Corpus Linguistics

Use Cases, Corpus Creation, Applications

Niko Schenk

`n.schenk@em.uni-frankfurt.de`

Applied Computational Linguistics Lab
Computer Science Department /
Department of English- and American Studies
Goethe University Frankfurt, Germany

April 24, 2019

1. Introduction

2. Corpus Properties, Text Digitization, Applications
   - Properties
   - Creation
   - A List of Available Corpora
   - Corpus Linguistics—Cases of Application

1 Introduction

2 Corpus Properties, Text Digitization, Applications

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

What is a Corpus?

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**.

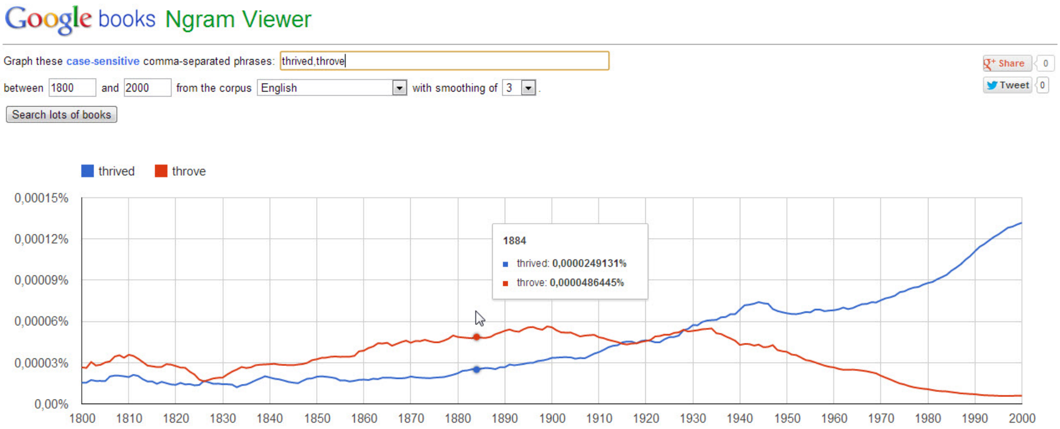
Last Session Revisited

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.

1884

■ thrived: 0,0000249131%

■ throve: 0,0000486445%

1 Introduction

2 Corpus Properties, Text Digitization, Applications

- 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?

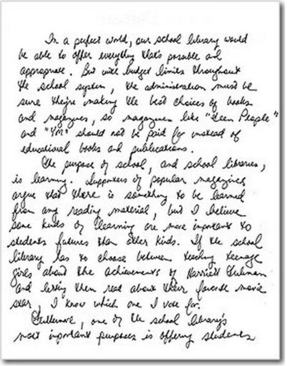
1. Introduction

2. Corpus Properties, Text Digitization, Applications
   - 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 they're 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 teaching teenage girls about the achievements of Harriett 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

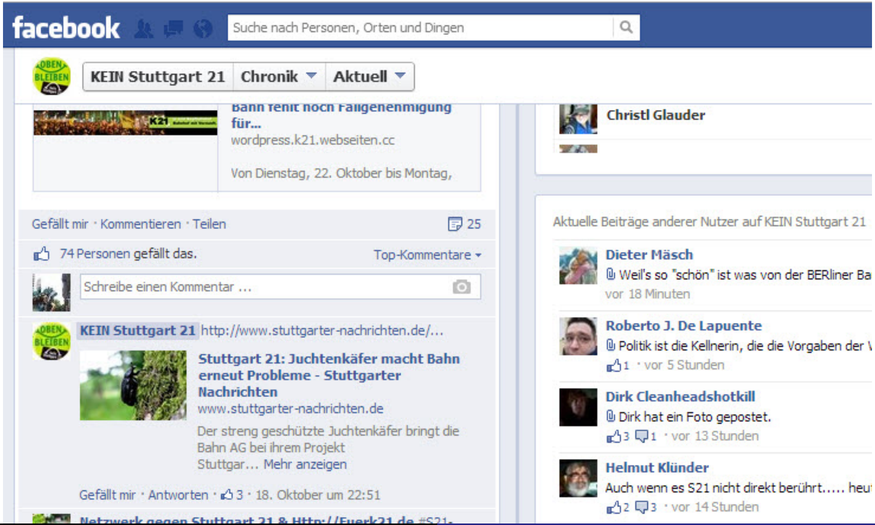
Figure: Student essays...

