Learner corpora and Second Language Acquisition research

Anke Lüdeling

CLiF, Bloomington, June 2016
1 Corpus design and collection
   - Corpus design: samples, learner-specific issues
   - Hands-on session 1

2 Error Annotation

3 Other annotation layers

4 Quantitative analysis

5 Summary
“The two main, well documented findings of SLA research of the past few decades are as follows:

- second language acquisition is highly systematic
- second language acquisition is highly variable

Although these two statements might appear contradictory at first sight, they are not. The first one primarily refers to what has been called the route of development (the nature of the stages all learners go through when acquiring the second language - L2). This route remains largely independent of both the learner’s mother tongue (L1) and the context of learning (e.g. whether instructed in a classroom or acquired naturally by exposure). The second statement usually refers to either the rate of the learning process (the speed at which learners are learning the L2), or the outcome of the learning process (how proficient learners become), or both. We all know that both speed of learning and range of outcomes are highly variable from learner to learner: some do much better much more quickly than others.” (Myles 2000)
Corpus evidence should be used to inform acquisition frameworks and models. Learner corpus data (if collected and prepared according to the relevant principles) can be useful for the study of

- systematicity
- variation.

cf., among many, many others, Bygate 2009; Granger 2002; Granger 2009; Johansson 2009; Lüdeling and Walter 2010; Vyatkina et al. 2015; Wulff and Gries 2010; Römer et al. 2014, and many articles in Granger et al. 2015
Learner corpora contain texts from learners of a second/foreign language (L2). These texts can be

- spoken (signal and/or transcription), written, multimodal
- produced by learners of any proficiency level and any L1
- produced freely or task-based
- longitudinal or cross-sectional
- “raw” or annotated

Note: Although it is highly relevant for acquisition theory, in this summer school we will not make a difference between Second Language Acquisition (untutored) and Foreign Language Acquisition (tutored). There is a large body of research on L3 etc. – here I will use L2 for any Ln. (And we are not interested in first language acquisition corpora.)

For a comprehensive list of learner corpora see the “Corpora around the world” page maintained by the Centre for English Corpus Linguistics at Louvain la Neuve, https://www.uclouvain.be/en-cecl-lcworld.html.
Learner data

Learner corpora provide one type of evidence in the study of acquisition. It can (and should) be complemented by other types of evidence (experimental, behavioral, ...).

Corpus data can be used to
- explore a variety or phenomenon
- find hypotheses that can then be tested in other corpora or with other methods
- as “experimental” data that is used in hypothesis testing (careful! more on this later)
- inform models

On combining different methods in acquisition studies, cf. Ellis and Simpson-Vlach (2009), Wulff et al. (2009), and Littré (2015).

Statistical modelling is actually one of the most interesting methods in studying learner data. We do not have time to go into this here but see Gries (2015b) and Gries and Deshors (2015).
Corpus design and corpus annotation depend primarily on the research question. Corpus architecture and formats depend on general considerations (availability, transparency, re-usability, reproducability).
1 Corpus design and collection
   - Corpus design: samples, learner-specific issues
   - Hands-on session 1

2 Error Annotation

3 Other annotation layers

4 Quantitative analysis

5 Summary
statistical aspects  Typically, one wants to analyze a population that is too big for exhaustive analysis and therefore, one has to draw a sample. Samples can be opportunistic, stratified, or representative. The sampling strategy determines the possibilities to extrapolate.

content  Which texts should the learner corpus contain? Which sampling parameters are relevant?

Research question: How much of Germany is covered by forest?
Research question: How much of Germany is covered by forest?

first we have to operationalize what we mean by “covered by forest”
Research question: How much of Germany is covered by forest?

First we have to operationalize what we mean by “covered by forest”

We cannot travel everywhere – how should we proceed?
Research question: How much of Germany is covered by forest?

First, we have to operationalize what we mean by “covered by forest”.

We cannot travel everywhere – how should we proceed?
Research question: How much of Germany is covered by forest?

First we have to operationalize what we mean by “covered by forest”

We cannot travel everywhere – how should we proceed?

Extrapolation from these opportunistic samples would lead to invalid results.
Many corpora (even some “reference” corpora) are opportunistic samples.
Opportunistic sampling

Many corpora (even some “reference” corpora) are opportunistic samples. Many learner corpora are opportunistic samples (my students’ essays, things I happen to get, the texts my test institute is collecting, ...).

Opportunistic corpora can be valuable sources of data. One has to be careful to generalize from the findings, however.
If we know (or suspect) parameters that influence the phenomenon we are interested in, we can build a stratified corpus in which we collect data according to each (combination of) parameters.

- for the forest example: urban/rural, type of ground, privately owned/publicly owned, mountainous/flat, etc.
- for corpus design (depending on the research question): mode, stage of acquisition, L1, task, etc.

Using a stratified sample, we can at the very least find out whether a given parameter has an influence.
Representative sampling

If we know the parameters and their distribution in the population, we can build a representative corpus.

- for the forest example: possible, since Germany is exhaustively mapped
- for corpus design: almost never possible – we typically don’t know the distributions in the population

A representative sample makes extrapolation of the results easier.

The situation is more complicated: Even if representative sampling is not possible, there are approaches that show how close a subcorpus matches a larger corpus. If a phenomenon approaches closure one could (perhaps) assume that the sampling is fairly complete (Cohen et al. 2016). But many linguistic phenomena follow a Zipfian distribution – closure is not expected (Baayen 2001; Baroni 2009).
The content of the learner corpus depends on the research question(s).

There are many possible parameters and collection strategies. Here I want to briefly talk about just three aspects that are often discussed with respect to learner corpora.

- level of proficiency
  (related: cross-sectional, quasi longitudinal, longitudinal)
- authenticity/naturalness; task dependency
- mode and register
It is often interesting (crucial!) to note the level of proficiency of the contributors for a learner corpus. How can the level of proficiency be established?

- in tutored acquisition: length of instruction

- test score (CEFR score, TOEFL score, assessment by professional testers, self assessment, etc.)
It is often interesting (crucial!) to note the level of proficiency of the contributors for a learner corpus. How can the level of proficiency be established?

- in tutored acquisition: length of instruction
  → problematic: different backgrounds (L1 ...Ln, differences in instruction, individual differences, etc.)

- test score (CEFR score, TOEFL score, assessment by professional testers, self assessment, etc.)
It is often interesting (crucial!) to note the level of proficiency of the contributors for a learner corpus. How can the level of proficiency be established?

- in tutored acquisition: length of instruction
  → problematic: different backgrounds (L1 ...Ln, differences in instruction, individual differences, etc.)

- test score (CEFR score, TOEFL score, assessment by professional testers, self assessment, etc.)
  → problematic: reliability, different scores test different abilities, comparability, etc.
“First the writing samples were holistically assessed for general proficiency as Ortega pointed out that studies that operationalize proficiency in terms of holistic ratings yielded more homogenous observations for the measures she studied as reflected in smaller standard deviations and narrower ranges than studies that operationalize proficiency in terms of naturally occurring classes or groups (2003, p. 502). Our procedure was carefully controlled to ensure high interrater reliability. A group of eight experienced ESL teachers (three native speakers of English, two of Dutch, and one each of Chinese, Portuguese, and Spanish) established scoring criteria as follows: Each rater first judged a set of six samples to see which they thought were the strongest and which ones the weakest in general English proficiency. These orders were discussed among the raters. From the discussions amongst the raters a range of factors emerged that are very much in line with general CAF measures: text length, sentence length, sentence complexity, use of different types of clauses, use of tense, aspect, voice and mood, vocabulary range, use of L1, use of idiomatic language, and accuracy.”
“However, many texts were difficult to rank. Some had no errors but contained only very simple constructions; others had a great many errors but contained different complex constructions. Some samples were discussed at length before agreement was reached. After the group reached agreement on the rank order, the texts were tentatively classified into proficiency levels. Following this procedure, the raters worked together with some 100 samples until they had settled on six proficiency levels (0–5). Assessment criteria were then established, which included the main characteristics of each level that had emerged from the discussions to help the raters classify the remainder of the samples. Then the raters were divided into two groups of four, first assessing the samples individually and then comparing the scores. For each text, the majority score (at least 3 out of 4) was taken and if there was no majority, differences were resolved through discussion. If a group was unable to reach consensus, the other group of raters was consulted.” (Verspoor et al. 2012, p. 243)
It is often interesting (crucial!) to note the level of proficiency of the contributors for a learner corpus.

How can the level of proficiency be established?

- in tutored acquisition: length of instruction
  → problematic: different backgrounds (L1 ...Ln, differences in instruction, individual differences, etc.)
- test score (CEFR score, TOEFL score, assessment by professional testers, etc.)
  → problematic: reliability, different scores test different abilities, comparability, etc.

Since it is unclear what is meant by “beginning”, “intermediate”, “advanced”, or other labels, the assessment criteria must be stated in the metadata for each text (and the assessment procedure must be described in the documentation).

See also Walter and Grommes (2008), or the discussion on criterial features (Hawkins and Filipović 2012). Regarding the validity of the CEFR scales, see e.g. Wisniewski (2012), see also the MERLIN project at http://merlin-platform.eu/.
A related design decision concerns the “experimental” setting:

**cross-sectional**: texts from learners with the same level of proficiency; typically used in comparison studies (such as Contrastive Interlanguage Analysis → Thursday)

**longitudinal**: texts produced by the same learners across several levels of proficiency; used for development studies – difficult to get, rare! (but see Byrnes (2009) and Vyatkina et al. (2015))

**quasi-longitudinal**: texts produced by different learners across several levels of proficiency; used for developmental studies – problematic because other parameters may intervene
For a long time, there was a (for me, entirely artificial and superfluous) split between learner corpora and learner data collected in SLA research (a good overview of the situation is given in Tracy-Ventura and Myles 2015).

**learner corpora** were supposed to be “authentic” or “natural”, large, to be analyzed statistically, often used by many people for many different questions*

**SLA datasets** were supposed to be small, tightly controlled, experimental, typically not re-used

*Some authors speak of “general purpose learner corpora” that are constructed “without a specific research agenda in mind” (Tracy-Ventura and Myles 2015, p. 60). Constructing a corpus without a specific agenda in mind is impossible. Purposes or questions can vary with respect to their specificity.
What would be natural?
What would be natural?

It has been argued that natural (authentic) texts are those that are not produced specifically for a corpus but would be produced anyway.
What would be natural?

It has been argued that natural (authentic) texts are those that are not produced specifically for a corpus but would be produced anyway.

What’s “natural” about essay writing?
Or test situation?
Or when a microphone is present?
What would be natural?

It has been argued that natural (authentic) texts are those that are not produced specifically for a corpus but would be produced anyway.

What’s “natural” about essay writing?
Or test situation?
Or when a microphone is present?

Again: What would be natural? And why is natural better?
More and more people understand that *every text* is produced in a specific set of circumstances and that this set influences the text (on every linguistic level). Less constrained settings (regarding topic, time, participants, etc.) can provide answers to *different* questions than more constrained/controlled situations. It is therefore important to understand the situation – and record it in the metadata (prompt, production circumstances, available aids, etc.). In recent years so-called task-based corpora are becoming more popular.

Compare register studies and variation studies – the more production parameters/functions/settings/etc. are studied, the more influencing factors are found (Labov 2004; Biber 2006; Biber and Conrad 2009; Eckert 2012; Kunz and Lapshinova–Koltunski 2015, among many others). For contrastive and learner corpus studies and register see Ädel (2008), Gilquin and Paquot (2008), Deshors (2015), and Neumann (2014).

About task-based corpora, see Quixal (2012), Gablasova et al. (2016), and Quixal and Meurers (2016). The term *task* is used in different ways; sometimes it is understood as referring only to classroom-based instruction.
Since each text is influenced by many factors (speaker factors, situational factors, etc.) in addition to the producers linguistic background, all of these have to be considered.* Which registers are to be included in a corpus should depend only on the research question(s).

Here I want to look at just one dimension of this decision: spoken** vs. written registers.

---

*Aside: Many existing learner corpora contain academic speech and language, and sometimes far-reaching conclusions about the acquisition process are drawn from these corpora. It is evident that students do not represent all learners.

