Introduction to Computational Linguistics
Session 4: Text as Data

Denise Löfflad

Universität Tübingen

November 22, 2023

Plan

- Corpus Linguistics / Corpora
  - Data collection
  - Data Modeling
  - Corpus Annotation
- How are words distributed in a text?
  - Zipf's Law
- Word meanings as vectors

History of (Computational) Linguistics
Meyer [2023], McEnery [2019]

- Before the 1960s, generative grammar dominated linguistics
- Linguistic analysis were created based on the intuitive knowledge of the researcher
- Corpus linguistics was criticized due to the restricted size & because corpora contain real-life, ill-formed data

Corpus Linguistics

Corpus Linguistics

What is a Corpus?
Structured collection of texts, collected with a specific question in mind.
Usually with metadata etc

- Structured collection of texts (typically with linguistic annotations)
- Often collected with a specific linguistic question in mind (to help to know what kind of data to collect))
- Usually digitized and represented with special formats (e.g., XML)
- Choice of format depends on purpose of corpus (e.g., NLP, philology)
- It is good practice to make corpora freely available to other researchers, but in reality that's not always done/not always easy

Definition Corpus

- structured collection of utterances
- machine readable
- components:
  - raw, original data (primary)
  - meta data
  - annotation guidelines

Example Corpus

Brown Corpus

- First modern Corpus
- Computer readable (available from different sources, but often as .txt)
- Approx. 1 milion words from written American sources (running test and English prose)
- 15 different categories, 500 texts, approx. 2k words per text, collection of text samples
- intended purpose: study word frequencies and distributions of everyday language use
- word class tagging (e.g. POS)
- Freely available

Brown Corpus - Data

```
I_PPSS answered_VBD the_AT routine_JJ question_
NN about_IN my_PP$ itinerary_NN ,_, rather_QL coolly_RB ._.
Chiang_NP spoke_VBD again_RB ,_, this_DT time_NN at_IN
greater_JJR length_NN ._.the_AT President_NN says_VBZ ,_,
the_AT translator_NN came_VBD in_RP ,_, that_CS the_AT
reason_NN he_PPS asked_VBD you_PPO where_WRB you_PPSS
were_BED going_VBG is_BEZ because_CS he_PPS hoped_VBD
you_PPSS would_MD be_BE visiting_VBG other_AP areas_NNS
 in_IN Southeast_JJ Asia_NP ,_, and_CC that_DT everywhere_RB
```

source: Brown corpus `http://www.sls.hawaii.edu/bley-vroman/browntag_nolines.txt`

Zipf's Law

Corpus linguistics can be used to analyze the frequencies of certain words, which can be useful for tasks like topic modeling, authorship attribution, (second) language development ...
Zipf's Power Law:

- Frequency is inversely proportional to its frequency rank.
- Type-token distinction explained: Unique word types vs. specific usages (tokens).
- Example from Brown Corpus: "the" (1st rank) accounts for 6% of tokens, "of" (2nd rank) for 3%, and so on

