Eine korpusbasierte Analyse von Adverbkategorien in Texten fortgeschrittener Lerner des DaF

16.10.2009
Lernerkorpustreffen
Tübingen-Berlin

Hagen Hirschmann

Gliederung

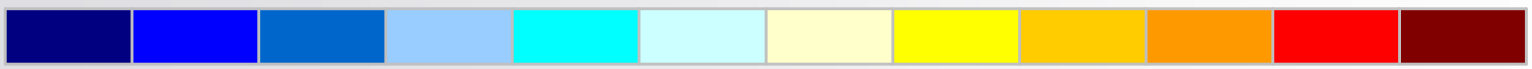
- ADV-Underuse in Falko L2
- Arbeitshypothesen
- das STTS-Tag "ADV" - Problemstellung
- syntaktische Ausdifferenzierung von "ADV"
- erste Ergebnisse

Contrastive Interlanguage Analysis (CIA)

- allgemein: Unterschiede zwischen Varietäten können durch Frequenzunterschiede ausgedrückt werden (u.a. Baayen 2001; Biber 2006)
- Over- und **Underuse**-Studien
  - (Underuse→Lernschwierigkeiten?)
- Frequenzvergleiche von bspw.
  - individuellen Lexemen (Inhalts- oder Funktionswörter) → Prototyp
    (→ semantische Kategorien)
  - heute: **(morpho-)syntaktische Kategorien**: Wortarten(-ketten)

CIA in Falko

- CIA und Underuse-Hypothese kann auf alle zählbaren Kategorien angewendet werden.
- Vergleich normalisierter Frequenzen von pos-n-Grammen zwischen Falko-Subkorpora (L2 unterschiedlicher MS sowie L1)
- Visualisierungsmethode Stärke des Over- bzw. Underuses wird farbkodiert (Amir Zeldes)



Vergleich der Frequenzen von pos-Ketten – Bsp: Bigramme

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

Aufeinanderfolgende ADVs werden von den Lernern unabhängig von ihrer L1 mindergebraucht.

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

Zusammenfassung (ADVs)

- Wenig individuelle Lexeme mit "ADV" werden von den Lernern zu häufig verwendet.
- "ADV" (STTS) am meisten mindergebrauchte Wortart
- Ketten von "ADV" (Bigramme und Trigramme) signifikant mehr mindergebraucht, als Unigramme vorhersagen
- Variable (fakultative) Elemente der Syntax
- ???

Zwischenüberlegungen

- "ADV"(-Ketten) → sehr unterschiedliche syntaktische Strukturen/Klassen verschiedener Komplexität
- Hypothese:

  Nur bestimmte ("schwierige") Einheiten bzw. Kombinationen werden mindergebraucht.
- Welches sind diese?

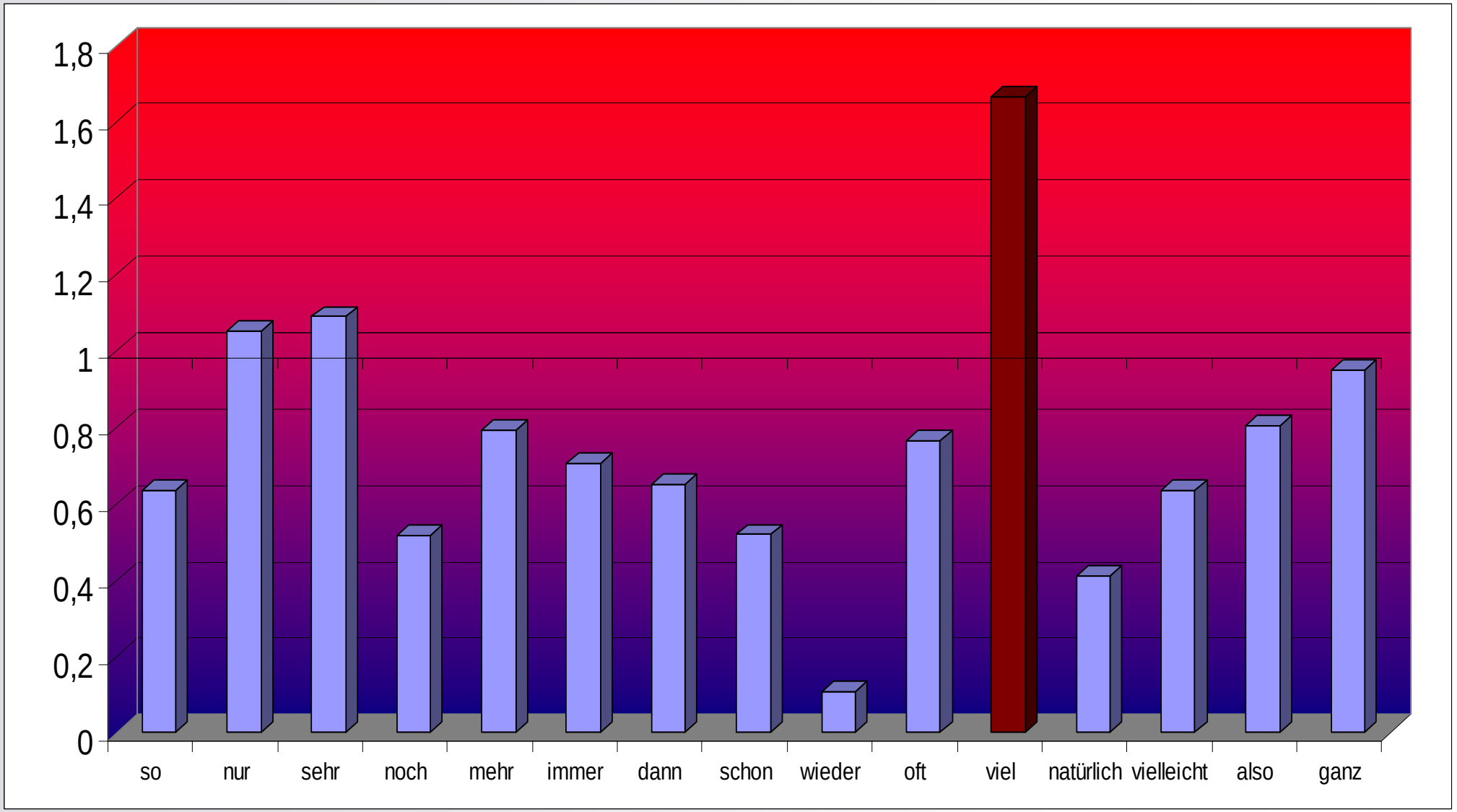
ADV-Studien

- unterschiedliche Auffassungen über Adverb-Subklassen; kaum systematische Klassifikationen, kaum konsequent syntaktische
- viele Arbeiten zu Adverbien, Partikeln, Modalpartikeln in DaF → Tradition: MP als Lernproblem (Zimmermann 1981)
- Korpusstudien ausschließlich zu individuellen Lexemen exempl.: "*I used the (…) tool to find each instance of each of the four focal words (ja, mal, denn, doch) with the potential of having an MP meaning*" (Vyatkina 2007)
- bedingt legitim bei der Untersuchung syntaktischer Kategorien (selbst MP keine geschlossene Klasse)

Bsp. lexikalische ADV-Studie

- "*viel*" (ADV!) in ADV-Frequenzliste das erste signifikant überrepräsentierte (#11)
- H1: Die Lerner greifen auf einen Default-Intensivierer zurück.
- H2: Die Lerner verwenden mehr Intensivierungen.
- [word="(*deutlich*|*drastisch*)"]. [pos="ADJ." & word=".*er(e[srmn]?)?"] stark mindergebraucht
- Favorisierung H1
- Entsprechende Schlussfolgerungen für Lehre…

Bsp. lexikalische ADV-Studie



Bsp. lexikalische ADV-Studie

- Aber:
  - Fazit "eingeschränkter Wortschatz" ist nicht unmittelbar interessant.
  - Wir können nicht alle Lexeme zählen…
  - Wir können deshalb keine statistisch valide Aussage über eine syntaktische Klasse machen.

