Routes of investigation on deeply annotated learner data
Falko – German learner corpus

Marc Reznicek
Hagen Hirschmann
Anke Lüdeling
Ines Rehbein
Amir Zeldes
research question

- How do heavily annotated second language learner corpora help in understanding interlanguage/acquisition processes?

- error annotation with target hypotheses
- overuse/underuse of multilevel items
- syntactical parsing of learner data
background: interlanguage

assumption:
- learners of a second/foreign language have a systematic internal grammar (interlanguage)
- different from the internal grammar of L1 speakers of the target language

interlanguage is influenced by
- the learners' L1 (transfer, interference)
- the structure of the L2
- general learning principles
- mode of acquisition / teaching method / learning strategies

Selinker (1972), Nickel (1998) and many others
interlanguage & data

further assumption:

- interlanguage can be researched through the analysis of (naturally occurring) learner data
- one type of data: learner corpora

analyses:

- error analysis (EA)
  → analysis of learner data with a 'correct' form
- contrastive interlanguage analysis (CIA)
  → analysis of the learner data wrt to another corpus
plan

• learner corpora / Falko
• annotation of learner corpora
• overuse and underuse statistics
• analysis of parsed learner data
plan

• learner corpora / Falko
• annotation of learner corpora
• overuse and underuse statistics
• analysis of parsed learner data
learner corpora

- principled and well-documented collections of learner language
- the design depends on the research question
  - written vs. oral data / text type / type of exercise
  - grade of advancedness
  - L1s of the learners
  - ...
- many learner corpora for English,
  more and more learner corpora for other languages

Granger/Hung/Petch-Tyson (2002), Cobb (2003), Tono (2003),
Myles/Mitchell (2004), Nesselhauf (2004), Tenfjord/Meurer/Hofland
(2004), Granger (2008), Lüdeling/Walter (2009) etc.
German error-annotated learner corpus

- freely available
- advanced learners ~B1-C1 (CEFR)
- written language / controlled, unaided writing
- several text types (sub-corpora)
  - essays (122,791 tokens) + WHIG (~120,000 tokens) soon
  - summaries (40,787 tokens)
- comparable native speaker corpora
  - essays (68,480 tokens)
  - summaries (21,184 tokens)
- meta-data for each learner (bibliographic data, linguistic history, c-test score)
  Lüdeling et al. (2008), Reznicek et al. (2010)

http://www.linguistik.hu-berlin.de/institut/professuren/-korpuslinguistik/forschung/falko/standardseite
plan

- learner corpora / Falko
- annotation of learner corpora
- overuse and underuse statistics
- analysis of parsed learner data
annotation of learner data: conceptual issues

- annotation of learner data is highly problematic

  - data is **unsystematic** (especially if there are different L1s) – difficult for automatic tools
  - for error analysis and contrastive interlanguage analysis:

→ data has to be **interpreted**
  
  (long discussion in acquisition research)

consider:

An der anderen Seite, wenn da kein Feminismus wäre, stünden wir noch nur in der Küche und köchten wir. (fkb034_2008_07)

~ "On the other hand, if there were no feminism, we would still only stand in the kitchen and cook."
consider:

An der anderen Seite, wenn da kein Feminismus wäre, stünden wir noch nur in der Küche und köchten wir. (fkb034_2008_07)

~ "On the other hand, if there were no feminism, we would still only stand in the kitchen and cook."
annotation of learner data: target hypothesis

- all error tags depend on an (at least implicit) correct version of a learner utterance → target hypothesis

- Falko: explicit target hypotheses
  often there are several ways of correcting an utterance

th1: Auf der anderen Seite, wenn da kein Feminismus wäre, stünden wir nur noch in der Küche und kochten.

th2: Andererseits, stünden wir wenn es keinen Feminismus gäbe, nur noch in der Küche und kochten.
annotation of learner data: target hypothesis in Falko

**th1:** sentence-based, very close to original text, mainly clear grammatical errors

**th2:** text-based, also stylistic errors

- the differences between a target hypothesis and the original data is automatically annotated with **edit tags** (change, insert, replace etc.)

