Approaching grammar

The lexicon-grammar continuum

Holger Keibel

March 18, 2008

TUFS, Global COE International Workshop
"New Approaches in Corpus Linguistics"

Overview

- Introduction: Goals and background
- Syntagmatic usage patterns:
  - *higher-order collocations*
  - *syntagmaic patterns*
- Paradigmatic usage patterns
  - *collocational schemas*
- Summary and conclusions

Goals and background

- grammatical modeling of language use
  - no complete grammar model
  - no system of abstract symbols and rules operating on them
  - but: local stochastic models
- inductive approach
  - bottom-up: starting from individual lexical items
  - generalize incrementally (in small steps)
  - data-driven: large corpora, automated methods, no/little a-priori theory
- but guided by deduction
  - psychological assumptions
  - operationalized in automated methods

Incremental generalizations

- syntagmatic usage patterns:
  - contiguous collocations (n-grams)
    `large majority, afraid of, once in a while,`
    `in the world, upside down, lost and found, ...`
  - discontinuous chunks
    `(He)` **`asked`** `(them)` **`whether`**
- paradigmatic usage patterns:
  - e.g. partially abstract schemas

```
      He   asked      them      whether
    Jane   asked  her mother    whether

→     *    asked       *        whether
```

General psychological hypothesis

- generalizations are psychologically real
  - as part of our procedural linguistic knowledge
  - as a result of linguistic experience
  - retrieved and processed as a whole
  - facilitate language processing and production
    → influence on language use (emergence)
- cf. Sinclair's idiom principle:
  "A language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments." (Sinclair 1991:110)

Related work

- *Collocational frameworks* (Renouf & Sinclair 1991)
- *Local Grammar* (Gross 1997)
- *Pattern Grammar* (Hunston & Francis 2000)
- *Local Grammar Patterns* (Mason 2004)
- *Linear Unit Grammar* (Sinclair & Mauranen 2006)
- "Multi-word units as a model of grammar" (Mason 2007)

Corpus data

The analyses presented in this talk are based on a 2.2 billion words subset of the *Mannheim German Reference Corpus* (DEREKO), the largest corpus archive of contemporary written German.

`http://www.ids-mannheim.de/kl/projekte/korpora/`

Syntagmatic usage patterns: Possible concepts

- *n-gram*
  - contiguous sequence of *n* words that co-occur together more frequently than would be expected by mere chance
  - examples:
    `large majority, afraid of, once in a while, in the world, upside down, lost and found, ask whether, ...`
- only of limited use here:

  many syntagmatic usage patterns are discontinuous (particularly in German!)
  - e.g., **`ask`** `<someone>` **`whether`**

Syntagmatic usage patterns: Possible concepts

- *positional n-gram*
  - a set of *n* words that co-occur with fixed relative positions more frequently than would be expected by mere chance
  - thus, the following examples constitute different positional n-grams
  - example 1:

    ask ____ whether

    as in `I came to ask you whether ...`
  - example 2:

    ask ____ _____ whether

    as in `I came to ask my brother whether ...`
- the concept is still too inflexible:
  examples 1 and 2 should rather be considered belonging to the same pattern!

Higher-order collocation

- a set of *n* words that co-occur more frequently than would be expected by mere chance
- potentially non-contiguous (unlike *n-grams*)
- may occur with varying relative positions (unlike *positional n-grams*)
- **Example:** A higher-order collocation of the words *why*, *reason*, and *the* would be instantiated by each of the following sentence fragments.
  - `This is the reason why you should always ...`
  - `This is the very best reason why you should ...`
  - `But the teacher found no good reason for why ...`

Higher-order collocation 2

In particular: A higher-order collocation may occur with different sequential word orders.

Example: The higher-order collocation of the collocates *why*, *reason*, and *the* is instantiated by each of the following sentence fragments.

