



# Quantitative Corpus Linguistics: A practical overview

Universität Leipzig  
*Typologisches Kolloquium*  
Summer 2009

Daniel Wiechmann  
FSU Jena

[www.daniel-wiechmann.net](http://www.daniel-wiechmann.net)



seit 1558

# Overview

---

- I. Introduction
  - I. What is Quantitative Corpus Linguistics (QCL)?
  - II. What is a linguistic corpus?
    - I. Types of corpora
- II. Fundamental notions in QCL
  - I. frequency lists
  - II. co-occurrences (ngrams)
  - III. concordances
- III. QCL meets psycholinguistics – Statistical associations
  - I. Bigrams and local syntactic ambiguity
  - II. Beyond bigrams: Patterns and processing difficulty





seit 1558

# What is Quantitative Corpus Linguistics?

---

- ▶ QCL is characterized by the *systematic* and *exhaustive analysis* of linguistic phenomena on the basis of a *linguistic corpus*.
- ▶ typically, QCL-research questions are formulated such that *conditional frequencies* constitute the *dependent variable*

Stefanowitsch (2005)





seit 1558

# What is Quantitative Corpus Linguistics?

- ▶ QCL is characterized by the *systematic* and *exhaustive analysis* of linguistic phenomena on the basis of a *linguistic corpus*.
- ▶ typically, QCL-research questions are formulated such that *conditional frequencies* constitute the *dependent variable*

e.g. Particle placement

*John picked up the book -> SUBJ V PRT DO*

*John picked the book up -> SUBJ V DO PRT*

Condition 1: “heavy” DO

Condition 2: not “heavy” DO



seit 1558

# What is a linguistic corpus?

---

- ▶ A corpus is a ...
  - ▶ collection of spoken or written text
  - ▶ machine readable





seit 1558

# What is a linguistic corpus?

---

- ▶ A corpus is a ...
  - ▶ collection of spoken or written text
  - ▶ machine readable
  - ▶ produced in a *natural* communicative setting
    - ecological validity
  - ▶ *representative* for a {variety | genre | register}
    - all parts of a {variety | genre | register} should be in it
  - ▶ *balanced* wrt a {variety | genre | register}
    - proportions the parts make up should be mirrored





seit 1558

# What types of corpora?

---

- ▶ **general vs. specific corpora**
- ▶ **synchronic vs. diachronic corpora**
- ▶ **monolingual vs. parallel corpora**
- ▶ **static vs. dynamic/monitor corpora**
- ▶ **plain vs. annotated (SGML/XML) corpora**





seit 1558

# What types of corpora?

---

- ▶ general vs. specific corpora
- ▶ **synchronic vs. diachronic corpora**
- ▶ monolingual vs. parallel corpora
- ▶ static vs. dynamic/monitor corpora
- ▶ plain vs. annotated (SGML/XML) corpora





seit 1558

# What types of corpora?

---

- ▶ general vs. specific corpora
- ▶ synchronic vs. diachronic corpora
- ▶ monolingual vs. parallel corpora
- ▶ static vs. dynamic/monitor corpora
- ▶ plain vs. annotated (SGML/XML) corpora





seit 1558

# What types of corpora?

---

- ▶ general vs. specific corpora
- ▶ synchronic vs. diachronic corpora
- ▶ monolingual vs. parallel corpora
- ▶ **static vs. dynamic/monitor corpora**
- ▶ plain vs. annotated (SGML/XML) corpora





seit 1558

# What types of corpora?

---

- ▶ general vs. specific corpora
- ▶ synchronic vs. diachronic corpora
- ▶ monolingual vs. parallel corpora
- ▶ static vs. dynamic/monitor corpora
- ▶ plain vs. annotated (SGML/XML) corpora
  - phonologically annotated
  - morphologically annotated
  - **POS-tagged**
  - Syntactically parsed





seit 1558

From:

Brown University Standard Corpus of Present-Day American English  
- **Brown** (1 million words)

```
C:\Documents and Settings\...
File Edit Search
A00 S1A-001.COR BROWN2.TXT
1 A01-0010-01-000000 The Fulton County Grand Jury said Friday an investigation
2 A01-0020-01-000000 of Atlanta's recent primary election produced "no evidence"
3 A01-0020-09-000000 that any irregularities took place.
4 A01-0030-05-000000 The jury further said in term-end presentments that
5 A01-0040-03-000000 the City Executive Committee, which had over-all charge
6 A01-0050-02-000000 of the election, "deserves the praise and thanks of
7 A01-0050-11-000000 the City of Atlanta" for the manner in which the election
8 A01-0060-11-000000 was conducted.
9 A01-0070-01-000000 The September-October term jury had been charged
10 A01-0070-09-000000 by Fulton Superior Court Judge Durwood Pye to investigate
11 A01-0080-08-000000 reports of possible "irregularities" in the hard-fought
12 A01-0090-06-000000 primary which was won by Mayor-nominate Ivan Allen
13 A01-0100-05-000000 Jr&.
14 A01-0100-06-000000 "Only a relative handful of such reports was received",
15 A01-0110-06-000000 the jury said, "considering the widespread interest
16 A01-0120-03-000000 in the election, the number of voters and the size
17 A01-0130-01-000000 of this city".
18 A01-0130-04-000000 The jury said it did find that many of Georgia's
19 A01-0140-01-000000 registration and election laws "are outmoded or inadequate
20 A01-0150-01-000000 and often ambiguous".
21 A01-0150-04-000000 It recommended that Fulton legislators act "to have
22 A01-0160-04-000000 these laws studied and revised to the end of modernizing
23 A01-0170-01-000000 and improving them".
24 A01-0170-04-000000 The grand jury commented on a number of other topics,
25 A01-0180-04-000000 among them the Atlanta and Fulton County purchasing
26 A01-0190-01-000000 departments which it said "are well operated and follow
```

Normal text nb char : 7920350 | Ln:1 Col:76 Sel:0 | Dos\Windows ANSI | INS



seit 1558

From:  
**British National Corpus**  
**BNC** (100 million words)

```
C:\Documents and Settings\Owner\Desktop\corpora\BNC Edited\Written Data\
File Edit Search View Format Language Settings Macro Run TextFX
A00 S1A-001.COR new 2
1 <bncDoc id=A00>
2 <teiHeader type="text" status="update" date.updated="2000-12-13"><fileDesc><
3 <text decls='CN004 HN001 QN000 SN000'><body>
4 <div1 n=1 type=item>
5 <head type=MAIN>
6 <s n="1"><w NN1>FACTSHEET <w DTQ>WHAT <w VBZ>IS <w NN1>AIDS<c PUN>?
7 </head>
8 <p>
9 <s n="2"><hi rend=bo><w NN1>AIDS <c PUL> (<w VVN-AJO>Acquired <w AJO>Immune <
10 <s n="3"><w DT0>This <w NN1>virus <w VVZ>affects <w AT0>the <w NN1>body<w PO;
11 </p>
12 <div2 type="u">
13 <head rend=bo type=MAIN>
14 <s n="4"><w AVQ>How <w VBZ>is <w NN1>infection <w VVN-VVD>transmitted<c PUN>
15 </head>
16 <list>
17 <item>
18 <s n="5"><w PRP>through <w AJO>unprotected <w AJO>sexual <w NN1>intercourse <
19 </item>
20 <item>
21 <s n="6"><w PRP>through <w AJO>infected <w NN1>blood <w CJC>or <w NN1>blood <
22 </item>
23 <item>
24 <s n="7"><w PRP>from <w AT0>an <w AJO>infected <w NN1>mother <w PRP>to <w DP;
25 </item>
```

exTensible nb char : 109707 Ln:1 Col:16 Sel:0 Dos\Windows ANSI INS



seit 1558

```
*new 2 - Notepad++
File Edit Search View Format Language Settings Macro Run TextFX Plugins
Window ?
A00 S1A-001.COR new 2
1  [<#30:1:B> ·<sent>]
2  PU, CL (main, cop, pres)
3  ·SU, NP (supersede)
4  [<> ·<->]
5  ·NPHD, PRON (dem, sing, ignore) ·{This}
6  [</-> ·<=>]
7  ·NPHD, PRON (dem, sing) ·{this}
8  ·VB, VP (cop, pres)
9  [</=> ·</>>]
10 ·MVB, V (cop, pres) ·{is}
11 ·CS, NP ()
12 ·DT, DTP ()
13 ·DTCE, ART (indef) ·{a}
14 ·NPHD, N (com, sing) ·{dance ·group}
15 ·NPPO, CL (depend, rel, montr, pres)
16 ·SU, NP ()
17 ·NPHD, PRON (rel) ·{which}
18 ·VB, VP (montr, pres, do)
19 ·OP, AUX (do, pres) ·{does}
20 ·A, AVP (ge)
21 ·AVHD, ADV (ge) ·{not}
22 ·MVB, V (montr, infin) ·{exclude}
23 ·OD, NP ()
24 ·NPHD, N (com, plu) ·{people}
nb Ln: 23 Col: 4 Sel: 0 Dos\Windows ANSI INS
```

