Methods in Computational Linguistics
Corpus Metadata & Analysis

Franziska Weeber

Master of Science *Computational Linguistics*
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart

November 17, 2025

Today's slides are based on the materials provided by:

- Eva Maria Vecchi
- Henry S. Thompson
- Hinrich Schütze
- Sabine Schulte im Walde
- Sharon Goldwater
- Diego Frassinelli
- Vivi Nastase

What is a Corpus?

- A collection of naturally occurring language text, chosen to characterize a state or variety of a language (Sinclair, 1991)
- Collection of electronic texts built according to explicit design criteria for a specific purpose (Atkins et al., 1992)
- Collection of pieces of language that are selected and ordered according to explicit linguistic criteria, in order to be used as a sample of the language (Sinclair, 1996)
- Any collection of more than one text (McEnery & Wilson, 2001)
- Large body of linguistic evidence typically composed of attested language use (McEnery, 2003)

What is a Corpus?

Fillmore (1992), “Corpus linguistics”: *I don’t think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore; ... Every corpus that I’ve had a chance to examine, however small, has taught me facts that I couldn’t imagine finding out about in any other way.*

Chomsky (1957), “Syntactic structures”: *Any natural corpus will be skewed. Some sentences won’t occur because they are obvious, others because they are false, still others because they are impolite.*

What is a Corpus?

A Corpus is a collection of written or spoken natural language utterances digitized and machine-readable.

- Data: raw/primary data, metadata, annotation
- Representativeness:
  - must be representative for the language/languages/genres it aims to cover
  - must be representative for the phenomena we want to investigate
- (usually) the bigger the better
- As a sample of language, we cannot assume balance or lack of bias!
  - incomplete
  - unbalanced
  - erroneous

What is a Corpus for?

- to study language
- to extract terminology and build dictionaries
- to study particular linguistic phenomena (possibly using annotations)
- to build language models for various applications
- to build word representations

What is a Corpus for?

- to study language
- to extract terminology and build dictionaries
- to study particular linguistic phenomena (possibly using annotations)
- to build language models for various applications
- to build word representations
- Purpose:
  - **General-purpose**: not built to study a specific phenomenon, but as a representative sample of a language/languages/genres/...
  - **Domain-specific**: built to capture language in a specific domain or genre – e.g. scientific publications in biology

Corpus Typology

Design Criteria

- functionality: purpose
- language: mono-/bi-/multilingual
- medium: written/spoken language, (multi-)modality
- size
- reference: language representative/variety specific

Corpus Preparation

- annotation

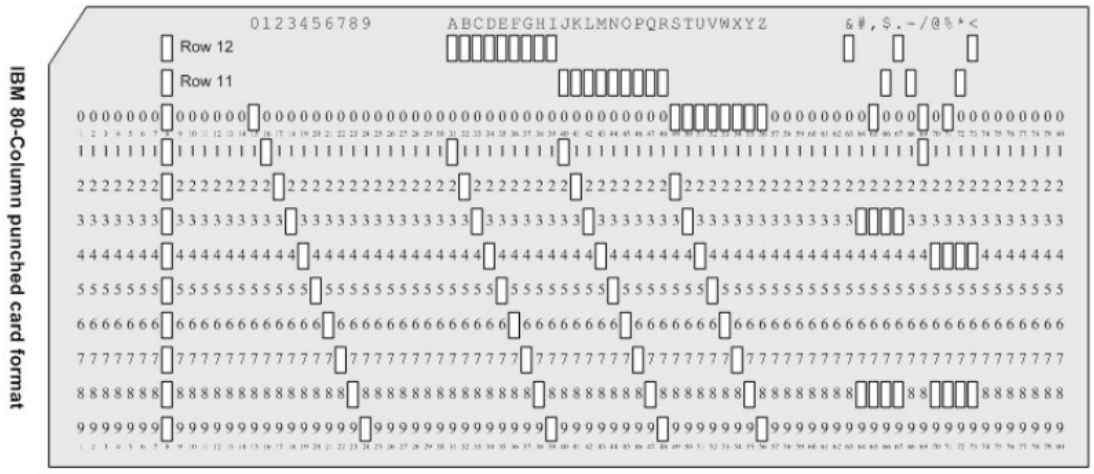
Physical Aspects

- persistency: static vs. dynamic
- availability: free vs. licensed

Examples of Corpora

1. The Brown Corpus

The Brown Corpus is a corpus of roughly 1 million words of American English text, balanced across 15 genres