- `This is the reason why you should always ...`
- `She asked why knowing the reason is so important.`
- `Now the students are wondering why you would need a reason to ...`

Higher-order collocation 3

- sentence fragments covered by one h.-o. collocation
  - `This is the reason why you should always ...`
  - `This is the very best reason why you should ...`
  - `She asked why knowing the reason is so important.`
  - `But the teacher found no good reason for why ...`
  - `Now the students are wondering why you would need a reason to ...`
- Is this concept too unrestricted?
- No!
- 3 counterarguments:
  practical, methodological, empirical

Why "higher-order"?

The algorithm by which higher-order collocations are detected is an iteratively applied extension to the family of standard algorithms for detecting simple collocations (such as n-grams and positional n-grams).

In this terminology, simple collocations are first-order collocations because they can be computed in one single computational cycle.

Syntagmatic pattern

- A higher-order collocation may occur with *different sequential word orders*.
- A *syntagmatic pattern* is one such word order, together with wild-card symbols indicating where other words may occur between the collocates.
- To improve legibility, a syntagmatic pattern is usually presented together with the words that occur most frequently in some of these wild-card positions.

Check it out!

The results of a systematic and large-scale detection of higher-order collocations and their dominant syntagmatic patterns in written German (viz., in the DeReKo corpus) can be browsed in our *research and development workbench* **CCDB** (Belica 2001-2007; Keibel & Belica 2007) at

`http://corpora.ids-mannheim.de/ccdb/`

Example

higher-order collocations for the word ***machen***
(English: *to make, to do*),
together with their dominant syntagmatic patterns

taken from CCDB at

`http://corpora.ids-mannheim.de/ccdb/`

-2	-1	11238	**aufmerksam** worden	31	96%	darauf aufmerksam gemacht [...] worden
-2	-1	11238	aufmerksam wollte	17	52%	aufmerksam machen wollte
-2	-1	11238	aufmerksam Öffentlichkeit	12	66%	die Öffentlichkeit [darauf|auf ...] aufmerksam [zu] machen daß der
-2	-1	11238	aufmerksam	1617	45%	aufmerksam [zu] machen
-1	5	6970	**Spaß** richtig viel	1	100%	macht ... richtig viel Spaß
-1	5	6970	Spaß richtig	53	50%	macht [...] richtig [...] Spaß
-1	5	6970	Spaß viel	172	37%	Es macht [... sehr] viel [...] Spaß
-1	5	6970	Spaß einfach	63	50%	Es macht [...] einfach [...] Spaß
-1	5	6970	Spaß	1459	40%	macht [...] Spaß
-2	4	6186	**deutlich** habe	59	76%	habe [...] deutlich gemacht daß|dass ...
-2	4	6186	deutlich	1965	23%	deutlich [...] gemacht dass|daß ...
-2	-1	6051	**geltend** werden können	26	69%	geltend gemacht werden [...] können
-2	-1	6051	geltend werden	106	90%	geltend [...] gemacht [...] werden
-2	-1	6051	geltend können	60	45%	geltend machen [...] können
-2	-1	6051	geltend hatten	12	83%	hatten [in ...] geltend gemacht daß ...
-2	-1	6051	geltend	810	44%	geltend [zu] machen
-2	-1	4188	**rückgängig** werden kann wird	1	100%	wird ... rückgängig gemacht werden kann
-2	-1	4188	rückgängig werden kann	15	73%	rückgängig gemacht werden kann
-2	-1	4188	rückgängig werden wird	2	50%	wird ... rückgängig gemacht werden
-2	-1	4188	rückgängig werden	86	94%	rückgängig gemacht [...] werden
-2	-1	4188	rückgängig kann wird	2	50%	wird ... rückgängig gemacht ... kann
-2	-1	4188	rückgängig kann	26	42%	rückgängig gemacht werden kann
-2	-1	4188	rückgängig wird	22	50%	die ... rückgängig gemacht [...] wird
-2	-1	4188	rückgängig	445	56%	rückgängig [zu] machen
-1	-1	3829	**Fehler** haben worden	1	100%	haben ... Fehler gemacht ... worden
-1	-1	3829	Fehler haben habe	1	100%	haben ... Fehler gemacht habe
-1	-1	3829	Fehler haben	109	45%	Wir haben [... einen ...] Fehler gemacht
-1	-1	3829	Fehler worden	43	95%	Fehler [...] gemacht [...] worden
-1	-1	3829	Fehler habe	83	61%	Ich habe [einen] Fehler gemacht
-1	-1	3829	Fehler	965	48%	einen Fehler [...] gemacht
-2	-1	3721	**Gebrauch** wird Angebot	1	100%	Angebot wird ... Gebrauch gemacht
-2	-1	3721	Gebrauch wird	25	48%	wird ... Gebrauch gemacht
-2	-1	3721	Gebrauch Angebot	25	40%	von dem|diesem Angebot [...] Gebrauch gemacht
-2	-1	3721	Gebrauch	595	42%	Gebrauch [zu] machen
-1	-1	3715	**verantwortlich** werden	106	90%	verantwortlich [...] gemacht [...] werden
-1	-1	3715	verantwortlich wird	29	68%	verantwortlich gemacht wird
-1	-1	3715	verantwortlich	767	50%	verantwortlich [...] gemacht werden

