

Underuse of Syntactic Categories in Falko

-

A Case Study on Modification

Hagen Hirschmann

Anke Lüdeling

Ines Rehbein

Marc Reznicek

Amir Zeldes

LEARNER CORPUS RESEARCH 2011

LOUVAIN-LA-NEUVE


research questions & approach

- how can syntactic analyses of L2 learner data help in understanding interlanguage/acquisition processes?
- what is the relationship between lexical elements and syntactic classes?
 - phenomenon: modification
 - data: dependency-parsed corpus of advanced L2 learners of German
 - CIA study (underuse statistics)



- freely available annotated learner corpus of German as a foreign language
- advanced learners (tutored acquisition)
- written language / controlled, unaided writing
- several text types (sub-corpora);
here essays (ca. 130000 tokens)
- comparable native speaker corpora (ca. 70000 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>

annotations in Falko

- standoff format (token annotation, span annotation, graphs, pointers etc.), annotation layers can be freely added (Lüdeling et al. 2005)
 - learner utterance
 - pos & lemma (automatic, manual correction)
(TreeTagger, Schmid 1994)
 - **target hypotheses** (manual, as many as necessary)
 - pos & lemma
 - error annotation (automatic)
 - parses (dependencies; automatic, manual correction)
 - manual error annotation of some phenomena
 - ...
- 

annotation of learner data: conceptual issues

- annotation of learner data is highly problematic
 - data is not systematic according to L1 grammar (especially if there are different L1s)
 - difficult for automatic tools (taggers, parsers)
 - for error analysis and contrastive interlanguage analysis: data has to be interpreted
- Corder (1981), Izumi/Uchimoto/Isahara (2005), Tenfjord/Hagen/Johansen (2004), Diaz-Negrillo et al. (2010) etc.

conceptual problems: pos

- word forms in L2 data sometimes correspond to different pos (Diaz-Negrillo et al. 2010)

Most	important	of	all	was	the	conscious	that
RBS	JJ	IN	DT	VBD	DT	JJ	IN/that
most	important	of	all	be	the	conscious	that

(ICLE)

- every assignment of a pos is an interpretation (*conscious*/NN?JJ → *consciousness*/NN)

conceptual problems: syntax

Most	important	of	all	was	the	conscious	that
RBS	JJ	IN	DT	VBD	DT	JJ	IN/that
most	important	of	all	be	the	conscious	that

- no possible/useful parse of this structure
 - utterance must be transformed into a canonical structure (Hirschmann et al. 2007)
- target hypothesis

parsing approach: target hypotheses

word	Most	important	of	all	was	the	conscious	that
POS	JJ	IN	DT	DT	VBD	DT	JJ	IN/that
lemma	most	important	of	all	be	the	conscious	that

- note: conflicting th may be formulated:

word	Most	important	of	all	was	the	conscious	thought	that
POS	JJ	IN	DT	DT	VBD	DT	JJ	NN	IN/that
lemma	most	important	of	all	be	the	conscious	thought	that
TH	Most	important	of	all	was	the	conscious	thought	that
TH_Diff								INS	
TH_POS	JJ	IN	DT	DT	VBD	DT	JJ	NN	IN/that

annotation of learner data: target hypothesis in Falko

- th1: sentence-based, very close to original text, mainly ‚genuine‘ 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.)
- (Lüdeling 2011, Reznicek et al. submitted)

target hypotheses ...

- are just as necessary for L1 data, btw

research question

- we want to find **structural** features/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

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

underuse

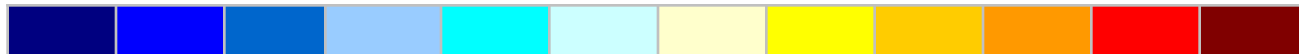
- L2 distributions are compared to L1 distributions
- overuse, underuse are defined as (statistically significant) differences between the varieties
- a category can be underused in L2 because
 - the learners do not know it
 - the learners do know it but (unconsciously) avoid it
 - a diagnostics for detecting structural acquisition problems

visualization of overuse and underuse

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

Underuse

Overuse



- Excel add in by Amir Zeldes available at <http://korpling.german.hu-berlin.de/~amir/uoadaddin.htm>

visualization of overuse and underuse: lexical categories

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	0.005465	0.001434	0.001426
aber	0.006048	0.003347	0.007309	0.006245	0.007315	0.003365	0.003831

sich (reflexive pronoun) is underused in all L1 groups

visualization of overuse and underuse: bigrams of pos-categories

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

adverb chains are underused in all L1 groups

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

- 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?)