STTS: "ADV"

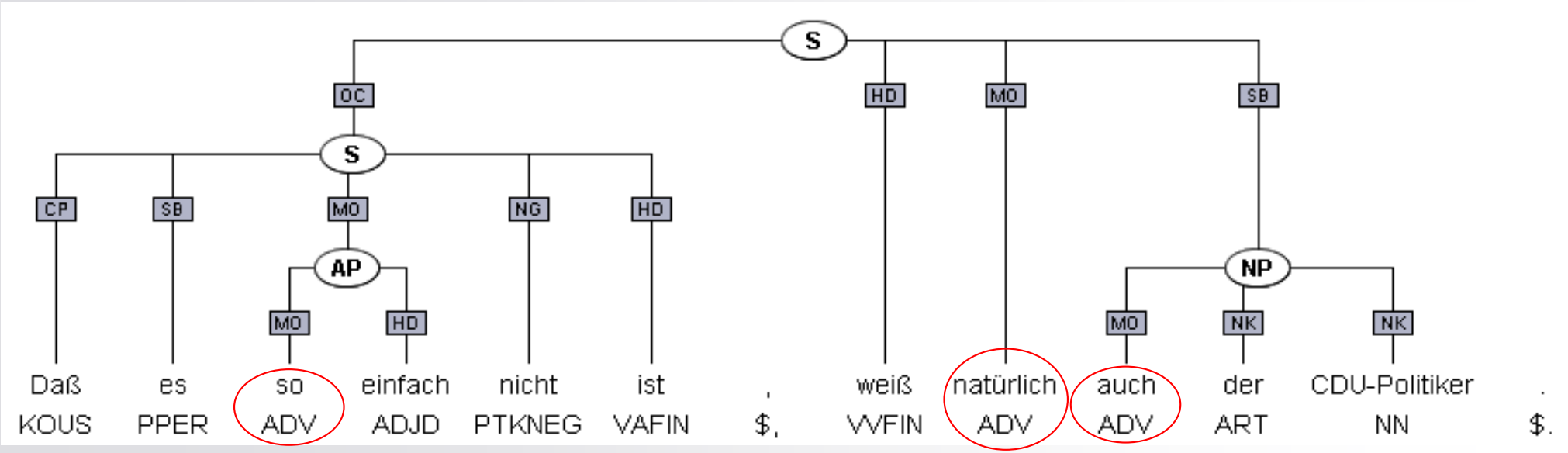
- pragmatische Lösung: systematische Taggerentscheidungen werden akzeptiert
- verschiedene 'streitbare' Festlegungen; z.B.: adverbial gebrauchte partizipiale Adjektive nicht "ADJD" (o. "ADV"), sondern "VVPP" (*gekonnt gemacht*)
- Subsumierung unflektierbarer Modifikatoren (unterschiedlichste syntaktische Eigenschaften)
- eher eine flexionsmorphologische Klasse als eine syntaktische (generelles Kategorisierungsproblem)

Bsp. Tigerkorpus

(http://www.ims.uni-stuttgart.de/projekte/TIGER/)

- *Daß es so einfach nicht ist , weiß natürlich auch der CDU-Politiker .*

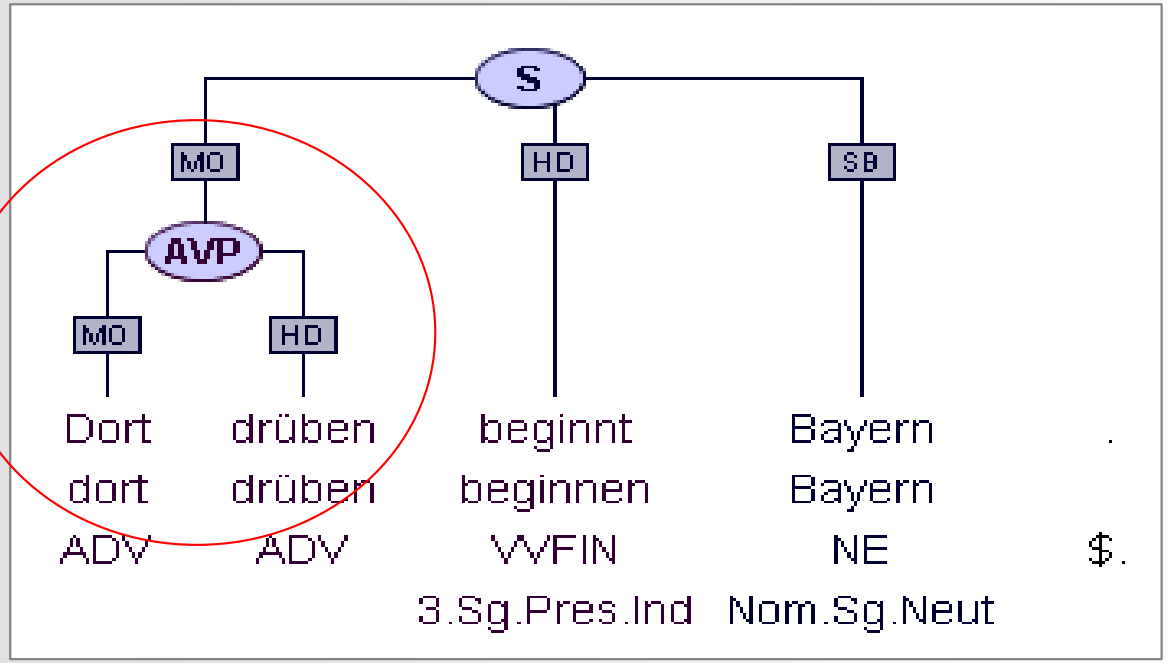
  (Tiger Release 07, S. 639)



