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.)

____

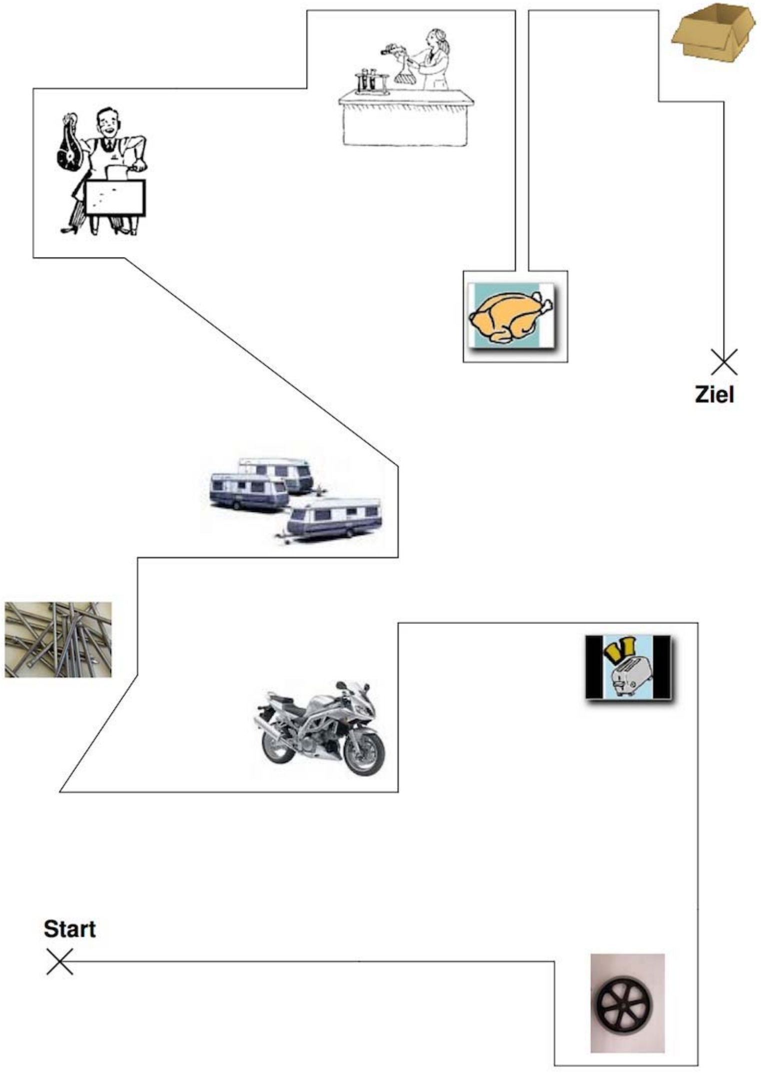
Gibbon, Mertins & Moore (2000), Wichmann (2008), Dahlmann & Adolphs (2009), etc.

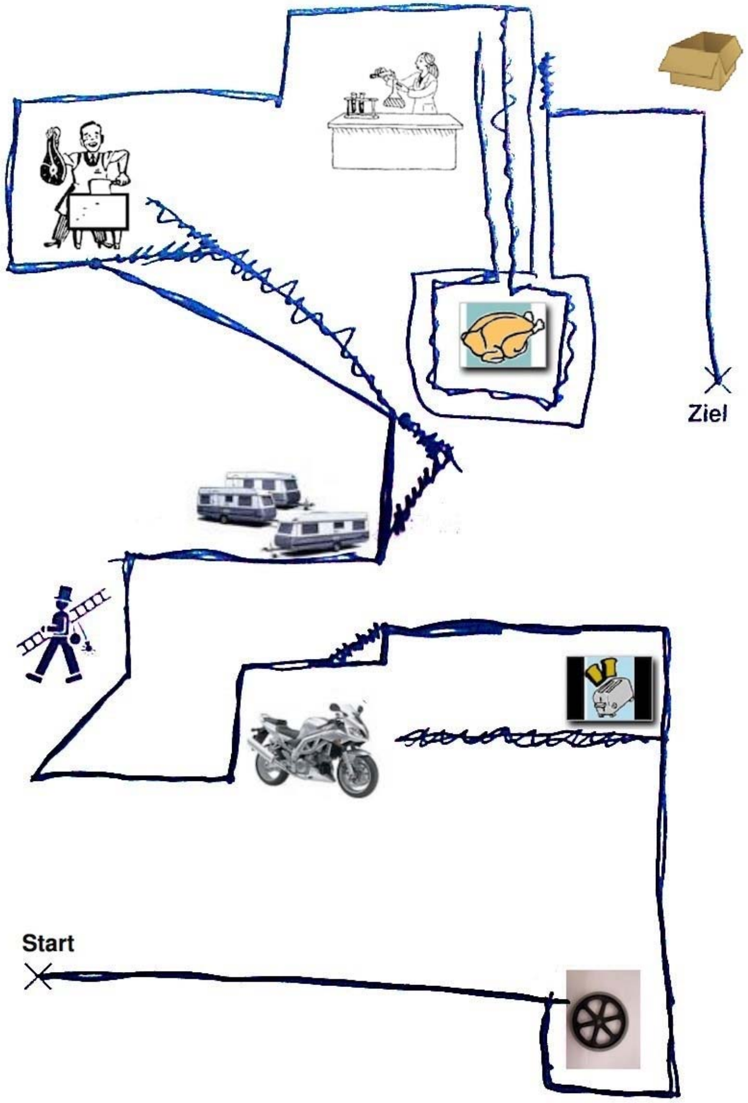
the Berlin Map Task Corpus (BeMaTaC)

