Computational Methods in Linguistics

Bender and Wassink 2012
University of Washington

Overview

- Course goals
- Course requirements
- Linguistic databases
- Why automate?
- Lab 1 preview

Course goals

- Be able to frame questions such that linguistic databases can be brought to bear on them
- Be familiar with existing linguistic databases and how to access them
- Understand how annotated corpora are created and how to cope with variability in annotations and other noise in the data

Course goals

- Be familiar with existing NLP tools for automatically adding certain annotations
- Be able to write simple computer programs to find and count things in large datasets
- Understand the issues that arise in making data reusable
- Understand the connection between data sampling and inferential stats and issues around choosing statistical tests for a given study

Your goals

- Interactions with lexical frequency; how to calculate lexical frequency and based on what data?
- How to use corpora/ DBs +3
- Frequency of t glottalization in spoken corpora
- Finding metalinguistic comments in collected texts
- Scripting languages (Python, Praat) +2
- Stats
- Evidence-based approach to semantics (not just ex, counter-ex)

Your goals

- What resources are available beyond raw corpora
- Text corpora for ESL/ CALL applications +1
- Unix, R
- Overcoming fear of command line
- Epicene pronouns in text
- Databases of other linguistic facts (e.g., attitudinal information)
- How do process data faster (transcription, annotation, tagging, alignment, phon measures, ...)

Course technology

- Canvas, Tegrity
- Patas cluster: access to corpora, access to NLP software, access to unix environment
- Lemur: version control
- Treehouse: Access to linux machines
- (BYO) Laptops: Wednesdays will include time for hands-on work on exercises.

Tech inventory

- OS: Who's using Mac, Windows, Linux?
- Text editors: What are you using?
- Version control
- Programming languages

ASK QUESTIONS!

- If we're doing it right, we'll be asking you to stretch beyond your comfort zone in this class
- Working with computers, there is a big potential for black hole time sinks
- Solution: 10 minute rule
    - If you've worked at something for 10 minutes and don't yet have the answer, that's when you post the question

Course requirements

- Weekly readings; be prepared to discuss these in class
    - Reading questions (due in Canvas by Midnight the night before class)
- Weekly labs/exercises, due 5pm on Fridays
    - Labs are individual
    - Practicum work is collaborative
- Term project, using computational methods and linguistic databases

Reading Questions

- Encourage careful reading of course materials
- Make class sessions more responsive to student questions
- Due by midnight the night before each class with a reading assignment

Reading Questions

- Post to Canvas discussion area with answers to:
  - What in the reading was most confusing? If you can, articulate a question about it. If nothing was confusing, what further questions does this reading raise for you?

Term project

- Apply computational methods to use existing linguistic databases to answer questions connected to your broader research progam.
- Not collecting, curating or annotating data

Write-up

- Hypothesis
- Resources
- Methodology
- Results
- Implications
- Future work

- Also turn in:
  - Scripts
  - Results files
  - LSA-style abstract

Final project schedule

- 4/6 project proposals due: define research questions
- 4/18 list of relevant resources
- 5/2 interim project report
- 5/23, 5/30 in-class presentations
- 6/6 final write up

Why isn't this course called "corpus linguistics"?

- Neither instructor is a corpus linguist!
- Goal is to work with large databases, not "corpora", strictly speaking
- Our focus is on existing computational resources including text and speech corpora, annotations over those corpora and software -- including scripts *you* write -- for manipulating them.

How is "corpus linguistics" defined by corpus linguists?

- McEnery and Wilson, 1996: "the study of language based on examples of real life language use"

- Biber et al., 1998: Corpus-based analyses of language. Empirical studies, which analyze the actual patterns of use in actual natural texts, use a large principaled collection of natural texts[1], make extensive use of computers, both interactive and automatic techniques, and depend on both quantitative and qualitative analytical techniques

- Baker, 2010: "the study of language using examples from real life"

- Baker, 2010: "Corpus linguistics looks at language...from a social perspective."

[1] Text: unannotated written language, or written-down language (transcribed speech)

Corpus-*hyphenated*

- *corpus-based analysis*: corpus is used as a source of examples to check researcher's intuitions, or examine the frequency or plausibility of the forms contained in some smaller dataset (Biber et al.)

