the following to get the total count of “gourmet” across all files in your directory
  - `cat * | grep -c gourmet`
- Use the `man` command with any Unix command to get documentation and a full list of options, flags, etc.  For example:
  - `man grep`

# Regular Expressions

- Use `egrep` rather than `grep` for more complex patterns
- Use . to match any character
  - `egrep “ h.t “ *`
  - Gets you sentences with *hat*, *hit*, *hot*, *h@t*, etc.
- Use `\w` to match any letter or number
- Use `\W` to match any other character
- Use ? to make a character optional
  - `egrep “travell?ed” *`
  - Gets you sentences with *traveled* or *travelled*.
- Use [] to match any one character within
  - `egrep “gr[ae]y” *`
  - Gets you sentences with *gray* or *grey*.
- Use (|) to choose between multiple 0-*n* character long choices in an expression
  - `egrep “jump(ing|ed|s|)\W” *`
  - Gets sentences with *jump*, *jumping*, *jumped*, *jumps*.

# More Regular Expressions

- `a+` matches one or more `a`
- `a*` matches zero or more `a`s
- `a{n}` matches exactly *n* `a`s
- `a{n,}` matches *n* or more `a`s
- `a{n,m}` matches between *n* and *m* `a`s

# Related Useful Commands

- Chaining commands using |
  - Search on the output of a query to restrict your results further – just like “refine search” in a library catalog
  - E.g. to get sentences with both *salt* and *pepper* somewhere in them
    - `grep salt * | grep pepper`
  - Useful with the `–v` switch, which returns all lines which *do not* match
    - `egrep "f[eo]{2}t" * | grep -v football`
- The > character prints the results to a file instead of to the screen (to a file in your home directory because of the ~):
  - `grep gourmet * > ~/results.txt`

# About tgrep2

- Generalizes the concept of `grep` to parsed data with tree structures
  - Allows you to search for particular syntactic-tree configurations involving relations of dominance, sisterhood, precedence, etc. between nodes in a tree.
- Replaces `tgrep`
- Written by Doug Rohde at MIT
- For instructions on how to set up your account to use `tgrep2` on AFS
  - http://www.stanford.edu/dept/linguistics/corpora/cas-tut-tgrep.html
- For good in-depth introductions and tutorials see
  - http://tedlab.mit.edu/~dr/Tgrep2/tgrep2.pdf
  - http://www.bcs.rochester.edu/people/fjaeger/teaching/tutorials/TGrep2/LabSyntax-Tutorial.html
- Look at the Penn Treebank tagset to get an idea of what the node labels mean
  - http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQP-HTMLDemo/PennTreebankTS.html
  - http://www.ling.ohio-state.edu/~hinrichs/course07/ptb.pdf

# More on tgrep2

- As with any corpus search, using `tgrep2` requires a lot of care and iterated refining till you are very sure that you have *all* and *only* the trees of interest to you
  - Start with a sample sentence of interest to you to get an initial `tgrep2` query set up and ensure that the query gets you that sentence
  - Then refine it
    - For example, did you use a node label of correct grain size for your purposes? (e.g. `NP-SBJ` instead of just `NP`)

Good luck and have fun!

# Some Useful Links

- Our corpus website -- http://www.stanford.edu/dept/linguistics/corpora/
- Chris Manning's corpus resources page -- http://www-nlp.stanford.edu/links/statnlp.html