From:  
International **C**orpus of **E**nglish  
(British Component)  
**ICE-GB** (1 million words)



seit 1558

# Fundamentals in QCL

---

## Frequency lists, Concordances and Co-occurrences

Wiechmann and Fuhs (2006) evaluate 10 concordancers



Daniel Wiechmann – FSU Jena



seit 1558

# Fundamentals in QCL

---

## Frequency lists, Concordances and Co-occurrences



Daniel Wiechmann – FSU Jena



seit 1558

# Fundamentals in QCL Frequency lists

Count	Pct	Word
5624123	6.1950%	the
2869761	3.1610%	of
2354005	2.5929%	to
2342302	2.5800%	and
1963840	2.1632%	a
1796346	1.9787%	in
891945	0.9825%	is
885588	0.9755%	that
829362	0.9135%	for
793700	0.8743%	it
787379	0.8673%	was
756040	0.8328%	bquo
738536	0.8135%	equo
641979	0.7071%	on
610358	0.6723%	as
608010	0.6697%	with
584470	0.6438%	be
578009	0.6367%	's
556413	0.6129%	he

Frequency list (excerpt): BNC written (3112 files ; ~ 90 million words)



seit 1558

# Fundamentals in QCL

## Frequency lists

Count	Pct	Word
5624123	6.1950%	the
2869761	3.1610%	of
2354005	2.5929%	to
2342302	2.5800%	and
1963840	2.1632%	a
1796346	1.9787%	in
891945	0.9825%	is
885588	0.9755%	that
829362	0.9135%	for
793700	0.8743%	it
787379	0.8673%	was
756040	0.8328%	bquo
738536	0.8135%	equo
641979	0.7071%	on
610358	0.6723%	as
608010	0.6697%	with
584470	0.6438%	be
578009	0.6367%	's
556413	0.6129%	he

Word frequency lists are useful for...

- choosing **experimental stimuli** correctly
  - e.g. lexical decision
- **corpus comparison**  
(-> stop lists)

Frequency list (excerpt): BNC written (3112 files ; ~ 90 million words)



seit 1558

# Fundamentals in QCL

---

## Frequency lists, Concordances and Co-occurrences



Daniel Wiechmann – FSU Jena



seit 1558

# Co-occurrences -> Collocations

---

LEXIS

## Collocation

(lexical associations)

<i>strong</i>	versus	<i>powerful</i>
<i>different from</i>	versus	<i>different than</i>
<i>alphabetic</i>	versus	<i>alphabetical</i>



seit 1558

# Co-occurrences -> Collocations

LEXIS

## Collocation

(lexical associations)

<i>strong</i>	versus	<i>powerful</i>
<i>different from</i>	versus	<i>different than</i>
<i>alphabetic</i>	versus	<i>alphabetical</i>

- ▶ Extract all occurrences of target construction from corpus
- ▶ Call up *concordance*

A Concordance of word  $w$  is a display of every occurrence of  $w$  in the corpus together with a user specified context (KWIC-display (key word in context display))

Concordance - [ $\langle w \rangle^*$ alphabetical

```

</p>
<p>
<s n="417"><w DT0>All
</p>
<p>
<s n="418"><w AV0>Meanwhile<c PUN>, <w NP0>Leonard <w NP0>Bernstein <w VBZ>is <w VBG>being <w VVN>celebrated <w PRP>by
<s n="419"><w DT0>This <w VBZ>is <w VVN>entitled <w AT0>the <c PUQ>&#39;<w AJ0>Royal <w NN1>Edition<c PUQ>&#39;; <w PR
<s n="420"><w AT0>The <w NN2>discs <w VBB>are <w T00>to <w VBI>be <w VVN>issued <w PRP>in <w AJ0>alphabetical <w NN1>or