**There are many studies concerning the speech signal itself (phonetic features of certain sounds, intonation, etc.). See Ballier and Martin (2015) and a forthcoming issue of the International Journal of Learner Corpus Research. We will come back to disfluencies on Wednesday.
The old distinction between mode and concept (introduced by Koch and Oesterreicher 1985, similar models elsewhere) is still helpful.

Abbildung 2.1: Schematische Anordnung diverser Äußerungsformen auf der Skala konzeptioneller Mündlichkeit/Schriftlichkeit nach Koch/Oesterreicher (a = familiäres Gespräch, b = Telefongespräch, c = Privatbrief, d = Vorstellungsgespräch, e = Zeitungsinterview, f = Predigt, g = wissenschaftlicher Vortrag, h = Leitartikel, i = Gesetzentext, a’ = verschriftetes Familiengespräch, i’ = verlautlichter Gesetzentext)
There are (infinitely) many spoken registers and (infinitely) many written registers.
It should be clear that it labels such as SPOKEN vs. WRITTEN, INFORMAL vs. FORMAL, DIALOGUE vs. MONOLOGUE, etc. are not sufficient to describe the texts in a corpus.
Summary corpus design

**sample**  almost all corpora are samples drawn from a larger population and almost always we are interested in the population rather than in just the sample – the way the sample is drawn determines whether and how and with which level of confidence we can extrapolate parameters we need to understand (and record) the parameters that influence the phenomena we are interested in

→ A **careful** and **well-documented** corpus design is crucial for most of our research questions.
Table of Contents

1. Corpus design and collection
   - Corpus design: samples, learner-specific issues
   - Hands-on session 1
     - Formats, architecture
     - Research questions, texts

2. Error Annotation

3. Other annotation layers

4. Quantitative analysis

5. Summary
Corpus design and corpus annotation depend primarily on the research question. **Corpus architecture and formats depend on general considerations** (availability, transparency, re-usability, reproducability).

Once we know which texts we want in the corpus we need to think about collection, formats and architectures.
When you collect texts (spoken or written) for your corpus, be careful to

- safely store the original data
- record all metadata (data about the learners/participants, task or setting, time, also information about pre-processing)
- get all necessary permissions from the participants
- decide on formats and architectures
We already saw that corpus design involves many decisions and that these decisions influence what we can do with a corpus. Many studies report only the results of a corpus study without giving us the data or the annotations. Why is that problematic?
We already saw that corpus design involves many decisions and that these decisions influence what we can do with a corpus. Many studies report only the results of a corpus study without giving us the data or the annotations.
Why is that problematic?

Because if we cannot see the data and its interpretation we cannot assess the validity of the results/conclusions of a study. → Tuesday, Wednesday
We already saw that corpus design involves many decisions and that these decisions influence what we can do with a corpus. Many studies report only the results of a corpus study without giving us the data or the annotations. Why is that problematic?

Because if we cannot see the data and its interpretation we cannot assess the validity of the results/conclusions of a study. → Tuesday, Wednesday

Because we might want to replicate the experiment. → Thursday
We already saw that corpus design involves many decisions and that these decisions influence what we can do with a corpus. Many studies report only the results of a corpus study without giving us the data or the annotations. Why is that problematic? Because if we cannot see the data and its interpretation we cannot assess the validity of the results/conclusions of a study. → Tuesday, Wednesday Because we might want to replicate the experiment. → Thursday Because we might want to keep and use the data (plus annotations) for a long time.
Transparency, availability, sustainability, ...

We already saw that corpus design involves many decisions and that these decisions influence what we can do with a corpus. Many studies report only the results of a corpus study without giving us the data or the annotations. Why is that problematic?

Because if we cannot see the data and its interpretation we cannot assess the validity of the results/conclusions of a study. → Tuesday, Wednesday
Because we might want to replicate the experiment. → Thursday
Because we might want to keep and use the data (plus annotations) for a long time.
Because we want to test our own (alternative) hypotheses on the same data or use the data for different research questions. But not on a piece of paper/in a Word file/in an Excel table/etc. because then we would again hide our interpretation.
We want a corpus that allows us to

- see the data and the interpretation (annotation)
- take the data and manipulate it
- store it so that it can be read in years to come
- further annotate the data on as many levels as one wants (and by as many people as are interested, independent of each other)
We want a corpus that allows us to

- see the data and the interpretation (annotation)
- take the data and manipulate it
- store it so that it can be read in years to come
- further annotate the data on as many levels as one wants (and by as many people as are interested, independent of each other)

This means that

- the corpus must be stored in well-understood sustainable formats that can be extended
- the corpus must be downloadable (not just available via an interface)
Sustainable formats are well documented, declarative, text-based (often XML). This does not mean that you have to type XML (yay!). You can use any number of tools for corpus creation and annotation. Each of these tools outputs a given format. In order to use different tools on the same data, we need a converter framework. There are several – the most powerful at the moment is SaltNPepper (http://corpus-tools.org/salt/, http://corpus-tools.org/pepper/)
Multi-layer architectures

The most powerful corpus architectures that are available today are multi-layer standoff architectures. The basic idea is as follows:
There is a layer of (tokenized) data.
The annotation layers are stored independently of the data – each tag points to one or many tokens.

The number of annotation layers is not limited – new annotation layers can be added at any point in time. Since the annotation layers are independent of each other, they do not conflict.

(There used to be inline architectures where the annotation was stored with the primary data. Not done anymore.)
On multi-layer architectures see Carletta et al. (2003), Ide et al. (2003), Dipper (2005), Wörner et al. (2006), Bański and Przepiórkowski (2009), Chiarcos et al. (2009), Zeldes et al. (2009), and Krause and Zeldes (2014), among many others. Krause and Zeldes (2014) introduces an even more powerful model which allows multiple tokenizations.
What are the legal issues involved in the publication of a (learner) corpus?

The participants must sign agreements that allows the corpus constructor to make the data available (anonymized, if necessary). (This can be difficult for children. It can also be problematic for spoken data.) The corpus creator should make the data available under a specified license. (Why is it not sufficient to just put your corpus on your homepage? If a user doesn’t know the terms she cannot be sure what she is allowed to do with the corpus. In essence, she cannot use it freely.) For corpora, the creative commons licenses (CC) are often used, https://creativecommons.org/. Under http://opensource.org/licenses you find the different licenses that are available for programs, scripts, etc. Consider using the most liberal license possible.
What are the legal issues involved in the publication of a (learner) corpus?

- The participants must sign agreements that allows the corpus constructor to make the data available (anonymized, if necessary).
  (This can be difficult for children. It can also be problematic for spoken data.)
What are the legal issues involved in the publication of a (learner) corpus?

- The participants must sign agreements that allow the corpus constructor to make the data available (anonymized, if necessary). (This can be difficult for children. It can also be problematic for spoken data.)

- The corpus creator should make the data available under a specified license. (Why is it not sufficient to just put your corpus on your homepage?)

For corpora, the creative commons licenses (CC) are often used, [https://creativecommons.org/](https://creativecommons.org/). Under [http://opensource.org/licenses](http://opensource.org/licenses) you find the different licenses that are available for programs, scripts, etc. Consider using the most liberal license possible.
What are the legal issues involved in the publication of a (learner) corpus?

- The participants must sign agreements that allows the corpus constructor to make the data available (anonymized, if necessary). (This can be difficult for children. It can also be problematic for spoken data.)

- The corpus creator should make the data available under a specified license. (Why is it not sufficient to just put your corpus on your homepage? If a user doesn’t know the terms she cannot be sure what she is allowed to do with the corpus. In essence, she cannot use it freely.)

For corpora, the creative commons licenses (CC) are often used, https://creativecommons.org/. Under http://opensource.org/licenses you find the different licenses that are available for programs, scripts, etc. Consider using the most liberal license possible.
What are your research questions?
We will be working with essays provided by Markus Dickinson (thank you, Markus!). Some metadata (country of origin, prompt) for each text is given in the header.
As for my goals and dreams, I'd like to say that I will contribute to the development of Korea's finance industry. In specific, I hope to found a professional finance academy which will provide the next generation with practical knowledge needed in the real field.
Markus:
How was this data collected?

How do we know the level of proficiency of the learner?

What do we know about the background of the learner?

Is the data freely available?

Is the data stored in sustainable formats?
Markus:
How was this data collected?
Intensive English Program administered exams at IU and offered usage to us. → opportunistic
How do we know the level of proficiency of the learner?

What do we know about the background of the learner?

Is the data freely available?

Is the data stored in sustainable formats?
Markus:
How was this data collected?
Intensive English Program administered exams at IU and offered usage to us. → opportunistic
How do we know the level of proficiency of the learner?
2-3 raters scored it, scores are averaged. Scoring process is documented.
What do we know about the background of the learner?

Is the data freely available?

Is the data stored in sustainable formats?
Markus:
How was this data collected?
Intensive English Program administered exams at IU and offered usage to us. → opportunistic
How do we know the level of proficiency of the learner?
2-3 raters scored it, scores are averaged. Scoring process is documented.
What do we know about the background of the learner?
Country.
Is the data freely available?

Is the data stored in sustainable formats?
Markus:
How was this data collected?
Intensive English Program administered exams at IU and offered usage to us. → opportunistic
How do we know the level of proficiency of the learner?
2-3 raters scored it, scores are averaged. Scoring process is documented.
What do we know about the background of the learner?
Country.
Is the data freely available?
No. It might be some day.
Is the data stored in sustainable formats?
Markus:
How was this data collected?  
Intensive English Program administered exams at IU and offered usage to us.  → opportunistic
How do we know the level of proficiency of the learner?  
2-3 raters scored it, scores are averaged. Scoring process is documented.
What do we know about the background of the learner?  
Country.
Is the data freely available?  
No. It might be some day.
Is the data stored in sustainable formats?  
Yes. Raw text, annotated files are available as txt, CoNLL, brat
Please go to https://box.hu-berlin.de/d/a6da03289e/.

What would/could you do with this text? What are your observations about the learner’s language?
Annotation means the assignment of a category to a corpus exponent (token, sequence of tokens).

The purpose of annotation is the categorization of similar or comparable instances of any phenomenon that is interesting for a given research question.

<mantra>
Corpus design and corpus annotation depend primarily on the research question. Corpus architecture and formats depend on general considerations (availability, transparency, re-usability, reproducability).
</mantra>
only some learner corpora are freely available

Tono 2000; Granger 2003; Rooy and Schäfer 2003; Izumi et al. 2005; Díaz-Negrillo and Fernández-Domínguez 2006; Dickinson and Ragheb 2009; Díaz-Negrillo et al. 2010; Ragheb and Dickinson 2011; Ballier and Martin 2015; Lüdeling and Hirschmann 2015; Rooy 2015, among many others
Annotation in learner corpora: Situation

- only some learner corpora are freely available
- many learner corpora are annotated with ‘generic’ error categories

Tono 2000; Granger 2003; Rooy and Schäfer 2003; Izumi et al. 2005; Díaz-Negrillo and Fernández-Domínguez 2006; Dickinson and Ragheb 2009; Díaz-Negrillo et al. 2010; Ragheb and Dickinson 2011; Ballier and Martin 2015; Lüdeling and Hirschmann 2015; Rooy 2015, among many others
Annotation in learner corpora: Situation

- only some learner corpora are freely available
- many learner corpora are annotated with ‘generic’ error categories
- only some learner corpora provide explicit target forms; only very few learner corpora provide several target hypotheses for a given utterance

Tono 2000; Granger 2003; Rooy and Schäfer 2003; Izumi et al. 2005; Díaz-Negrillo and Fernández-Domínguez 2006; Dickinson and Ragheb 2009; Díaz-Negrillo et al. 2010; Ragheb and Dickinson 2011; Ballier and Martin 2015; Lüdeling and Hirschmann 2015; Rooy 2015, among many others
Annotation in learner corpora: Situation

- only some learner corpora are freely available
- many learner corpora are annotated with ‘generic’ error categories
- only some learner corpora provide explicit target forms; only very few learner corpora provide several target hypotheses for a given utterance
- only very few learner corpora have further annotation layers (parts of speech, lemma, syntax, ...)

