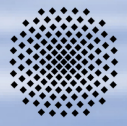


University of Stuttgart
Germany

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

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
  $\Rightarrow$ cf. e.g. [Lemnitzer and Zinsmeister, 2006]
- include hands-on instructions for one specific corpus, database or query tool
  $\Rightarrow$ 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
  ⇒ 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

Overview

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.

Translated from: Lothar Lemnitzer, Heike Zinsmeister. 2006. Korpuslinguistik, Eine Einführung. Definition 1, page 7.

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.

Translated from: Lothar Lemnitzer, Heike Zinsmeister. 2006. Korpuslinguistik, Eine Einführung. Definition 1, page 7.

In this course we take a broad pragmatic view:

- A corpus is a resource, based on primary data (textual, audio, video, multi-modal, ...) $\Rightarrow$ 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.

Translated from: Lothar Lemnitzer, Heike Zinsmeister. 2006. Korpuslinguistik, Eine Einführung. Definition 1, page 7.

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.

Translated from: Lothar Lemnitzer, Heike Zinsmeister. 2006. Korpuslinguistik, Eine Einführung. Definition 1, page 7.

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.

Translated from: Lothar Lemnitzer, Heike Zinsmeister. 2006. Korpuslinguistik, Eine Einführung. Definition 1, page 7.

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

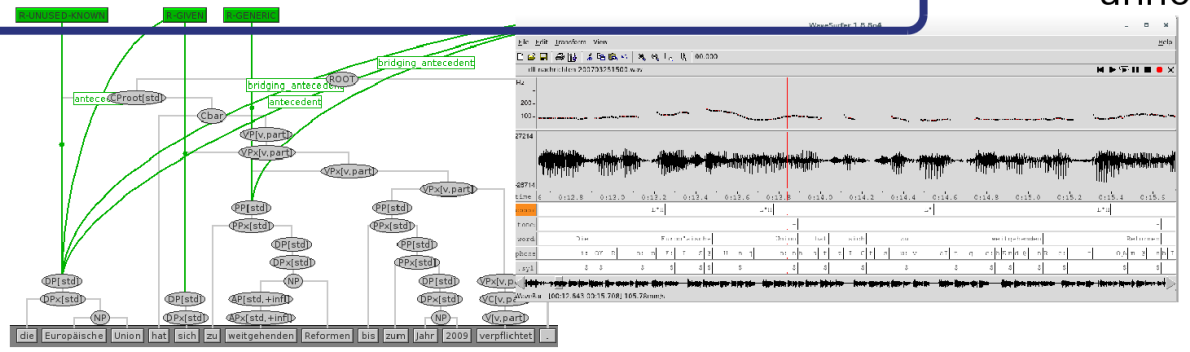
Example: The DIRNDL corpus

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
Spaker Demographics	professional speakers 5 male, 4 female

primary data

annotations



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
  - . . .

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

historical	vs.	**contemporary**
Fuerstinnen-korrespondenz 1.1	**example corpora**	Corpus of Comparisons in Product Reviews
Analysis of social networks	**example studies**	Sentiment analysis

Classification criteria

- Temporal origin of primary data

historical	vs.	**contemporary**
Fuerstinnen-korrespondenz 1.1	**example corpora**	Corpus of Comparisons in Product Reviews
Analysis of social networks	**example studies**	Sentiment analysis

Fuerstinnenkorrespondenz 1.1

- letters of 15 princesses and their male correspondence partners
- from 1546 to 1756

Classification criteria

- Temporal origin of primary data

historical	vs.	**contemporary**
Fuerstinnen-korrespondenz 1.1	**example corpora**	Corpus of Comparisons in Product Reviews
Analysis of social networks	**example studies**	Sentiment analysis

Corpus of Comparisons in Product Reviews

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

Classification criteria

- Temporal origin of primary data

diachronic	vs.	**synchronic**
several (historical) stages of a language		one stage
Deutsche Diachrone Baumbank (DDB)	**example corpus**	British National Corpus (BNC)
language change	**example studies**	parser training