- all layers are automatically annotated with POS tags & lemma (TreeTagger, Schmid 1994)

- additionally – **manual error tags** for some phenomena
<table>
<thead>
<tr>
<th>LT</th>
<th>TH1</th>
<th>TH1Diff</th>
<th>TH2</th>
<th>TH2Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>An</td>
<td>Auf</td>
<td>CHA</td>
<td>Andererseits</td>
<td>MERGE</td>
</tr>
<tr>
<td>der</td>
<td>der</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>anderen</td>
<td>anderen</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seite</td>
<td>Seite</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>,</td>
<td>,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>,</td>
<td>,</td>
<td>stünden</td>
<td>MOVT</td>
<td></td>
</tr>
<tr>
<td>wenn</td>
<td>wenn</td>
<td></td>
<td></td>
<td>INS</td>
</tr>
<tr>
<td>da</td>
<td>da</td>
<td></td>
<td>wend</td>
<td>DEL</td>
</tr>
<tr>
<td>kein</td>
<td>kein</td>
<td></td>
<td>es</td>
<td>INS</td>
</tr>
<tr>
<td>Feminismus</td>
<td>Feminismus</td>
<td>Feminismus</td>
<td>Feminismus</td>
<td></td>
</tr>
<tr>
<td>wäre</td>
<td>wäre</td>
<td></td>
<td>gäbe</td>
<td>CHA</td>
</tr>
<tr>
<td>,</td>
<td>,</td>
<td></td>
<td>,</td>
<td></td>
</tr>
<tr>
<td>stünden</td>
<td>stünden</td>
<td></td>
<td>MOVS</td>
<td></td>
</tr>
<tr>
<td>wir</td>
<td>wir</td>
<td></td>
<td>MOVS</td>
<td></td>
</tr>
<tr>
<td>nur</td>
<td>nur</td>
<td>MOVT</td>
<td>nur</td>
<td>MOVS</td>
</tr>
<tr>
<td>noch</td>
<td>noch</td>
<td>MOVS</td>
<td>noch</td>
<td>MOVS</td>
</tr>
<tr>
<td>nur</td>
<td>nur</td>
<td>MOVS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>in</td>
<td>in</td>
<td></td>
<td>in</td>
<td></td>
</tr>
<tr>
<td>LT</td>
<td>pos</td>
<td>lemma</td>
<td>TH1</td>
<td>TH1Diff</td>
</tr>
<tr>
<td>-----</td>
<td>------</td>
<td>-------</td>
<td>-------</td>
<td>---------</td>
</tr>
<tr>
<td>An</td>
<td>APPR</td>
<td>an</td>
<td>Auf</td>
<td>CHA</td>
</tr>
<tr>
<td>der</td>
<td>ART</td>
<td>d</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anderen</td>
<td>ADJA</td>
<td>andere</td>
<td>anderen</td>
<td>ADJA</td>
</tr>
<tr>
<td>Seite</td>
<td>NN</td>
<td>Seite</td>
<td>Seite</td>
<td>NN</td>
</tr>
<tr>
<td>,</td>
<td>$,</td>
<td>,</td>
<td>,</td>
<td>$,</td>
</tr>
<tr>
<td>wenn</td>
<td>KOUS</td>
<td>wenn</td>
<td>wenn</td>
<td>KOUS</td>
</tr>
<tr>
<td>da</td>
<td>PAV</td>
<td>da</td>
<td>da</td>
<td>PAV</td>
</tr>
<tr>
<td>kein</td>
<td>PIAT</td>
<td>kein</td>
<td>kein</td>
<td>PIAT</td>
</tr>
<tr>
<td>Feminismus</td>
<td>NN</td>
<td>Feminismus</td>
<td>Feminismus</td>
<td>NN</td>
</tr>
<tr>
<td>wäre</td>
<td>VAFIN</td>
<td>sein</td>
<td>wäre</td>
<td>VAFIN</td>
</tr>
<tr>
<td>,</td>
<td>$,</td>
<td>,</td>
<td>,</td>
<td>$,</td>
</tr>
<tr>
<td>stünden</td>
<td>VVFIN</td>
<td>stehen</td>
<td>stünden</td>
<td>VVFIN</td>
</tr>
<tr>
<td>wir</td>
<td>PPER</td>
<td>wir</td>
<td>wir</td>
<td>PPER</td>
</tr>
<tr>
<td>nur</td>
<td>ADV</td>
<td>nur</td>
<td>nur</td>
<td>ADV</td>
</tr>
<tr>
<td>noch</td>
<td>ADV</td>
<td>noch</td>
<td>noch</td>
<td>ADV</td>
</tr>
<tr>
<td>nur</td>
<td>ADV</td>
<td>nur</td>
<td>nur</td>
<td>ADV</td>
</tr>
<tr>
<td>in</td>
<td>APPR</td>
<td>in</td>
<td>in</td>
<td>APPR</td>
</tr>
</tbody>
</table>
### Automatic error annotation: edit tags