Zipf's Power Law

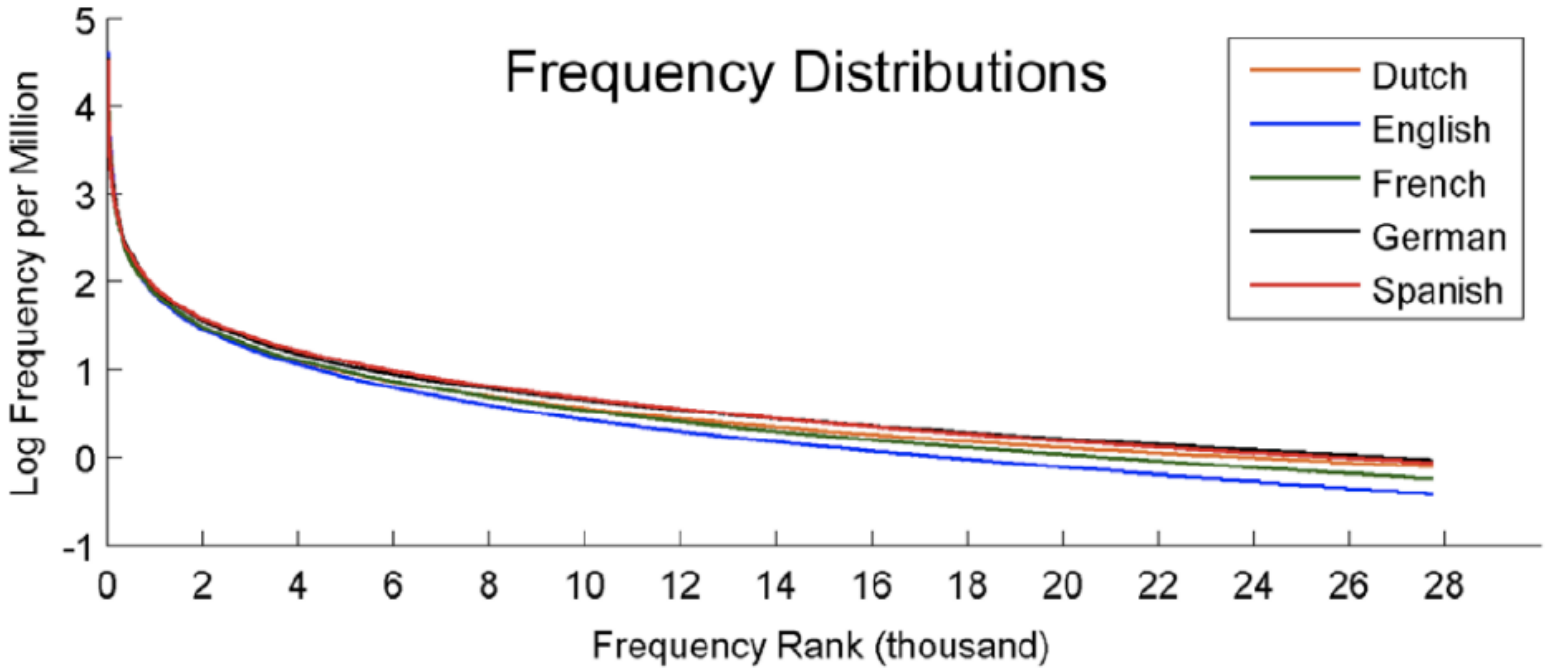


Figure: Log frequency of words as a function of their frequency rank (ordered from left to right as the first quent word, the second most frequent, and so on). Figure from Marian et al. [2012].

Mandelbrot [1961] Carrol [1967]