Classification criteria

- Temporal origin of primary data

diachronic	vs.	**synchronic**
several (historical) stages of a language		one stage
Deutsche Diachrone Baumbank (DDB)	**example corpus**	British National Corpus (BNC)
language change	**example studies**	parser training

DDB

- three sub corpora
- Old High German, Middle High German, Early New High German

Classification criteria

- Temporal origin of primary data

diachronic	vs.	**synchronic**
several (historical) stages of a language		one stage
Deutsche Diachrone Baumbank (DDB)	**example corpus**	British National Corpus (BNC)
language change	**example studies**	parser training

BNC

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

Classification criteria

- Temporal origin of the primary data

static	vs.	**monitor corpora**
TIGER Corpus	**example corpora**	Corpus of Contemporary American English (COCA)
grammar extraction/ induction	**example studies**	identification of new words

Classification criteria

- Temporal origin of the primary data

static	vs.	monitor corpora
TIGER Corpus	**example corpora**	Corpus of Contemporary American English (COCA)
grammar extraction/ induction	**example studies**	identification of new words

TIGER Corpus

- newspaper texts
- 1990s

Classification criteria

- Temporal origin of the primary data

static	vs.	monitor corpora
TIGER Corpus	**example corpora**	Corpus of Contemporary American English (COCA)
grammar extraction/ induction	**example studies**	identification of new words

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

	en	de	el	pl
29733339	It is your task , Commissioner Frattini - especially as the Commissioner responsible for human rights - to defend the principles of the European Union and not to promote yourself again as a **minister** in a future Italian Government when required .	Ihre Aufgabe , Herr Kommissar Frattini , ist es - gerade als Grundrechtekommissar - , die Prinzipien der Europäischen Union zu verteidigen und nicht sich bei Bedarf als Minister einer zukünftigen italienischen Regierung wieder empfehlen zu wollen .	Έχετε υποχρέωση , Επίτροπε Frattini -δεδομένου ιδίως ότι είστε αρμόδιος Επίτροπος για τα ανθρώπινα δικαιώματα- να προασπίζετε τις αρχές της Ευρωπαϊκής Ένωσης και να μην προωθείτε πάλι τον εαυτό σας ως υποψήφιο υπουργό μιας μελλοντικής ιταλικής κυβέρνησης εφόσον χρειαστεί .	Pańskim zadaniem jest , panie komisarzu Frattini , szczególnie jako komisarza odpowiedzialnego za prawa człowieka , bronić zasad Unii Europejskiej , jeżeli to konieczne , a nie promować się na przyszłego ministra w rządzie włoskim .
30278006	Although this ambitious scheme seems to have fitted into a proper technocratic mould , I	Obwohl für dieses ehrgeizige Vorhaben die passende technokratische Form gefunden worden zu sein scheint , frage ich	Ενώ όλο αυτό το μεγαλεπήβολο σχέδιο φαίνεται να έχει μπει σε ένα σωστό τεχνοκρατικό καλούπι ,	Choć ten ambitny plan wpasował się w odpowiedni wzorzec technokratyczny ,

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 . . .

