Introduction to corpora and phonetic databases and how to access them

Katrin Schweitzer  Kerstin Eckart  Markus Gärtner

Institut für Maschinelle Sprachverarbeitung
Pfaffenwaldring 5b
70569 Stuttgart

MGK academic courses
April 17th, 2015
The INF-project: Information Infrastructure

- PIs: Jonas Kuhn, Grzegorz Dogil, Sebastian Pado
- Researchers: Kerstin Eckart, Markus Gärtner, Katrin Schweitzer

Project responsibilities

- accumulate large collection of language data
- support optimal exploitation of these data
- making them available to the research community
Thematic focus

This course will not...

• provide a general introduction to corpus linguistics theory
  ⇒ cf. e.g. [Lemnitzer and Zinsmeister, 2006]

• include hands-on instructions for one specific corpus, database or query tool
  ⇒ follow-up training possible

This course intends however, ...

• to provide you with the ability to find available corpora/phonetic databases and assess their usefulness for your task

• to make you come up with new ideas regarding what can be done using corpus data

• to at least provide you with some useful links
Thematic focus

This course will not . . .

• provide a general introduction to corpus linguistics theory
  ⇒ cf. e.g. [Lemnitzer and Zinsmeister, 2006]

• include hands-on instructions for one specific corpus, database or query tool
  ⇒ follow-up training possible
Thematic focus

This course will not . . .

• provide a general introduction to corpus linguistics theory
  ⇒ cf. e.g. [Lemnitzer and Zinsmeister, 2006]
• include hands-on instructions for one specific corpus, database or query tool
  ⇒ follow-up training possible

This course intends however, . . .

• to provide you with the ability to find available corpora/phonetic databases and assess their usefulness for your task
• to make you come up with new ideas regarding what can be done using corpus data
Thematic focus

This course will not . . .

• provide a general introduction to corpus linguistics theory
generated from [Lemnitzer and Zinsmeister, 2006]

• include hands-on instructions for one specific corpus,
database or query tool
generated from follow-up training possible

This course intends however, . . .

• to provide you with the ability to find available corpora/
phonetic databases and assess their usefulness for your task

• to make you come up with new ideas regarding what can be
done using corpus data

• to at least provide you with some useful links
Overview

1 Introduction
2 Classification
3 Motivation: Example studies
4 Finding a corpus
5 Using a corpus
6 Tips
7 References
Defining corpora

A corpus is a collection of written or spoken utterances. Corpus data have typically been digitized, i.e. they are machine-readable and stored on a computer. A part of a corpus, i.e. a text, consists of [primary] data as well as possible metadata to describe the [primary] data and linguistic annotations attached to the [primary] data.

Defining corpora

A corpus is a collection of written or spoken utterances. Corpus data have typically been digitized, i.e. they are machine-readable and stored on a computer. A part of a corpus, i.e. a text, consists of [primary] data as well as possible metadata to describe the [primary] data and linguistic annotations attached to the [primary] data.

In this course we take a broad pragmatic view:

- A corpus is a resource, based on primary data (textual, audio, video, multi-modal, . . .) ⇒ including phonetic databases
Defining corpora

A corpus is a collection of written or spoken utterances. Corpus data have typically been digitized, i.e. they are machine-readable and stored on a computer. A part of a corpus, i.e. a text, consists of [primary] data as well as possible metadata to describe the [primary] data and linguistic annotations attached to the [primary] data.

In this course we take a broad pragmatic view:

- A corpus is a resource, based on primary data (textual, audio, video, multi-modal, . . .) ⇒ including phonetic databases
Defining corpora

A corpus is a collection of written or spoken utterances. Corpus data have typically been digitized, i.e. they are machine-readable and stored on a computer. A part of a corpus, i.e. a text, consists of [primary] data as well as possible metadata to describe the [primary] data and linguistic annotations attached to the [primary] data.