Corpus data can be collected from various sources

Corpus data can be collected from various sources

Today's News
Up to the minute information
Departments
INTERNATIONAL
UNITED STATES
LOCAL
BUSINESS
SPORTS
POLITICS
ENTERTAINMENT
TECHNOLOGY
LIFESTYLE
HEALTH
WEATHER
YOUR WORD
Headlines
President signs Crime Bill into Law
Space Shuttle Returns
Democrats on the Move
Terror Grips Jerusalem
Stocks Rise Again
Wildfires Spread in Midwest
New Social Networking Craze Hits Web
Gasoline Prices Drop For Second Week
Housing Stabilizes
This Weekend's Box Office
Single and Loving It
Students Beginning the Return to School
Across the USA, children are readying themselves for another year of learning.
Go

Corpus data can be collected from various sources



Bahn fehlt noch Fällgenehmigung für...
wordpress.k21.webseiten.cc

Von Dienstag, 22. Oktober bis Montag,

Gefällt mir · Kommentieren · Teilen 25

74 Personen gefällt das. Top-Kommentare

Schreibe einen Kommentar ...

KEIN Stuttgart 21 http://www.stuttgarter-nachrichten.de/...

Stuttgart 21: Juchtenkäfer macht Bahn erneut Probleme - Stuttgarter Nachrichten
www.stuttgarter-nachrichten.de

Der streng geschützte Juchtenkäfer bringt die Bahn AG bei ihrem Projekt Stuttgar... Mehr anzeigen

Gefällt mir · Antworten · 3 · 18. Oktober um 22:51

Netzwerk gegen Stuttgart 21 & Http://Fuerk21.de #S21-

Christl Glauder

Aktuelle Beiträge anderer Nutzer auf KEIN Stuttgart 21

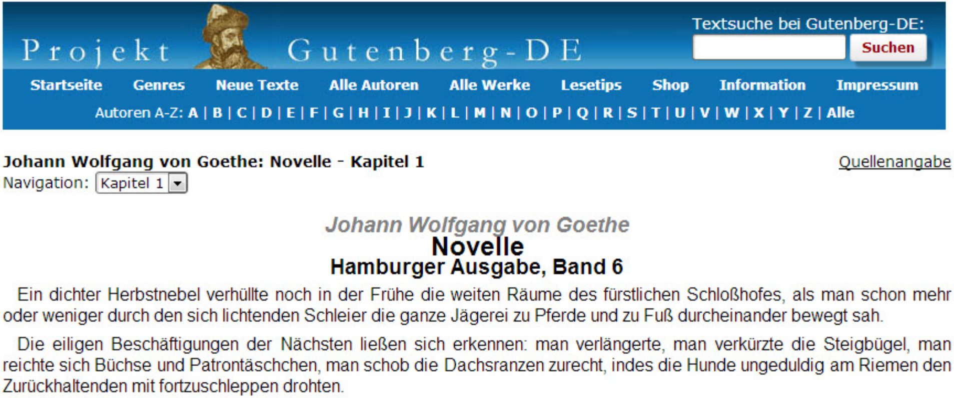
Dieter Mäsch
Weil's so "schön" ist was von der BERliner Ba
vor 18 Minuten

Roberto J. De Lapuente
Politik ist die Kellnerin, die die Vorgaben der \
1 · vor 5 Stunden

Dirk Cleanheadshotkill
Dirk hat ein Foto gepostet.
3 1 · vor 13 Stunden

Helmut Klünder
Auch wenn es S21 nicht direkt berührt..... heu
2 3 · vor 14 Stunden

Corpus data can be collected from various sources


Projekt Gutenberg-DE
Textsuche bei Gutenberg-DE:
Suchen
Startseite
Genres
Neue Texte
Alle Autoren
Alle Werke
Lesetips
Shop
Information
Impressum
Autoren A-Z: A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z | Alle
Quellenangabe
Navigation: Kapitel 1