- *corpus-driven analysis:* inductive method of analysis. the corpus itself is the data. Linguistic theories are built upon what is learned from this corpus. Note: cannot approach a corpus entirely naively.

- *corpus-assisted analysis:* corpus is used as data in order to carry out linguistic analysis, but can also involve other forms of data or analysis (interviews, historical research, etymologies, etc.) (Partington 2006)

What is a corpus, then?

- Baker: "a body of language, or more specifically, a (usually) very large collection of naturally occurring language, stored as computer files." (p. 6)

- Typically, corpora are texts represented in written form (as words)

  ...very few contain sound files, pictures, or video data (alone or in combination)

- Intended to serve as a representative sample of some language variety
  -instructed with care to ensure representation
  -balanced
  -in principle, allows us to make claims not on a sample from the population, but from the population itself (too strong)
  -can compare with less carefully sampled collections of texts (textual databases)

- Allows extrapolation of linguistic frequencies and patterns

- Baker: "Additionally, within large corpora rare or unusual cases of language use are likely to occur, which may not be so readily attained via introspection...or examining smaller samples"

- Often useless unless we have the tools in hand to analyze it.

- *Often, the "stuff we analyze" are measurements over or annotations added to some corpus of naturally-occurring language...a database.*

Types of Corpora

(source: Gries, 2010)

General vs. *specific* corpora

General: representative and balanced for the language as a whole

Specific: restricted to a particular variety, genre, register, etc.

Annotated vs. *raw*

Raw: contain only corpus material

Annotated: contain corpus material ... plus annotations over the data*

(*plus plus = metadata in a header or readme file)

Monolingual vs. *parallel*

Monolingual: provide information about just one language

Parallel: same text in several different languages (e.g., EU Parliament debates are translated into 23 languages or Canadian Parliament debates in English and French)

Types of Corpora, cont.