```

... DREN<w POS>'S <w NN1>CHOICE <w DT0>This <w AJ0>alphabetical <w NN1>list <w PRF>of <w NP0>Britain<w ...  
 ... lp<c PUN>. <s n="288"><c PUL>[<w AT0>An <w AJ0>alphabetical <w NN1>directory <w PRF>of <w AJ0-NN1>s ...  
 ... lp<c PUN>. <s n="210"><c PUL>[<w AT0>An <w AJ0>alphabetical <w NN1>directory <w PRF>of <w AJ0-NN1>s ...  
 ... ; <w AV0>either <w PRP-CJS>as <w AT0>an <w AJ0>alphabetical <w NN1>list<c PUN>, <w CJC>or <w AV0>as ...  
 ... 0>more <w AJ0>useful<c PUN>; <w DPS>its <w AJ0>alphabetical <w NN1>list <w PRF>of <w NN2>foods <w V ...  
 ... 00>to <w VBI>be <w VVN>issued <w PRP>in <w AJ0>alphabetical <w NN1>order <w PRF>of <w NN1>composer< ...  
 ... re <w XX0>not <w AV0>entirely <w PRP>in <w AJ0>alphabetical <w NN1>order<c PUN>. <s n="1163"><w AV0 ...  
 ... 200 <w NN1>pupil <w NN2>files <w PRP>in <w AJ0>alphabetical <w NN1>order<c PUN>. <s n="407"><w CJS> ...  
 ... ek </hi><c PUN>, <c PUQ>&#39;<w AT0>AN <w AJ0>ALPHABETICAL <w NN1>CATALOGUE <w PRF>OF <w NN2>Names ...  
 ... w PRP>from <w AT0>the <w AJ0>felicitous <w AJ0>alphabetical <w NN2>conjunctions <w PRF>of <w NN1>Sc ...  
 ... rds <w PNP>I <w VVD>filed <w PRP-AVP>in <w AJ0>alphabetical <w NN1>order<c PUN>, <w CJC>and <w PNP> ...  
 ... ing <w DPS>my <w NN2>listings <w PRP>in <w AJ0>alphabetical <w NN1>order<c PUN>. </p> <p> <s n="122 ...  
 ... <w NN1>book <w VVZ>comprises <w AT0>an <w AJ0>alphabetical <w NN1>dictionary <c PUQ>&#39;<w PRP>i ...  
 ... 0>the <w NN1>basis <w PRP>for <w AT0>an <w AJ0>alphabetical <w NN1>sort <w PRP>within <w DT0>each < ...  
 ... n="702"><w AV0>Also<c PUN>, <w AT0>the <w AJ0>alphabetical <w NN1>author/title <w NN1>arrangement ...  
 ... ORD>last <w NN1>year<c PUN>, <w PRP>in <w AJ0>alphabetical <w NN1>artist <w NN1>order<c PUN>. <s n ...  
 ... <w AV0>now <w VVN-VVD>indexed <w PRP>in <w AJ0>alphabetical <w NN1>order <w PRF>of <w NN1>name <w P ...  
 ... <w NN2>counties <w VVN>taken <w PRP>in <w AJ0>alphabetical <w NN1>order <w PRF>of <w NN1>rotation< ...  
 ... AT0>the <w ORD>last <w PRP>in <w AT0>an <w AJ0>alphabetical <w NN1>list <w PRF>of <w NN2>MPs<c PUN> ...  
 ... CRD>two <w NN2>words<c PUN>, <w PRP>in <w AJ0>alphabetical <w NN1>order<c PUN>, <c PUQ>&#39;<w NN ...  
 ... <w CJC>and <w AJ0>new<c PUN>, <w PRP>in <w AJ0>alphabetical <w NN1>order<c PUN>, <w AV0>only <w VVG ...  
 ... >been <w VVN>put <w ORD>first <w PRP>by <w AJ0>alphabetical <w NN1>chance<c PUN>, <w XX0>not <w PRP ...  
 ... 0-NN1>GAWOR <w VVZ>continues <w DPS>his <w AJ0>alphabetical <w NN1>advice <w PRP>on <w NN0>fish <w ...  
 ... UN>, <w PRP>with <w AT0>a <w AJ0>proper <w AJ0>alphabetical <w NN1>index <w CJC>and <w DT0>much <w ...

24 matches      Original text order      Strings matching:  $\langle w \rangle^*$ alphabetical



seit 1558

# A (partial) concordance of *alphabetical*

L1	search term	R1
This	<w AJ0>alphabetical	list
An	<w AJ0>alphabetical	directory
An	<w AJ0>alphabetical	directory
an	<w AJ0>alphabetical	list
its	<w AJ0>alphabetical	list
in	<w AJ0>alphabetical	order
in	<w AJ0>alphabetical	order
in	<w AJ0>alphabetical	order
an	<w AJ0>ALPHABETICAL	CATALOGUE
felicitous	<w AJ0>alphabetical	conjunctions
in	<w AJ0>alphabetical	order
in	<w AJ0>alphabetical	order
an	<w AJ0>alphabetical	dictionary
an	<w AJ0>alphabetical	sort
the	<w AJ0>alphabetical	author/title
in	<w AJ0>alphabetical	artist
in	<w AJ0>alphabetical	order
in	<w AJ0>alphabetical	order
an	<w AJ0>alphabetical	list
in	<w AJ0>alphabetical	order
in	<w AJ0>alphabetical	order
by	<w AJ0>alphabetical	chance
his	<w AJ0>alphabetical	advice
proper	<w AJ0>alphabetical	index



seit 1558

# Collocate display of *alphabetical*

Table 2.3 A collocate display of *alphabetical* based on the BNC

Word at L1	Freq L1	Node word	Freq Node	Word at R1	Freq R1
<w prp>in	77	<w aj0>alphabetical	234	<w nn1>order	89
<w at0>an	36			<w nn1>index	15
<w at0>the	23			<w nn1>list	13
<w prf>of	6			<w nn1>indexing	12
<w cjc>and	6			<w nn1>subject	12
<c pun>.	6			<w nn1>sequence	11
<c pun>,	6			<w nn1>listing	9
<w aj0>ascending	5			<w nn1>guest	6
<w cjc>or	5			<w cjc>and	5
<w aj0>strict	4			<w nn1>description	2