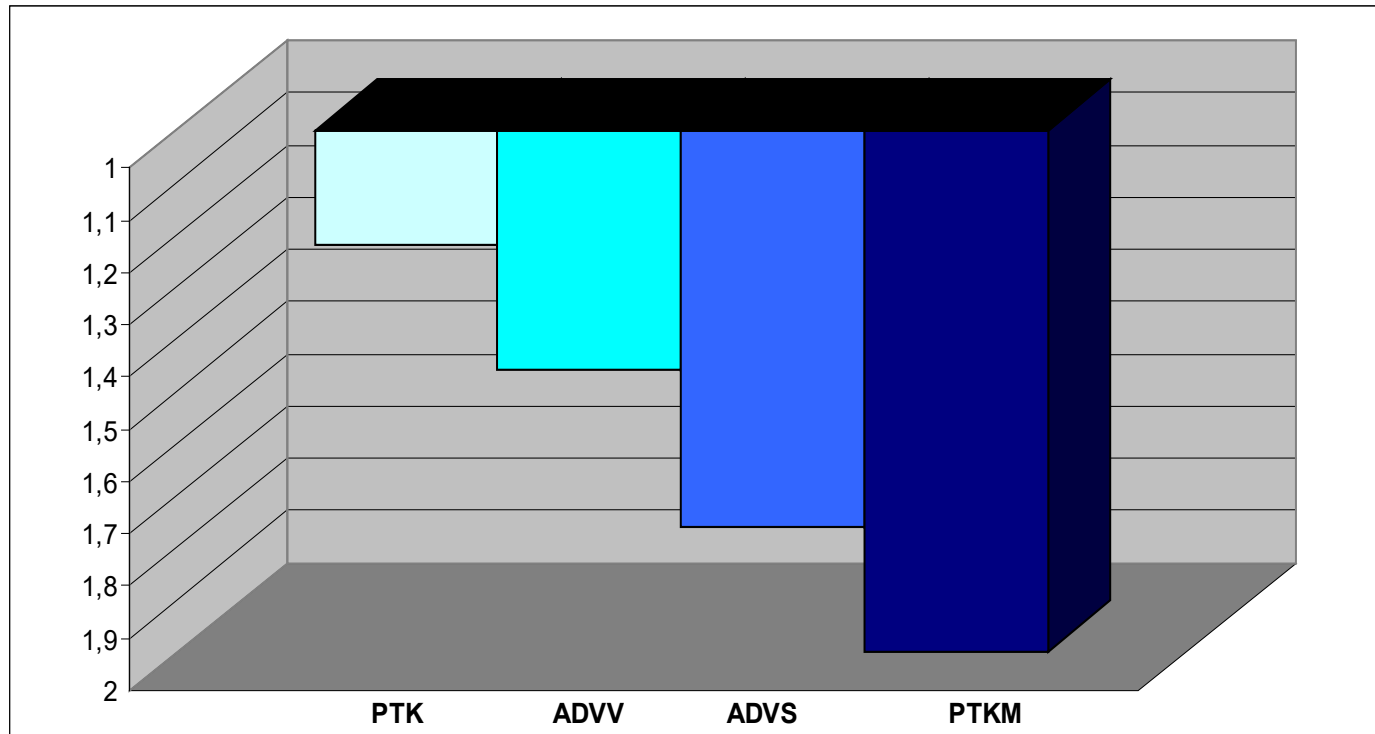
modification

- are the effects form-based or function-based?
 - are all adverbs underused?
no; *auch, noch* etc. overused
 - are certain adverbs (forms) underused?
yes
 - are certain adverbial functions underused?
 - are certain adverbs (forms) underused in certain functions?
 - 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, 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 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*)

underuse of adverbs: function

- underuse differences between different adverbial functions
- but classification still word based
- compensation strategies?
- necessity to code syntactic functions independent of filler category

Falko – syntactic annotation

- **target hypothesis1** of Falko L1 and L2 corpora
- manually corrected pos tags
- semi-automatic sentence segmentation
- dependency parser by 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

The screenshot displays the ANNIS² interface. On the left is the 'Search Form' with the following fields:

- AnnisQL: POS="VVFİN" & POS="APPR" & POS="ADV" & #1 ->dep #2 & #2 ->dep #3
- Query Builder: Show >>
- Result: 110
- History: Query History