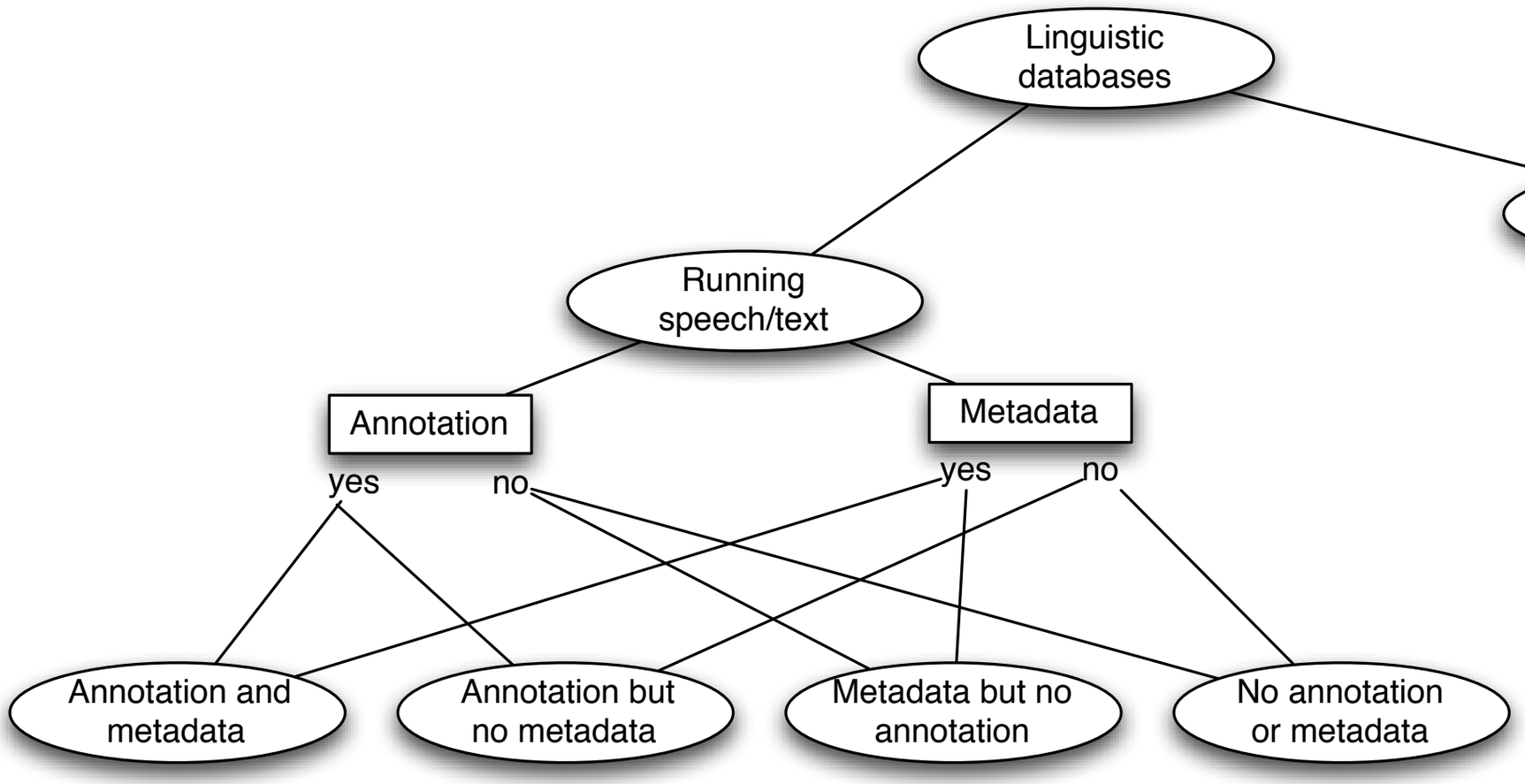
Static vs. *dynamic (or "monitor")*

Static: fixed size

Dynamic or Monitor: extended with new material (e.g., the Bank of English)

