What's Hard?

Quantitative Evidence for Difficult Constructions in German Learner Data

Humboldt-Universität zu Berlin

Amir Zeldes
amir.zeldes@rz.hu-berlin.de
Anke Lüdeling
anke.luedeling@rz.hu-berlin.de
Hagen Hirschmann
hirschhx@rz.hu-berlin.de

QITL-3, Helsinki, 2-4 June 2008

Research questions

- What's hard/easy for L2 German learners, and how can we find this out?
- What do (advanced) learners do differently from natives?
- Why?

Overview

- Operationalizing L2 difficulties
- Learner data and the Falko corpus
- Analysis of two case studies
- Summary and conclusions

Approaches to L2 difficulty

- Use intuition / introspection as learner, teacher or native speaker
- Compose questionnaires for students or teachers (Diehl et al. 1991)
- Gather corpus data:
  - Learner corpora (see Pravec 2002; Tono 2003; Granger, to appear)
  - Comparable L1 corpora

Corpus Data

- Learner corpora contain L2 learner data from essays, exercises etc. (see Granger 2002, to appear)
- Usually give metadata on learner level and background
- Some contain explicit error annotations (Corder 1981)

Error annotation

- Essentially based on a target hypothesis:
  "what should the learner have said?"

  John goed home > *John went home*
  anno=[irregular past tense form error]

- But things are not always so simple…

Ambiguity of error annotation

	was		der	Novelle	oder	der	Ode	nicht	betrift
	what		the	novella	or	the	ode	not	applies
	which does not apply to the novella or the ode								
1	was	auf		Novelle	oder		Ode	nicht	zutrifft
2	was	auf	die	Novelle	oder	die	Ode	nicht	zutrifft
3	was	bei	der	Novelle	oder	der	Ode	nicht	der Fall ist
4	was	für	die	Novelle	oder	die	Ode	nicht	zutrifft
5	das		die	Novelle	oder	die	Ode	nicht	betrifft

Lüdeling (2008)

Target hypothesis: experiment

- 5 annotations for 17 sentences (one text)
- target hypothesis differs, annotation scheme identical

content words	function words
15	13
24	26
17	25
16	12
14	22

Working with raw learner data

- Frequencies of word forms etc. in learner data
- Work on lexical density as an index of L2 competence (Halliday 1989; Laufer/Nation 1999)
- Studies using underuse/overuse compared to native data in the framework of Contrastive Interlanguage Analysis (see Selinker 1972; Ringbom 1998; Granger et al. 2002)

Underuse and Overuse

- Simplified model of target language competence
- Learner's interlanguage distributions as opposed to L1 distributions
- Underuse and overuse defined as statistically significant deviations from L1 control frequencies

Underuse as an index of difficulty

- Phenomena that are underrepresented can either be:
  - Unknown to learners (e.g. probably the word *forthwith*)
  - Known but (more or less consciously) avoided (e.g. the past perfect progressive)
- No attempt is made here to distinguish between these cases

L1 Independence

- Some errors are strongly L1 dependent, i.e. transfer errors:

  is beautiful! (Italian pro-drop transfer)

- We are interested in phenomena that present difficulties to German learners independently of L1

➢ Use L1 metadata to rule out interference and other language dependent effects

Our data – the Falko corpus

- **F**ehler**a**nnotiertes **L**erner**ko**rpus des Deutschen (Lüdeling et al. 2008)
- Advanced learners (c-test, university exam)
- Summaries and essays written by learners, total of 262230 tokens
- ca. 50 different L1s represented
- Control corpus of native Germans, total of 101404 tokens

Corpus available at:
http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko

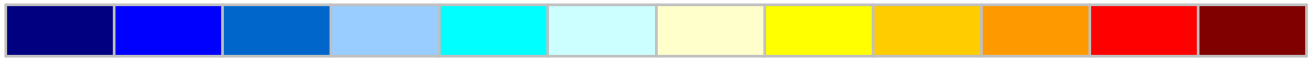
Our data – the Falko corpus

- We examine 5 sub-corpora of L1: Danish, English, French, Polish & Russian speakers
- Comparable native corpus
- Other L1s left as unseen data (58210 tokens)