In this course we take a broad pragmatic view:

- A corpus is a resource, based on primary data (textual, audio, video, multi-modal, . . .) ⇒ including phonetic databases
- There are no restrictions with regard to context (phoneme, sentence, utterance, text, . . .)
Defining corpora

A corpus is a collection of written or spoken utterances. Corpus data have typically been digitized, i.e. they are machine-readable and stored on a computer. A part of a corpus, i.e. a text, consists of [primary] data as well as possible metadata to describe the [primary] data and linguistic annotations attached to the [primary] data.


In this course we take a broad pragmatic view:

- A corpus is a resource, based on primary data (textual, audio, video, multi-modal, ...) ⇒ including phonetic databases
- There are no restrictions with regard to context (phoneme, sentence, utterance, text, ...)
- The primary data has or has not been annotated (yet) ⇒ details on linguistic annotation layers to come in the next courses
Primary data, metadata, annotations

Example: The DIRNDL corpus

metadata

- ResourceTitle: Discourse Information Radio News Database for Linguistic analysis
- Modalities: spoken, written
- Topic: radio news
- Project: SFB 732 A1
- TotalSize: 5 hours, 3221 sentences
- Sparker: professional speakers
- Demographics: 5 male, 4 female

primary data

Die Europäische Union hat sich zu weitgehenden Reformen bis zum Jahr 2009 verpflichtet. Bei einem Festakt in Berlin zum fünfzigsten Jahrestag der
Classification criteria

What is the “right” corpus for my study?
Define requirements based on classification criteria:

- Modality
- Temporal origin of the primary data
- Language
- Genre
- Balance
- Degree of Planning (for spoken data)
- Size
- Context size
- State of annotation
- Technical representation
Classification criteria

- Modality
  - speech
  - text
  - gesture
  - multimodal
  - ...

MGK: Corpora
Classification criteria

- Temporal origin of primary data

**Independent dimensions:**
- historical vs. contemporary
- diachronic vs. synchronic
- static vs. monitor corpus
Classification criteria

- Temporal origin of primary data

<table>
<thead>
<tr>
<th>historical</th>
<th>vs.</th>
<th>contemporary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fuerstinnen-korrespondenz 1.1</td>
<td>example</td>
<td>Corpus of Comparisons in Product Reviews</td>
</tr>
<tr>
<td>Analysis of social networks</td>
<td>corpora</td>
<td>Sentiment analysis</td>
</tr>
</tbody>
</table>
Classification criteria

- Temporal origin of primary data

<table>
<thead>
<tr>
<th>historical</th>
<th>vs.</th>
<th>contemporary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fuerstinnenkorrespondenz 1.1</td>
<td>example</td>
<td>Corpus of Comparisons in Product Reviews</td>
</tr>
<tr>
<td>Analysis of social networks</td>
<td>corpora</td>
<td>Sentiment analysis</td>
</tr>
<tr>
<td></td>
<td>example</td>
<td></td>
</tr>
<tr>
<td></td>
<td>studies</td>
<td></td>
</tr>
</tbody>
</table>

Fuerstinnenkorrespondenz 1.1
- letters of 15 princesses and their male correspondence partners
- from 1546 to 1756
Classification criteria

- Temporal origin of primary data

<table>
<thead>
<tr>
<th>historical</th>
<th>vs.</th>
<th>contemporary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fuerstinnen-korrespondenz 1.1</td>
<td>example</td>
<td>Corpus of Comparisons in Product Reviews</td>
</tr>
<tr>
<td>Analysis of social networks</td>
<td>example</td>
<td>Sentiment analysis</td>
</tr>
</tbody>
</table>

Corpus of Comparisons in Product Reviews

- camera reviews (1707 annotated sentences)
- written between 2005 and 2009
Classification criteria

- Temporal origin of primary data

<table>
<thead>
<tr>
<th>diachronic</th>
<th>vs.</th>
<th>synchronous</th>
</tr>
</thead>
<tbody>
<tr>
<td>several (historical) stages of a language</td>
<td></td>
<td>one stage</td>
</tr>
<tr>
<td>Deutsche Diachrone Baumbank (DDB)</td>
<td>example corpus</td>
<td>British National Corpus (BNC)</td>
</tr>
<tr>
<td>language change</td>
<td>example studies</td>
<td>parser training</td>
</tr>
</tbody>
</table>
Classification criteria