Tono 2000; Granger 2003; Rooy and Schäfer 2003; Izumi et al. 2005; Díaz-Negrillo and Fernández-Domínguez 2006; Dickinson and Ragheb 2009; Díaz-Negrillo et al. 2010; Ragheb and Dickinson 2011; Ballier and Martin 2015; Lüdeling and Hirschmann 2015; Rooy 2015, among many others
Annotation in learner corpora: Situation

- only some learner corpora are freely available
- many learner corpora are annotated with ‘generic’ error categories
- only some learner corpora provide explicit target forms; only very few learner corpora provide several target hypotheses for a given utterance
- only very few learner corpora have further annotation layers (parts of speech, lemma, syntax, ...)
- many learner corpora are not stored in architectures or provided in formats that allow further annotation through the researcher

Tono 2000; Granger 2003; Rooy and Schäfer 2003; Izumi et al. 2005; Díaz-Negrillo and Fernández-Domínguez 2006; Dickinson and Ragheb 2009; Díaz-Negrillo et al. 2010; Ragheb and Dickinson 2011; Ballier and Martin 2015; Lüdeling and Hirschmann 2015; Rooy 2015, among many others
Annotation in learner corpora: Situation

- only some learner corpora are freely available
- many learner corpora are annotated with ‘generic’ error categories
- only some learner corpora provide explicit target forms; only very few learner corpora provide several target hypotheses for a given utterance
- only very few learner corpora have further annotation layers (parts of speech, lemma, syntax, ...)
- many learner corpora are not stored in architectures or provided in formats that allow further annotation through the researcher
- for spoken learner corpora: there is often a division between corpora that are used for phonetic and speech purposes and those that are used to do research on properties of ‘spoken/oral’ registers

Tono 2000; Granger 2003; Rooy and Schäfer 2003; Izumi et al. 2005; Díaz-Negrillo and Fernández-Domínguez 2006; Dickinson and Ragheb 2009; Díaz-Negrillo et al. 2010; Ragheb and Dickinson 2011; Ballier and Martin 2015; Lüdeling and Hirschmann 2015; Rooy 2015, among many others
Annotation in learner corpora: Problems

Annotation of learner corpora is problematic and difficult.

- Conceptual problems: notion of ‘error’, comparative fallacy, applicability of grammatical models
- (related) Methodological problems: taggers, parsers, etc. do not work with high accuracy on data that differs from their training data (spelling, word order, patterns, ...) and are inconsistent

(1) learner: und man deshalb immer auf der flucht sein wird
lexical/spelling/(tagger): and one therefore always in the curse
syntactic/semantic: and one therefore always on the flight
‘and one would therefore always be on the run’

Corder 1981; Bley-Vroman 1983; Lennon 1991; Tenfjord et al. 2006; Granger 2008; Dickinson and Ragheb 2009; Ellis and Barkhuizen 2009; Díaz-Negrillo et al. 2010; Rehbein et al. 2012; Lüdeling and Hirschmann 2015, and many others
Table of Contents

1 Corpus design and collection

2 Error Annotation
   - Target hypotheses
   - Specific error tags
   - Hands-on session 2

3 Other annotation layers

4 Quantitative analysis

5 Summary
Target Hypotheses

**target hypothesis** Every annotation of a learner text (just like every annotation of any other text) involves interpretation of the data (necessary loss of information). The construction of a target hypothesis (TH) is influenced by the research question. Target hypotheses do not code ‘the truth’, or even the ‘correct’ utterance. They are a necessary prerequisite step for error annotation but also for any other type of annotation.

**error** Here I define ‘error’ as the difference between a learner utterance and an (implicit or explicit) target hypothesis. This is a technical definition, and the notion of error does not necessarily have anything to do with other error definitions.
In the following I will argue that

- target hypotheses should always be explicit. It should be possible to code several target hypotheses for the same learner utterance (and it is interesting to compare THs).
- it should be possible to freely add annotation layers to a given learner corpus.
Corpora

**Falko** is a deeply annotated corpus of written texts produced by advanced learners of German as a Foreign Language and comparable texts by native speakers (Lüdeling et al. 2008; Reznicek et al. 2010; Reznicek et al. 2013, https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko). The Falko-Essay corpus used here consists of essays covering four different topics, acquired in a controlled setting. Annotation layers include different target hypotheses, error annotation, part-of-speech annotation, and syntactic annotation (dependency syntax).

**BeMaTaC** is a small spoken learner corpus consisting of Map Task Dialogues by advanced learners of German and native speakers of German (http://u.hu-berlin.de/bematac). The corpus is transcribed, tokenized (aligned with the audio and video signal), annotated on many layers, among them parts of speech and disfluencies. A part of BeMaTaC-L1 is syntactically annotated.

Both corpora are freely available in a number of formats under CC licenses. They can be searched through the ANNIS search interface.
BeMaTaC, Instructor map and instructee map
TH: Explicitness

bevor man überhaupt anfangen kann, sich neues Wissen zu erlernen
before one even start can REFL new knowledge to learn
‘before one can even start to acquire new knowledge’
TH: Explicitness

bevor man überhaupt anfangen kann, sich neues Wissen zu erlernen
‘before one can even start to acquire new knowledge’

bevor man überhaupt anfangen kann, neues Wissen zu erlernen

bevor man überhaupt anfangen kann, sich neues Wissen anzueignen

bevor man überhaupt anfangen kann, neues Wissen zu erwerben

[Falko L2, cbs003-2007-10-L2v2.4]
bevor man überhaupt anfangen kann, sich neues Wissen zu erlernen

→ argument structure error

bevor man überhaupt anfangen kann, sich neues Wissen anzueignen

→ lexical error

bevor man überhaupt anfangen kann, neues Wissen zu erwerben

→ lexical error and argument structure error
Interpretation of ‘errors’

In spoken German, verb-final schwas are often omitted (Kohler and Rodgers 2001):

(ich) [laʊf] instead of (ich) [laʊfə] 'I walk',
(ich) [hap] instead of (ich) [ha:bə] 'I have', etc.

BeMaTaC has a narrow transcription (not entirely phonetic) as well as a 'standard' German written TH. Using these and the part-of-speech layer is possible to find all instances of omitted schwa in verbs.

<table>
<thead>
<tr>
<th>dipl</th>
<th>das</th>
<th>hab</th>
<th>ich</th>
<th>nicht</th>
<th>gesagt</th>
</tr>
</thead>
<tbody>
<tr>
<td>TH</td>
<td>das</td>
<td>habe</td>
<td>ich</td>
<td>nicht</td>
<td>gesagt</td>
</tr>
<tr>
<td>pos</td>
<td>PDS</td>
<td>VAFIN</td>
<td>PPER</td>
<td>PTKNEG</td>
<td>VVPP</td>
</tr>
<tr>
<td>gloss</td>
<td>that</td>
<td>have</td>
<td>I</td>
<td>not</td>
<td>said</td>
</tr>
</tbody>
</table>

See Lüdeling et al. 2015.
If the TH is seen as the ‘correct’ form, it would seem that native speakers make more ‘errors’ than learners. [side remark: Strangely, this is the way that learners and native speakers think about it.]

<table>
<thead>
<tr>
<th></th>
<th>$\emptyset$-forms (dipl)</th>
<th>-e-forms (norm)</th>
<th>% elided schwa</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>67</td>
<td>108</td>
<td>62</td>
</tr>
<tr>
<td>L2</td>
<td>44</td>
<td>107</td>
<td>41</td>
</tr>
</tbody>
</table>
If the TH is seen as the ‘correct’ form, it would seem that native speakers make more ‘errors’ than learners. [side remark: Strangely, this is the way that learners and native speakers think about it.]

<table>
<thead>
<tr>
<th></th>
<th>$\emptyset$-forms (dipl)</th>
<th>-e-forms (norm)</th>
<th>% elided schwa</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>67</td>
<td>108</td>
<td>62</td>
</tr>
<tr>
<td>L2</td>
<td>44</td>
<td>107</td>
<td>41</td>
</tr>
</tbody>
</table>

This is, of course, not the way we want to interpret this data. Schwa-elision is more natural and therefore should be acquired. 41% of elided schwas seem to suggest that learners are not too bad. If we look more closely at the verbs which are shortened the picture is different.
Interpretation of ‘errors’


Elision
- elided
- non-elided

Percentage

Construction
Interpretation of ‘errors’

Schwa elision does not only depend on the phonological properties of the verb. The learners mainly elide schwa with haben ’to have’ in the forms ich hab X, X hab ich ‘I have X’. They have learned a form but have not learned that schwa-elision is productive.
In Falko we have two target hypotheses that can be used to see different properties of learner language:

**TH1** codes mainly grammatical problems and is sentence-based. There are extensive guidelines, and the inter-annotator agreement is okay (still many problematic decisions).

**TH2** codes stylistic as well as grammatical problems and is text-based. There are guidelines but the inter-annotator agreement is a lot worse.

The target hypotheses are provided manually. There is then automatic lemmatisation, pos-tagging, and automatic distance-based error tagging on all levels. (In addition, there can be as many linguistically motivated error annotation layers as one needs - more on this later.)
Target hypotheses are equally necessary for native speaker data. [Falko data shows that native speakers produce significantly more orthography (punctuation) errors than learners.]

In Falko and BeMaTaC the L1 and L2 subcorpora are treated in the same way - comparisons on all layers are possible.
(2) Solche Literatur nützt den Interessen der Frauen, weil sie entdeckt eine Aspekte der Gesellschaft, die oft nicht bekannt ist.

[Falko Essay fk007-2006-07-L2v2.4]
(3) Solche Literatur nützt den Interessen der Frauen, weil sie entdeckt eine Aspekte der Gesellschaft, die oft nicht bekannt ist. which often not known is
weil sie entdeckt eine Aspekte der Gesellschaft

weil sie einen Aspekt der Gesellschaft darstellt

[Note: Tagging is done automatically with the TreeTagger Schmid 1994, using the STTS Tagset Schiller et al. 1999; one tagger error corrected here]
## TH and error tagging in Falko: Example

<table>
<thead>
<tr>
<th>LU:</th>
<th>weil</th>
<th>sie</th>
<th>entdeckt</th>
<th>eine</th>
<th>Aspekte</th>
<th>der</th>
<th>Gesellschaft</th>
</tr>
</thead>
<tbody>
<tr>
<td>posLU:</td>
<td>KOUS</td>
<td>PPER</td>
<td>VVFIN</td>
<td>ART</td>
<td>NN</td>
<td>ART</td>
<td>NN</td>
</tr>
<tr>
<td>lemmaLU</td>
<td>weil</td>
<td>sie</td>
<td>entdecken</td>
<td>ein</td>
<td>Aspekt</td>
<td>d</td>
<td>Gesellschaft</td>
</tr>
<tr>
<td>TH1:</td>
<td>weil</td>
<td>sie</td>
<td>einen</td>
<td>Aspekt</td>
<td>der</td>
<td>Gesellschaft</td>
<td>entdeckt</td>
</tr>
<tr>
<td>posTH1:</td>
<td>KOUS</td>
<td>PPER</td>
<td>ART</td>
<td>NN</td>
<td>ART</td>
<td>NN</td>
<td>VVFIN</td>
</tr>
<tr>
<td>lemmaTH1:</td>
<td>weil</td>
<td>sie</td>
<td>ein</td>
<td>Aspekt</td>
<td>d</td>
<td>Gesellschaft</td>
<td>MOVS</td>
</tr>
<tr>
<td>TH1diff:</td>
<td>MOV</td>
<td>CHA</td>
<td>CHA</td>
<td>CHA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TH1posdiff:</td>
<td>DEL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TH1lemmadiff:</td>
<td>MOV</td>
<td>CHA</td>
<td>CHA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Note: Tagging is done automatically with the TreeTagger Schmid 1994, using the STTS Tagset Schiller et al. 1999; one tagger error corrected here]
**TH and error tagging in Falko: Example**