Natives		Learners	
de	74280	da	15593
		en	21600
		fr	7786
		pl	18100
		ru	11203
subtotal	74280	subtotal	88736
total 163016			

Visualizing Underuse/Overuse

- Normalized frequencies are collected from all subcorpora for:
  - lexical categories (lemmas)
  - grammatical categories (POS *n*-grams)
- Degree of deviation from native frequency is represented in progressively warmer or colder colors

Underuse  **Overuse**

Visualization of Lexical Data

lemma	tot_norm	de	da	en	fr	pl	ru
in	0.013188	0.012261	0.014041	0.014247	0.015272	0.012135	0.009534
es	0.010897	0.011945	0.010900	0.011379	0.013347	0.008163	0.012385
sie	0.010618	0.008193	0.010643	0.008835	0.010909	0.006067	0.005613
man	0.010164	0.007900	0.012438	0.008742	0.009754	0.006950	0.007306
dass	0.009522	0.007404	0.012823	0.008789	0.009625	0.008880	0.009890
von	0.007982	0.007122	0.007309	0.006846	0.007315	0.010259	0.007930
auch	0.007028	0.008362	0.008527	0.005828	0.005775	0.005461	0.004455
für	0.006683	0.007201	0.006091	0.007216	0.006802	0.005736	0.004188
sind	0.006465	0.004271	0.008976	0.007308	0.006930	0.004964	0.005346
sich	**0.006309**	**0.011697**	**0.006283**	**0.006291**	**0.006930**	**0.007170**	**0.005435**
ich	0.006262	0.003877	0.013272	0.005366	0.003465	0.001434	0.001426
aber	0.006048	0.003347	0.007309	0.006245	0.007315	0.003365	0.003831

Reflexive *sich* ‘self’ is used too rarely!

Underuse of reflexive *sich* in all L1s

- Underuse ratio ~0.5 (half as frequent in learner data: 479:1038)
- Very significant difference between natives and learners in post-hoc test of equal proportions

	de:learner	de:da	de:en	de:fr	de:pl	de:ru
p-val.	< 2.2e-16	3.314e-9	8.518e-12	1.849e-4	1.595e-7	3.465e-9

- Confirmed pre-hoc in unseen L1s (p-val. < 2.2e-16)
- No difference between learner L1s (p-val. 0.4478)

Possible explanations

- Interference: learners use *sich* under the influence of their native reflexives
- But:
  - Interference is L1-dependent and should produce different results in each L1
  - Learner L1s differ substantially in this respect (e.g. no reflexive in English, very similar one in Danish, and likewise in non IE languages)

Possible explanations

- Word order complexity
  - German word order varies depending on syntactic construction
  - Difficult to acquire (cf. Clahsen 1984, Parodi 1998)

Four positions for *sich*

1. *die Stadt ändert* ***sich***
   the city changes [refl]
   the city changes
2. *dass* ***sich*** *die Stadt ändert*
   that [refl] the city changes
   that the city changes
3. *dass die Stadt* ***sich*** *ändert*
   that the city [refl] changes
   that the city changes
4. ***sich*** *zu ändern*
   [refl] to change
   to change

Possible explanations

- Word order complexity

  > but no difference between clause types (χ^2 p-val. of 0.354)

***sich* is similarly underused independent of L1 and embedding clause type**

Where is *sich* not underused?