Zipf's Brevity Law

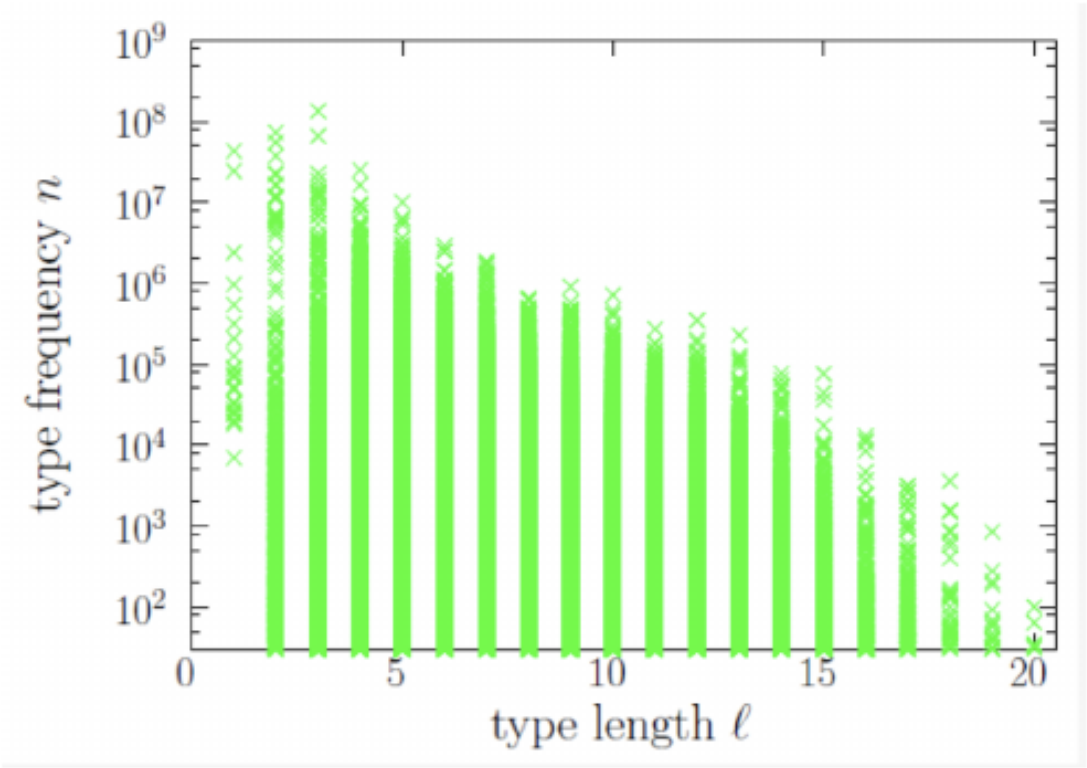
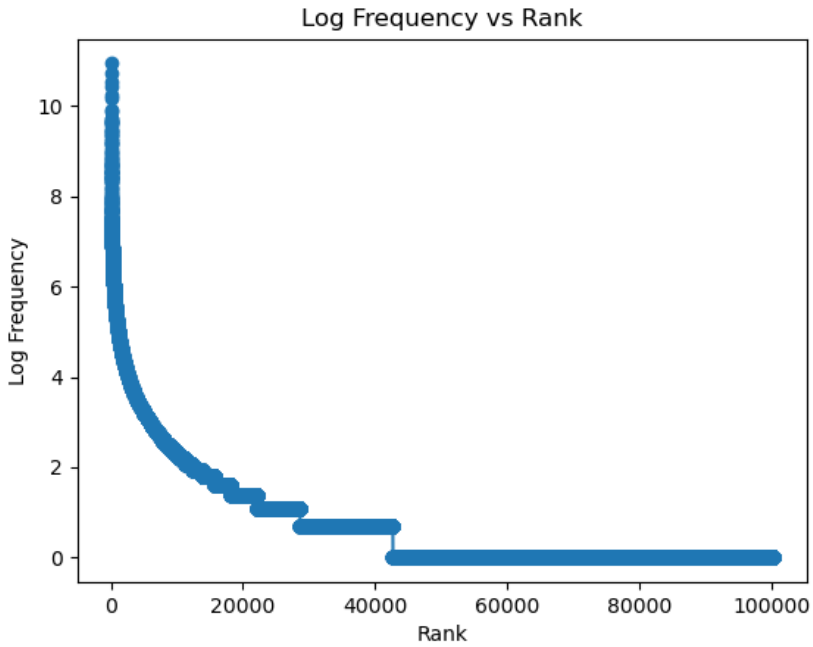
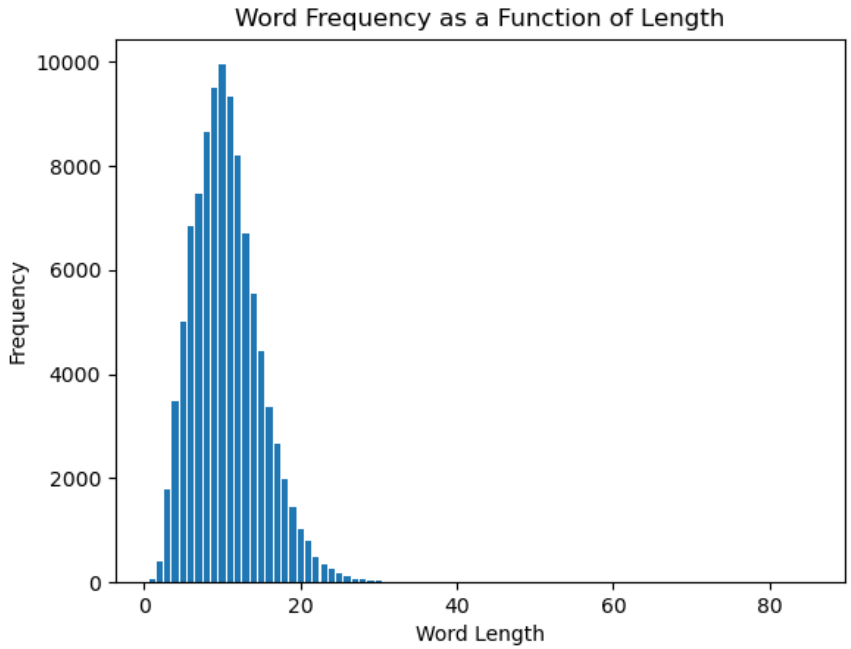