<table>
<thead>
<tr>
<th>LU:</th>
<th>weil</th>
<th>sie</th>
<th>entdeckt</th>
<th>eine</th>
<th>Aspekte</th>
<th>der</th>
<th>Gesellschaft</th>
</tr>
</thead>
<tbody>
<tr>
<td>posLU:</td>
<td>KOUS</td>
<td>PPER</td>
<td>VVFIN</td>
<td>ART</td>
<td>NN</td>
<td>ART</td>
<td>NN</td>
</tr>
<tr>
<td>lemmaLU</td>
<td>weil</td>
<td>sie</td>
<td>entdecken</td>
<td>ein</td>
<td>Aspekt</td>
<td>d</td>
<td>Gesellschaft</td>
</tr>
<tr>
<td>TH1:</td>
<td>weil</td>
<td>sie</td>
<td>entdecken</td>
<td>einen</td>
<td>Aspekt</td>
<td>der</td>
<td>Gesellschaft</td>
</tr>
<tr>
<td>posTH1:</td>
<td>KOUS</td>
<td>PPER</td>
<td>ART</td>
<td>NN</td>
<td>ART</td>
<td>NN</td>
<td>Gesellschaft</td>
</tr>
<tr>
<td>lemmaTH1:</td>
<td>weil</td>
<td>sie</td>
<td>CHA</td>
<td>CHA</td>
<td>CHA</td>
<td>CHA</td>
<td>CHA</td>
</tr>
<tr>
<td>TH1diff:</td>
<td>MOV5</td>
<td>DEL</td>
<td>CHA</td>
<td>CHA</td>
<td>MOVT</td>
<td>INS</td>
<td>MOVT</td>
</tr>
<tr>
<td>TH1posdiff:</td>
<td>DEL</td>
<td>MOV5</td>
<td>CHA</td>
<td>CHA</td>
<td>MOVT</td>
<td>INS</td>
<td>MOVT</td>
</tr>
<tr>
<td>TH2:</td>
<td>weil</td>
<td>sie</td>
<td>entdecken</td>
<td>einen</td>
<td>Aspekt</td>
<td>der</td>
<td>Gesellschaft</td>
</tr>
<tr>
<td>posTH2:</td>
<td>KOUS</td>
<td>PPER</td>
<td>ART</td>
<td>NN</td>
<td>ART</td>
<td>NN</td>
<td>Gesellschaft</td>
</tr>
<tr>
<td>lemmaTH2:</td>
<td>weil</td>
<td>sie</td>
<td>CHA</td>
<td>CHA</td>
<td>CHA</td>
<td>CHA</td>
<td>CHA</td>
</tr>
<tr>
<td>TH2diff:</td>
<td>DEL</td>
<td>DEL</td>
<td>CHA</td>
<td>CHA</td>
<td>INS</td>
<td>INS</td>
<td>INS</td>
</tr>
<tr>
<td>TH2posdiff:</td>
<td>DEL</td>
<td>DEL</td>
<td>CHA</td>
<td>CHA</td>
<td>INS</td>
<td>INS</td>
<td>INS</td>
</tr>
<tr>
<td>TH2lemmadiff:</td>
<td>DEL</td>
<td>DEL</td>
<td>CHA</td>
<td>CHA</td>
<td>INS</td>
<td>INS</td>
<td>INS</td>
</tr>
</tbody>
</table>

[Note: Tagging is done automatically with the TreeTagger Schmid 1994, using the STTS Tagset Schiller et al. 1999; one tagger error corrected here]
Multiple target hypotheses in Falko

If the target hypotheses are assigned consistently, the combination of annotation layers lets us find specific error classes:

- MOVS/MOVT points to word order errors
- MOVS/MOVT with verbs points to verb order errors (very common in GFL learners)
- CHA in TH and pos but not in lemma point to grammatical errors
- MOVT in TH1lemmadiff and INS in TH2lemmadiff points to lexical errors
- ...

These error classes can then be investigated further. It is always possible to add error tags for specific linguistic errors.
Aside: Syntactic annotation

Syntactic annotation is done on TH1, and there is a mapping of differences between TH1 and the sentence in the original word order but with spelling and case errors corrected (TH0).
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Corpus design and collection</td>
</tr>
<tr>
<td>2</td>
<td>Error Annotation</td>
</tr>
<tr>
<td></td>
<td>• Target hypotheses</td>
</tr>
<tr>
<td></td>
<td>• Specific error tags</td>
</tr>
<tr>
<td></td>
<td>• Hands-on session 2</td>
</tr>
<tr>
<td>3</td>
<td>Other annotation layers</td>
</tr>
<tr>
<td>4</td>
<td>Quantitative analysis</td>
</tr>
<tr>
<td>5</td>
<td>Summary</td>
</tr>
</tbody>
</table>
Error tags

In addition to general error tags (be they linguistic or distance-based), for a given research question it might be necessary to categorize the data further. This is often done away from the corpus (the analysis itself is not available, only the results of a study are presented).

However, just like we cannot always agree on target hypotheses or general error tags we can also not agree on specific error tags. Making them explicit is equally necessary.
Study on the use and productivity of complex verbs in German.

Complex verbs might be formally difficult because they come in two types that behave differently. The formal properties are taught extensively in all textbooks.

<table>
<thead>
<tr>
<th>prefix verbs</th>
<th>particle verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>stehen – verstehen</td>
<td>stehen – aufstehen</td>
</tr>
<tr>
<td>”stand” , ”understand”</td>
<td>”stand”, ”get up”</td>
</tr>
<tr>
<td>dass ich das Problem verstehe</td>
<td>dass ich morgens aufstehe</td>
</tr>
<tr>
<td>ich verstehe das Problem</td>
<td>ich stehe morgens auf</td>
</tr>
<tr>
<td>zu verstehen</td>
<td>aufzustehen</td>
</tr>
<tr>
<td>verstanden</td>
<td>aufgestanden</td>
</tr>
</tbody>
</table>
The semantic specifics of complex verbs are not taught (morphological productivity is almost never mentioned in textbooks).

<table>
<thead>
<tr>
<th>prefix verbs</th>
<th>particle verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>entscheiden ”decide”, begreifen ”understand”, versuchen ”try, tempt”</td>
<td>anfangen ”begin”, aufhören ”stop”, zustimmen ”agree”</td>
</tr>
<tr>
<td>entkernen ”deseed”, entkalken ”decalcify”, entkleiden ”undress”, ...</td>
<td>anlesen ”begin reading, read partly”, andenken ”start thinking about”, andiskutieren ”start discussing”, ...</td>
</tr>
<tr>
<td>pattern: ent- + denominal verb: take away N</td>
<td>pattern: an + Verb: begin V, V partly</td>
</tr>
</tbody>
</table>

many lexicalized cases

many productive patterns
What needs to be acquired

Learners of L2 German have to acquire

- the formal differences between prefix verbs and particle verbs (especially the formal peculiarities of particle verbs)
- the fact that there are lexicalized cases and productive patterns and how to distinguish between the two for productive patterns: restrictions and degree of productivity

Teaching materials cover the formal properties but usually do not even mention patterns/productivity – these have to be acquired through the input.
Productivity

- productivity in acquisition: generalization from one form (pivot, item, teddy bear, construction, ...) to other forms in L1 or L2 acquisition, studied qualitatively and quantitatively in learner corpus data
- productivity in morphological (and recently also syntactic) theory (many different notions): the chance of a rule/process to produce more items, studied qualitatively and quantitatively in corpus data
- desideratum: combine different views and methods

here: vocabulary growth curves [many caveats apply: too little data, comparability issues, etc.]

Baayen 1992; Baayen 1993; Evert and Lüdeling 2001; Hay and Baayen 2003; MacWhinney 2004; Plag 2006; Baayen 2009; Berth 2009; Ellis and Ferreira-Junior 2009; Ellis and Simpson-Vlach 2009; Ellis 2012; Zeldes 2012, and many others
Complex verbs: Vocabulary growth curve
Complex verbs: Vocabulary growth curve
A better analysis is needed

The VGCs suggest that learners use complex verbs productively but slightly less so than native speakers. In order to understand the acquisition of complex verbs in L2 German (and perhaps more generally something about morphological productivity in acquisition) we need a more fine-grained analysis of formal (separability, orthography, inflection) and semantic properties.

- construction of a new target hypothesis only for complex verbs (based on TH1, without changing any other issues)
- addition of new annotation layers regarding form, tense, semantics, orthography, etc.; interpretation on all these layers (guidelines)
  - delimitation from resultative constructions, depictive constructions, etc. (Lüdeling 2001), delimitation from nontransparent forms
- new form: every form I did not know
- new use: every use I did not know
- ...

- quantitative analysis
We find only very few (six) formal errors that have to do with separation issues. All other form-based errors are independent of separation.

(4) nichts vorbereitet einen so gut auf das wirkliche Welt, wie nichts bereitet einen so gut auf die wirkliche Welt vor, wie ‘nothing prepares one for the real world as well as’

(5) sie suchen nach der Lösung und klagen die Universitäten an sie suchen nach der Lösung und anklagen die Universitäten an ‘they look for the solution and accuse the universities’
New forms:

(6) die darin bestehen, die Briefsysteme aufzudatieren
   ‘that consist in updating the letter systems’

(7) Aber in Situationen der körperlicher Arbeit wie auszubauern und in der Armee zu arbeiten
   ‘but in situations of physical work like ?? and working in the army’

(8) aber die Zeit verwachselt sich schnell
   ‘but time changes(?) fast’

New uses:

(9) kann man dann Erfahrung erreichen?
   ‘can one gain experience?’

(10) Die Studenten lesen eine Menge Bücher von Theorie etc., üben aber diese Theorie selten in Praxis aus.
    ‘the students read many books about theory but rarely put this theory into practice’
Particle verbs: Semantic errors

complex verb form errors

L1

as

L2

sem

neo

other
Productivity

Productivity: A more detailed study of the new forms and new uses shows some interesting patterns - the study needs to be extended to other data.

In this study (as in the very few other studies on morphological productivity and learners) we see

- learners use morphological patterns/rules productively, often more so than native speakers (since they know fewer words)
- learners do often not understand the restrictions and the degrees of productivity (including blocking)
Target hypotheses should be explicit because it is impossible *not* to interpret
because the error exponent depends on them
because the error class depends on them
because other annotation layers might depend on them

which means that the construction of the TH is the base for further analysis.

It should be possible to add multiple target hypotheses
because someone might disagree with my interpretation of the data
because (see above) they reflect the research question
because the comparison between target hypotheses might lead to interesting results.

It should be possible to add annotation layers at any point in time
because (again) someone might disagree with my interpretation of the data
because it helps me to understand the phenomenon better (annotation is research!)
Table of Contents

1. Corpus design and collection

2. Error Annotation
   - Target hypotheses
   - Specific error tags
   - Hands-on session 2
     - Annotation experiment
     - Annotation of target hypotheses

3. Other annotation layers

4. Quantitative analysis

5. Summary
Please go to
http://korpling.german.hu-berlin.de/~felix/indiana/ and
download the Table experiment.xls.
In the second column: Whenever you find a non-appropriate/wrong word, write down the correct form. If the word/phrase spans several tokens, please write down the exact form in each corresponding cell.

<table>
<thead>
<tr>
<th>girl</th>
<th>girlfriend</th>
</tr>
</thead>
<tbody>
<tr>
<td>friend</td>
<td>girlfriend</td>
</tr>
</tbody>
</table>

Please do not change the tokens/insert tokens/delete tokens in the first **column**. We want to evaluate this experiment automatically (using R) and the evaluation and need the exact tokens.
In the third column, please mark errors (again, if an error concerns several consecutive tokens, please mark it in each corresponding cell) using the following error tags:

<table>
<thead>
<tr>
<th>exponent</th>
<th>error</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>explre</td>
<td>SP</td>
<td><em>to explre different culture</em> (essay 173)</td>
</tr>
<tr>
<td>different</td>
<td>ART</td>
<td><em>to explre different culture</em> (mark missing article on the word before the relevant noun)</td>
</tr>
<tr>
<td>the</td>
<td>ART</td>
<td><em>I specialized on the biology</em> (TOEFL, GER, 416) (mark extra articles on the article)</td>
</tr>
<tr>
<td>don’t whether</td>
<td>WORDMISS</td>
<td><em>I don’t whether I can go or not.</em> (essay 148) (mark missing word error on both words surrounding the missing word)</td>
</tr>
<tr>
<td>childrens</td>
<td>NUM</td>
<td><em>taking care of childrens</em> (TOEFL, KOR, 2206) (mark number errors on the corresponding noun)</td>
</tr>
<tr>
<td>can heard</td>
<td>VV</td>
<td><em>over hundred thousands students can heard about that</em> (TOEFL, KOR, 138) (mark verb inflection errors on the corresponding verb(s))</td>
</tr>
<tr>
<td>with</td>
<td>PREP</td>
<td><em>If someone specialize in one specific things, he cannot talk with other things freely.</em> (TOEFL, KOR, 1204) (mark wrong prepositions on the preposition)</td>
</tr>
</tbody>
</table>

Ignore all other errors.
Please save your file with the annotations under any name and upload it: http://korpling.german.hu-berlin.de/~felix/indiana/

The filename will be changed in the upload.
We will look at the inter-annotator agreement tomorrow.
We pos-tagged and lemmatized the essays Markus gave us. Please download a tagged and lemmatized table of your essay from https://box.hu-berlin.de/d/17d1843bff/.

