Corpus Linguistics

John Gamboa

Resources

This discussions is mostly based on

- Gries, S. T. (2009). What is corpus linguistics?. *Language and linguistics compass*, 3(5), 1225-1241.
- Gries, S. T. (2012). Corpus linguistics, theoretical linguistics, and cognitive/psycholinguistics: Towards more and more fruitful exchanges. In *Corpus Linguistics and Variation in English* (pp. 41-63). Brill Rodopi.

A lot is also based on the Chapter 2 of the NLTK book.

Resources

Example applications of Corpus Linguistics can be found in

- J Garcia-Lopez, L., B Díez-Bedmar, M., Perez-Paredes, P., & Tornero, E. (2011). Treatment change in adolescents with social anxiety disorder: Insights from corpus linguistics. *Ansiedad y Estrés*, 17.
- Al-Mosaiwi, M., & Johnstone, T. (2018). In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical Psychological Science*, *6*(4), 529-542.

What is Corpus Linguistics?

What is Corpus Linguistics?

- A relatively new field

What is Corpus Linguistics?

- A relatively new field

- For some people, just a methodology used in Linguistics

What is Corpus Linguistics?

- A relatively new field
- For some people, just a methodology used in Linguistics
- For others

What is Corpus Linguistics?

- A relatively new field

- For some people, just a methodology used in Linguistics

- For others
  - *Computer Corpus Linguistics* "defines not just a newly emerging methodology for studying language, but a new research enterprise, and in fact a new philosophical approach to the subject"

    (Leech, G. (1992). Corpora and theories of linguistic performance. *Directions in corpus linguistics*, 105-122)

What is Corpus Linguistics?

What is Corpus Linguistics?

- Following Gries (2009), Corpus Linguistics is

What is Corpus Linguistics?

- Following Gries (2009), Corpus Linguistics is

  A *methodology* that analyses **corpora** to **address linguistic questions**

What is Corpus Linguistics?

- Following Gries (2009), Corpus Linguistics is

A *methodology* that analyses **corpora** to **address linguistic questions**

What is Corpus Linguistics?

- Following Gries (2009), Corpus Linguistics is

  A *methodology* that analyses **corpora** to **address linguistic questions**

What is Corpus Linguistics?

- Following Gries (2009), Corpus Linguistics is

A *methodology* that analyses **corpora** to **address linguistic questions**

What is a corpus?

What is Corpus Linguistics?

- Following Gries (2009), Corpus Linguistics is

A *methodology* that analyses **corpora** to **address linguistic questions**

What is a corpus?

How?

What is Corpus Linguistics?

- Following Gries (2009), Corpus Linguistics is

A *methodology* that analyses **corpora** to **address linguistic questions**

What is a corpus?

How?

Corpus (plural: Corpora)

Corpus (plural: Corpora)

- A collection of “texts”

Corpus (plural: Corpora)

- A collection of ~~“texts”~~ language

Corpus (plural: Corpora)

- A collection of ~~“texts”~~ language
- Representative

Corpus (plural: Corpora)

- A collection of ~~“texts”~~ language

- Representative
  - “*different parts of the linguistic variety I’m interested in are all manifested in the corpus*”

Corpus (plural: Corpora)

- A collection of ~~“texts”~~ language

- Representative
  - “*different parts of the linguistic variety I’m interested in are all manifested in the corpus*”

Corpus (plural: Corpora)

- A collection of ~~“texts”~~ language

- Representative
  - “*different parts of the linguistic variety I’m interested in are all manifested in the corpus*”

Example goal:
understand phonological reduction by Californian adolescents

Corpus (plural: Corpora)

- A collection of “~~texts~~” language

- Representative
  - *“different parts of the linguistic variety I’m interested in are all manifested in the corpus”*

Example goal:
understand phonological reduction by Californian adolescents

- *Include conversations of adolescents with their peer group?*

Corpus (plural: Corpora)

- A collection of ~~“texts”~~ language

- Representative
  - “*different parts of the linguistic variety I’m interested in are all manifested in the corpus*”

Example goal:
understand phonological reduction by Californian adolescents

- *Include conversations of adolescents with their peer group?*
- *Include their conversations with parents, teachers, … ?*

Corpus (plural: Corpora)

Corpus (plural: Corpora)

- Balanced

Corpus (plural: Corpora)

- Balanced
  - “*the proportion with which a particular part is represented in a corpus should reflect the proportion the part makes up in this variety and/or the importance of the part in this variety*”

Corpus (plural: Corpora)

- Balanced
  - “*the proportion with which a particular part is represented in a corpus should reflect the proportion the part makes up in this variety and/or the importance of the part in this variety*”

Corpus (plural: Corpora)

- Balanced
    - “*the proportion with which a particular part is represented in a corpus should reflect the proportion the part makes up in this variety and/or the importance of the part in this variety*”

Californian adolescents

Corpus (plural: Corpora)

- Balanced
    - “*the proportion with which a particular part is represented in a corpus should reflect the proportion the part makes up in this variety and/or the importance of the part in this variety*”

Californian adolescents

- How much of the corpus should consist of speech?