- 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

———

Boersma (2010), Giesel et al. (2013), Hedeland & Schmidt (2012), Schmid (1994), Sauer & Lüdeling (2013), Zeldes et al. (2009)


Ziel
Start


Ziel
Start

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?

____

Stegmann, Telljohann & Hinrichs (2000), Dickinson & Meurers (2006), Hennig (2006), etc.

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

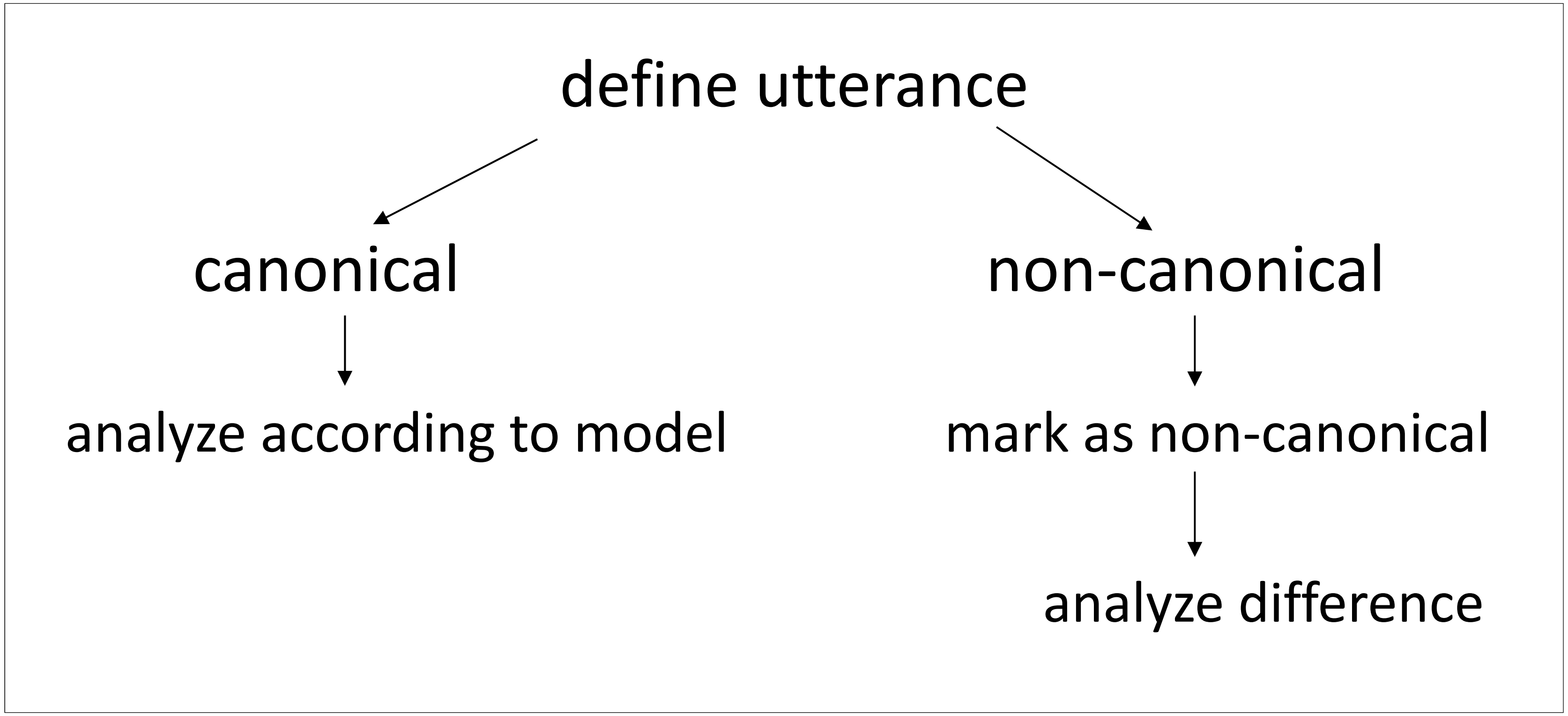
____

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)

analysis



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 $_{\text{(infinitive)}}$ 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