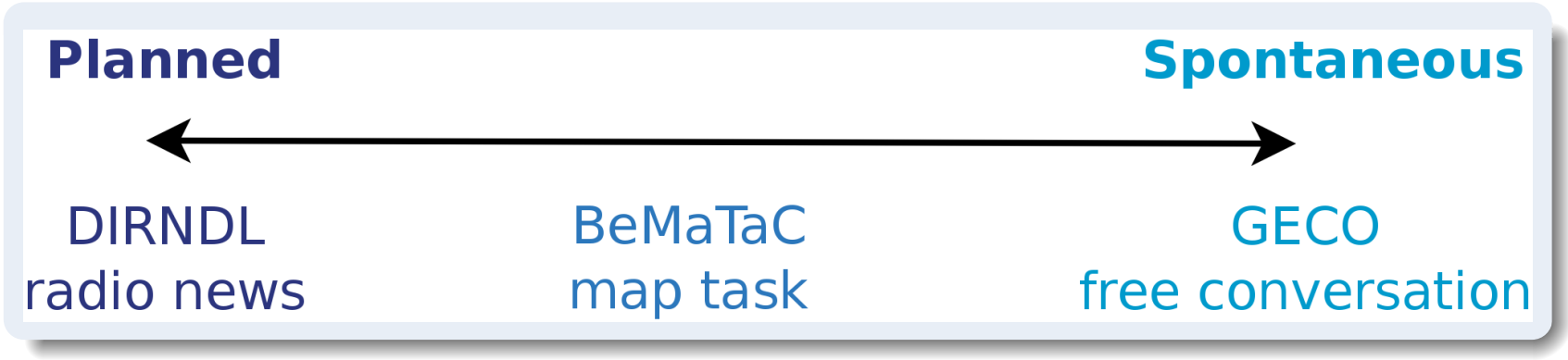
Classification criteria

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

Classification criteria

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

- Degree of Planning (for spoken data)



Classification criteria

- Size

Text based

English Gigaword Fifth Edition		**TIGER Corpus 2.1**
4,032,686,000	(word) tokens	~900,000

Sound based

GECO		**DIRNDL**
~19	hours	~5

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) [Zarrieß et al., 2013]

- 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.'

Motivation for using a corpus

Passives of reflexives (PoR)

[Zarrieß et al., 2013]

- 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
  $\Rightarrow$ 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

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

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 → 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

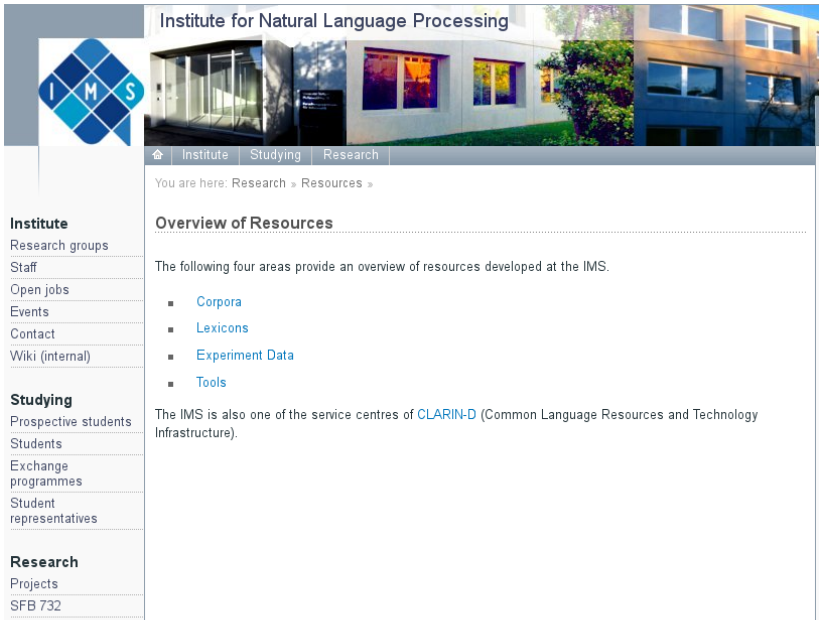
Other catchy corpus studies 

- 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

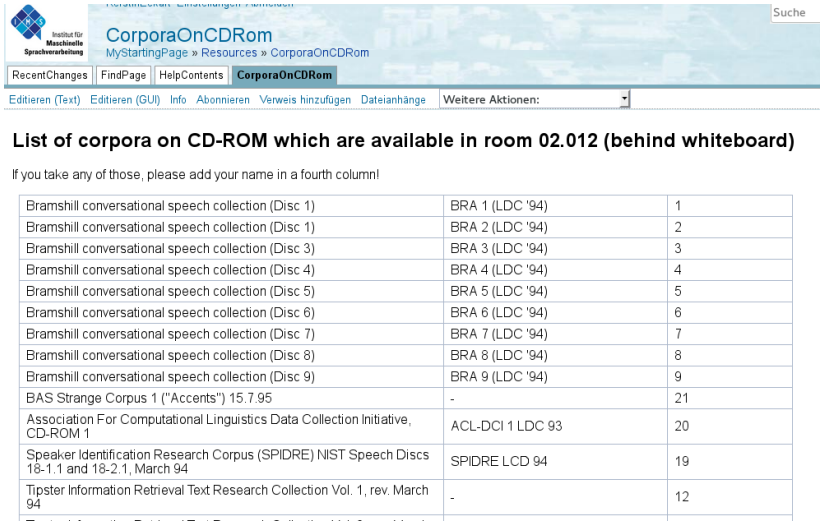
Corpora at the IMS

- 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`



CorporaOnCDRom
MyStartingPage » Resources » CorporaOnCDRom

RecentChanges | FindPage | HelpContents | CorporaOnCDRom

Editieren (Text) Editieren (GUI) Info Abonnieren Verweis hinzufügen Dateianhänge Weitere Aktionen:

List of corpora on CD-ROM which are available in room 02.012 (behind whiteboard)

If you take any of those, please add your name in a fourth column!

Bramshill conversational speech collection (Disc 1)	BRA 1 (LDC '94)	1
Bramshill conversational speech collection (Disc 1)	BRA 2 (LDC '94)	2
Bramshill conversational speech collection (Disc 3)	BRA 3 (LDC '94)	3
Bramshill conversational speech collection (Disc 4)	BRA 4 (LDC '94)	4
Bramshill conversational speech collection (Disc 5)	BRA 5 (LDC '94)	5
Bramshill conversational speech collection (Disc 6)	BRA 6 (LDC '94)	6
Bramshill conversational speech collection (Disc 7)	BRA 7 (LDC '94)	7
Bramshill conversational speech collection (Disc 8)	BRA 8 (LDC '94)	8
Bramshill conversational speech collection (Disc 9)	BRA 9 (LDC '94)	9
BAS Strange Corpus 1 ("Accents") 15.7.95	-	21
Association For Computational Linguistics Data Collection Initiative, CD-ROM 1	ACL-DCI 1 LDC 93	20
Speaker Identification Research Corpus (SPIDRE) NIST Speech Discs 18-1.1 and 18-2.1, March 94	SPIDRE LCD 94	19
Tipster Information Retrieval Text Research Collection Vol. 1, rev. March 94	-	12

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 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

External Catalogues/Initiatives

- LDC – Linguistic Data Consortium
  `https://www.ldc.upenn.edu/`
- ELRA – European Language Resources Association
  `http://catalog.elra.info/`

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 – COrpora 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, Colibri^2. . .)
- 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, Colibri^2. . .)
- 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
    $\Rightarrow$ sustainability of results
    $\Rightarrow$ 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
    $\Rightarrow$ sustainability of results
    $\Rightarrow$ 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)

  $\Rightarrow$ 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

AAC-FACKEL Austrian Academy Corpus DIE FACKEL
http://www.aac.ac.at/apps_digied_fackel.html

BeMaTaC Berlin Map Task Corpus http://u.hu-berlin.de/bematac
[Giesel et al., 2013, Sauer and Rasskazova, 2014]

BNC The British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium.
http://www.natcorp.ox.ac.uk/

CHILDES Child Language Data Exchange System http://childes.psy.cmu.edu/

Corpus of Comparisons in Product Reviews http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/reviewcomparisons/index.html
[Kessler and Kuhn, 2014]

Corpus of Contemporary American English (COCA) http://corpus.byu.edu/coca/
[Davies, 2010]

DCEP Digital Corpus of the European Parliament
https://ec.europa.eu/jrc/en/language-technologies/dcep
[Hajlaoui et al., 2014]

DDB Deutsche Diachrone Baumbank
Hirschmann, Hagen; Linde, Sonja; DDB (Version 1.0), Humboldt-Universität zu Berlin.
http://korpling.german.hu-berlin.de/ddb-doku/index.htm
metadata: http://hdl.handle.net/11022/0000-0000-2107-3

deWaC http://wacky.sslmit.unibo.it/doku.php?id=corpora#german
[Baroni et al., 2009]

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/korpuslinguistik/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.indogermanistik.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

References

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

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009).
The WaCky wide web: a collection of very large linguistically processed web-crawled corpora.
Language Resources and Evaluation, 43(3):209–226.

Brants, S., Dipper, S., Eisenberg, P., Hansen-Schirra, S., König, E., Lezius, W., Rohrer, C., Smith, G., and Uszkoreit, H. (2004).
TIGER: Linguistic interpretation of a German corpus.
Research on Language and Computation, 2(4):597–620.

Davies, M. (2010).
The corpus of contemporary american english as the first reliable monitor corpus of english.
Literary and Linguistic Computing, 25(4):447–464.

Eckart, K., Riester, A., and Schweitzer, K. (2012).
A discourse information radio news database for linguistic analysis.
In Chiarcos, C., Nordhoff, S., and Hellmann, S., editors, *Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata*, pages 65–75. Springer, Heidelberg.

Faaß, G. and Eckart, K. (2013).
Sdewac – a corpus of parsable sentences from the web.
In Gurevych, I., Biemann, C., and Zesch, T., editors, *Language Processing and Knowledge in the Web*, volume 8105 of *Lecture Notes in Computer Science*, pages 61–68. Springer Berlin Heidelberg.

References II

Giesel, L., Klapi, M., Krüger, D., Nunberger, I., Rasskazova, O., and Sauer, S. (2013).

Berlin Map Task Corpus – A deeply annotated multimodal map-task corpus of spoken learner and native German.

Poster presentation at DGfS-CL 2013.

Gregory, S. W. and Gallagher, T. J. (2002).

Spectral analysis of candidates' nonverbal communication: Predicting U.S. presidential election outcomes.

Social Psychology Quarterly, 65(3):298–308.

Hajlaoui, N., Kolovratnik, D., Väyrynen, J., Steinberger, R., and Varga, D. (2014).

Dcep -digital corpus of the european parliament.

In Chair), N. C. C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Harrington, J., Palethorpe, S., and Watson, C. I. (2000).

Does the queen speak the queen's english?

Nature, 408:927–928.

Kessler, W. and Kuhn, J. (2014).

A corpus of comparisons in product reviews.

In *In Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)*, Reykjavik, Iceland.

References III

Koehn, P. (2005).
Europarl: A parallel corpus for statistical machine translation.
In *The Tenth Machine Translation Summit, Proceedings of Conference*, pages 79–86, Phuket, Thailand.

Lemnitzer, L. and Zinsmeister, H. (2006).
Korpuslinguistik – Eine Einführung.
narr Studienbücher. Narr Francke Attempto Verlag, Tübingen, Germany.

Reznicek, M., Lüdeling, A., and Schwantuschke, F. (2012).
Das Falko-Handbuch: Korpusaufbau und Annotationen: Version 2.01.
Berlin.

Rösner, D., Kunze, M., Otto, M., and Frommer, J. (2012).
Linguistic analyses of the LAST MINUTE corpus.
In Jancsary, J., editor, *Proceedings of KONVENS 2012*, pages 145–154. ÖGAI.
Main track: oral presentations.

Sauer, S. and Rasskazova, O. (2014).
BeMaTaC – eine digitale multimodale Ressource für Sprach- und Dialogforschung.
Poster presentation at the workshop Grenzen überschreiten – Digitale Geisteswissenschaft heute und morgen, Digital Humanities Berlin 2014.

Schweitzer, A. and Lewandowski, N. (2012).
Accommodation of backchannels in spontaneous speech (abstract).
Booklet of the International Symposium on Imitation and Convergence in Speech.

References IV

Schäfer, R. and Bildhauer, F. (2012).
Building large corpora from the web using a new efficient tool chain.
In *Proceedings of LREC 2012*, Istanbul. ELRA.

Zarrieß, S., Schäfer, F., and Schulte im Walde, S. (2013).
Passives of reflexives: a corpus study.
Abstract at LinguisticEvidence – Berlin Special.