Basic concepts and corpus types

Introduction to Corpus Linguistics

Adam J.R. Tallman

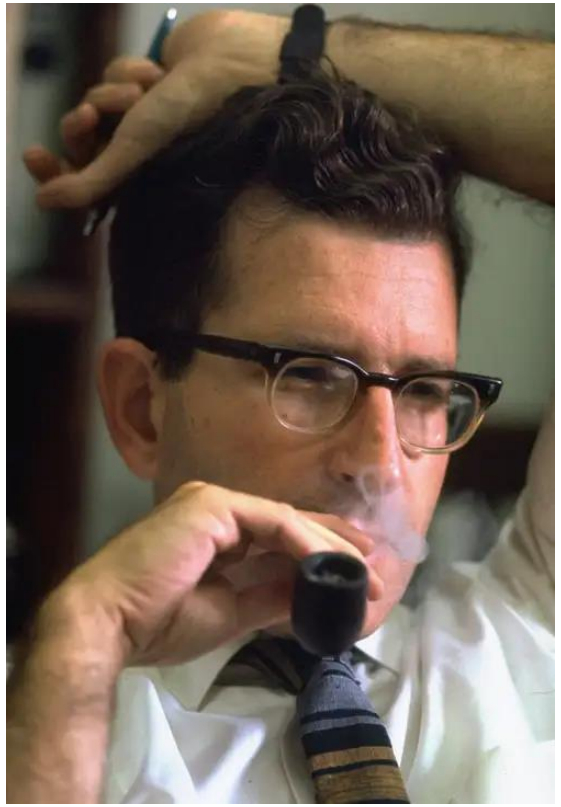
2023-10-27

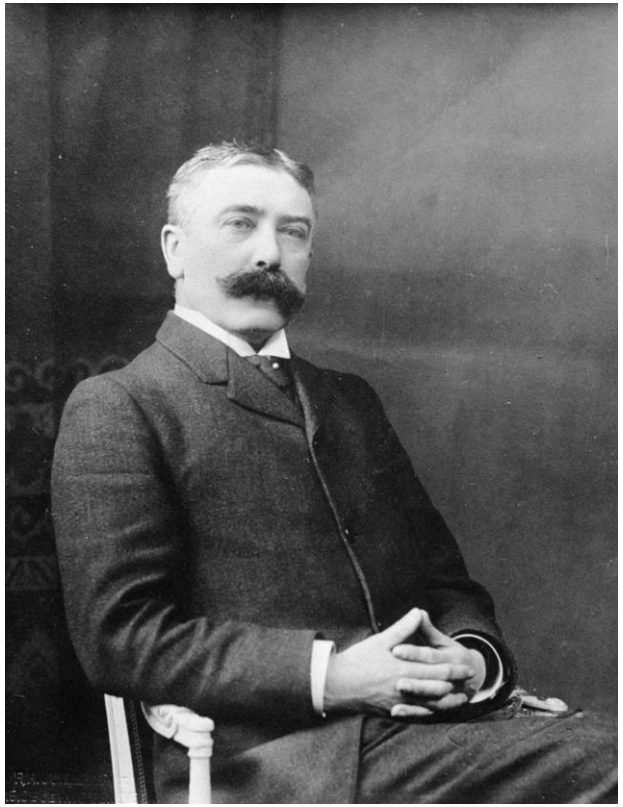
Corpus & text

- Corpus: collection of texts
- Text: an instance of recorded language
- Modern concern: corpus should be *machine-readable*
- That means you can search it

Language knowledge vs. language use

- There’s a distinction in linguistics between knowledge and use
- Competence and performance (Chomsky)
- Langue and parole (Saussure)





Knowledge and use

- So what's the relationship?
- To what extent does language use condition linguistic knowledge?
- To what extent does linguistic knowledge constrain language use?
- Linguists are *very much divided* on this issue (at least on the surface)

Corpus Linguistics

- Corpus linguistics
- “regularities of language use” (Barth & Schnell 2022)
- what people are most likely to say rather than what the possibilities are of what they could say.

Corpora

- There’s not a very precise definition of what a corpus is – people vary according to what they consider a corpus.
- Only ‘authentic’ texts (McEnery & Wilson 2001; Stefanowitsch 2020)
- Experimental corpora are not allowed for example.
- Main problem is that its hard to draw a line between authentic and nonauthentic

Corpora

- What does **not** matter for inclusion
- The difference between proper and improper speech
- Corpus linguistics is not ***prescriptivist***

Text

- Strings of ‘words’ ... but structured into a coherent and meaningful series and recognized by speakers of the language as such.

