Corpus Linguistics

Introduction to Corpus Linguistics
Corpora, Creation & Applications

Niko Schenk

Institut für England- und Amerikastudien
Goethe-Universität Frankfurt am Main
Winter Term 2017/18

November 1, 2017

1 Introduction

2 Corpora
- Properties
- Creation
- A List of Available Corpora
- Corpus Linguistics—Cases of Application

1 Introduction

2 Corpora
- Properties
- Creation
- A List of Available Corpora
- Corpus Linguistics—Cases of Application

What is a Corpus?

An informal definition:

In linguistics,
a **corpus** *(plural:* **corpora***) is a large collection of texts.*

- *Usually, a corpus consists of smaller units which are called* **documents***.*

Corpora We Have Seen So Far

1. Google Books Corpus
   - `http://googlebooks.byu.edu/x.asp`
     1.3 million books (155 billion words) for American English
     "How many books are there in the world?" [1]
   - Software to search the books: Google NGram Viewer
     (`https://books.google.com/ngrams`)
2. "The Web"

[1] `http://www.fastcompany.com/1678254/how-many-books-are-there-world`

What is Corpus Linguistics?

- The objective is to **use corpora** to
  - investigate (compare) interesting linguistic phenomena
  - to find useful patterns in the data
- Usually, you differentiate between two approaches (cf. previous lecture slides)
  - Hypothesis-**testing** methods.
  - Hypothesis-**generating** methods.
- **Software** is used by linguists to **analyze corpora**.
  - The primary method applied to texts is **SEARCH**.
  - As a result, we obtain instances of the desired phenomena + **frequencies**.

Last Session Revisited

Google offers specialized (exploratory) search as a corpus linguistic application **for digitized books**:

Google Ngram Viewer[2]

- We inspected a particular linguistic phenomenon: *thrived vs. throve*

[2] `http://books.google.com/ngrams/`

Google books Ngram Viewer

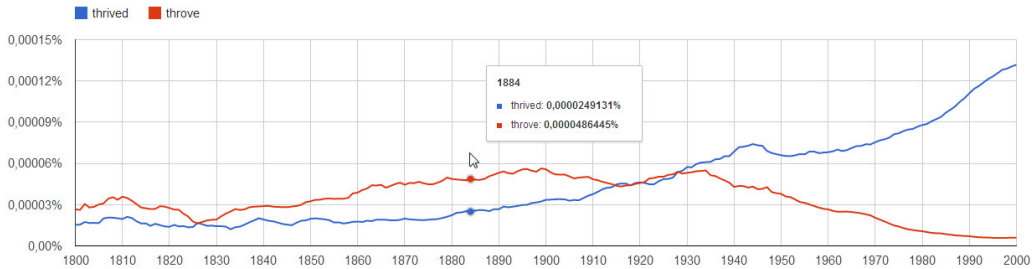
Graph these case-sensitive comma-separated phrases: thrived,throve

between 1800 and 2000 from the corpus English with smoothing of 3 .

Search lots of books

Share 0

Tweet 0



1 Introduction

2 Corpora
- Properties
- Creation
- A List of Available Corpora
- Corpus Linguistics—Cases of Application

Corpus Properties

Requirement:

- The texts should be **electronically stored** (**as text(!) files**).
  - → efficiently **processable by a computer** (search).
  - 1. fast
  - 2. space-efficient
  - 3. accurate
  - 4. deterministic

Corpus Properties

Requirement:

- The collection should be **large**. (What counts as "large"?)
  - → **quantitative**, instead of theoretical analysis of language.
    (you can count the phenomena that you see in the corpus)
  - We want to verify/falsify linguistic theories based on large amounts of linguistic data.