- Temporal origin of primary data

<table>
<thead>
<tr>
<th>diachronic</th>
<th>vs.</th>
<th>synchronous</th>
</tr>
</thead>
<tbody>
<tr>
<td>several (historical) stages of a language</td>
<td></td>
<td>one stage</td>
</tr>
<tr>
<td>Deutsche Diachrone Baumbank (DDB)</td>
<td>example corpus</td>
<td>British National Corpus (BNC)</td>
</tr>
<tr>
<td>language change</td>
<td>example studies</td>
<td>parser training</td>
</tr>
</tbody>
</table>

DDB
- three sub corpora
- Old High German, Middle High German, Early New High German
Classification criteria

- Temporal origin of primary data

<table>
<thead>
<tr>
<th>diachronic</th>
<th>vs.</th>
<th>synchronous</th>
</tr>
</thead>
<tbody>
<tr>
<td>several (historical) stages of a language</td>
<td></td>
<td>one stage</td>
</tr>
<tr>
<td>Deutsche Diachrone Baumbank (DDB)</td>
<td>example corpus</td>
<td>British National Corpus (BNC)</td>
</tr>
<tr>
<td>language change</td>
<td>example studies</td>
<td>parser training</td>
</tr>
</tbody>
</table>

BNC

- only texts starting from 1975
- some imaginative works with continued popularity date back till 1964
Classification criteria

- Temporal origin of the primary data

<table>
<thead>
<tr>
<th>static</th>
<th>vs.</th>
<th>monitor corpora</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIGER Corpus</td>
<td>example</td>
<td>Corpus of Contemporary American English (COCA)</td>
</tr>
<tr>
<td>grammar extraction/</td>
<td>corpora</td>
<td>identification of new words</td>
</tr>
<tr>
<td>induction</td>
<td>studies</td>
<td></td>
</tr>
</tbody>
</table>
Classification criteria

• Temporal origin of the primary data

<table>
<thead>
<tr>
<th>static</th>
<th>vs.</th>
<th>monitor corpora</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIGER Corpus</td>
<td>example corpora</td>
<td>Corpus of Contemporary American English (COCA)</td>
</tr>
<tr>
<td>grammar extraction/induction</td>
<td>example studies</td>
<td>identification of new words</td>
</tr>
</tbody>
</table>

TIGER Corpus

• newspaper texts

• 1990s
Classification criteria

- Temporal origin of the primary data

<table>
<thead>
<tr>
<th>static</th>
<th>vs.</th>
<th>monitor corpora</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIGER Corpus</td>
<td>example corpora</td>
<td>Corpus of Contemporary American English (COCA)</td>
</tr>
<tr>
<td>grammar extraction/induction</td>
<td>example studies</td>
<td>identification of new words</td>
</tr>
</tbody>
</table>

COCA

- content equally divided among spoken data, fiction, popular magazines, news papers, academic texts

- 20 million words per year
Classification criteria

- Language
  - monolingual
  - multilingual: parallel/comparable/other
### Classification criteria

