CC BY NC SA

Corpora: definition and rights

Karën Fort

karen.fort@univ-lorraine.fr / https://members.loria.fr/KFort/

Institut des sciences du Digital Management & Cognition
COMPOSANTE DE L'UNIVERSITÉ DE LORRAINE

UNIVERSITÉ DE LORRAINE

Sources of inspiration

- Corpus Linguistics [McEnery and Wilson, 1996],
- Slides from Cédrick Fairon and Anne Catherine Simon (Université de Louvain): *Méthodologie de l'analyse de corpus en linguistique*, with their permission.
- Bruno Guillaume's course on the same subject, with his permission

Why talking about it?

You'll use corpora in many of your courses and in NLP in general:

- define properly the object you're working with
- know the possibilities and the limits of its usage
- know the most well-known corpora in the domain

What is a corpus?

Definition

Corpus

A corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic [and/or extra-linguistic] criteria in order to be used as a sample of the language [Sinclair, 1996]

?

text

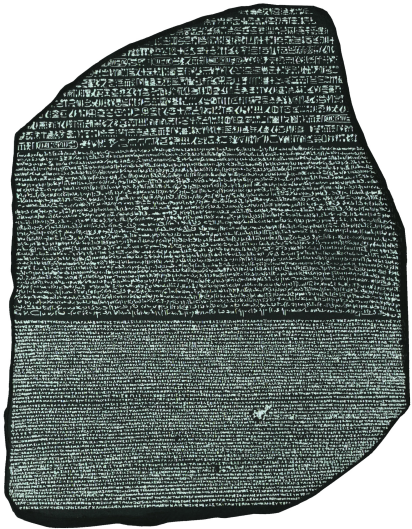
?

text speech

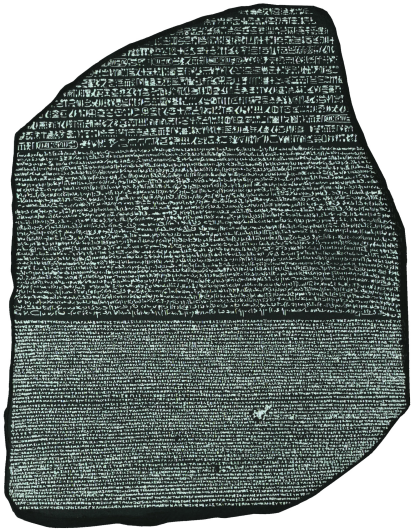
?

text speech video

Monolingual *vs* Multilingual

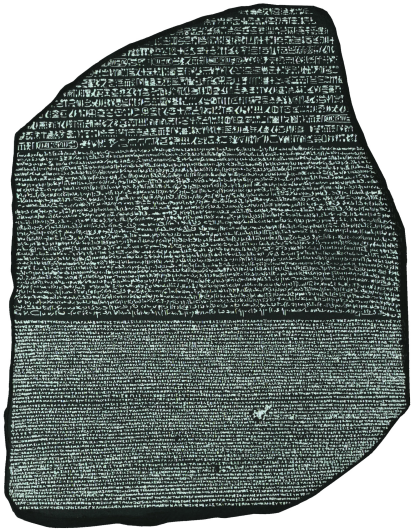


Monolingual *vs* Multilingual



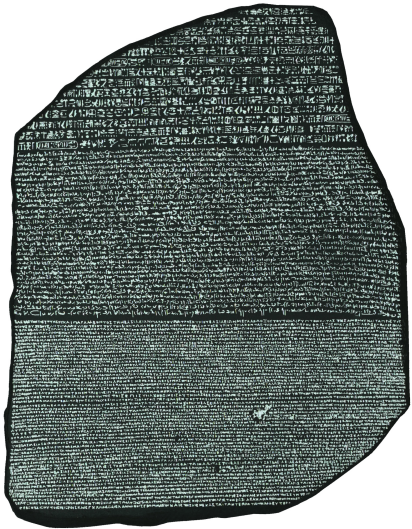
"The Rosetta Stone is a stele of granodiorite inscribed with three versions of a decree issued in 196 BC during the Ptolemaic dynasty of Egypt, on behalf of King Ptolemy V Epiphanes. The top and middle texts are in Ancient Egyptian using hieroglyphic and Demotic scripts, respectively, while the bottom is in Ancient Greek."
[Wikipedia, August 20th 2025]

