

Corpus Tools: Introduction to corpora (and to the course)

Max Ionov
max.ionov@gmail.com

Goethe-Universität Frankfurt

24.03.2017

Table of Contents

- 1 Course information
- 2 What is a corpus and why do we need it
- 3 Corpus properties
- 4 Types of corpora

Course information

Topics:

- 1 Corpora basics: information, properties, types and examples
- 2 Practical matters: getting texts, search methodologies, using corpora

Course information

Topics:

- 1 Corpora basics: information, properties, types and examples
- 2 Practical matters: getting texts, search methodologies, using corpora
- 3 Corpus tools: UAM Corpus Tools, BRAT, and others

Course information

Topics:

- 1 Corpora basics: information, properties, types and examples
- 2 Practical matters: getting texts, search methodologies, using corpora
- 3 Corpus tools: UAM Corpus Tools, BRAT, and others
- 4 Corpus research: examples of how to do corpus research
- 5 FLex tutorial

Course information

Topics:

- 1 Corpora basics: information, properties, types and examples
- 2 Practical matters: getting texts, search methodologies, using corpora
- 3 Corpus tools: UAM Corpus Tools, BRAT, and others
- 4 Corpus research: examples of how to do corpus research
- 5 FLex tutorial

Seminar task:

- Create a (small) corpus

Course information

Topics:

- 1 Corpora basics: information, properties, types and examples
- 2 Practical matters: getting texts, search methodologies, using corpora
- 3 Corpus tools: UAM Corpus Tools, BRAT, and others
- 4 Corpus research: examples of how to do corpus research
- 5 FLex tutorial

Seminar task:

- Create a (small) corpus
- Annotate it (a part of it)

Course information

Topics:

- 1 Corpora basics: information, properties, types and examples
- 2 Practical matters: getting texts, search methodologies, using corpora
- 3 Corpus tools: UAM Corpus Tools, BRAT, and others
- 4 Corpus research: examples of how to do corpus research
- 5 FLex tutorial

Seminar task:

- Create a (small) corpus
- Annotate it (a part of it)
- Demonstrate its possibilities with a few queries (a small corpus study)

Table of Contents

- 1 Course information
- 2 What is a corpus and why do we need it
- 3 Corpus properties
- 4 Types of corpora

What

- Language learning

What

- Language learning
- Language studying (studies of some phenomena in language(s))

What

- Language learning
- Language studying (studies of some phenomena in language(s))
- Language description

What

- Language learning
- Language studying (studies of some phenomena in language(s))
- Language description
- Language processing (computational linguistics)

What

- Language learning
 - Language studying (studies of some phenomena in language(s))
 - Language description
 - Language processing (computational linguistics)
- ⇒ There is a need for corpora

Data for corpora

- Annotation

Data for corpora

- Annotation
- Search functionality

Data for corpora

- Annotation
- Search functionality
- A (graphical) interface to use it

Data for corpora

- Annotation
- Search functionality
- A (graphical) interface to use it
- In reality, not necessarily :)

What is corpus data

- Two approaches:
language use vs. language competence \approx experiment vs. observation
(\approx E-language vs. I-language)

What is corpus data

- Two approaches:
language use vs. language competence \approx experiment vs. observation
(\approx E-language vs. I-language)
- Experiment:
 - Reproducibility
 - Control
 - Reliability (stability)

What is corpus data

- Two approaches:
language use vs. language competence \approx experiment vs. observation
(\approx E-language vs. I-language)
- Experiment:
 - Reproducibility
 - Control
 - Reliability (stability)
- Observation:
 - Non-reproducibility
 - No negative data

What is corpus data

- Two approaches:
language use vs. language competence \approx experiment vs. observation
(\approx E-language vs. I-language)
- Experiment:
 - Reproducibility
 - Control
 - Reliability (stability)
- Observation:
 - Non-reproducibility
 - No negative data

Corpora contains observed data

What is corpus data

- Two approaches:
language use vs. language competence \approx experiment vs. observation
(\approx E-language vs. I-language)
- Experiment:
 - Reproducibility
 - Control
 - Reliability (stability)
- Observation:
 - Non-reproducibility
 - No negative data

