Corpora of Spoken Language

Anke Lüdeling
Humboldt-Universität zu Berlin
introduction

‘Corpus‘ can mean many different things – it is, however, important to know about the corpus design to know what one can do with a corpus. Corpus annotation makes the interpretation of the data transparent. It is also important to know about the corpus architecture & format to understand how a corpus can be searched and stored.
what is a corpus?

„A corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language.” (EAGLES, emphasis added)

“Words such as collection and archive refer to sets of texts that do not need to be selected, or do not need to be ordered, or the selection and/or ordering do not need to be on linguistic criteria. They are therefore quite unlike corpora.” (EAGLES, emphasis added)
corpora

• can contain any language variety one needs to answer one's research question (genre, time, place, situation, etc.)
• can be large (Web corpora, billions of tokens) and small (a dialogue, a poem, etc.)
• can be fixed (reference corpora) or growing (monitor corpora)
• can be monomodal (written) or multimodal (spoken & written & gestures & ...) (→ sign-language corpora)
corpora & annotation

while it is often useful to have a digitally available text, one of the biggest advantages of using corpora is that the primary data can be explicitly and transparently annotated

• it is not possible not to interpret a text in research
• interpretation depends on many issues (research question, similarity measure, tradition, etc.)
• even for the same type of category we can have many different ways of interpreting the same data (think part-of-speech but also categories like 'loud' or 'long')
• annotation makes the interpretation visible and only if the interpretation is accessible with the data is it possible to understand and replicate results

Leech (1997), Atwell et al. (2000), Lüdeling (2011), ...
research questions

qualitative research
• editing
• hermeneutic research
• example bank

quantitative research
• exploration
• experiments
• modelling
spoken corpora

(in contrast to speech corpora which are huge collections of spoken data used for technological purposes)

• spoken corpora are typically small(ish)
• in addition to the sound file they contain at least one written layer (transliteration or transcription, often additional normalization layers)
• sometimes spoken corpora contain additional layers of annotation which can range from 'standard' annotation layers (part of speech, lemma, etc.) to specific layers (phonetic annotation, disfluency, etc.)

the Berlin Map Task Corpus (BeMaTaC)

- small dialogue corpus, 12 map task dialogues
- video (hand of instructee) & audio
- transcribed, tokenized & aligned (Praat), annotated with pos & lemma (TreeTagger)
- multi-layer format: converted to RelAnnis, freely available in Annis

research questions for spoken corpora

qualitative research
• hermeneutic research
• example bank

quantitative research
• exploration
• experiments
• modelling

with regard to
• phonetic issues
• communication issues
• rhetorical structure
• register
• grammar
• lexicon
• processing
• ...
grammatical phenomenon

VERBLESS UNITS
syntactic analysis of spoken language

research topic: syntactic analysis of spoken utterances

problems:
• most grammars deal with highly idealized (written) language
• in most grammars the category 'sentence' is the basic unit of analysis – a sentence or clause depends on a (finite) verb

What can we do to analyze verbless units?

___________
grammatical / canonical

- grammatical – a basic category in many grammar theories
  (can be generated by the (internal) grammar)
- canonical – here used as a technical term
  can be analyzed by a given grammar

Verbless units are non-canonical for most grammars (traditional grammars as well as the more theoretic/formal grammars)
syntactic analysis of non-canonical units

What can be done with non-canonical units?

• change grammar
• ignore non-canonical structure by either not annotating it or using an unsuitable structure
• mark as non-canonical and analyze in a different way

Marcus, Marcinkiewicz & Santorini 1993, Sampson 1995, Granger 2009, etc.
syntactic analysis of non-canonical utterances

• change grammar
  – reflects the idea that different registers may have different grammars
  – makes it difficult to compare varieties

• ignore non-canonical structure by either not annotating it or using an unsuitable structure
  – non-canonical structures cannot be found and studied systematically

• mark as non-canonical and analyze in a different way
  – makes it possible to identify and analyze non-canonical structures
  – makes it possible to compare varieties (qualitatively and quantitatively)
define utterance

- canonical
  - analyze according to model

- non-canonical
  - mark as non-canonical
    - analyze difference

(Hirschmann, Doolittle & Lüdeling 2007)
example

also äh oben so ne Art Rahmen zeichnen von von dem Bild ja äh dann gehste rechts [BeMaTaC_L1_2013-01]
"well eh above some kind of frame to draw (infinitive) of of the picture yes eh then you go right"

also äh oben so ne Art Rahmen zeichnen von von dem Bild
→ mark as non-canonical
ja äh dann gehste rechts
→ canonical, analyze according to framework
example

also äh oben so ne Art Rahmen zeichnen von von dem Bild

formulate a target hypothesis (here a minimal change to make the sentence canonical)

also äh oben [musst du] so ne Art Rahmen zeichnen von von dem Bild

also äh oben [MODAL VERB 3rd Sg. du] so ne Art Rahmen zeichnen von von dem Bild

analyze the difference between the original unit and the target hypothesis
aside: target hypothesis

• analysis of differences only possible in contrast to a target hypothesis
• often different target hypotheses possible (long discussion in analysis of learner corpora)
• the target hypothesis has **no theoretical status** (it does not make a sentence 'correct')
  – it is merely a technical step in the analysis

---

Lüdeling (2007), Reznicek et al. (2013)
relevant?