Monolingual *vs* Multilingual



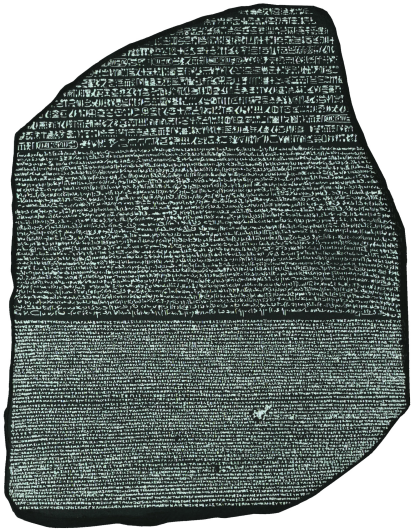
aligned *vs* comparable

Monolingual *vs* Multilingual



1 or 2 (3) corpora?

Monolingual *vs* Multilingual



1 or 2 (3) corpora? depends on the application!

Finite *vs* Open *vs* Dynamic [Baude, 2007]

- Finite / closed: built once and for all as a "complete" corpus [Corpus de référence du français parlé 1; Brown corpus]
- Open: built to integrate new data, excepted or not [Web, online press, Twitter/X]
- Dynamic: sub-category of open corpora, that includes monitor corpora [COBUILD/BOE] and tank corpora [VALIBEL]

Exhaustive *vs* Representative *vs* Balanced *vs* Gold/Reference [Baude, 2007]

- Exhaustive: finite corpus contening all the texts corresponding to a specific usage (eg. from an author)
- Representative: vague notion, by genre, by sociological sampling, by communicational situation
- Balanced: samples of texts (Brown corpus)
- Gold/Reference : built to provide information on a language, of large size and diversity (French TreeBank, Penn TreeBank)

More about balanced corpora

Balanced corpus: (fixed-size) text samples from several sources (eg Brown corpus)

- different types of texts (based on a typology of genres)
- different periods
- different authors
- different sociological categories
- etc

→ There is no valid *scientific measure* for checking the balance of texts in a corpus

From Bruno Guillaume, with his approval

More about representative corpora

No corpus can represent **the** language!

Representativeness is considered from the specific (as opposed to general) perspective of the application for which it is collected.

> *"Every corpus assumes a detailed knowledge of the application for which it is collected, even if this is a simple documentary application: it not only determines the way texts are selected, but also cleaned up, encoded, tagged and finally the structure of the corpus itself. [...] a corpus is relevant to a task according to which one can determine the criteria for its representativeness and homogeneity." [Rastier, 2004]*

From Bruno Guillaume, with his approval

More about reference corpora

"I argue that although it is impossible actually to define a representative corpus in our present state of knowledge, nevertheless the important point is that you try to make it as representative as you can, and so that it will be accepted by the community of speakers of the language, and particularly those studying the language, as some kind of standard reference point."

The Importance of Reference Corpora [Leech, 2002]

From Bruno Guillaume, with his approval

Ecological data *vs* provoked data [Baude, 2007]

natural *vs* **provoked** data (interviews, etc)

Small *vs* big size [Baude, 2007]

What is **big** for a corpus today (in NLP)?

Organized collection *vs* data bank [Baude, 2007]

Selection?

[Joe Crawford from Moorpark, California, USA]

Bags of words *vs* texts collections [Baude, 2007]

- **structured** texts or lists of *independent* words?
- complete texts or parts of texts (samples)?

Raw *vs* annotated [Baude, 2007]

Is speech **transcription** annotation?

Short term *vs* long term [Baude, 2007]

- corpus create for **a** research project
- corpus usable in **several** research projects
- corpus including annotations that can be **shared** (standards)

Conclusion

- corpora are built
- variety of points of view
- not just text!