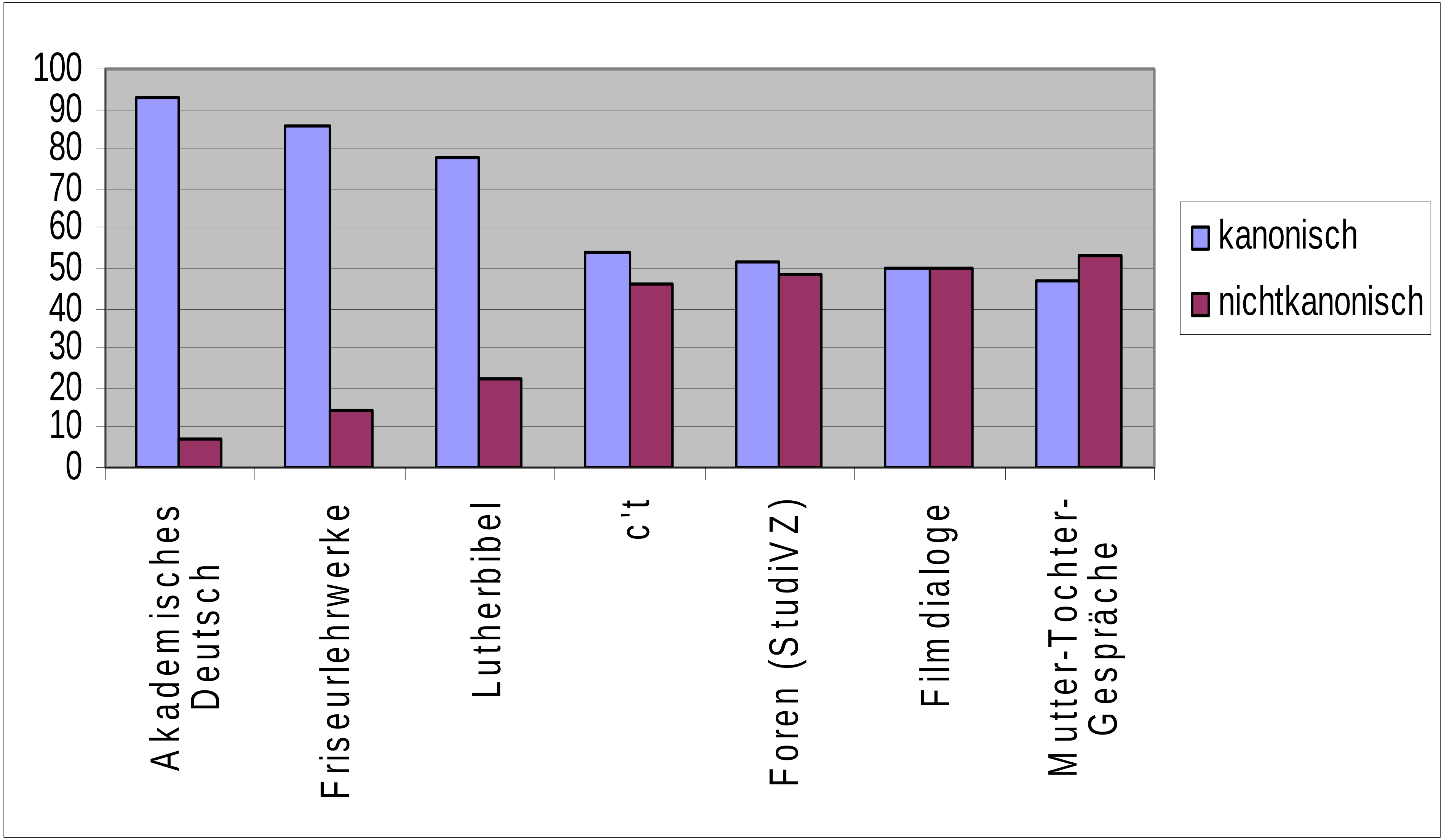
aside: target hypothesis

- 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


100
90
80
70
60
50
40
30
20
10
0
kanonisch
nichtkanonisch
Akademisches Deutsch
Friseurlehrwerke
Lutherbibel
c't
Foren (StudiVZ)
Filmdialoge
Mutter-Tochter-Gespräche

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