- British National Corpus -- http://sara.natcorp.ox.ac.uk/lookup.html
- BYU corpus of American English – 360+ million words of American English from 1990-2007, with a variety of search possibilities, http://www.americancorpus.org/
- CHILDES (Child Language Data Exchange System) – child language data from various languages, http://childes.psy.cmu.edu/
- COBUILD Corpus – a balanced corpus used to create Collins dictionaries, http://www.collins.co.uk/corpus/CorpusSearch.aspx
- FrameNet – Lexical resource based on frame semantics, http://framenet.icsi.berkeley.edu/index.php
- Lexis-Nexis Academic – http://www.lexis-nexis.com/universe/
- MRC Psycholinguistic Database Web Interface – a word list annotated by a range of properties, including part of speech, frequency, length, etc., http://www.psy.uwa.edu.au/mrcdatabase/uwa_mrc.htm
- OED (Oxford English Dictionary) – the example sentences make quite a corpus, http://dictionary.oed.com/
- Oxford Text Archive – a source for over 2500 electronic texts in over 25 different languages. http://ota.ahds.ac.uk/
- Stanford Humanities Digital Information Service – full text of lots of literature, http://library.stanford.edu/depts/hasrg/hdis/text.html
- WebCorp – uses the web as a corpus, http://www.webcorp.org.uk
- WordNet – lexical database with words grouped into synonym-sets interlinked by lexical and conceptual relations, http://wordnet.princeton.edu/

# Background Material on Corpora

Barlow, Michael and Suzanne Kemmer, eds. 2000. Usage-Based Models of Language. Stanford, CA: CSLI Publications. [a recent collection of studies drawing on usage data]

Barnbrook, Geoff. 1996. Language and computers: A Practical Introduction to the Computer Analysis of Language. Edinburgh: Edinburgh University Press.

Biber, Doug. 1993. Representativeness in Corpus Design. Literary and Linguistic Computing 4: 243-257. [criteria to consider in designing a corpus]

Biber, Douglas, Susan Conrad, and Randi Reppen. 1998. Corpus Linguistics: Investigating Language Structure and Use. Cambridge, UK: Cambridge University Press.

Edwards, Jane A. and Martin D. Lampert, eds. 1993. Talking Data: Transcription and Coding in Discourse Research. Hillsdale, NJ: Lawrence Erlbaum. [good for analysis of transcripts and other kinds of corpora on languge use; useful set of web-sites at the end]

Fillmore, Charles J. 1991. `Corpus Linguistics' or `Computer-aided Armchair Linguistics'. In J. Svartvik, ed., Directions in Corpus Linguistics, 35-60. Berlin: Mouton de Gruyter. [on Fillmore's conversion to corpus linguistics; includes a case study of {\it risk}]

Kennedy, Graeme. 1998. An Introduction to Corpus Linguistics. London: Longman. [this book provides a comprehensive introduction and guide to all aspects of corpus linguistics, from the various types of electronic corpora that are available to instructions on how to design and compile a corpus]

# Background Material on Corpora (cont.)

Lehmann, H.M., P. Schneider, and S. Hoffmann. 2000. BNCweb. In J.M. Kirk, ed., Corpora Galore: Analyses and Techniques in Describing English, 249-266. Amsterdam: Rodopi. [a description of a web interface to the British National Corpus, which we hope to have available soon]

Nelson, G., S. Wallis, and B. Aarts. 2002. Exploring Natural Language: Working with the British Component of the International Corpus of English. Amsterdam: John Benjamins. [description of the corpus plus a manual of the ICE corpus utility program]

Rickford, John R., Tom Wasow, Norma Mendoza-Denton, and Juli Espinoza. 1995. Syntactic Variation and Change in Progress: Loss of the Verbal Coda in Topic-Restricting as far as Constructions. Language 71: 102-131.

Stubbs, Michael. 1996. Text and Corpus Analysis: Computer-assisted Studies of Language and Culture. Oxford: Blackwell.

Stubbs, Michael. 2001. Words and Phrases: Corpus Studies of Lexical Semantics. Oxford: Blackwell. [a diverse series of case studies using corpus data; includes some discussion of methodological issues]

Wasow, Thomas. 2002. Postverbal Behavior. Stanford, CA: CSLI Publications. [an exploration of the order of postverbal elements in English through corpus studies and psycholinguistic experiments; also includes discussion of how studies of language use bear on issues of linguistic theory]