Corpus Properties

Requirement:

- The texts should contain **authentic + representative language examples**.
  - → basis for **linguistic analysis**.
    (researchers do not have to make up their own artificial examples)

Corpus Properties cont'd

- **Language**
  - mono-lingual, bilingual, multi-lingual
- **Contents, type**
  - literature, newspaper, contemporary data, spoken, written, learner data, etc.
- **Time period of the data**
  - Historical novels vs. WhatsApp chat history
  - Note that the time period of a corpus is different from the creation time of a corpus. e.g., a 17th century novel digitized by state-of-the art corpus tools.
- **Licenses, member fee**
- **Availability** (online vs. local)

Corpus Properties cont'd

- ...
- **Meta data** (title, document description, linguistic annotations such as verbs, nouns, etc.)
- **Corpus tools** (yes, no), data format
  - searchable for words, synonyms, collocations, etc.
  - export format / compatibility with other tools
- **Balanced vs. not balanced**
  - i.e. an equal amount of all different phenomena researchers are interested in.
  - (It does not make sense to collect spoken language data only from children if one is interested in an overall picture including young and old speakers.)
- **Automatically vs. manually generated**
  - automatically vs. manually post-processed
  - book scanner/character recognition involved?

1 Introduction

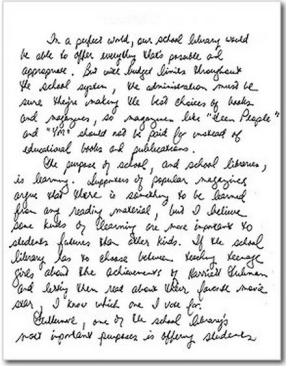
2 Corpora
- Properties
- Creation
- A List of Available Corpora
- Corpus Linguistics—Cases of Application

Corpus data can be collected from various sources



Figure: Books and literature...

Corpus data can be collected from various sources



In a perfect world, our school library would be able to offer everything that's possible and appropriate. But with budget limits throughout the school system, the administration must be sure that in making the best choices of books and magazines, so magazines like "Teen People" and "YM" should not be paid for instead of educational books and publications.

The purpose of school, and school libraries, is learning. Supporters of popular magazines argue that there is something to be learned from any reading material, but I believe some kinds of learning are more important to students' futures than other kinds. If the school library has to choose between reading teenage girls' about the achievements of Harriet Tubman and letting them read about their favorite movie star, I know which one I vote for.

Furthermore, one of the school library's most important purposes is offering students

Figure: Student essays...

Corpus data can be collected from various sources



Corpus data can be collected from various sources



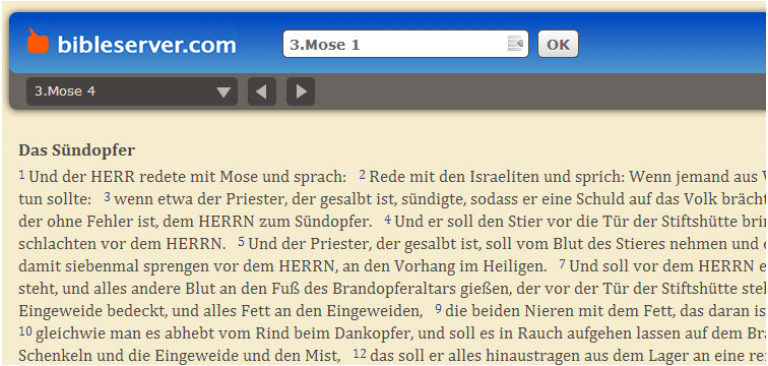
Corpus data can be collected from various sources



Corpus data can be collected from various sources



Corpus data can be collected from various sources



Corpus data can be collected from various sources

Songwriters: BLAIR, PAUL EDWARD / GERMANOTTA, STEFANI J. / BRESSO, MARTIN / MONSON, NICK / ZISIS, DINO

I stand here waiting for you to bang the gong
To crash the critic saying, "is it right or is it wrong?"
If only fame had an IV, baby could I bear
Being away from you; I found the vein, put it in here

I live for the applause, applause, applause
I live for the applause-plause, live for the applause-plause
Live for the way that you cheer and scream for me
The applause, applause, applause