____

Fox Tree & Clark (1997), Bortfeld et al. (2001),Eklund (2004, 2012), Gilquin & de Cock (2011), etc.

example

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"

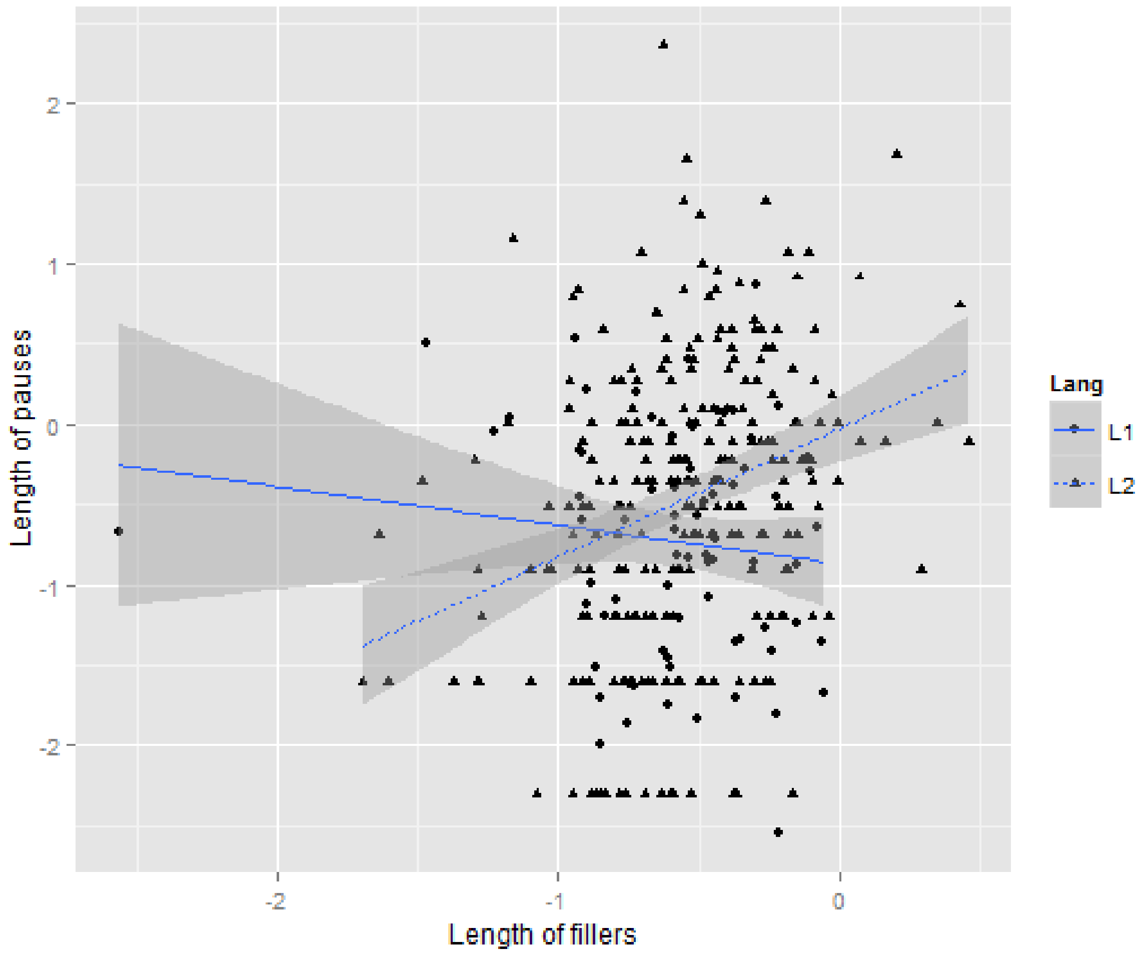
---

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 forein language)