The Brown corpus

Brown University Standard Corpus of Present-Day American English

- released in **1961**
- 500 samples of 2,000+ words each (edited in 1961)
- 15 genres
- all written by native speakers of American English
- tagged with POS

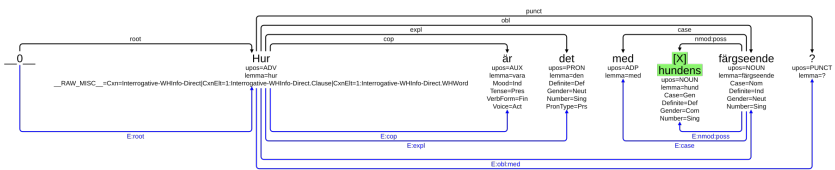
('The', 'AT'), ('Fulton', 'NP-TL'), ('County', 'NN-TL'), ('Grand', 'JJ-TL'), ('Jury', 'NN-TL'), ('said', 'VBD') [Brown corpus]

From Bruno Guillaume, with his approval

Talbanken Swedish corpus

Constructed at Lund University in the 1970s by Jan Einarsson et al.

- 350,000 tokens
- Syntactically annotated corpus, containing both written and spoken Swedish
- Stored initially on punch cards!
- Still available today (see The five lives of Talbanken)



From Bruno Guillaume, with his approval

The Penn Treebank (PTB)

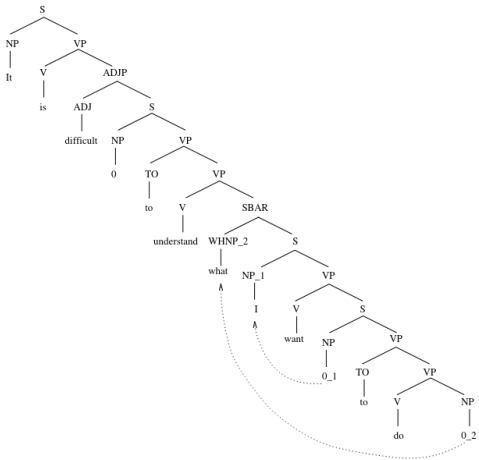
- 40,000 sentences of WSJ newspaper text annotated with POS and phrase structure trees (constituents)
- Trees include some predicate-argument information and traces
- Created in the early 90s
- Produced by automatically parsing the newspaper sentences followed by manual correction
- Took around 3 years to create
- Largely used in many NLP studies for years and years

→ led some commentators to describe "the last 10 years of NLP as the study of the WSJ"

From Bruno Guillaume, with his approval

The Penn Treebank (PTB)

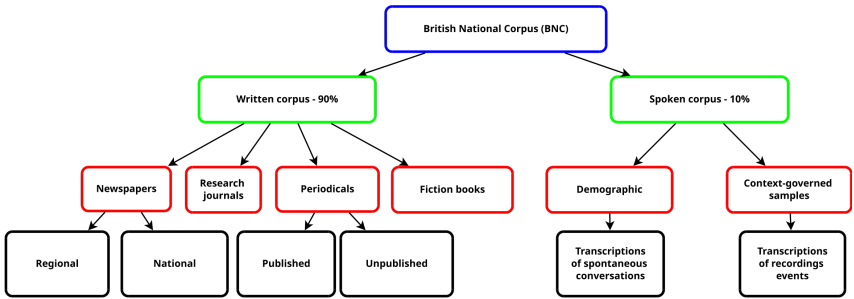
An Example Penn Treebank Tree



http://www.cs.ox.ac.uk/files/552/stat_parse2.pdf

British National Corpus (BNC)

- 100 millions word collection of British English from the 20th century
- started in 1991
- 3 publishers and 2 universities