Corpus data can be collected from various sources

Run-Length Compressed Indexes Are Superior for Highly Repetitive Sequence Collections

Jouni Sirén[1*], Niko Välimäki[1**], Veli Mäkinen[1**], and Gonzalo Navarro[2***]

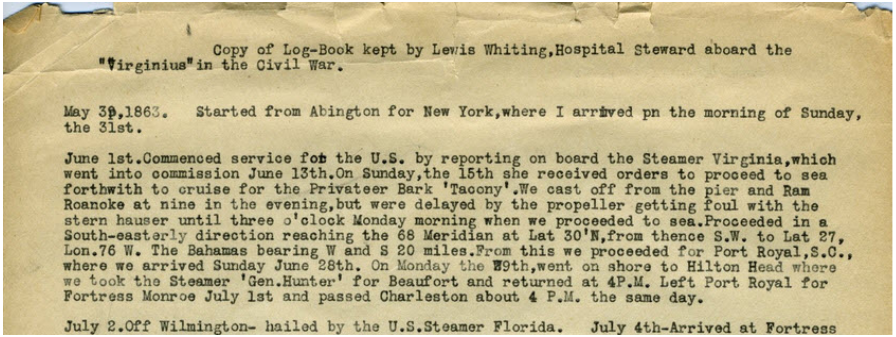
[1] Dept. of Computer Science, Univ. of Helsinki, Finland.
{jltsiren,nvalimak,vmakinen}@cs.helsinki.fi
[2] Dept. of Computer Science, Univ. of Chile. gnavarro@dcc.uchile.cl

Abstract. A repetitive sequence collection is one where portions of a *base sequence* of length n are repeated many times with small variations, forming a collection of total length N. Examples of such collections are version control data and genome sequences of individuals, where the differences can be expressed by lists of basic edit operations. This paper is devoted to studying ways to store massive sets of highly repetitive sequence collections in space-efficient manner so that retrieval of the content as well as queries on the content of the sequences can be provided time-efficiently. We show that the state-of-the-art entropy-bound full-text *self-indexes* do not yet provide satisfactory space bounds for this specific task. We engineer some new structures that use run-length encoding and give empirical evidence that these structures are superior to the current structures.

1 Introduction

Self-indexing [9, 5, 24, 20] is a new algorithmic approach to storing and retrieving sequential data. The idea is to represent the text (a.k.a. sequence or string)

Corpus data can be collected from various sources



Copy of Log-Book kept by Lewis Whiting,Hospital Steward aboard the "Virginius" in the Civil War.

May 30,1863. Started from Abington for New York,where I arrived pn the morning of Sunday, the 31st.

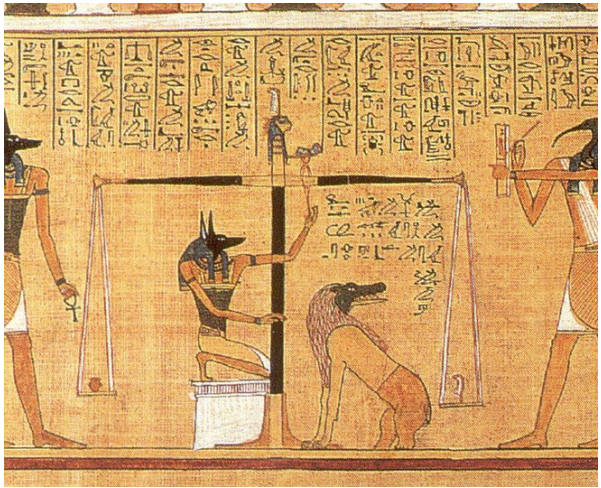
June 1st.Commenced service for the U.S. by reporting on board the Steamer Virginia,which went into commission June 13th.On Sunday,the 15th she received orders to proceed to sea forthwith to cruise for the Privateer Bark 'Tacony'.We cast off from the pier and Ram Roanoke at nine in the evening,but were delayed by the propeller getting foul with the stern hauser until three o'clock Monday morning when we proceeded to sea.Proceeded in a South-easterly direction reaching the 68 Meridian at Lat 30'N,from thence S.W. to Lat 27, Lon.76 W. The Bahamas bearing W and S 20 miles.From this we proceeded for Port Royal,S.C., where we arrived Sunday June 28th. On Monday the 29th,went on shore to Hilton Head where we took the Steamer 'Gen.Hunter' for Beaufort and returned at 4P.M. Left Port Royal for Fortress Monroe July 1st and passed Charleston about 4 P.M. the same day.