Corpus (plural: Corpora)

- Balanced
  - “*the proportion with which a particular part is represented in a corpus should reflect the proportion the part makes up in this variety and/or the importance of the part in this variety*”

Californian adolescents

- How much of the corpus should consist of speech?
- How much should consist of text?

Corpus (plural: Corpora)

- Balanced
    - “*the proportion with which a particular part is represented in a corpus should reflect the proportion the part makes up in this variety and/or the importance of the part in this variety*”

Californian adolescents

- How much of the corpus should consist of speech?
- How much should consist of text?
- How much should be conversations with peers?

Corpus (plural: Corpora)

- Balanced
  - “*the proportion with which a particular part is represented in a corpus should reflect the proportion the part makes up in this variety and/or the importance of the part in this variety*”

Californian adolescents

- How much of the corpus should consist of speech?
- How much should consist of text?
- How much should be conversations with peers?
- How much should be conversations with parents?

Corpus (plural: Corpora)

- Balanced
  - “*the proportion with which a particular part is represented in a corpus should reflect the proportion the part makes up in this variety and/or the importance of the part in this variety*”

Californian adolescents

- How much of the corpus should consist of speech?
- How much should consist of text?
- How much should be conversations with peers?
- How much should be conversations with parents?

...

Corpus (plural: Corpora)

- Balanced
  - “*the proportion with which a particular part is represented in a corpus should reflect the proportion the part makes up in this variety and/or the importance of the part in this variety*”

- Naturally occurring

Corpus (plural: Corpora)

- Balanced
  - “*the proportion with which a particular part is represented in a corpus should reflect the proportion the part makes up in this variety and/or the importance of the part in this variety*”

- Naturally occurring
  - “*the texts were spoken or written for some authentic communicative purpose, but not for the purpose of putting them into a corpus*”

Additional Corpus data

Additional Corpus data

- **Metadata:**

Additional Corpus data

- **Metadata:**

  “identifies the source(s) of the data in the corpus file...”

Additional Corpus data

- **Metadata:**

  “identifies the source(s) of the data in the corpus file...”

  “... number of interlocutors…”

Additional Corpus data

- **Metadata:**

  “identifies the source(s) of the data in the corpus file...”

  “... number of interlocutors…”

  “... age, profession, and highest degree”

Additional Corpus data

- **Metadata:**

  “identifies the source(s) of the data in the corpus file...”

  “... number of interlocutors…”

  “... age, profession, and highest degree”

  copyright information

Additional Corpus data

- **Metadata:**

  “identifies the source(s) of the data in the corpus file...”

  “... number of interlocutors…”

  “... age, profession, and highest degree”

  copyright information

  etc.

Additional Corpus data

Additional Corpus data

- **Annotations**

Additional Corpus data

- **Annotations**

  Word classes (article, adjective, verb, …)

Additional Corpus data

- **Annotations**

  Word classes (article, adjective, verb, …)

  Stem

Additional Corpus data

- **Annotations**

  Word classes (article, adjective, verb, …)

  Stem

  Passive/active voice

Additional Corpus data

- **Annotations**

  Word classes (article, adjective, verb, …)

  Stem

  Passive/active voice

  Constituent structure of the sentences

Additional Corpus data

- **Annotations**

  Word classes (article, adjective, verb, …)

  Stem

  Passive/active voice

  Constituent structure of the sentences

  ...

Types of Corpora

Types of Corpora

general:

vs.

specific:

Types of Corpora

general: *“representative and balanced for the language as a whole”*

vs.

specific:

Types of Corpora

general: *“representative and balanced for the language as a whole”*

vs.

specific: *“by design restricted to a particular variety, register, genre, …”*

Types of Corpora

general: *“representative and balanced for the language as a whole”*

vs.

specific: *“by design restricted to a particular variety, register, genre, …”*

diachronic:

vs.

synchronic:

Types of Corpora

general: *“representative and balanced for the language as a whole”*

vs.

specific: *“by design restricted to a particular variety, register, genre, …”*

diachronic: *captures “how a language/variety changes over time”*

vs.

synchronic:

Types of Corpora

general: *“representative and balanced for the language as a whole”*
vs.
specific: *“by design restricted to a particular variety, register, genre, …”*

diachronic: *captures “how a language/variety changes over time”*
vs.
synchronic: *“snapshot of a language/variety at one particular point of time”*

Types of Corpora

Types of Corpora

monolingual:

vs.

parallel:

Types of Corpora

monolingual: *“compiled to provide information about one particular language/variety”*

vs.

parallel:

Types of Corpora

monolingual: *“compiled to provide information about one particular language/variety”*

vs.

parallel: *“ideally provide the same text in several different languages”*

Types of Corpora

monolingual: *“compiled to provide information about one particular language/variety”*

vs.

parallel: *“ideally provide the same text in several different languages”*

static:

vs.

dynamic/monitor:

Types of Corpora

monolingual: *“compiled to provide information about one particular language/variety”*

vs.

parallel: *“ideally provide the same text in several different languages”*

static: *“have a fixed size”*

vs.

dynamic/monitor:

Types of Corpora

monolingual: *“compiled to provide information about one particular language/variety”*

vs.

parallel: *“ideally provide the same text in several different languages”*

static: *“have a fixed size”*

vs.

dynamic/monitor: *“may be constantly extended with new material”*

Types of Corpora

Types of Corpora

raw:

vs.

annotated:

Types of Corpora

raw: *“files only containing the corpus material … (and maybe metadata and markup)”*

vs.

annotated:

Types of Corpora

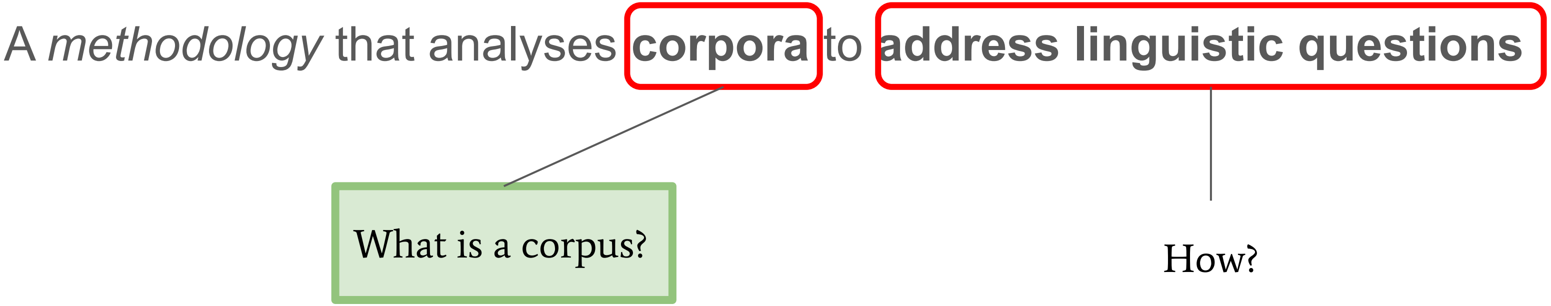
raw: *“files only containing the corpus material … (and maybe metadata and markup)”*

vs.

annotated: *“also contain information about the language data in the corpus part, information that represents a particular linguistic analysis”*

What is Corpus Linguistics?

- Following Gries (2009), Corpus Linguistics is



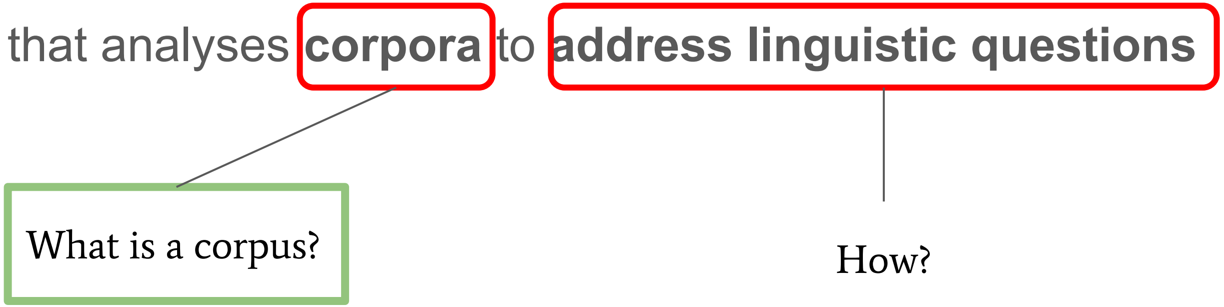
A *methodology* that analyses **corpora** to **address linguistic questions**

What is a corpus?

How?

What is Corpus Linguistics?

- Following Gries (2009), Corpus Linguistics is



A *methodology* that analyses **corpora** to **address linguistic questions**

What is a corpus?

How?

What is Corpus Linguistics?

- Following Gries (2009), Corpus Linguistics is

A *methodology* that analyses **corpora** to **address linguistic questions**

What is a corpus?

How?