Corpora contains observed data

Still, the data may be from an experiment

Observation: cons

- Language is infinite, corpora are finite

Observation: cons

- Language is infinite, corpora are finite
- No introspection

Observation: cons

- Language is infinite, corpora are finite
- No introspection
- It's harder (or more expensive)

Observation: pros

- Research can be verified

Observation: pros

- Research can be verified
- Statistical metrics

Observation: pros

- Research can be verified
- Statistical metrics
- Search is easier

Table of Contents

- 1 Course information
- 2 What is a corpus and why do we need it
- 3 Corpus properties**
- 4 Types of corpora

Size

- Consists of texts

Size

- Consists of texts
- What is an elementary unit?

Size

- Consists of texts
- What is an elementary unit?
 - Not text

Size

- Consists of texts
- What is an elementary unit?
 - Not text
 - Not page

Size

- Consists of texts
- What is an elementary unit?
 - Not text
 - Not page
 - \Rightarrow A word

Words in a corpus

- Language contains:
 - Wordforms
 - Lexemes

Words in a corpus

- Language contains:
 - Wordforms
 - Lexemes
- A text consist of words

Words in a corpus

- Language contains:
 - Wordforms
 - Lexemes
- A text consist of words \approx wordforms

Words in a corpus

- Language contains:
 - Wordforms
 - Lexemes
- A text consist of words \approx wordforms
 - A wordform — element of language

Words in a corpus

- Language contains:
 - Wordforms
 - Lexemes
- A text consist of words \approx wordforms
 - A wordform — element of language
 - A word — element of a text

Size

1-million-word corpus — is it enough?

Таблица 1: The presence of a lexeme *imaginable* in corpora of various sizes (in millions of words)

Size	Corpus	Absolute frequency	Words per million
1	Brown Corpus	0	0
1	Bible	0	0
2	Shakespeare	0	0
7	World Street Journal	41	5.9
18	Hansard	15	0.8

Size

What is 1 million of words

Size

What is 1 million of words

- A standard book page contains approx. 215 words

Size

What is 1 million of words

- A standard book page contains approx. 215 words
- 64 500 in a book that is 300 pages long

Size

What is 1 million of words

- A standard book page contains approx. 215 words
- 64 500 in a book that is 300 pages long
- *Approx. 15 books* of that size

Requirements for a corpus

Requirements for a corpus

- Representativeness

Requirements for a corpus

- Representativeness
- Coverage

Requirements for a corpus

- Representativeness
- Coverage
- Economy

Requirements for a corpus

- Representativeness
- Coverage
- Economy
- Structure

Requirements for a corpus

- Representativeness
- Coverage
- Economy
- Structure
- Computer aid to use a corpus

Requirements: Representativeness

- A corpus is a subset of language

Requirements: Representativeness

- A corpus is a subset of language
- It should contain **all relevant phenomena** (relevant for a specific research)

Requirements: Representativeness

- A corpus is a subset of language
- It should contain **all relevant phenomena** (relevant for a specific research)
- A frequency of a phenomena in a corpus should be the same as its frequency *in general*

National corpora and representativeness

- Genres
- Styles
- Time periods
- Authors

National corpora and representativeness

- Genres
- Styles
- Time periods
- Authors

In this case, representativeness = balance

National corpora and representativeness

Douglas Biber: fractions of language types in a corpus

National corpora and representativeness

Douglas Biber: fractions of language types in a corpus

- 90% speech
- 3% notes and letters
- 7% published texts of various genres

National corpora and representativeness

It's not the case usually

National corpora and representativeness

It's not the case usually

- Fiction is easier to find and add

National corpora and representativeness

It's not the case usually

- Fiction is easier to find and add
- 20–40% fiction
- The rest depends on what is present

National corpora and representativeness

It's not the case usually

- Fiction is easier to find and add
- 20–40% fiction
- The rest depends on what is present

For mostly written corpora it should be at least like this:

National corpora and representativeness

It's not the case usually

- Fiction is easier to find and add
- 20–40% fiction
- The rest depends on what is present

