Text Corpora and Linguistic Annotation

Marcos Garcia
marcos.garcia.gonzalez@udc.gal
grupolys.org/~marcos

LyS Group, UdC

Contents

Contents

Introduction

- Objectives:
  - Text corpora.
  - Brief introduction to corpus linguistics.
  - Types of corpora.
  - Corpus design and annotation.
  - Utilities: corpus linguistics and NLP.

What is a corpus?

- What is a corpus?

What is a corpus?

- What is a corpus?
  - Group of texts in a machine-readable format.
  - For linguistic analysis.
  - For natural language processing (NLP).
  - For other purposes.
- Text *versus* Speech corpora.

Contents

Corpus Linguistics: a brief introduction

- Corpus linguistics (CL) as a methodology for language analysis.
  - Allows linguists to observe *real language data*.
  - Fast extraction of statistical information.

Corpus Linguistics: a brief introduction

- Corpus linguistics (CL) as a methodology for language analysis.
  - Allows linguists to observe *real language data*.
  - Fast extraction of statistical information.
- *Early corpus linguistics:*
  - Many linguistic studies used corpora before the consolidation of CL.
  - Käding, 1897: German corpus of 11 milion words.
  - Studies in language acquisition and learning, pedadogy, etc.

Generativism and Corpus Linguistics

- Chomsky, 1957. *Syntactic Structures*.
- Generative linguistics:
  - Rationalism *versus* empiricism.
  - Competence *versus* performance.
  - Introspection *versus* data analysis.

Generativism and Corpus Linguistics

- Chomsky, 1957. *Syntactic Structures*.
- Generative linguistics:
  - Rationalism *versus* empiricism.
  - Competence *versus* performance.
  - Introspection *versus* data analysis.
- Criticism agains corpus-based studies of language.

Generativism and Corpus Linguistics (II)

- Corpus-driven studies started to *reemerge* after several years.
- Generative linguists lowered their criticism against CL.
- Growth of studies in CL both in UK and US (Quirk, Harris, Halliday, Sinclair, etc.).
- Emergence of personal computers allowed linguists to rapidly obtain information from language data (corpora).

Why?

- Why using corpora for language analysis?
  - Data tells us what is *normal* in real-life language use, tendencies of change, etc.
  - Corpora allows linguists to find rare cases (difficult to observe in smaller texts).
  - Computers count faster (and better) than humans.

Types of corpora

- Depending on the aims:
  - General corpus: very large.
  - Specialized:
    - Geographical: only texts from Jamaica, etc.
    - Chronological: e.g., 2000 to 2018.
    - Genre: only fiction / only newspapers, technical, etc.
  - Bilingual / Multilingual:
    - Paralell *versus* comparable.
  - Diachronic (historical) corpus.
  - Learner corpus.

Contents

Corpora (in English)

- English:
  - Brown corpus (American English): 1960. 500 samples, about 1 million words.
  - LOB: Lancaster-Oslo/Bergen Corpus (British English): 1970s. Designed to match the Brown corpus.
  - PTB: Penn Treebank (1990s): Wall Street Journal (most), transcribed speech. More than 4.5 million words.
  - BNC: British National Corpus: 100 million words. Newspapers, research, fiction books, etc. 10% transcribed speech.

Corpora (in other languages)

- Portuguese and Galician:
  - CRPC (311M words), NILC (BP), CORGA (GAL, 36M words)...
- Spanish:
  - CREA, CREA-oral, Corpus del Español de América...
- Russian:
  - Russian National Corpus (>600M words), GICR...
- Kazakh:
  - Kazakh-KTB (*Universal Dependencies*, 10k words).

A few links

- BNC: `https://corpus.byu.edu/bnc/`
- Russian NC: `http://www.ruscorpora.ru/en/`
- GICR: `http://www.webcorpora.ru/en/`

Contents

Corpus design and construction

- Criteria:
  - Size: large body of text.
  - Representativity: of a language / variety / epoch, etc. (findings can be generalized to the variety)
  - Balance: different text categories (to represent the selected variety).
  - Sampling: select samples which reproduce the characteristics of the population.
- Format: machine-readable (e.g., *.txt* files).
- Annotation: enriched with linguistic annotation.