- Examine *n*-grams with *sich*
- *sich* is not underused:
  - When the subject is *man* ‘one’ (ratio ~0.9)

    *Wenn **man sich** bemüht*

    if one [refl] exerts

    If one makes the effort
  - When the verb is *lassen* ‘allow, let’ (ratio ~1.5)

    *Anhand dieses Beispiels **läßt sich** erschließen*

    using this example allows [refl] conclude

    Using this example it is possible to conclude

Possible explanations

- Learners overuse *man* and *lassen*
  > not true: underuse of 0.95 and 0.56
- These bigrams are especially common
  - *man* is the 3rd most common word form preceding *sich* in the native corpus
  - *lassen* is the 4th most common verb preceding *sich*, and 2nd most associated with *sich* (MI)

Possible explanations

- Word order is simpler/more constant
  - Word order (2) is impossible with *man*
  - *sich* always follows *man*
  - *lassen* is most common in main sentences with *sich* following

> *sich* is underused except in frequent, consistent constructions

POS Chains

bigram	tot_norm	de	da	en	fr	pl	ru
$.-PPER	0.042384	0.005297	0.009748	0.007963	0.006166	0.005801	0.007409
VVFIN-$,	0.042131	0.006457	0.00776	0.006343	0.006937	0.006243	0.008391
PPOSAT-NN	0.041739	0.008058	0.007247	0.007269	0.007066	0.006298	0.005802
ADV-ADV	**0.041604**	**0.012858**	**0.010518**	**0.006111**	**0.006166**	**0.003094**	**0.002856**
ADV-APPR	0.039742	0.009117	0.008016	0.005324	0.007837	0.004807	0.004642
PDAT-NN	0.03956	0.005409	0.004233	0.005509	0.007837	0.007735	0.008837
ADV-ART	0.037125	0.007629	0.006349	0.006898	0.005653	0.006133	0.004463

Multiple adverb chains are underused in all learner subcorpora

Underuse of ADV-ADV *n*-grams

- Underuse very significant, larger ratio the longer the chain:
  - ADV x 2: 1141:432 ~45% ($p < 2.2e-16$)
  - ADV x 3: 162:36 ~27% ($p = 1.776e-14$)
  - ADV x 4: 19:1
  - ADV x 5: 2:0
  - ADV x 6: 1:0
- Confirmed pre-hoc in other L1s (ADV x 2: $p < 2.2e-16$, ADV x 3: $p = 2.060e-12$)

Underuse of ADV-ADV *n*-grams

- High type-token ratio
  > can't statistically contrast specific chains
- Division of the 30 most common types into four categories:
  - Adverbs belong to different phrases
  - Adverbs belong to same phrase
    - Left-headed
    - Right-headed
    - lexicalized

ADV-ADV examples

1. *Es ist* [*doch*] [*auch*] *statistisch belegt*
   it is indeed also statistically proven
   Furthermore, it is indeed statistically proven
2. *ein Kampf, dass bis* [***heute*** *noch*] *andauert*
   a fight that until today still endures
   a fight which has lasted until today
3. *wo es* (...) [[*viel* ***mehr***] *Arbeitsplätze*] *gibt*
   where it much more jobs gives
   where there are many more jobs
4. *und* [*immer noch*] *kann man eine unzufriedenheit spüren*
   and always still can one a discontentment sense
   and still one can sense some discontentment

Separate phrases

- Sentence level chains very rare in learner data:

Es ist [*doch*] [*auch*] *statistisch belegt*
it is indeed also statistically proven
Furthermore, it is indeed statistically proven

- Sentence ADVs before DP-modifying ADVs are not uncommon in learner data:

[*schon*] [[*ziemlich viele*] *Lebenserfahrungen*]
already quite many life-experiences
already quite a lot of life experience

Possible explanations

- Word order in sentence ADVs is variable:

  Doch ist es auch statistisch belegt
  indeed is it also statistically proven

- DP-ADVs cannot be moved or separated:

  * *schon viele ziemlich Lebenserfahrungen*
  already many quite life-experiences

Possible explanations

- Fixed chains have one realization which:
  - covers all occurrences
  - potentially appears more frequently
- Invariable position and unambiguous order facilitate learning
- Topologically flexible elements are less easily acquired or avoided due to uncertainty

Same phrase chains

- Left-headed rare overall (34:10)
- Right-headed common in learners & natives (105:78, e.g. *viel mehr* 'much more')
  > fixed order
  > resemble ADJ intensifiers (*sehr schön* 'very pretty')