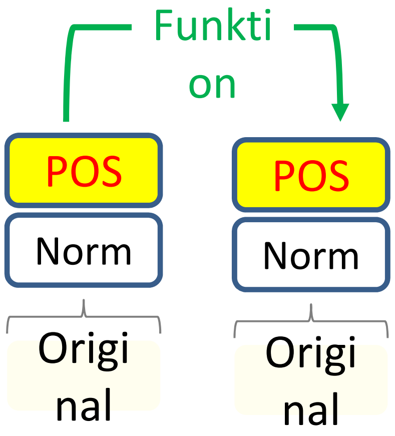
Belz & Klapi (2013) show that learners of German as a forein language make longer pauses followed by longer fillers than native speakers (data BeMaTaC L1 & L2)

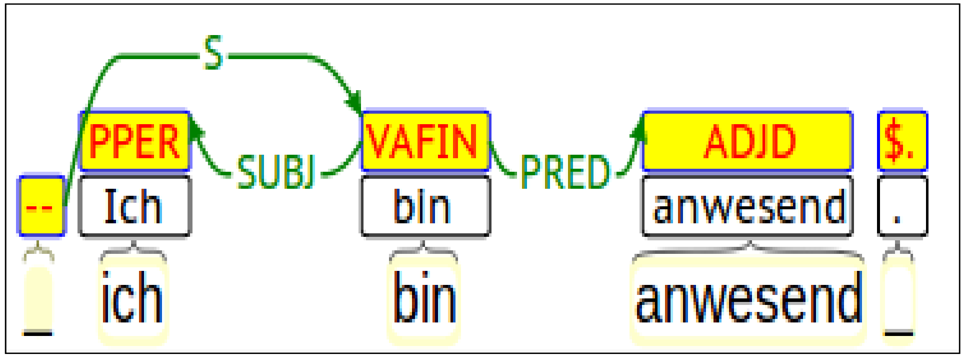
multi-layer annotation

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

____

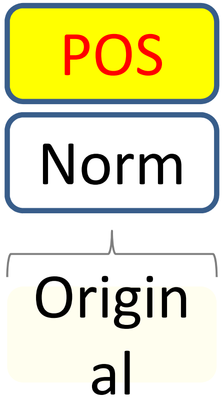
Dipper, Lüdeling & Reznicek (to appear),
Webanno: http://code.google.com/p/webanno/





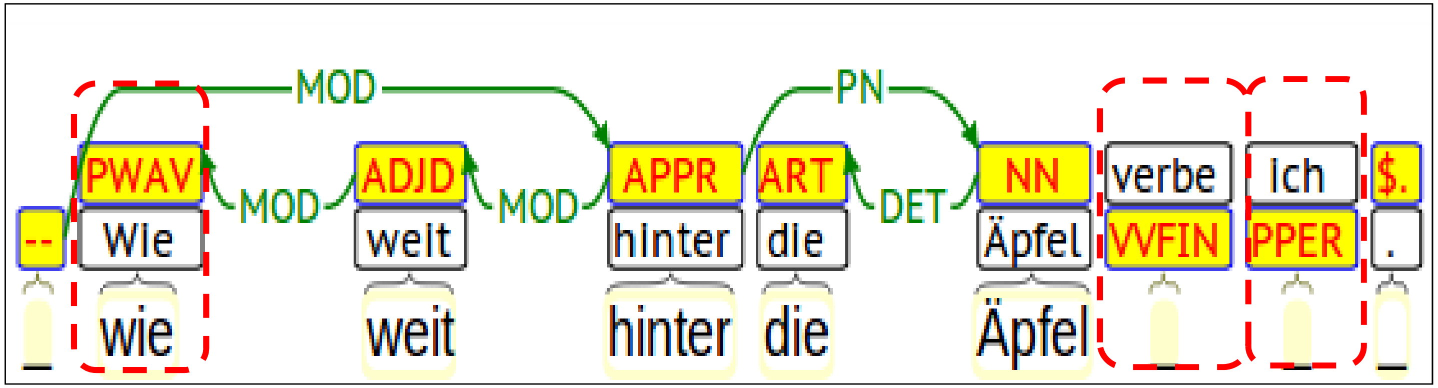
Bematac_2012-11-02-B

how far beyond the apples [do I go]?