Johann Wolfgang von Goethe: Novelle - Kapitel 1

Johann Wolfgang von Goethe

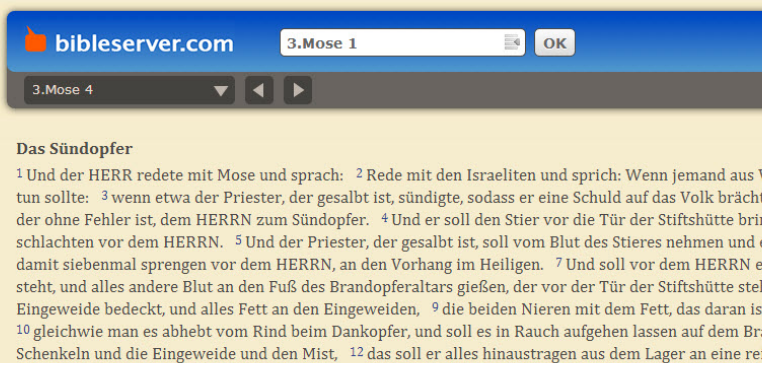
Novelle

Hamburger Ausgabe, Band 6

Ein dichter Herbstnebel verhüllte noch in der Frühe die weiten Räume des fürstlichen Schloßhofes, als man schon mehr oder weniger durch den sich lichtenden Schleier die ganze Jägerei zu Pferde und zu Fuß durcheinander bewegt sah.

Die eiligen Beschäftigungen der Nächsten ließen sich erkennen: man verlängerte, man verkürzte die Steigbügel, man reichte sich Büchse und Patrontäschchen, man schob die Dachsranzen zurecht, indes die Hunde ungeduldig am Riemen den Zurückhaltenden mit fortzuschleppen drohten.

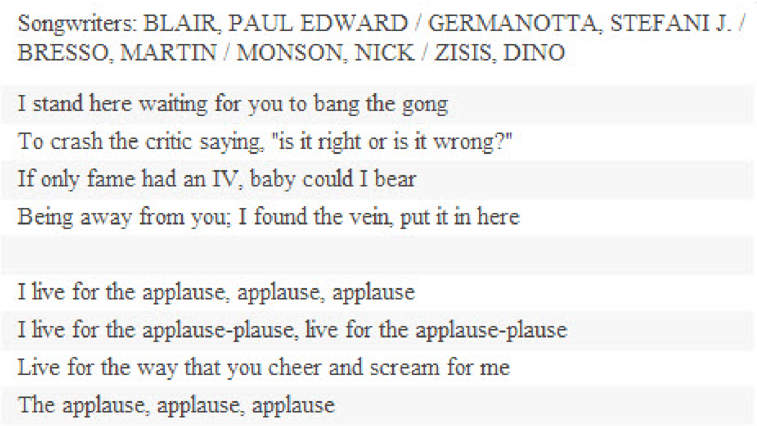
Corpus data can be collected from various sources



Das Sündopfer

[1] Und der HERR redete mit Mose und sprach: [2] Rede mit den Israeliten und sprich: Wenn jemand aus V
tun sollte: [3] wenn etwa der Priester, der gesalbt ist, sündigte, sodass er eine Schuld auf das Volk brächt
der ohne Fehler ist, dem HERRN zum Sündopfer. [4] Und er soll den Stier vor die Tür der Stiftshütte bri
schlachten vor dem HERRN. [5] Und der Priester, der gesalbt ist, soll vom Blut des Stieres nehmen und e
damit siebenmal sprengen vor dem HERRN, an den Vorhang im Heiligen. [7] Und soll vor dem HERRN e
steht, und alles andere Blut an den Fuß des Brandopferaltars gießen, der vor der Tür der Stiftshütte stel
Eingeweide bedeckt, und alles Fett an den Eingeweiden, [9] die beiden Nieren mit dem Fett, das daran is
[10] gleichwie man es abhebt vom Rind beim Dankopfer, und soll es in Rauch aufgehen lassen auf dem Br
Schenkeln und die Eingeweide und den Mist, [12] das soll er alles hinaustragen aus dem Lager an eine re