From Gries (2009)



seit 1558

# Collocate display of *alphabetic*

Table 2.2 A collocate display of *alphabetic* based on the BNC

Word at L1	Freq L1	Node word	Freq Node	Word at R1	Freq R1
<w prf>of	8	<w aj0>alphabetic	42	<w nn1>literacy	7
<w at0>the	6			<w nn1>writing	5
<w at0>an	5			<w nn1>order	3
<w prp>in	2			<w nn1>character	3
<w prp>such as	2			<w cjc>and	2
<w dps>our	2			<w nn1>system	2
<w cjs>when	2			<w nn2>characters	2
<w aj0 widespread>	1			<w nn1>culture	2
<w nn2>systems	1			<w prp>in	1
<w aj0>varying	1			<c pun>.	1

From Gries (2009)



seit 1558

# Fundamentals in QCL

## Distinctive collocates

		Modified noun		
		+ order	- order	
Target word	+ alphabetic	<b>O 11</b>		<b>R 1</b>
	- alphabetic			
		<b>C 1</b>		<b>N (Grand Total)</b>

Frequency signature of a pair = { O11, R1, C1, N }



seit 1558

# Fundamentals in QCL

## Distinctive collocates

		Modified noun		
		+ order	- order	
Target word	+ alphabetic	<b>O 11</b>	O 12	<b>R1</b>
	- alphabetic	O 21	O 22	R2
		<b>C1</b>	C2	<b>N (Grand Total)</b>

Frequency signature of a pair = { O11, R1, C1, N }



seit 1558

# Fundamentals in QCL

## Distinctive collocates

		Modified noun		
		+ order	- order	
Target word	+ alphabetic	O 11 E 11	O 12 E 12	R1
	- alphabetic	O 21 E 21	O 22 E 22	R2
		C1	C2	Grand Total

- ▶ Compare observed with expected frequencies via some measure of statistical association
- ▶ Identify statistically significant collocates  
(Maybe *alphabetical* and *alphabetic* modify different types of nouns)

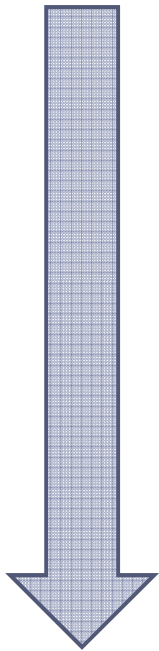


seit 1558

# Fundamentals in QCL

## Co-occurrences

LEXIS



GRAMMAR

### Collocation

(lexical associations)

<i>strong</i>	versus	<i>powerful</i>
<i>different from</i>	versus	<i>different than</i>
<i>alphabetic</i>	versus	<i>alphabetical</i>

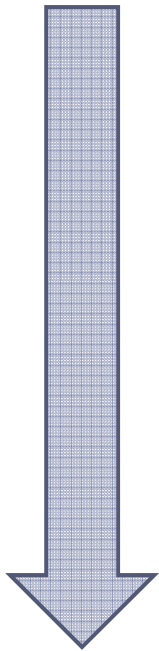


seit 1558

# Fundamentals in QCL

## Co-occurrences

LEXIS



GRAMMAR

### Collocation

(lexical associations)

<i>strong</i>	versus	<i>powerful</i>
<i>different from</i>	versus	<i>different than</i>
<i>alphabetic</i>	versus	<i>alphabetical</i>

### Colligation

(functional associations)

e.g. preferred gram. function or POS  
of a word

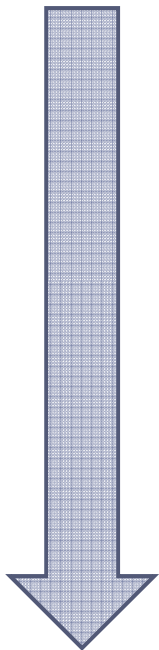


seit 1558

# Fundamentals in QCL

## Co-occurrences

LEXIS



GRAMMAR

### Collocation

(lexical associations)

*strong*                      versus    *powerful*  
*different from*        versus    *different than*  
*alphabetic*            versus    *alphabetical*

### Colligation

(functional associations)

e.g. preferred gram. function or POS  
of a word

### Collostruction

(*associations between word and syn. pattern*)

e.g. *give* in ditransitive construction



# QCL – More domains of application

## Language comprehension

---

Corpus-derived predictions about  
human language processing

### I. Bigram information

- I. Association strength between verbs  
and complementation patterns





seit 1558

# Associations and probabilistic parsing

## NP/S local syntactic ambiguity



Inspector Clouseau suspected [the phantom ]...  
SUBJECT  $V_{trans}$  [ NP ]...

