HG351 Corpus Linquistics

Introduction to Corpus Linguistics
Main Issues

Francis Bond
Division of Linguistics and Multilingual Studies
http://www3.ntu.edu.sg/home/fcbond/
bond@ieee.org

Lecture 1
https://github.com/bond-lab/Corpus-Linguistics

Introduction

- Administrivia
- What is Corpus Linguistics
- What this course is (and isn't)
- Getting to know each other (what do you want?)

Corpora I have been involved with

- Semantic markup of the LDC *Call Home Corpus*
  sense tagging of Japanese telephone transcripts

- *Hinoki Treebank of Japanese*
  HPSG parses of Japanese definitions, examples and newspaper text
  sense tagging of same

- *Tanaka Corpus* of aligned Japanese-English text
  Now the *Tatoeba multilingual project*: `www.tatoeba.org`

- NICT English learner corpus (advisor)

- *Japanese WordNet gloss corpus*, jSEMCOR corpus
  aligned Japanese-English text, sense tagging

Corpora I am building now

- **NTU Multilingual Corpus: NTU-MC**
  with help from Tan Liling, HG2002, HG8011 and many more
  - Arabic, Chinese, English, Indonesian, Japanese, Korean, Vietnamese
    - Essays
    - Short Stories (Sherlock Holmes)
    - News Text
    - Singapore Tourist Web Sites
  - Wordnet sense tagging, HPSG parses
  - Cross lingual alignment
  - Tagging various phenomena
  - Used in URECA and FYP research, we will use it in this course
- **NTU Corpus of Learner English: NTUCLE**
  - With help from LCC

(Tan and Bond, 2012; Winder et al., 2017)

100% Continuous Assessment

- Individual Lab Work (4x10%)
- Individual Project (20%)
  - Describe some linguistic phenomenon quantitatively in a 6-page paper (ACL format)
  - The paper must motivate both the choice of phenomenon and corpus
- Group Project (30%) One of:
  - A program to perform some substantial corpus processing task
  - The collection and annotation of a new (sub)corpus

  \+ 8-page paper (ACL full paper format with extra page for references) describing your approach
- Class Participation (10%)

Guidelines for Written Work in HG3051

- All assignments must follow the *(Computational) Linguistic Style Guidelines: a guide for the perplexed.*
  http://www3.ntu.edu.sg/home/fcbond/data/ling-style.pdf

- Proper citation is important
  — failure to cite is plagiarism — **zero** or **fail**

- Local Rules
  - ACL format for paper submission (No need for LMS title page)
    only the first n pages will be marked
  - Late assignments get **zero**
  - I expect some quantitative analysis
  - I will try to give you real problems to work on

Extra Credit

- If you submit a patch[1] that gets accepted to a corpus or tool we use
  - you can get 1-5% extra credit (depending on the size/difficulty) typically 10^{n-1} where n is the number of lines you changed
  - you can't go over 100%
- A patch can involve
  - extending the corpus/code with new capabilities
  - fixing a bug in annotation/code
  - fixing a bug in or extending documentation
    - fixing a spelling error; rewording for clarity; translating to a new language
- Has to be for this course (not overlap with URECA, project, HG2051, . . .)

[1] a short set of commands to correct a bug in a computer program

The goal of this course

Master the uses of text corpora in linguistics research and applications.

- Selecting text
- Marking up extra information
- The range of existing corpora
- How to build your own corpus
- Using corpora to test linguistic hypotheses
- Using corpora to train language tools
- Extracting knowledge from corpora

HG3051 Prerequisites

- Some linguistic knowledge assumed
  - You know what a **lexeme** is
  - You know what an **inflectional paradigm** is
  - You know what a **constituent** is

  If you don't know these, you will have to do a little background reading, I recommend Huddleston (1988)

- A little computational knowledge (not required but useful)
  - You will learn some very simple techniques here
  - You will learn to use some corpus programs
  - If you can program a little I encourage you to use your skills

What do you learn?

On completion of this module, students should be able to:

- Understand the uses of text corpora in language research
  Be able to manipulate them with simple tools
- Use a concordance program to extract data from a corpus
- Design and build a corpus for some task
- Understand how to analyse corpus data through basic statistical methods

Textbook and Readings

- I haven't found a good text book, so we won't use one.
  - Stubbs, Michael, *Text and Corpus Analysis*. Blackwell Publishers, 1996 is not bad
- Readings will be assigned, I will try to choose works that are on-line.
- All Wikipedia articles cited have been checked by me, and I will watch them for changes. (extend the web of trust)

Student Responsibilities

By remaining in this class, the student agrees to:

1. Make a good-faith effort to learn and enjoy the material.
2. Read assigned texts and participate in class discussions and activities.
3. Submit assignments on time.
4. Attend class at all times, barring special circumstances (see below).
5. Get help early: approach me when you first have trouble understanding a concept or homework problem rather than complaining about a lack of understanding afterward.
6. Treat other students with respect in all class-related activities, including on-line discussions.