Diachronic vs. *synchronic*

Typology of Databases

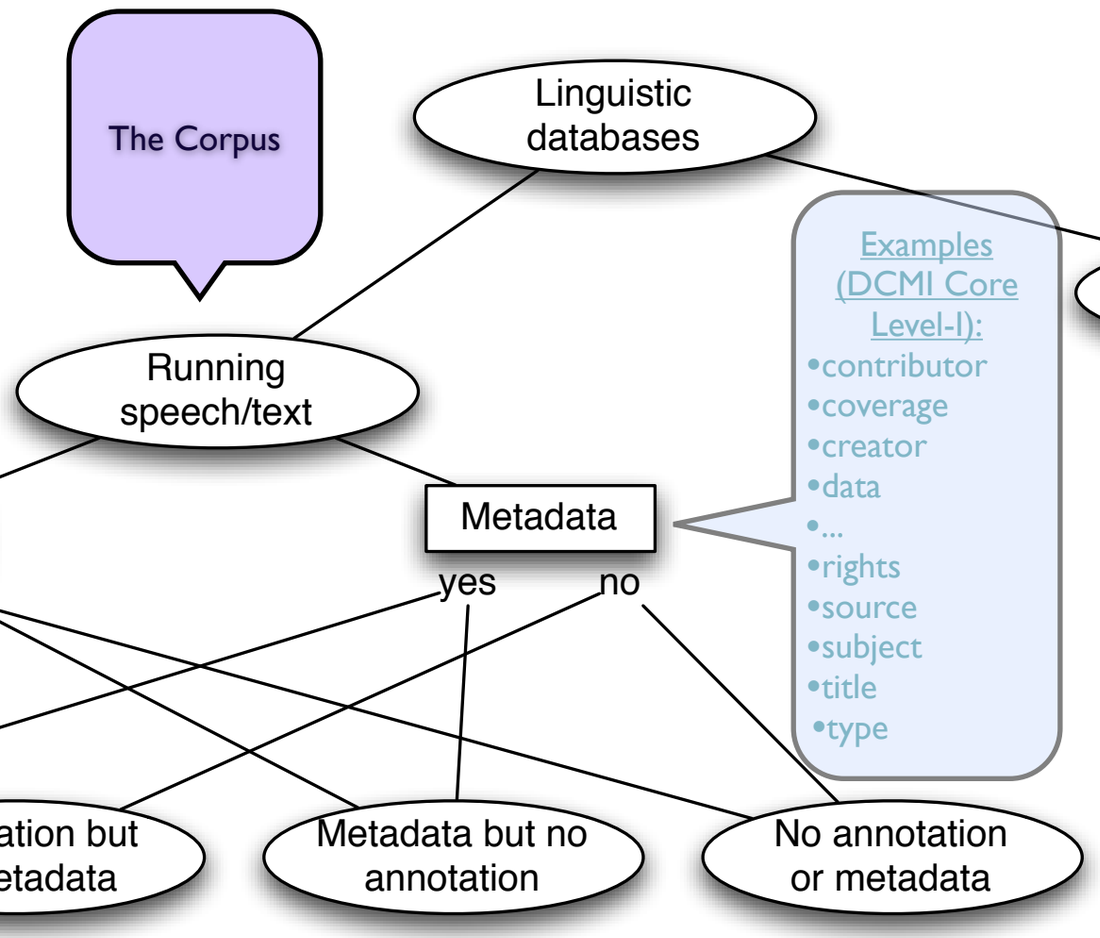


Databases of other linguistic facts

- WALS
- WordNet
- Pronouncing dictionaries
- PanLex
- ...

Typology of Databases

Examples:
- Speaker or Lg ID
- Orthographic transcription
- IPA transcription
- Tokenization (segmentation)
- Lemmatization
- Gloss
- Word senses
- POS tags
- Morpheme segmentation
- Syntactic parses
- Events (TimeML)
- Dialogue acts
- Prosodic (stress, pitch peak)
- Phonetic events (release burst tags, sociolinguistic variant)



Examples (DCMI Core Level-I):
- contributor
- coverage
- creator
- data
- ...
- rights
- source
- subject
- title
- type

Databases of other linguistic facts

- WALS
- WordNet
- Pronouncing dictionaries
- PanLex
- ...

Disagreements: is corpus linguistics a methodology, a theory of language, both?

- McEnery and Wilson: "conceived as nothing but a methodology (created from theoretical principles about language)

    - counting things (Lg X has 2-letter, 3-letter, 4-letter words...Ns, Vs, Adjs, etc.)
    - determining the frequencies of things (e.g. the most common words)
    - later... "association patterns": the systematic ways in which linguistic features are used in association with other linguistic and non-linguistic features (Biber et al. 1998)
        lexical associations: systematic associations between words
            (e.g., *big* tends to frequently co-occur with *toe; large* with *number)*
        grammatical associations: systematic associations between grammatical items in immediate contexts
            (e.g., *verbs* like "hope" can occur in either *that*-clauses or *to*-clauses, as in *I hope that I can go, vs. I hope to go.)*

- Teubert: not a method, but an insistence on working only with real language data taken from discourse in a principled way"

- Tognini-Bonelli (2001): corpus linguistics has gone well-beyond its methodological role to become an independent discipline

- Baker (2010): gone well beyond its methodological role, but is NOT an independent discipline in the same way as phonetics syntax, semantics or pragmatics

What Database Resources exist?

- First, need to consider that the resources we find must be related to our research questions.
- Biber et al. 1998 divide studies of language into two types:
  - Structure: possible forms that occur in a language
  - Use: forms used (or not) by actual speakers
- Many research questions in studies of language use originate in prior structural analyses:
  - Hypothetical relation b/w spontaneous speech and sentence complexity
  - Differences in different styles of academic writing
  - Common types of L2 errors
- To these we can add questions about patterns of use

What do we wish to accomplish with our database?

- What are you working on now?
- How can we use linguistic databases to further those goals?

Why automate?

- Automation = Reproducibility/Verifiability
- Automation = Flexibility
- Automation = Extensibility, by self or others
- But also: Automation requires quality assurance

That's Science!

- Automation = Reproducibility/Verifiability
- Automation = Flexibility
- Automation = Extensibility, by self or others
- But also: Automation requires quality assurance

Ideal scenario

- Conceptual work: literature review, hypothesis generation
- Data gathering: new data or identification of existing resources
- Annotation: work is recorded as part of database
- Analysis: work is recorded as scripts that take database and produce numerical results and graphs
- Result: Reproducibility!

The command line

- What is scary about the command line?
- Why should you use it anyway?
  - More flexible access to data (contents of files)
  - Extremely powerful light-weight tools
  - Batch processing

Lab 1 preview

- http://courses.washington.edu/cmling/lab1.html

Things to do now

- Request patas account: https://vervet.ling.washington.edu/db/accountrequest-form.php
- Request lemur account: email linghelp@u with UWNetID and statement of affiliation
- NB: Use a strong password for patas, use a password you don't rely on elsewhere for lemur