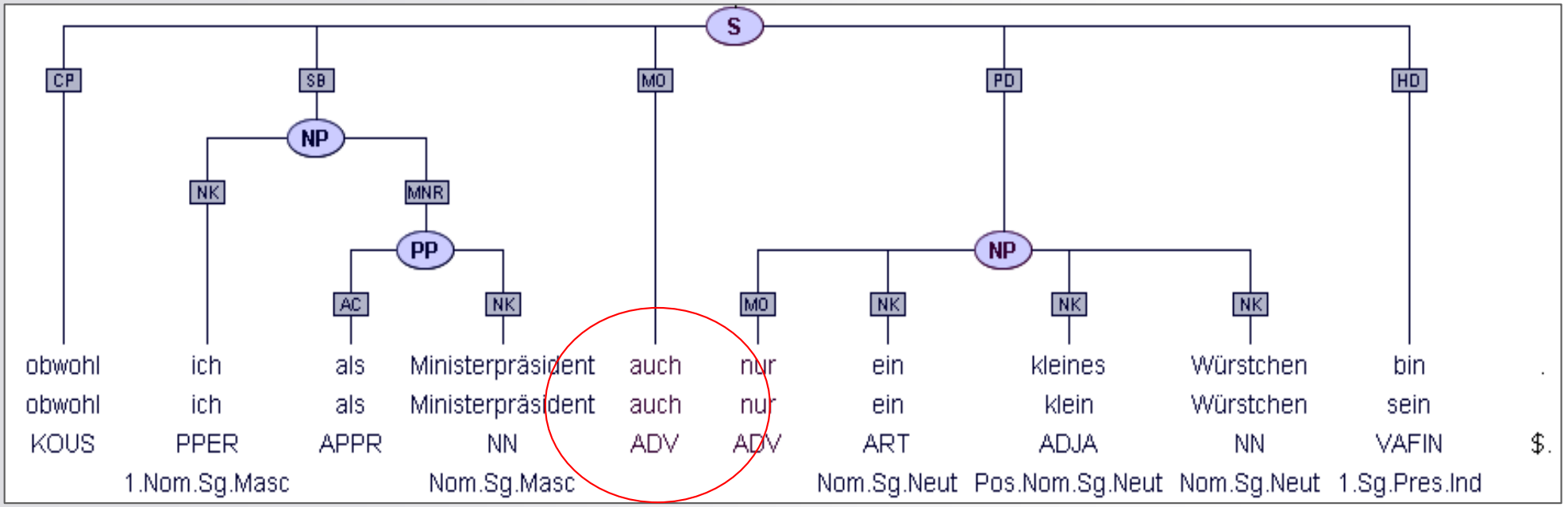
Komplexitätsstufen bei Bigrammen

- 3 (grobe) Komplexitätsstufen von ADV-ADV ohne Probleme in Tiger suchbar:

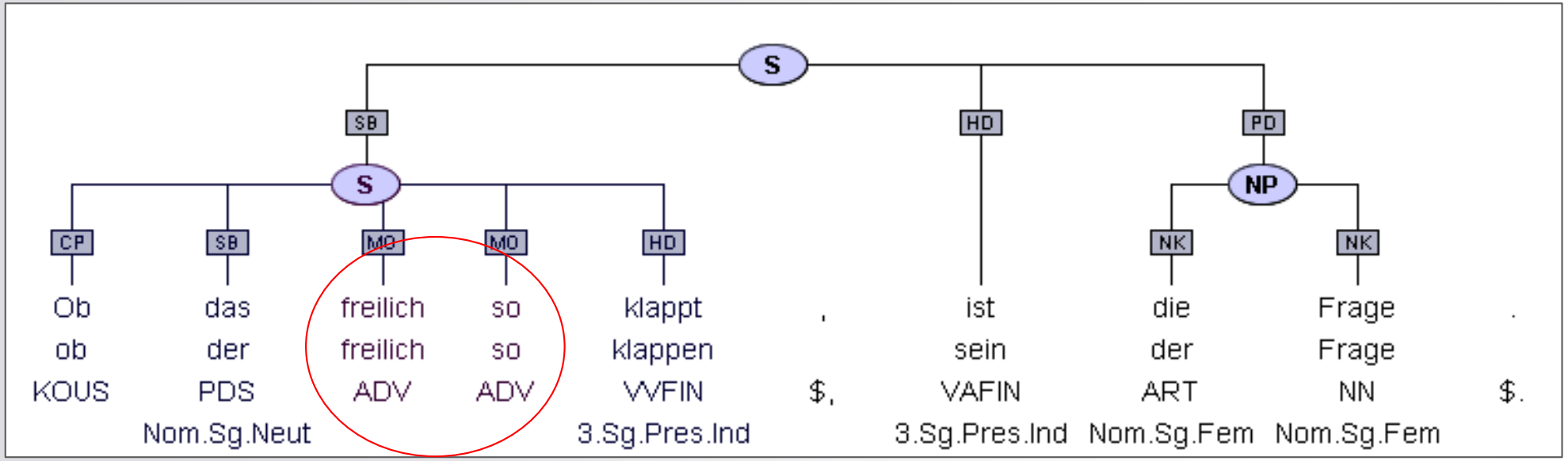
1. [ADV-ADV]


S
MO
HD
SB
AVP
MO
HD
Dort
drüben
beginnt
Bayern
.
dort
drüben
beginnen
Bayern
ADV
ADV
VVFIN
NE
$.
3.Sg.Pres.Ind
Nom.Sg.Neut

2. [ADV][ADV+X]


S
CP
SB
MO
PD
HD
NP
NK
MNR
PP
AC
NK
MO
NK
NK
NK
obwohl obwohl KOUS
ich ich PPER
als als APPR
Ministerpräsident Ministerpräsident NN
auch auch ADV
nur nur ADV
ein ein ART
kleines klein ADJA
Würstchen Würstchen NN
bin sein VAFIN
. $.
1.Nom.Sg.Masc
Nom.Sg.Masc
Nom.Sg.Neut
Pos.Nom.Sg.Neut
Nom.Sg.Neut
1.Sg.Pres.Ind

3. [ADV][ADV]


S
SB
HD
PD
S
NP
CP
SB
MO
MO
HD
NK
NK
Ob
das
freilich
so
klappt
,
ist
die
Frage
.
ob
der
freilich
so
klappen
sein
der
Frage
KOUS
PDS
ADV
ADV
VVFIN
$,
VAFIN
ART
NN
$.
Nom.Sg.Neut
3.Sg.Pres.Ind
3.Sg.Pres.Ind
Nom.Sg.Fem
Nom.Sg.Fem

ADV-Klassifikation für nicht tief annotierte Korpora

- Unterschiede im Mindergebrauch dieser Klassen?
- Wie können diese syntaktischen Klassen möglichst effizient in Falko annotiert werden?
- (automatisches Parsing (noch?) nicht erfolgreich)

ADV-Klassifikation für nicht tief annotierte Korpora

- zu klassifizieren: Einheiten, die gemeinhin als Adverbien, Partikeln, Modalpartikeln bezeichnet werden
- höchst widersprüchliche Definitionen in unterschiedlichen Grammatiken ("Modalwort"-"Diskurspartikel"-"Modalpartikel"-"Satzpartikel" usw.)
- unterschiedlich motivierte Klassen ("*Fokus*partikel"-"*Steigerung*spartikel"-"*Abtönung*spartikel" usw.)
- hier: *syntaktische* Klassifikation

syntaktische Subkategorisierung – ein 2-Ebenen-Ansatz

- Anspruch:
  - Umsetzung *eines syntaktischen* Konsens
  - rein syntaktische Kategorien
  - Annotation gut operationalisierbar
  - DaF-Relevanz
- 2 Annotationsebenen:
  - Wortart (grob, aber einheitlich)
  - syntaktische Funktion (funktionale Ausdifferenzierung; **ist pos-universell**)

syntaktische Adverbklassen

- Grundlage: weitester Adverbbegriff (flexionsmorphologische Klasse; "alle unflektierbaren Modifizierer")
- ! aktueller Äußerungskontext eines Lexems
- Klassifikationskriterien:
  - topologischer Rahmen (wo kann das Lexem im Satz platziert werden?)
  - Modifikationsbereich/Skopus (ausdrückbar bspw. in Phrasenstrukturmodellen)

syntaktische Adverbklassen

- Unter diesen Voraussetzungen nur 3 Oberklassen (Wortartenklassen):
  - "Adverb" (Satz- oder VP-modifizierend)
  - "Partikel
  - "Modalpartikel"

