

Initial Parsing Decisions & Lexical Bias

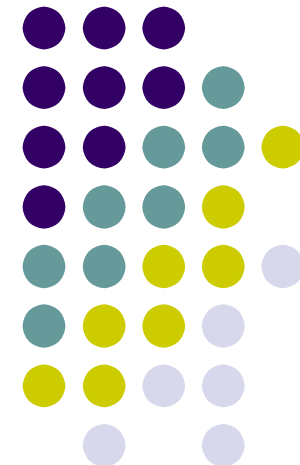
Second International
Conference of the
German Cognitive
Linguistics
Association

October 2006



seit 1558

Daniel Wiechmann
Friedrich-Schiller-Universität Jena





seit 1558

structure of the talk

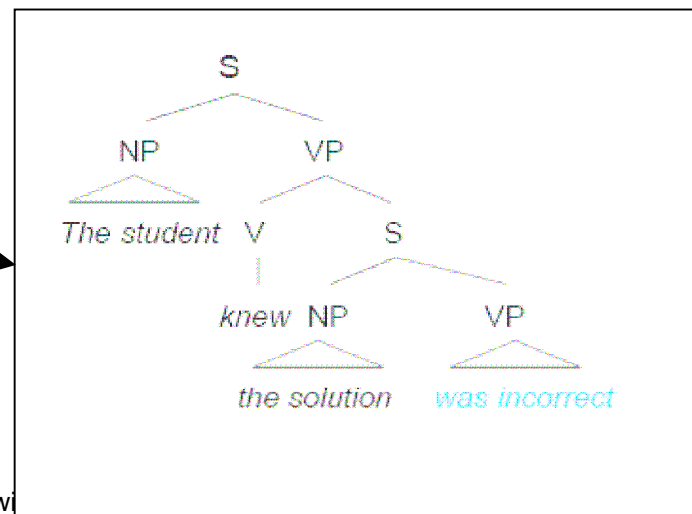
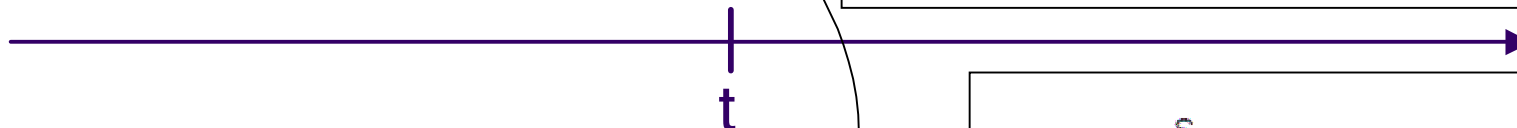
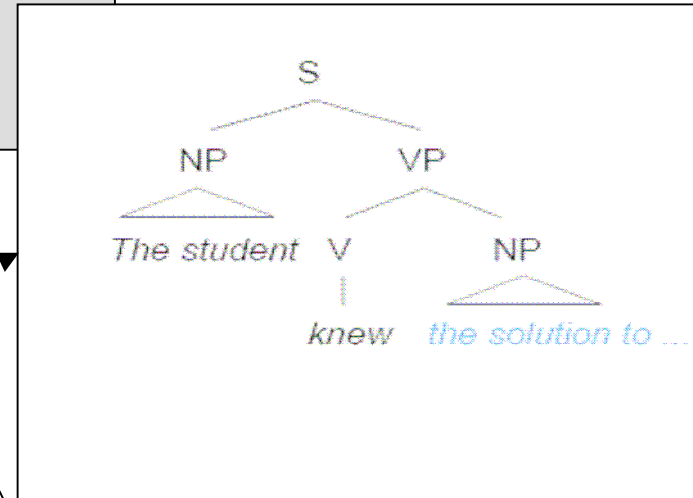
- ❑ Phenomenon:
 - ❑ Local NP/S-ambiguity
- ❑ Hypothesis
 - ❑ Sense-contingent lexical guidance
- ❑ Methods and Results
 - ❑ Corpus-based estimation of lexical bias (form & sense)
 - ❑ Comparison with experimental data (Hare et al. 2003)
- ❑ Conclusion and afterthoughts
 - ❑ Relevance for processing oriented variants of CxGs (e.g. Embodied Construction Grammar)



seit 1558

local syntactic ambiguity: NP/S-ambiguity

The student knew the solution...





seit 1558

Processing local syntactic ambiguities

Observation:

Some sentences of this type cause more processing difficulties than others.