- experiment using 500 sentences each from 7 varieties
verbless units

What types of verbless units do we find in BeMaTaC?
• interjections, short answers, formulae, etc.
• disfluencies
• infinitives ("cook book style")
• (sometimes long) sequences of adverbial phrases (missing modals)
• ...

each of these needs a different analysis – with target hypotheses different verbless units can be found systematically
verbless units – research questions

• How do the elements in verbless units combine?
• How can 'arguments' be assigned without a verb?
• How can temporal information be assigned?
• What is the theoretical status of a finite verb if finite verbs are not always necessary?
verbless units – summary

• very common in spoken language
• interesting syntactically – grammar might be different from what is often assumed
• interesting processually – it is unproblematic to understand them
• a relevant register feature
processing

DISFLUENCIES
“The Watergate tapes are the most famous and extensive transcripts of real-life speech ever published. When they were released, Americans were shocked, though not all for the same reason. Some people – a very small number – were surprised that Nixon had taken part in a conspiracy to obstruct justice. A few were surprised that the leader of the free world cussed like a stevedore. But one thing that surprised everyone was what ordinary conversation looks like when it is written down verbatim. Conversation out of context is virtually opaque.” [Pinker 1995, 224]
disfluencies - forms

• unfilled pauses
• lengthening
• repetitions (sounds, syllables, words)
• repairs
  (complex: reparandum, interregnum, reparans)
• filled pauses (äh, ähm, perhaps also: sozusagen 'so to say', ich mein 'I mean', etc.)
disfluencies

interesting because

• different forms

• different functions
  processing issues as well as communication issues (signals for turn holding, turn relinquishing, etc.)

• difficult to integrate into grammatical theories (no theoretical status in current theories)
disfluencies

• again: using target hypotheses helps find the different types of disfluencies (form) in a systematic way

• then a layer/several layers of disfluency annotation can be added

• disfluencies often appear together - it is interesting to see the interaction of different types of disfluencies

ja okay gut darauf läufst du geradeaus zu und ähm machst äh rechts oder beziehungsweise gehst rechts herum einmal [BeMaTaC L1_2013-01]
"yes okay well you go straight in that direction and ehm make eh right or rather go right once"

---

genau du gehst b/ gehst bis geh/ gehst dieses gehst dieses Wohnwagenbild hoch [BeMaTaC L1_2013-01]
"exactly you go t/ go to go go this go this picture of a caravan up"
disfluency annotation

• tagset and guidelines for disfluencies, several levels
  (already finished for BeMaTaC L1, done by Malte Belz and Myriam Klapi)

• systematic disfluency annotation is useful for qualitative studies as well as quantitative studies (here comparison of native German speakers and learners of German as a foreign language)
Belz & Klapi (2013) show that learners of German as a foreign language make longer pauses followed by longer fillers than native speakers (data BeMaTaC L1 & L2)
multi-layer annotation

• disfluency annotation and target hypotheses make it possible to add a syntactic annotation layer (here dependency annotation)
• work in progress – no (quantitative) results yet

Dipper, Lüdeling & Reznicek (to appear),
Webanno: http://code.google.com/p/webanno/
how far beyond the apples [do I go]?
example with disfluencies

da bist du hast du in der Mitte des Blattes diese Äpfel und dann gehst du von dem Punkt auf dem du da warst gehst du zu den Äpfeln hoch

[Bematac_2012-11-02-B]

"there you are you have in the middle of the sheet these apples and then you go from the point at which you were you go up to the apples"
example with disfluencies


"there you are you have in the middle of in the middle of the sheet these apples and then you go from the point at which you were you go up to the apples"
categories in red circles are 'regular' categories (used in dependency schemes), categories in blue circles are added to deal with disfluencies (COR – self correction)
disfluency classes (self repair)

"you are you have in the middle of in the middle of the sheet these apples"
summary

• spoken language differs in many interesting ways from 'canonical' (modelled after written) language (and this is not reducible to the difference between 'competence' and 'performance')

• spoken corpora are interesting resources for the study of spoken language if they are well designed, transparently annotated & publicly available

• we have the technical means (multi-layer architectures, annotation tools, search tools, etc.) but we have to think much more about the conceptual issues
thank you
danke

a special thank you to Malte Belz, Marc Reznicek & Simon Sauer

contact: anke.luedeling@rz.hu-berlin.de
BeMaTaC: http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/bematac
references


references

references

• Stegmann, Rosmary, Telljohann, Heike & Hinrichs, Erhard W. 2000. *Stylebook for the German Treebank in VERBMOBIL.*