Example: raw text

One football pitch of forest lost every second in 2017, data reveals.

Global deforestation is on an upward trend, jeopardising efforts to tackle climate change and the massive decline in wildlife.

The world lost more than one football pitch of forest every second in 2017, according to new data from a global satellite survey, adding up to an area equivalent to the whole of Italy over the year.

Example: headers and paragraphs

<head type=title>
One football pitch of forest lost every second in 2017, data reveals.
</head>

<head type=subtitle>
Global deforestation is on an upward trend, jeopardising efforts to tackle climate change and the massive decline in wildlife.
</head>

The world lost more than one football pitch of forest every second in 2017, according to new data from a global satellite survey, adding up to an area equivalent to the whole of Italy over the year.

Example: sentences

```
<head type=title>
<sent id=1>One football pitch of forest lost every second in 2017, data reveals.</sent>
</head>

<head type=subtitle>
<sent id=2>Global deforestation is on an upward trend, jeopardising efforts to tackle climate change and the massive decline in wildlife.</sent>
</head>

<p>
<sent id=3>The world lost more than one football pitch of forest every second in 2017, according to new data from a global satellite survey, adding up to an area equivalent to the whole of Italy over the year.</sent>
</p>
```

Example: words and tokens

One football pitch of forest lost every second in 2017, data reveals.

```
<head type=title>
<sent id=1><w>One</w> <w>football</w> <w>pitch</w>
<w>of</w> <w>forest</w> <w>lost</w> <w>every</w>
<w>second</w> <w>in</w> <w>2017</w> <pt>,</pt>
<w>data</w> <w>reveals</w> <pt>.</pt></sent>
</head>
```

Example: words and tokens

One football pitch of forest lost every second in 2017, data reveals.

```
<head type=title>
<sent id=1><w=DET>One</w> <w=NOUN>football</w>
<w=NOUN>pitch</w> <w=ADP>of</w> <w=NOUN>forest</w>
<w=VERB>lost</w> <w=DET>every</w> <w=NOUN>second</w>
<w=ADP>in</w> <w=NUM>2017</w> <pt>,</pt>
<w=NOUN>data</w> <w=VERB>reveals</w> <pt>.</pt></sent>
</head>
```

Process

1. Design (aims, language/varieties, text typologies, size, etc.).
2. Compilation + permissions (important: legal issues!).
3. Sample selection + text capture.
4. Organization + markup.
5. Annotation (automatic? / manual?)

- Internet corpus: BootCaT (`https://bootcat.dipintra.it/`).

Contents

Corpus annotation

- Raw *versus* annotated corpora.
- Layers:
  - Metadata, headers, paragraphs.
  - Tokens.
  - Lemmas.
  - Part-of-Speech tags.
  - Morphological features.
  - Syntactic analysis.
  - ...

Format and annotation

- Corpus format:
  - SGML-XML format.
  - CoNLL-like format.
- Annotation guidelines and tagsets:
  - EAGLES.
  - Universal Dependencies.

CoNLL

One football pitch of forest lost every second in 2017, data reveals.

```
# text = One football pitch of forest lost every second in 2017, data reveals.
1	One
2	football
3	pitch
4	of
5	forest
6	lost
7	every
8	second
9	in
10	2017
11	,
12	data
13	reveals
14	.
```

CoNLL

One football pitch of forest lost every second in 2017, data reveals.

```
# text = One football pitch of forest lost every second in 2017, data reveals.
1	One	one	NUM	NumType=Card
2	football	football	NOUN	Number=Sing
3	pitch	pitch	NOUN	Number=Sing
4	of	of	ADP	_
5	forest	forest	NOUN	Number=Sing
6	lost	lose	VERB	Tense=Past|VbForm=Part
7	every	every	DET	_
8	second	second	NOUN	Number=Sing
9	in	in	ADP	_
10	2017	2017	NUM	NumType=Card
11	,	,	PUNCT	_
12	data	data	NOUN	Number=Sing
13	reveals	reveal	VERB	Tense=Pres
14	.	.	PUNCT	_
```