- **Language**
  - monolingual
  - multilingual: parallel/comparable/other

**parallel** translations in different languages, often aligned at sentence or word level

**e.g. Europarl v7, DCEP**

<table>
<thead>
<tr>
<th>en</th>
<th>de</th>
<th>el</th>
<th>pl</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>2973329</strong> It is your task,</td>
<td>Ihre Aufgabe, Herr Kommissar Frattini,</td>
<td>Έχετε υποχρέωση, Επίτροπε Frattini,</td>
<td>Pańskim zadaniem jest,</td>
</tr>
<tr>
<td>Commissioner Frattini - especially as the</td>
<td>ist es gerade als</td>
<td>είστε αρμόδιος Επίτροπος για</td>
<td>panie komisarzu Frattini,</td>
</tr>
<tr>
<td>Commissioner responsible for human</td>
<td>Grundrechtkommissar - die</td>
<td>τα ανθρώπινα δικαιώματα - να</td>
<td>szczególnie jako</td>
</tr>
<tr>
<td>rights - to defend the</td>
<td>Prinzipien der Europäischen Union</td>
<td>προασπίζετε τις αρχές της</td>
<td>komisarza</td>
</tr>
<tr>
<td>principles of the European Union and</td>
<td>zu verteidigen und nicht sich bei</td>
<td>Ευρωπαϊκής Ένωσης και να</td>
<td>odpowiedzialnego za</td>
</tr>
<tr>
<td>not to promote yourself</td>
<td>Bedarf als Minister einer horizontal</td>
<td>μην προωθείτε πάλι τον εαυτό σας ως</td>
<td>prawa człowieka,</td>
</tr>
<tr>
<td>again as a minister in</td>
<td>italienischen Regierung</td>
<td>υπουργό μιας μελλοντικής ιταλικής</td>
<td>bronić zasad Unii Europejskiej,</td>
</tr>
<tr>
<td>a future Italian Government when</td>
<td>wieder empfehlen zu wollen.</td>
<td>ευβέμενης εφορίας χρειάστε.</td>
<td>jeżeli to konieczne,</td>
</tr>
<tr>
<td>required.</td>
<td></td>
<td></td>
<td>a nie</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>promować się na</td>
</tr>
<tr>
<td><strong>30278006</strong> Although this ambitious</td>
<td>Obwohl für dieses ehrgeizige</td>
<td>Ενώ όλο αυτό το</td>
<td>przyszego ministra w</td>
</tr>
<tr>
<td>scheme seems to have</td>
<td>Vorhaben die passende</td>
<td>μεγαλεπιβόλο σχέδιο φαίνεται</td>
<td>rządzie włoskim.</td>
</tr>
<tr>
<td>fitted into a proper</td>
<td>technokratische Form gefunden</td>
<td>να έχει μπει σε ένα σωστό</td>
<td></td>
</tr>
<tr>
<td>technocratic mould, i</td>
<td>worden zu sein scheint, Frage ich</td>
<td>technokratisches Kalòyìì,</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**OPUS multilingual search interface, Europarl**

http://opus.lingfil.uu.se
Classification criteria

- Language
  - monolingual
  - multilingual: parallel/comparable/other

**parallel** translations in different languages, often aligned at sentence or word level

**comparable** similar in different languages
wrt size, context size, genre, (topic)

- e.g. MULTEXT-East cesDoc corpus
  Comparable components: fiction, news
  wrt number and size of texts
Classification criteria

- Genre
  - news corpus: few errors, fixed register  
    TIGER 2.1, DIRNDL
  - web corpus: noise, missing context, user created content  
    deWaC, ENCOW14
  - literary corpus: author corpus, historical journal, ...  
    Goethe-Korpus, AAC-FACKEL
  - conversation corpus  
    SCoSE, GECO
  - language acquisition data / learner corpus  
    data produced by L1 or L2 learners, often includes target hypotheses  
    CHILDES, Falko
  - specialized language: instruction manuals, law texts, medical texts, ...  
    fieldwork data: e.g. under-resourced languages ...  
    experiment data: e.g. eye-tracking, reaction times ...
Classification criteria

• Balance
  with respect to language varieties, genre, topic, size . . .
  ⇒ reference corpora, e.g. BNC
Classification criteria

- Balance with respect to language varieties, genre, topic, size...
  ⇒ reference corpora, e.g. BNC

- Degree of Planning (for spoken data)

Planned

- DIRNDL radio news
- BeMaTaC map task

Spontaneous

- GECO free conversation
Classification criteria

- Size

### Text based

<table>
<thead>
<tr>
<th>English Gigaword</th>
<th>TIGER Corpus 2.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fifth Edition</td>
<td></td>
</tr>
<tr>
<td>4,032,686,000</td>
<td>~900,000</td>
</tr>
<tr>
<td>(word) tokens</td>
<td></td>
</tr>
</tbody>
</table>