English example

higher-order collocations for the word ***why***,
together with their dominant syntagmatic patterns,

computed on a
small (2.5 million words)
web-based corpus
of written English

-1	-1	1805	**reason** one main	1	100% reason ... main one why
-1	-1	1805	reason one	64	96% is one reason [...] why the ...
-1	-1	1805	reason main	21	90% The|the main reason why the ...
-1	-1	1805	reason One	24	100% One reason [...] why the ...
-1	-1	1805	reason	181	100% is one reason [...] why the ...
-1	-1	1282	**explain** helps	23	100% This helps [to] explain [...] why ... the
-1	-1	1282	explain may help	3	100% may help [to] explain why
-1	-1	1282	explain may	28	96% This may [...] explain why the ...
-1	-1	1282	explain help	13	100% may|might help [to] explain why
-1	-1	1282	explain	113	100% helps may|to explain [...] why
-1	-1	575	**is** That	78	83% That [...] is [...] why the ...
-1	-1	575	is easy **It**	13	92% It is easy to see why
-1	-1	575	is easy	19	89% It|it is [...] easy to see why
-1	-1	575	is **It**	21	90% It is easy to see why the
-1	-1	575	is	393	71% That is [... reason] why the ...
-1	-1	543	**explains** This partly	3	100% This [...] partly explains why
-1	-1	543	explains This	12	100% This [partly] explains why
-1	-1	543	explains partly	7	100% This partly explains why the ...
-1	-1	543	explains	49	100% This explains [...] why the ...
-1	-1	528	**reasons** There are several	7	85% There are several reasons why
-1	-1	528	reasons There are	23	78% There are several|two reasons why
-1	-1	528	reasons There	24	100% There are several|many reasons why
-1	-1	528	reasons are several	9	88% There are several reasons why
-1	-1	528	reasons are	34	82% There|there are [several|two] reasons why the ...
-1	-1	528	reasons several	11	100% There|there are several reasons why
-1	-1	528	reasons	57	100% are ... reasons [...] why the ...
-4	-2	527	**That** Fund	3	100% That is why [the] Fund
-4	-2	527	That	106	100% That is ... why the ...
-2	-2	302	**one**	76	93% is one [reason] why the ...
-1	-1	247	**see** hard **It**	1	100% It ... hard ... see why
-1	-1	247	see	43	95% easy|hard to see [...] why
-4	-2	235	**This**	75	100% This is ... explain why
1	3	226	**should** no	14	92% there is no reason why [...] should not ...
1	3	226	should be	16	93% why [...] should [...] be
1	3	226	should	65	98% reason why [...] should be ...
			helps	24	95% This helps to explain why ... the

Collocational schema

- generalization across collocations
  or across syntagmatic patterns
- **example:**

```
    there are [...] several reasons why ...
    there are [...]  some   reasons why ...
 →  there are [...]    *     reasons why ...
```

- *slots* and *fillers*:
  The set of fillers for a given slot is no predefined and language-general abstract category
  (as in generative grammars);
  instead it is specific to this very schema and this very slot
  (similar to some construction grammar approaches,
  e.g., Croft 2001).