Attendance

1. You are expected to attend all classes.

2. Be on time - lateness is disruptive to your own and others' learning.

3. Valid reasons for missing class include the following:

   (a) A medical emergency (including mental health emergencies)
   (b) A family emergency (death, birth, natural disaster, etc).

   You must provide documentation to me and the student office.

4. There will be significant material covered in class that is not in your readings. You cannot expect to do well without coming to class.

5. If you miss a class, it is your responsibility to get the notes, any handouts you missed, schedule changes, etc. from a classmate.

Remediation and Academic Integrity

1. No late work will be accepted, except in the case of a documented excuse.

2. For planned, justified, absences on class days or days on which assignments are due, advance notice must be provided.

3. Cheating will not be tolerated. Violations, including plagiarism, will be seriously dealt with, and could result in **a failing grade for the entire course**.

4. For all other issues of academic integrity, refer to the University Honour Code: http://www.ntu.edu.sg/sao/Pages/HonourCode.aspx

5. As always, use your common sense and conscience.

Why do you do HG351?

- Language Poll (What do you speak and/or study?)
  - Natural
    - Mandarin
    - Bahasa Malay
    - Tamil

    ...
  - Corpus Type
    - Text
    - Speech
    - Other

    ...

What is a Corpus?

corpus (pl: ***corpora***):

1. A collection of texts, especially if complete and self-contained: the corpus of Anglo-Saxon verse.

2. In linguistics and lexicography, a body of texts, utterances, or other specimens considered more or less representative of a language, and usually stored as an electronic database. Currently, computer corpora may store many millions of running words, whose features can be analyzed by means of tagging (the addition of identifying and classifying tags to words and other formations) and the use of concordancing programs. Corpus linguistics studies data in any such corpus . . .

(from *The Oxford Companion to the English Language*, ed. McArthur & McArthur, 1992)

Definition of a corpus

- In principle, any collection of more than one text can be called a **corpus**
- Characteristics of modern corpora:
  - machine-readable (i.e., computer-based)
  - authentic (i.e., naturally occurring)
  - sampled (bits of text taken from multiple sources)
  - representative of a particular language or language variety.
- Sinclair (1991, 171):

  A corpus is a collection of naturally-occurring language text, chosen to characterize a state of variety of language.

Why Are Electronic Corpora Useful?

- as a collection of examples for linguists
- as a data resource for lexicographers
- as instruction material for language teachers and learners
- as training material for natural language processing applications
  - training of speech recognizers
  - training of statistical part-of-speech taggers and parsers
  - training of example-based and statistical machine translation systems
- “Big Data” is just another corpus, to analyze it wisely requires the same techniques

Examples for Linguists

Give examples for English noun phrases . . .

Examples for Linguists

Examples from the Penn treebank:

(1) *USX 's transition from Big Steel to Big Oil*

(2) *Pittsburgh instead of New York or Findlay, Ohio, Marathon 's home*

(3) *his concern about boosting shareholder value*

(4) *the modest goal of becoming tax manager by the age of 46*

(5) *a move that, in effect, raised the cost of a $7.19 billion Icahn bid by about $3 billion*

(6) *an undistinguished college student who dabbled in zoology until he concluded that he couldn't stand cutting up frogs*

(7) *the sale of the reserves of Texas Oil & Gas, which was acquired three years ago and hasn't posted any significant operating profits since*

Some Linguists dismiss Corpus Linguistics

. . . it is obvious that the set of grammatical sentences cannot be identified with any particular corpus of utterances . . .

. . . a grammar mirrors the behavior of the speaker, who, on the basis of a finite and accidental experience with language, can produce or understand an indefinite number of new sentences.

. . . ones's ability to produce and recognize grammatical utterances is not based on notions of statistical approximations or the like. . . . If we rank the sequences of a given length in order of statistical approximation to English, we will find both grammatical and ungrammatical sequences scattered throughout the list; there appears to be no particular relation between the order of approximations and grammatical.

Chomsky (1957, pp15–17) *Syntactic Structures*

Can grammaticality be predicted?