Figure: Frequency of English words as a function of their length in orthographic letters. Figure from Corral and Serra [2020].



(a) Rank vs Log Frequency



(b) Word Frequency as a Function of Length

Figure: Testing Zipf's Law on our data. Corpus included about 3k texts, 11208627 tokens, and 100327 types

```
[('die', 56495),
 ('der', 45026),
 ('und', 36994),
 ('in', 33694),
 ('das', 27532),
 ('ist', 26199),
 ('sie', 19982),
 ('es', 19491),
 ('den', 16443),
 ('ein', 16354),
 ('nicht', 15697),
 ('auch', 15541),
 ('zu', 15392),
 ('mit', 14928),
 ('„', 14841),
 ('für', 14226),
 ('ich', 12935),
 ('eine', 12647),
 ('im', 12593),
 ('aber', 11805)]
```

Figure: 20 most frequent words in Spotlight Corpus.

Type Token Ratio

- #words = #tokens
- #different words = #types
- TTR = $\frac{\#types}{\#tokens}$
- high TTR indicates high lexical variation
- BUT affected by text length!

Word vectors

visualization of word vectors

- we use number vectors to represent words
- for computers, strings are only sequences of characters with no meaning
- there are pre-trained models you can use to vectorize words, e.g. Word2Vec in python
- vectors allow to represent words in a coordinate space and represent relations between words

