Interaction between Colligation, Register and Surface Variability in German Learners and Natives

Humboldt-Universität zu Berlin

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

31. Jahrestagung der DGfS,
Osnabrück, 5 März 2009

Research questions

- Do learners distinguish register?
- If so, how much?
- Similarly or differently to natives?
- What is particularly difficult for learners in the acquisition of registers?

Overview

- Studying learner language
- Operationalizing interlanguage differences quantitatively
- Case study: adverbs and adverb chains in L1 and L2 registers

Data for L2 Studies

- Intuition / introspection (learner or teacher)
- Questionnaires (Diehl et al. 1991)
- Corpus data:
  - Learner corpora (Pravec 2002; Tono 2003; Granger 2008) and comparable L1 corpora
  - Metadata – reference to L2 proficiency, learner's L1...
  - Annotation – pos, lemmatization, possibly error annotation (Corder 1981; Granger 2008)

Working with raw learner data

- Frequencies of word forms, annotated categories, or colligations using both
  - 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** (Selinker 1972; Ringbom 1998; Granger et al. 2002)

Underuse and Overuse

- Simplified model of target register 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)

L1 Independence

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

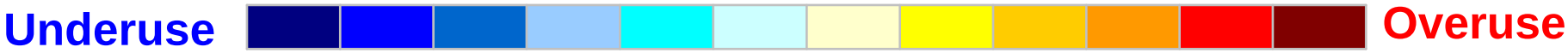
  is beautiful! (Italian pro-drop transfer)

- We are interested in phenomena that apply to GFL learners independently of L1

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

Visualizing Underuse/Overuse

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



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
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	0.006465	0.001434	0.001426
aber	0.006048	0.003347	0.007309	0.006245	0.007315	0.003365	0.003831

Reflexive *sich* ‘self’ is underused

Excel Under/Overuse Addin: http://korpling.german.hu-berlin.de/~amir/uoaddin.htm

Underuse of pos-chains in L2 data

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 generally underused

used tagset: STTS; http://www.sfs.uni-tuebingen.de/Elwis/stts/stts-guide

ADVs in registers and learner language

- ADV-underuse characteristic of advanced learner variety
- Biber 2009: adverb type and token frequencies relevant for measuring register differences
- Independent or interacting factors?

Registers in L2 data?

- "lack of register awareness" (Gilquin/Paquot 2007)
- this predicts:
  - a general underuse of ADVs and ADV chains
  - no significant ADV-differences between registers
- Production of L2 ADV-ADV-chains dependent on (syntactic) complexity

  (Zeldes, Hirschmann & Lüdeling 2008)

Study/approach

a) Comparing ADV-n-grams:
- ADV
- ADV-ADV
- ADV-ADV-ADV

in L1 and L2 data with different registers

b) Comparing different syntactic structures of consecutive ADVs

Corpora for this study

L1	L2
academic theses 1,804,993	
law texts 5,896,940	
Falko Essays L1 67,529	Falko Essays L2 91,112
Falko Summaries L1 21,211	Falko Summaries L2 41,075
parliament debates 36,723,139	