(8) *Colorless green ideas sleep furiously.*

(9) **Furiously sleep ideas green colorless.* (Chomsky, 1957: as (1) and (2))

It is fair to assume that neither sentence (8) nor (9) (nor indeed any part of these sentences) has ever occurred in an English discourse. Hence, in any statistical model for grammaticalness, these sentences will be ruled out on identical grounds as equally 'remote' from English. Yet (8), though nonsensical, is grammatical, while (9) is not.

Not really: Using a simple probabilistic model (based only on the probability of a word occurring given the two proceeding words) Pereira (2000) showed that P(8) $\gg$ P(9) ($\times 200,000$).

Context helps

It can only be the thought of verdure to come, which prompts us in the autumn to buy these dormant white lumps of vegetable matter covered by a brown papery skin, and lovingly to plant them and care for them. It is a marvel to me that under this cover they are labouring unseen at such a rate within to give us the sudden awesome beauty of spring flowering bulbs. While winter reigns the earth reposes but these colourless green ideas sleep furiously. C.M Street (1985)

Why do Linguists need Corpora?

Chomksy The verb ***perform*** cannot be used with mass word objects: one can *perform a task* but not *perform labour*.

Hatcher How do you know, if you don't use a corpus and have not studied the verb ***perform***?

Chomksy How do I know? Because I am a native speaker of the English Language.

Hill (1962:29) cited in McEnery and Wilson (2001, 11)

This is why

From the BNC (search for "`perform [nn1*]`")

```
PERFORM MUSIC  4
PERFORM WORK  4
PERFORM SURGERY  3
PERFORM EUTHANASIA  2
PERFORM RESEARCH  2
```

many Continental musicians, and it can not be doubted that professional English singers often perform music which they have not had time to " learn " in any sense of

Not only do " Saxtet " perform music previously unassociated with the saxophone, but they include a selection of their own

Linguists' intuitions are unreliable: Explanations of languages based on false data are not very valuable.

Examples for Lexicographers

How many senses does the word ***line*** have?

Examples for Lexicographers

The noun line has 30 senses according to WordNet (first 23 from tagged texts):

1. (51) ***line*** — (a formation of people or things one beside another; *the line of soldiers advanced with their bayonets fixed*; *they were arrayed in line of battle*; *the cast stood in line for the curtain call*)

2. (20) ***line*** — (a mark that is long relative to its width; *He drew a line on the chart*)

3. (15) ***line*** — (a formation of people or things one behind another; *the line stretched clear around the corner*; *you must wait in a long line at the checkout counter*)

4. (13) ***line*** — (a length (straight or curved) without breadth or thickness; the trace of a moving point)

5. (11) ***line*** — (text consisting of a row of words written across a page or computer screen; *the letter consisted of three short lines*; *there are six lines in every stanza*)

6. (10) ***line*** — (a single frequency (or very narrow band) of radiation in a spectrum)

7. (10) ***line*** — (a fortified position (especially one marking the most forward position of troops); *they attacked the enemy's line*)

8. (10) argumentation, logical argument, argument, line of reasoning, ***line*** — (a course of reasoning aimed at demonstrating a truth or falsehood; the methodical process of logical reasoning; *I can't follow your line of reasoning*)

9. (9) cable, ***line***, transmission line — (a conductor for transmitting electrical or optical signals or electric power)

10. (8) course, ***line*** — (a connected series of events or actions or developments; *the government took a firm course*; *historians can only point out those lines for which evidence is available*)

11. (6) ***line*** — (a spatial location defined by a real or imaginary unidimensional extent)

12. (5) wrinkle, furrow, crease, crinkle, seam, ***line*** — (a slight depression in the smoothness of a surface; *his face has many lines*; *ironing gets rid of most wrinkles*)

13. (4) pipeline, ***line*** — (a pipe used to transport liquids or gases; *a pipeline runs from the wells to the seaport*)

14. (4) ***line***, railway line, rail line — (the road consisting of railroad track and roadbed)

15. (3) telephone line, phone line, telephone circuit, subscriber line, ***line*** — (a telephone connection)

16. (3) ***line*** — (acting in conformity; *in line with*; *he got out of line*; *toe the line*)

17. (2) lineage, ***line***, line of descent, descent, bloodline, blood line, blood, pedigree, ancestry, origin, parentage, stemma, stock – (the descendants of one individual; *his entire lineage has been warriors*)

18. (2) ***line*** — (something (as a cord or rope) that is long and thin and flexible; *a washing line*)

19. (2) occupation, business, job, line of work, ***line*** — (the principal activity in your life that you do to earn money; *he's not in my line of business*)