- Usually corpora are annotated for different properties (e.g. part of speech, noun, verb, adjective etc.)

- They could also be annotated with **glosses**, translation equivalents from another language

Lexeme vs wordform

- There is an important distinction between **wordform** and **lexeme**
- **Lexeme** often means the same thing as **lemma**
- **Wordforms** are directly observable manifestations of **lexemes** in particular contexts
- **e.g. *GO* = LEXEME,** which has the wordforms, *going, gone, goes etc.*

Lexeme vs. wordform

- The relationship between lexemes and wordforms is usually thought to be mediated by *inflection*

- For instance, *runs* is the present tense, third person singular subject inflected form of the lexeme RUN

Lemmatization

- We might want to search lexemes rather than wordforms

- Lemmatization is the process whereby wordforms are given lexeme representations (or annotations) – *running* would just be *run*

- Caveat: This discussion is all very Eurocentric, since a clear distinction between inflection and other types of modification, not to mention a distinction between words and phrases is primarily a property of the European grammatical tradition.

Types versus tokens

- A related distinction but broader in application is between types versus tokens

- **Types**: a type of construction or word-form defined broadly than lexeme

- **Token**: specific instance of a wordform

Token vs. type

- A lexeme is a type, but a type is not necessarily a lexeme
- Because a type can be defined more broadly
- For instance, a type could be PAST TENSE FORMS or PAST TENSE FORMS marked with *–ed* or PREFIX-ROOT combinations

Type vs. token

- Its the relationship between type and token that is thought to be important in linguistics

- Type/token ratio: A low type/token ratio is associated with irregularity (broadly).

- What does this mean?

Frequency measures (central tendency)

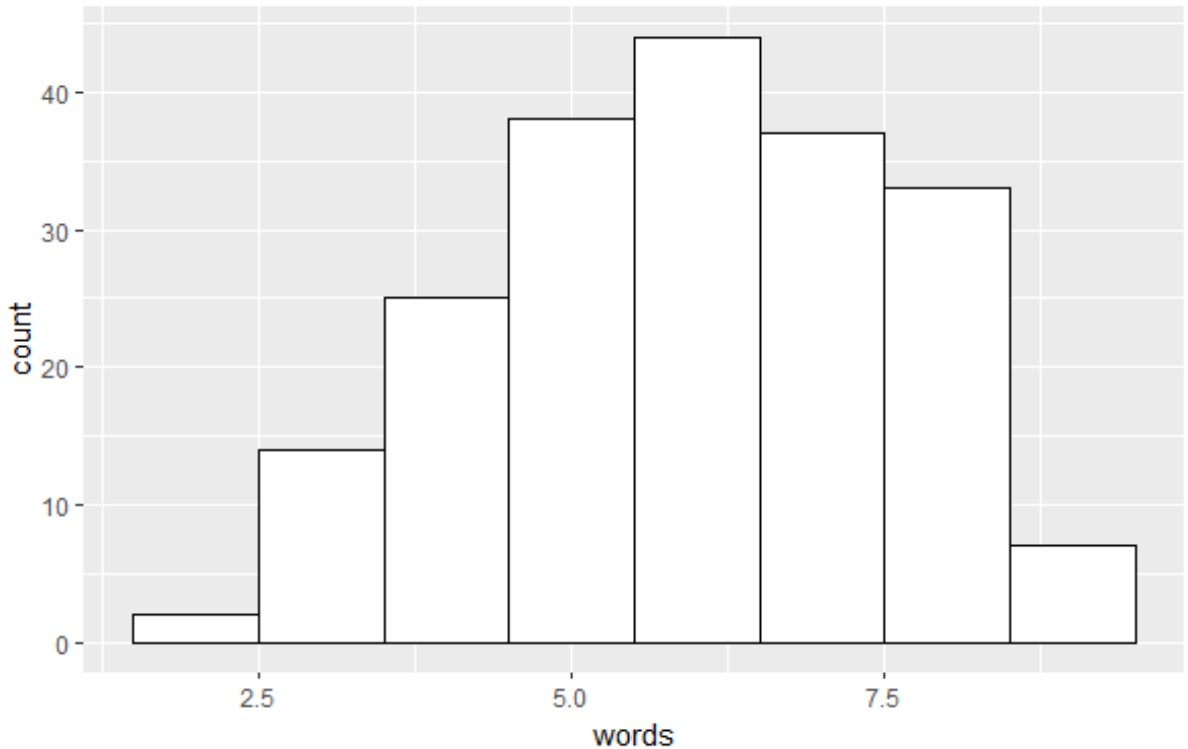
- **arithmetic mean:** Take the sum of the frequencies you have and divide by the number of frequency measurements
- **mode:** the most value for a frequency measurement
- **range**: the difference between the most frequent and the least frequent
- These are measurements of ‘central tendency’