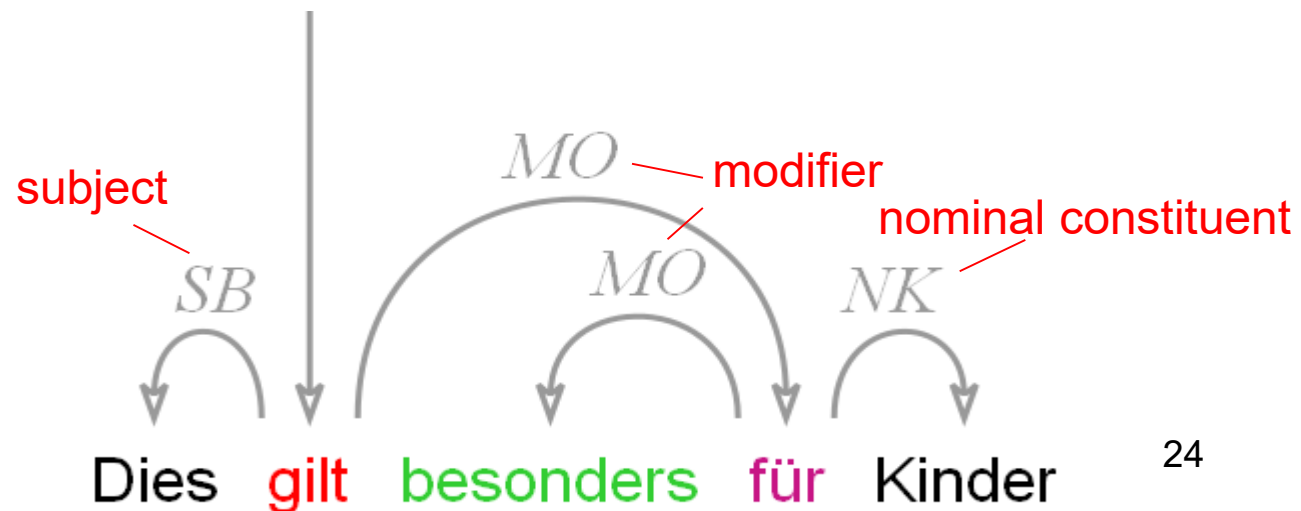
Below the search form is a table for 'More Corpora':

Name	Texts	Tokens
<input type="checkbox"/> I1_0509_2	94	68940
<input checked="" type="checkbox"/> I2_0609	248	124524

On the right is the 'Search Result' for the query. It shows the sentence: Dies gilt besonders für Kinder. The words are color-coded: 'gilt' is red, 'besonders' is green, and 'für' is blue. Below the sentence is a dependency parse diagram with arcs labeled with syntactic functions: SB (Subject) from 'Dies' to 'gilt', MO (Modifier) from 'gilt' to 'besonders', MO from 'besonders' to 'für', and NK (Noun Kernel) from 'für' to 'Kinder'. The diagram also shows morphological and syntactic information for each token, such as 'Nom|Sg|Neut 3|Sg|Pres|Ind' for 'gilt' and 'Acc|Pl|Masc' for 'Kinder'.