The Tagset (Penn Treebank) can be found here: https://corpling.uis.georgetown.edu/ptb_tags.html.

We used the TreeTagger (Schmid 1994), http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.
Do a quick first check. Are there pos-tagging problems?
Write a target hypothesis (minimal, only grammatical errors, sentence-based).
Table of Contents

1. Corpus design and collection

2. Error Annotation

3. Other annotation layers
   - Grammatical annotation
   - Hands-on session 3
   - Disfluency annotation
   - Variationist annotation

4. Quantitative analysis

5. Summary
Table of Contents

1. Corpus design and collection

2. Error Annotation

3. Other annotation layers
   - Grammatical annotation
   - Hands-on session 3
   - Disfluency annotation
   - Variationist annotation

4. Quantitative analysis

5. Summary
By grammatical annotation I mean annotation of categories that pertain to the structure of the language (part-of-speech, syntax, inflection, word formation etc.). All grammatical annotation of learner language is in danger of the comparative fallacy (the assignment of grammatical categories that are made for one linguistic system to another linguistic system, Bley-Vroman 1983).

This is not problematic if annotation categories are assigned carefully and consistently and if they are understood as classes of similar instances (where similarity is defined by the research question) which help us find relevant instances for our research question. It is, however, problematic if the categories are understood as meaning the same thing that they mean in the target language grammar. “As for annotation, to avoid the comparative fallacy is to provide annotation which does not overcompare a learner’s language system to either the L2 or the L1, thereby avoiding the comparative fallacy by not viewing learner production as a “degenerate form” of the target language. One such type of annotation would be a linguistic description of learner production applied to all the text, marking linguistic phenomena in all their occurrences, whether target like or non target like.” (Ragheb and Dickinson 2011, p. 117)
Grammatical annotation

By grammatical annotation I mean annotation of categories that pertain to the structure of the language (part-of-speech, syntax, inflection, word formation etc.). All grammatical annotation of learner language is in danger of the comparative fallacy (the assignment of grammatical categories that are made for one linguistic system to another linguistic system, Bley-Vroman 1983).

This is not problematic if annotation categories are assigned carefully and consistently and if they are understood as classes of similar instances (where similarity is defined by the research question) which help us find relevant instances for our research question.
By grammatical annotation I mean annotation of categories that pertain to the structure of the language (part-of-speech, syntax, inflection, word formation etc.). All grammatical annotation of learner language is in danger of the comparative fallacy (the assignment of grammatical categories that are made for one linguistic system to another linguistic system, Bley-Vroman 1983).

This is not problematic if annotation categories are assigned carefully and consistently and if they are understood as classes of similar instances (where similarity is defined by the research question) which help us find relevant instances for our research question.

It is, however, problematic if the categories are understood as meaning the same thing that they mean in the target language grammar.
Grammatical annotation

By grammatical annotation I mean annotation of categories that pertain to the structure of the language (part-of-speech, syntax, inflection, word formation etc.). All grammatical annotation of learner language is in danger of the comparative fallacy (the assignment of grammatical categories that are made for one linguistic system to another linguistic system, Bley-Vroman 1983).

This is not problematic if annotation categories are assigned carefully and consistently and if they are understood as classes of similar instances (where similarity is defined by the research question) which help us find relevant instances for our research question.

It is, however, problematic if the categories are understood as meaning the same thing that they mean in the target language grammar.

“As for annotation, to avoid the comparative fallacy is to provide annotation which does not overcompare a learner’s language system to either the L2 or the L1, thereby avoiding the comparative fallacy by not viewing learner production as a “degenerate form” of the target language. One such type of annotation would be a linguistic description of learner production applied to all the text, marking linguistic phenomena in all their occurrences, whether target like or non target like.” (Ragheb and Dickinson 2011, p. 117)
Why is grammatical annotation of learner language relevant?

- testing the robustness of parsers/tagger
- information extraction
- L1 identification
- the study of learner language properties (interlanguage, cf. Selinker 1972; Angelis and Selinker 2001)
Why is grammatical annotation of learner language relevant?

- testing the robustness of parsers/tagger
- information extraction
- L1 identification
- the study of learner language properties (interlanguage, cf. Selinker 1972; Angelis and Selinker 2001)

Each research question requires a different approach. I will argue/show that if you want to study interlanguage/acquisition processes

- grammatical annotation is often (always?) based on an implicit target hypothesis – and just like with error tags it would be best to make the TH explicit
- it is easier to understand learner language if tags describe one type of information
The most widespread annotation for corpora is part-of-speech annotation (cf. Stefanie’s lecture yesterday). A word class (pos) is assigned to each token in a corpus. We heard yesterday that there are many different tagsets and each codes different information. Pos tagging is problematic/difficult for native language because

- the original definitions (and definitions based on these) often mix lexical, syntactic, morphological information
- many tagsets mix pos information and inflectional information

The AMALGAM project compared different pos tagsets for English, see http://www.scs.leeds.ac.uk/amalgam/amalgam/amalghome.htm.
Today there are a lot of international issues we have to face.

Example from an essay collected at IU in a summer course (provided by Markus), tagged with Treetagger (Schmid 1994), http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/, using the extended Penn-Treebank tagset, see https://corpling.uis.georgetown.edu/ptb_tags.html
<table>
<thead>
<tr>
<th>lu</th>
<th>pos</th>
<th>lemma</th>
</tr>
</thead>
<tbody>
<tr>
<td>Todays</td>
<td>NNS</td>
<td>today</td>
</tr>
<tr>
<td>,</td>
<td>,</td>
<td>,</td>
</tr>
<tr>
<td>there</td>
<td>EX</td>
<td>there</td>
</tr>
<tr>
<td>are</td>
<td>VBP</td>
<td>be</td>
</tr>
<tr>
<td>a</td>
<td>DT</td>
<td>a</td>
</tr>
<tr>
<td>lot</td>
<td>NN</td>
<td>lot</td>
</tr>
<tr>
<td>international</td>
<td>JJ</td>
<td>international</td>
</tr>
<tr>
<td>issues</td>
<td>NNS</td>
<td>issue</td>
</tr>
<tr>
<td>we</td>
<td>PP</td>
<td>we</td>
</tr>
<tr>
<td>have</td>
<td>VHP</td>
<td>have</td>
</tr>
<tr>
<td>to</td>
<td>TO</td>
<td>to</td>
</tr>
<tr>
<td>face</td>
<td>VV</td>
<td>face</td>
</tr>
</tbody>
</table>

An example from an essay collected at IU in a summer course (provided by Markus), tagged with Treetagger (Schmid 1994), [http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/](http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/), using the extended Penn Treebank tagset, see [https://corpling.uis.georgetown.edu/ptb_tags.html](https://corpling.uis.georgetown.edu/ptb_tags.html).
All tags in the example are assigned correctly – this hides the problems that the learner has.
Part-of-speech annotation

All tags in the example are assigned correctly – this hides the problems that the learner has.

For many research questions, this might be exactly what we want. If we want to study interlanguage, however, we cannot find the interesting phenomena. A TH (and DIFF tags, as shown yesterday) might help.

<table>
<thead>
<tr>
<th>lu</th>
<th>pos</th>
<th>lemma</th>
<th>th</th>
<th>thpos</th>
<th>thlemma</th>
</tr>
</thead>
<tbody>
<tr>
<td>Todays</td>
<td>NNS</td>
<td>today</td>
<td>today</td>
<td>NN</td>
<td>today</td>
</tr>
<tr>
<td>,</td>
<td>,</td>
<td>,</td>
<td>,</td>
<td>,</td>
<td>,</td>
</tr>
<tr>
<td>there</td>
<td>EX</td>
<td>there</td>
<td>there</td>
<td>EX</td>
<td>there</td>
</tr>
<tr>
<td>are</td>
<td>VBP</td>
<td>be</td>
<td>are</td>
<td>VBP</td>
<td>be</td>
</tr>
<tr>
<td>a</td>
<td>DT</td>
<td>a</td>
<td>a</td>
<td>DT</td>
<td>a</td>
</tr>
<tr>
<td>lot</td>
<td>NN</td>
<td>lot</td>
<td>lot</td>
<td>NN</td>
<td>lot</td>
</tr>
<tr>
<td>of</td>
<td>IN</td>
<td>of</td>
<td>of</td>
<td>IN</td>
<td>of</td>
</tr>
<tr>
<td>issues</td>
<td>NNS</td>
<td>issue</td>
<td>issues</td>
<td>NNS</td>
<td>issue</td>
</tr>
<tr>
<td>we</td>
<td>PP</td>
<td>we</td>
<td>we</td>
<td>PP</td>
<td>we</td>
</tr>
<tr>
<td>have</td>
<td>VHP</td>
<td>have</td>
<td>have</td>
<td>VHP</td>
<td>have</td>
</tr>
<tr>
<td>to</td>
<td>TO</td>
<td>to</td>
<td>to</td>
<td>TO</td>
<td>to</td>
</tr>
<tr>
<td>face</td>
<td>VV</td>
<td>face</td>
<td>face</td>
<td>VV</td>
<td>face</td>
</tr>
</tbody>
</table>
Part-of-speech annotation

In the example above we (and the parser) knew which tags to assign for each token. Many learner utterances contain tokens for which this is not as clear. Díaz-Negrillo et al. (2010) argue that it is helpful to tease apart the different aspects hidden in many pos-tag definitions – they distinguish between stem (lexical knowledge), inflectional morphology, and distribution (syntactic knowledge). This more fine-grained analysis allows a better understanding of which hypotheses the learner may have about the target language.

internet have some “pages” that contents something so horrible (Ex 11)
mismatch between stem (noun) and distribution/morphology (verb)

for almost every jobs nowadays (Ex 13)
mismatch between inflection (plural) and distribution (singular)

Examples are from Díaz-Negrillo et al. (2010).
The notion part of speech (pars orationis) goes back to Dionysius Thrax (ca. 100 BC). Thrax distinguished between eight parts of speech. His definitions were already “mixed” definitions. This has been discussed many times; some authors suggest different tagsets according to each level, cf. Sütterlin (1923), Robins (1986), Knobloch and Schaeder (2000), and Haspelmath (2008)
Syntactic annotation builds on pos annotation. There are many different theories, schemes, and parsers. Some syntactic annotation is done manually, some automatically.

There are only very few learner corpora that are syntactically annotated.

On syntactic annotation of learner language, see Dickinson and Ragheb (2009), Rosén and Smedt (2010), Rehbein et al. (2012), Hahn and Meurers (2014), and Ragheb and Dickinson (2014). See Lu (2011) for a survey of syntactic complexity measures for learner English. There is a large amount of work in computational linguistics on the robustness of parsers.
What would the syntactic structure of the following sentences be?

The age changed, and we should change to know broad things and start with get broad knowledge.

First of all, various knowledge make people can talk with others freely and flexible.

They are worried about live the moments, no matters the consecuences of their acts and the family is only the group of persons that wait for they at home, and give them money.

The cost the product is determined by the materials that was used in the build, of the instruments uses in the company.

so old man is power fainally, It become a money.

