Corpus Linguistics

Why Corpus Linguistics?

Markus Dickinson
Department of Linguistics, Indiana University
Spring 2013

Corpus Linguistics

Why Corpus Linguistics?

Corpora

Corpus linguistics

Applications

Computational linguistics

Theoretical linguistics

Approaches

Objectives

Training of speech recognizers

Training of example-based and statistical machine translation systems

Potentially representativeness could be based on linguistic or collection of examples for linguists

Training material for natural language processing (NLP)

Data resource for lexicographers

Instruction material for language teachers and learners

Methodology: a way to gather relevant data

Theory: an independent discipline in its own right

Corpus linguistics means many things to many people

Corpus linguistics

A corpus is a body of naturally-occurring text

(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)

Why use computers to study language?

Computers offer a variety of benefits for handling text:

- manipulate data easily & rapidly (searching, sorting, etc.)
- process data accurately & consistently
- process data reliably, without human bias
- automatically annotate data, i.e., allow for further processing downstream

Why are electronic corpora useful?

Some purposes that corpora serve:

- collection of examples for linguists
- data resource for lexicographers
- instruction material for language teachers and learners
- training material for natural language processing (NLP) applications
  - training of speech recognizers
  - training of statistical part-of-speech taggers and parsers
  - training of example-based and statistical machine translation systems

What is a corpus?

To fit these purposes, corpora need some organization; general qualities:

- machine readable
- authentic texts
- texts which have been sampled to form a body of text
- representative of language, or a particular aspect of language
  - representativeness could be based on linguistic or non-linguistic criteria
  - potentially annotated with linguistic information

There are also specialized subcorpora which can meet certain research needs

Corpus linguistics

Corpus linguistics has a wide range of applications

...
Applications

Linguistic searching

Corpora can investigate questions such as:
- How does one order different types of adjectives in English?
- In what contexts are split constituents allowed in German?
- With what frequency do parasitic gaps occur in academic language?

Applications

Lexicography

How many senses does the word line have?
1. a comparatively strong slender cord
2. a cord, wire, or tape used in measuring and leveling
3. piping for conveying a fluid
4. a row of words, letters, numbers or symbols that are written, printed, or displayed
5. something that is distinct, elongated, and narrow
6. a state of agreement (bring ideas into line)
7. a course of conduct, action, or thought (a political line)
8. limit, restraint (overstep the line of good taste)...

A corpus can provide examples & help re-define senses

Applications

Corpora for computational linguistics

Corpora are useful for linguistics research, but have also revolutionized computational linguistics (CL)
- With annotation, CL can train and evaluate new algorithms
- Technology has become more robust and more efficient since the early 1990s
- All sorts of new annotations (with practical focuses—e.g., biomedical annotation) have taken off

We will investigate computational applications at various points this semester

Applications

Corpus Linguistics

Why Corpus Linguistics?

Corpora

Corpus linguistics

Applications

Computational linguistics

Approaches

Objectives

Data

Methods

Probabilistic language

Advantages

History

4.

8.

7.

5.

2.

3.

6.

1.

Corpus-based

Corpus-driven

Corpus-based research: corpora expound upon theories that were formulated before corpora
- May have to do away with particular pieces of evidence

Corpus-driven research: strictly committed to corpus data

Approaches to corpora

Corpus-based & Corpus-driven

For any of the questions mentioned above, we can compare for different language groups
- Do Indian speakers of English reduplicate words (more than other groups)?
- With what frequency do older speakers in the midwest use cool?
  - vs. younger speakers
  - vs. in the south
  - vs. written language

Applications

Language variation

What do you say in English: think about or think on?

According to Google (1/3/13):
290,000,000 hits for “think about” 8,050,000 hits for “think on”

Applications

Language learning

▶ How does one order different types of adjectives in English?
▶ In what contexts are split constituents allowed in German?
▶ With what frequency do parasitic gaps occur in academic language?
Corpus-based vs. Corpus-driven

1. Type of corpus data
   - representativeness: important for corpus-based approaches
   - corpus size: very large corpora (supposed to be balanced) argued for in corpus-driven approaches
   - annotation: corpus-driven approaches want to be pre-theoretical (which annotation is not) and derive categories completely from corpus

2. Attitude towards existing theories & intuitions
   - Corpus-based approaches uses existing theory as a starting point

3. Research focus:
   - Corpus-based: uses standard linguistic levels
   - Corpus-driven: holistic view, with a functional view of meaning

A brief history

There is a long history of using empirical, observed data, even before the advent of computers

- 1940s: structuralism, ‘shoebox corpora’
- late 1950s, 1960s: generativism, almost no corpus linguistics
  - Chomsky had several arguments against corpora (see next slides), some geared towards shoebox corpora
  - notable exception: Brown corpus
- 1980s and beyond: increased interest in corpus linguistics
  - opened new areas of research

Bad start for corpus linguistics

Noam Chomsky (1957) *Syntactic Structures*:

- p. 15: "... 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."