For mostly written corpora it should be at least like this:

- Limit the time frame (synchronic corpus)

National corpora and representativeness

It's not the case usually

- Fiction is easier to find and add
- 20–40% fiction
- The rest depends on what is present

For mostly written corpora it should be at least like this:

- Limit the time frame (synchronic corpus)
- Choose a dimension for a representativeness:

National corpora and representativeness

It's not the case usually

- Fiction is easier to find and add
- 20–40% fiction
- The rest depends on what is present

For mostly written corpora it should be at least like this:

- Limit the time frame (synchronic corpus)
- Choose a dimension for a representativeness:
 - Styles / genres
 - Specific language phenomena

National corpora and representativeness

It's not the case usually

- Fiction is easier to find and add
- 20–40% fiction
- The rest depends on what is present

For mostly written corpora it should be at least like this:

- Limit the time frame (synchronic corpus)
- Choose a dimension for a representativeness:
 - Styles / genres
 - Specific language phenomena
- We can limit this to 'culturally important texts'

Representativeness in national corpora

But: author's language vs. normal language:

Representativeness in national corpora

But: author's language vs. normal language:

- Culturally important texts are not *normal*

Representativeness in national corpora

But: author's language vs. normal language:

- Culturally important texts are not *normal*
- 'Real' language is not represented in texts

Representativeness in national corpora

But: author's language vs. normal language:

- Culturally important texts are not *normal*
- 'Real' language is not represented in texts
- Example: corpus study of discourse markers

Representativeness in national corpora

But: author's language vs. normal language:

- Culturally important texts are not *normal*
- 'Real' language is not represented in texts
- Example: corpus study of discourse markers
- *well*

Representativeness in national corpora

But: author's language vs. normal language:

- Culturally important texts are not *normal*
- 'Real' language is not represented in texts
- Example: corpus study of discourse markers
- *well*
- *like*

Requirements: Coverage

- A phenomena under research should be presented *fully*

Requirements: Coverage

- A phenomena under research should be presented *fully*
- Including very rare cases

Requirements: Coverage

- A phenomena under research should be presented *fully*
- Including very rare cases
- ... even if this contradicts the representativeness requirement

Requirements: Economy

- Corpus is a subset

Requirements: Economy

- Corpus is a subset
- The most representative corpus is a language itself

Requirements: Economy

- Corpus is a subset
- The most representative corpus is a language itself
- **But:** It is hard (or even impossible) to work with the 'full language'

Requirements: Economy

- Corpus is a subset
- The most representative corpus is a language itself
- **But:** It is hard (or even impossible) to work with the 'full language'
- Corpus should be economic

Requirements: Structure

- Corpus → subcorpora

Requirements: Structure

- Corpus → subcorpora
- Structure depends on a goal

Requirements: Structure

- Corpus → subcorpora
- Structure depends on a goal
 - time frames
 - authors

Requirements: Structure

- Corpus → subcorpora
- Structure depends on a goal
 - time frames
 - authors
 - text topics (News360)

Requirements: Computer aid

Two sides:

Requirements: Computer aid

Two sides:

- Tools for using a corpus

Requirements: Computer aid

Two sides:

- Tools for using a corpus
- Format of a corpus

Requirements: Computer aid

Two sides:

- Tools for using a corpus
- Format of a corpus
- There are many standards

Requirements: Computer aid

Two sides:

- Tools for using a corpus
- Format of a corpus
- There are many standards
- Using standards makes reuse possible

Table of Contents

- 1 Course information
- 2 What is a corpus and why do we need it
- 3 Corpus properties
- 4 Types of corpora**

Depends on annotation

Depends on annotation

- 1 Morphological

Depends on annotation

- 1 Morphological
- 2 Syntactic

Depends on annotation

- 1 Morphological
- 2 Syntactic
- 3 Coreference

Depends on annotation

- 1 Morphological
- 2 Syntactic
- 3 Coreference
- 4 Corpora of language errors

Depends on annotation

- 1 Morphological
- 2 Syntactic
- 3 Coreference
- 4 Corpora of language errors
- 5 ...