But why?

(i.e. what is the (set of) underlying mechanism(s)?)



seit 1558

constraints on constructing an interpretation

- Phrase-formation constraints
 - minimal attachment
- Contextual constraints
 - Plausibility
 - Referential contexts
- Computational resource constraints
 - Locality (memory cost constraints)
- Phrase-level contingent frequency constraints
- Prosodic constraints
- Lexically-based constraints
 - Grammatical category of perceived element
 - Subcategorization preferences
- ...





seit 1558

constraints on constructing an interpretation

- Phrase-formation constraints
 - minimal attachment
- Contextual constraints
 - Plausibility
 - Referential contexts
- Computational resource constraints
 - Locality (memory cost constraints)
- Phrase-level contingent frequency constraints
- Prosodic constraints
- Lexically-based constraints
 - Grammatical category of perceived element
 - Subcategorization preferences
- ...

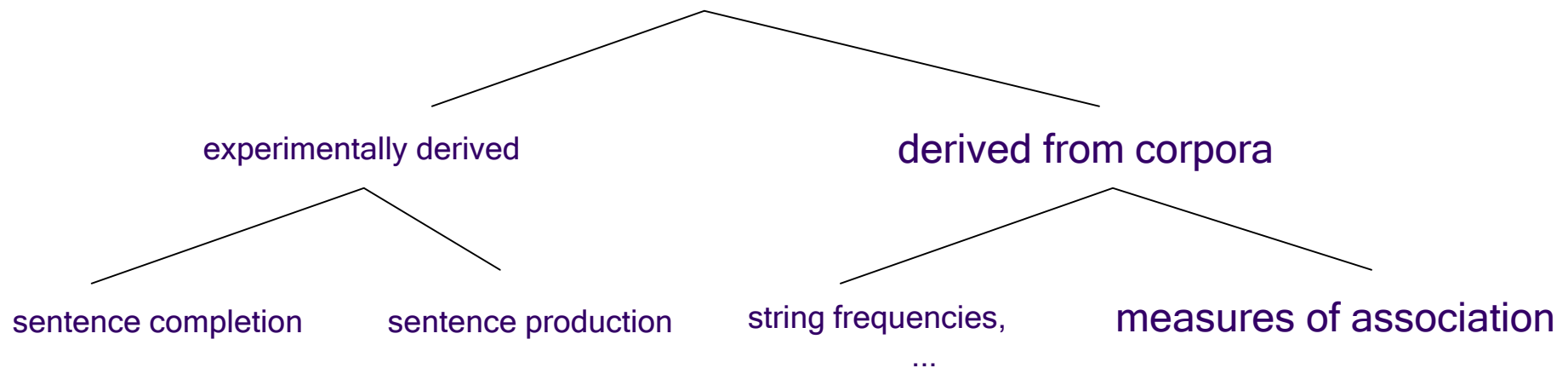




seit 1558

estimating lexical preferences

Means to estimate lexical preferences





seit 1558

lexical preferences: form vs. sense

- Estimations of lexical bias on the level of verb-form are problematic
 - Verbs are (conventionally) used to express a number of semantic relations and each of these have their own preferences
 - i. Peter_{VP} [V admitted₁ NP [his ex-girlfriend] PP [to the club]].
 - ii. Peter_{VP} [V admitted₂ S [NP [his ex-girlfriend] was nicer than his current one]].
 - iii. Peter_{VP} [V admitted₂ NP [his error]]
- $\rightarrow LB_{Verb(form)} \neq LB_{Verb(sense)}$

(cf. Roland et al. 2000: experimental vs. corpus-based norms)



seit 1558

lexical guidance hypothesis

Parsing is guided by structural expectations derived from sense-contingent lexically specific preferences.



seit 1558

procedure

1. Corpus-based estimation of LB both on the level of verb form and sense
2. LBs are compared with reading-time latencies in self-paced moving window experiment
(Hare et al. 2003)



seit 1558

data

corpus: 17 mio word sample (BNC)
[compiled as to mirror structural properties of ICE-GB corpus]
(Nelson 1996)

Step 1: Extraction of all [V NP]-constructions for 20 verbs

Number of tokens per type:	N > 3000	: random sample: 10%
	300 < N < 3000	: random sample: n=300
	N < 300	: all tokens

token_{total}: 4960



seit 1558

data

Step 2 → All instances of the relevant construction were coded with respect to

- the grammatical role played by the postverbal NP
- the verb sense instantiated by the verb in the matrix clause using a lexical database (WordNet 2.0, Fellbaum 1998)

Step 3 → Lexical bias was calculated both on the level of form and sense using Distinctive Collexeme Analysis (DCA)





seit 1558

distinctive collexeme analysis (DCA)

- DCA is a variant of ‘Collostructional Analysis’ (Stefanowitsch and Gries 2003)
 - Adopts the perspective of construction grammar
 - Assesses the degree of association between two constructions of arbitrary degrees of specificity
- DCA measures the relationship of a lexical construction towards more abstract constructions it can occur in (Gries & Stefanowitsch 2004)
 - Lexical preferences are expressed in terms of association scores
 - → outputs a gradual measure



seit 1558

distinctive collexeme analysis (DCA)

contingency table DCA	find	other verbs	row totals
NP-complementation	observed (expected)	observed (expected)	R1
S-complementation	observed (expected)	observed (expected)	R2
column totals	C1	C2	N

1. Collection of data above
2. Statistical analysis of the observed distribution;
 → Application of discounted odds-ratio as a measure of association

(cf. Evert 2004 for a discussion of association measures)

$$odds - ratio_{disc} = \log \frac{[(O_{11} + \frac{1}{2})(O_{22} + \frac{1}{2})]}{[(O_{12} + \frac{1}{2})(O_{21} + \frac{1}{2})]}$$



seit 1558

lexical bias: form-based

type verb	coll. strength form
find	4,279385367
report	3,377376751
admit	2,616279933
declare	2,362509375
add	2,222934376
grasp	1,402012303
project	1,402012303
observe	1,384432064
insert	1,299333558
indicate	0,885152033
recognize	0,64118265
claim	0,593529934
reveal	0,375739745
bet	0,295568701
reflect	0,270618368
acknowledge	0,114698963
recall	-1,203840186
anticipate	-1,350352611
feel	-2,041734805
confirm	-3,657014769

= Bias towards DO_{NP}





seit 1558

lexical bias: form vs. sense

type	preferences towards nominal complementation (discounted lg _e odds ratio)		
	formal	sense1	sense2
bet	0,30	-4,38	1,39
admit	2,62	1,08	0,87
feel	-2,04	-2,15	-0,96
claim	0,59	-0,53	1,53
confirm	-3,66	-1,63	-3,22
recall	-1,20	-0,35	-1,22
acknowledge	0,11	-0,35	1,76
indicate	0,89	-0,25	0,91
anticipate	-1,35	-0,21	-2,55
observe	1,38	0,98	1,33
declare	2,36	-0,75	-0,39
reveal	0,38	0,38	0,21
find	4,28	-0,02	-1,04
recognize	0,64	-0,91	1,61
report	3,38	-1,47	-1,04
reflect	0,27	-1,82	1,57
grasp	1,40	-0,07	0,85
add	2,22	1,27	-0,98
project	1,40	-0,73	2,39
insert	1,30	0,93	0,79

quant. difference

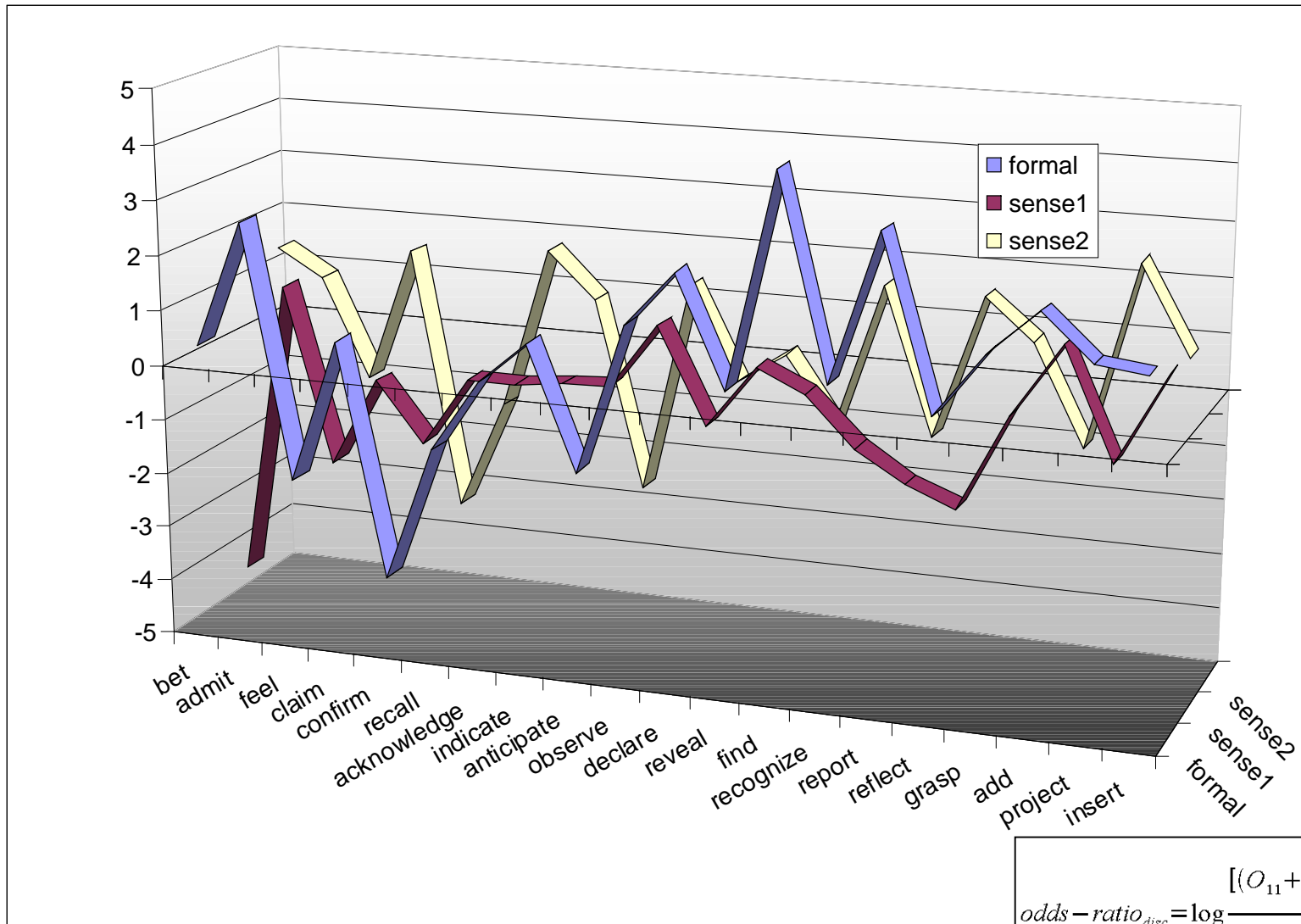
comparable

qual. difference





seit 1558



$$odds - ratio_{disc} = \log \frac{[(O_{11} + \frac{1}{2})(O_{22} + \frac{1}{2})]}{[(O_{12} + \frac{1}{2})(O_{21} + \frac{1}{2})]}$$





seit 1558

form-based vs. sense-contingent preferences

There are substantial differences between form-based and sense-contingent preferences.





seit 1558

predictive power of lexical preferences

If lexical bias is in fact an indicative factor for the subconscious expectations that guide early parsing, this should be reflected in experimental findings (e.g. reading time experiments)

→ Correlational analysis:
association scores - reading time delta

(Hare et al. 2003)



seit 1558

Hare et al. 2003: Self-paced reading

Materials:

- 20 polysemous verbs (2 senses each → source: WordNet 2.0)
 - sense 1 → preference DO_{NP}
 - sense 2 → preference DO_S

Procedure

- Subject read 2 sentences: 1 prime- & 1 target sentence
 1. Subjects read prime (→ activate sense 1 or 2)
 2. Subjects read target (always sentential complementation)
- Reading time at each word is measured
 - Ambiguity effect is assessed



seit 1558

example: find

CONDITION 1

Context evokes a scenario compatible with REALIZE-sense of find (prefers S):

“The intro psychology students hated having to read the assigned text because it was boring.”

CONDITION 2

Context evokes a scenario compatible with LOCATE-sense of find (prefers NP):

“Allison and her friends had been searching for John Grisham’s new novel for a week, but yesterday they finally were successful.”

TARGET

They found _s[(that) the book was written poorly and difficult to understand.]

condition	reading time (ms)						
	<i>that</i>	<i>the</i>	<i>book</i>	<i>was</i>	<i>written</i>	<i>poorly</i>	<i>and</i>
lemma S, no <i>that</i>		342	356	369	359	366	373
lemma S, with <i>that</i>	333	332	349	350	342	350	360
delta (ambiguity effect)		10	7	19	17	16	13
lemma NP, no <i>that</i>		339	350	372	422	369	374
lemma NP, with <i>that</i>	357	345	362	364	356	370	374
delta (ambiguity effect)		-6	-12	8	66	-1	0

Area of interest

(second word of the disambiguation region (DR_{POS2}))





seit 1558

correlation association scores_{disc}OddsRatio - RT delta_{DR Pos2}

Prediction Lexical Guidance:

The greater the strength of association in direction DO_S , the smaller the ambiguity effect

(i.e. the greater the (log) odds ratio_{disc}, the smaller the reading time delta at DR_{POS2})

Method: Spearman's Rho ($\rho \equiv r_s$)

$$\rho = 1 - \frac{6 \sum D^2}{N(N^2 - 1)}$$

$$t = \frac{r_s}{\sqrt{(1 - r_s^2)/(n - 2)}}$$





seit 1558

correlation: form-based association scores_{discOddsRatio} - RT delta_{DR Pos2}

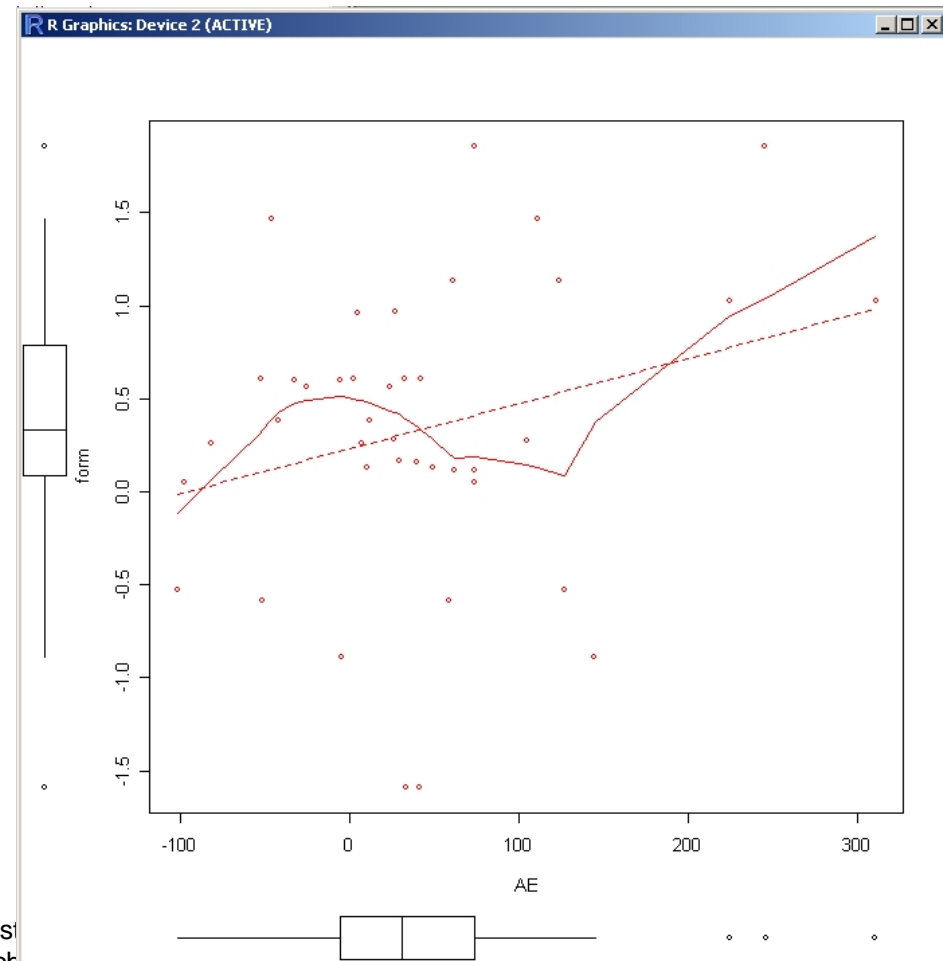
Results:

correlation_{spearman} is weak

$$r_s = 0,1597145$$

$$t = 0,99734582; p > .05 \text{ (ms)}$$

→ cannot predict reading times





seit 1558

correlation: sense-based association scores_{discOddsRatio} - RT delta_{DR Pos2}

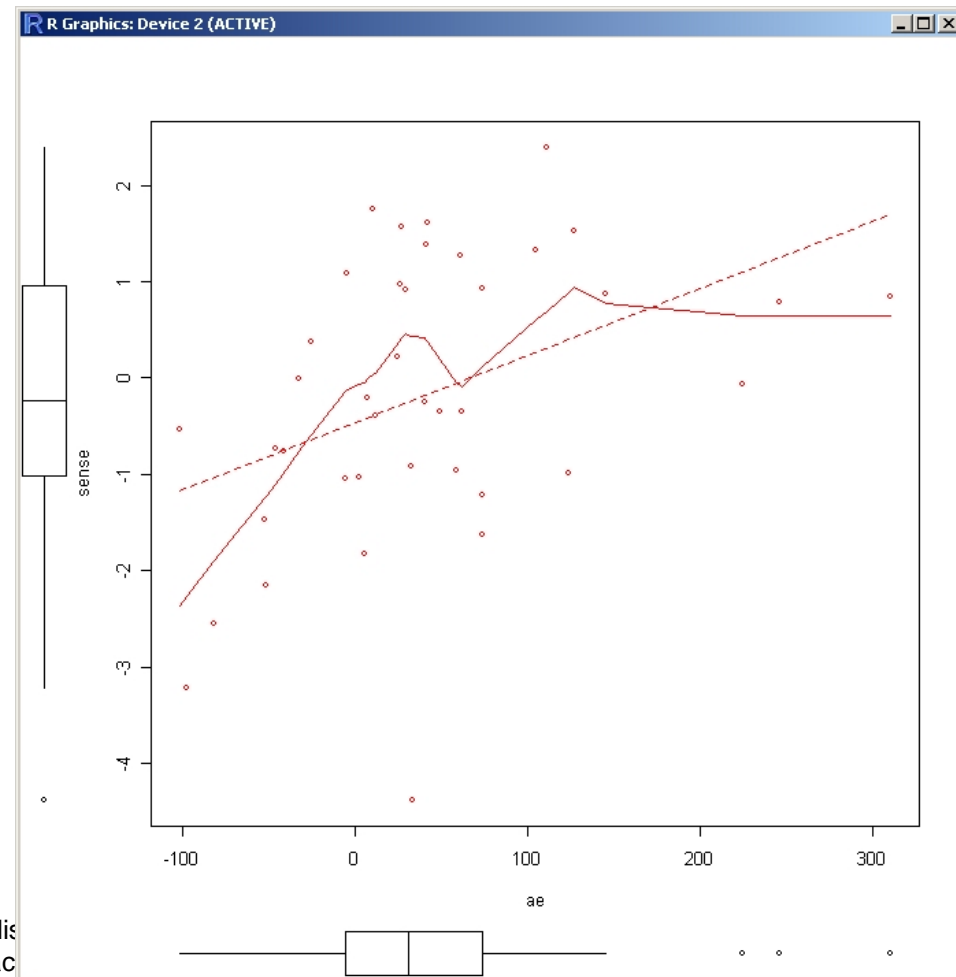
Results:

correlation_{spearman} is moderate

$r_s = 0,4116543$

$t = 2,86145612; p < .001$

→ sense-contingent lexical bias can
predict reading times



conclusion



seit 1558

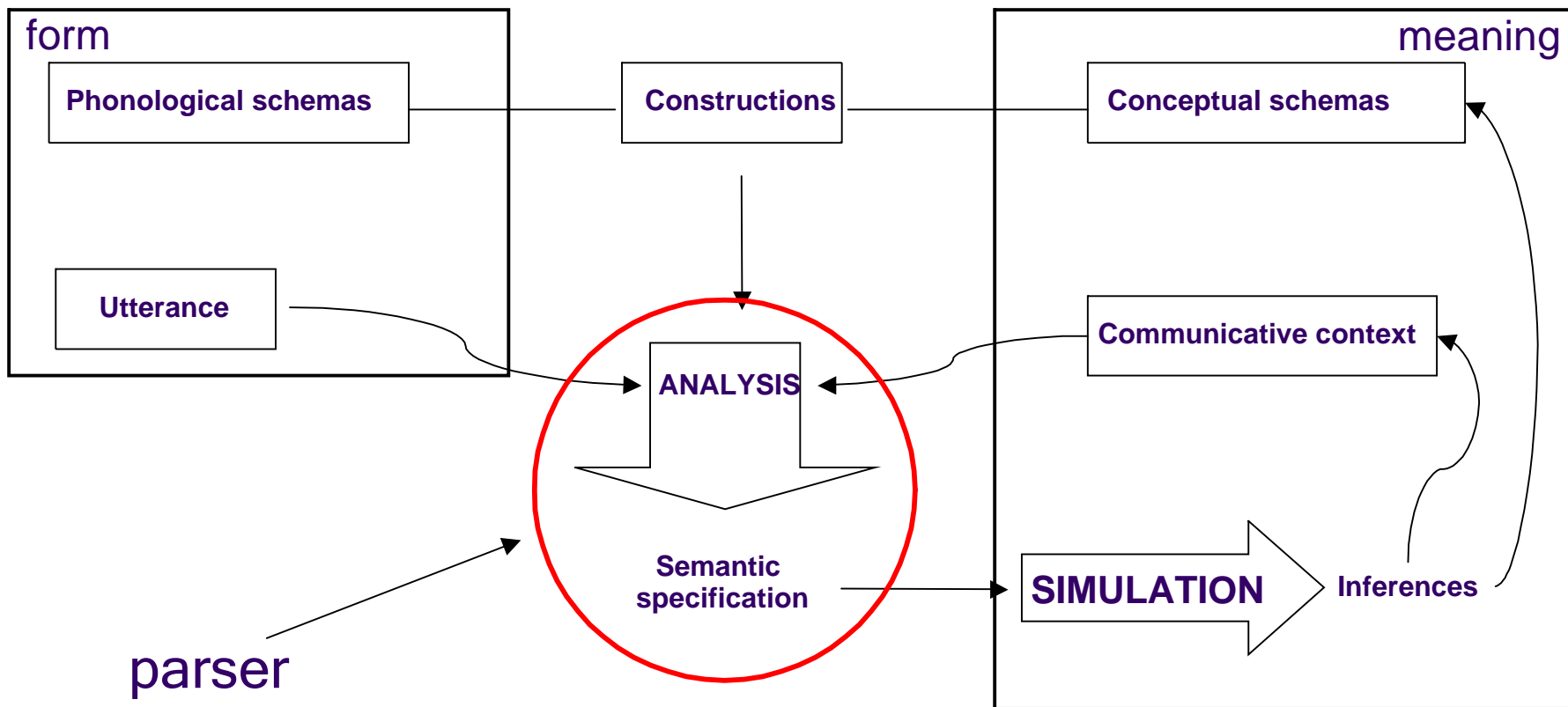
- The present study presented corpus-based evidence for a dominant role of sense-contingent lexical preferences for the resolution of local syntactic NP/S-ambiguities...
- ...and, by implication, for a sense-contingent lexically driven comprehension system