POS

Norm

Origin al



Bematac_2012-11-02-B

change

PWAV

wie

Wie

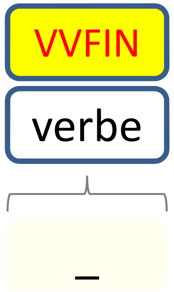
insertion

PPER

ich

_

'abstract' verb


VVFIN

verbe

_

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]

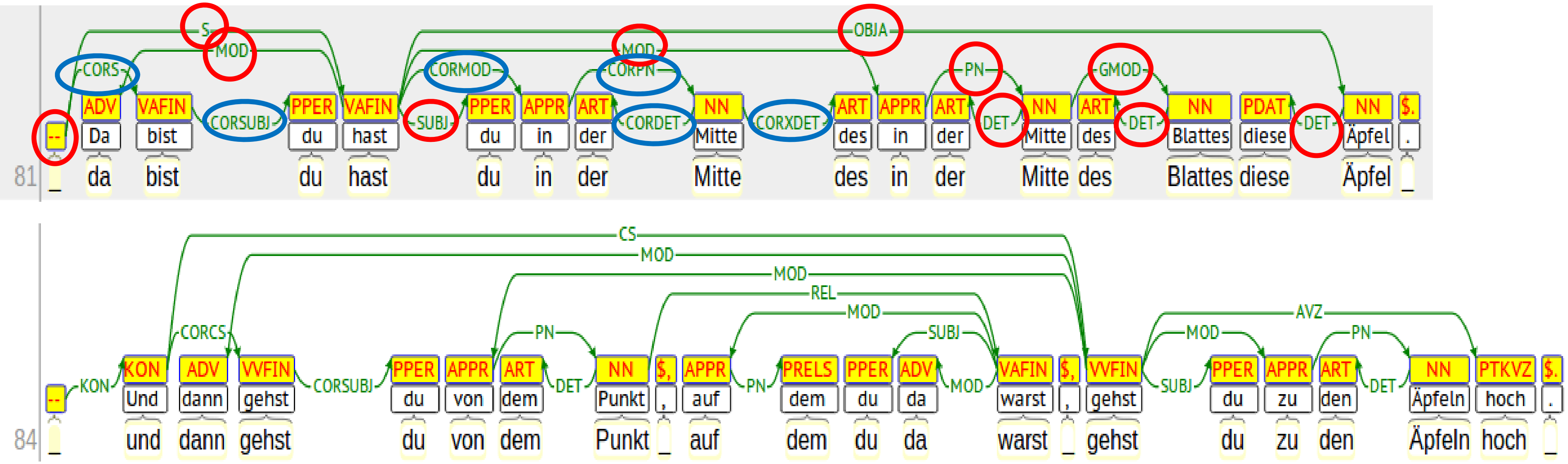
"there you are you have in the middle of the sheet these apples and then you go from the point at which you were you go up to the apples"

example with disfluencies

da bist du hast du in der Mitte des 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]

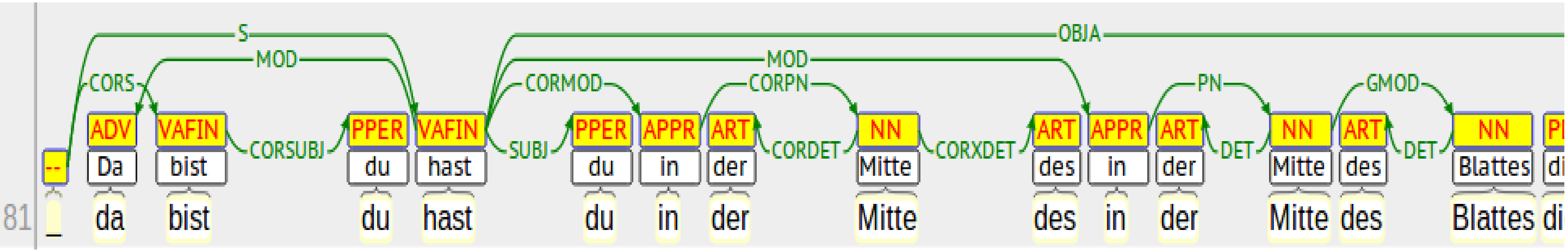
"there you are you have in the middle of in the middle of the sheet these apples and then you go from the point at which you were you go up to the apples"

categories in red circles are 'regular' categories (used in dependency schemes),
categories in blue circles are added to deal with disfluencies (COR – self correction)