Word vectors

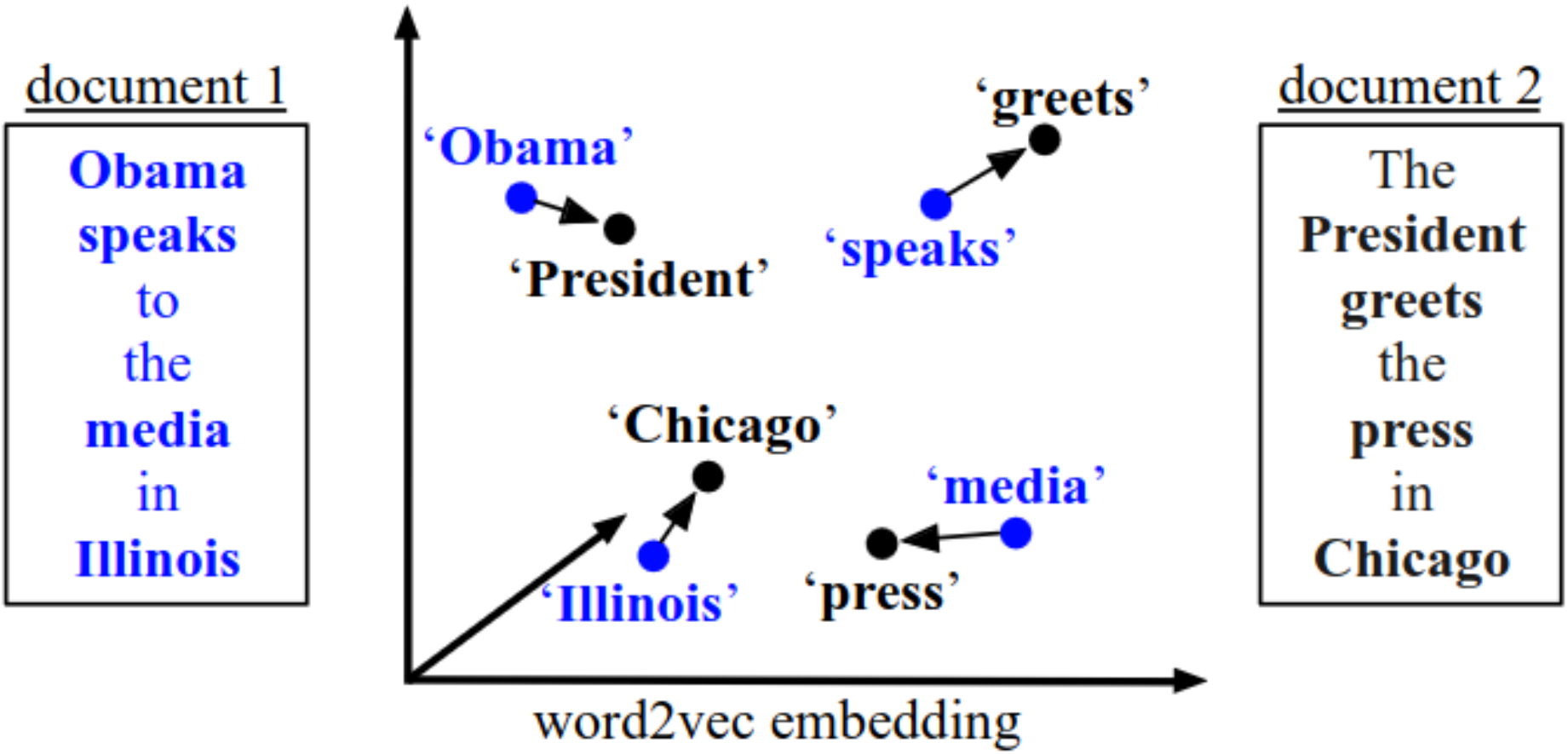


Figure: Word Vector Embedding from Kusner et al. [2015]

Role of Corpora

- testing (linguistic?) theories
- searching relevant examples
- but:
  - corpus is never fully representative or balanced
  - absence/infrequency of construction not necessarily indicator of ungrammaticality, etc.
  - negative data: no data in search results doesn't mean it doesn't exist

Role of Corpora

- training/testing systems
- many different applications, e.g.
  - extracting frequencies of parts-of-speech for words
  - estimating n-gram frequencies from sentence-segmented text
  - training probabilities of parsing rules from trees
  - extracting errors in learner language
  - extracting markers for sentiments in corpora
  - ...

→ Many uses for corpora, depending on the task at hand

Design - Selecting Data

Questions to answer before collecting data:

- What domain/task is the data needed for ?
- From what sources can the data be accessed ? Is it legally and technically possible ?
- Is it ethically acceptable?
- How to select data representative for this domain ?
- How can the data be selected in a balanced fashion ?