For an introduction to the Falko corpus, Lüdeling et al. 2008

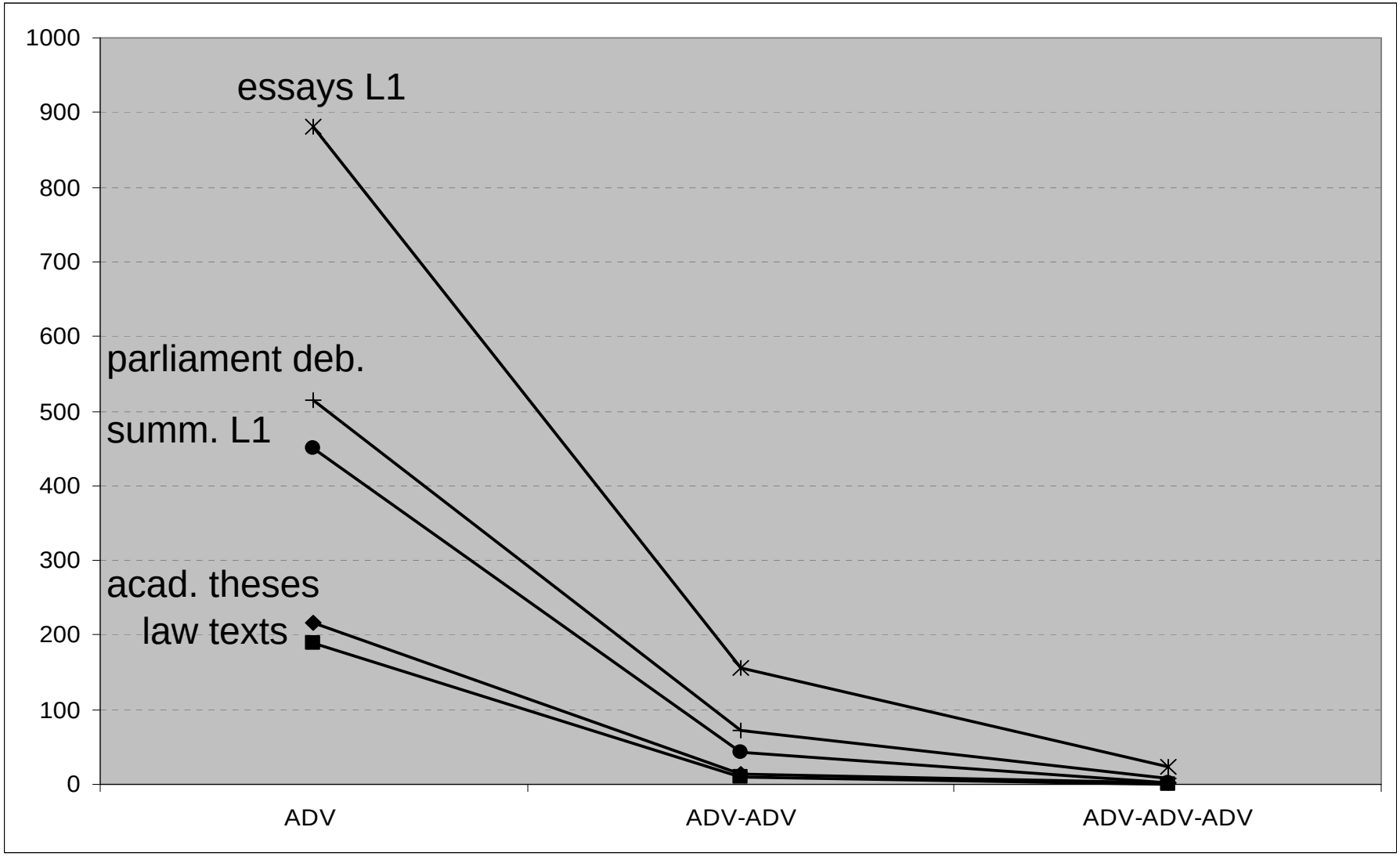
Study a): ADV-n-gram comparison

Raw L1 data:

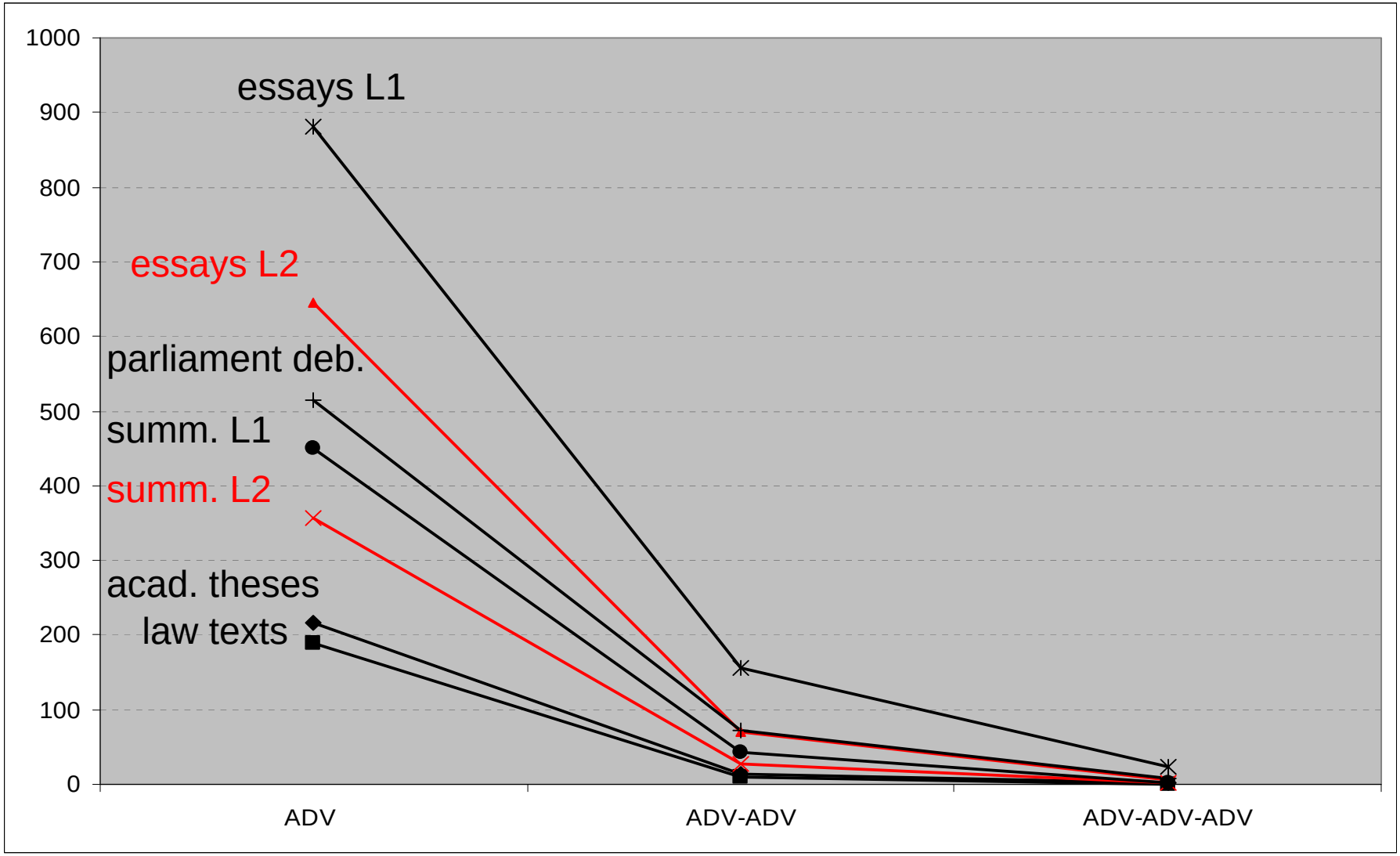
Corpus	**ADV**
Falko Essays L1	881,8
parliament debates	514,8
Falko Summaries L1	450,7
academic theses	215,6
law texts	189,6

numbers normalized to 10,000 tokens

ADV-ADV-chains (L1)



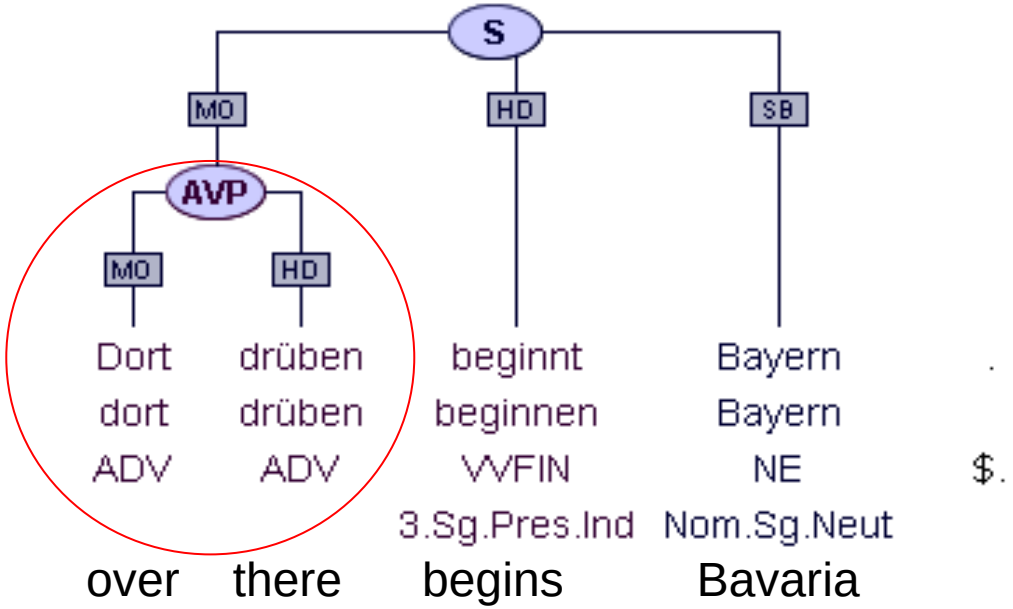
ADV-ADV-chains (incl. L2)