The examples are taken from Essays written for the TOEFL text (TOEFL11 corpus, https://catalog.ldc.upenn.edu/LDC2014T06). KOR is for L1 Korean, SPA is for L1 Spanish.
What would the syntactic structure of the following sentences be?

_The age changed, and we should change to know broad things and start with get broad knowledge._ (TOEFL, KOR, 1204)
What would the syntactic structure of the following sentences be?

*The age changed, and we should change to know broad things and start with get broad knowledge.* (TOEFL, KOR, 1204)

*First of all, various knowledge make people can talk with others freely and flexible.* (TOEFL, KOR, 1204)

The examples are taken from Essays written for the TOEFL text (TOEFL11 corpus, https://catalog.ldc.upenn.edu/LDC2014T06). KOR is for L1 Korean, SPA is for L1 Spanish.
Syntactic annotation | examples | What should the syntactic structure be?

What would the syntactic structure of the following sentences be?

The age changed, and we should change to know broad things and start with get broad knowledge. (TOEFL, KOR, 1204)

First of all, various knowledge make people can talk with others freely and flexible. (TOEFL, KOR, 1204)

They are worried about live the moments, no matters the consequences of their acts and the family is only the group of persons that wait for they at home, and give them money. (TOEFL, SPA 1973)

The examples are taken from Essays written for the TOEFL text (TOEFL11 corpus, https://catalog.ldc.upenn.edu/LDC2014T06). KOR is for L1 Korean, SPA is for L1 Spanish.
What would the syntactic structure of the following sentences be?

*The age changed, and we should change to know broad things and start with get broad knowledge.* (TOEFL, KOR, 1204)

*First of all, various knowledge make people can talk with others freely and flexible.* (TOEFL, KOR, 1204)

*They are worried about live the moments, no matters the consequences of their acts and the family is only the group of persons that wait for they at home, and give them money.* (TOEFL, SPA 1973)

*The cost the product is determined by the materials that was used in the build, of the instruments uses in the company.* (TOEFL, SPA, 5666)

The examples are taken from Essays written for the TOEFL text (TOEFL11 corpus, https://catalog.ldc.upenn.edu/LDC2014T06). KOR is for L1 Korean, SPA is for L1 Spanish.
What would the syntactic structure of the following sentences be?

The age changed, and we should change to know broad things and start with get broad knowledge. (TOEFL, KOR, 1204)

First of all, various knowledge make people can talk with others freely and flexible. (TOEFL, KOR, 1204)

They are worried about live the moments, no matters the consequences of their acts and the family is only the group of persons that wait for they at home, and give them money. (TOEFL, SPA 1973)

The cost the product is determined by the materials that was used in the build, of the instruments uses in the company. (TOEFL, SPA, 5666)

so old man is power fainally, It become a money. (TOEF, KOR, 11)

---

The examples are taken from Essays written for the TOEFL text (TOEFL11 corpus, https://catalog.ldc.upenn.edu/LDC2014T06). KOR is for L1 Korean, SPA is for L1 Spanish.
Example of a parsed sentence from Ragheb and Dickinson (2014)

**Figure:** Example annotation using SALLE scheme
Syntactic annotation

What should the syntactic structure of a problematic utterance be? How do we model learner syntax (in all its variations)?
What should the parser do?

Depends on the research question.

If the robustness of the parser is to be tested or if the syntactic annotation is to be used for NLP purposes (L1 identification, information extraction, etc.) the parser should probably provide a plausible (whatever that may mean) parse. Note that this parse is based on an interpretation (implicit TH) of the learner utterance. If the parser is to be evaluated against a gold standard – what would that be?

In a study of interlanguage/acquisition processes it might again be interesting to see where the learner utterance differs from a “canonical” sentence (where canonical sentence means: a sentence that can be described by the target language grammar).
What should the syntactic structure of a problematic utterance be? How do we model learner syntax (in all its variations)?

What should the parser do?

Depends on the research question. 😊

If the robustness of the parser is to be tested or if the syntactic annotation is to be used for NLP purposes (L1 identification, information extraction, etc.) the parser should probably provide a plausible (whatever that may mean) parse. Note that this parse is based on an interpretation (implicit TH) of the learner utterance. If the parser is to be evaluated against a gold standard – what would that be?

In a study of interlanguage/acquisition processes it might again be interesting to see where the learner utterance differs from a “canonical” sentence (where canonical sentence means: a sentence that can be described by the target language grammar).
Research question: How do learners of GFL acquire modification?
Corpus: Falko

It is well known from the literature on GFL that learners have problems acquiring/using (certain) adverbs and particles. In an overuse/underuse table of pos-bigrams we see that ADV-ADV combinations are underused by GFL learners, independent of their L1 background.

Excel plugin for visualizing overuse/underuse written by Amir Zeldes. Blue cells indicate underuse, red cells indicate overuse, the intensity of the color shows the strength of the overuse/underuse.
Previous studies on adverbs in learner German are all form based (certain adverbs (forms) are analyzed). But is this the whole story? Are the effects form based or function based?

- Are all adverbs underused?

- Are certain adverbs (forms) underused?

- Are certain adverbs (forms) underused in certain functions?
- Are certain adverbial functions underused?
- Is modification generally underused (or do learners make up for the underuse of adverbs by other means of modification)?
Previous studies on adverbs in learner German are all form based (certain adverbs (forms) are analyzed). But is this the whole story? Are the effects form based or function based?

- Are all adverbs underused? No. (Some forms like *auch* “too, as well”, *noch* “still, even” are overused)
- Are certain adverbs (forms) underused?
  - Are certain adverbs (forms) underused in certain functions?
  - Are certain adverbial functions underused?
  - Is modification generally underused (or do learners make up for the underuse of adverbs by other means of modification)?
Previous studies on adverbs in learner German are all form based (certain adverbs (forms) are analyzed). But is this the whole story? Are the effects form based or function based?

- Are all adverbs underused? No. (Some forms like *auch* “too, as well”, *noch* “still, even” are overused)

- Are certain adverbs (forms) underused? Yes. The pos tag ADV is not fine-grained enough because ADVs show different distributions and have different functions (Hirschmann 2015).

- Are certain adverbs (forms) underused in certain functions?
- Are certain adverbial functions underused?
- Is modification generally underused (or do learners make up for the underuse of adverbs by other means of modification)?
The ADV classes are used by learners in different ways.

PTK: particles *(sehr gut - very good)*

ADVV: modal adverbs *(Bald schneit es – Soon it will snow)*

ADVS: sentence adverbs *(Bestimmt schneit es bald – Certainly, it will snow soon)*

PTKM: modal particles *(Es schneit *wohl* gerade – It is *apparently* snowing now)*
Using syntactic annotation, we can explore this in more detail:

- Is the syntactic function 'modification' underused?
- What is the target of the modification?
- What are the categories used for modification?
The analysis shows that

- the syntactic function MO is underused by all groups of learners.
- all syntactic relations possible for modification are underused; modifiers of adverbs show the strongest underuse
- categories of different complexity (lexemes to sentences) are used for modification; some categories are underused by the learners, two categories are slightly overused; adverbs and (adverbially used) adjectives show the strongest underuse
Are the effects form based or function based?

- Are all adverbs underused? No.
- Are certain adverbs (forms) underused? Yes. The pos tag ADV is not fine-grained enough because ADVs show different distributions and have different functions (Hirschmann 2015).
- Are certain adverbs (forms) underused in certain functions? Yes.
- Are certain adverbial functions underused? Yes.
- Is modification generally underused (or do learners make up for the underuse of adverbs by other means of modification)? Yes.
This case study shows that modification is a difficult category for learners of GFL. Whereas the previous evidence was form based we can now show that the function modification is underused (independent of the L1 of the learners). This means that previous hypotheses that refer to transfer phenomena or the “polyfunctionality” of certain elements cannot be sufficient.

Methodologically, this study shows that

- it is useful to separate form and function in annotation
- parsing the data (here: the target hypotheses) can help find the interesting cases
- sometimes the pos tags in a given tagset are not fine-grained enough to describe a phenomenon; in a multi-layer architecture it is possible to add new pos-tag layers
Table of Contents

1 Corpus design and collection

2 Error Annotation

3 Other annotation layers
   - Grammatical annotation
   - Hands-on session 3
     - Evaluation annotation experiment
   - Disfluency annotation
   - Variationist annotation

4 Quantitative analysis

5 Summary
Parts of speech

Onoma (noun): a part of speech inflected for case, signifying a person or thing
Rhema (verb): a part of speech without case inflection, but inflected for tense, person, number, signifying an activity or process performed or undone
Metoche (participle): a part of speech sharing the features of the verb and the noun
Arthron (article): a part of speech inflected for case and preposed or postposed to nouns
Antonymia (pronoun): a part of speech substitutable for a noun and marked for person
Prothesis (preposition): a part of speech placed before other words in composition and in syntax
Epirrhema (adverb): a part of speech without inflection, in modification of or in addition to a verb
Syndesmos (conjunction): a part of speech binding together the discourse and filling gaps in its interpretation
Thank you to Felix Golcher for writing the evaluation program (in R) and helping with the evaluation.
Inter-Annotator Agreement (IAA; sometimes also Inter-Rater Reliability, Inter-Rater Agreement) quantifies how well the annotators agree. There are several measures (e.g. Cohen’s Kappa, siehe Carletta 1996; Artstein and Poesio 2008). In order to assess IAA we have to ask:

- Did we annotate the same exponents?
  (problem: false negatives!)
- Did we assign the same categories?

In our experiment we worked on one annotation layer with a fixed tagset (errors) and one annotation layer with an open tagset (lexical). IAA is difficult to calculate for open tagsets (but see Bollmann et al. 2016).
To quantify the agreement of (for now 2) annotators:

\( \kappa = 1 \): Both annotators agree on everything.

\( \kappa = 0 \): There are only random agreements.

- Definition (Cohen 1960):
  \[
  \kappa = \frac{\text{relative agreement} - \text{random baseline}}{1 - \text{random baseline}}
  \]

- The measure was extended to many raters in Fleiss (1971)
- Implemented by Gamer et al. (2012) in R-package irr.
The data was uploaded via a script, converted to a text-based format (batch) and evaluated with R (R Core Team 2014)
  - programmable (scripts)
  - platform independent
  - free ...
This evaluation helps us to find the interesting places in the text (places where we disagree):

LU: 

Actually, I have girlfriend with whom I have planned bright future.
Right now, I have a girlfriend with whom I have planned a bright future.

LU: So we have always talked about our marriage life, which I and girlfriend would have two babies in our sweet house, kiss, and hug each other everyday and talk about our normal days.
So we have always talked about our married life, (where/in which) I and my girlfriend would have two babies in our sweet home. We will kiss, and hug each other every day and talk about our (regular working day, day, life).

This is due to the fact that I did not specify the guidelines – there was no well-understood common research question.
We disagreed in many cases. However, there were a few instances where we at least agreed on the exponent. There is no case where all eight analyzers agreed on the exponent and the tag.

Again, the guidelines should have been better. For cases with a finite tagset, it would also help to have a tool that controls the options (ART WORDMISS | WORDMISS ART).
because we so often disagree, it is necessary to make all annotations explicit

annotation is an iterative process (tagset and guidelines $\rightarrow$ annotation $\rightarrow$ evaluation $\rightarrow$ improved tagset and guidelines $\rightarrow$ annotation $\rightarrow$ evaluation $\rightarrow$ .... until evaluation is satisfactory). **Annotation is research!**
Now we need to formulate the target hypothesis for our texts. We will do this in the Excel files that were provided. Since only one person works on each text, it is possible to introduce empty lines when you want to add an element. You can also merge cells in the annotation column.

Background note: Excel is no annotation tool. However, many people have it and it works well for token and span annotation. Should we use Excel as an annotation tool?
Now we need to formulate the target hypothesis for our texts. We will do this in the Excel files that were provided. Since only one person works on each text, it is possible to introduce empty lines when you want to add an element. You can also merge cells in the annotation column.

Background note: Excel is no annotation tool. However, many people have it and it works well for token and span annotation. Should we use Excel as an annotation tool?