By Alexchuvak - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=47478326

From Bruno Guillaume, with his approval

Europarl

- proceedings of the European Parliament from 1996 to 2011
- 1st release in 2001
- The latest version includes 21 European languages: Romanic (French, Italian, Spanish, Portuguese, Romanian), Germanic (English, Dutch, German, Danish, Swedish), Slavic (Bulgarian, Czech, Polish, Slovak, Slovene), Finno-Ugric (Finnish, Hungarian, Estonian), Baltic (Latvian, Lithuanian), and Greek.
- After sentence splitting and tokenization the sentences were aligned across languages with the help of an algorithm developed by Gale & Church
- See Europarl Webpage (size of corpora, parallel corpora)

From Bruno Guillaume, with his approval

Universal Dependencies (UD):

- Since 2014, 2 versions / year
- Collaboratif project for the production of corpora annotated in dependency syntax
- http://universaldependencies.org

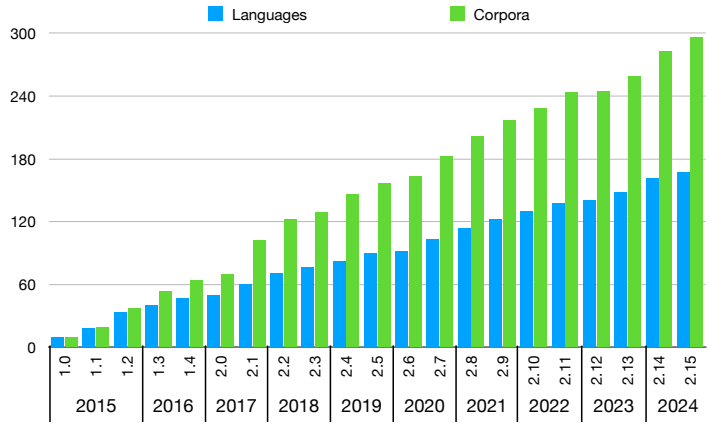
Universal Dependencies in a nutshell

Skolt Sami:

- 245 sentences
- 350 speakers in 2016

UD German-HDT:

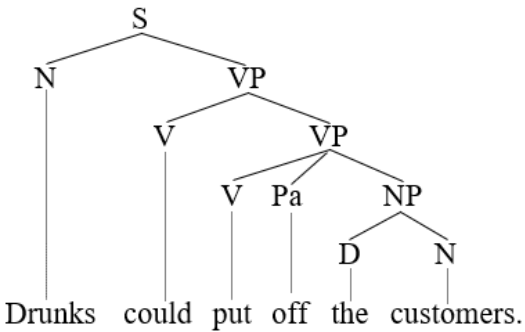
- 189,928 sentences



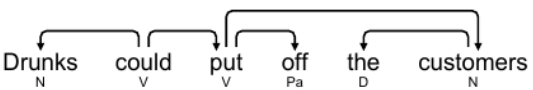
Difference between dependencies (UD) and constituents (PTB)?

Difference between dependencies (UD) and constituents (PTB)

Constituents



Dependencies



From Bruno Guillaume, with his approval

Is a corpus the same as a dataset?

Reminder

Corpus

A corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic [and/or extra-linguistic] criteria in order to be used as a sample of the language [Sinclair, 1996]

Corpus or dataset?

- The Penn TreeBank?

Corpus or dataset?

- The Penn TreeBank?
- MMLU?

Corpus or dataset?

- The Penn TreeBank?
- MMLU?
- texts used to pretrain ChatGPT?

Definition

Dataset

A data set (or dataset) is a collection of data. In the case of tabular data, a data set corresponds to one or more database tables, where every column of a table represents a particular variable, and each row corresponds to a given record of the data set in question. The data set lists values for each of the variables, such as for example height and weight of an object, for each member of the data set. Data sets can also consist of a collection of documents or files [Wikipedia, consulted on August 20th, 2025]

Why is it important?