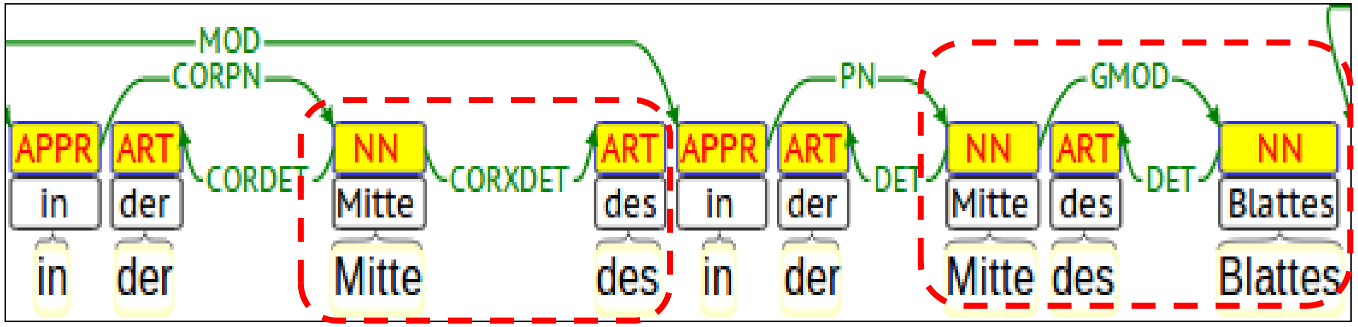

S
MOD
CORS
OBJA
CORMOD
MOD
CORPN
PN
GMOD
CORSUBJ
SUBJ
CORDET
CORXDET
DET
DET
DET
-- ADV VAFIN PPER VAFIN PPER APPR ART NN ART APPR ART NN ART NN PDAT NN $.
-- Da bist du hast du in der Mitte des in der Mitte des Blattes diese Äpfel .
81 _ da bist du hast du in der Mitte des in der Mitte des Blattes diese Äpfel _
CS
MOD
MOD
REL
MOD
AVZ
CORCS
PN
SUBJ
MOD
PN
KON
CORSUBJ
DET
PN
MOD
SUBJ
DET
-- KON ADV VVFIN PPER APPR ART NN $, APPR PRELS PPER ADV VAFIN $, VVFIN PPER APPR ART NN PTKVZ $.
-- Und dann gehst du von dem Punkt , auf dem du da warst , gehst du zu den Äpfeln hoch .
84 _ und dann gehst du von dem Punkt _ auf dem du da warst _ gehst du zu den Äpfeln hoch _

Bematac_2012-11-02-B

disfluency classes (self repair)


S
OBJA
MOD
MOD
CORS
CORMOD
CORPN
PN
GMOD
CORSUBJ
SUBJ
CORDET
CORXDET
DET
DET
-- ADV VAFIN PPER VAFIN PPER APPR ART NN ART APPR ART NN ART NN PI
Da bist du hast du in der Mitte des in der Mitte des Blattes di
81 _ da bist du hast du in der Mitte des in der Mitte des Blattes di

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


MOD
CORPN
PN
GMOD
CORDET
CORXDET
DET
DET
APPR ART NN ART APPR ART NN ART NN
in der Mitte des in der Mitte des Blattes
in der Mitte des in der Mitte des Blattes

Bematac_2012-11-02-B:

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

- Anderson, Anne H., Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry Thompson & Regina Weinert (1991) The HCRC Map Task Corpus. *Language and Speech* 34, 351-366.
- Atwell, Eric, Demetriou, George, Hughes, John, Schiffrin, Amanda, Souter, Clive & Wilcock, Sean 2000. 'A comparative evaluation of modern English corpus grammatical annotation schemes', *ICAME JOURNAL* 24: 7–24.
- Belz, Malte & Myriam Klapi (2013) Pauses following Fillers in L1 and L2 German Map Task Dialogues. In: *Proceedings of the 6th Workshop on Disfluency in Spontaneous Speech.* Stockholm, Sweden.
- Boersma, Paul 2010. 'Praat, a system for doing phonetics by computer', *Glot International* 5(9/10): 341–345.
- Bortfeld, Heather, Leon, Silvia D., Bloom, Jonathan E., Schober, Michael F. & Brennan, Susan E. 2001. 'Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender', *Language and Speech* 44(2): 123–147.
- Dahlmann, Irina & Adolphs, Svenja 2009. 'Spoken Corpus Analysis: Multi-modal Approaches to Language Description', *Contemporary Corpus Linguistics. London/New York: Continuum* 1: 125–139.
- Dickinson, Markus & Meurers, W. Detmar 2006. 'Detecting Annotation Errors in Spoken Language Corpora', *Copenhagen studies in language* 32: 53–67.
- Dipper, Stefanie, Lüdeling, Anke & Reznicek, Marc 2013. 'NoSta-D: A Corpus of German Non-Standard Varieties', in Zampieri, Marcos (ed.) *Non-Standard Data Sources in Corpus-Based Research*. Shaker Verlag.

