

Introduction/Overview

LING 5200/6200

Corpus Linguistics

Kevin Cohen & Bert Xue

© Kevin Bretonnel Cohen, 2008

[http://verbs.colorado.edu/~xuen/
teaching/ling5200/](http://verbs.colorado.edu/~xuen/teaching/ling5200/)

What's a corpus?

- Meyer:
 - "a collection of texts or parts of texts upon which some general linguistic analysis can be conducted"
 - "a body of text made available in computer-readable form for purposes of linguistic analysis"

What's a corpus?

- McEnery & Wilson:
 - (i) (loosely) any body of text
 - (ii) (most commonly) a body of machine-readable text
 - (iii) (more strictly) a finite collection of machine-readable text, sampled to be maximally representable of a language or variety

What's corpus linguistics?

- "the study of language based on examples of 'real life' language use" (McEnery & Wilson)
- Biber et al.:
 - Uses computers
 - "Natural" texts
 - Large & principled collection
 - Both quantitative and qualitative

What's corpus linguistics not?

- Phonetics
- Phonology
- Morphology
- Syntax
- Semantics
- Pragmatics
- Psycholinguistics
- Computational Lx
- Descriptive Lx
- Historical Lx
- Sociolinguistics

What's a corpus linguistics research project like?

- Empiricist, not rationalist (mostly)
- Concerned with use, not structure
- Quantitative, not qualitative (mostly)
- Embraces variability, doesn't reject it
- "Descriptive" adequacy, not "explanatory" adequacy
- NOT mutually exclusive with rationalist, qualitative, "explanatory" work
- Some things you can't do without a corpus-based approach: phonetics, CLA

Some good corpus linguistics projects

- Are the patterns of occurrence of common syntactic phenomena like coordination and negation the same in news/speech and scientific writing? (*They're not.*) Why?
- Write code to determine how well a given data set fits the sublanguage model.
- Here are two corpora. Do anything you can to demonstrate to me that they're different.
- Do experiencer-subject (*admire*) and experiencer-object (*amuse*) verbs have different frequencies in the past tense?

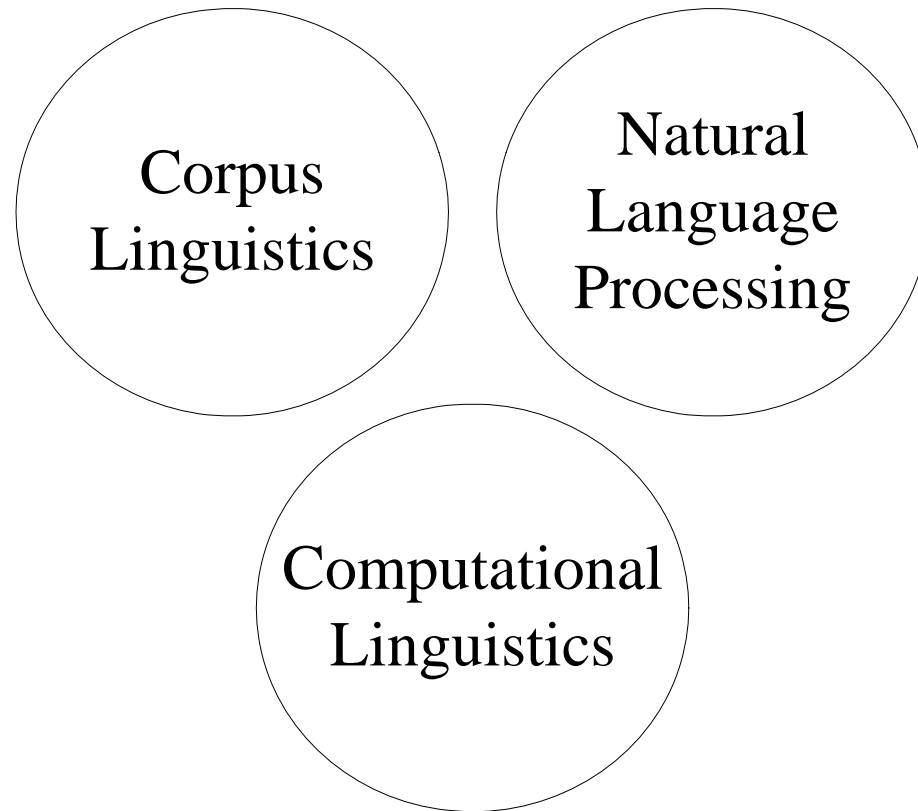
What isn't corpus linguistics?