- Lexicalized phrases overall more common in natives (122:55), but vary as any lexeme:
  - *(und) so weiter* '(and) so on' overused
  - *schon einmal* 'already' underused

Summary

- Investigation of difficult constructions based on underuse in learners vs. natives
- Strong cases of underuse hypothesized to be connected to surface variability
- Less variable environments show significantly less underuse for same items

Conclusion

- Frequent, fixed surface forms and fixed topological structures promote use and acquisition of constructions in L2 German (cf. Ellis 2002; Cobb 2003; De Cock et al. 1998; Ringbom 1998)
- Conversely variability has a ‘destructive’ effect (cf. restrictedness of Eng. collocations in Nesselhauf 2003)
- Natives embed and fill arguments in these constructions more independently of surface realization and lexical items

Outlook

- No data like more data
- Better theoretical understanding of L1 vs. L2 acquisition processes
- Replication of paradigm in other L2s
- Can variability predict underuse?
- External sources of evidence

References (1/2)

- Clahsen, H. (1984) The acquisition of German word order: a test case for cognitive approaches to L2 development. In: Andersen, R.W. (ed.), *Second Languages*. Rowley, MA: Newbury House, 219–242.
- Cobb, T. (2003) Analyzing late interlanguage with learner corpora: québec replications of three european studies. In: *The Canadian Modern Language Review/La Revue canadienne des langues vivantes* 59(3), 393-423.
- Corder, S.P. (1981) *Error Analysis and Interlanguage*. Oxford: OUP.
- De Cock, S./Granger, S./Leech, G./McEnery, T. (1998) An automated approach to the Phrasicon of EFL learners. In: Granger, S. (ed.), *Learner English on Computer*. London/New York: Addison Wesley Longman, 67-79.
- Diehl, E./Albrecht, H./Zoch, I. (1991) *Lernerstrategien im Fremdsprachenerwerb. Untersuchungen zum Erwerb des deutschen Deklinationssystems*. Tübingen: Niemeyer.
- Ellis, N.C. (2002) Frequency effects in language processing. *Studies in Second Language Acquisition* 24, 143-188.
- Granger, S. (2002) A bird's-eye view of learner corpus research. In: Granger/Hung/Petch-Tyson 2002, 3-33.
- Granger, S. (to appear) Learner Corpora. In: Lüdeling, A./Kytö, M. (eds.), Corpus Linguistics. An International Handbook. Berlin: Mouton de Gruyter.
- Granger, S./Hung, J./Petch-Tyson, S. (eds.) (2002) *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins.

References (2/2)

- Halliday, M.A.K. (1989) *Spoken and Written Language*. Oxford: OUP.
- Laufer, B./Nation, P. (1999) A vocabulary-size test of controlled productive ability. *Language Testing* 16(1), 33-51.
- Lüdeling, A. (2008) Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. In: Grommes, P./Walter, M. (eds.), *Fortgeschrittene Lernervarietäten*. Niemeyer, Tübingen, 119-140.
- Lüdeling, A./Doolittle, S./Hirschmann, H./Schmidt, K./Walter, M. (2008) Das Lernerkorpus Falko. *Deutsch als Fremdsprache* 2/2008.
- Nesselhauf, N. (2003), The use of collocations by advanced learners of English and some implications for teaching. Applied Linguistics 24(2), 223-242.
- Parodi, T. (1998) *Der Erwerb funktionaler Kategorien im Deutschen*. Tübingen: Narr.
- Pravec, N. A. (2002) Survey of learner corpora. *ICAME Journal 26*, 81-114.
- Ringbom, H. (1998) Vocabulary frequencies in advanced learner English: a cross-linguistic approach. In: Granger, S. (ed.), *Learner English on Computer*. London/New York: Addison Wesley Longman, 41-52.
- Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics* 10, 209-231
- Tono, Y. (2003) Learner corpora: design, development, and applications. In: *Pre-conference workshop on learner corpora, Corpus Linguistics 2003, Lancaster.*