CoNLL

One football pitch of forest lost every second in 2017, data reveals.

```
# text = One football pitch of forest lost every second in 2017, data reveals.
1	One	one	NUM	NumType=Card	3	nummod
2	football	football	NOUN	Number=Sing	3	compound
3	pitch	pitch	NOUN	Number=Sing	6	nsubj
4	of	of	ADP	_	5	case
5	forest	forest	NOUN	Number=Sing	3	nmod
6	lost	lose	VERB	Tense=Past|VbForm=Part	0	root
7	every	every	DET	_	8	det
8	second	second	NOUN	Number=Sing	6	nmod
9	in	in	ADP	_	10	case
10	2017	2017	NUM	NumType=Card	6	obl
11	,	,	PUNCT	_	13	punct
12	data	data	NOUN	Number=Sing	13	nsubj
13	reveals	reveal	VERB	Tense=Pres	6	parataxis
14	.	.	PUNCT	_	13	punct
```

Automatic/manual annotation and correction

- Automatic annotation (NLP):
  - Stanford CoreNLP.
  - FreeLing.
  - LinguaKit.
  - UDPipe.
  - NLTK.
  - ...
- Correction (with/without automatic annotation):
  - bart: `http://brat.nlplab.org/`
  - WebAnno: `https://webanno.github.io/webanno/`
  - atomic: `http://corpus-tools.org/atomic/`
  - GATE: `https://gate.ac.uk/`

...and many others!

Contents

Uses of language corpora

1. Linguistic analyses:
   - Lexicon, syntax, phraseology, morphology, etc.
   - Comparative studies.
2. Social / Cultural / Historical analyses.
3. Natural language processing:
   - Evaluate the performance of NLP tools.
   - Train NLP tools with annotated data.

Corpus queries

- Frequencies.
- Concordances (KWIC).
- Collocations.
- Keyword extraction.
- ...

- BNC: `https://corpus.byu.edu/bnc/`
- LinguaKit: `https://linguakit.com/en`

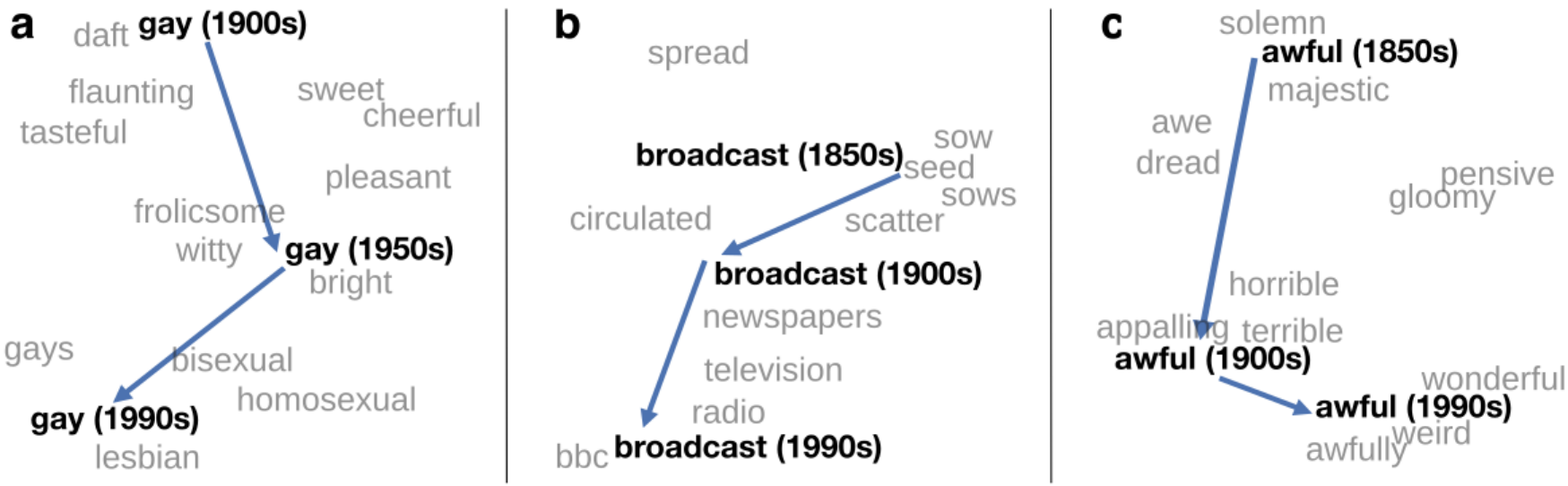
Textbooks and dictionaries

- Textbooks may include the most frequent (and useful) words (and how to use them), instead of rare words which are less useful.
- Learn how to use and combine already known words.

