Corpus Linguistics (L615)

Why Corpus Linguistics?

Markus Dickinson
Department of Linguistics, Indiana University
Fall 2015
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 ...
A practical definition

A corpus provides texts in form of linguistically meaningful and retrievable units in a reusable way.

(from Kübler & Zinsmeister, 2014, Corpus Linguistics and Linguistically Annotated Corpora)
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 means many things to many people

▶ Tool: a way to gather relevant evidence
  ▶ Can be used for any layer of linguistic information
    (phonetic, phonological, morphological, syntactic, ...)
▶ Theory: an independent discipline in its own right
  ▶ Has an independent method & philosophical approach
to language analysis

Corpus linguistics has a wide range of applications
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?
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
Lexicography

How many senses does the word *line* have? 14 (*Webster’s New Encyclopedic Dictionary*, 1994):

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

According to Google (8/20/15):

261,000,000 hits for "think about"

8,560,000 hits for "think on"
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
A brief history

End of 1800s: Friedrich Wilhelm Kaeding wanted to improve shorthand
  ▶ Studied statistics of word, syllable, & character distributions
  ▶ Hundreds of volunteers examined 11 million words over the course of five years

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 & ways of looking at corpora
Approaches to corpora
Corpus-based & Corpus-driven

- **Corpus-based** research: corpora expound upon theories that were formulated before corpora
- **Corpus-driven** research: strictly committed to corpus data
Corpus-based vs. Corpus-driven

Differences

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
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.”
Bad start for corpus linguistics (2)

Noam Chomsky (1957) *Syntactic Structures*:

- p. 16/17: “...one’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 grammaticalness.”
Chomsky more 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.”
question: "Think of the occurrence of 'Can you . . .' or, 'Could you . . .' rather than 'Are you able to . . .' in polite requests in given communicative situations (a domain studied extensively by speech act theorists). Such chunks of linguistic expressions can be traced by the researcher via the application of corpus linguistic methods. It is from a corpus that one can identify their frequency and trace shifts in their meaning and use. Would you attribute significance to such data in your approach to linguistic analysis and description?"

answer: "People who work seriously in this particular area do not rely on corpus linguistics. They may begin by looking at facts about frequency and shifts in frequency and so on, but if they want to move on to some understanding of what’s happening they will very quickly, and in fact do, shift to the experimental framework. Where you design situations, you enquire into how people will act in those situations. You design them within a framework of theoretical inquiry which has already suggested that these are likely to be important questions and I want the answers to them. But that’s not corpus linguistics."
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 corpora 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.
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 intuition cannot provide

Not every research question needs corpus data: use judgment
Seeing how this plays out

Consider ditransitive verbs: when & why do people use each of the variants?

(1) a. Leslie gave the CD to Brent.
    b. Leslie gave Brent the CD.

What does it tell us if we find for *give*, in 50,000 sentences of newspaper text:

- 192 NP-NP sequences, &
- 44 NP-PP sequences?
Seeing how this plays out (2)

Hypothesis: NP-NP for change in state, NP-PP for change in space:

(2) a. That movie gave me the creeps.
   b. *That movie gave the creeps to me.

What do we then make of this corpus example?

(3) This story is designed to give the creeps to people who hate spiders, but is not true. [NP-PP]
A nice feature of corpus linguistics is that it forces one to be precise about a research question

- Operationalization translates a hypothesis into something that can be recovered from a corpus
- The results are then reusable

Corpora also allow one to naturally investigate variability & gradability in linguistic phenomena
There are several discussion points in the Teubert (2005) reading.

Some starting questions:

- What points were confusing?
- What points contradict traditional generative linguistics? What is your opinion of such points?
- What is your definition of meaning?
- What do you see as the goal of using corpus data?