Study b):
Different types of ADV-ADV-bigrams

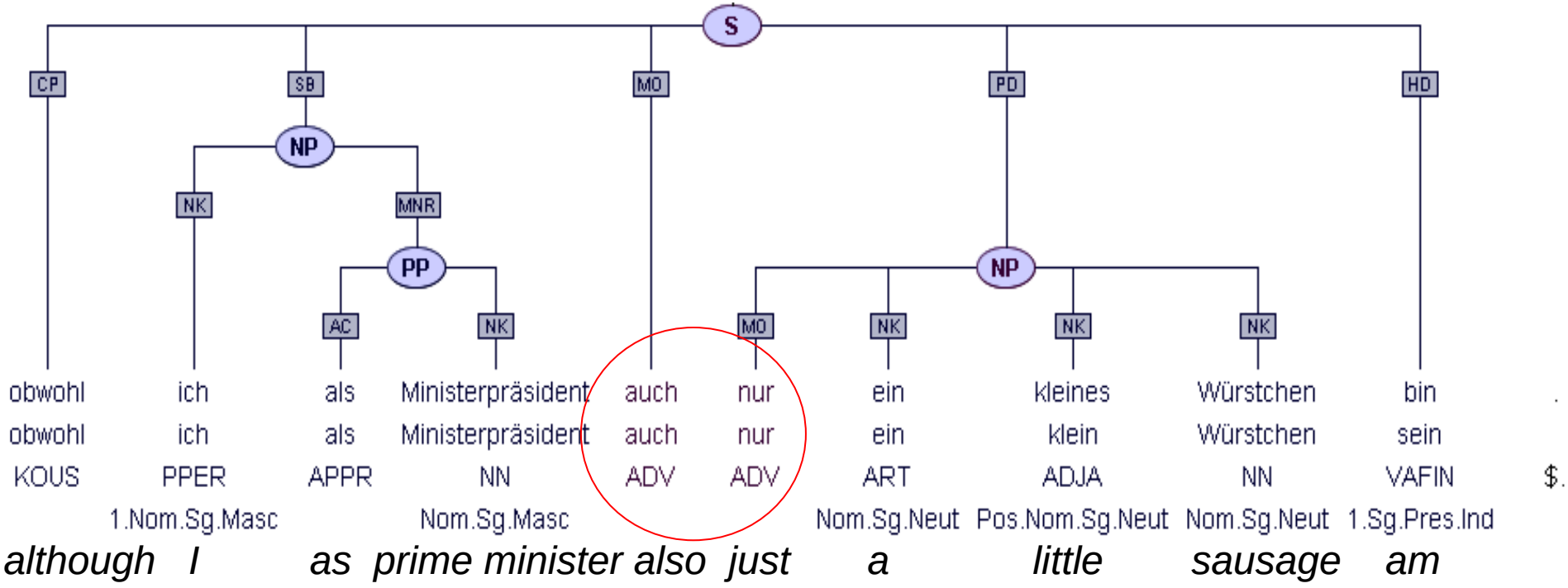
- Method:
  - Syntacticly classify ADV-ADV-bigrams
  - Token frequencies for each class from a Treebank (Tiger)
  - Compare frequencies in L1 & L2 registers

Category [ADV-ADV]