<table>
<thead>
<tr>
<th>LT</th>
<th>pos</th>
<th>lemma</th>
<th>TH1</th>
<th>TH1Diff</th>
<th>TH1pos</th>
<th>TH1posDiff</th>
</tr>
</thead>
<tbody>
<tr>
<td>An</td>
<td>APPR</td>
<td>an</td>
<td>Auf</td>
<td>CHA</td>
<td>APPR</td>
<td></td>
</tr>
<tr>
<td>der</td>
<td>ART</td>
<td>d</td>
<td>der</td>
<td>ART</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anderen</td>
<td>ADJA</td>
<td>andere</td>
<td>anderen</td>
<td>ADJA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seite</td>
<td>NN</td>
<td>Seite</td>
<td>Seite</td>
<td>NN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>,</td>
<td>$,$</td>
<td>,</td>
<td>,</td>
<td>$,$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>wenn</td>
<td>KOUS</td>
<td>wenn</td>
<td>wenn</td>
<td>KOUS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>da</td>
<td>PAV</td>
<td>da</td>
<td>da</td>
<td>PAV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>kein</td>
<td>PIAT</td>
<td>kein</td>
<td>kein</td>
<td>PIAT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feminismus</td>
<td>NN</td>
<td>Feminismus</td>
<td>Feminismus</td>
<td>NN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>wäre</td>
<td>VAFIN</td>
<td>sein</td>
<td>wäre</td>
<td>VAFIN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>,</td>
<td>$,$</td>
<td>,</td>
<td>,</td>
<td>$,$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>stünden</td>
<td>VVFIN</td>
<td>stehen</td>
<td>stünden</td>
<td>VVFIN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>wir</td>
<td>PPER</td>
<td>wir</td>
<td>wir</td>
<td>PPER</td>
<td></td>
<td></td>
</tr>
<tr>
<td>noch</td>
<td>ADV</td>
<td>noch</td>
<td>noch</td>
<td>ADV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nur</td>
<td>ADV</td>
<td>nur</td>
<td>nur</td>
<td>ADV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>in</td>
<td>APPR</td>
<td>in</td>
<td>in</td>
<td>APPR</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Search in Annis

- **Search window**
- **Match count**
- **Corpus selection**
- **Metadata text**

http://korpling.german.hu-berlin.de/falko-suche/
Search in Annis

Original text and token based annotations

show or export results

Choose left and right context

partitur with spans

Target hypothesis 2
error annotations

Target hypothesis 1
error annotations
Search in Annis

nur noch
only still
just

MOVT = MOVEDtarget
token should appear here

tokens in complete text

http://korpling.german.hu-berlin.de/falko-suche
plan

• learner corpora / Falko
• annotation of learner corpora
• overuse and underuse statistics
• analysis of parsed learner data
research question

- We want to find **structural problems** in German L2 interlanguage

- structural problems are those problems that
  - occur **independent of the learners' L1**
  - and are therefore **attributed to the structure of the target grammar**
how can we detect acquisition problems?

