Morphological Productivity in Learner German: Complex Verbs

Anke Lüdeling
Humboldt-Universität zu Berlin

MIMS, Hamburg, Oct 2010

morphological productivity

- research question: Do advanced learners of German as a foreign language use productive word formation mechanisms in the same way that native speakers do?
- test case: complex verbs
- corpus: Falko

plan

- description of Falko corpus
  - design
  - pre-processing,
    especially: target hypotheses
- productivity
- complex verbs in German
- study of complex verbs in Falko

Falko - design

- freely available annotated corpus of learner German
  (Lüdeling et al. 2008, Reznicek et al. 2010)
- several subcorpora, here: essay corpus
  - advanced learners (c-test)
  - argumentative texts, four topics
  - controlled collection
  - detailed metadata
  - L2 data and L1 'control' data

Falko - preprocessing

- maximally flexible multi-layer standoff architecture
- *every* error analysis is based on a target hypothesis
- for transparent analyses the target hypothesis needs to be made explicit
- in Falko at the moment:
  three target hypotheses
  → annotations to the original learner data

Falko – target hypotheses

- target hypothesis 1: very close to original data, clear grammatical errors, sentence-based
- target hypthesis 2: additionally style errors, text-based
- target hypothesis verb: similar to th2, special emphasis on complex verbs

Falko – target hypotheses - example

learner	th1	th2
Meine	Meine	Mein
wichtigsten	wichtigsten	
Interresse	Interessen	Hauptinteresse
sind	sind	gilt
die	die	den
Sache	Sachen	Dingen
,	,	,
die	die	die
nicht	nicht	nicht
"	"	"
wirklich	wirklich	wirklich
"	"	"
sind	sind	sind
.	.	.

~ My main interest are those things that are not 'real'. (nz002_2007_06)

Falko – annotation

- original text plus all target hypotheses are automatically annotated with pos and lemma (TreeTagger, Schmid 1994)
- all target hypotheses are automatically annotated with edit errors
  – edit differences between learner utterance and target hypothesis
  (CHA, INS, DEL, MOVS - MOVT, SPLIT, MERGE)

learner	pos	lemma	th2	error tags
Meine	PPOSAT	mein	Mein	CHA
wichtigsten	ADJA	wichtig		
Interresse	NN	[UNKNOWN]	Hauptinteresse	MERGE
sind	VAFIN	sein	gilt	CHA
die	ART	d	den	CHA
Sache	NN	Sache	Dingen	CHA
,	$,	,	,	
die	PRELS	d	die	
nicht	PTKNEG	nicht	nicht	
"	$(	"	"	
wirklich	ADJD	wirklich	wirklich	
"	$(	"	"	
sind	VAFIN	sein	sind	
.	$.	.	.	

Falko – annotation

- original text plus all target hypotheses are automatically annotated with pos and lemma all target hypotheses are automatically annotated with edit errors
  (CHA, INS, DEL, MOVS - MOVT, SPLIT, MERGE)
- structural annotation (spans) for titles, comments etc.
- thverb is manually annotated with verb information

Falko – search

- all annotation layers can be searched simultaneously in Annis2
  (Zeldes et al. 2009)
- meta-data filters (ad hoc corpora)

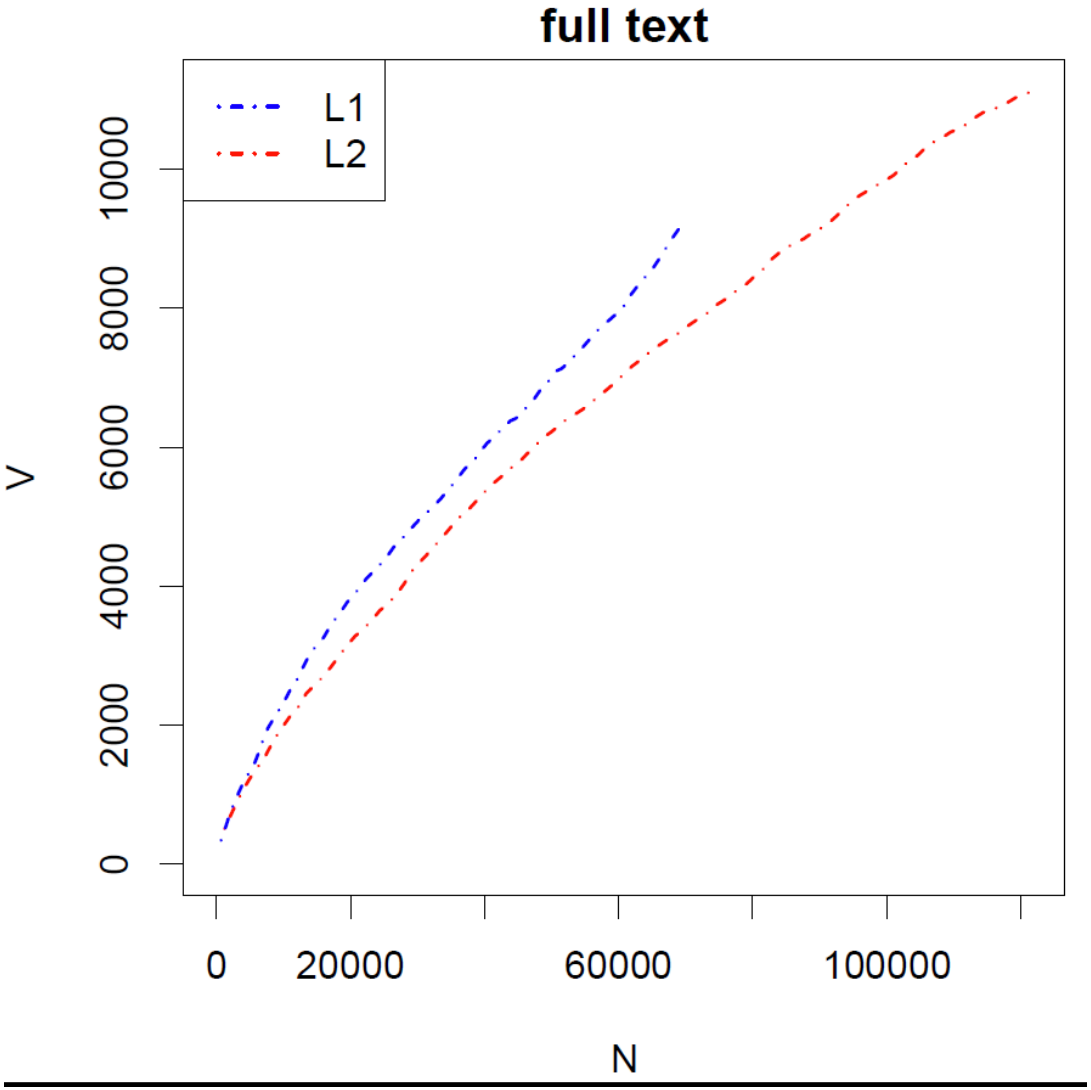
back to the research question: productivity

- from many studies we know that native speakers often form new words,
  i.e. they are (implicitly) aware of regular word formation rules,
  probably abstracted from linguistic input
  (Baayen 1992, 2001 etc., Plag 1998, Bauer 2001, Lüdeling & Evert 2005, Kiss 2007, Zeldes to appear etc.)
- do learners do the same?
- do learner corpora help us to find out?
- not many studies yet
  (Berth 2009)

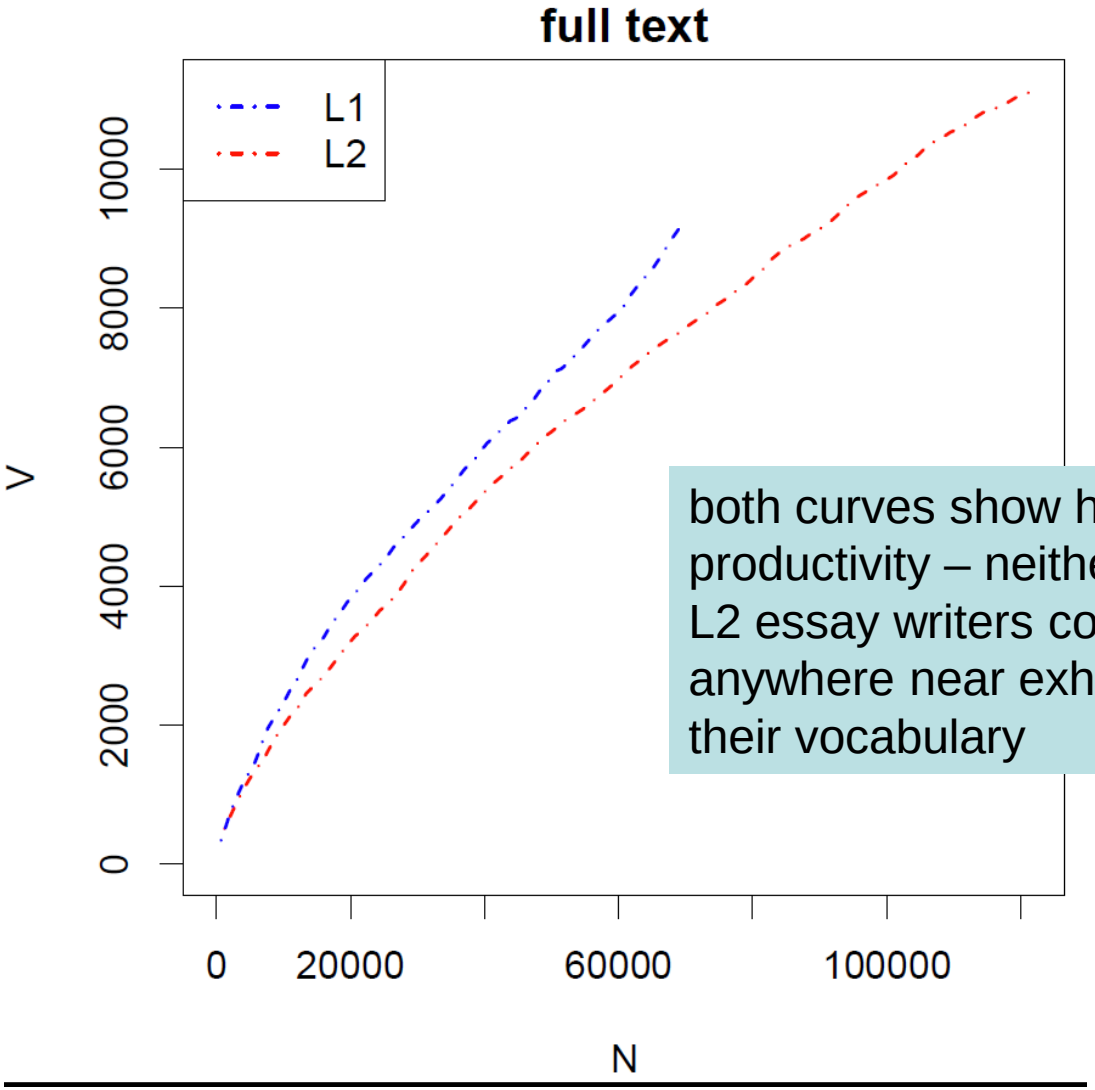
vocabulary growth curves

- many corpus-based measures for morphological productivity
  (Baayen 2001, 2009 etc.)
- here only: vocabulary growth curves, N (tokens) vs. V (types)
- they show whether the sample (the corpus) exhausts the full vocabulary or whether there are many more (possible) words in the population

vocabulary growth curve for Falko L1 and L2



vocabulary growth curve for Falko L1 and L2



both curves show high productivity – neither L1 nor L2 essay writers come anywhere near exhausting their vocabulary

productivity of complex verbs

- the high productivity of the full text can be due to
  (a) a good lexical memory
  (b) the use of productive mechanisms
  (probably both)
- complex verbs in German are a good test case for this because they have many lexicalized forms and also productive patterns
- complex verbs here:
  prefix verbs and particle verbs

complex verbs in German: form

prefix verbs

- *ver•kaufen* ‚sell‘
- *[…] dass Peter Schokolade verkauft.* ‚that Peter sells chocolate‘
- *Peter verkauft Schokolade.*
- infinitive: *zu verkaufen*
- participle: *verkauft*

particle verbs

- *auf•essen* ‚eat up‘
- *[…] dass Peter die Schokolade aufisst.* ‚that Peter finishes the chocolate‘
- *Peter isst die Schokolade auf.*
- infinitive: *aufzuessen*
- participle: *aufgegessen*

complex verbs in German: semantics

- both prefix verbs and particle verbs can be lexicalized (completely non-transparent)
- both form productive patterns
- prefixes and particles can have similar effects: argument structure changes, aspectual changes, changes wrt causation, direction etc.
  Stiebels & Wunderlich 1994, Olsen 1996, Lüdeling 2001, Zeller 2001, Müller 2002, Kolehmainen 2005, …
- productive patterns are not explictly taught in learner grammars / instructions – they have to be implicitly learned from the input

complex verbs in Falko

- thverb concentrates on complex verbs – every complex verb is annotated (manually) wrt
  - type (particle verb vs. prefix verb)
  - lemma
  - error type (orthography, inflection, argument structure, separability etc.)
  - form (finite, infinite, participle, separated etc.)

aside: errors in complex verbs

- all error differences are significant:

➢ the learners make significantly more structural and semantic errors

➢ the native speakers make significantly more orthographic errors ☺

descriptive comparison

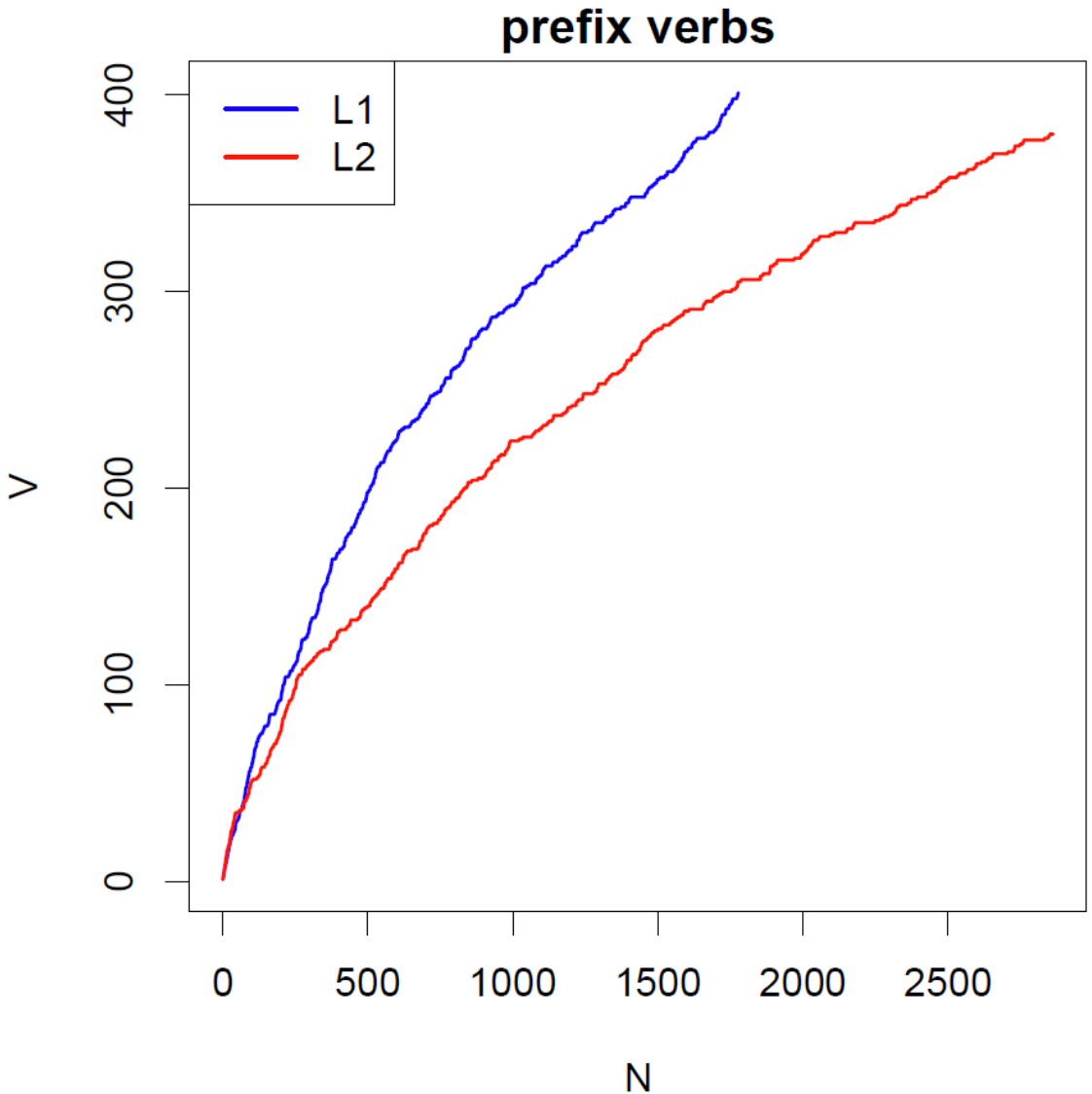
L2

- corpus size: 122778 tokens
- particle verbs: 1337
- prefix verbs: 2863
- all main verbs: 11247
  - 9,1 % of all tokens in the L2 texts are verbs
  - of these **37,3%** are complex

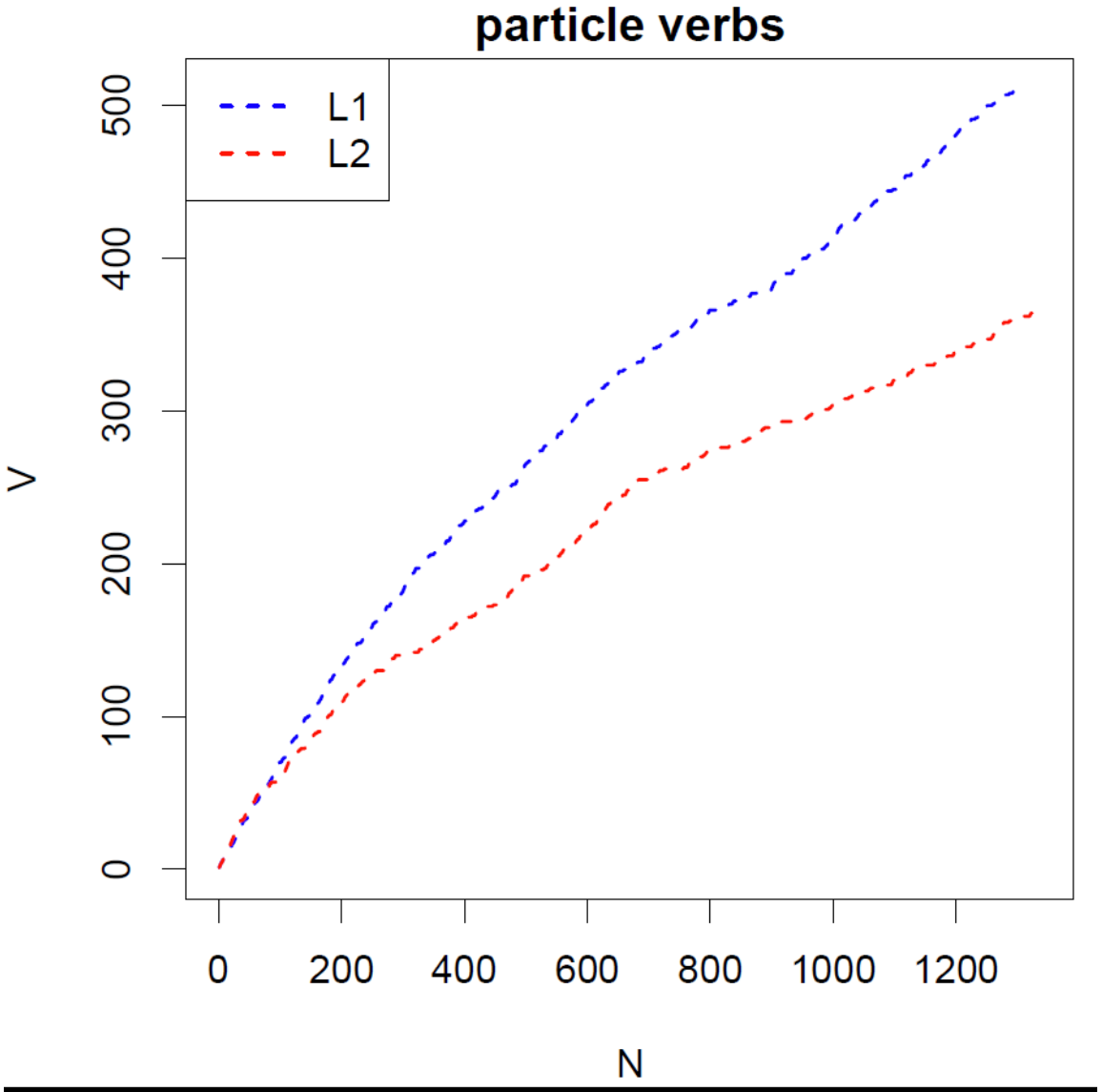
L1

- corpus size: 68491 tokens
- particle verbs: 1303
- prefix verbs: 1778
- all main verbs: 6629
  - 9,7% of all tokens in the L1 texts are verbs
  - of these **46,5%** are complex

vocabulary growth curves: prefix verbs



vocabulary growth curves: particle verbs



complex verbs

- native speakers use a higher proportion of complex verbs
- but both prefix verbs and particle verbs show growing vocabulary growth curves (for L1 and L2)
- is this simply due to the fact that native speakers *know* more complex verbs or do they know how to *form* complex verbs?

➢ analysis of productive patterns

➢ analysis of clearly new-formed words/readings

productive patterns: *hinein+verb*

hineinbewegen – move into

hineingehen – go into

hineintreten – step into

hineinbringen – bring into

hineinfallen – fall into

hineingeraten – get into

hineinpassen – fit into

hineinversetzen – put yourself in someones position

hineinwachsen – grow into

hineinziehen – pull into

productive patterns

- for many possible patterns: native speakers use more different types
- native speakers use more different patterns
- it remains unclear which types are productively formed – regularly formed new words are not noticable and indistinguishable from memorized transparent members of a pattern

analysis of 'unknown' words

- 34 'new' forms (unknown to me, tokens)

  … was eine sozialere oder gerechtigere Entlohnung benachträchtigt.
  "… what goes against a more social or more just compensation"
  (hu_005_2006_09)

- 274 new readings of existing forms (tokens)

  Viele Leute sich bewundern, ob …
  "Many people wonder if …"
  (hu_006_2006_10)

- 23 'new' forms (unknown to me, tokens)

  Sie vollrichten in ihrer Arbeitszeit fast doppelt so viel …
  "They accomplish double the amount during their work hours …" (dhw_031_2007-06)

- 34 new readings of existing forms (tokens)

  Die Regierung kann dieses große Netz der Machenschaften meist nicht entspinnen.
  "The government usually cannot disentangle this large web of machinations." (dew_03_2007_09)

analysis of 'unknown' words

- learners form 'new' words significantly more often and use 'new' readings more often than native speakers
- they often do not get the pattern quite right

➢ they understand *that* they can use the word formation means productively but have not (yet) discovered all the restrictions for the patterns

summary

- do learners use productive word formation like native speakers?
- quantitative and qualitative corpus study – Falko corpus – multi-layer architecture with specific target hypotheses
- similarities:
  - both the L1 corpus and the L2 corpus have growing vocabulary growth curves
  - many rare types, many regular formations → indication of productive word formation
- differences:
  - the native speakers use complex verbs in a much larger proportion than the learners
  - the learners try and fail in forming 'new' words (or senses) much more often

outlook

- generally: tests on variation within the samples
- for productivity studies on learner language
  - more patterns
  - a better understanding of how to describe patterns and to distinguish between productively formed words and memorized words
  - experimental studies – learning and production
- for Falko
  - more annotation layers …
  - more data …
  - all of you can help ☺!

Thanks - Danke

- to Marc Reznicek and Felix Golcher for a lot of help in analyzing the data!
- to the Falkos – Torsten Andreas, Hagen Hirschmann, Marc Reznicek, Maik Walter, Amir Zeldes – for preparing the corpora and building Annis
- Falko: http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko/standardseite
- Annis: http://www.sfb632.uni-potsdam.de/d1/annis/

references

- Baayen, R. Harald (1992) Quantitative aspects of morphological productivity. In: G. E. Booij and J. van Marle (eds), *Yearbook of Morphology 1991*, Kluwer Academic Publishers, Dordrecht, 109-149.
- Baayen, R. Harald (2001) *Word Frequency Distributions*. Kluwer, Dordrecht.
- Baayen, R. Harald (2009) Corpus linguistics in morphology: morphological productivity. In A. Lüdeling and M. Kytö (eds), *Corpus Linguistics. An International Handbook*. Vol 2. Mouton De Gruyter, Berlin, 899-919.
- Bauer, Laurie (2001), *Morphological Productivity*. Cambridge: Cambridge University Press.
- Berth, Michael (2009) *Treffungen, Sinkung* und *Benützung* - Korpuslinguistische Untersuchung des Erwerbs von derivationsmorphologischen Wortbildungsregularitäten bei fortgeschrittenen Lernern des Deutschen als Fremdsprache. Magisterarbeit, Humboldt-Universität zu Berlin
- Kiss, Tibor (2007). Produktivität und Idiomatizität von Präposition-Substantiv-Sequenzen. *Zeitschrift für Sprachwissenschaft, 26* (2), 317-345.
- Kolehmainen, Leena (2006) *Präfix- und Partikelverben im deutsch-finnischen Kontrast*. Peter Lang, Frankfurt.
- Lüdeling, Anke (2001) *On Particle Verbs and Similar Constructions in German*. CSLI, Stanford.
- Lüdeling, Anke; Doolittle, Seanna; Hirschmann, Hagen; Schmidt, Karin & Walter, Maik (2008) Das Lernerkorpus Falko. In: *Deutsch als Fremdsprache 2(2008)*, 67-73.
- Lüdeling, Anke/Evert, Stefan (2005), The Emergence of Productive Non-medical *-itis*. Corpus Evidence and Qualitative Analysis. In: Kepser, Stephan/Reis, M. (eds.), *Linguistic Evidence. Empirical, Theoretical, and Computational Perspectives.* Berlin/New York: Mouton de Gruyter, 351-370.

references

- Müller, Stefan (2002) *Complex Predicates: Verbal Complexes, Resultative Constructions and Particle Verbs in German. CSLI, Stanford.*
- Olsen, Susan (1996) Über Präfix- und Partikelverbsysteme. *FAS Papers in Linguistics* 3, 1995, 86-112.
- Plag, Ingo (2003), *Word-formation in English*. Cambridge: Cambridge University Press.
- Reznicek, Marc; Walter, Maik; Schmid, Karin; Lüdeling, Anke; Hirschmann, Hagen; Krummes, Cedric (2010) Das Falko-Handbuch. Korpusaufbau und Annotationen. Online at http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko/standardseite
- Schmid, Helmut (1994) Probabilistic Part-of-Speech Tagging Using Decision Trees. In: *Proceedings of International Conference on New Methods in Language Processing. September 1994.* Online at http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf.
- Stiebels, Barbara/Wunderlich, Dieter (1994) Morphology feeds Syntax: the Case of Particle Verbs. In: Linguistics 32, S. 913–968.
- Zeldes, Amir (to appear), "On the Productivity and Variability of the Slots in German Comparative Correlative Constructions". In: Konopka, Marek, Kubczak, Jacqueline, Mair, Christian, Štícha, František & Waßner, Ulrich H. (Hrsg.), *Grammar & Corpora / Grammatik und Korpora 2009. Third International Conference / Dritte Internationale Konferenz, Mannheim, 22.-24.09.2009*. Tübingen: Gunter Narr Verlag.
- Zeldes, Amir, Ritz, Julia, Lüdeling, Anke & Chiarcos, Christian (2009), "ANNIS: A Search Tool for Multi-Layer Annotated Corpora". In: *Proceedings of Corpus Linguistics 2009*, July 20-23, Liverpool, UK.
- Zeller, Jochen (2001) *Particle Verbs and Local Domains*. John Benjamins, Amsterdam