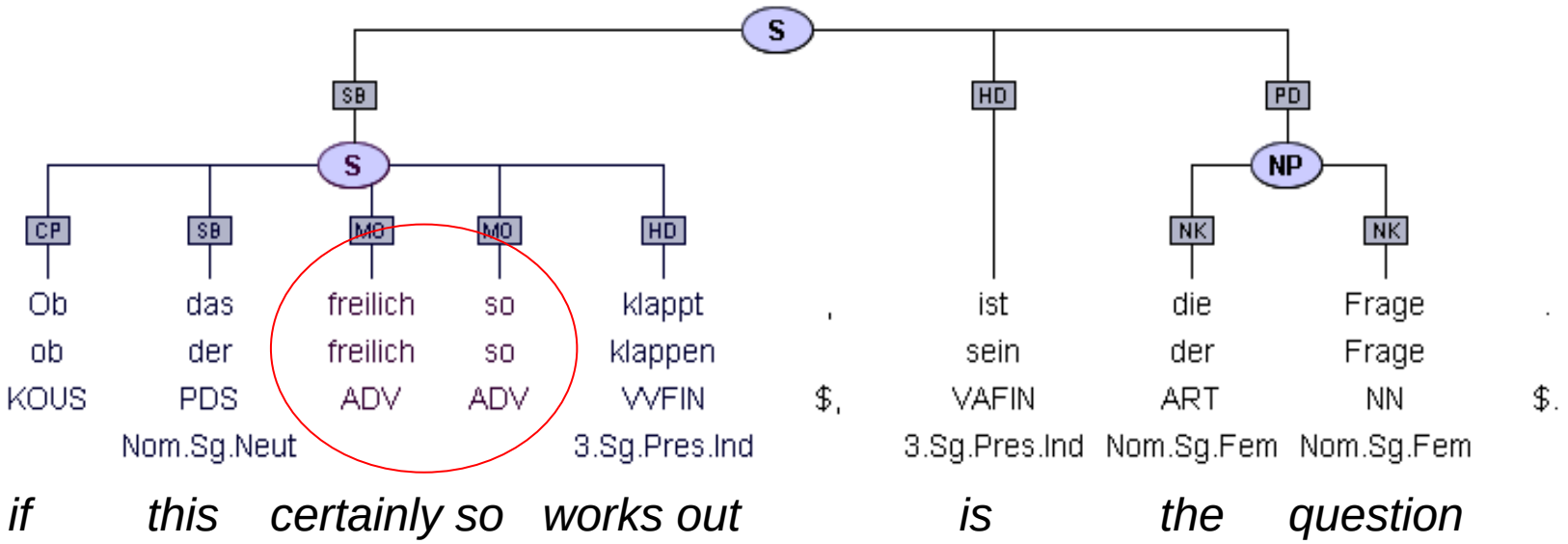
- Least complex category
- Lexicalized pairs (*immer noch* – still) or left headed (***morgen*** *früh* – tomorrow early) or right headed (*sehr* ***bald*** – very soon) AdvPs
- Temporal adverbials, local adverbials

Category [ADV][ADV+X]