ADV-Klassen

"ADV" (STTS)

vorfeldfähig
phrasenbildend
→Adverb
→ "ADV"

nicht vorfeldfähig
nicht phrasenbildend
phrasengebunden
→Partikel
→ "PTK"

nicht vorfeldfähig
nicht phrasengebunden
→Modalpartikel
→ "PTKM"

(u.a. Pittner 1999)

syntaktische Adverbklassen

- weitere funktional-syntaktische Klassen:
  - "Adverb"
    - "Satzadverb" (CP-modifizierend)
    - "Modaladverb" (VP-modifizierend)
  - "Partikel"
    - gebunden in übrigen lexikalischen Phrasen (bspw. AP-modifizierend)
  - "Modalpartikel" (immer CP-modifizierend)

Adverb (ADV)

- *morgen; vielleicht* (s.u.)
- alleine topikalisierbar

- syntakt. Subklassifikation:

(1) 'Modaladverb' (angelehnt an Pittner 1999)
  - VP-Anbindung
  - ➢ Propositionsmodifikation; Skopus ist VP

(2) 'Satzadverb' (angelehnt an Pittner 1999)
  - S-Anbindung/CP-Anbindung
  - ➢ ohne Propositionsmodifikation; Skopus ist S

- gute Testmöglichkeiten: Erfragbarkeit, Substituierbarkeit, Extraponierbarkeit

Partikel (PTK)

- Kriterium: in Mutterphrase fixiert; kann nur mit Kopf permutiert werden
  (*Das kann* [*nur Peter*] *sagen.*)

- syntakt. Subklassifikation:

  Spezifizierung des Skopus (z.B. AP)
  (zu Skopus- und Fokusrahmen, vgl. Dimroth/Klein 1996)

Modalpartikel (PTKM)

- Exempl.: ***nicht*** Definition der IDS-Grammatik (Zifonun et al. 1997) ("illokutionstangierende Bewertung oder Einschränkung der Geltung eines Sachverhalts als Modifikation des Modus dicendi") (*bedauerlicherweise, sicherlich, vielleicht*)

  - ***Sicherlich/bedauerlicherweise/vielleicht*** *wird sie mit mir bei sich eine Suppe kochen.*
    (Bsp. aus Zifonun et al. 1997)
  - ***Morgen*** *wird sie mit mir bei sich eine Suppe kochen.*

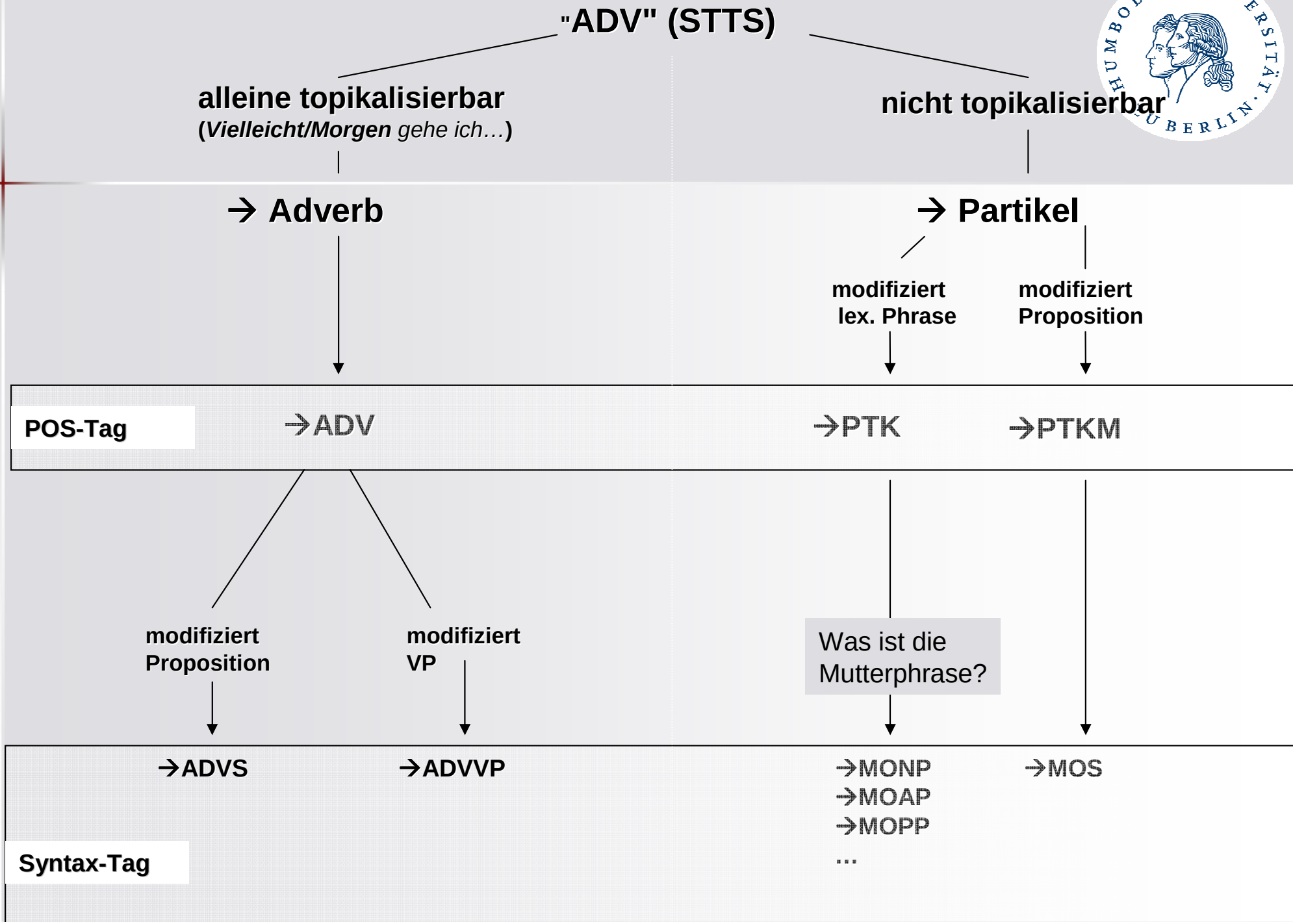
  → gleiche Permutierbarkeit; Satzgliedposition

  - **Vgl.:** ***Morgen*** *wird sie* ***ja*** *mit mir bei sich eine Suppe kochen.*