Frequency measures (spread)

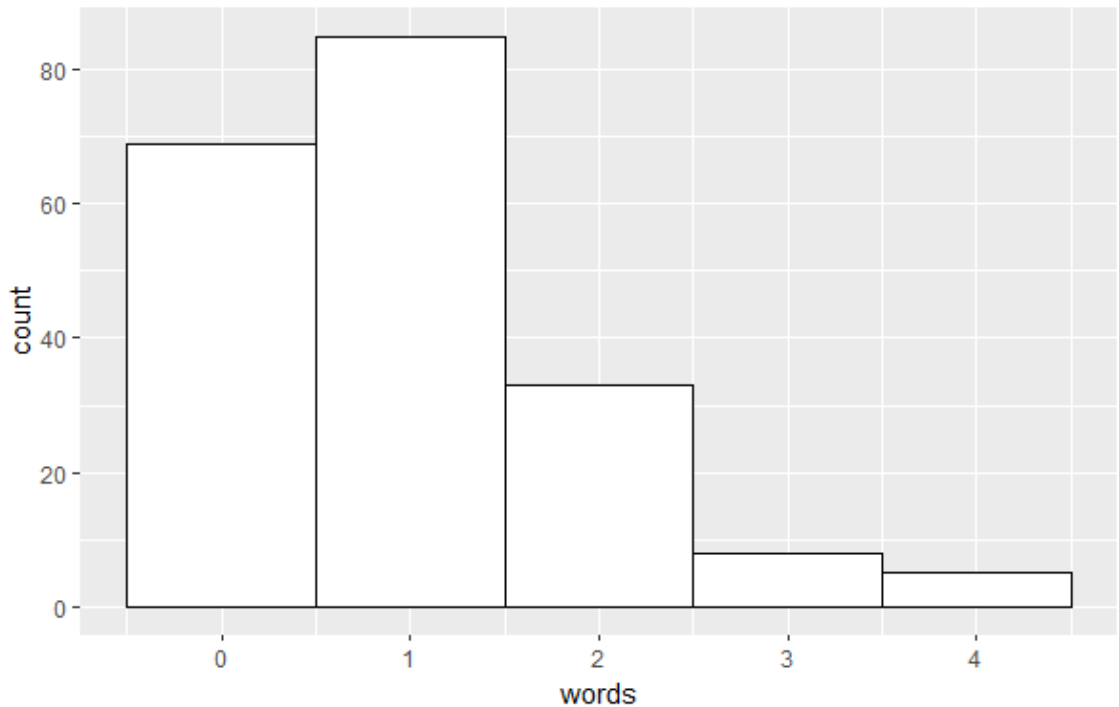
- A *distribution* is defined by a combination of measurements for central tendency and for ***spread***

- Spread refers to how diverse the data are, how different they are from one another.