Design - Data Modeling

- decide on a model and use it for annotating/creating the corpus
- data **model**ing
- About 'models':
  - they include some aspects of objects
  - other aspects are omitted
  - can include extended information: e.g. metadata not present in original data

Data Modeling - Labels and Structure

- For aspects included in model:
  what categories / labels (system) to use ?
- How is the system organized ?
- Note: there often is more than one valid model

⇒ **Document your decisions !**

Annotation Manual

- documentation of decisions: annotation manual
- lists all possible annotations used in the annotation process plus explanations and examples
- describes procedure's "grey areas"
- usually built by previous knowledge, applied to pilot data, refined from pilot data, then applied to main data
- often useful to CL for building NLP tools

Annotation Manual Examples

- Manual for Brown Corpus:
  http://icame.uib.no/brown/bcm.html
- Penn Treebank tag set:
  https://catalog.ldc.upenn.edu/docs/LDC99T42/tagguid1.pdf

Next Stage: Corpus Annotation

- Raw data collection is completed
- Next step: Add *meta data* to original data

Metadata

Metadata

What is metadata ?

- data describing the primary, raw data
  (e.g. author, creation date, tags, . . .)
- often general meta data (created when, by whom, where, . . .)
- computational linguistics: often linguistic annotations (sentence/word boundaries, POS tags, parse trees, . . .) $\rightarrow$ different levels/layers

Workflow in a nutshell

Typical Workflow:

1. Decide which data to use in the corpus.
2. Collect the data.
3. Decide on an annotation scheme.
4. Pilot annotation: test and refine annotation scheme on data subset.
5. Annotate the rest of the data.
6. Validate the annotation.
7. Assign a license and publish the data and guidelines.

Annotating Data with Metadata

further considerations:

- size of corpus: manual, computer-assisted, automatic annotation
- annotation accuracy of humans and tools

Data Serialization

- corpus and annotations should be digital
- machine readable (easily processable by machines)
- well-formed (system strictly following defined rules)
- non-proprietary, non-binary format (better)
- archivable

Data Serialization

How to store the information *in practice* ?

- structured text (e.g. inline, table)
- markup (e.g. XML)
- Javascript Object Notation (JSON)

→ More on this in Text Technology

In Summary

Why annotate metadata ?

- systematic assignment of metadata enriches corpus
- search using abstractions (data classes) instead of compiling out all surface forms and context
- allows to learn from the data/gain insights
- annotated data can be reused by others (who possibly couldn't perform the annotation)

Annotations allow for a qualitative and quantitative analysis !

Attention Span

Further Reading

If you want to learn more about corpus linguistics, or if you have to create your own corpus or analyze a corpus, I recommend Meyer [2023] and McEnery [2019] (available through the university website `https://uni-tuebingen.de/einrichtungen/universitaetsbibliothek/suchen-ausleihen/`). These books are a good start, but as they are from the early 2000s, the methods might not be up-to-date, so look for more recent sources as well.

Acknowledgements

These slides are largely based on Dickinson et al. [2012] and partly based on slides for Text Technology, 2023 by Stephen Bodnar and on Detmar Meurers' slides on Second Language Acquisition.

References and Acknowledgments

JB Carrol. On sampling from a lognormal model of word frequency distribution. *Computational analysis of present-day American English*, pages 406–424, 1967.

Álvaro Corral and Isabel Serra. The brevity law as a scaling law, and a possible origin of zipf's law for word frequencies. *Entropy*, 22(2):224, 2020.

Markus Dickinson, Chris Brew, and Detmar Meurers. *Language and computers*. John Wiley & Sons, 2012.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR, 2015.

Benoit Mandelbrot. On the theory of word frequencies and on related markovian models of discourse. *Structure of language and its mathematical aspects*, 12:190–219, 1961.

Viorica Marian, James Bartolotti, Sarah Chabal, and Anthony Shook. Clearpond: Cross-linguistic easy-access resource for phonological and orthographic neighborhood densities. 2012.