Collocational schema 2

Q: Is this a feasible approach? Are there any interesting collocational schemas to be discovered in the data?

A: We do not know yet.

Q: If so, how exactly would one find them?

A: We do not know yet.

The core problems to be solved:

- devise a clever method that groups together *similar* collocations/patterns and generalizes them to a schema
- devise a set of measures that assess complementary notions of *similarity* between any two collocations/patterns
- devise a suitable evaluation heuristics

Where to begin?

Some syntagmatic patterns one would want to group together in one schema.

nicht um Apologie sondern um
geht|es nicht [...] um [die|eine] Bestrafung [von|der ...] sondern um eine
nicht um ... blindwütige ... sondern um
nicht um [...] Effekthascherei sondern [...] um
nicht ... um Ehrverletzung sondern ... um
nicht um die Feststellung ... sondern um
geht|es nicht [...] um [...] Ideologie [...] sondern [...] um
es dabei nicht [nur] um [den] Kommerz [geht] sondern [auch] um
nicht um Kuriositäten [...] sondern um
nicht um mildtätige ... sondern [...] um
geht|es nicht um Parteipolitik [...] sondern um die ...
nicht [...] um [...] Parteitaktik [...] sondern um
nicht [...] um Pietät sondern um
es nicht wie gewohnt um Retuschen sondern um
es|hier nicht [...] um [eine] sachliche [...] sondern [...] um
es nicht um [eine] Verteufelung [des gehen ...] sondern um
es nicht [...] um [die] Wahrheitsfindung [...] sondern [...] um

How would a schema look like?

nicht *um* * *sondern* *um*

			nicht		um			Apologie			sondern		um		
geht	es			[...]		[die|eine]		Bestrafung	[von|der ...]					eine	
						...	blindwütige	...							
							[...]	Effekthascherei				[...]			
				...				Ehrverletzung				...			
						die		Feststellung ...							
geht	es			[...]			[...]	Ideologie	[...]			[...]			
	es	dabei		[nur]		[den]		Kommerz		[geht]		[auch]			
								Kuriositäten		[...]					
							mildtätige	...				[...]			
geht	es							Parteipolitik	[...]					die	...
				[...]			[...]	Parteitaktik	[...]						
				[...]				Pietät							
	es			wie gewohnt				Retuschen							
	es	hier		[...]		[eine]	sachliche	[...]				[...]			
	es					[eine]		Verteufelung	[des	gehen ...]					
	es			[...]		[die]		Wahrheitsfindung	[...]			[...]			

What is (in) a schema?

- A collocational schema is a quantitative preference-relational structure that is partly abstract (→ the slot).
- It captures the complex range of similarity relations between the underlying collocations/patterns.
- a starting point for this specific example
  - Use the collocates as fixed points.
  - Analyze the internal structure of the discontinuities/gaps between these collocates
    in terms of the lexical items observed in these gaps.
  - Analyze the lexical variation in the slot(s) and try to characterize the paradigmatic class of (observed) fillers.