20. (1) ***line*** — (in games or sports; a mark indicating positions or bounds of the playing area)

21. (1) channel, communication channel, ***line*** — ((often plural) a means of communication or access; *it must go through official channels*; *lines of communication were set up between the two firms*)

22. (1) ***line***, product line, line of products, line of merchandise, business line, line of business – (a particular kind of product or merchandise; *a nice line of shoes*)

23. (1) ***line*** — (a commercial organization serving as a common carrier)

24. agate line, ***line*** — (space for one line of print (one column wide and 1/14 inch deep) used to measure advertising)

25. credit line, line of credit, bank line, ***line***, personal credit line, personal line of credit – (the maximum credit that a customer is allowed)

26. tune, melody, air, strain, melodic line, ***line***, melodic phrase – (a succession of notes forming a distinctive sequence; *she was humming an air from Beethoven*)

27. ***line*** — (persuasive but insincere talk that is usually intended to deceive or impress; *'let me show you my etchings' is a rather worn line*; *he has a smooth line but I didn't fall for it*; *that salesman must have practiced his fast line of talk*)

28. note, short letter, ***line***, billet – (a short personal letter; *drop me a line when you get there*)

29. ***line***, dividing line, demarcation, contrast – (a conceptual separation or distinction; *there is a narrow line between sanity and insanity*)

30. production line, assembly line, ***line*** — (mechanical system in a factory whereby an article is conveyed through sites at which successive operations are performed on it)

Instruction for Language Learning

Which do you say in English: *think about* or *think on*?

Instruction for Language Learning

Which do you say in English: *think about* or *think on*?

If in doubt, ask google:	36,300,000 hits	*think about*
	738,000 hits	*think on*

Types of Corpora

- mono-lingual versus multi-lingual corpora
- special-purpose, domain-specific corpora versus general-purpose, large-scale corpora
- spoken language corpora versus collections of written text
- ad-hoc corpus collections versus balanced, representative corpora
- raw text versus marked-up documents
- unannotated versus annotated corpora
- Web as a corpus

What does a corpus consist of?

- A collection of ordinary text files (Raw Corpus)
- Annotated corpora
  - Raw corpora with html/xml tags (genre, date, subject, . . .)
  - Annotated corpora (part of speech, syntactic structures, etc.)

The British National Corpus (BNC)

- 100 million words of written and spoken British English (Burnard, 2000)
- Designed to represent a wide cross-section of British English from late 20th century: balanced and representative
- POS tagging (2 million word sampler hand checked)

Written	Domain	Date	Medium
(90%)	Imaginative (22%)	1960-74 (2%)	Book (59%)
	Arts (8%)	1975-93 (89%)	Periodical (31%)
	Social science (15%)	Unclassified (8%)	Misc. published (4%)
	Natural science (4%) . . .		Misc. un-pub (4%)
Spoken	Region	Interaction type	Context-governed
(10%)	South (46%)	Monologue (19%)	Informative (21%)
	Midlands (23%)	Dialogue (75%)	Business (21%)
	North (25%) . . .	Unclassified (6%)	Institutional (22%) . . .

General vs. specialized corpora

- General corpora (such as national corpora) are a huge undertaking. These are built on an institutional scale over the course of many years.

- Specialized corpora (ex: corpus of English essays written by Japanese university students, medical dialogue corpus) can be built relatively quickly for the purpose at hand, and therefore are more common

- Characteristics of corpora:

  1. Machine-readable, authentic
  2. Sampled to be balanced and representative

- Trend: for specialized corpora, criteria in (2) are often weakened in favor of quick assembly and large size
  - Do-it-yourself corpora
  - World-Wide Web as a corpus
  - Google 1T corpus

  Rare phenomena only show up in large collections

A short list of well-known corpora

- National corpora:
  - The British National Corpus
  - The American National Corpus
  - The German National Corpus
  - King Sejong the Great Corpus

  Chinese, Greek, Italian, Hungarian, Polish, Czech . . .