July 2.Off Wilmington- hailed by the U.S.Steamer Florida. July 4th-Arrived at Fortress

Corpus data can be collected from various sources



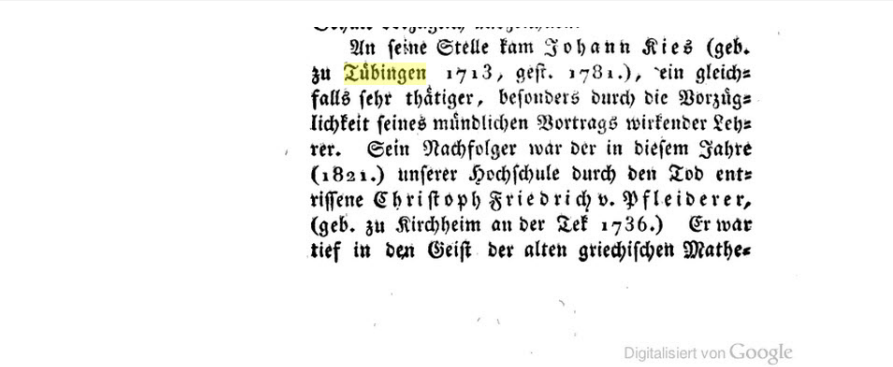
Corpus data can be collected from various sources



Corpus data can be collected from various sources



Corpus data can be collected from various sources



An seine Stelle kam Johann Kies (geb. zu Tübingen 1713, gest. 1781.), ein gleichfalls sehr thätiger, besonders durch die Vorzüglichkeit seines mündlichen Vortrags wirkender Lehrer. Sein Nachfolger war der in diesem Jahre (1821.) unserer Hochschule durch den Tod entrissene Christoph Friedrich v. Pfleiderer, (geb. zu Kirchheim an der Teck 1736.) Er war tief in den Geist der alten griechischen Mathe-

Digitalisiert von Google

Figure: *"Beschreibung und Geschichte der Universität und Stadt Tübingen."* as a *Google* Books document

Corpus Data

Corpus data can be collected from various sources:

E.g., books, papers, letters, news feeds from the internet, **spoken language**, dialogues, reports, twitter data, Facebook posts, customer reviews, chat data, historical texts, homework exercises, student exams, academic literature, song lyrics, bible verses, biological data, etc.

Remember that they need to be **electronically available**.
Why? $\rightarrow$ **Only digitized texts are efficiently searchable!**

How are Corpora Created?

→ All of the previously introduced "text types" are interesting language data.
Goal: Generate computer-processable (electronically-stored) text files / **a corpus**.
Question: How would you proceed?

How are Corpora Created?—Conversion Examples

1. electronically available
   1. text file → done
   2. e.g., PDF/image → *OCR*[3] → text file
   3. e.g., audio file → *speech-to-text*[4] → text file
2. **not** electronically available
   1. manually written/printed texts
      - e.g., student essays on paper → manually typewrite / *handwriting recognition* → text file
      - e.g., historical books → digitize (cf. *books scanner*[5] [6]) → image → OCR → text file
   2. spoken language
      - e.g., radio interview → manually typewrite → text file
      - e.g., phone conversation → *speech-to-text* → text file

[3] Optical Character Recognition
[4] cf. *Siri*
[5] Google Books Scanner, 03:35min
[6] Another scanner

How are Corpora Created?