- Dictionaries may include examples from real-life data as well as collocational, phraseological, and other semantic information extracted from corpora.

Cultural and social analyses

- Historical corpora can be useful to find cultural changes:
  - Food (google ngrams).
  - Technology (google ngrams).



Annotated corpora for NLP: evaluation

- Some NLP tasks:
  - PoS-tagging.
  - Lemmatization.
  - Syntactic analysis.
- NLP tools need to be evaluated in human-labeled corpus.

Annotated corpora for NLP: evaluation

- Some NLP tasks:
  - PoS-tagging.
  - Lemmatization.
  - Syntactic analysis.
- NLP tools need to be evaluated in human-labeled corpus.

“The cat chases the mouses”.

1 The
2 cat
3 chases
4 the
5 mouses

Annotated corpora for NLP: evaluation

- Some NLP tasks:
  - PoS-tagging.
  - Lemmatization.
  - Syntactic analysis.
- NLP tools need to be evaluated in human-labeled corpus.

“The cat chases the mouses”.

1 The the
2 cat cat
3 chases chase
4 the the
5 mouses mouse

Annotated corpora for NLP: evaluation

- Some NLP tasks:
  - PoS-tagging.
  - Lemmatization.
  - Syntactic analysis.
- NLP tools need to be evaluated in human-labeled corpus.

“The cat chases the mouses”.

1 The the DET

2 cat cat NOUN

3 chases chase VERB

4 the the DET

5 mouses mouse NOUN

Annotated corpora for NLP: evaluation

- Some NLP tasks:
  - PoS-tagging.
  - Lemmatization.
  - Syntactic analysis.
- NLP tools need to be evaluated in human-labeled corpus.

“The cat chases the mouses”.

1 The the DET _
2 cat cat NOUN Numb=Sing
3 chases chase VERB Tense=Pres
4 the the DET _
5 mouses mouse NOUN Numb=Plur

Annotated corpora for NLP: evaluation

- Some NLP tasks:
  - PoS-tagging.
  - Lemmatization.
  - Syntactic analysis.
- NLP tools need to be evaluated in human-labeled corpus.

“The cat chases the mouses”.

1 The the DET _ det
2 cat cat NOUN Numb=Sing nsubj
3 chases chase VERB Tense=Pres root
4 the the DET _ det
5 mouses mouse NOUN Numb=Plur obj

Annotated corpora for NLP (II): learning

- NLP tools:
  - Symbolic (rule-based).
  - Statistical (machine-learning).
  - Hybrid.
- Statistical and hybrid methods need corpora to learn linguistic models.
- Symbolic approaches can be designed after corpus analyses.

References

- References:

  - Chomsky, Noam, 1957. *Syntactic structures*. Mouton. L'Aia.
  - Firth, J. R., 1951. *Papers in linguistics, 1934-1951*. Oxford: OUP.
  - Käding, F. W., 1897. *Häufigkeitswörterbuch der deutschen Sprache*. Steglitz (unpublished).

- Further reading:

  - Leech, Geoffrey, 1997. "Introducing Corpus Annotation", in *Corpus Annotation*. Longman.
  - McEnery, Tony & Andrew Wilson, 1996. *Corpus Linguistics. An Introduction*. Edinburgh Textbooks in Empirical Linguistics.
  - McEnery, Tony, Richard Xiao & Yukio Tono, 2006. *Corpus-Based Language Studies. An advanced resource book.* Routledge.
  - Lu, Xiaofei, 2014. *Computational Methods for Corpus Annotation and Analysis*. Springer.

- Useful links:
  - EAGLES guidelines: `http://www.ilc.cnr.it/EAGLES/browse.html`
  - TEI: `http://www.tei-c.org/`
  - UD: `http://universaldependencies.org/`

Thanks! :-)

Questions?