- structures that are unique for the L2 or different from the learners' L1s (transfer)

- structures that are judged to be difficult by the learners

- structures that contain many errors

- underused structures
how can we detect acquisition problems?

- grammatical analysis
- proved to be extremely problematic no straightforward transfer

- intuition of the learners
  (unsystematic, dependent on teaching)

- experiments

- intuition of the teachers (unsystematic)

- corpus analysis, error analysis

- corpus analysis, Contrastive Interlanguage Analysis
how can we detect acquisition problems?

- **grammatical analysis**
  - proved to be extremely problematic no straightforward transfer

- **intuition of the learners**
  - (unsystematic, dependent on teaching)

- **experiments**

- **intuition of the teachers** (unsystematic)

- **corpus analysis, error analysis**

**underused structures**

- **corpus analysis, Contrastive Interlanguage Analysis**
underuse

- L2 distributions are compared to L1 distributions

- Overuse & underuse are defined as (statistically significant) differences between the varieties

- A category can be underused in L2 because
  - the learners do not know it
  - the learners do know it but (unconsciously) avoid it

→ diagnostics for detecting structural acquisition problems
visualization of overuse and underuse

- underuse: cold colours
- overuse: warm colours
- intensity of colour signals strength of overuse/underuse

- Excel add in by Amir Zeldes available at http://korpling.german.hu-berlin.de/~amir/uoaddin.htm
visualization of overuse and underuse: lexical categories

<table>
<thead>
<tr>
<th>lemma</th>
<th>tot_norm</th>
<th>deu</th>
<th>dan</th>
<th>eng</th>
<th>fra</th>
<th>pln</th>
<th>rus</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>0.004964</td>
<td>0.005346</td>
</tr>
<tr>
<td>für</td>
<td>0.006683</td>
<td>0.007201</td>
<td>0.006091</td>
<td>0.006802</td>
<td>0.007216</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>0.006309</td>
<td>0.011887</td>
<td>0.006283</td>
<td>0.006291</td>
<td>0.006930</td>
<td>0.007170</td>
<td>0.005435</td>
</tr>
<tr>
<td>ich</td>
<td>0.006262</td>
<td>0.003877</td>
<td>0.013272</td>
<td>0.005366</td>
<td>0.003465</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* (reflexive pronoun) is underused in all L1 groups
visualization of overuse and underuse: bigrams of pos-categories

<table>
<thead>
<tr>
<th>bigram</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>$.-PPER</td>
<td>0.042384</td>
<td>0.005297</td>
<td>0.009748</td>
<td>0.007963</td>
<td>0.006166</td>
<td>0.005801</td>
<td>0.007409</td>
</tr>
<tr>
<td>VVFIN-$,</td>
<td>0.042131</td>
<td>0.006457</td>
<td>0.00776</td>
<td>0.006343</td>
<td>0.006937</td>
<td>0.006243</td>
<td>0.008391</td>
</tr>
<tr>
<td>PPOSAT-NN</td>
<td>0.041739</td>
<td>0.008058</td>
<td>0.007247</td>
<td>0.007269</td>
<td>0.007066</td>
<td>0.006298</td>
<td>0.005802</td>
</tr>
<tr>
<td><strong>ADV-ADV</strong></td>
<td><strong>0.041604</strong></td>
<td><strong>0.012848</strong></td>
<td><strong>0.010518</strong></td>
<td><strong>0.006111</strong></td>
<td><strong>0.006166</strong></td>
<td><strong>0.003094</strong></td>
<td><strong>0.002856</strong></td>
</tr>
<tr>
<td>ADV-APPR</td>
<td>0.039742</td>
<td>0.009117</td>
<td>0.008016</td>
<td>0.005324</td>
<td>0.007837</td>
<td>0.004807</td>
<td>0.004642</td>
</tr>
<tr>
<td>PDAT-NN</td>
<td>0.03956</td>
<td>0.005409</td>
<td>0.004233</td>
<td>0.005509</td>
<td>0.007837</td>
<td>0.007735</td>
<td>0.008837</td>
</tr>
<tr>
<td>ADV-ART</td>
<td>0.037125</td>
<td>0.007629</td>
<td>0.006349</td>
<td>0.006898</td>
<td>0.005653</td>
<td>0.006133</td>
<td>0.004463</td>
</tr>
</tbody>
</table>

