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
Introduction

Corpus Properties, Text Digitization, Applications

- Properties
- Creation
- A List of Available Corpora
- Corpus Linguistics—Cases of Application
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](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](https://books.google.com/ngrams))

2. “The Web”

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

Google Ngram Viewer

- We inspected a particular linguistic phenomenon:
  
  \textit{thrived vs. thaw}
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?
Introduction

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.

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.

Das Sündopfer

1 Und der HERR redete mit Mose und sprach: 2 Rede mit den Israeliten und sprich: Wenn jemand aus ihnen tun sollte: 3 wenn etwa der Priester, der gesalbt ist, sündigte, sodass er eine Schuld auf das Volk bringt, der ohne Fehler ist, dem HERRN zum Sündopfer. 4 Und er soll den Stier vor die Tür der Stiftshütte bringen und schlachten vor dem HERRN. 5 Und der Priester, der gesalbt ist, soll vom Blut des Stieres nehmen und damit siebenmal sprengen vor dem HERRN, an den Vorhang im Heiligen. 7 Und soll vor dem HERRN ein Standbild stehen, und alles andere Blut an den Fuß des Brandopferaltars gießen, der vor der Tür der Stiftshütte steht. 8 Die Eingeweide bedeckt, und alles Fett an den Eingeweiden, 9 die beiden Nieren mit dem Fett, das daran ist, gleichwie man es abhebt vom Rind beim Dankopfer, und soll es in Rauch aufgehen lassen auf dem Brennholz. Schenken und die Eingeweide und den Mist, 12 das soll er alles hinaustragen aus dem Lager an eine rei...
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 Sirén1*, Niko Välimäki1**, Veli Mäkinen1**, and Gonzalo Navarro2***

1 Dept. of Computer Science, Univ. of Helsinki, Finland. {jlsiren, nvalimak, vmakinen}@cse.helsinki.fi
2 Dept. of Computer Science, Univ. of Chile. gnnavarro@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.
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

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!**
→ 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 → text file
   - 3. e.g., audio file → speech-to-text → 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 → 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)
Introduction

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

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Language</th>
<th>Words</th>
<th>Time Period</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google’s N-Gram Corpus&lt;sup&gt;7&lt;/sup&gt;</td>
<td>English</td>
<td>1.024 trillion</td>
<td>-</td>
<td>web data</td>
</tr>
<tr>
<td>Google Books Corpus</td>
<td>AE/BE</td>
<td>155/34 billion</td>
<td>1500s-2000s</td>
<td>historical, contemporary books</td>
</tr>
<tr>
<td>Global Web-Based English (GloWbE)</td>
<td>20 countries</td>
<td>1.9 billion</td>
<td>2012-2013</td>
<td>web pages</td>
</tr>
<tr>
<td>Corpus of Contemporary AE (COCA)</td>
<td>AE</td>
<td>450 million</td>
<td>1990-2012</td>
<td>spoken, fiction, magazines, news, acad texts</td>
</tr>
<tr>
<td>British National Corpus (BYU-BNC)</td>
<td>BE</td>
<td>100 million</td>
<td>1980s-1993</td>
<td>representative sample of written/spoken BE</td>
</tr>
<tr>
<td>Corpus of American Soap Operas</td>
<td>AE</td>
<td>100 million</td>
<td>2001-2012</td>
<td>film dialogues</td>
</tr>
<tr>
<td>My S-21 Facebook Corpus</td>
<td>German</td>
<td>50 million</td>
<td>2010-2013</td>
<td>UGC, web data</td>
</tr>
<tr>
<td>Corpus do Português</td>
<td>Portuguese</td>
<td>45 million</td>
<td>1300s-1900s</td>
<td>newspaper</td>
</tr>
<tr>
<td>Canadian Hansard Corpus</td>
<td>English, French</td>
<td>26 million</td>
<td>1986-1987</td>
<td>academic texts, parallel corpus, parliament debates</td>
</tr>
<tr>
<td>International Corpus of Learner English</td>
<td>English</td>
<td>3.7 million</td>
<td>2002</td>
<td>essays written by learners of English</td>
</tr>
</tbody>
</table>

<sup>7</sup> *Web 1T 5-gram Version 1*, only n-grams available, not the corpus itself

<sup>8</sup> No exhaustive list, sorted by size; references: 1, 2
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)
- \(\ldots\)
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?
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.\(^{10}\)
  - 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
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?
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
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)
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 & collocation 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