Corpus Linguistics

Jean Mark Gawron

Linguistics 503
San Diego State University

January 18, 2013
Outline

1. Introduction
2. Limitations of armchair linguistics
3. Intrinsic problems
4. Corpus resources
5. Conclusion
What is a corpus?

Corpus

A corpus is a (usually computer readable) collection of spoken or written texts or conversations that is representative of a particular area of language use, by virtue of its size or composition.

Jespersen (1938)

I am above all an observer: I quite simply cannot help making linguistic observations. In conversations at home and abroad, in railway compartments, when passing people in streets and on roads, I am constantly noticing oddities of pronunciation, forms, and sentence constructions... For these notes I have found it practical to use small slips of paper...  

\(^a\)This passage is cited and translated by Jan Svartvik in Svartvik (1992).
General corpus

Basic properties

Representative of language as a whole (and therefore LARGE). Examples: Brown Corpus (Francis and Kučera 1964), British National Corpus.

1. Seeks balance: statistics of samples should reflect statistics of language as a whole.

2. Seeks completeness: All the major phenomena of the language should appear. Therefore large [BNC → 100 million words]
Specialized corpus

Hunston (2002)

... a corpus of texts of a particular types, such as newspaper editorials, geography textbooks, academic articles in a particular subject, lectures, casual conversations, essays written by students, etc. It aims to be representative of a given type of text. It is used to investigate a particular type of language.

1. Representative of a specific text type, such as the Wikipedia corpus (Denoyer and Gallinari 2006), register, genre, or population (dialect corpora, CHILDES acquisition corpus).

2. May also be representative of a single speaker/author.

3. In extreme cases, can be a single document (Prince 1992)
Corpus studies: What can be studied

1. Occurrence and re-occurrence of particular linguistic features (how and where do they occur?): Frequencies of particular items (words) or of sequences of items (ngrams, lexical bundles); Collocations: sets of words that typically occur together

2. Language of a particular domain: spoken academic discourse (MICASE, Michigan academic corpus of spoken English)

3. Language of a particular genre: university tutorial discussion (a fixed communicative purpose)

4. Language of a particular population (Survey of English Dialects), CHILDES

5. Translation (parallel corpora, Hansard, Europarl)

6. Grammar of language as a whole, or of a particular language type
Corpus studies: Who benefits

I. For the past 50 years, at least
   a. Applied linguistics: studies of specific text/speech types, pedagogically oriented studies
   b. Lexicography (at least since Johnson)
   c. Dialectology

II. Recent decades
   a. Descriptive linguistics
   b. Theoretical linguistics
   c. Language technology
   d. Social network studies [Enron email corpus]
Example corpus studies (general corpus)

a. Words that collocate with *girl* and *lady* (Sigley and Holmes 2002)

b. Compare the use of *hedges* (*kind of, sort of*) in English in general with their use in academic texts (Poos and Simpson 2002)

Two kinds of linguist

Caricatured

Fillmore (1992)
“... the armchair linguist ... sits in a deep soft comfortable armchair, with his eyes closed and his hands clasped behind his head. Once in a while, he opens his eyes, sits up abruptly shouting, ‘Wow, what a neat fact!', grabs his pencil, and writes something down.”
Two kinds of linguist

Caricatured

Fillmore (1992)

“... the armchair linguist ... sits in a deep soft comfortable armchair, with his eyes closed and his hands clasped behind his head. Once in a while, he opens his eyes, sits up abruptly shouting, ‘Wow, what a neat fact!’, grabs his pencil, and writes something down.”