adverb chains are underused in all L1 groups
form vs. function
example study on modification
modification

- corpus-based studies of adverbs in GFL
  - typically based on lexical items and (rarely) word classes (form-based)
  - typically for one language pair (Möllering 2004, Vyatkina 2007 etc.)

- ADV underuse points to a more general phenomenon: modification
modification

- 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?)
underuse of adverbs: function

- pos tag ADV is not fine-grained enough

- better classification and different functions
  - classes show different distributions
  - only some of these classes are underused by the learners

Hirschmann (2011, in preparation)
strength of underuse of different syntactic ADV classes

PTK: particles *sehr* gut – *very* good

ADVV: modal adv *Bald* schneit es – *Soon* it will snow

ADVS: sentence adv. *Bestimmt* regnet es bald – *Certainly*, it will snow soon

PTKM: modal part. *Er mag sie* eben – *He just* likes her.
modification

- effects form-based or function-based?
  - are all adverbs underused?
    - → No: auch, noch etc. over
  - are certain adverbs (forms) underused?
    - → Yes
  - 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?)
underuse of adverbs: function

- underuse shows differences between different adverbial functions

- but classification is still word based

→ necessity to code syntactic functions independent of filler category
Falko – syntactic preprocessing

- Parses should be built on
  - grammatical sentences
  - learner near utterances

- for CIA Falko L1 & L2 corpus
- manually corrected pos tags
- semi-automatic sentence segmentation
Falko – syntactic annotation

- dependency parser (Bernd Bohnet 2010; Syntactic Analyser)
- training data: TiGer dependency bank
  - (derived from ~50000 trees of the TiGer treebank)
- result: very accurate dependency parses
  - with syntactic functions
syntax schema (very briefly)

- every word is connected with its dependent(s)
- arrows point to hierachically lower dependent
- each arrow (dependency) has a function label
searching for modification in Falko

- different aspects of the problem
  - is the syntactic function 'modification' underused?
  - what is the target of the modification?
  - what are the categories used for modification?