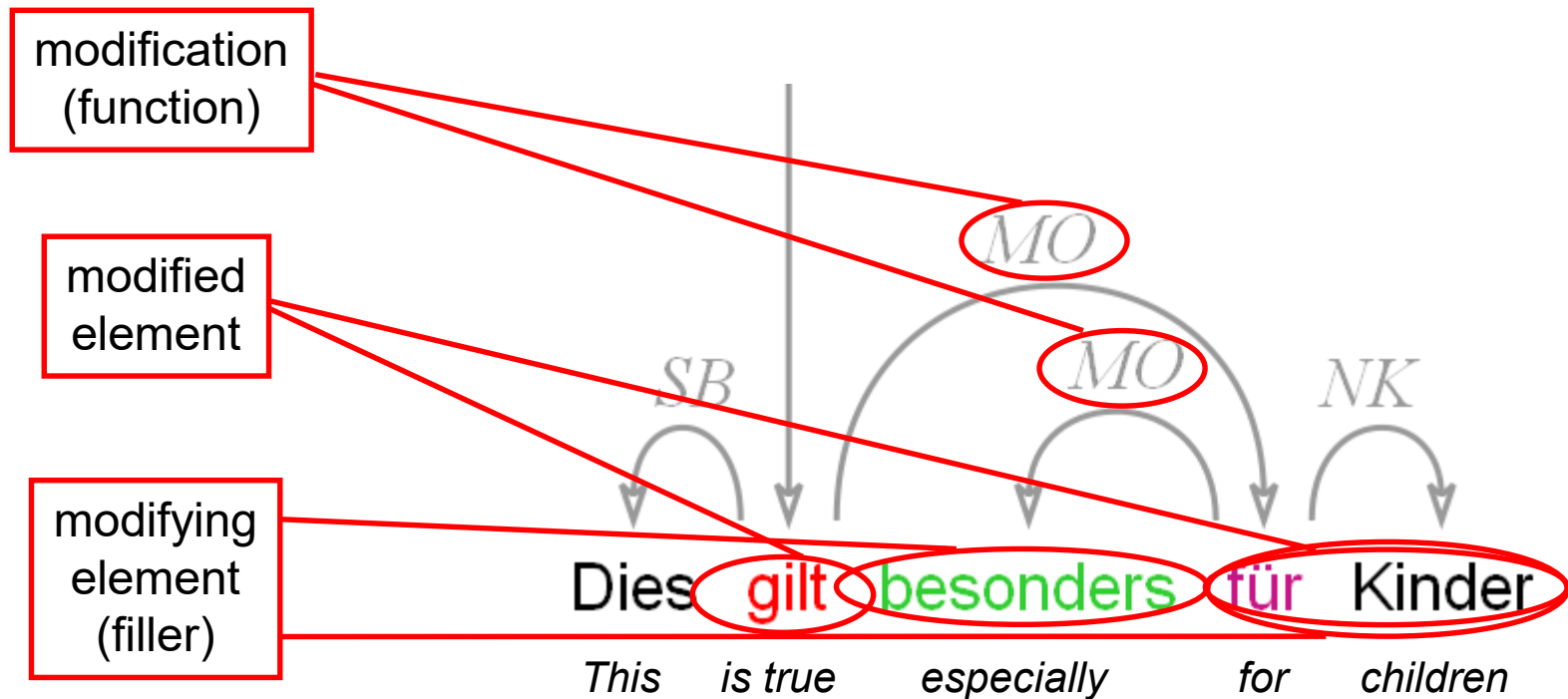
syntax schema (very briefly)

- every word is connected with its dependent(s)
- arrows point to hierarchically lower dependent
- each arrow (dependency) has a function label

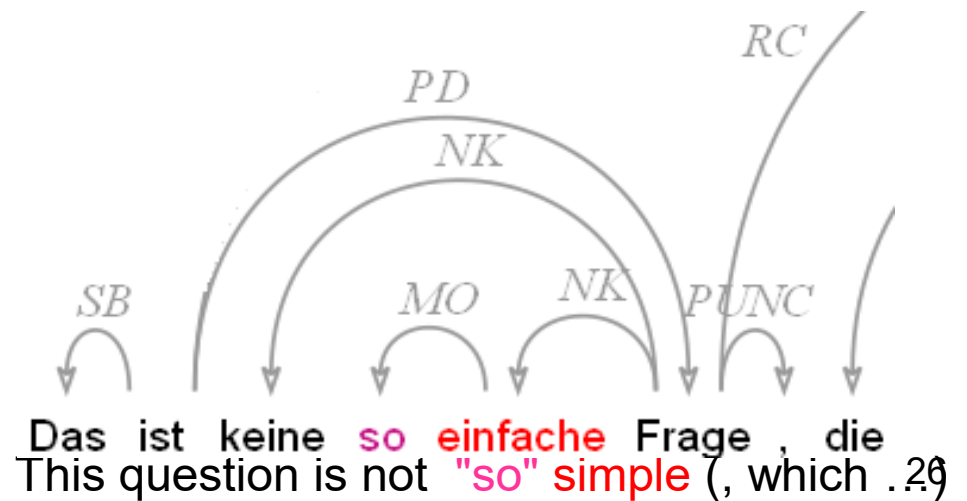
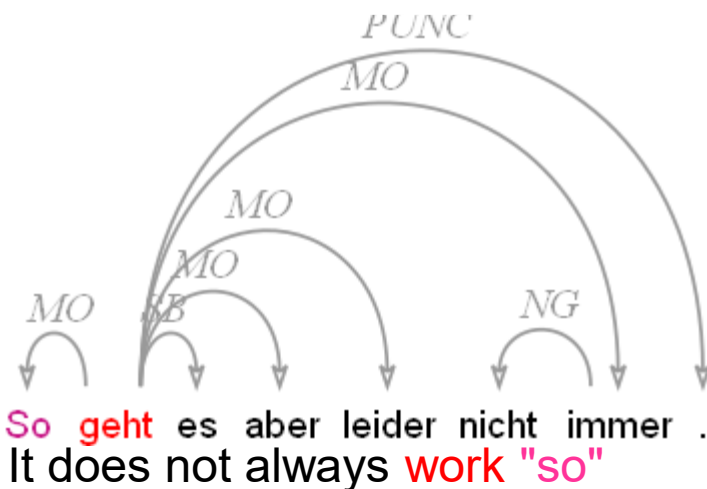
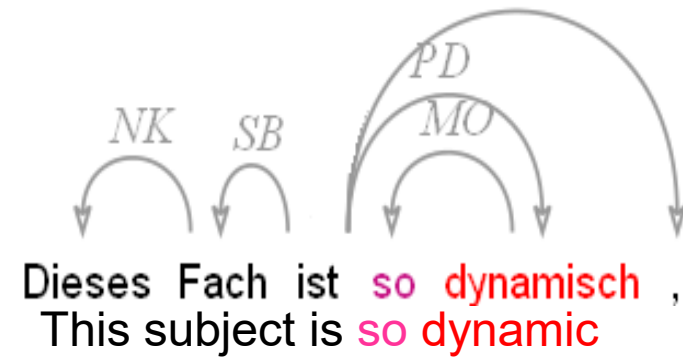