references

- Eklund, Robert (2004) Disfluency in Swedish human-human and human-machine travel booking dialogues. PhD thesis. Linköping : Linköpings Universitet. url: http://www.ida.liu.se/~g-robek/pdf/Eklund_2004_PhD_Thesis_Corrected.pdf
- Gibbon, Dafydd, Mertins,Inge & Moore, Roger K. (eds.) 2000. *Handbook of Multimodal and Spoken Dialogue Systems. Resources, Terminology and Product Evaluation.* Dordrecht: Kluwer Academic Publishers.
- Giesel, Linda, Myriam Klapi, Daisy Krüger, Isabelle Nunberger, Oxana Rasskazova & Simon Sauer (2013). *A deeply annotated multimodal map-task corpus of spoken learner and native German*. DGfS Potsdam. url: http://www.linguistik.huberlin.de/institut/professuren/korpuslinguistik/forschung/bematac
- Gilquin, Gaëtanelle & Cock, Sylvie De 2011. 'Errors and disfluencies in spoken corpora. Setting the scene.', *International Journal of Corpus Linguistics* 16(2): 141-172.
- Hedeland, Hanna & Schmidt, Thomas . 2012. 'Technological and Methodological Challenges in Creating, Annotating and Sharing a Learner Corpus of Spoken German', in Schmidt, Thomas & Wörner, Kai (eds.) *Multilingual Corpora and Multilingual Corpus Analysis*. Amsterdam: John, pp. 25-46.
- Hennig, Mathilde 2006. *Grammatik der gesprochenen Sprache in Theorie und Praxis*. Kassel University Press.
- Leech, Geoffrey N. 1997. 'Introducing Corpus Annotation', in Garside, Roger, Leech, Geoffrey N. & McEnery, Tony (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman, pp. 1–18.
- Lüdeling, Anke 2011. 'Corpora in Linguistics: Sampling and Annotation', in Grandin, Karl (ed.) *Going Digital. Evolutionary and Revolutionary Aspects of Digitization.* (Nobel Symposium 147.) New York: Science History Publications/USA, pp. 220–243.

references

- Pinker, Steven (1995) *The Language Instinct. How the Mind Creates Language*. HarperPerennial, New York.
- Sauer, Simon & Anke Lüdeling (2013) BeMaTaC. A Flexible Multi-Layer Spoken Corpus for Contrastive SLA Analysis. Talk atICAME 34, Santiago de Compostela, Mai 2013. (publication in preparation)
- Schmid, Helmut 1994. 'Probabilistic Part-of-Speech Tagging Using Decision Trees', in *Proceedings of the Conference on New Methods in Language Processing*. Manchester, UK, pp. 44–49.
- Schmidt, Thomas, Hanna Hedeland, Timm Lehmberg & Kai Wörner (2010). *HAMATAC - The Hamburg MapTask Corpus*. url: http://www.exmaralda.org/files/HAMATAC.pdf.
- Stegmann, Rosmary, Telljohann, Heike & Hinrichs, Erhard W. 2000. *Stylebook for the German Treebank in VERBMOBIL*.
- Wichmann, Anne 2008. 'Spoken corpora and speech corpora', in Lüdeling, Anke & Kytö¶, Merja (eds.) *Corpus Linguistics. An International Handbook*. (HSK.) Vol. 1. Berlin: Mouton de Gruyter, pp. 187–207.
- Zeldes, Amir, Ritz, Julia, Lüdeling, Anke & Chiarcos, Christian 2009. 'ANNIS: A Search Tool for Multi-Layer Annotated Corpora', in *Proceedings of Corpus Linguistics 2009, Liverpool, July 20-23, 2009.*