This is true especially for children
overuse / underuse of syntactic functions

<table>
<thead>
<tr>
<th>label</th>
<th>de</th>
<th>da</th>
<th>en</th>
<th>fr</th>
<th>ru</th>
<th>usb</th>
</tr>
</thead>
<tbody>
<tr>
<td>NK</td>
<td>0.264067</td>
<td>0.278546</td>
<td>0.284881</td>
<td>0.303271</td>
<td>0.29552</td>
<td>0.295136</td>
</tr>
<tr>
<td>HD</td>
<td>0.156192</td>
<td>0.155622</td>
<td>0.157178</td>
<td>0.154275</td>
<td>0.15809</td>
<td>0.156483</td>
</tr>
<tr>
<td>MO</td>
<td>0.141968</td>
<td>0.12789</td>
<td>0.113704</td>
<td>0.110112</td>
<td>0.112513</td>
<td>0.108707</td>
</tr>
<tr>
<td>SB</td>
<td>0.07398</td>
<td>0.078506</td>
<td>0.077099</td>
<td>0.075093</td>
<td>0.078852</td>
<td>0.085512</td>
</tr>
<tr>
<td>CJ</td>
<td>0.059604</td>
<td>0.053397</td>
<td>0.056411</td>
<td>0.050632</td>
<td>0.059274</td>
<td>0.072183</td>
</tr>
<tr>
<td>AC</td>
<td>0.057051</td>
<td>0.059317</td>
<td>0.057215</td>
<td>0.054796</td>
<td>0.054012</td>
<td>0.04916</td>
</tr>
<tr>
<td>OC</td>
<td>0.050335</td>
<td>0.053039</td>
<td>0.050008</td>
<td>0.049888</td>
<td>0.047125</td>
<td>0.040679</td>
</tr>
<tr>
<td>OA</td>
<td>0.044213</td>
<td>0.042352</td>
<td>0.044097</td>
<td>0.043643</td>
<td>0.046119</td>
<td>0.046218</td>
</tr>
<tr>
<td>CD</td>
<td>0.026549</td>
<td>0.024632</td>
<td>0.025639</td>
<td>0.022156</td>
<td>0.024917</td>
<td>0.030466</td>
</tr>
<tr>
<td>CP</td>
<td>0.017653</td>
<td>0.021732</td>
<td>0.020325</td>
<td>0.018141</td>
<td>0.017256</td>
<td>0.014887</td>
</tr>
<tr>
<td>PD</td>
<td>0.014435</td>
<td>0.014462</td>
<td>0.015943</td>
<td>0.015019</td>
<td>0.016947</td>
<td>0.018002</td>
</tr>
<tr>
<td>NG</td>
<td>0.011065</td>
<td>0.011561</td>
<td>0.010914</td>
<td>0.00974</td>
<td>0.00975</td>
<td>0.011252</td>
</tr>
<tr>
<td>MNR</td>
<td>0.010995</td>
<td>0.013707</td>
<td>0.013429</td>
<td>0.013383</td>
<td>0.010679</td>
<td>0.009521</td>
</tr>
<tr>
<td>RC</td>
<td>0.010051</td>
<td>0.008979</td>
<td>0.009385</td>
<td>0.011375</td>
<td>0.006268</td>
<td>0.005366</td>
</tr>
</tbody>
</table>
overuse / underuse of syntactic functions – significant results

<table>
<thead>
<tr>
<th>label</th>
<th>de</th>
<th>da</th>
<th>en</th>
<th>fr</th>
<th>ru</th>
<th>usb</th>
</tr>
</thead>
<tbody>
<tr>
<td>NK</td>
<td>0,264067</td>
<td>0,278546</td>
<td>0,284881</td>
<td>0,303271</td>
<td>0,29552</td>
<td>0,295136</td>
</tr>
<tr>
<td>HD</td>
<td>0,156192</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MO</td>
<td>0,141968</td>
<td>0,12789</td>
<td>0,113704</td>
<td>0,110112</td>
<td>0,112513</td>
<td>0,108707</td>
</tr>
<tr>
<td>SB</td>
<td>0,07398</td>
<td>0,078506</td>
<td></td>
<td>0,078852</td>
<td>0,085512</td>
<td></td>
</tr>
<tr>
<td>CJ</td>
<td>0,059604</td>
<td>0,053397</td>
<td>0,056411</td>
<td>0,050632</td>
<td></td>
<td>0,072183</td>
</tr>
<tr>
<td>AC</td>
<td>0,057051</td>
<td></td>
<td></td>
<td>0,050335</td>
<td></td>
<td>0,04916</td>
</tr>
<tr>
<td>OC</td>
<td>0,050335</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0,040679</td>
</tr>
<tr>
<td>OA</td>
<td>0,044213</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CD</td>
<td>0,026549</td>
<td></td>
<td>0,022156</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP</td>
<td>0,017653</td>
<td>0,021732</td>
<td>0,020325</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PD</td>
<td>0,014435</td>
<td>0,015943</td>
<td></td>
<td>0,016947</td>
<td>0,018002</td>
<td></td>
</tr>
<tr>
<td>NG</td>
<td>0,011065</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MNR</td>
<td>0,010995</td>
<td>0,013707</td>
<td>0,013429</td>
<td>0,013383</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RC</td>
<td>0,010051</td>
<td></td>
<td></td>
<td>0,006268</td>
<td>0,005366</td>
<td></td>
</tr>
</tbody>
</table>