“... the corpus linguist ... has all of the primary facts he needs, in the form of a corpus of one zillion running words, and he sees his job as that of deriving secondary facts from these primary facts. At the moment he is busy determining the frequencies of eleven parts of speech as the first word word of a sentence versus as the second word...”
Four sources of data

Placing what Labov calls “wooly-minded” introspection in context (Fillmore 1992)

<table>
<thead>
<tr>
<th>Type</th>
<th>Method</th>
<th>Primary data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Armchair</td>
<td>“wooly-minded” introspection</td>
<td>native speaker intuitions</td>
</tr>
<tr>
<td>Corpus</td>
<td>text/speech collection</td>
<td>text tokens</td>
</tr>
<tr>
<td>Experimental</td>
<td>experiment</td>
<td>experiment measurements</td>
</tr>
<tr>
<td>Simulation</td>
<td>computer simulation</td>
<td>computational performance</td>
</tr>
</tbody>
</table>
## Object of study

<table>
<thead>
<tr>
<th>Type</th>
<th>Object of study</th>
<th>Idealization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Armchair</td>
<td>Native speaker knowledge</td>
<td>competence</td>
</tr>
<tr>
<td>Corpus</td>
<td>Language in context</td>
<td>Various</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Created</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental</td>
<td>Language and brain</td>
<td>computational models</td>
</tr>
<tr>
<td>Simulation</td>
<td>Language and brain</td>
<td>computational models</td>
</tr>
<tr>
<td></td>
<td>Sociohistorical language patterns</td>
<td>speakers are “agents”</td>
</tr>
</tbody>
</table>
Other legitimate objects of study I

1. Linguistic properties of particular registers

Example

- Halliday and Matthiessen (1999)
- Dale (1990)
- Teng et al. (2012)

- Recipes and weather reports
- Computer recipe generator
- Building ingredient networks from recipes

2. Endangered languages or dead languages or older forms of living languages
Underlying human computational system (intuitions unavailable).

Example

*Speech errors: key evidence for computational system*
Other legitimate objects of study III

4 Discourse function/syntax interface: Much less accessible to native speaker intuition

Example

Prince (1978) Corpus-based study of It- and Wh-clefts
Prince (1992) Study of subject & definiteness based on a single letter
Bresnan et al. (2007) Corpus-based study of dative alternation
Other legitimate objects of study IV

Soft constraints/variation: Significant reproducible results for statistically dominant patterns

Example

Labov (1972)   Variation with social class
Bresnan et al. (2001)   Voice and person in English and Lummi
Jelinek and Demers (1983)
Other legitimate objects of study V

- Detailed analysis of lexical patterns (e.g., lexicography)

**Example**

**Johnson (1825)**
- Samuel Johnson’s dictionary plan

**James A. H. Murray**
- Oxford English Dictionary 1928 (4 million citation slips)

**George & Charles Merriam**
- Webster’s New International (1934, 2nd Ed.)

**Sinclair (1987)**
- Collins COBILD English Language dictionary
Experimental digression

Kaiser and Trueswell (2004)

1. Various online studies of flexible word order languages: Hyönnä and Hujanen (1997) showing noncanonical word orders are harder to process

2. Kaiser and Trueswell (2004) did an experiment which controlled for discourse context which showed that much of the difficulty in processing noncanonical word orders goes away in the appropriate discourse contexts.

3. Experimental method can play a role in many of the kinds of inquiry discussed above.
The role of experimentation

1. Can experiments play a role in historical linguistics?

2. Can experiments play a role in studying native speaker intuitions?
The role of experimentation

1. Can experiments play a role in historical linguistics?
   - de Boer (2000) presents a computational simulation using agents of the evolution of vowel systems

2. Can experiments play a role in studying native speaker intuitions?
The role of experimentation

1. Can experiments play a role in historical linguistics?
   - de Boer (2000) presents a computational simulation using agents of the evolution of vowel systems
   - Briscoe (2000) presents a computational model of acquisition for populations of agents from which some selectional pressures on language evolution emerge.

2. Can experiments play a role in studying native speaker intuitions?
The role of experimentation

1. Can experiments play a role in historical linguistics?
   1. de Boer (2000) presents a computational simulation using agents of the evolution of vowel systems.
   2. Briscoe (2000) presents a computational model of acquisition for populations of agents from which some selectional pressures on language evolution emerge.

2. Can experiments play a role in studying native speaker intuitions?
   Magnitude estimation is a technique originally used in psychophysics (for assigning measures to things like perceived loudness or perceived brightness). In the linguistic variant, subjects are asked to assign numbers reflecting their estimate of the degree of acceptability of various sentences (Bard et al. 1996).
We focus on the problem of characterizing *lexical competence*, in particular, on characterizing what we know when we know the meaning of a word, and what we know when we know how to use a word correctly.