seit 1558

afterthoughts: simulation-based language understanding model

Embodied Construction Grammar (Bergen & Chang 2002)





seit 1558

afterthoughts: parsing in Embodied Construction Grammar

- Parsing = analysis process that takes an **input utterance in context** and **determines the set of constructions** most likely to be **responsible** for it
 - Cued construction potentially **supply top-down constraints** on their constituents
 - Constructions and their constraints should be regarded **not** as **deterministic** but as fitting a given utterance and context to some **quantifiable** degree
 - Constructions and their constraints could be equipped with connection weights
 - cf. Bryant (2003, 2004) for a computational model
 - → **association strengths (collostruction strength)**



references

- Agresti, Alan. 1992. A Survey of Exact Inference for Contingency Tables. *Statistical Science*, 7(1), 131-177.
- Bergen, Benjamin K. and Nancy C. Chang. 2003. Embodied Construction Grammar in Simulation-Based Language Understanding. In Jan Ola Ostman and Mirjam Fried (eds.), *Construction Grammar(s): Cognitive and Cross-Language Dimensions*. John Benjamins: Amsterdam.
- Bryant, John. 2004. Scalable Construction-Based Parsing and Semantic Analysis. In: *Proceedings of the Second International Workshop on Scalable Natural Language Understanding*. Boston 2004.
- Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge: MIT Press.
- Fisher, R.A. 1922. "On the interpretation of χ^2 from contingency tables, and the calculation of P". *Journal of the Royal Statistical Society* 85(1):87-94.
- Gries, St.Th., and Stefanowitsch, A. 2004. Extending colostruational analysis: A corpus-based perspectives on 'alternations'. *International Journal of Corpus Linguistics*, 9, 97-129.
- Hare, M. L., McRae, K., and Elman, J.L. 2003. Sense and structure: Meaning as a determinant of verb subcategorization preferences. *Journal of Memory and Language*, 48(2), 281-303.
- Nelson, Gerald. 1996. The Design of the Corpus. In Greenbaum, Sidney (ed.) *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon Press, 27-35.
- Roland, D., and Jurafsky, D. 2002. Verb sense and verb subcategorization probabilities. In S. Stevenson, and P. Merlo (Eds.), *The Lexical Basis of Sentence Processing: Formal, Computational, and Experimental Issues*. Amsterdam: John Benjamins.
- Stefanowitsch, Anatol and Stefan Th. Gries. 2003. Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics*, 8(2), 209-243.

seit 1558





seit 1558

Thank you.

And special thanks to Mary Hare, Ken McRae and Jeff Elman for providing me with their original data