- Other Well Known Corpora:
  - Brown Corpus
  - Corpus of Contemporary American English
  - Michigan Corpus of Academic Spoken English

- 1st Language acquisition:
  - CHILDES (Child Language Data Exchange System)
- 2nd Language acquisition (mostly English)
  - ICLE (the International Corpus of Learner English) and LOCNESS (the Louvain Corpus of Native English Essays)
  - Longman Learners' Corpus
  - CLC (Cambridge Learner Corpus)
- Multilingual Corpora
  - Canadian Hansard
  - Hong Kong Hansard
  - Europarl
- Parsed Corpora
  - Penn Treebank (WSJ, Brown, Chinese)
  - Czech Dependency Bank
  - Redwoods HPSG corpus of English

See Also

- Linguist list corpora page
  https://www.linguistlist.org/sp/GetWRListings.cfm?wrtypeid=1
- ACL Siglex Links to the CORPORA Mailing List Archive
- Linguistics Data Consortium (LDC)
- European Language Resources Association (ELRA)
- *Gengo Shigen Kyouyuukikou* Language Resource Consortium (GSK)
- Chinese Linguistic Data Consortium (CLDC)

Corpora at NTU

- Cantonese Corpus (KK)
  Now at https://github.com/fcbond/hkcancor
- Tatoeba Japanese-English (FCB)
- Various small corpora (AC, FK)
- NTU Multilingual Corpus (under construction: FCB)
- NTU Learner Corpus (under construction: FCB)
- We may add to these in this class

Let's Explore

Go to the BYU interface to the BNC (see web page for logon details).

MORPHOLOGY : Look for words starting with the prefix *dis-* (e.g. *dissent*). What are the three most common singular nouns (`dis*.[nn1]`), the three most common adjectives (`dis*.[j*]`), and the three most common infinitival verbs (`dis*.[vvi]`)

LEXICAL : Search for *robot* [using CHARTS] and then compare the frequency in the five main genres. In which genre is it the most/least common? In which sub-genre is it the most common (click on [SEE ALL SECTIONS]

COLLOCATIONS : What are the 5 most frequent adjectives with *curry* as a noun (`curry.[nn*]`)? (CONTEXT = [j*], [4] [4], [SORT] = [FREQUENCY]). Now change to [SORT] = [RELEVANCE]. What are the five most highly-ranked adjectives. What has changed, and why?

GRAMMATICAL : In which genre is the present perfect (`has[vvn*]`) and the past perfect (`had[vvn*]`) most common? Any idea why?

LEXICO-GRAMMAR : Look at the top five adjectives following *come* and *go* (use [COMPARE WORDS]; WORD(S) = come , go; CONTEXT = [j*] [0] [2]). Is there any pattern in terms of which adjectives occur with the two verbs?

SEMANTICS : Compare the collocates of *find* and *discover* (use [COMPARE WORDS]; WORD(S) = find.[v*] , discover.[v*]; CONTEXT = [nn*] [0] [2]). Any patterns here?

LEXICO-GRAMMAR : Compare the five most common phrases with `we[v*]` in SPOKEN vs ACADEMIC. What is the major difference between the two registers?

LEXICAL : Compare the most frequent singular and plural nouns (`[nn1*]` and `[nn2*]`) in MAGAZINE vs ACADEMIC). Which types are more common in each register?

Acknowledgments

- Thanks to Na-Rae Han for inspiration for some of the slides (from *LING 2050 Special Topics in Linguistics: Corpus linguistics*, U Penn) and also for the Student Policies (adapted).
- Thanks to Sandra Kübler for some of the slides from her *RoCoLi*[2] *Course: Computational Tools for Corpus Linguistics*
- Thanks to Mark Davies (BYU) for the exploration ideas.
- Definitions from WordNet 3.0

[2] *Romania Computational Linguistics Summer School*

*

References

Lou Burnard. 2000. *The British National Corpus Users Reference Guide*. Oxford University Computing Services.

Noam Chomsky. 1957. *Syntactic Structures*. Mouton.

Rodney Huddleston. 1988. *English grammar: an outline*. Cambridge University Press, Cambridge.

Tony McEnery and Andrew Wilson. 2001. *Corpus Linguistics*. Edinburgh UP, second edition.

Fernando Pereira. 2000. Formal grammar and information theory: together again? *Philosophical Transactions of the Royal Society*, 358(1769):12391253. http://dx.doi.org/10.1098/rsta.2000.0583.

John Sinclair. 1991. *Corpus, Concordance, Collocation*. Oxford UP.

Liling Tan and Francis Bond. 2012. Building and annotating the linguistically diverse NTU-MC (NTU-multilingual corpus). *International Journal of Asian Language Processing*, 22(4):161–174.

Roger V. P. Winder, Joe MacKinnon, Shu Yun Li, Benedict Lin, Carmel Heah, Lus Morgado da Costa, Takayuki Kuribayashi, and Francis Bond. 2017. NTUCLE: Developing a corpus of learner English to provide writing support for engineering students. In *The 4th Workshop on NLP Techniques for Educational Applications (NLPTEA 2017)*. (IJCNLP 2017 Workshop).