More spread out values

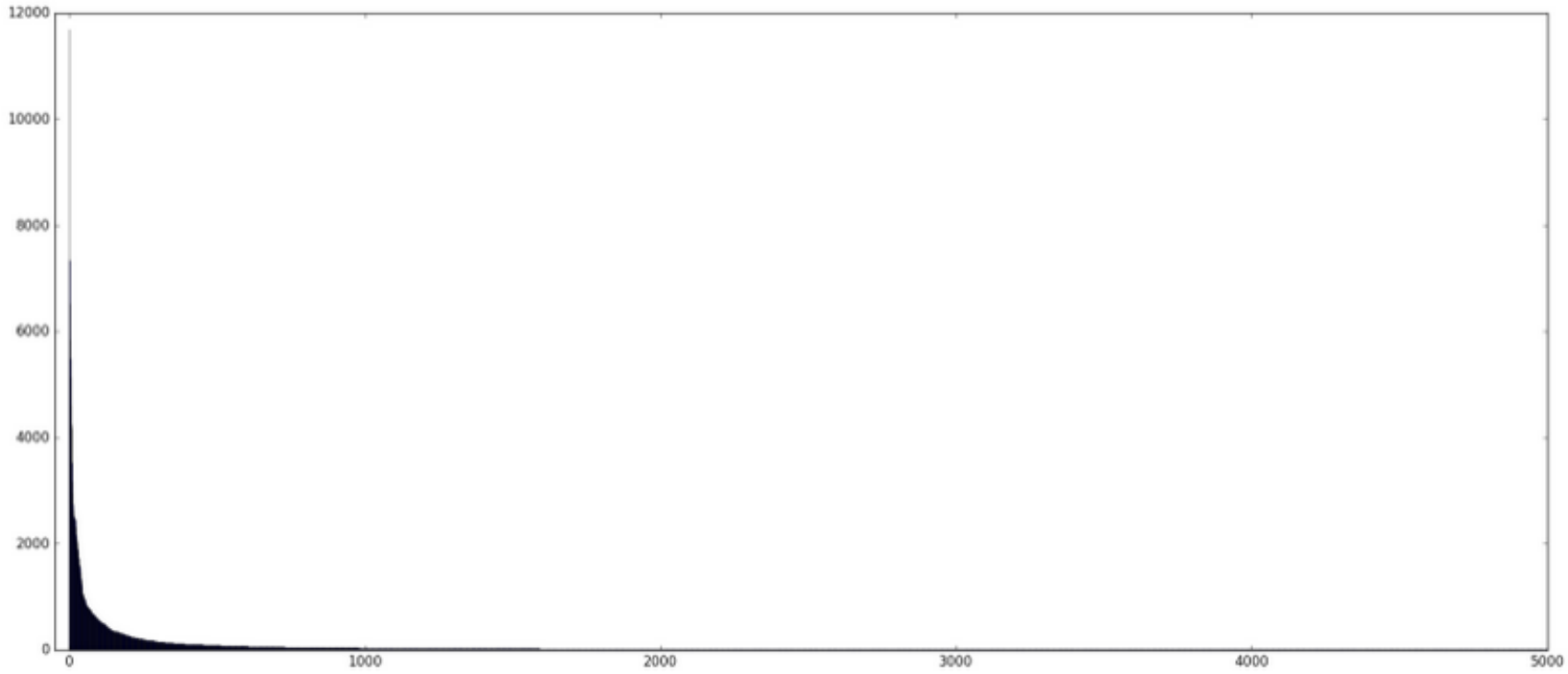


Less spread out values



Zipfian distribution

- There’s a famous distribution for word frequencies



Contexts

- What is relevant for interpretation of the text as a whole?
- **Language-internal**: derived by other properties internal to the text
- **Language-external**: background knowledge, the audience etc.

Language-internal

- **Syntagmatic context**: which words (morphs or elements) come before or after a word-form, morph etc.

- **Constructional context**: what position it has in a higher order construction

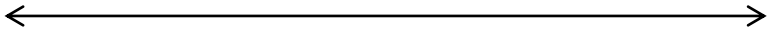
Syntagmatic context

- There's a distinction between paradigmatic and syntagmatic in linguistics.
- Imagine grammar as a set of **positions** and elements that can fit out those positions, then the relationship between elements that could occur in a specific position is a paradigmatic one and the relationship between elements in a different position is a syntagmatic one.

The	student		read	his	textbook	yesterday
	Adam		prepared	the	lecture	today
The	Dean		gave	a	speech	last week
My	cat	will	eat	the	plants	

Syntagmatic context

Syntagmatic



Paradigmatic

The	student		read	his	textbook	yesterday
	Adam		prepared	the	lecture	today
The	Dean		gave	a	speech	last week
My	cat	will	eat	the	plants	

Constructional context

- Construction is a very broad term – but it just refers to any type of ‘frame’ that you can put sentences in

- We can talk about a **passive construction**

SUBJECT	BE	VERB-ed	by NP	
The lecture	was	given	by	the teacher

Constructional context

- We could say that *the lecture* is a subject in a passive construction
- We could say that *The* is a determiner of a subject in a passive construction etc.
- These are statements about the constructional context

SUBJECT	BE	VERB-ed	by NP	
The lecture	was	given	by	the teacher

‘Concordancing’

- McEnery & Hardie (2012)

‘The procedures themselves are still developing , and remina an unclearly delineated set – though some of them, such as concordancing are well established and viewed as central to the approach’ (p.1)

Concordanding just means looking that the syntagmatic context of a form...