...right away  
NP is DO

...was nobody else but ...  
NP is subject of embedded clause

TIME

$t_i$





seit 1558

# Approaching the issue from a QCL point of view

- Task 1: Corpus choice
- Task 2: Extract all [V NP]-sequences from a corpus
- Task 3: Annotate -> develop data frame
- Task 4: Statistical analysis



seit 1558

## Approaching the issue from a QCL point of view

### Task1: Corpus choice

	BNC	ICE-GB
Size	100 million	1 million
POS tagged	+	+
Parsed	-	+



seit 1558

## Approaching the issue from a QCL point of view

### Task1: Corpus choice

	BNC	ICE-GB
Size	100 million	1 million
POS tagged	+	+
Parsed	-	+

### Task 2: Extract all [V NP]-sequences from a corpus

This involves the design of a search string (**regular expression**) that matches all and only those patterns that are instances of the target construction



seit 1558

## Approaching the issue from a QCL point of view

### Task 3: Annotate -> develop data frame

case	example.or.corpusID	VERB_LEMMA	SYNTAX
1	bla1	suspect	DO
2	bla2	acknowledge	DO
3	bla3	admit	S
4	bla4	know	S
5	bla5	understand	S
6	bla6	discover	DO
7	bla7	announce	DO
8	bla8	write	S
13	bla13	suggest	DO
14	bla14	realize	DO
15	bla15	believe	S
16	bla16	notice	DO
17	bla17	feel	S
18	bla18	deny	DO
...	...	...	...
4209	bla4209	believe	DO



seit 1558

## Approaching the issue from a QCL point of view

► Task 4: Statistical analysis:  
Measure association strength  
Choose test and compute lexical bias towards complementation  
patterns from frequency signatures

	sentential object	nominal object	
+ <i>suspected</i>	O 11 E 11	O 12 E 12	R1
- <i>suspected</i>	O 21 E 21	O 22 E 22	R2
	C1	C2	Grand Total