- The first publicly available electronic corpus
- Produced initially in 1964 by Henry Kucera and Nelson Francis of the Linguistics Department of Brown University
- Composed of 500 texts of roughly 2000 words each
- Part-of-speech information was added in 1979

Examples of Corpora

2. The British National Corpus

In 1995 the BNC was designed to be a representative corpus of contemporary British English

- Data:
  - written (90 million words)
  - spoken (10 million words)
  - i.e., a hundred times the size of the Brown corpus
- two subsequent editions have been issued (SGML, XML)
- The spoken part was particularly ambitious involving the recruiting of participants to record their daily conversation

The BNC effort, jointly funded by the British government and a consortium of dictionary publishers

- Novelty: who has data is willing to donate it with acceptable license terms

Examples of Corpora

3. The Web as a Corpus

Advantages

- huge amount of already digitized texts
- any kind of text available (w.r.t. language, genre, structure, etc)

Disadvantages

- noise – typos, html tags, inner structure (e.g. tables, figures, captions)
- various degrees of structure
- different languages on the same page
- lack/inconsistent metadata

Examples of Corpora

3. The Web as a Corpus

WaCky – the Web-As-Corpus Kool Yinitiative: crawl the web and collect web pages, using seed words

ukWaC (2B words): crawled the .uk domain; seeds: medium-frequency words from the BNC

deWaC (1.7B words): crawled the .de/.at domain; seeds: medium-frequency words from the SüdDeutsche Zeitung corpus and basic German vocabulary lists

itWaC (2B words): crawled the .it domain; seeds: medium-frequency words from the Repubblica corpus and basic Italian vocabulary lists

Examples of Corpora

3. The Web as a Corpus

COW – Corpora from the Web: collect data that is not biased towards certain hosts, basic cleanup and duplicate removal

- multilingual (English, Dutch, German, French, Spanish, Swedish)
- does not consist of a collection of documents, but a collection of sentences! (sentences have been shuffled for copyright purposes)

Examples of Corpora

3. The Web as a Corpus

Common Crawl:[1] free and openly accessible dynamic corpus of web scraped data that exists since 2007

- over 300 billion pages
- 3–5 billion new pages added each month
- used in the pretraining data of many LLMs
- multiple petabytes of data

[1] https://commoncrawl.org

Examples of Corpora

4. Parallel Corpora

- The same sentence in different languages
- Different applications:
  - Training data for Machine Translation systems
  - Test of linguistic theories (Universals?)

Europarl debates of the European Parliament from 1996 to 2011

- 21 parallel corpora
- sentence aligned

Hansard debates of the Canadian Parliament

- English and French (later also Inuktitut)
- approx 2000k/2000k aligned sentences

Many more recent resources out there, in a variety of languages (https://www.clarin.eu/resource-families/parallel-corpora)

Examples of Corpora

5. Comparable Corpora

Comparable Corpora: contain texts covering the same topics in different languages, e.g.

- Wikipedia pages corpus
- The Coronavirus corpus

What kind of information is there in a Corpus?

- the texts themselves
- metadata
- optional: annotations

What kind of information is there in a Corpus?

Metadata

authorship information

- Who wrote it
- When
- Who published it
- What language it is in
- and all the other things you would expect in a bibliographic record

corpus building information

- Who processed it
- What tools they used
- Criteria for filtering data
- etc

Properties and statistics of corpora

Basic Corpus Statistics

- Number of words including repetitions (token count)
- Number of unique words (type count)
- Number of lexemes (lemma/lexeme count)
- Sentence lengths (crucial for language modeling)
- Document lengths

More Statistics

- Variability measure: token/type ratio
- Recall: Zipf's Law (Zipf, 1949)

Co-occurrences statistics

Gries and Durrant (2021)