Imagine you had to build up your own corpus. How would you proceed?

Some guidelines:

- Corpora should be built using (semi-)automated processes.
  - E.g., copying news feeds **manually** from the Internet is not elegant. Use web crawlers instead.
- Corpora should be balanced.
- Corpora should contain real world examples.
- Corpora should be very large.
- Corpora should have a proper format. (advanced)

1 Introduction

2 Corpora
- Properties
- Creation
- A List of Available Corpora
- Corpus Linguistics—Cases of Application

A List of Available Corpora[a]

[a] no exhaustive list, sorted by size; references: 1, 2

	Properties			
Corpus	language	words	time period	type
Google's N-Gram Corpus[7]	English	1.024 trillion	-	web data
Google Books Corpus	AE/BE	155/34 billion	1500s-2000s	historical, contemporary books
Global Web-Based English (GloWbE)	20 countries	1.9 billion	2012-2013	web pages
Corpus of Contemporary AE (COCA)	AE	450 million	1990-2012	spoken, fiction, magazines, news, acad texts
British National Corpus (BYU-BNC)	BE	100 million	1980s-1993	representative sample of written/spoken BE
Corpus of American Soap Operas	AE	100 million	2001-2012	film dialogues
Strathy Corpus	Canadian English	50 million	1970s-2000	spoken, fiction, magazines, newspapers, academic texts.
My S-21 Facebook Corpus	German	50 million	2010-2013	UGC, web data
Corpus do Português	Portuguese	45 million	1300s-1900s	newspaper academic texts
Canadian Hansard Corpus	English, French	26 million	1986-1987	parallel corpus, parliament debates
International Corpus of Learner English	English 16 native langs	3.7 million	2002	essays written by learners of English

[7] (*Web 1T 5-gram Version 1*, only n-grams available, not the corpus itself)

More Corpora...[a]

[a] source

There exist specialized corpora for almost all commonly known languages...

- Bergen Corpus of London Teenager Language
- KidPub, ("Collection of stories written by kids from all over the planet")
- Movie Review Corpus
- Facebook Status Messages Corpus
- Enron Email Corpus
- Japanese Speech Corpora of Major City Dialects
- The Complete Corpus of Old English
- The Blog Authorship Corpus
- CoRD, Corpus of Early English Medical Writing (CEEM)
- The York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE)
- ...

1 Introduction

2 Corpora
- Properties
- Creation
- A List of Available Corpora
- Corpus Linguistics—Cases of Application

Typical Applications of Corpus Linguistics

Having this large collection of digitized texts, books, etc...

What can you do with it?

...with a Focus on *Linguistic* Research

Typical research questions:

- Is passive tense used more often in spoken language or in academic writing?
- What properties have adjectives which co-occur with "rather" compared to those co-occurring with "fairly"?
- What are the most frequent word categories in the German *Vorfeld*? Is there a difference to English?
- Is the German *dative -e* still present in 2013? (*Wie es im Buche steht.*)
- Do emails contain more spelling mistakes than newspaper texts?
- Is "*ain't*" more frequently used in BE or in AE?
- Topicalized vs. non-topicalized constructions (*All these foreign cars I drive...*)
- Comparing syntactic constructions in song texts among song writers.

...with a Focus on *Linguistic* Research

- **Diachronic corpus data**
  - See how frequency of word usage changes over time.[8]
  - Check which syntactic constructions or word combinations become more prominent/less frequent.[9]
- **Lexicography/language use**
  - Find new words which appeared recently.
  - Find words and phrases which co-occur. (idiomatic expressions)
  - Compare slang to formal language, etc.
- **Analyze word meaning**
  - Lookup a word and its contexts—depending on the context, a word can have different meanings.