Two problems arise: **Completeness** and **Correctness**.

We start with completeness, using the example of a complete account of the meaning of the word *risk*.

We illustrate correctness with the problem of knowing the syntax of the verb *give*.
Frame semantics

Word Meaning and event types

The meanings of most words can best be understood on the basis of a semantic frame: a description of a type of event, relation, or entity and the participants in it. For example, the concept of cooking typically involves the following:

\[
\begin{bmatrix}
\text{cook} & \text{person doing the cooking} \\
\text{food} & \text{the food that is to be cooked} \\
\text{heating\_instrument} & \text{source of heat} \\
\text{container} & \text{what holds the food during cooking}
\end{bmatrix}
\]

Words that evoke this frame:

fry, bake, boil, poach, and broil.
Completeness

Fillmore’s 1992 *risk* examples, illustrating elements of the *risk frame*.

**Risk frame**

- You would risk death doing what she did
- He decided to risk the venture
- Now he was prepared to risk his good name
Completeness

Fillmore’s 1992 *risk* examples, illustrating elements of the *risk frame*.

**Risk frame**

- You would risk **death** doing what she did
- He decided to risk the venture
- Now he was prepared to risk his good name
Completeness

Fillmore’s 1992 *risk* examples, illustrating elements of the *risk frame*.

Risk frame

You would risk death doing what she did

He decided to risk the *venture*

Now he was prepared to risk his good name

Harm

Deed
Completeness

Fillmore’s 1992 *risk* examples, illustrating elements of the *risk frame*.

**Risk frame**

You would risk death doing what she did

He decided to risk the venture

Now he was prepared to risk his good name

Harm

Deed

Valued Possession
Completeness

Fillmore’s 1992 *risk* examples, illustrating elements of the *risk frame*.

**Risk frame**

You would risk death doing what she did

He decided to risk the venture

Now he was prepared to risk his good name

Roosevelt risked fifty thousand dollars *in* Dakota ranch lands.

You risked a month’s earnings *on* that stupid horse!

The captain risked his ship *to* torpedo attack.

Harm

Deed

Valued Possession

Invest

Gamble

Expose
Correctness

To characterize the syntax of the word *give*, many linguists have assumed TWO meanings. Bresnan et al. (2007)

**Meaning to Structure Hypothesis**

causing a change of state (possession) $\Rightarrow$ V NP NP

Susan [\textit{\text{v}} gave] [\textit{\text{NP the children}}] [\textit{\text{NP toys}}].

cauing a change of place (movement to goal) $\Rightarrow$ V NP [to NP]

Susan [\textit{\text{v}} gave] [\textit{\text{NP toys}}] [\textit{\text{PP to the children}}].
Evidence

Examples 1

i. That movie gave me the creeps.

ii. *That movie gave the creeps to me.

iii. The lighting here gives me a headache.

iv. *The lighting here gives a headache to me.
Evidence

Examples 1

i. That movie gave me the creeps.

ii. *That movie gave the creeps to me.

iii. The lighting here gives me a headache.

iv. *The lighting here gives a headache to me.

Examples 2

v. I carried/pulled/pushed the box to John.

vi. *I carried/pulled/pushed John the box.
Counterexample 1

Corpus: The web (Bresnan et al. 2007)

i. . . .Orson Welles, who as the radio character, The Shadow, used to give the creeps to countless child listeners. . .

ii. This story is designed to give the creeps to people who hate spiders, but is not true.

iii. She found it hard to look at the Sages form for long. The spells that protected her identity also gave a headache to anyone trying to determine even her size...

iv. Design? Well, unless you take pride in giving a headache to your visitors with a flashing background? no.
Counterexample 2

a. Karen spoke with Gretchen about the procedure for registering a complaint, and hand-carried her a form, but Gretchen never completed it.

b. As player A pushed him the chips, all hell broke loose at the table.

c. Nothing like heart burn food. I have the tums. Nick joked. He pulled himself a steaming piece of the pie. Thanks for being here.

d. “Well. . . it started like this. . . “. Shinbo explained while Sumomo dragged him a can of beer and opened it for him, “We were having dinner together and. . . “
What is going on?

a. * That movie gave the creeps to me.
b. Stories like these must give the creeps to people whose idea of heaven is a world without religion.
c. ?? Stories like these must give people whose idea of heaven is a world without religion the creeps.
d. That movie gave me the creeps.

“the longer phrase is placed at the end by the principle of end weight (Wasow 2002)” give X the creeps has a strong bias toward V NP NP, but the principle of end weight can override that bias.

Multidimensionality/Context-dependence

The valid forms of a language are the results of compromises between many weighted constraints evaluated in context.
Summary thus far

- There are legitimate objects of study outside the scope of armchair linguistics (the object of study isn’t linguistic competence)

- **Even if** the object of study is linguistic knowledge (competence), different kinds of linguistic knowledge vary considerably in their accessibility to introspection: register, socially sensitive variables, discourse constraints on syntactic constructions

- Completeness: Generalizations we would never have found by consulting our intuitions

- Correctness: False generalizations we make because of our inability to manipulate all the many variables of linguistic acceptability in our heads.

- So even if you want to study competence, limiting yourself to armchair linguistics is a mistake.
Discussion

False dichotomy

Corpus resources: Getting started

- American Association for Corpus Linguistics Conference 2013
- Compling Lab
- Comp Ling Lab corpora
- Online Corpora
British National Corpus

The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of current British English, both spoken and written.

1. Tokenized, lemmatized

The dogs barked.

⇒

The/the dogs/dog barked/bark ./.
Part of speech tagged, supporting queries that use parts of speech.

The dogs barked.

⇒

The_DT/the dogs_NN2/dog barked_VBD/bark ./.

Sample texts, paragraphs, sentences, separated by XML boundaries

Text headers for extraction of subcorpora conforming to certain text types
Introduction

Limitations of armchair linguistics

Intrinsic problems

Corpus resources

Conclusion

References

XML: Typical document header

Jean Mark Gawron

Linguistics 503
San Diego State University
XML: Typical sentence

You need to involve your friends collecting jumble.

<s n="52">
  <w c5="PNP" hw="you" pos="PRON">You</w>
  <w c5="VVB" hw="need" pos="VERB">need</w>
  <w c5="TO0" hw="to" pos="PREP">to</w>
  <w c5="VVI" hw="involve" pos="VERB">involve</w>
  <w c5="DPS" hw="you" pos="PRON">your</w>
  <w c5="NN2" hw="friend" pos="SUBST">friends</w>
  <w c5="VVG" hw="collect" pos="VERB">collecting</w>
  <w c5="NN1" hw="jumble" pos="SUBST">jumble</w>
  <c c5="PUN">
</c>
</s>
More BNC info

David Lee’s genre classification scheme
BNC home page
BNCweb

BNCweb is a pretty intuitive web-based interface to BNC you can use with your web browser.

BNCweb in CompLing Lab
http://bulba.sdsu.edu/bncweb

Requirements
Password and account name
Simple queries

Queries

introvertedness [no matches]
introverted

Results

Your query "introverted" returned 84 hits in 69 different texts.

1 A06 1335 But besides this more obvious point, there are subtler connections between voice and body: Cicely Berry observes that an introverted and thoughtful person often finds more difficulty in speaking and does not carry the thought through into the physical process of making speech.

2 A18 429 Razumikhin himself may or may not have come from the country, but he is certainly a member of the floating, unbelonging population of students and ex-students, and he records in simple puzzlement that Raskolnikov has been growing increasingly moody and suspicious and introverted; he has no time for anything, people are always in his way, and yet he lies about and does nothing, a confirming echo of Raskolnikov on his bed telling Nastasya the maid that he is working, by which he means thinking.
Lemmas, intervening stuff I

Lemmas

\{kick/V\} curly braces signal lemma, \{kick/N\} and \{kick/V\} are different lemmas

\{kick\} All instances of all lemmas with the form kick

Intervening stuff

day ⟦⟩ 2 ⟦⟩ night day followed by night within a 2 word window (excludes by day and by night)
night ⟦⟩ 2 ⟦⟩ day day followed by night within a 2 word window, but night will be the highlighted word
day ⟦⟩ 2 ⟦⟩ night day and night within 2 words in either order
Lemmas, intervening stuff II

Plus and Star

\( \text{this} \ (-\{A\})^{*} \ \{\text{day}\} \)  \( \text{this} \) followed by any form of \( \text{day} \) with any number of adjectives intervening

\( \text{this} \ (-\{A\})^{+} \ \{\text{day}\} \)  \( \text{this} \) followed by any form of \( \text{day} \) with at least one adjective intervening

\( \text{this} + \ \text{day} \)  \( \text{this} \) followed by any form of \( \text{day} \) with exactly one word intervening

\( \text{this} + + \ \text{day} \)  \( \text{this} \) followed by any form of \( \text{day} \) with exactly two words intervening
Lemma/partof speech/sentences boundary

Query

kick/V <<s>> bucket_NN1

“All tokens of the LEMMA kick and the singular noun bucket occurring within a single sentence (in either order)”

Results

Your query "{kick/V} <<s>> bucket_NN1" returned 24 hits in 17 different texts

No Filename
1 A6W1120 At any speed, in any gear on the mile straight there is enough power to bury you hard into the thin bucket seats; every quick-fire gearchange though the massively solid Borg-Warner box kicks at the back end.
3 AC4 2431 Jinny was so startled that she nearly kicked the bucket over.
6 ATE 787 ‘‘Did you think I’d kicked the bucket, Ma?’’
<table>
<thead>
<tr>
<th>Hoffmann et al. (2008)</th>
<th>Book flyer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>Book extract</td>
</tr>
<tr>
<td>Getting started</td>
<td>Quick tutorial</td>
</tr>
<tr>
<td>ISBN</td>
<td>978-3-631-56315-1</td>
</tr>
</tbody>
</table>
Jean Mark Gawron

**Downloading results**

Your query "whether <<s>> (or not)" returned 5321 hits in 1941 different texts (98,313,429 words [4,048 texts]; frequency: 54.12 instances per million words)

<table>
<thead>
<tr>
<th>No</th>
<th>Filename</th>
<th>Hits 1 to 50</th>
<th>Page 1 / 107</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A04 1325</td>
<td>From a critic's point of view, a label, whether ending in 'ism' or not, is convenient.</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>A05 943</td>
<td>Shakespeare's play has an arranged duel which miscarries, and which takes off a divided, gambling man who has wondered whether or not it might be better to end his life.</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>A05 1138</td>
<td>Whether it is or not, the poem can be called distinctive — distinctive both of Larkin and of Amis.</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>A05 1318</td>
<td>If you were to tell me that there are people, like the man upstairs to whom you now threaten to turn yourself in, who actually do have a strong sense of themselves, I would have to tell you that they are only impersonating people with a strong sense of themselves — to which you could correctly reply that since there is no way of proving whether I'm right or not, this is a circular argument from which there is no escape.</td>
<td></td>
</tr>
</tbody>
</table>
Download concordance

Output format options

Choose operating system on which you will be working with the file: UNIX (incl. OS X)

Print codes (numbers) or full values for metatextual categories: Full values

Mark query result in sentence (format: <<< result >>>): Yes

Size of context: 1 <s>-unit

Download both tagged and untagged version of your results: Yes

Write information about order of categories at the beginning of file: No

Format of output: KWIC or list: List

Include corpus positions (required for re-import): Yes

Include URL to context display: Yes

Enter name for the downloaded file: gawron
From a critic’s point of view, a label, whether ending in 'ism' or not, is convenient.