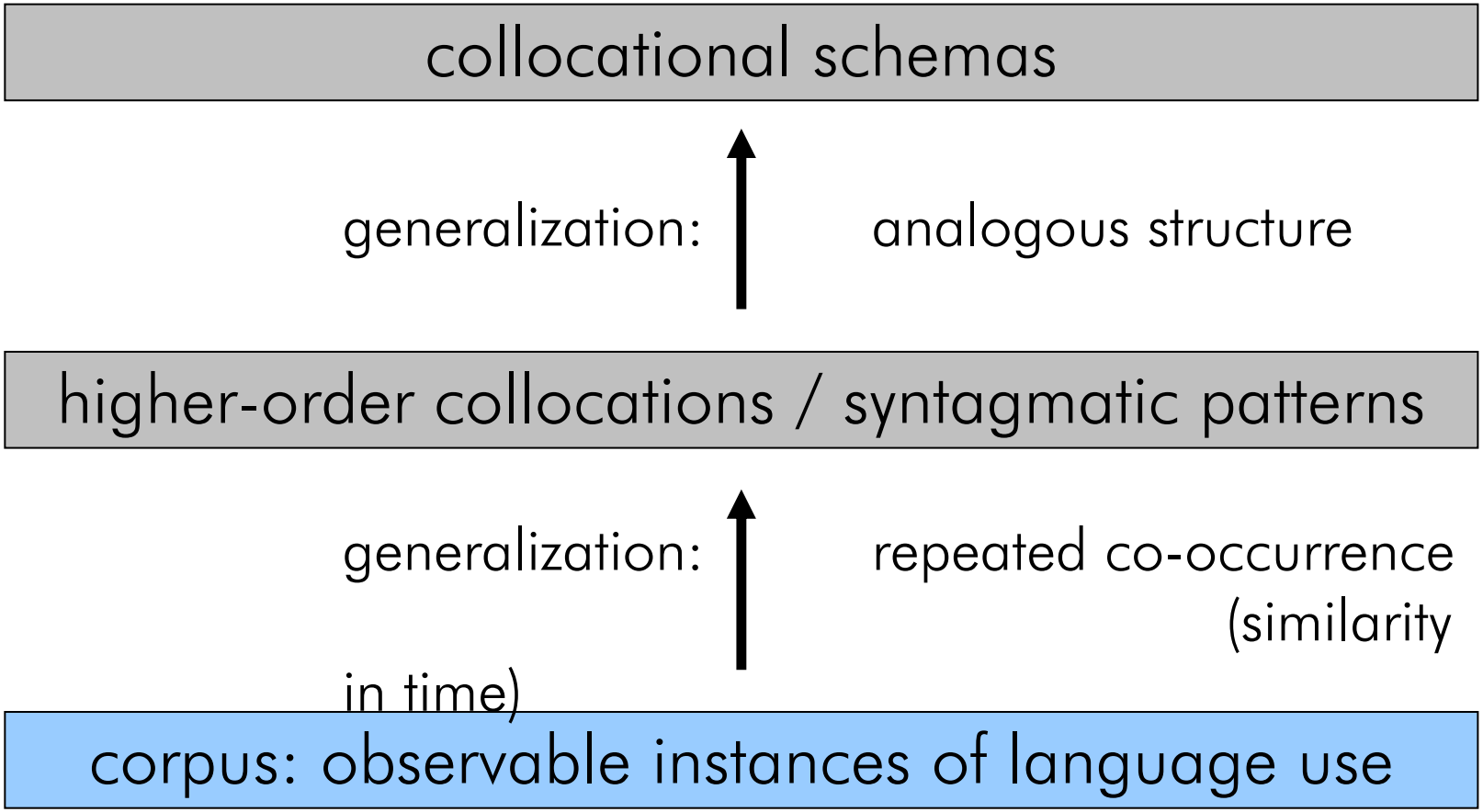
Are we on the right track?

- a very primitive ad-hoc search for collocational schemas across syntagmatic patterns
- results: very reassuring
- some examples of schema *names*:
  - `nicht nur * sondern auch`
  - `mit dem * ausgezeichnet worden`
  - `der größte * der Welt`
  - `mit einem * mit dem`
  - `Die * Regierung hat die`
  - `als Sohn eines * geboren`
  - `* in den Griff bekommen`