### Sound based

<table>
<thead>
<tr>
<th>GECO</th>
<th>DIRNDL</th>
</tr>
</thead>
<tbody>
<tr>
<td>~19</td>
<td>~5</td>
</tr>
<tr>
<td>hours</td>
<td></td>
</tr>
</tbody>
</table>
Classification criteria

- Context size
  - document, sentence, utterance, word, ...
  - some corpora are shuffled for licensing reasons
Classification criteria

- Context size
  
  document, sentence, utterance, word, ...
  → some corpora are shuffled for licensing reasons

- State of annotation
  
  - only primary data
    → allows for studies on frequency, context, ...
  - annotations:
    phonological, prosodic, morpho-syntactic, semantic, ...
Classification criteria

- Context size
  - document, sentence, utterance, word, . . .
  - some corpora are shuffled for licensing reasons

- State of annotation
  - only primary data
    - allows for studies on frequency, context, . . .
  - annotations:
    - phonological, prosodic, morpho-syntactic, semantic, . . .

- Technical representation
  - sampling rate
  - character encoding / scripts
  - specific format (XML, plain text, .wav)
  - inline / stand-off annotation
Motivation for using a corpus

Passives of reflexives (PoR)

• German is one of the few languages where reflexive verbs can passivize
  BUT: sparse phenomenon

(1) Erst wird sich geküsst, dann wird geheiratet.
  first is REFL kissed, then is married
  ’First one kisses, then one marries.’

[Zarrieß et al., 2013]
Motivation for using a corpus

Passives of reflexives (PoR)

- German is one of the few languages where reflexive verbs can passivize
  BUT: sparse phenomenon

  (1) Erst wird sich geküßt, dann wird geheiratet.
  first is REFL kissed, then is married
  ’First one kisses, then one marries.’

- Which verbs can form a PoR?

SdeWaC-v3

- web corpus
- 880 million (word) tokens
- automatic syntactic annotations available
Motivation for using a corpus

Passives of reflexives (PoR) [Zarrieß et al., 2013]

SdeWaC-v3

- web corpus
- 880 million (word) tokens
- automatic syntactic annotations available

- 134 verbs with PoR were identified
- General proportions of reflexive uses were extracted
- Evidence from the corpus study: formation of PoRs seems to be lexically restricted
  ⇒ only available for inherently and naturally reflexive verbs
Motivation for using a corpus

Study on backchannels in spontaneous conversations

[Schweitzer and Lewandowski, 2012]

**GECO database**

- 46 dialogs (~25mins)
- unacquainted female speakers
- ratings for mutual liking and competence

→ more sympathy towards interlocutor / more assumed competence of the interlocutor

but: if someone produces more backchannels, they are not rated as more likeable or competence

more backchannels produced → rated less competent! (statistical tendency)
Motivation for using a corpus

Study on backchannels in spontaneous conversations

[Schweitzer and Lewandowski, 2012]

GECO database

- 46 dialogs (~25mins)
- unacquainted female speakers
- ratings for mutual liking and competence

results

- more sympathy towards interlocutor / more assumed competence of the interlocutor $\rightarrow$ more backchannels
Motivation for using a corpus

Study on backchannels in spontaneous conversations

[Schweitzer and Lewandowski, 2012]

GECO database

- 46 dialogs (~25mins)
- unacquainted female speakers
- ratings for mutual liking and competence

results

- more sympathy towards interlocutor / more assumed competence of the interlocutor $\rightarrow$ more backchannels
- but: if someone produces more backchannels, they are not rated as more likeable or competence
Motivation: Example studies

Study on backchannels in spontaneous conversations

[Schweitzer and Lewandowski, 2012]

GECO database

- 46 dialogs (~25mins)
- unacquainted female speakers
- ratings for mutual liking and competence

• results
  - more sympathy towards interlocutor / more assumed competence of the interlocutor → more backchannels
  - but: if someone produces more backchannels, they are not rated as more likeable or competence
  - more backchannels produced → rated less competent! (statistical tendency)
Motivation for using a corpus