[8] https://books.google.com/ngrams/

[9] http://members.unine.ch/martin.hilpert/motion.html

...with a Focus on Language Learning

- **As a learner:**
  - Foreign language learning technique (Google!)
  - Check which constructions are correct and which are incorrect
    e.g., *****_ten items or_ **less** vs. _ten items or_ **fewer**
    e.g., _make a speech_ vs. _give a speech_
    e.g., *****_more strict_ vs. _stricter_
  - Get to know different meanings of same word
  - Get to know correct word position within the sentence (_"yet"_)

...with a Focus on Language Learning

- **As a teacher:**
  - Is a certain construction "grammatical"?
    (avoid answers like: "it just sounds better...")
  - Propose appropriate synonyms for a particular word
  - Make students learn most frequent constructions first (broader coverage)
  - What are the most typical errors by learners of German?

...with a Focus on Language Learning

- **As a researcher:**
  - Do English students (learning German) have the same problems with to-infinitives compared to native speakers of Spanish?
  - What are the most prominent/problematic grammatical constructions for language learners in their 2nd year?
  - Sociolinguistics, dialectology—e.g., comparison of European and Brazilian Portuguese

...with a Focus on (Computational) Information Retrieval

- **Authorship detection**
  - Which linguistic properties are relevant to identify the author of a particular text fragment? Is the average sentence length indicative of a particular author? How about the average number of noun phrases? Vocabulary? Function words?
- **"NSA-related"**
  - Email corpus: which keywords in a particular email could potentially be relevant/alarming regarding terrorism prevention.
  - Email corpus: spam detection/priority inbox
- **Advertisement**
  - Which words/phrases of your Facebook status messages are relevant indicators for sending appropriate advertisement to you?
  - Given your previous Google search history, what are you likely to type in/search next? (*Golf fahren* vs. *Golf spielen*)

...with a Focus on (Computational) Information Retrieval

Automated statistical methods:

- Find long repetitions (e.g., plagiarism detection[10], biological data analysis)
- Keyword extraction, terminology detection
- Automatically find synonyms, antonyms, etc.
- Spell checkers (propose alternative/next words/**autocomplete**).
- Speech recognition, *Apple's Siri*
- Machine translation (cf. aligned corpora)
- Dialectometry
- Collocation & collostruction analysis
  i.e. word–word and word–syntax associations
- Word clustering (Monday, Tuesday, . . . , automatically find semantically related words)
- Ontology creation (e.g., WordNet)

[10] Gutenplag forum

Homework Assignment

Task 1: Assume, you are given a diverse set of language data, e.g.,

- a set of your homework assignments produced on the computer
- a collection of newspaper articles
- a list of student essays from your own class
- a *WhatsApp* history of conversations with your best friends on your mobile phone
- a political speech recorded from the radio program
- a section of the *"Egyptian Book of the Dead"* written on papyrus
- a collection of PDF user manuals from the automobile sector

Task: You are supposed to digitize the data. (Only this way, you can search it by means of a computer). **For each item on the list**, how would you proceed? Also, describe the type of language data, their **linguistic characteristics** in closer detail.

Homework Assignment

Task 2: Read through the materials on the Google Ngram Viewer page: `http://books.google.com/ngrams/info#advanced` and use the software (`http://books.google.com/ngrams`) to come up with **two** linguistically interesting examples showing differences in the distributions of terms. You should come up with a detailed explanation for the trend you see.

For example, the following illustrates that math and biology have been traditional disciplines whereas computational linguistics, for example, is quite new: `http://books.google.com/ngrams/graph?content=Linguistik%2CInformatik%2CBiologie%2CGermanistik%2CComputerlinguistik%2CMathematik&year_start=1800&year_end=2000&corpus=20&smoothing=3&share=`

Moreover, interpret these two examples: `http://books.google.com/ngrams/graph?content=Marc+Chagall&year_start=1800&year_end=2000&corpus=20&smoothing=3&share=`

`https://books.google.com/ngrams/graph?content=Eminem&year_start=1800&year_end=2000&corpus=15&smoothing=3&share=&direct_url=t1%3B%2CEminem%3B%2Cc0`