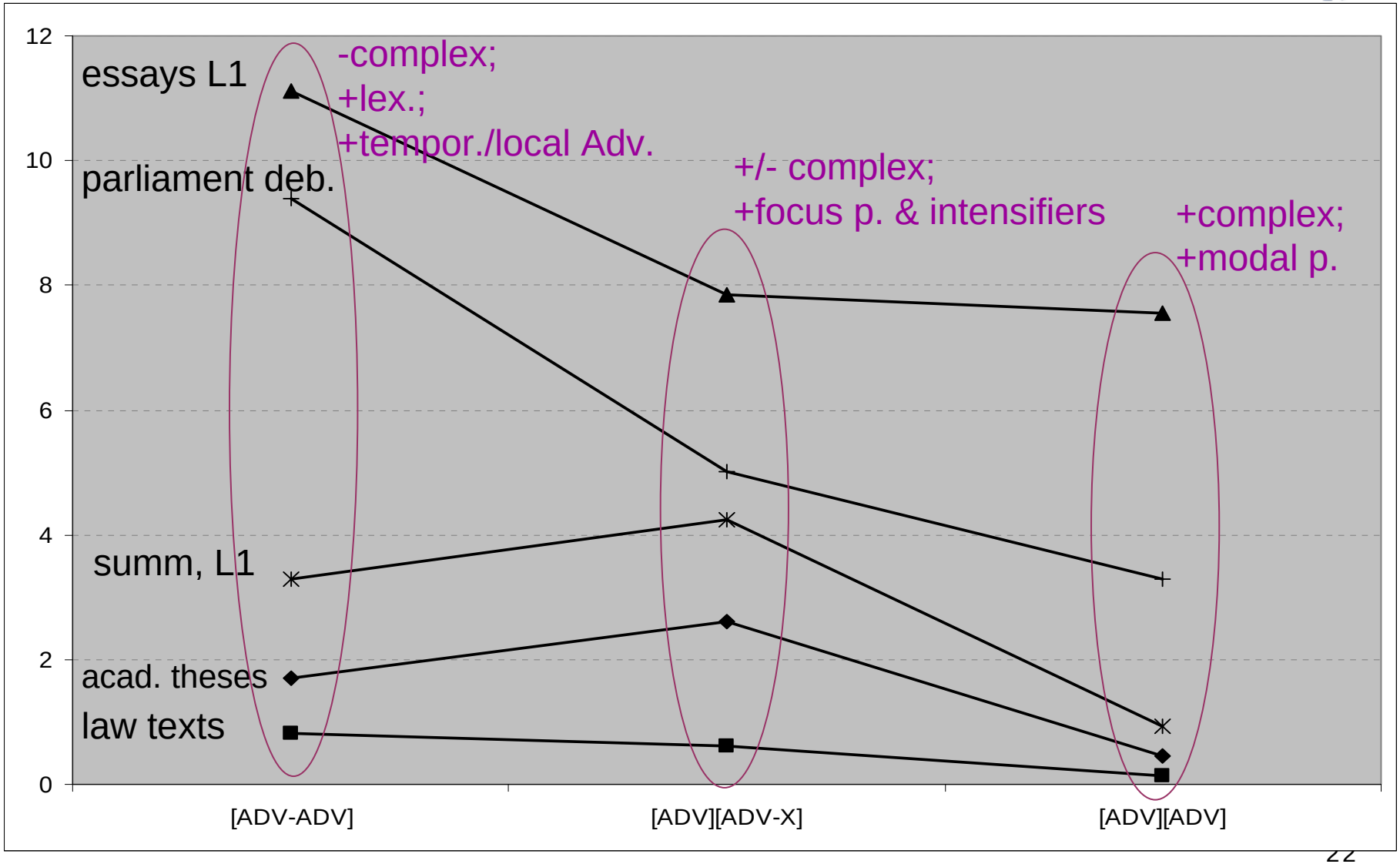
- sentence adverb, (coincidentally) followed by phrase-internal adverb
- More complex category
- Many focus particles & intensifiers

Category [ADV][ADV]