Collocations

- **Collocations:** A combination of adjacent forms.
- **N-gram:** We can talk about bigram, trigram etc. collocations
- **Collocates:** A forms collocates are the forms it occurs beside
- Let’s do a search on the COCA corpus

Colligation

- A special form of collocation but refer to more abstract categories (e.g. part of speech categories)

- We can ask often *the* occurs before a noun rather than an adjective

- This is a question about **colligation**.

Language-external context

- Background knowledge
- Who the audience is
- Aspects about the speaker, or writer etc.
- All texts require background knowledge (about a culture for example) in order to understand

Situational context

- **Critical discourse analysis** is concerned with the ideologies embedded in texts and their relationship to particular social groups

- The use of words **frame** particular situations beyond the literal meaning of the text

- For example, what do the use of words in one newspaper as opposed to another say about the ideology of the newspaper producers (or their view of their intended audience)

Situational context

- How do different newspaper report about the conflicts in Israel/Palestine and Russia/Ukraine ... for instance.

Guardian
vs.
The Daily Telegraph

Hamas and Israel at war: what we know on day 20

Netanyahu says Israeli military 'getting prepared' for ground invasion; WHO asks Hamas for proof of life for hostages

As mobs cheer on Hamas, Western civilisation is imploding around us

Shocking anti-Semitism and widespread anti-Israel bias show how sick our societies have become

Types of corpus

Sample & Population

- We often make a distinction between a **sample** and its **population**
- **Population:** All instances of something
- **Sample:** some subset of that
- A lot of science is based on **sampling** from a population and making inferences about the population from a sample

Sample and Population

- Buts there's of ways that the inference from sample to population can be problematic
- Samples vary in terms of how representative they are of the population
- We talk about **sampling procedures** in science in general that are designed to overcome or attenuate these problems
- Samples can be biased ('sampling bias' refers to a property of a sampling procedure which is thought to result in systematic bias)

Representativeness

- A corpus is a **sample** of some **population**
- How representative is it of that population?
- This requires us to think or have some theory about the structure of the population.

‘Skewing’

- Let’s say I want to study contemporary English
- I tell you that the verb *begot* is extremely frequent in English.
- But then I tell you that I only relied on Biblical texts to come to that conclusion.
- Maybe this is just a property of biblical texts not of English in general

Balance

- We should try to find balance in the sample, but we do this according to some theory of what the population is like
- What proportion of our corpus should contain science fiction comopared to speeches given by politicians?
- The important point is that its obvious that language use varies according to genre and register etc.

Parameters of a corpus

- **Representativeness**: how representative is the corpus of its population
- **Spontaneity**: Texts can be more or less spontaneous (a planned speech versus a conversation)
- **Routinized**: Texts can be more or less routinized (e.g. prayers vs. creative writing)
- **Saturation**: The property whereby as a text grows, the probability of seeing a new construction decreases (only true if we are looking at different types of text)

Spontaneity

- How might a spontaneous conversation differ from a planned one?

Spontaneity

- How might a spontaneous narrative differ from a planned one?
    - More speech errors
    - Less complex sentences?
    - More reduction
    - Different vocabulary
    - Less repetition?

Routinization

- How might routinized speech compare to less routine speech?

Routinization

- How might routinized speech compare to less routine speech?
  - -routinized speech has more reduction?
  - more idiomatic expressions?
  - Routinized speech is likely faster ...

Register

- **mode:** spoken versus signed
- **participants:** who are the people involved and what relationships do they have to one another
- **communicative purpose:** why are you talking to one another

https://en.wikipedia.org/wiki/American_Sign_Language

Genre

- A genre is defined by linguistic features that are conventionally used with it.
- Prose writing, poetry, fairy tales
- Note that genres are culturally specific (they usually have names)
- **Genre marker:** A specific property that marks off a genre

Genre

- In Chácobo there is a specific ‘male joking genre’ of speech, which is characterized heavily by the liberal use of adjectivalizers.
- Typically between men drinking chicha
- Some sentences which appear in this context are judged unacceptable or ungrammatical outside of it.

jënë pohi=xëni mia
chicha shit-ter you
“You are a habitual chicha shitter.”

Style

- Like genre but about individuals:
- Some people just use certain constructions more than others
- These can be markers of social identity, but they can also be very specific.

Linked data

- Rawa vs. primary data
- Time alignment
- Metadata