Hm. Depends. Some people do not have access to it. It is always better to use free software than proprietary software. The biggest problem is the output format - xls is problematic, can be changed without notice, etc. We actually use Excel sometimes in teaching. But we always convert the data to sustainable and well described formats (using SaltNPepper).
1 Coropus design and collection

2 Error Annotation

3 Other annotation layers
   - Grammatical annotation
   - Hands-on session 3
   - Disfluency annotation
   - Variationist annotation

4 Quantitative analysis

5 Summary
(Spontaneous, informal) spoken language is full of disfluencies. Fox Tree (1995) defines disfluencies as phenomena that do not add propositional content to an utterance.

Consider the following example from the Trinity Lancaster Corpus (http://cass.lancs.ac.uk/?page_id=1327) that consists of spoken texts collected in assessment situations (by Trinity College London). The corpus is not yet freely available – Dana Gablasova gave me the excerpt (thanks!). Words in angle brackets indicate that the transcriber was unsure of what she heard.
S: okay er my topic is mental attitude erm <.> it is known that mental attitude is er mental co= erm <classed as> a mental state involving beliefs er feelings values and dispositions erm to act in certain ways erm in a general way mental attitudes can be divided into two classes positive and negative much of your behaviours depends on your attitudes if your attitudes are negative erm it is er if your attitudes are negative you er you can expect to be vulnerable to addictions and er <sub somatic> er <set somatic> er <.> expects to er I’m sorry <set somatic> disorders and that erm may <.> cause a a bad influence s= erm for e-every ar= every area of your life however er if your at if you have a positive attitude can lead you to erm <.> brighter can bring you er erm <.> oh <.> erm <.> can can can give you confidence boost your self confidence and erm erm <.> erm <.> I’m sorry too nervous can I
E: you can use your points to help you that’s why you have a plan
S: oh sorry mm a positive attitude can er improve your health and er er <.> booster your self confidence and contribute to your success in pro= social and professional settings in addition to improving your outlook on life and erm a positive attitude can <.>
Types of disfluencies

There are many different types of disfluencies

- unfilled pauses
- filled pauses
- repairs
- repetitions
- truncations

In the Trinity Lancaster Corpus filled pauses are represented either as *er* or *erm*. This already involves a decision – phonetically the filled pauses differ from each other in many respects.


Transcription of spoken data and transliteration of handwritten data involve numerous decisions that should be documented.
Types of disfluencies

There are many different types of disfluencies

- unfilled pauses
- filled pauses
- repairs
- repetitions
- truncations

In the Trinity Lancaster Corpus filled pauses are represented either as *er* or *erm*. This already involves a decision – phonetically the filled pauses differ from each other in many respects.

Transcription of spoken data and transliteration of handwritten data involve numerous decisions that should be documented.
S: okay er my topic is mental attitude erm <.> it is known that mental attitude is er mental co= erm <classed as> a mental state involving beliefs er feelings values and dispositionserm to act in certain ways erm in a general way mental attitudes can be divided into two classes positive and negative much of your behaviours depends on your attitudes if your attitudes are negative erm it is er if your attitudes are negative you er you can expect to be vulnerable to addictions and er <sub somatic> er <set somatic> er <.> expects to er I’m sorry <set somatic> disorders and that erm may <.> cause a a bad influence s= erm for e-every ar= every area of your life however er if your at if you have a positive attitude can lead you to erm <.> brighter can bring you er erm <.> oh <.> erm <.> can can can give you confidence boost your self confidence and erm erm <.> erm <.> I’m sorry too nervous can I
E: you can use your points to help you that’s why you have a plan
S: oh sorry mm a positive attitude can er improve your health and er er <.> booster your self confidence and contribute to your success in pro= social and professional settings in addition to improving your outlook on life and erm a positive attitude can <.>
Why are disfluencies interesting?

- disfluencies are frequent
- disfluencies are multifunctional: they may indicate planning, they may signal turn taking, prepare the hearer for surprise, mark social status, etc.

Whereas disfluencies seem natural (and are often unnoticed) in spoken L1, they tend to be noticed more in learner language, and are sometimes seen as signs of “non-fluency”.
Case study filled pauses and repairs

Research question: How do advanced learners of German and native speakers in their disfluency production?

Corpus: BeMaTaC

This is joint work with Malte Belz, Simon Sauer and Christine Mooshammer. Here I will only speak about a few aspects of this study. For results and further aspects, see Belz et al. (2015) and Belz et al. (submitted).
Frequency: In a corpus study of 50 interviews with advanced German learners of English, 84% use significantly more filled pauses than native speakers (Brand and Götz 2013)

Form: äh [ɛː] in German, euh [œː] in L2 German with a background of L1 French (Gick et al. 2004)

Position: Strong inclination of learners to transfer pause profiles from L1 to L2 (Raupach 1980, for German)
Repairs

Here, we will look specifically at repairs.
Repair scheme: Reparandum, interregnum, reparans (Shriberg 1994)

Show flights \([from Boston on]_{RD} [uh]_{IR} [from Denver on]_{RS} Monday\)

Motivation

- Speech production is less automatic in L2 than in L1 (Wiese 1984; Kormos 1999)
- Repairs reflect automatization and monitoring processes (Bygate 1996)
- Repairs indicate that something needs to be corrected (Eklund 2004)
  → insight into how L1 and L2 choose to repair

Which repair information should be represented? At the moment, we have no good (operationalizable) theory for the functions of repairs/repair types. We can only reliably talk about form features.
Repair annotation is done on several layers. Again, it is useful to tease apart the different types of information.

- Reparanda (*rd*), interregna (*ir*) and reparantia (*rs*) on *instructor_repair*
- Substitutions (*s*), repetitions (*r*) and insertions (*i*) on *instructor_subrep*
## Repair annotation in BeMaTaC

<table>
<thead>
<tr>
<th>Repair type</th>
<th>Reparandum</th>
<th>Interregnum</th>
<th>Reparans</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fresh start</td>
<td><em>du hast</em> ‘you have’</td>
<td></td>
<td><em>links kannst doch gar nicht mehr vorbei</em> ‘you cannot pass left by any means’</td>
</tr>
<tr>
<td>Insertion</td>
<td><em>nach 0.8s links</em> ‘to the left’</td>
<td>0.5 s</td>
<td><em>waagerecht nach links</em> ‘horizontally to the left’</td>
</tr>
<tr>
<td>Substitution</td>
<td><em>dem</em> ‘the.DAT.MASC’</td>
<td>0.2 s äh</td>
<td><em>der Sanduhr</em> ‘the.DAT.FEM hourglass.DAT.FEM’</td>
</tr>
<tr>
<td>Repetition</td>
<td>*des Kai/‘of the Kai/’</td>
<td></td>
<td><em>des Kaiserbrötchens</em> ‘of the Kaiser roll’</td>
</tr>
</tbody>
</table>
Annotation of disfluencies

Using the different annotation layers for disfluencies (there are more), we can answer questions about frequency and position (in an utterance, in front of certain words, etc.) of disfluencies. We can also answer questions about those form attributes that are in the data. In order to answer questions about the phonetic properties, we would have to do a signal analysis. In order to answer questions about transfer, we would also have to do a study on disfluencies in the learner’s L1.

There are many phenomena for which we cannot use a given tagset but have to develop one (based on our research question and the theoretical background we want to work in). In multi-layer architectures, we can “play with” different types of information until we have found the representation we want.
Table of Contents

1 Corpus design and collection

2 Error Annotation

3 Other annotation layers
   - Grammatical annotation
   - Hands-on session 3
   - Disfluency annotation
   - Variationist annotation

4 Quantitative analysis

5 Summary
Variationism

We have seen that sometimes it is not the form (a word) that we want to study but rather the function of something (remember the modification study).

Consider a variable ("the same") and (all) its variants.

Sometimes it is possible to enumerate the variants of a variable, sometimes it isn’t.
A basic assumption is that the speaker can (in principle) choose between the variants.
We can then describe the distribution of the variants in different varieties.
Co-varying variables

The choice of a given variant is not arbitrary. Variationist research is concerned with describing influencing factors (co-varying variables), quantifying their impact and explaining the situation in a model.

- relationship between speaker and hearer
- mode of production
- task
- level of proficiency
- L1 of the speaker

...
The choice of a given variant is not arbitrary. Variationist research is concerned with describing influencing factors (co-varying variables), quantifying their impact and explaining the situation in a model.
Variationist annotation is difficult and time consuming because we have to interpret the data on a functional level (what did somebody want to do?) rather than on a formal level (what do we see?). There are a variationist studies on learner language (e.g. Gablasova et al. 2016 on stance in L2 English) but often the functions are not annotated in the data and we do not know how exactly a function was assigned.

Another problem: A seemingly variationist comparison of categories that are not functionally equivalent.

On the definition of variables (and all the problems associated with it) see e.g. (Lavandera 1978; Labov 1978; Croft 2010).
**Example: Overuse/underuse studies**

<table>
<thead>
<tr>
<th>lemma</th>
<th>tot_norm</th>
<th>de</th>
<th>da</th>
<th>en</th>
<th>fr</th>
<th>pl</th>
<th>ru</th>
</tr>
</thead>
<tbody>
<tr>
<td>in</td>
<td>0.013188</td>
<td>0.012261</td>
<td>0.014041</td>
<td>0.014247</td>
<td>0.015272</td>
<td>0.012135</td>
<td>0.009534</td>
</tr>
<tr>
<td>es</td>
<td>0.010897</td>
<td>0.011945</td>
<td>0.010900</td>
<td>0.011379</td>
<td>0.013347</td>
<td>0.008163</td>
<td>0.012385</td>
</tr>
<tr>
<td>sie</td>
<td>0.010618</td>
<td>0.008193</td>
<td>0.010643</td>
<td>0.008835</td>
<td>0.010909</td>
<td>0.006067</td>
<td>0.005613</td>
</tr>
<tr>
<td>man</td>
<td>0.010164</td>
<td>0.007900</td>
<td>0.012438</td>
<td>0.008742</td>
<td>0.009754</td>
<td>0.006950</td>
<td>0.007306</td>
</tr>
<tr>
<td>dass</td>
<td>0.009522</td>
<td>0.007404</td>
<td>0.012823</td>
<td>0.008789</td>
<td>0.009625</td>
<td>0.008880</td>
<td>0.009890</td>
</tr>
<tr>
<td>von</td>
<td>0.007982</td>
<td>0.007122</td>
<td>0.007309</td>
<td>0.006846</td>
<td>0.007315</td>
<td>0.010259</td>
<td>0.007930</td>
</tr>
<tr>
<td>auch</td>
<td>0.007028</td>
<td>0.008362</td>
<td>0.008527</td>
<td>0.005828</td>
<td>0.005775</td>
<td>0.005461</td>
<td>0.004455</td>
</tr>
<tr>
<td>für</td>
<td>0.006683</td>
<td>0.007201</td>
<td>0.006091</td>
<td>0.007216</td>
<td>0.006802</td>
<td>0.005736</td>
<td>0.004188</td>
</tr>
<tr>
<td>sind</td>
<td>0.006465</td>
<td>0.004271</td>
<td>0.008976</td>
<td>0.007308</td>
<td>0.006930</td>
<td>0.004964</td>
<td>0.005346</td>
</tr>
<tr>
<td>sich</td>
<td><strong>0.006309</strong></td>
<td><strong>0.011690</strong></td>
<td><strong>0.006283</strong></td>
<td><strong>0.006291</strong></td>
<td><strong>0.006930</strong></td>
<td><strong>0.007170</strong></td>
<td><strong>0.005435</strong></td>
</tr>
<tr>
<td>ich</td>
<td>0.006262</td>
<td>0.003877</td>
<td>0.013272</td>
<td>0.005366</td>
<td>0.008346</td>
<td>0.001434</td>
<td>0.001426</td>
</tr>
<tr>
<td>aber</td>
<td>0.006048</td>
<td>0.003347</td>
<td>0.007309</td>
<td>0.006245</td>
<td>0.007315</td>
<td>0.003365</td>
<td>0.003831</td>
</tr>
</tbody>
</table>

*sich* ist in allen L1 deutlich unterrepräsentiert.
Example: Overuse/underuse studies

What do we count? Variables? Variants? Both?
What do we count? Variables? Variants? Both?