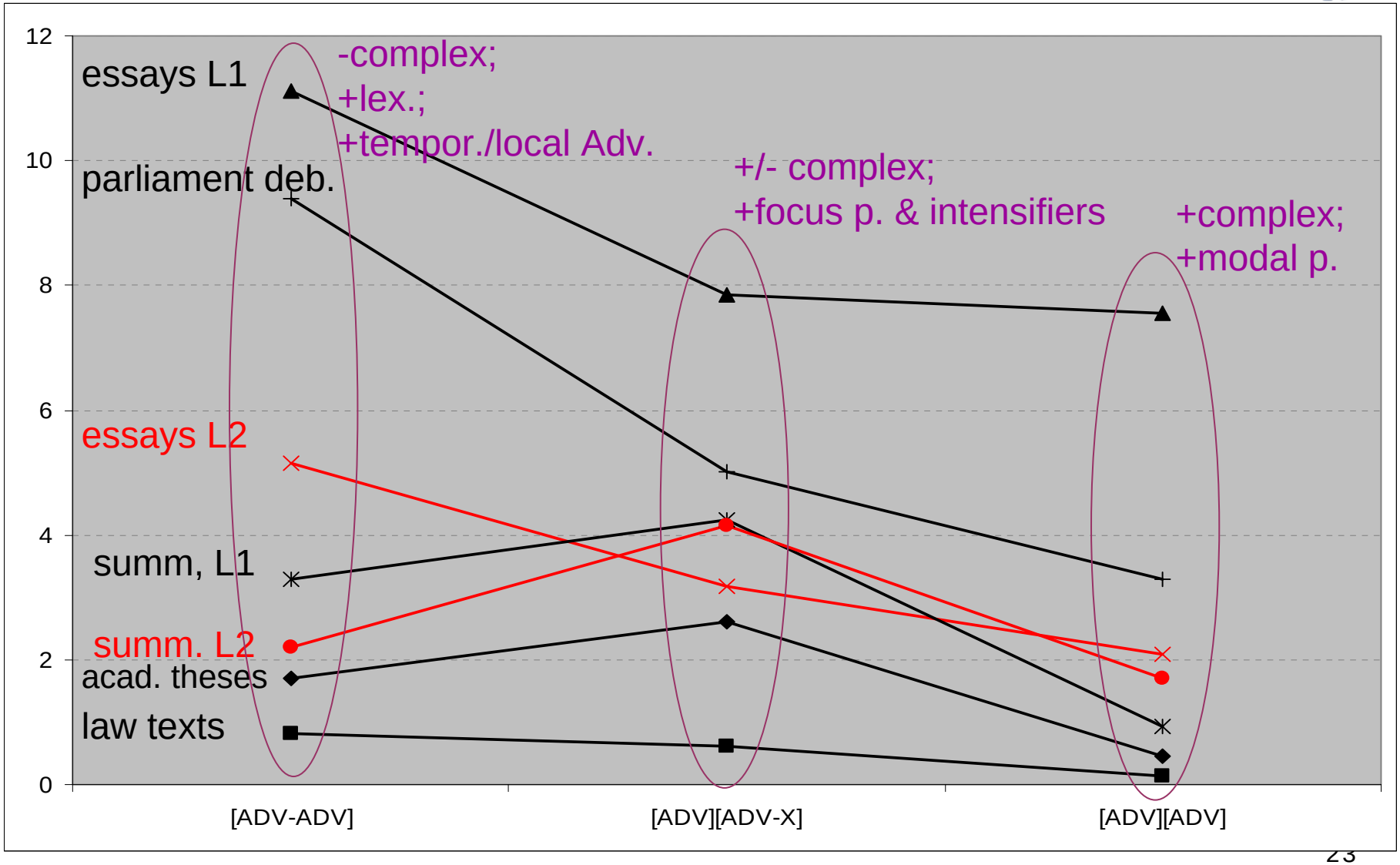


- Independent consecutive sentence adverbs
- Most complex category
- Many modal particles

ADV-ADV-types



ADV-ADV-types



Summary

- Highly significant register dependent ADV use in German L1 texts
- ADV-underuse in L2 data is dependent on register
- ADV-ADV-categories register dependent but do not correlate with underuse
- The generally higher frequencies in L1 essays are more difficult for learners than the lower frequencies in L1 summaries

Outlook

- Behavior of certain individual lexemes and lexeme groups (in progress; *'xxx einmal'*)
- More granularity than STTS offers
- More registers (ideally also spoken data)
- Underuse / overuse beyond surface statistics (syntactic categories, phrase structures)

Thank you!

- Falko is freely available at http://korpling.german.hu-berlin.de/falko/index.jsp

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.
- 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.
- Gilquin, Gaëtanelle & Paquot, Magali (2007), “Spoken Features in Learner Academic Writing: Identification, Explanation and Solution”. In: Proceedings of Corpus Linguistics 2007. Birmingham, UK, 27-30 July, 2007.
- Granger, S. (2008) Learner Corpora. In: Lüdeling, A./Kytö, M. (eds.), Corpus Linguistics. An International Handbook. Berlin: Mouton de Gruyter, 259-275.
- 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)

- Gries, Stefan Th. & Wulff, Stefanie (2005), “Do Foreign Language Learners also Have Constructions? Evidence from Priming, Sorting, and Corpora”. Annual Review of Cognitive Linguistics 3, 182-200.
- Halliday, M.A.K. (1989) *Spoken and Written Language*. Oxford: OUP.
- 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.*

Examples from learner data (Falko)

1. *und* [*immer noch*] *kann man eine*
   and still can one an
   unzufriedenheit spüren
   dissatisfaction feel
2. *muss man* [*eigentlich*] [*nur bis ungefähr*
   must one actually only till about
   achtzehn] *uberleben*
   eighteen survive
3. *Es ist* [*doch*] [*auch*] *statistisch belegt*
   it is also statistically proven

Error annotation and register

- Some learner data has obvious errors:

  Je viel liest, desto mehr weißt (usb013_2006_10)
  The much read, the more know
- Error analysis hard to apply to register:
  Es kommen auch Leute nach Skandinavien nur um dort "vom Staat" zu leben. Das tolle "Staats-model" hat sich herumgesprochen, und jetzt haben die Skandinavier ein Problem. (hu012_2006_09)
  People come to Scandinavia too, just to “live off the state”. Word of the cool “state-model” has gotten out, and now the Scandinavians have a problem.