Other catchy corpus studies

- Does the Queen speak the Queen’s English?
  [Harrington et al., 2000]
  (spoiler: Nope)
- Predicting U.S. Presidential Election Outcomes
  [Gregory and Gallagher, 2002]
  (spoiler: Losers’ speech converges to winners’)
- Natural dialog behaviour in user companion interaction?
  [Rösner et al., 2012]
  (spoiler: Somewhere between human and machine)
Corpora at the IMS

- Online information about corpora created or adapted at the IMS: http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/index.en.html
- Corpora available at the IMS (see licensing notes)
  - List of corpora on CD-ROM: http://wiki.ims.uni-stuttgart.de/CorporaOnCDRom
  - File system: /resources/
Corpora at the IMS

- Online information about corpora created or adapted at the IMS:
  http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/index.en.html
Corpora at the IMS

- Online information about corpora created or adapted at the IMS:
  http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/index.en.html

- Corpora available at the IMS (see licensing notes)
  - List of corpora on CD-ROM: http://wiki.ims.uni-stuttgart.de/CorporaOnCDRom
Corpora at the IMS

- Online information about corpora created or adapted at the IMS:

- Corpora available at the IMS (see licensing notes)
  - List of corpora on CD-ROM: [http://wiki.ims.uni-stuttgart.de/CorporaOnCDRom](http://wiki.ims.uni-stuttgart.de/CorporaOnCDRom)
  - File system: `/resources/`
Corpora at the IMS

- Online information about corpora created or adapted at the IMS:
  http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/index.en.html

- Corpora available at the IMS (see licensing notes)
  - List of corpora on CD-ROM: http://wiki.ims.uni-stuttgart.de/CorporaOnCDRom
  - File system: /resources/

In case that you want to use the corpora on the IMS servers but do not have access yet, please drop by the administration office: 00.008.
To get an account for the IMS Wiki, see http://wiki.ims.uni-stuttgart.de/HowToGetAnAccount
External Catalogues/Initiatives

- LDC – Linguistic Data Consortium
  https://www.ldc.upenn.edu/
- ELRA – European Language Resources Association
  http://catalog.elra.info/
- CLARIN Virtual Language Observatory:
  Overview on sets of metadata
  https://catalog.clarin.eu
- Web corpora
  - WaCky, up to now: de, en, fr, it
    http://wacky.sslmit.unibo.it
  - COW – COrpora from the Web, up to now: de, en, es, fr, nl, sv
    http://corporafromtheweb.org/
- CHILDES – Child Language Data Exchange System
  http://childes.psy.cmu.edu/
External Catalogues/Initiatives

- LDC – Linguistic Data Consortium
  [https://www.ldc.upenn.edu/](https://www.ldc.upenn.edu/)

- ELRA – European Language Resources Association
External Catalogues/Initiatives

- LDC – Linguistic Data Consortium
  https://www.ldc.upenn.edu/

- ELRA – European Language Resources Association
  http://catalog.elra.info/

- CLARIN Virtual Language Observatory: Overview on sets of metadata
  https://catalog.clarin.eu
External Catalogues/Initiatives

- LDC – Linguistic Data Consortium
  https://www.ldc.upenn.edu/

- ELRA – European Language Resources Association
  http://catalog.elra.info/

- CLARIN Virtual Language Observatory:
  Overview on sets of metadata
  https://catalog.clarin.eu

- Web corpora
  - WaCky, up to now: de, en, fr, it
    http://wacky.sslmit.unibo.it
  - COW – COrrpora from the Web, up to now: de, en, es, fr, nl, sv
    http://corporafromtheweb.org/
External Catalogues/Initiatives

- LDC – Linguistic Data Consortium
  https://www.ldc.upenn.edu/
- ELRA – European Language Resources Association
  http://catalog.elra.info/
- CLARIN Virtual Language Observatory:
  Overview on sets of metadata
  https://catalog.clarin.eu
- Web corpora
  - WaCky, up to now: de, en, fr, it
    http://wacky.sslmit.unibo.it
  - COW – COrpora from the Web, up to now: de, en, es, fr, nl, sv
    http://corporafromtheweb.org/
- CHILDES – Child Language Data Exchange System
  http://childes.psy.cmu.edu/
How to use