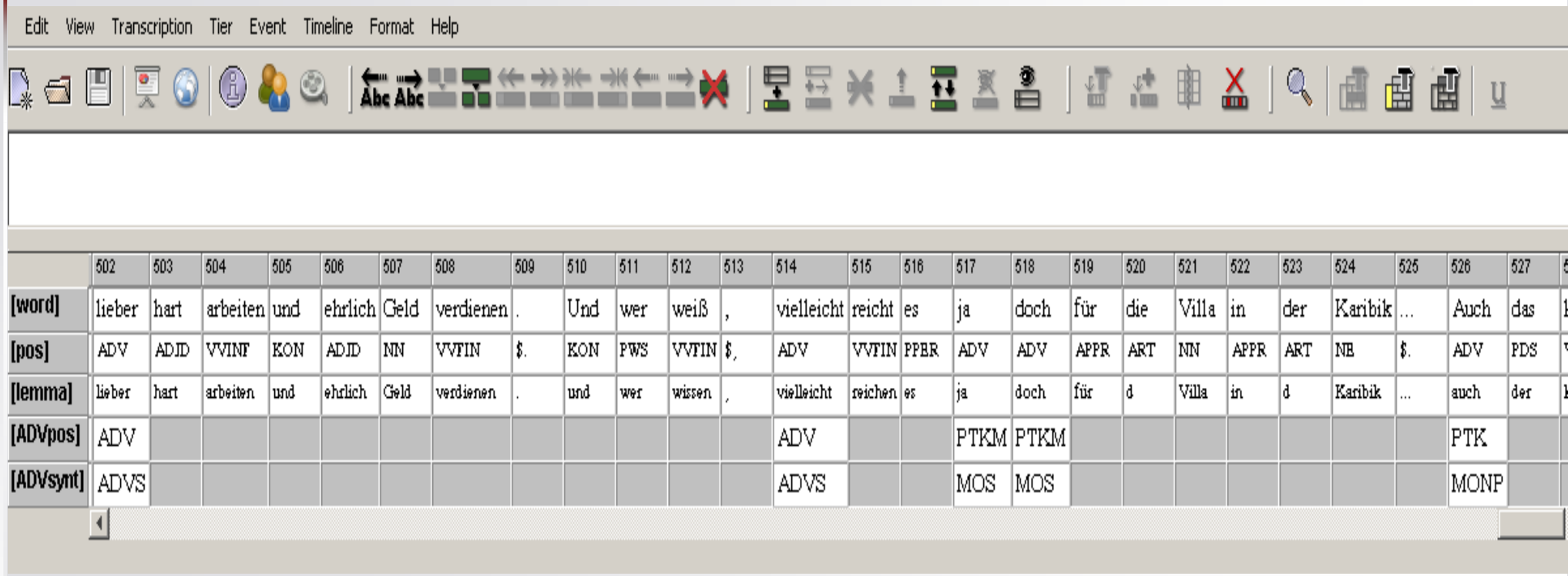
Modalpartikel (PTKM)

- strikt syntaktisch-topologische Kriterien:
  - Nicht-Vorfeldfähigkeit
  - keine lexikalische Mutterphrase
  - (Relevanz für Fremdspracherwerb)

ADV-Schema


"ADV" (STTS)
alleine topikalisierbar
(*Vielleicht/Morgen* gehe ich…)
nicht topikalisierbar
→ Adverb
→ Partikel
modifiziert lex. Phrase
modifiziert Proposition
POS-Tag
→ADV
→PTK
→PTKM
modifiziert Proposition
modifiziert VP
Was ist die Mutterphrase?
→ADVS
→ADVVP
→MONP
→MOAP
→MOPP
…
→MOS
Syntax-Tag

Annotation



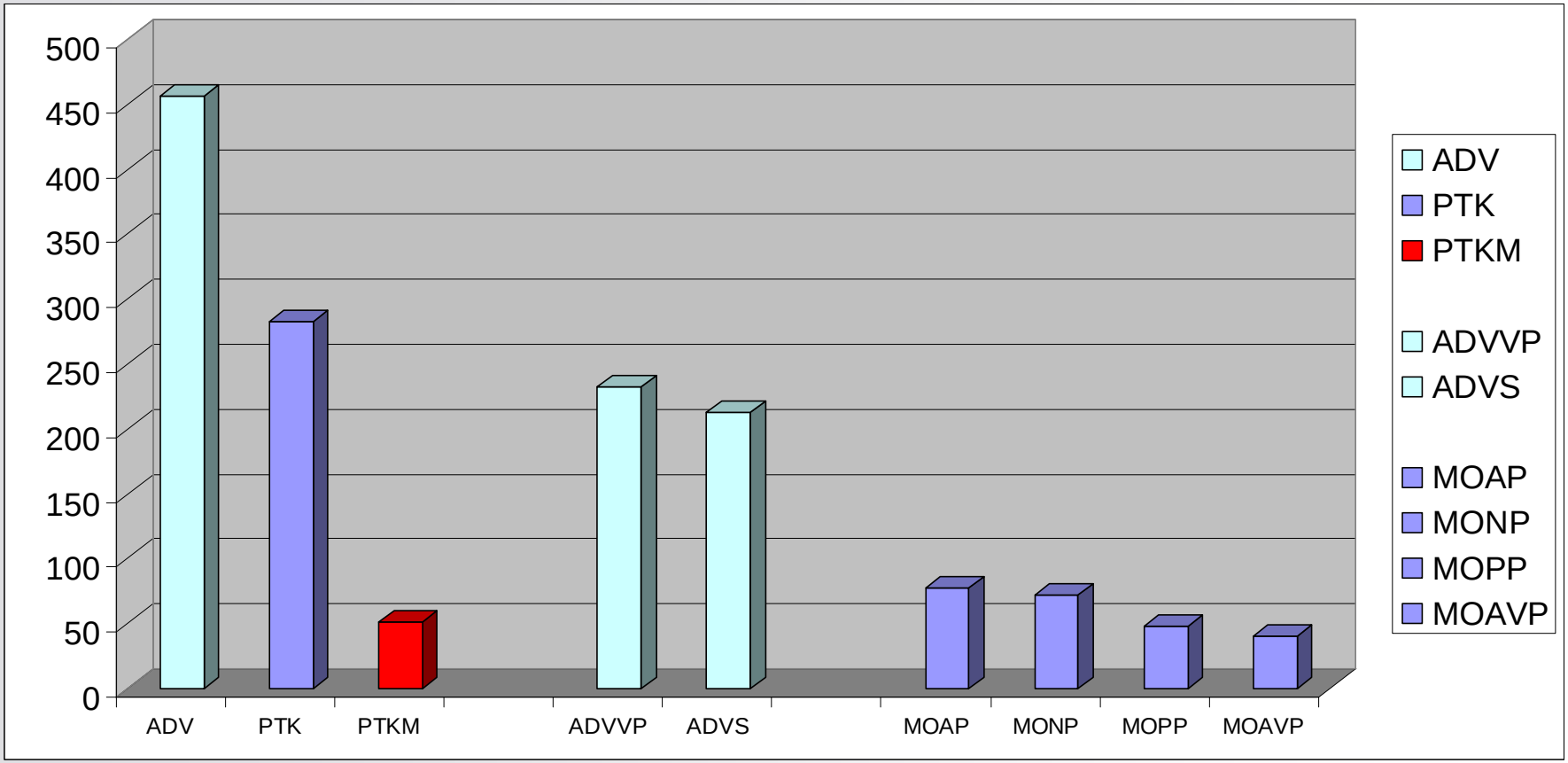
	502	503	504	505	506	507	508	509	510	511	512	513	514	515	516	517	518	519	520	521	522	523	524	525	526	527
[word]	lieber	hart	arbeiten	und	ehrlich	Geld	verdienen	.	Und	wer	weiß	,	vielleicht	reicht	es	ja	doch	für	die	Villa	in	der	Karibik	...	Auch	das
[pos]	ADV	ADJD	VVINF	KON	ADJD	NN	VVFIN	$.	KON	PWS	VVFIN	$,	ADV	VVFIN	PPER	ADV	ADV	APPR	ART	NN	APPR	ART	NE	$.	ADV	PDS
[lemma]	lieber	hart	arbeiten	und	ehrlich	Geld	verdienen	.	und	wer	wissen	,	vielleicht	reichen	es	ja	doch	für	d	Villa	in	d	Karibik	...	auch	der
[ADVpos]	ADV												ADV			PTKM	PTKM								PTK	
[ADVsynt]	ADVS												ADVS			MOS	MOS								MONP	

- Annotationstool: Exmaralda
  (vgl. Schmidt 2004; http://www.exmaralda.org)

aktueller Stand

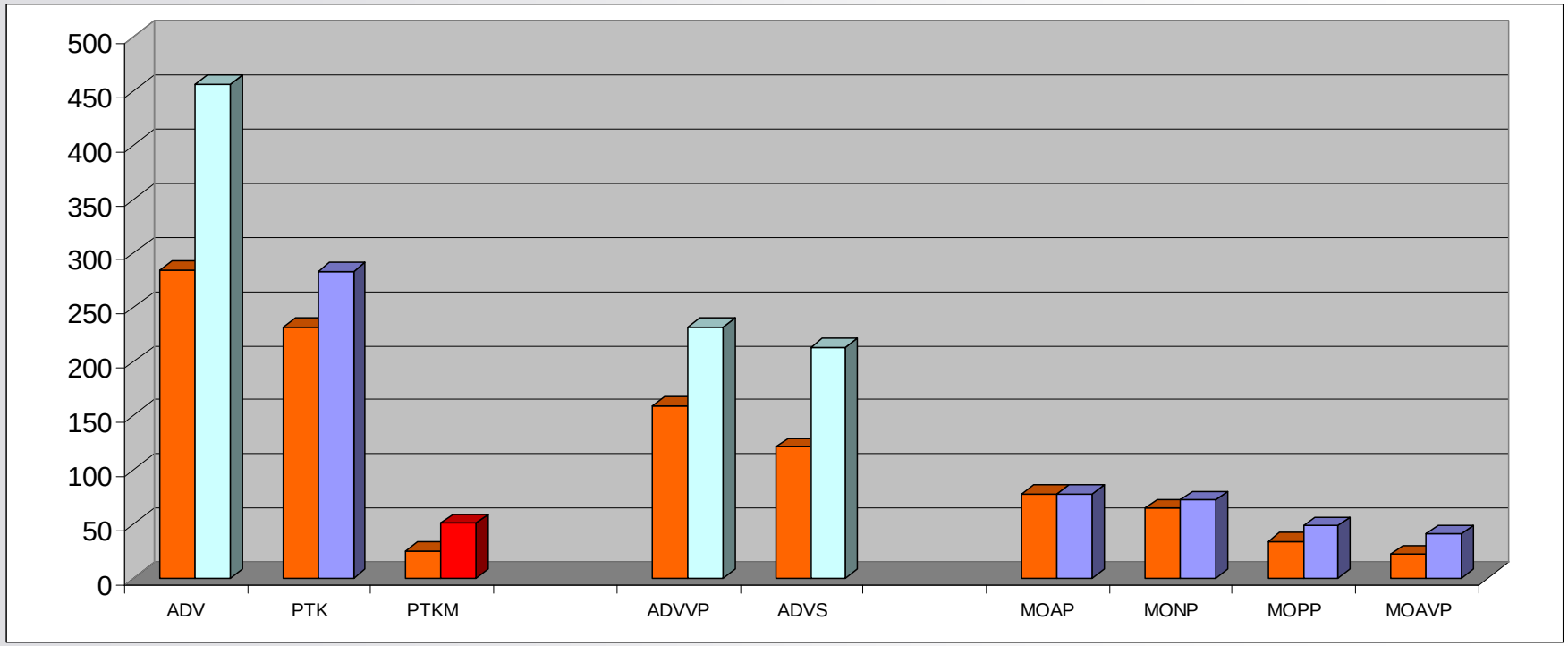
- Überarbeitung der Falko-Korpora auf diese Weise (großer Vorteil: Mehrebenenarchiktur: Sämtliche bisherige Arbeit bleibt erhalten, wird nicht beeinträchtigt).
- Gleichzeitig Markierung der unkanonischen Strukturen
- Bislang jeweils 30000 Token für Essay L1 und L2 (absolut vergleichbare Korpora)
- Pilotstudienergebnisse:

Verteilung von "ADV"-Kategorien in L1 Essays

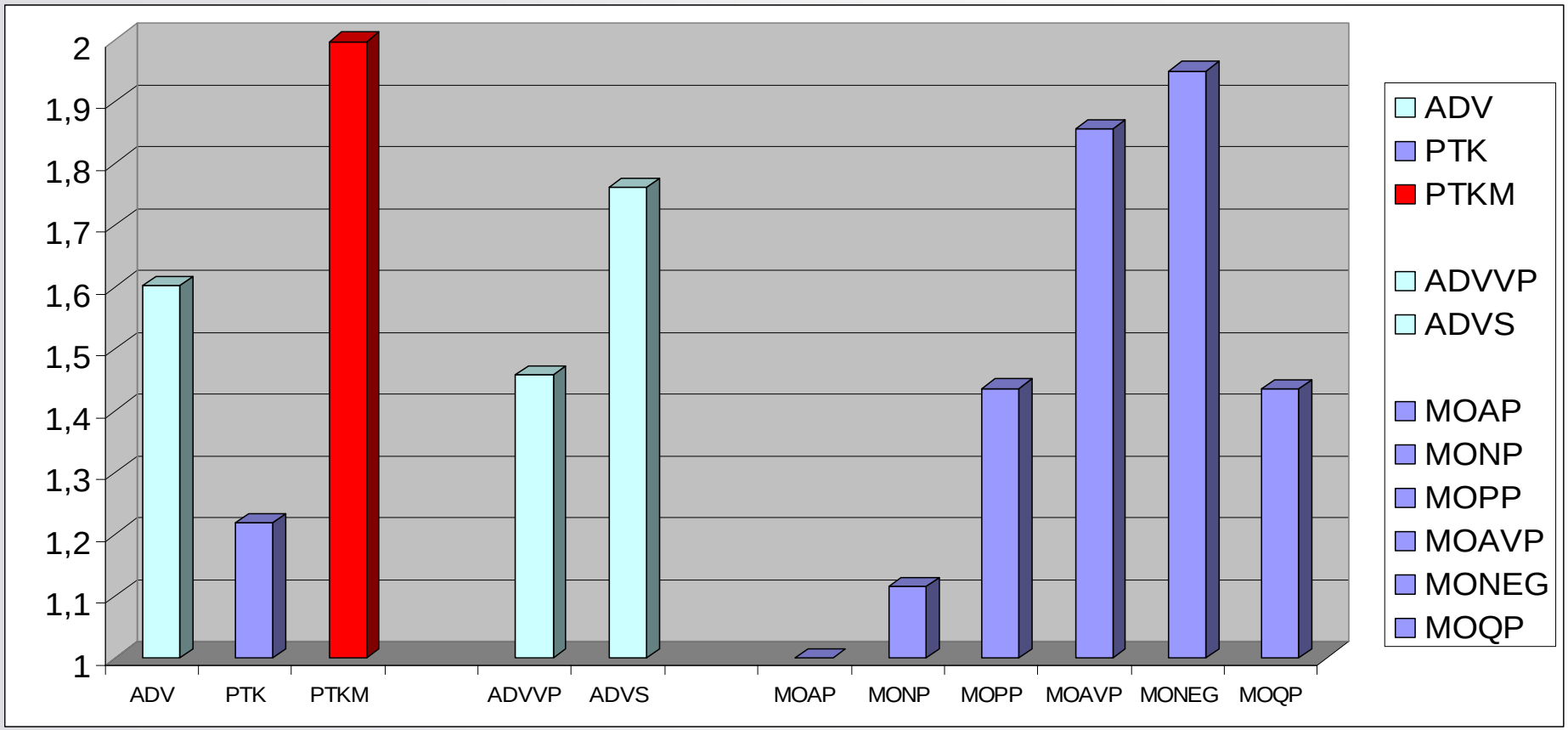


(normalisiert auf 10000 Token)

Verhältnis "ADV"-Kategorien in L2 (links) und L1 (rechts)



Underuse-Stärke (L2)



ADV-ADV-Ketten

- Alle für uns interessanten Klassen lassen sich definieren und finden:

[ADV-ADV]$_{AdvP}$

(rechtsköpfig; L1)

11	Was bewegt so viele Menschen **dazu immer wieder auf** Gewalt oder derartige Mittel zurückzugreifen
12	Polizei zu tun , da **sie immer wieder mit** dem Gesetz in Konflikt geraten
13	. Abgesehen von den Strafen **die immer wieder auf** einen warten ist es allgemein
14	In den letzten Jahren hört **man immer wieder verstärkt** von Kriminalität und Gewalt in
15	erhofft hat , ist es **doch noch lange kein** Grund Kriminäl zu werden und
16	keiner Lösung , weshalb die **Kriminalitätsrate immer weiter ansteigt** . Es wir nie einer
17	berücksichtig werden würde , würde **es weniger oft zu** Diskussionen um das liebe Geld
18	bestimmten Branchen aufgestellt werden . **Aber auch so würde** es schwer werden , die
19	die Villa in der Karibik **... Immer wieder bekommt** man in den Zeitungen und
20	) leidet , unmöglich , **sich ebenso lange auf** seine Arbeit zu konzentrieren wie
21	zu begehen , können sie **sich sehr wohl lohnen** (vorrausgesetzt natürlich sie werden
22	ja den Chefs von großen **Konzernen noch ein bisschen** zuwinken , wenn sie an
23	Material etc. im Zuge des **Scheinerwerbs immer wieder gefordert** und geübt worden , weshalb
24	Sinne die Antwort auf diese **Frage schon lange zu** spüren - durch Kürzungen ,
25	Klasse sogar ausprobiert und reflektiert **und immer wieder machen** sich Schülerinnen und Schüler auf

Modalpartikel, gefolgt von phrasengebundener Partikel…

L1:

HU Korpuslinguistik - CQP-Webinterface - Mozilla Firefox

https://korpling.german.hu-berlin.de/cqpwi/results.php?query=[pos%3D"PTKM"%26word!%3D"(aber|auch)"][pos%3D"PTK"]&corpus=ADV_L1&output=mi

print

Your Query	
query	[pos="PTKM"][pos="PTK"]
corpus	ADV_L1 i
result set	all
CQP commands	set autoshow no; set PrintOptions number; set leftContext 5 word; set rightContext 5 word;

Nr.	Match	Position
1	Leute die in Grenznähe wohnen **wohl eher** nach Polen fahren zum Tanken	2122
2	wirklich was davon hat ist **doch sehr** zweifelhaft . Deswegen bin ich	3239
3	erhofft hat , ist es **doch noch** lange kein Grund Kriminäl zu	3807
4	Haushalt . Mädchen gehen wenn **überhaupt nur** bis sie 14 Jahre sind	4781
5	passieren . Doch man kennt **ja auch** die andere Seite , die	5954
6	Man geriet als JUgendlicher , **ja schon** als Kind , an die	5974
7	auch nicht sagt : " **ja so** und so viele Kriminelle sind	6203
8	trinken bekommen , holt uns **doch ganz** schnell wieder auf den Boden	9319
9	. Heutzutage ist es sowieso **schon sehr** schwer einen schönen Beruf und	9725
10	etwas Schwaches anzusehen- vielmehr sind **doch gerade** die Unterschiede zwischen den Geschlechtern	11535
11	dürfte , erscheint es selbstverständlich **schon weitaus** sinnvoller , sich mit großer	11777
12	Aussagen , überspitzt formuliert , **ja erst** im Referendariat in einer Phase	12109
13	Rat : " Verjessen se **mal janz** schnell , wat se an	13175

Modalpartikel, gefolgt von phrasengebundener Partikel...

L2:

zurück zu "*viel*"...

- eine der am wenigsten akkuraten Desambiguierungen des Treetaggers ist ADV|PIS bei "*viel*" und "*wenig*"
- manuelle pos-Korrektur beim Annotieren
- In den annotierten Daten gibt es zwei Klassen
  - (Intensivierungs-)Partikel (*viel besser*)
  - → unsignifikanter Frequenzunterschied
  - VP-Adverb:

viel als ADVVP

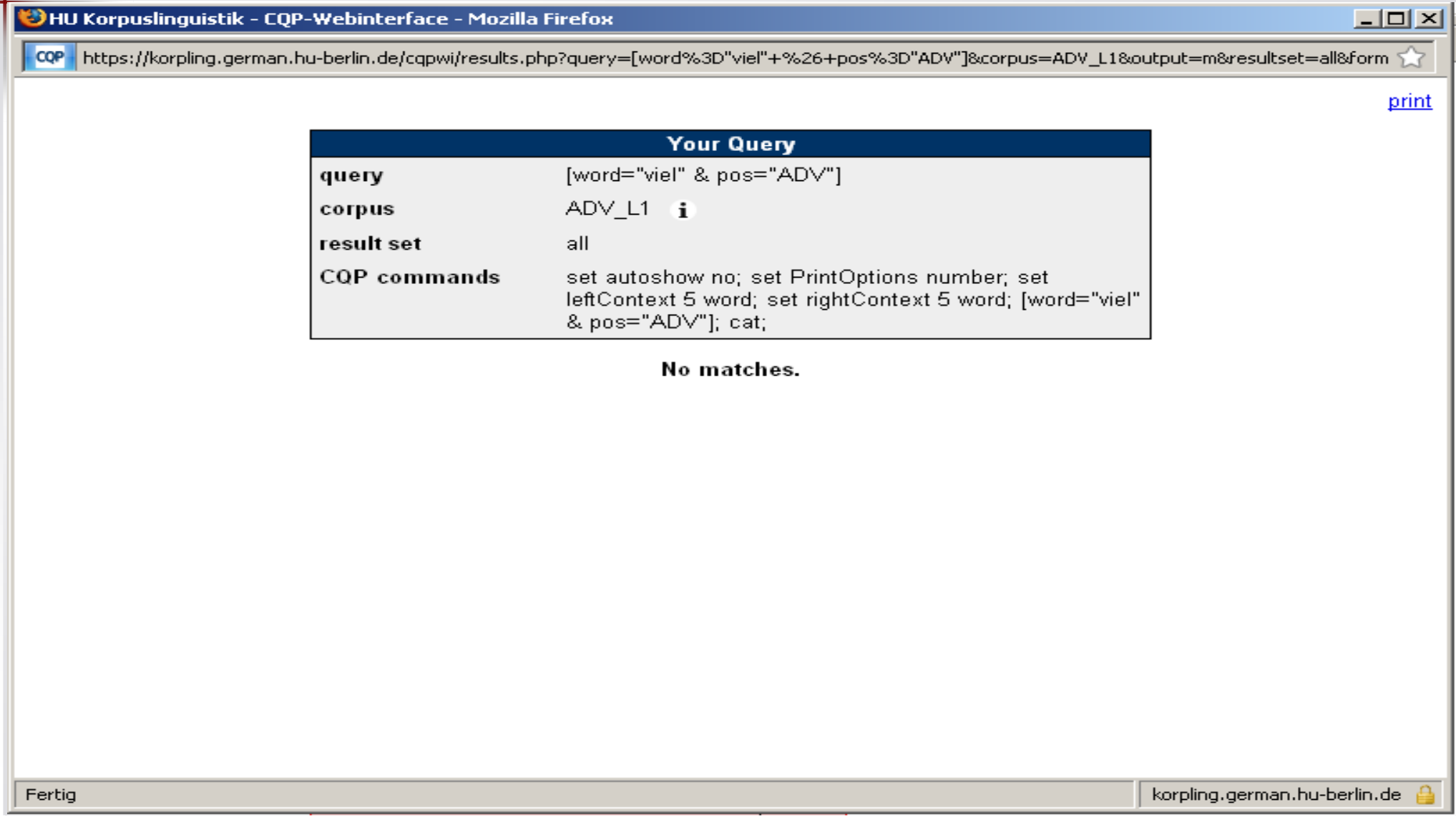
L2:

Your Query	
query	[word="viel" & pos="ADV"]
corpus	ADV_L2
result set	all
CQP commands	set autoshow no; set PrintOptions number; set leftContext 5 word; set rightContext 5 word; [word="viel" & pos="ADV"]; cat;

Match
Schiller usw. , aber dafür **viel** mit relevanten Themen , wofür
Gestaltung von Briefen wird ja **viel** in der Praxis verwendet ,
diejenigen , die für sie **viel** arbeiten werden (weil sie
hat den Interessen der Frauen **viel** genützt . Jedoch könnten die
Probleme liegen heute nicht so **viel** in Gesetzen und Kvotierungen ,
hat seit die fünfzigen Jahren **viel** evoluiert . Heute könnte es
mehr , es wird so **viel** übertrieben . Wenn ein Mann
den Interessen der Frauen sehr **viel** nützen könne , aber es
Universitätssystem in der UK ziemlich **viel** verändert . Von dem Jahr zu dem Jahr

viel als ADVVP

L1:



Zusammenfassung/Fazit

- STTS-Tagset im Bereich "ADV" ungenügend syntakt. Aussagekraft f. Lernerstudie
- Treetagger dennoch gut zur Vorverarbeitung von Annotation syntaktischer Kategorien
- Subklassifikation von "ADV" in Mehrebenkorpus mit relativ geringem Aufwand
- Syntaktisch relevante "ADV"-Klassen (Unigramme u. Ketten) können gefunden werden.
- neue Erkenntnisse ableitbar:

Zusammenfassung/Fazit ADV-Studie

- deutliche Unterschiede in der Stärke des Underuses in ADV-Klassen (traditionelle DaF-Aussage in Daten bestätigt)
- keine Overuse-Klasse
- fortgeschrittene Lerner des Deutschen als Fremdsprache in argumentativen Texten bei Modalpartikeln am wenigsten authentisch, bei phrasengebundenen Partikeln am meisten
- ➢ Unterstützung der Variabilitätshypothese
- Erklärung lexematischen Overuses mithilfe von ADV-Klassen möglich

viel zu tun…

- sämtliche Falkodaten auf diese Weise annotieren
- andere Modifikatoren (ADJD, PROAV) miteinbeziehen (syntaktisch dieselben Werte!)
- Falko nach ANNIS
- Identifikation spezifischer syntaktischer ADV-Underuse-Klassen (noch zu wenig annotierte Daten für komplexere Strukturen)

Herzlichen Dank!

wichtige Falkos:
Jia Wei Chan, Thomas Krause,
Cedric Krummes, Anke Lüdeling,
Marc Reznicek, Maik Walter, Amir Zeldes

email: hirschhx@hu-berlin.de

References

- Admoni, Wladimir (1980): Der deutsche Sprachbau, München; Beck.
- Baayen, R. Harald (2008) Corpus Linguistics in Morphology: Morphological Productivity. In: Lüdeling, A./Kytö, M. (eds.), Corpus Linguistics. An International Handbook. Berlin: Mouton de Gruyte
- Biber, D. (2006), *University Language: A Corpus-based Study of Spoken and Written Registers*. Amsterdam: John Benjamins.
- Dimroth, Christine; Klein, Wolfgang (1996) Fokuspartikeln in Lernervarietäten. Ein Analyserahmen und einige Beispiele. *Zeitschrift für Literaturwissenschaft und Linguistik. 104.* S. 73-114
- Granger, S. (2008) Learner Corpora. In: Lüdeling, A./Kytö, M. (eds.), Corpus Linguistics. An International Handbook. Berlin: Mouton de Gruyter, 259-275.
- Lüdeling, Anke; Doolittle, Seanna; Hirschmann, Hagen; Schmidt, Karin; Walter, Maik (2008) Das Lernerkorpus Falko. In: *Deutsch als Fremdsprache 2(2008)*, 67-73.
- Pittner, Karin (1999) Adverbiale im Deutschen. Untersuchungen zu ihrer Stellung und Interpretation. Tübingen: Stauffenburg
- Rapp, Reinhard; Lezius, Wolfgang (2001) Statistische Wortartenannotierung für das Deutsche *Sprache und Datenverarbeitung* 25(2):5-21.
- Schmid, Helmut (1994) Probabilistic Part-of-Speech Tagging Using Decision Trees. In: *Proceedings of the International Conference on New Methods in Language Processing.* Manchester, 44-49. [extended version available at http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf].
- Schmidt, Thomas (2004) EXMARaLDA - ein System zur computergestützten Diskurstranskription. In: Mehler, A. & Lobin, H. (ed.): Automatische Textanalyse. Systeme und Methoden zur Annotation und Analyse natürlichsprachlicher Texte, 203-218. Wiesbaden: Verlag für Sozialwissenschaften.
- Zifonun, Gisela; Hoffmann, Ludger; Strecker, Bruno (1997): Grammatik der deutschen Sprache. Band 1. Berlin/New York; Walter de Gruyter.