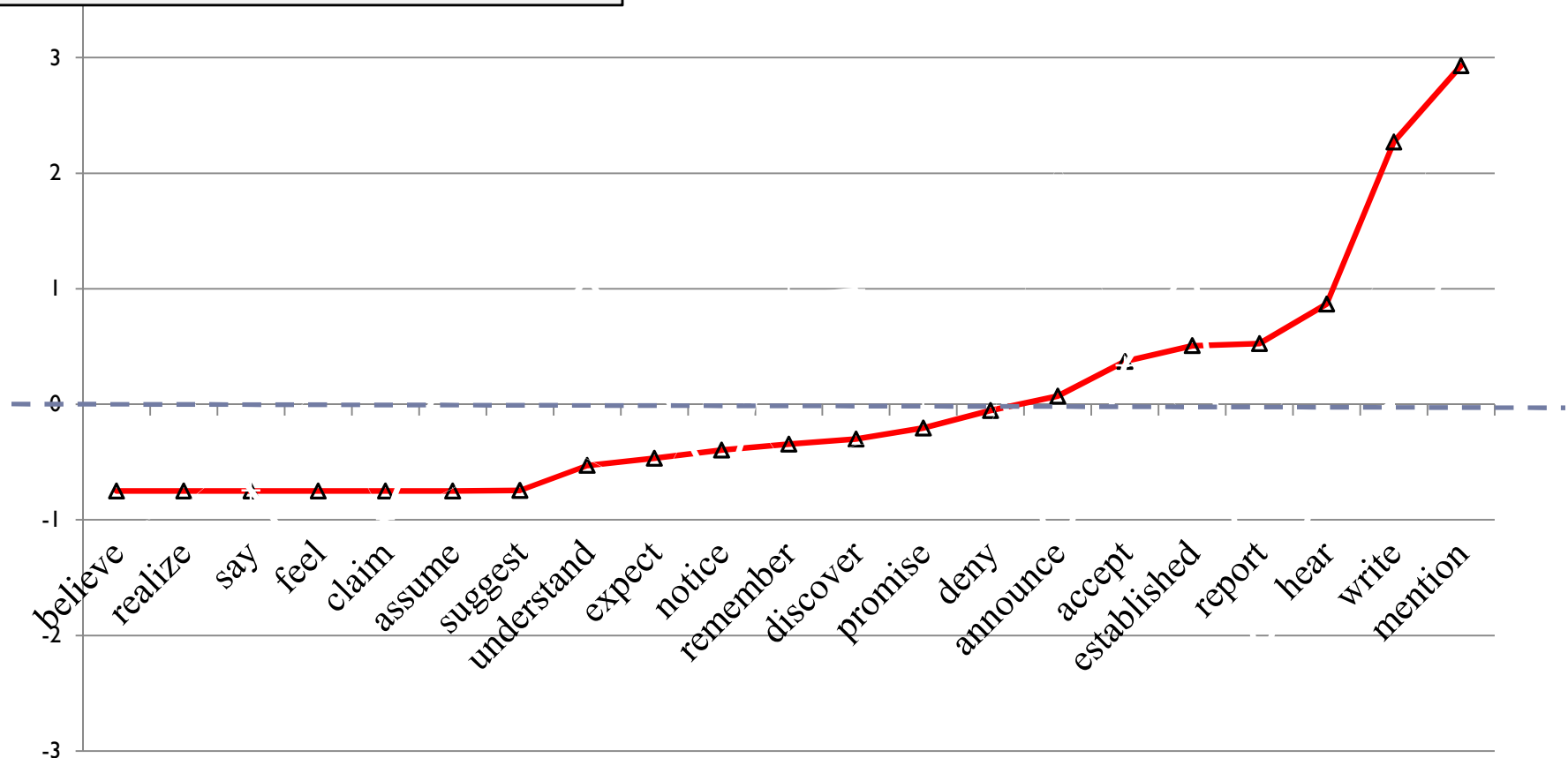


# Fundamentals in QCL Statistical Associations

Bias towards NP complements  
Measure: *Fisher exact test*

minimum sensitivity (MS)