- Access and query

- Additional preparation steps
How to use

• Access and query
  ⇒ depend on
  • available annotation layers and their granularity:
    “Only what has been annotated can be searched for.”
  • tagsets (i.e. vocabulary, granularity)
  • representation format
  • query language

• Additional preparation steps
How to use

• Access and query
  ⇒ depend on
    • available annotation layers and their granularity:
      “Only what has been annotated can be searched for.”
    • tagsets (i.e. vocabulary, granularity)
    • representation format
    • query language
  ⇒ can be done by
    • command line tools (UNIX, CQP, Praat scripts (course to come), Festival...)
    • database management systems (XML, SQL...)
    • visualisation tools (vpf, TigerSearch, ANNIS, ICARUS...)
    • web interfaces (CQPweb, OPUS, COSMAS II, Colibri²...)

• Additional preparation steps
How to use

- Access and query
  - depend on
    - available annotation layers and their granularity:
      "Only what has been annotated can be searched for."
    - tagsets (i.e. vocabulary, granularity)
    - representation format
    - query language
  - can be done by
    - command line tools (UNIX, CQP, Praat scripts (course to come), Festival...)
    - database management systems (XML, SQL...)
    - visualisation tools (vpf, TigerSearch, ANNIS, ICARUS...)
    - web interfaces (CQPweb, OPUS, COSMAS II, Colibri²...)

- Additional preparation steps
  - annotation (manual/automatic)
  - extraction of (phenomenon-specific) sub corpus
Showcases

- CQPweb
- ICARUS

Help and training provided by project INF

– please contact us!
Tips

• When utilizing a corpus
  • cite the resource (e.g. via PID) along with its exact version
    ⇒ sustainability of results
    ⇒ reproducibility
  • have a look at the licensing information (if available)
Tips

• When utilizing a corpus
  • cite the resource (e.g. via PID) along with its exact version
    ⇒ sustainability of results
    ⇒ reproducibility
  • have a look at the licensing information (if available)

• When preparing a corpus
  • do not change the primary data (use additional layers, stand-off annotation)
  • document your workflow
  • make it available and persistent (if possible)
    ⇒ be as restrictive as needed for your task, but keep in mind that this could be a helpful resource for others as well
Corpora in Phase 3 of the SFB

Silver-standard collection

- non-canonical
- dynamic
- suitable for both large-scale and instance-based exploration
Corpora in Phase 3 of the SFB

Silver-standard collection

- non-canonical
- dynamic
- suitable for both large-scale and instance-based exploration

Up to now:
- radio interviews (DE)

To come:
- radio conversations (FR)
- web data
Further information on the mentioned corpora I

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Name</th>
<th>URL</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAC-FAKEL</td>
<td>Austrian Academy Corpus DIE FACKEL</td>
<td><a href="http://www.aac.ac.at/apps_digied_fackel.html">http://www.aac.ac.at/apps_digied_fackel.html</a></td>
<td></td>
</tr>
<tr>
<td>BeMaTaC</td>
<td>Berlin Map Task Corpus</td>
<td><a href="http://u.hu-berlin.de/bematac">http://u.hu-berlin.de/bematac</a></td>
<td>[Giesel et al., 2013, Sauer and Rasskazova, 2014]</td>
</tr>
<tr>
<td>CHILDES</td>
<td>Child Language Data Exchange System</td>
<td><a href="http://chiltes.psy.cmu.edu/">http://chiltes.psy.cmu.edu/</a></td>
<td></td>
</tr>
<tr>
<td>Corpus of Contemporary American English (COCA)</td>
<td><a href="http://corpus.byu.edu/coca/">http://corpus.byu.edu/coca/</a></td>
<td>[Davies, 2010]</td>
<td></td>
</tr>
<tr>
<td>DDB</td>
<td>Deutsche Diachrone Baumbank</td>
<td><a href="http://korpling.german.hu-berlin.de/ddb-doku/index.htm">http://korpling.german.hu-berlin.de/ddb-doku/index.htm</a></td>
<td>metadata: <a href="http://hdl.handle.net/11022/0000-0000-2107-3">http://hdl.handle.net/11022/0000-0000-2107-3</a></td>
</tr>
<tr>
<td>deWaC</td>
<td></td>
<td><a href="http://wacky.sslmit.unibo.it/doku.php?id=corpora#german">http://wacky.sslmit.unibo.it/doku.php?id=corpora#german</a></td>
<td>[Baroni et al., 2009]</td>
</tr>
</tbody>
</table>
Further information on the mentioned corpora II