In an academic environment, it's more and more common to try to build reusable data:

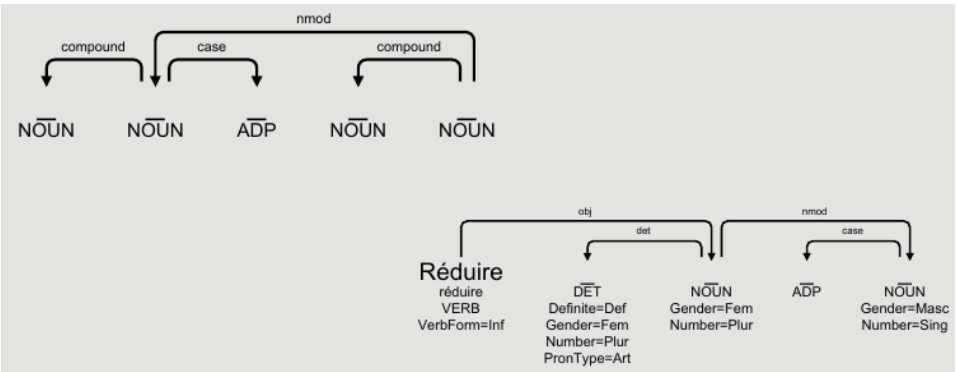
- Free software
- Open access edition

If you plan to publish your corpus and to make it available for others researchers, you need to consider these aspects at the beginning of the project!

From Bruno Guillaume, with his approval

What could happen (in real life)?

Exemples from the French TreeBank



From Bruno Guillaume, with his approval

Finding texts on the Web 1/2

Search for the following in your favorite search engine:

"hitchhicker's guide to the galaxy pdf"

What do you find? Are you allowed to read it? To redistribute it?

Finding texts on the Web 2/2

Search for the following in your favorite search engine:

"victor hugo les misérables pdf"

What do you find? Are you allowed to read it? To redistribute it?

Authors' rights

in the EU

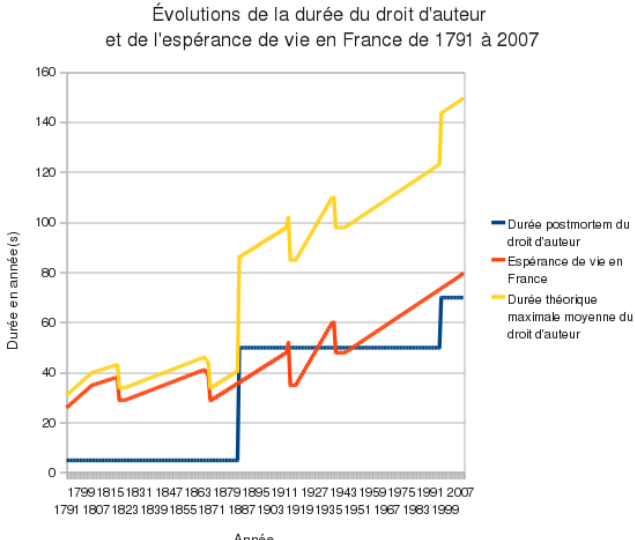
Protection of the authors' creations limited in time:

- Copyright and Information Society Directive (2001)
- but it does not seem that the adaptation in the French law system allows for the same exceptions (see: Wikipédia FR)

70 years after the death of the author

Authors' rights: evolution of the duration

for France



Exceptions

fair use in the USA (else, 95 years duration) and *fair dealing* in the Commonwealth countries

- for non-commercial usages
- for education or research purposes

In France (code de la propriété intellectuelle), exceptions only include:

- free and private presentation within the family circle
- private copy (personal usage or for close relatives)
- short citation

So?

The Hitchhiker's Guide to the Galaxy

- Douglas Adams
- English
- died in 2001

Les Misérables

- Victor Hugo
- French
- died in 1885

So?

The Hitchhiker's Guide to the Galaxy

- Douglas Adams
- English
- died in 2001

Les Misérables

- Victor Hugo
- French
- died in 1885

So?

The Hitchhiker's Guide to the Galaxy

- Douglas Adams
- English
- died in 2001

Les Misérables

- Victor Hugo
- French
- died in 1885

So?

The Hitchhiker's Guide to the Galaxy

- Douglas Adams
- English
- died in 2001

Les Misérables

- Victor Hugo
- French
- died in 1885

depends on the published version...