Written books and periodicals W:ac:humanities_arts

1985-1993 Academic prose Beginning sample Book
Informative: Arts
High unknown UK and Ireland Male Sole
Adult Mixed Medium n/a n/a n/a n/a n/a n/a n/a n/a
n/a n/a n/a n/a n/a n/a
http://bulba.sdsu.edu/bncweb-cgi/context.pl?text=A04&qname=nosol&refnum=0&theData=0&len=0&showTheTag=0&color=0&begin=1325&spids=1&interval=11&first=yes&urlTest=yes
77845 77845
1 A04 1325 From a critic’s point of view, a label, whether ending in ‘ism’ or not, is convenient.
2 A05 943 Shakespeare’s play has an arranged duel which miscarries, and which takes off a divided, gambling man who has wondered whether or not it might be better to end his life.
3 A05 1138 Whether it is or not, the poem can be called distinctive — distinctive both of Larkin and of Amis.
4 A05 1318 If you were to tell me that there are people, like the man upstairs to whom you now threaten to turn yourself in, who actually do have a strong sense of themselves, I would have to tell you that they are only impersonating people with a strong sense of themselves; to which you could correctly reply that since there is no way of proving whether I’m right or not, this is a circular argument from which there is no escape.
FrameNet I

Intro

The FrameNet project is building a lexical database of English that is both human- and machine-readable, based on annotating examples of how words are used in actual texts. From the student’s point of view, it is a dictionary of more than 10,000 word senses, most of them with annotated examples that show the meaning and usage. For the researcher in Natural Language Processing, the more than 170,000 manually annotated sentences provide a unique training dataset for semantic role labeling, used in applications such as information extraction, machine translation, event recognition, and sentiment analysis.

FrameNet link
FrameNet II

Apply_Heat

In the FrameNet project, cooking event types are represented as a frame called **Apply_heat**, and the Cook, Food, Heating_instrument and Container are called **frame elements** (FEs).

\[
\begin{bmatrix}
\text{cook} & \text{person doing the cooking} \\
\text{food} & \text{the food that is to be cooked} \\
\text{heating_instrument} & \text{source of heat} \\
\text{container} & \text{what holds the food during cooking}
\end{bmatrix}
\]

Words that evoke this frame:

fry, bake, boil, poach, *and* broil.

Such words are called **lexical units** (LUs) of the Apply_heat frame.
An Avenger performs a Punishment on a Offender as a consequence of an earlier action by the Offender, the Injury. The Avenger inflicting the Punishment need not be the same as the Injured_Party who suffered the Injury. The Injured_Party can be an abstract concept such as honor.

Revenge

\[
\begin{array}{|c|}
\hline
\text{Avenger} & \text{person performing the punishment for the Injury} \\
\text{Offender} & \text{person performing the Injury} \\
\text{Injury} & \text{the wrong perpetrated by the offender} \\
\text{Injured_Party} & \text{person or abstract concept injured by the Injury} \\
\text{Punishment} & \text{action performed by the Avenger} \\
\hline
\end{array}
\]
Revenge frame II

LUs: avenge.v, avenger.n, get_back.(at).v, get_even.v, payback.n, retaliate.v, retaliation.n, retribution.n, retributive.a, retributory.a, revenge.n, revenge.v, revengeful.a, revenger.n, sanction.n, vengeance.n, vengeful.a, vindictive.a

i. They took revenge for the deaths of two men. Avenger Injury

ii. Lachlan sought to avenge them. Avenger Injured Party

iii. Later the Romans took revenge on their enemies. Avenger Offender
Summary

Two excellent reasons to use corpora

1. To do linguistic research that is not competence/grammar oriented
2. To do competence/grammar-based research in a more complete and correct way.

FrameNet and BNC (BNCweb) are existing corpus resources that provide tools for a variety of different kinds of corpus studies.

FrameNet and BNC annotate different and complementary kinds of information.
Conclusion

1. There are variety of reasons to stop being a linguist who does only armchair linguistics.

2. At the same time, linguistics that studies competence (the grammar in people’s heads) is alive and well, and corpus-based linguistics and armchair linguistics are not incompatible.