MO (modification) is significantly underused independent of L1
modification

- effects form-based or function-based?
  - are all adverbs underused? → No: auch, noch etc. over
  - are certain adverbs (forms) underused? → Yes
  - are certain adverbs (forms) underused in certain functions? → Yes
  - are certain adverbial functions underused?
  - is modification generally underused?
In my opinion this statement holds.

the often very theoretical approach

especially in Denmark where ...

...and exactly for this reason ...

Perhaps not when

Only then do they develop...

frequencies normalized per 1000 edges
modified element – results

- all categories are frequently modified in both L1 and L2

- but *all syntactic relations* possible for modification are underused

- modifiers of adverbs show the strongest underuse
If she makes her career, …

Some have success [with this] …

One can, as mentioned above …

To make money on a criminal basis …

… criminality increases steadily …

which still exists …

<table>
<thead>
<tr>
<th>func</th>
<th>L2 (norm)</th>
<th>L1 (norm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>14,6162802</td>
<td>12,8218827</td>
</tr>
<tr>
<td>PROADV</td>
<td>7,41011413</td>
<td>6,73148841</td>
</tr>
<tr>
<td>COMPARE</td>
<td>0,26343296</td>
<td>0,27475463</td>
</tr>
<tr>
<td>PREP</td>
<td>44,8600831</td>
<td>48,5857769</td>
</tr>
<tr>
<td>ADJ</td>
<td>12,7722495</td>
<td>17,5842962</td>
</tr>
<tr>
<td>ADV</td>
<td>61,8302642</td>
<td>87,7230473</td>
</tr>
</tbody>
</table>

frequencies normalized per 1000 edges
modifier – results

- categories of different complexity (lexemes to sentences) are used for modification; modification is frequent in L2 and L1

- some categories are underused by the learners, two categories are slightly overused

- adverbs and (adverbially used) adjectives show the strongest underuse
modification

- effects form-based or function-based?
  - are all adverbs underused?
    → No: auch, noch etc. overused
  - are certain adverbs (forms) underused?
    → Yes
  - are certain adverbs (forms) underused in certain functions?
    → Yes
  - are certain adverbial functions underused?
    → Yes
  - is modification generally underused?
    → Yes
summary: modification in Falko

- modification is a difficult category for learners of GFL
  - previous evidence: form-based
  - previous hypotheses: 'transfer', polyfunctionality

- additional syntactic evidence shows the syntactic function 'modification' is underused
  - independent of form
  - independent of the learners L1
Summary: learner corpora

- Learner corpora like Falko can help:
  - find abstract error patterns and relate them to meta-data
  - contrast frequencies in learner language structures on very different levels of abstraction to
    - to native speakers
    - between different learner groups

Parsed data empowers investigators to study differences in the use of specific functions rather than simply of forms or form classes.
謝謝
Thank you!
Danke!

Falko:
http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko

contact: marc.reznicek@staff.hu-berlin.de


Bibliography


Tenfjord, Kari; Hagen, Jon Erik; Johansen, Hilde (2006) The hows and whys of coding categories in a learner corpus (or How and why an error-tagged learner corpus is not ipso facto one big comparative fallacy). In: Rivista di Psicolinguistica Applicata (RiPLA) VI.(3), 93-108