Chomsky recently


- "Corpus linguistics doesn’t mean anything. It’s like saying suppose a physicist decides, suppose physics and chemistry decide that instead of relying on experiments, what they’re going to do is take videotapes of things happening in the world and they’ll collect huge videotapes of everything that’s happening and from that maybe they’ll come up with some generalizations or insights."
Advantages
Responses to Chomsky

There are good points about the limitations of corpus-based research, but corpora should not be dismissed.

1. Existence in corpus ≠ grammatical.
   - Response: Intuition is necessary, but existence in corpus can point out new assumptions & reduce some biases (see next slide).
2. Finite corpus cannot capture all possible sentences.
   - Response: A corpus can supplement the sentences your brain can generate (show appropriate context).
3. Grammaticality is not statistical.
   - Response: This point is arguable (see later slide), and grammaticality is not everything (cf. language use).
4. Corpora are observational, not experimental.
   - Response: Both are worth investigating: controlled studies and real-world use.

Is language probabilistic?
Degrees of grammaticality

Are sentences completely grammatical or completely ungrammatical?

1. a. John I believe Sally said Bill believed Sue saw.
    b. What did Sally whisper that she had secretly read?
    c. Who did Jo think said John saw him?
    d. The boys read Mary’s stories about each other.
    e. I considered John as a good candidate.

⇒ Probabilistic modeling gives a degree of grammaticality

See Manning and Schütze (1999), Foundations of Statistical Natural Language Processing and Abney (1996), Statistical Methods and Linguistics.

Advantages
Corpus-based & Intuition-based approaches

Being empirical (i.e., using corpora [& experiments]) has advantages over intuition on its own:

- Intuition can be influenced by ideolect or dialect
  - corpus-based approach is free of overt judgments
- Intuition is based on a conscious monitoring of one’s production
  - generated sentences may not be typical language use
- Intuition-based examples are difficult to verify

Additionally, corpus-based approaches can show differences that intuition cannot provide

NB: Not every research question needs corpus data, so use judgment

More on objections to corpora

Spectrum of viewpoints on the usefulness of corpora

- Strong view: “without a corpus (or corpora) there is no meaningful work to be done” (Murison-Bowie 1996)
- Weak view: corpora provide viewpoints previously unavailable for language study & applications

It is rare for people to think that corpora are completely unusable

Is language probabilistic?
Language Acquisition, Change, and Variation

- Language Acquisition: child uses grammatical constructions with varying frequency
  - Rule possibilities are tried with different frequencies
- Language Change & Variation: gradual changes
  - Some proportion of the population uses new constructions
  - Dialect continua can make judgments difficult

Widdowson 2000

Benefits of corpora:

- Quantitative analysis from corpora is not accessible by intuition
  - regular patterns of collocational co-occurrence
  - 3rd person observed data different from 1st person intuition or 2nd person elicited data

Limitations of corpora:

- “it cannot represent the reality of first person awareness”
  - Corpora do not reveal what people know, nor what they think they know, about language
Other points from Widdowson

- Corpora are textual *products* which do not reveal the process underlying it
  - Decontextualized language which has to be re-contextualized for use in, e.g., language teaching
  - Corpora are texts, not discourse
- The “real” nature of corpus data may not fit the purpose at hand
  - It must be justified that, e.g., language learners should be exposed to real, native language

Process and product

Stubbs agrees that we can only view the product of language in a corpus
- But this is generally true in empirical disciplines (e.g., geology)
- Corpora can still provide a particular level of objectivity previously unavailable

Data and methods of corpus linguistics

- Corpora reveal what frequently and typically occurs
  - This is only a small proportion of what is possible
  - Corpora do not capture all possibilities: that’s why bigger corpora are always needed
- Corpora reveal how divergent intuition & usage can be
- Corpora can reveal both syntagmatic (co-occurrence) and paradigmatic (recurrence) relations