3. Important questions remain as to the content and design of corpora:
   - What kind of annotation should the corpus I use for my research contain?
   - What kind of data should the corpus I use for my research contain?
Course outline

I. Syntax and morphology
   a. Corpus linguistics motivations and methods
   b. Word structure
   c. Constituent structure
   d. Semantic roles and grammatical relations
   e. Lexical entries and well formed clauses

II. Phonetics/phonology
    a. Acoustic phonetics
    b. Cross linguistic phonetic variation
    c. Speech perception
    d. Phonological patterns

III. Information structure
    a. Topic/focus
    b. Given/new
Assignment, slide 1

the prefix *out-*

Consider the prefix *out-*, which attaches to verbs and produces a verb. Restrict your attention to cases in which the resulting verb is transitive, and in which the meaning of the prefixed verb involves comparing the subject and object on some scale relevant to the verb. Here are some examples:

1. *This bell outweighs that one.*
2. *The Jets outscored the Patriots.*
3. *The Texans outlasted Santa Anna.*

We call the prefix morpheme in these examples *comparative out-*.

Find more examples of this morpheme by doing a BNC web search to find all instances of verbs forms beginning with *out-*. Answer the questions on
Assignment, slide II

the following slides. Whenever making a claim about the data, give the entire example and the **Filename** of the example.

1. What was your query?
2. How many examples did your search return?
3. In some of examples returned, the highlighted *out*- word is not in fact a verb. Find one such part of speech error. (Note: There are are errors within the first 500 examples returned).
4. In all the sentences returned, was *out*- a morpheme? If not give examples.
5. In the cases in which *out*- is a morpheme, is it always comparative *out*-? That is, is it the same morpheme as in examples (i)-(iii) above? If not, give examples. If multiple examples exist, give at least three.
Evaluate the following claim: Comparative *out*- attaches to verbs which inherently involve something being measured and its measurement on some scale. We’ll call the thing being measured the **object** and we’ll call the result of the measurement the **measurement**. [inspired by some Framenet examples]. Here are some examples of inherent measurement verbs:

- Some bells weigh more than a ton
- The Jets scored 30 points
- The trip lasted 4 hours

The comparison described by the prefixed verb is always on this scale.
Assignment, slide IV

(outweigh, outscore, outlast). If you think this hypothesis is wrong, give 3 counterexamples from the data.

7 If the previous hypothesis is wrong, try to make some generalizations about where the scale used by out- is coming from. Give examples.

8 Evaluate the original hypothesis, Hypothesis A. Is it right? Is it complete and correct? Is it specific?
References


References II

Grammatical acquisition: Inductive bias and coevolution of language and the language acquisition device.

Dale, R. 1990.
Generating recipes: An overview of epicure.

de Boer, Bart. 2000.
Self organization in vowel systems.


References IV


Effects of case marking and word order on sentence parsing in finnish: An eye fixation analysis.

The agent hierarchy and voice in some coast salish languages.

Jespersen, O. 1938.
*En Sprogmands levned (Memoirs).*
Copenhagen: Gyldendal.
This is Jespersen’s autobiography.
References VI

Johnson, Samuel. 1825.
   Plan of an english dictionary.

   The role of discourse context in the processing of a flexible word-order language.

   Sociolinguistic patterns.
References VII

Building a large annotated corpus of english: The penn treebank. 

Cross-disciplinary comparisons of hedging: Some findings from the michigan corpus of academic spoken english.
In Reppen et al. (Reppen et al. 2002), 3–23.

The zpg letter: Subjects, definiteness, and information-status.
References VIII

Prince, Ellen. 1978.
A comparison of wh-clefts and it-clefts in discourse. 

The linguist and the English language.
London: Edward Arnold.

*Using corpora to explore linguistic variation.*
Amsterdam: John Benjamins.

Looking at *girls* in corpora of english. 
References IX

*Looking up: An account of the COBUILD project in lexical computing and the development of the Collins COBUILD English language dictionary.*  
Collins ELT.

*Directions in corpus linguistics: Proceedings of the Nobel symposium 82, New York.* Mouton de Gruyter.

Recipe recommendation using ingredient networks.  
In *Proc. 4th International ACM Web Science Conference, 447–456, Evanston.* ACM.
*Postverbal Behavior.*
Stanford: CSLI.