- Help in identifying patterns and relationships between words
  - tendency of words to be found together
  - both content and function words
- Frequently co-occurring content words form collocations
  - *World Cup, prison term/sentence, climate crisis*
- Content words co-occurring with function words form constructions
  - *to be regarded as, to be done with X*, etc

Co-occurrences statistics

Gries and Durrant (2021)

- More general pattern of co-occurrence can elucidate word meanings
  - the *distributional hypothesis*, more on this in a couple weeks
- Common applications
  - Collocation analysis
  - Word similarity measurement
  - Contextual frequency evaluation

Co-occurrence Measures

1. G^2 Statistic (Log-likelihood Ratio):

- Used to measure the association between two terms by comparing observed and expected frequencies

$$G^2 = 2 \sum_{i=1}^{4} O_i \ln \frac{O_i}{E_i}$$

where:

- O_i: Observed frequency in cell i of the contingency table
- E_i: Expected frequency in cell i under independence assumption

- contingency table: word a present/absent, word b present/absent
- higher score means stronger association

Co-occurrence Measures

2. *t*-score:

- Highlights co-occurrence strength; higher *t*-scores suggest a stronger association

$$t = \frac{O - E}{\sqrt{O}}$$

where:

- O: Observed frequency of the co-occurrence
- E: Expected frequency of the co-occurrence under independence

Co-occurrence Measures

3. Pointwise Mutual Information (PMI):

- Measures how much more likely two words co-occur compared to what would be expected by chance
- Probably the most often used

$$\text{PMI}(x, y) = \log \frac{p(x, y)}{p(x) \cdot p(y)}$$

where:

- $p(x, y)$: Joint probability of observing x and y together
- $p(x)$ and $p(y)$: Marginal probabilities of x and y, respectively

Co-occurrence Measures

4. Odds Ratio (OR):

- Compares the likelihood of two words appearing together versus separately
- commonly used in corpus linguistics

$$\text{OR} = \frac{O_{11} \cdot O_{22}}{O_{12} \cdot O_{21}}$$

where:

- O_{11}: Observed frequency of both x and y occurring
- O_{22}: Observed frequency of neither x nor y occurring
- O_{12}: Observed frequency of x without y
- O_{21}: Observed frequency of y without x

Adding Layers: Annotations

Adding one (or more) extra level(s) of information to the raw data

- Aim: linguistically enrich the raw data
  - e.g. word types, syntactic categories, semantic roles
- Procedure: manually vs. (semi-) automatically
  - Often requires the use of several annotators
  - Time consuming and expensive

Level	**Annotation (examples)**
Morpho-Syntax	part-of-speech
Morphology	lemmatization
Syntax	constituents, dependencies
Semantics	semantic roles, named entities
Pragmatics	coreference, discourse structure
Others	orthography, time, emotions, gestures

Further Readings on Corpus Processing

- Manning and Schütze (1999). Foundations of statistical natural language processing. MIT press.
  - **Chapter 1** – Introduction
- Schütze et al. (2008). Introduction to information retrieval. Cambridge: Cambridge University Press.
  - **From Section 2.0 to Section 2.2**
- Jurafsky and Martin (2024). Speech and Language Processing. Online Manuscript.
  - **Chapter 1** – Introduction
  - **Chapter 2** – Regular Expressions, Tokenization, Edit Distance

References

Noam Chomsky. 1957. Syntactic structures. *Cambridge, Mass.: MIT Press.(1981) Lectures on Government and Binding, Dordrecht: Foris.(1982) Some Concepts and Consequences of the Theory of Government and Binding. LI Monographs* 6(12):1–52.

Charles J Fillmore. 1992. Corpus linguistics or computer-aided armchair linguistics. *Directions in Corpus Linguistics/Mouton de Gruyter* .

Stefan T. Gries and Philip Durrant. 2021. Analyzing co-occurrence data. In *A practical handbook of corpus linguistics*, Springer, pages 141–159.

Daniel Jurafsky and James H. Martin. 2024. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd edition. Online manuscript released August 20, 2024. https://web.stanford.edu/ jurafsky/slp3/.

Christopher Manning and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. MIT press.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.

George K Zipf. 1949. *Human behavior and the principle of least effort*. Addison-Wesley.