Answering linguistic questions

Answering linguistic questions

- It’s all about frequencies

Answering linguistic questions

- It’s all about frequencies
    - *“corpora are (usually text) files and all you can get out of such files is distributional (or quantitative/statistical) information”*

Answering linguistic questions

- It’s all about frequencies
  - *“corpora are (usually text) files and all you can get out of such files is distributional (or quantitative/statistical) information”*

- **“Frequencies of what?”**

Answering linguistic questions

- It’s all about frequencies
  - *“corpora are (usually text) files and all you can get out of such files is distributional (or quantitative/statistical) information”*

- **“Frequencies of what?”**
  - ***occurrences*** of linguistic elements (e.g., words, syntactic structures, …)

Answering linguistic questions

- It’s all about frequencies
  - *“corpora are (usually text) files and all you can get out of such files is distributional (or quantitative/statistical) information”*

- **“Frequencies of what?”**
  - ***occurrences*** of linguistic elements (e.g., words, syntactic structures, …)
  - ***co-occurences*** of these elements

Answering linguistic questions

- It’s all about frequencies
  - *“corpora are (usually text) files and all you can get out of such files is distributional (or quantitative/statistical) information”*

- **“Frequencies of what?”**
  - ***occurrences*** of linguistic elements (e.g., words, syntactic structures, …)
  - ***co-occurences*** of these elements

- In particular, you often analyse:

Answering linguistic questions

- It’s all about frequencies
  - *“corpora are (usually text) files and all you can get out of such files is distributional (or quantitative/statistical) information”*

- **“Frequencies of what?”**
  - ***occurrences*** of linguistic elements (e.g., words, syntactic structures, …)
  - ***co-occurences*** of these elements

- In particular, you often analyse:
  - ***Existence***: is a frequency 0 or larger?

Answering linguistic questions

- It’s all about frequencies
    - *“corpora are (usually text) files and all you can get out of such files is distributional (or quantitative/statistical) information”*

- **“Frequencies of what?”**
    - ***occurrences*** of linguistic elements (e.g., words, syntactic structures, …)
    - ***co-occurences*** of these elements

- In particular, you often analyse:
    - ***Existence***: is a frequency 0 or larger?
    - ***Relative frequencies***:

Answering linguistic questions

- It’s all about frequencies
    - *“corpora are (usually text) files and all you can get out of such files is distributional (or quantitative/statistical) information”*

- **“Frequencies of what?”**
    - ***occurrences*** of linguistic elements (e.g., words, syntactic structures, …)
    - ***co-occurences*** of these elements

- In particular, you often analyse:
    - ***Existence***: is a frequency 0 or larger?
    - ***Relative frequencies***:
        - Is a frequency larger/smaller than you’d expect by chance?

Answering linguistic questions

- It’s all about frequencies
    - *“corpora are (usually text) files and all you can get out of such files is distributional (or quantitative/statistical) information”*

- **“Frequencies of what?”**
    - ***occurrences*** of linguistic elements (e.g., words, syntactic structures, …)
    - ***co-occurences*** of these elements

- In particular, you often analyse:
    - ***Existence***: is a frequency 0 or larger?
    - ***Relative frequencies***:
        - Is a frequency larger/smaller than you’d expect by chance?
        - Is a frequency larger than another?

From Statistics to Linguistics

From Statistics to Linguistics

- This “conceptual leap” has been stated with different names

From Statistics to Linguistics

- This “conceptual leap” has been stated with different names
    - **Bohlinger (1968):** *“A difference in syntactic form always spells a difference in meaning”*

From Statistics to Linguistics

- This “conceptual leap” has been stated with different names
  - **Bohlinger (1968):** *“A difference in syntactic form always spells a difference in meaning”*
  - **Goldberg (1995):** *“If two constructions are syntactically distinct, they must be semantically or pragmatically distinct”*

(references from Gries, 2009)

From Statistics to Linguistics

- This “conceptual leap” has been stated with different names
  - **Bohlinger (1968):** *“A difference in syntactic form always spells a difference in meaning”*
  - **Goldberg (1995):** *“If two constructions are syntactically distinct, they must be semantically or pragmatically distinct”*

    (references from Gries, 2009)

“different frequencies of (co-)occurrences of formal elements … are assumed to reflect functional regularities, and ‘functional’ is … anything … that is intended to perform a particular communicative function”

This conceptual leap is everywhere

This conceptual leap is everywhere

- First language acquisition

This conceptual leap is everywhere

- First language acquisition
- Phonology

This conceptual leap is everywhere

- First language acquisition
- Phonology
- Morphology

This conceptual leap is everywhere

- First language acquisition
- Phonology
- Morphology
- Syntax

This conceptual leap is everywhere

- First language acquisition
- Phonology
- Morphology
- Syntax
- Semantics/pragmatics

This conceptual leap is everywhere

- First language acquisition
- Phonology
- Morphology
- Syntax
- Semantics/pragmatics
- …

This conceptual leap is everywhere

- First language acquisition
- Phonology
- Morphology
- Syntax
- Semantics/pragmatics
- …

We control for all sorts of frequencies in psycholinguistic studies