searching for modification in Falko

- 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?



polyfunctional lexemes: **so**



modification

- are the 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?
 - is modification generally underused?
(or do learners make up for the underuse of adverbs by other means of modification?)

overuse / underuse of syntactic functions

label	de	da	en	fr	ru	usb
NK	0,264067	0,278546	0,284881	0,303271	0,29552	0,295136
HD	0,156192	0,155622	0,157178	0,154275	0,15809	0,156483
MO	0,141968	0,12789	0,113704	0,110112	0,112513	0,108707
SB	0,07398	0,078506	0,077099	0,075093	0,078852	0,085512
CJ	0,059604	0,053397	0,056411	0,050632	0,059274	0,072183
AC	0,057051	0,059317	0,057215	0,054796	0,054012	0,04916
OC	0,050335	0,053039	0,050008	0,049888	0,047125	0,040679
OA	0,044213	0,042352	0,044097	0,043643	0,046119	0,046218
CD	0,026549	0,024632	0,025639	0,022156	0,024917	0,030466
CP	0,017653	0,021732	0,020325	0,018141	0,017256	0,014887
PD	0,014435	0,014462	0,015943	0,015019	0,016947	0,018002
NG	0,011065	0,011561	0,010914	0,00974	0,00975	0,011252
MNR	0,010995	0,013707	0,013429	0,013383	0,010679	0,009521
RC	0,010051	0,008979	0,009385	0,011375	0,006268	0,005366

overuse / underuse of syntactic functions – significant results

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

MO (modification) is significantly underused independent of L1

modified element

func	L2 (norm)	L1 (norm)	
V	117,635562	139,407446	In my opinion this statement holds .
ADJ	11,8629809	14,5772595	the often very theoretical approach
PREP	4,24891865	6,05986598	especially in Denmark where ...
PROADV	0,08497837	0,15264146	...and exactly for this reason ...
NEG	1,22368857	2,57964068	Perhaps not when
ADV	2,85527333	5,08296063	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

modifiers

func	L2 (norm)	L1 (norm)	
V	14,6162802	12,8218827	If she makes her career, ...
PROADV	7,41011413	6,73148841	Some have success [with this] ...
COMPARE	0,26343296	0,27475463	One can, as mentioned above ...
PREP	44,8600831	48,5857769	To make money on a criminal basis
ADJ	12,7722495	17,5842962	... criminality increases steadily ...
ADV	61,8302642	87,7230473	which still exists ...

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

- are the effects form-based or function-based?
 - are all adverbs underused?
no; *auch, noch* etc. overused
 - are certain adverbs (forms) underused?
yes
 - are certain adverbial functions underused?
yes
 - are certain adverbs (forms) underused in certain functions?
yes
 - is modification generally underused?
(or do learners make up for the underuse of adverbs by other means of modification?)
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 L1 of the learners

methodological conclusions

- in annotation **separation of form and function** necessary
- **parsing** of learner data necessary to find syntactic functions
- **explicit target hypotheses**: making interpretation visible and learner language parsable
- **multi-layer architectures**

Thank you!

Merci!

Danke!

Falko:

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

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