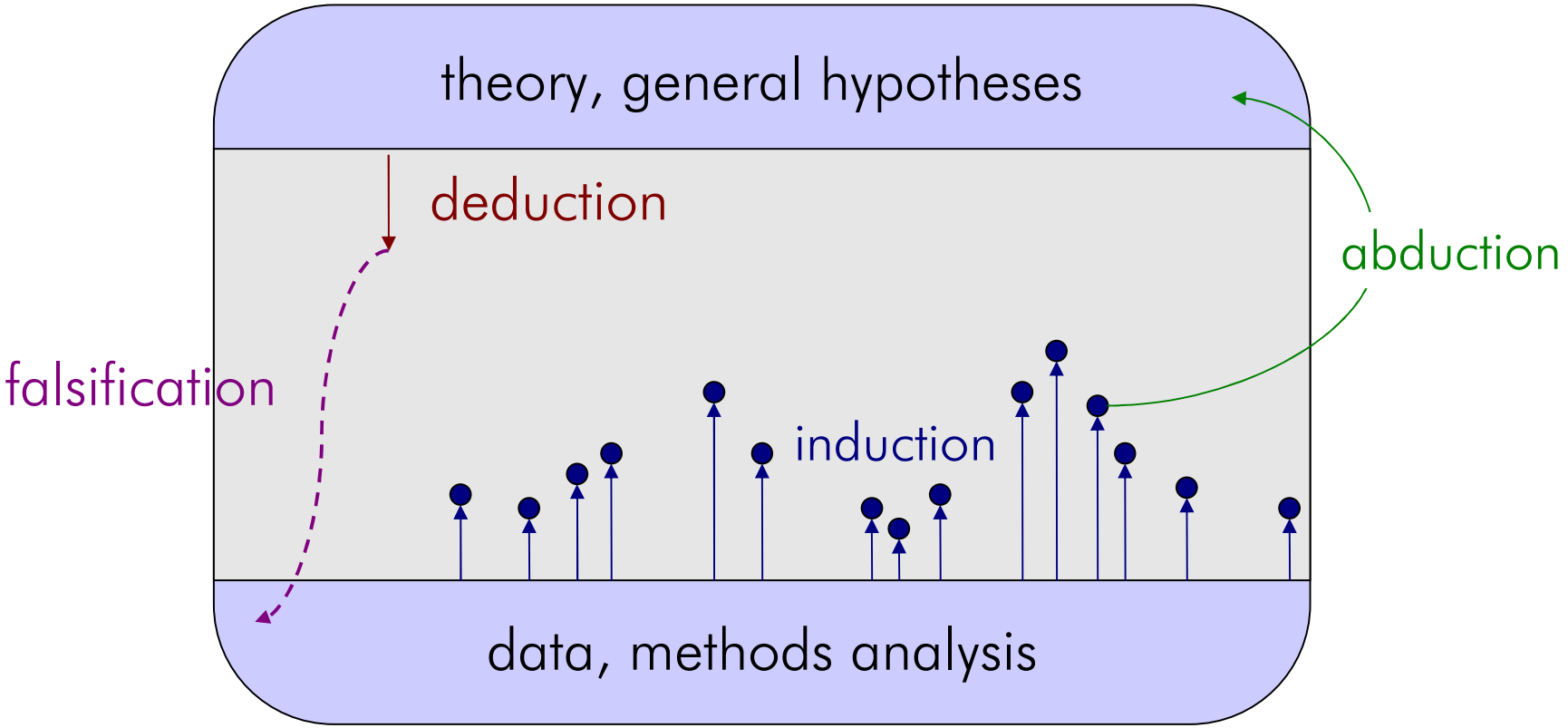
Examples with English translation

- nicht nur * sondern auch
  not only * but also
- mit dem * ausgezeichnet worden
  was awarded the *
- der größte * der Welt
  the greatest/largest * in the world
- mit einem * mit dem
  with a * with the
- Die * Regierung hat die
  The * government has the
- als Sohn eines * geboren
  born as the son of a *
- * in den Griff bekommen
  get * under control

Summary of generalizations



Local Models and the Explanatory Gap



Conclusions

- elaborate concepts and mature operationalization: *higher-order collocation* and *syntagmatic pattern*
- very preliminary operationalization of the naïve concept of *collocational schema*
  → already yields meaningful generalizations
- first steps from lexis towards syntax
  → new view on syntax?
- highly relevant for:
  - linguistic theory
  - (first/second) language acquisition
  - lexicography

Relevance for language acquisition

- general hypothesis: *higher-order collocations*, *syntagmatic patterns*, and *collocational schemas* are psychologically real
  - as part of our procedural linguistic knowledge
  - as a result of linguistic experience
  - retrieved and processed as a whole
  - facilitate language processing and production
    → influence on language use (emergence)
- cf. Sinclair's idiom principle (again):
  "A language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments." (Sinclair 1991:110)

Thank you!

korpuslinguistik@ids-mannheim.de