**DIRNDL**  http://hdl.handle.net/11858/00-247C-0000-0022-F7B2-8  
*metadata*: http://hdl.handle.net/11858/00-247C-0000-0022-F7B1-A  
[Eckart et al., 2012]

**ENCOW14**  http://corporafromtheweb.org/encow14/  
[Schäfer and Bildhauer, 2012]

**Europarl**  (Release v7) http://www.statmt.org/europarl/  
[Koehn, 2005]

**Falko**  https://www.linguistik.hu-berlin.de/institut/professuren/korporuslinguistik/forschung/falko  
[Reznicek et al., 2012]

**Fuerstinnenkorrespondenz 1.1**  Lühr, Rosemarie; Faßhauer, Vera; Prutscher, Daniela; Seidel, Henry;  
Fuerstinnenkorrespondenz (Version 1.1), Universität Jena, DFG.  
http://www.indogerianistik.uni-jena.de/Web/Projekte/Fuerstinnenkorr.htm  
*metadata*: http://hdl.handle.net/11022/0000-0000-2E44-1

**GECO**  http://hdl.handle.net/11858/00-247C-0000-0023-5137-2  
*metadata*: http://hdl.handle.net/11858/00-247C-0000-0023-512E-7

**Goethe-Korpus (goe)**  http://www1.ids-mannheim.de/kl/projekte/korpora/archiv/goe.html

**MULTEXT-East cesDoc corpus**  (Version 4)  http://nl.ijs.si/ME/

**SCoSE**  Saarbrücken Corpus of Spoken English  
http://www.uni-saarland.de/lehrstuhl/engling/scose.html
Further information on the mentioned corpora III

**SdeWaC**
http://hdl.handle.net/11858/00-247C-0000-0022-F7BA-7
metadata: http://hdl.handle.net/11858/00-247C-0000-0022-F7B9-9
[Faaß and Eckart, 2013]

**TIGER Corpus** (Version 2.1)
http://hdl.handle.net/11858/00-247C-0000-000D-FFB5-1
metadata: http://hdl.handle.net/11858/00-247C-0000-000D-FFB7-E
[Brants et al., 2004]
References I

The WaCky wide web: a collection of very large linguistically processed web-crawled corpora.

Brants, S., Dipper, S., Eisenberg, P., Hansen-Schirra, S., König, E., Lezius, W., Rohrer, C., Smith, G.,
TIGER: Linguistic interpretation of a German corpus.

Davies, M. (2010).
The corpus of contemporary american english as the first reliable monitor corpus of english.

A discourse information radio news database for linguistic analysis.

Sdewac – a corpus of parsable sentences from the web.


*Korpuslinguistik – Eine Einführung.*
narr Studienbücher. Narr Francke Attempto Verlag, Tübingen, Germany.

*Das Falko-Handbuch: Korpusaufbau und Annotationen: Version 2.01.*
Berlin.

Linguistic analyses of the LAST MINUTE corpus.
Main track: oral presentations.

BeMaTaC – eine digitale multimodale Ressource für Sprach- und Dialogforschung.

Accommodation of backchannels in spontaneous speech (abstract).
Building large corpora from the web using a new efficient tool chain.
In Proceedings of LREC 2012, Istanbul. ELRA.

Passives of reflexives: a corpus study.
Abstract at LinguisticEvidence – Berlin Special.