We see that the reflexive *sich* is underused. Does this mean that

- learners leave out the *sich* in contexts where *sich* would be necessary? → variable “verbs with (subcategorized) reflexives”

- learners underuse structures that subcategorize for a reflexive? → variable “classes of semantically equivalent verbs”

- learners do not speak about anything that would require a reflexive verb? variable? no idea.

Each category (form, annotation category) can be the variant of different variables.
A very nice study that is variationist in spirit is the modality study by Ursula Maden Weinberger. She looked at modality in German, and instead of doing it form based (by looking at e.g. modal verbs) she annotated all modal expressions (modal verbs, adverbials, particles, mood, etc.) in several learner corpora, together with many co-varying variables.
Example: Modality | Maden-Weinberger (2009) | Form
FIGURE 7: EPISTEMIC MODAL VERB FREQUENCIES
PME = periphrastic modal expressions
We looked at

- grammatical annotation and discussed the comparative fallacy
- disfluency annotation and discussed the need for the development of new layers
- modality and variationist annotation and discussed the need for functional analysis

There are, of course, many more aspects of learner language that can be analyzed and annotated, such as the phonetic features, lexical properties, text structure, communicative aspects, politeness, etc.
Table of Contents

1 Corpus design and collection

2 Error Annotation

3 Other annotation layers

4 Quantitative analysis
   - Comparing Corpora

5 Summary
Table of Contents

1. Corpus design and collection

2. Error Annotation

3. Other annotation layers

4. Quantitative analysis
   - Comparing Corpora

5. Summary
We have spoken about corpus design and annotation. What can we do with an annotated learner corpus?

For many research questions we want to do a quantitative comparison of categories (forms, annotation categories) between two corpora. Each quantitative analysis is based on a qualitative analysis.

“corpora contain nothing but frequencies/probabilities — of occurrence or of co-occurrence” Gries (2015a, p. 93)

We can compare

- between two or more learner corpora
  - cross sectional studies: comparison between learners with different L1s, different ages, backgrounds, etc.
  - comparison between different registers
  - longitudinal: comparison between learners with different levels of proficiency (sometimes: quasi longitudinal)
- between learners and native speakers

Such comparisons are often subsumed under the label Contrastive Interlanguage Analysis (Granger 1996; Granger 2015).

There are many aspects that are relevant for the quantitative comparison of (learner) corpora – see Amir’s class and Gries (2015a).
vocabulary between Japanese adult learners of English, in-class conversation with spoken component of BNC (Shirato and Stapleton 2007)

certain words (stance, academic words) between ICLE and spoken and academic components of BNC (Gilquin and Paquot 2008)

parts of speech in a longitudinal corpus of learner German in KanDeL (Vyatkina et al. 2015)

modification between different subcorpora of Falko (different L2 corpora, L1) (Hirschmann et al. 2013)

disfluencies in L1 and L2 German, using BeMaTaC (Belz et al. submitted)

...
Comparing corpora: Prerequisites

- What’s the research question?
  → exploration, experiment, model

- Which varieties do we want to study? How is the variety sampled?
  → corpus design, documentation

- Which categories are analyzed? How can they be defined and annotated?
  → categorization, annotation, evaluation, documentation

- Which statistical methods are appropriate? How are the relevant categories distributed?
  → distributions, statistical methods (descriptive, inferential)
The quantitative comparison of varieties (corpora) is based on three (often implicit) assumptions

- the varieties are relatively homogenous
- the varieties differ only in the relevant parameters
- the categories we want to study are assigned in comparable ways
Comparing corpora: Prerequisites

The quantitative comparison of varieties (corpora) is based on three (often implicit) assumptions:

- The varieties are relatively homogenous.
- The varieties differ only in the relevant parameters.
- The categories we want to study are assigned in comparable ways.

In most cases, none of these assumptions holds. Here we will look at the first two assumptions.
Within-group variation

The assumption is that the varieties we want to compare are relatively homogenous (between-group variation is stronger than within-group variation). This is often not the case.

This is a real problem for our acquisition models. More and more studies are aware of this DoernyeiSkehan2003,Vyatkina2013-Specific,Aedel2015,HaniqueErnestusBoves2015.
Within-group variation | example | (Vyatkina et al. 2015)

Fig. 8. Individual trajectories of ADV for all participants. Smoothened lines symbolize the individual developments. Black dots represent the mean frequencies per time of measurement, black vertical lines are bootstrapped 95% confidence intervals. Frequencies are normalized by token frequencies.
Within-group variation | example | (Belz et al. submitted, from)

Figure 1: Frequency of silent and filled pauses within utterances per speaker for L1 and L2, normalized on speaker tokens.
Within-group variation | example | (Belz et al. submitted, from)

Figure 3: Frequency of repairs per speaker for L1 and L2, normalized on speaker tokens.
Another assumption is that the varieties differ only in the relevant parameters. In essence: we can attribute the differences we find to the parameters we study.

Corpus data is typically not tightly controlled. Any variable is influenced by many co-varying variables.
Comparison | studies

- vocabulary between Japanese adult learners of English, in-class conversation with spoken component of BNC (Shirato and Stapleton 2007)

- certain words (stance, academic words) between ICLE and spoken and academic components of BNC (Gilquin and Paquot 2008)

- parts of speech in a longitudinal corpus of learner German in KanDeL (Vyatkina et al. 2015)

- modification between different subcorpora of Falko (different L2 corpora, L1) (Hirschmann et al. 2013)

- disfluencies in L1 and L2 German, using BeMaTaC (Belz et al. submitted)
Comparison | studies

- vocabulary between Japanese adult learners of English, in-class conversation with spoken component of BNC (Shirato and Stapleton 2007) → additional (confounding) differences in situation, age, topic, ...

- certain words (stance, academic words) between ICLE and spoken and academic components of BNC (Gilquin and Paquot 2008)

- parts of speech in a longitudinal corpus of learner German in KanDeL (Vyatkina et al. 2015)

- modification between different sub corpora of Falko (different L2 corpora, L1) (Hirschmann et al. 2013)

- disfluencies in L1 and L2 German, using BeMaTaC (Belz et al. submitted)
Comparison | studies

- vocabulary between Japanese adult learners of English, in-class conversation with spoken component of BNC (Shirato and Stapleton 2007) → additional (confounding) differences in situation, age, topic, ...
- certain words (stance, academic words) between ICLE and spoken and academic components of BNC (Gilquin and Paquot 2008) → additional (confounding) differences in situation, age, topic, writing proficiency, ...
- parts of speech in a longitudinal corpus of learner German in KanDeL (Vyatkina et al. 2015)
- modification between different subcorpora of Falko (different L2 corpora, L1) (Hirschmann et al. 2013)
- disfluencies in L1 and L2 German, using BeMaTaC (Belz et al. submitted)
vocabulary between Japanese adult learners of English, in-class conversation with spoken component of BNC (Shirato and Stapleton 2007) → additional (confounding) differences in situation, age, topic, ...

certain words (stance, academic words) between ICLE and spoken and academic components of BNC (Gilquin and Paquot 2008) → additional (confounding) differences in situation, age, topic, writing proficiency, ...

parts of speech in a longitudinal corpus of learner German in KanDeL (Vyatkina et al. 2015) → additional (confounding) differences in task, topic, age ...

modification between different subcorpora of Falko (different L2 corpora, L1) (Hirschmann et al. 2013)

disfluencies in L1 and L2 German, using BeMaTaC (Belz et al. submitted)
Comparison | studies

- vocabulary between Japanese adult learners of English, in-class conversation with spoken component of BNC (Shirato and Stapleton 2007) → additional (confounding) differences in situation, age, topic, ...

- certain words (stance, academic words) between ICLE and spoken and academic components of BNC (Gilquin and Paquot 2008) → additional (confounding) differences in situation, age, topic, writing proficiency, ...

- parts of speech in a longitudinal corpus of learner German in KanDeL (Vyatkina et al. 2015) → additional (confounding) differences in task, topic, age ...

- modification between different subcorpora of Falko (different L2 corpora, L1) (Hirschmann et al. 2013) → additional (confounding) differences in background and experience, (sometimes age), ...

- disfluencies in L1 and L2 German, using BeMaTaC (Belz et al. submitted)
- vocabulary between Japanese adult learners of English, in-class conversation with spoken component of BNC (Shirato and Stapleton 2007) → additional (confounding) differences in situation, age, topic, ...

- certain words (stance, academic words) between ICLE and spoken and academic components of BNC (Gilquin and Paquot 2008) → additional (confounding) differences in situation, age, topic, writing proficiency, ...

- parts of speech in a longitudinal corpus of learner German in KanDeL (Vyatkina et al. 2015) → additional (confounding) differences in task, topic, age ...

- modification between different subcorpora of Falko (different L2 corpora, L1) (Hirschmann et al. 2013) → additional (confounding) differences in background and experience, (sometimes age), ...

- disfluencies in L1 and L2 German, using BeMaTaC (Belz et al. submitted) → additional (confounding) differences in background and experience, (sometimes age), ...
Often we find strong within-group variation.

Often we have several co-varying variables.

(I did not show it here but very often the assignment of categories differs between the corpora that are to be compared.)

The comparison of varieties can be difficult and conclusions must be formulated carefully. Alternative explanations cannot simply be discounted.
Learner language is systematic and, at the same time, variable. Learner corpora are one data source (among many) that can be used to explore, analyze, and model learner language.

We looked at

- corpus design – statistical properties, content decisions (level of proficiency, mode, naturalness)
- error annotation – general error tag sets, target hypotheses, specific tag sets
- other types of annotation – form annotation, functional annotation
- comparison of corpora – comparability, within-group variation
For each topic, we saw that it is necessary to think about methodology.

- In corpus design and representation as well as in annotation there are many (potentially controversial) decisions – document everything (metadata, tagsets, guidelines).

- Make your data and analysis available for transparency and re-use – use sustainable, well-understood open formats, architectures and open licenses.

<mantra>
Corpus design and corpus annotation depend primarily on the research question. Corpus architecture and formats depend on general considerations (availability, transparency, re-usability, reproducability).
</mantra>
Danke!  
Thank you!  

contact: anke.luedeling@rz.hu-berlin.de  

Thank you to the Korpling group at Humboldt-Universität.


Belz, Malte, Simon Sauer, Anke Lüdeling, and Christine Mooshammer (submitted). “Fluently disfluent? Pauses and repairs of advanced learners and native speakers of German”. In:


Chiarcos, Christian, Julia Ritz, and Manfred Stede (2009). “By all these lovely tokens... Merging Conflicting Tokenizations”. In: *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*. Singapore, pp. 35–43.


Lavandera, Beatriz (1978). “Where Does the Sociolinguistic Variable Stop?” In: 

Lüdeling, Anke and Hagen Hirschmann (2015). “Error Annotation”. In: The 
Cambridge Handbook of Learner Corpus Research. Ed. by Sylviane Granger, 
Gaëtanelle Gilquin, and Fanny Meunier. Cambridge: Cambridge University 
Press.

Lüdeling, Anke and Maik Walter (2010). “Korporallenguistik für Deutsch als 
Fremdsprache”. In: Deutsch als Fremd- und Zweitsprache. Ein internationales 
Handbuch. Ed. by Hans-Jürgen Krumm, Christian Fandrych, Britta Hufeisen, 

“Error Annotation in Spoken Learner Corpora”. In: Extended Abstracts of the 
Workshop on Phonetic Learner Corpora, Satellite workshop of the 18th 
International Congress of Phonetic Sciences. Ed. by Jürgen Trouvain, 
Frank Zimmerer, Mária Gósy, and Anne Bonneau, pp. 29–32. URL: 


Rastelli, Stefano (2009). “Learner corpora without error tagging”. In: Linguistik online 38.2/09, pp. 57–66.


Schiller, Anne, Simone Teufel, Christine Stöckert, and Christine Thielen (1999). “Guidelines für das Tagging deutscher Textcorpora mit STTS”. In: Universität Stuttgart, Institut für maschinelle Sprachverarbeitung; Universität Tübingen, Seminar für Sprachwissenschaft.


Tavakoli, Parvaneh. “Pausing Patterns: Differences Between L2 Learners and Native Speakers”. In:


