Spoken Corpora

Malte Belz

CoLAB Workshop on Corpus Linguistics
Humboldt-Universität zu Berlin
June 4–7, 2018

@inproceedings{Belz.2018,
    author = {Belz, Malte},
    crossref = {colab.2018},
    title = {Spoken Corpora},
    booktitle = {CoLAB Workshop on Corpus Linguistics},
    year = {2018}
}

@proceedings{colab.2018,
    year = {2018},
    title = {CoLAB Workshop on Corpus Linguistics},
    location = {Humboldt-Universität zu Berlin}
}
Overview

1. Spoken corpora
   - Definition and contexts
   - Designs
   - Accessability and findability

2. Corpus preparation
   - Recording and transcription
   - Segmentation and tokenization
   - Multi-layer architecture
   - Annotation and agreement

3. Corpus analysis
   - Method
   - Two microstudies

4. Corpus documentation
1. Spoken corpora

1.1 Definition and contexts

- A corpus\(^1\)
  - is a collection of machine-readable ‘text’
  - allows users to search through them rapidly and reliably
  - allows users to look at words in context
    - frequencies
    - concordances
  - Corpus data must match the research question
  - Corpora can vary internally

---

\(^1\)Cf. McEnery and Hardie (2012)
1. Spoken corpora

1.1 Definition and contexts

- **Types of studies (settings, design)**
  - Mode of communication (written, spoken, signs)
  - Corpus-based (verify/falsify informed research question) vs. corpus-driven (empirical detection of phenomena)
  - Data collection regime (balanced?)
  - Annotated vs. unannotated
  - Total accountability vs. data selection (do not select your favourite subset of a corpus)
  - Multilingual vs. monolingual
1. Spoken corpora
1.1 Definition and contexts

- **Research contexts**
  - **Language acquisition**
    - **Child language** (cf. McCarthy and O’Keeffe 2009: 1008)
    - **Learner corpora and L2 research** (for an overview on annotation of learner corpora cf. Ballier and Martin 2015; for fluency and corpora cf. Götz 2013)
  - **Sociolinguistic research**
    - **Level of education, age, sex, power, dialect** (compare the designs of Bortfeld et al. 2001; Laserna et al. 2014; Tottie 2011; Wieling et al. 2016)
  - **Discourse/interactional research**
    - **Task-based vs. task-free** (for example, GECO\(^2\) vs. BeMaTaC\(^3\))

---

\(^2\) Lemnitzer and Zinsmeister (2010),
http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/IMS-GEC0.en.html

\(^3\) Sauer (2014) and Sauer and Lüdeling (2013, 2016), hu-berlin.de/bematac
1. Spoken corpora
1.1 Definition and contexts

- **Linguistic dimensions**
  - Phonetics, Phonology, Intonation, Pausology
    - Ex.: Frequency and position of glottal pauses in L1 German (Belz 2017)
    - Ex.: Convergence in spontaneous speech (Schweitzer and Lewandowski 2013)
  - Morphology, Syntax, Morpho-Syntax, Phono-Morphology
    - Ex.: Level of verb-final schwa elision in L1/L2 German (Lüdeling et al. 2015)
    - Ex.: Priming of syntactic structure in spontaneous dialogue in English (Healey et al. 2013)
  - Semantics, Pragmatics, Discourse, Interactional Linguistics
    - Ex.: Frequency and structure of self-repair in L1/L2 German (Belz et al. 2017)
    - Ex.: Persistence in turn-taking structure (Szmrecsanyi 2005)
1. Spoken corpora
1.1 Definition and contexts

Problems (McCarthy and O’Keeffe 2009)

- spoken corpora are much more expensive (time-consuming) to create than written corpora
- spoken corpora for phonetic studies depend on a reasonable or high recording quality
- spoken corpora need to be transcribed first
- there can be many different normalizations
- Participants must allow you to save and publish their data (anonymization of real names is mandatory)
1. Spoken corpora

1.2 Designs

- Corpus design\(^4\) – depends on research question!
- Does the corpus suit the relevant criteria needed for answering the research question?
  - Modality
    - Read vs. spontaneous speech, monologue vs. dialogue
  - Sampling criteria
  - Authenticity of ‘texts’
  - Balancing
    - L1 and L2
    - Selection of participants (male and female)
  - Setting

---

\(^4\)Cf. McCarthy and O’Keeffe (2009)
1. Spoken corpora
1.2 Designs: What kind of spoken language?

GECO – Gesprächscorpus\(^5\)
- German native speakers; dialogues; task-free spontaneous speech
- 13 female speakers; 41.6h of recorded material, studio
- Free dialogues about topics such as university, classes, studying, moving out, celebrating, vacation

BeMaTaC – Berlin Map Task Corpus\(^6\)
- German learners, German native speakers; dialogues; task-based spontaneous speech (Map Task)
- 16 L1 speaker, 12 dialogues; 6 L2 speaker, 5 dialogues; 2.4h of recorded material, studio
- Task-based dialogues: conceptualizing, directions, feedback questions, checking of common ground

\(^5\) (Schweitzer and Lewandowski 2013), http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/IMS-GECO.en.html

\(^6\) (Sauer 2014; Sauer and Lüdeling 2013, 2016), http://hu-berlin.de/bematac, searchable via ANNIS (Krause and Zeldes 2016): https://korpling.german.hu-berlin.de/annis3/?id=db2dfc4e-b7ce-4d26-bcf9-0ebdbfb8d07a
1. Spoken corpora
1.2 Designs: BeMaTaC maps
1. Spoken corpora
1.3 Accessability and findability

- Meta search engines
  - Virtual Language Observatory (VLO)\(^7\)
  - Linguistic Data Consortium (LDC)\(^8\)
  - Open Language Archives Community (OLAC)\(^9\)

- Repositories
  - HZSK Repository (Hamburg Center for Language Corpora)\(^10\)
  - Find repositories: re3data Consortium\(^11\)

- Handbook articles
  - For example

\(^7\) https://vlo.clarin.eu  
\(^8\) https://catalog.ldc.upenn.edu  
\(^9\) http://www.language-archives.org  
\(^10\) https://corpora.uni-hamburg.de/hzsk/en  
\(^11\) https://www.re3data.org
Corpus access and reusability  (Simons and Bird 2008)

- Availability: Is the sound available or only the transliteration?
- Is the annotation available?
- Is the documentation and metadata available?
- Is the data freely accessible and expendable?

Sustainability – document everything in the corpus and every preparation step
2. Corpus preparation

2.1 Recording and transcription

- Set up your design
- Gather relevant/important metadata about participants
- Agreement from participants to publish their data anonymously

Record data

- wav-format
- 48 000 or 44 100 Hz sampling frequency
- ideally in a studio, otherwise in a quiet room
- beware of background noises (trains, buzzing light bulbs, etc.)
- for dialogues: use separate microphones (2 channels)
- rod mics or head-mounted bullet mics or portable recorders

Software

- Praat\textsuperscript{12}: possible, but prone to crash for long recordings
- Audacity\textsuperscript{13}: can handle long recordings, different channels

\textsuperscript{12}Boersma (2001), \url{http://praat.org}
\textsuperscript{13}Audacity Team (2017), \url{https://www.audacityteam.org/}
2. Corpus preparation

2.1 Recording and transcription: Example

Clip of a menu task recorded in our studio in Berlin (not yet published as a corpus)

Two channels in Praat

![Graph showing two channels in Praat](image)
2. Corpus preparation

2.1 Recording and transcription: Example

Transcription

- Transcription vs. transliteration
- Speech is a continuous signal in time
- Segmentation is difficult
- Usually orthographic word layer with varying degrees of normalization (e.g., de-cliticization)
- This can be understood as the first annotation of the data and needs to be documented
2. Corpus preparation

2.1 Recording and transcription

- **Tools**
  - Plain text in UTF-8 with a free editor\textsuperscript{14}
  - OCTRA\textsuperscript{15}
  - FOLKER and OrthoNormal\textsuperscript{16}

\textsuperscript{14}For example Notepad++, https://notepad-plus-plus.org/
\textsuperscript{15}https://www.phonetik.uni-muenchen.de/apps/octra/octra
2. Corpus preparation

2.1 Recording and transcription

Example of a word segmentation with Notepad++

‘Yes of course, the boletes go into the gravy’
2. Corpus preparation
2.2 Segmentation and tokenization

- A token is the smallest technical unit in a corpus.
- There are no tokens in a speech signal, only a continuous time signal.
- The continuous signal is tokenized into discrete entities.
- Tokens created through segmentation.
- Corpora can have multiple tokenizations and segmentations on different layers (cf. Odebrecht et al. 2017).
2. Corpus preparation
2.2 Segmentation and tokenization

“A segmentation is an interpretation of the primary data [...] segmentation directly influences the annotation” (Odebrecht et al. 2017: 699)

<table>
<thead>
<tr>
<th>tok_a</th>
<th>4 units</th>
<th>we</th>
<th>can't</th>
<th>do</th>
<th>that</th>
</tr>
</thead>
<tbody>
<tr>
<td>tok_b</td>
<td>5 units</td>
<td>we</td>
<td>can</td>
<td>t</td>
<td>do</td>
</tr>
<tr>
<td>tok_c</td>
<td>5 units</td>
<td>we</td>
<td>can</td>
<td>'t</td>
<td>do</td>
</tr>
<tr>
<td>tok_d</td>
<td>6 units</td>
<td>we</td>
<td>can</td>
<td>' t</td>
<td>do</td>
</tr>
<tr>
<td>tok_e</td>
<td>5 units</td>
<td>we</td>
<td>can</td>
<td>n't</td>
<td>do</td>
</tr>
</tbody>
</table>

*Fig. 1* Different tokenizations for *we can’t do that* (ibid.: 700)
2. Corpus preparation

2.3 Multi-layer architecture

Orthographic transcription (ORT), and automatically added canonical tier (KAN) and segmental tier (MAU)

‘Yes of course, the boletes go into the gravy’

Kisler et al. (2017), WebMAUS http://hdl.handle.net/11858/00-1779-0000-000C-DA82-F
Annotation = data encoding (McEnery and Hardie 2012)
Annotation is always an interpretation of the data (Lüdeling 2011)
Annotation on multiple layers needs a multi-layer architecture
Store annotations separately (stand-off annotation)
  • Do not use inline annotations (boletes_noelision)
  • Use stand-off annotation

Table: Example for stand-off annotation

<table>
<thead>
<tr>
<th>Word layer</th>
<th>boletes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elision</td>
<td>no</td>
</tr>
</tbody>
</table>
2. Corpus preparation
2.4 Annotation and agreement

- One phenomenon – one annotation layer (do not mix categories)
  - Basics
    - Word tokenization
    - Lemmatization
    - Part-of-speech
    - Phonetic segmentation
  - ⇒ There is no single right tagset! Tagsets can be adapted or created
  - Content
    - Research question 1 (e.g., elision phenomena)
    - Research question 2 (e.g., disfluencies)
    - Research question 3 (e.g., morpho-syntactic phenomena)
2. Corpus preparation

2.4 Annotation and agreement

- Annotation is research
- Document the guidelines for every new annotation layer (document for yourself and others)
  - Variable (name of the annotation layer)
  - Variants (tagset)
- Tools
  - Praat\textsuperscript{18}, EXMARaLDA\textsuperscript{19}, ...

\textsuperscript{18} Boersma (2001), http://praat.org
\textsuperscript{19} Schmidt and Wörner (2009), http://exmaralda.org
Layer 4 – variable: part-of-speech, variants: article (ART), noun (NN), preposition (APPR), finite non-auxiliary verb (VVFIN), adverb (ADV)\(^{20}\)

Layer 5 – variable: elision, variants: elided form (e), full form (f)

‘Yes of course, the boletes go into the gravy’

2. Corpus preparation
2.4 Annotation and agreement

- Improvement via inter-annotator agreement\textsuperscript{21}
  - Sample a small set of your data and let someone else annotate it using your guidelines
  - Check inter-annotator agreement
  - Identify problems and difficulties
    - Which tags are difficult to allocate? Why?
    - Is the tagset too fine-grained/too coarse-grained for covering the phenomenon?
    - Always have tags for “not interpretable (signal)”, “ambiguous content for the tagset”, “no tag of the tagset assignable”
  - Improve your guidelines

\textsuperscript{21} Cf. Artstein (2017) for an overview; one of the most common measures is Cohen’s $\kappa$ (Cohen 1960)
3. Corpus analysis
3.1 Method

- Best practice in corpus linguistics: variationist approach \(\text{(Rissanen 2008)}\)
  \(\Rightarrow\) make sure you can identify the contexts in which your phenomenon does \textbf{not} show up

- Tools: Corpus search engines
  - ANNIS\(^{22}\)
  - EMU\(^{23}\)

- Let's look at the same phenomenon in GECO and BeMaTAC: \textit{ähm} (Engl. \textit{uhm})

---

\(^{22}\) Krause and Zeldes (2016), \url{https://korpling.german.hu-berlin.de/annis3/}, \url{http://corpus-tool.org}

\(^{23}\) Harrington (2010)
3. Corpus analysis
3.2 Two microstudies

Example of *ähm* in BeMaTaC, queried with ANNIS (frequency analysis)

<table>
<thead>
<tr>
<th>tok</th>
<th>96</th>
<th>97</th>
<th>98</th>
<th>99</th>
<th>100</th>
<th>101</th>
<th>102</th>
<th>103</th>
<th>104</th>
</tr>
</thead>
<tbody>
<tr>
<td>instructor:dipl</td>
<td>hinter</td>
<td>dem</td>
<td>Toaster</td>
<td>muss</td>
<td>rechts</td>
<td><strong>ähm</strong></td>
<td>tschuldigung</td>
<td>links</td>
<td>abgebogen</td>
</tr>
<tr>
<td>instructor:utt</td>
<td>utt</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>instructor:df</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>instructor:repair</td>
<td>type</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>instructor:repair</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>instructor:subrep</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>len</td>
<td>0.270</td>
<td>0.156</td>
<td>0.487</td>
<td>0.529</td>
<td>0.393</td>
<td>0.376</td>
<td>0.386</td>
<td>0.337</td>
<td>0.434</td>
</tr>
</tbody>
</table>

‘Behind the toaster one needs to go to the right *ähm* sorry to the left’

---

24 Query reference link: https://korpling.german.hu-berlin.de/annis3/?id=5b4256e8-72a0-4c1a-b364-d8f635fa67ed
3. Corpus analysis
3.2 Two microstudies: outcome of pipeline 1)

Percentage of ähm for ähm vs. äh in L1 and L2 German in BeMaTaC

(cf. Belz et al. 2017: 135)
3. Corpus analysis

3.2 Two microstudies

Example of *ähm* in GECO (my annotation), queried with *emuR*\(^{25}\) (acoustic analysis)

\(^{25}\) Winkelmann et al. (2017, 2016)
3. Corpus analysis

3.2 Two microstudiesoutcome of pipeline 2)

Extraction of formants for ähm with emuR and plotted in R (R Core Team 2016) with ggplot2 (Wickham 2009)
4. Corpus documentation

Documentation is the key to accessibility, useability and reuseability

- Information about corpus design, preparation, formats, content, ...
- **Metadata** (Odebrecht 2014)
  - Within corpus: authors, license, release version, speakers, task, ...
  - Within documents: authors, segmentation, annotations, speaker, ...
  - Within annotation: authors, segmentation, reference layer, tagset, ...

- Who will host your data once the project is finished?
  - **sustainability** (Simons and Bird 2008)

- Open access with Creative Commons\(^{26}\) licenses

\(^{26}\) https://creativecommons.org/
Summary/take-home

- Corpus preparation
  - Design & Recording
  - Segmentation & Annotation
- Corpus analysis
  - Tools depend on research question
    (Analysis of frequencies or analysis of acoustic measures)
- Documentation
  - Best case: open access
Thank you very much!
References I


Lüdeling, Anke; Sauer, Simon; Belz, Malte, and Mooshammer, Christine (2015): Error annotation in spoken learner corpora. In: Trouvain, Jürgen; Zimmerer, Frank; Gósy, Mária, and Bonneau, Anne (eds.): *IFCASL Workshop on Phonetic Learner Corpora*. Glasgow.


Sauer, Simon (2014): BeMaTaC. A deeply annotated multimodal map-task corpus of spoken learner and native German. URL: http://u.hu-berlin.de/bematac (visited on 06/16/2014).


Wieling, Martijn; Grieve, Jack; Bouma, Gosse; Fruehwald, Josef; Coleman, John, and Liberman, Mark (2016): Variation and change in the use of hesitation markers in Germanic languages. In: Language Dynamics and Change 6 (2). 199–234.


Winkelmann, Raphael; Jaensch, Klaus; Cassidy, Steve, and Harrington, Jonathan (2016): emuR: Main Package of the EMU Speech Database Management System. Version 0.2.1 (R package).