Corpus data can be collected from various sources

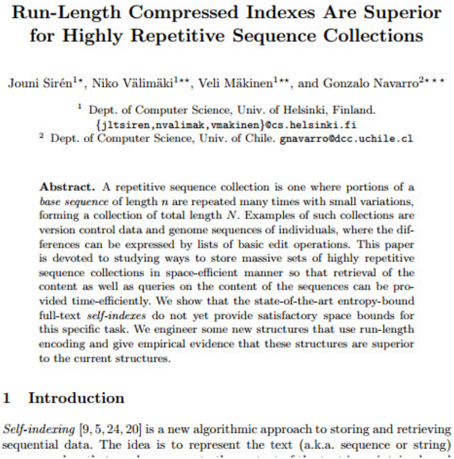


Songwriters: BLAIR, PAUL EDWARD / GERMANOTTA, STEFANI J. / BRESSO, MARTIN / MONSON, NICK / ZISIS, DINO

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]***

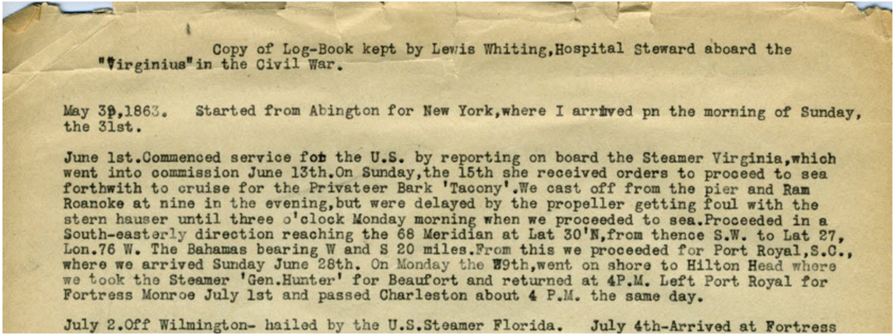
[1] Dept. of Computer Science, Univ. of Helsinki, Finland.
{jltsiren,nvalimak,vmakinen}@cs.helsinki.fi
[2] Dept. of Computer Science, Univ. of Chile. gnavarro@dcc.uchile.cl

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



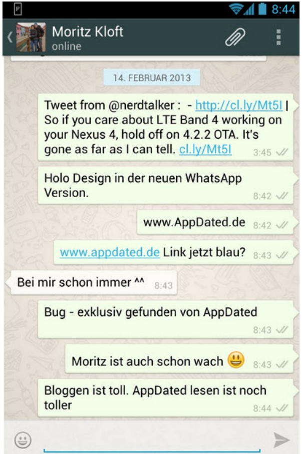
Copy of Log-Book kept by Lewis Whiting,Hospital Steward aboard the "Virginius"in the Civil War.

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



14. FEBRUAR 2013

Tweet from @nerdtalker : - http://cl.ly/Mt5I | So if you care about LTE Band 4 working on your Nexus 4, hold off on 4.2.2 OTA. It's gone as far as I can tell. cl.ly/Mt5I 3:45

Holo Design in der neuen WhatsApp Version. 8:42

www.AppDated.de 8:42

www.appdated.de Link jetzt blau? 8:43

Bei mir schon immer ^^ 8:43

Bug - exklusiv gefunden von AppDated 8:43

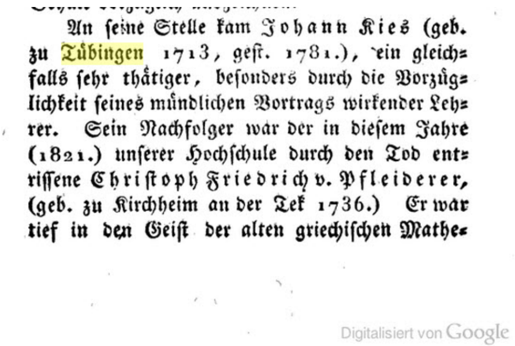
Moritz ist auch schon wach 😃 8:43

Bloggen ist toll. AppDated lesen ist noch toller 8:44

Corpus data can be collected from various sources

Corpus data can be collected from various sources

Corpus data can be collected from various sources



[illegible]

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 Tek 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? → **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. Corpus Properties, Text Digitization, Applications
   - Properties
   - Creation
   - A List of Available Corpora
   - Corpus Linguistics—Cases of Application

A List of Available Corpora[8]

	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)

[8]

More Corpora...[9]

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)
- . . .

[9]

1. Introduction

2. Corpus Properties, Text Digitization, Applications
   - 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 Buch***e** *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.[10]
  - Check which syntactic constructions or word combinations become more prominent/less frequent.[11]
- **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.

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

[11] 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[12], 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)

[12] 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