✓ 1900-1960

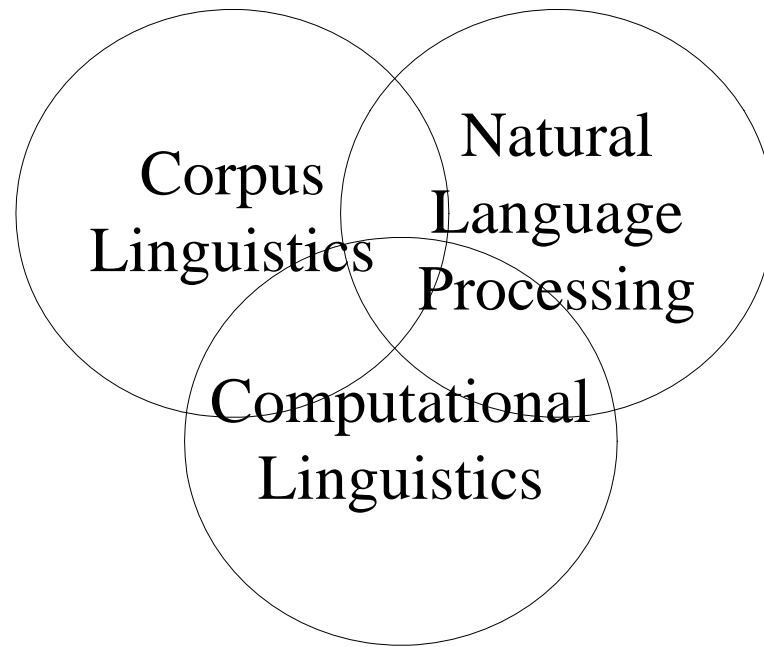
X 1960-1980

✓ 1980-present

Corpus linguistics in context



Corpus linguistics in context



What's LING 5200 Corpus Linguistics?

- Tools
- Techniques

Overview

- Intro to Unix (Kevin)
- A little corpus design (Kevin)
- Quick tour of corpora (Kevin)
- Practical techniques for building a corpus (Bert)
- Programming in Python (Bert)
- Tools for working with corpora (Kevin)

Why Perl?

- It works
- Many advantages
- Gentle learning curve

Why Perl?

- It works
 - Stemmers, POS taggers, text alignment tools...
- Many advantages
- Gentle learning curve

Why Perl?

- It works
 - Stemmers, POS taggers, text alignment tools...
- Many advantages
 - Phenomenal text-handling abilities, regular expressions...
- Gentle learning curve

Why Perl?

- It works
 - Stemmers, POS taggers, text alignment tools...
- Many advantages
 - Phenomenal text-handling abilities, regular expressions...
- Gentle learning curve
 - Linguists, MD's, software testers...

Why Python?

- It works
- Required for CSCI 5832 (Natural language processing)

Why not Java/C++?

```
public class HelloWorld {  
    public static void main(String[]  
        args) {  
        System.out.println("hello,  
        world");  
    }  
}
```

```
print 'hello, world'
```

Why not something else?

LINGUIST-L				monster.com			
	2003	2004	2008		2003	2004	2008
C++	51	82		Java	4,134	>5,000	>5,000
Java	40	63		C++	2,570	>5000	>5,000
Perl	34	66		Perl	1,404	3,243	3,790
Python	6	12		Python	114	270	1,039
Lisp	4	6		Lisp	15	22	30

LINGUIST-L shows jobs submitted since 1/1/2002.

Administrative things

- Textbooks
- Office hours
- Prerequisites
- Grades
- Accounts on babel/magellan
- Mailing list
- Web page

Me

- Corpora for:
 - information extraction & entity identification
 - evaluation/testing
 - building statistical language models
- Corpora types:
 - written
 - Un-, lightly, special-purpose annotated
 - ontologies/vocabularies
- Domains:
 - molecular biology
 - medical

Bert

- **Corpus creation:**
 - Chinese Treebank
 - Chinese Propbank
 - Temporal and discourse annotation
 - Sense tagging
 - OntoNotes
- **Using Corpora:**
 - Chinese word segmentation
 - Parsing
 - Semantic Role labeling
 - Word sense disambiguation
 - Temporal inference

You

- Grad student in linguistics dept.
- Want to develop skills in:
 - Unix
 - Programming
 - Corpus manipulation tools
- Survey
 - Don't need your name on first page;
please include your email address on second page

[http://verbs.colorado.edu/~xuen/
teaching/ling5200/](http://verbs.colorado.edu/~xuen/teaching/ling5200/)