Then: public domain (*copyleft*)

All the books available for download on the `Project Gutenberg` website are copyleft

`http://www.gutenberg.org/`

What about?

images found on the Web

What about?

images found on the Web

Check the rights
(and always cite the source)

What about?

blogs

What about?

blogs

What about?

Doctissimo

What about?

Doctissimo

What about?

`Twitter/X`

What about?

`Twitter/X`

Can a tweet be considered as a artistic creation?

- originality
- form

https://scinfolex.com/2015/07/30/twitter-le-micro-plagiat-et-la-physique-quantique-du-copyright/

What about?

`Facebook`

Facebook

Date: Tue, 27 Oct 2015 22:31:12 -0700
From: "Eric Ringger" <ringger@cs.byu.edu>
Subject: [Corpora-List] Facebook's policy with regard to sharing
content
To: "'Manuel Burghardt'" <manuelburghardt@gmx.de>,<corpora@uib.no>

Greetings, Manuel.

Thanks for checking regarding the Facebook dataset you have gathered.

I followed up with folks here at Facebook and found that the short answer to
your question is that publishing Facebook data is essentially not
permissible. The written policy for external developers can be found here:

https://developers.facebook.com/policy

Two points from the policy regarding the API for data access are worth
highlighting: one must both "Obtain consent from people before publishing
content on their behalf" and "Protect the information you receive from us
against unauthorized access, use, or disclosure."

As I'm sure you understand, user privacy is a top priority for everyone at
Facebook. I also know the team is open to feedback from the research
community regarding these policies. I'd be happy to pass along any feedback
you may have.

Regards,
--Eric
Research Scientist, Facebook
Associate Professor, CS, BYU

Principles inherited from free software

"Gnu's Not Unix"

"the users have the freedom to run, copy, distribute, study, change and improve the software"

https://www.gnu.org/philosophy/free-sw.en.html

Examples of licences for corpora

Inherited from free software

- `LGPL-LR`: inherited from LGPL, usable in proprietary applications/resources (L for lesser)
- `CeCILL`: French licence (CEA CNRS INRIA Logiciel Libre)

Specific to artistic creations and to documents:

`Creative Commons`: family of licences

Creative Commons

CC licences: example

CC licences: example

`Creative Commons`:

- by: attribution
- nc: *non commercial*
- sa: *share alike* → viral

`https://creativecommons.org/licenses/by-nc-sa/4.0/deed.en`

Some places to find freely available corpora

- The Gutenberg project (literature)
- Ortolang (mainly FR)
- The LRE map
- LINDAT
- The Corpora mailing list

Some freely available aligned corpora

- Europarl
- Open subtitles
- tatoeba

The 3 Ds

- Describe the objective of the corpus
- Do not try to reinvent the wheel!
- Document the corpus:
  - Purpose of the corpus
  - Metadata:
    - Source of the texts
    - Authors
    - Date of extraction in case of web reference (Wikipedia...)
    - Licenses
  - How all the choices were made

From Bruno Guillaume, with his approval

Storing your corpus

- Avoid personal web page!
- GitLab from the Université de Lorraine
- Ortolang (mainly FR)
- LINDAT
- Nakala (API)

From Bruno Guillaume, with his approval

- Definition of a corpus
- Different views on corpora
- `Creative Commons` licences
- Good practises

TODO during practise session: Build a corpus

- Create groups of 3-4 students having a language in common (not English, if possible)
- Build a balanced, representative corpus for the language
- Planned application: named-entity recognition (names, places, addresses, numbers, etc)
- Keep some time at the end of the session to explain to the entire class what you did and the problems you encountered

Baude, O. (2007).
Contribution des corpus oraux à la linguistique de corpus : une démarche réflexive intégrée.
In Actes de Journées de Linguistique de Corpus, Lorient, France.

McEnery, T. and Wilson, A. (1996).
Corpus linguistics.
Edinburgh University Press.

Sinclair, J. (1996).
Preliminary recommendations on corpus typology.
Technical report, Eagles.