discounted odds ratio

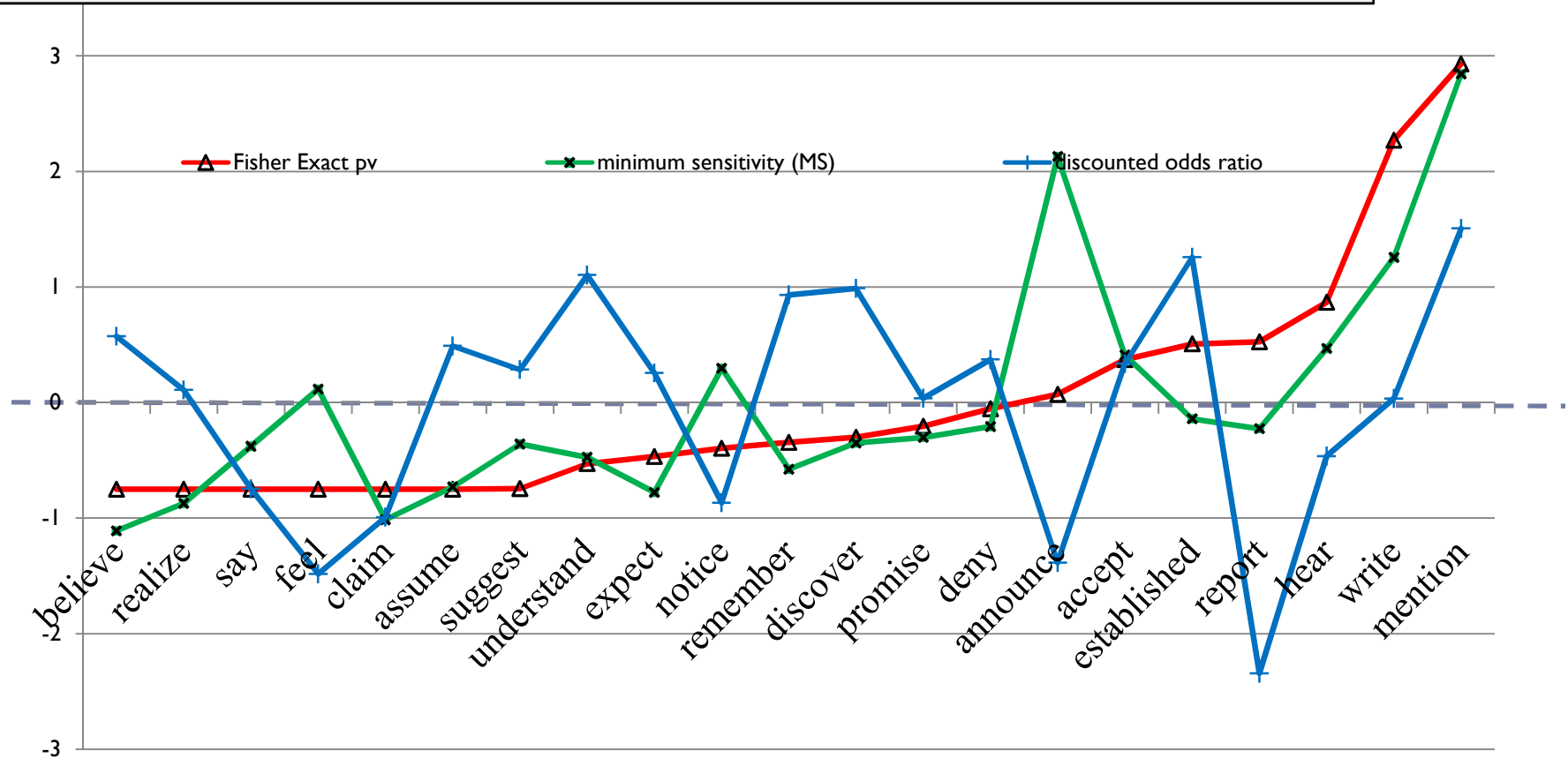




# Fundamentals in QCL Statistical Associations

Bias towards NP complements

Measures: *Fisher exact*, *minimum sensitivity*, *discounted odds ratios*

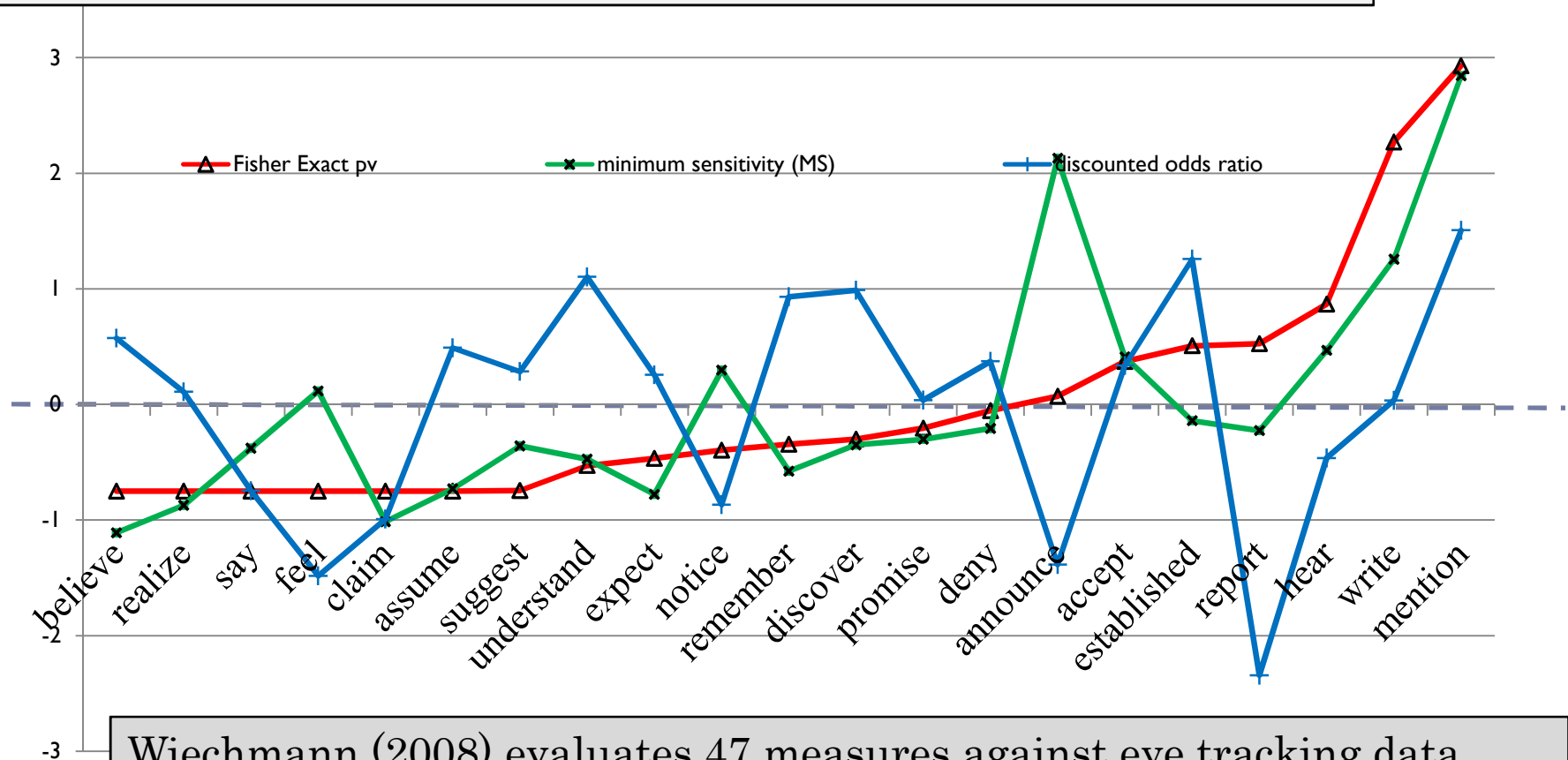




# Fundamentals in QCL Statistical Associations

Bias towards NP complements

Measures: *Fisher exact*, *minimum sensitivity*, *discounted odds ratios*



Wiechmann (2008) evaluates 47 measures against eye tracking data.  
Regression modeling; Coefficients of determination ( $R^2$ ) vary between .07 and .34



# QCL – Domains of application Complex Statistical Associations

---

Corpus-derived predictions about  
human language processing II

## I. Complex associations (ngrams with $n > 2$ )





# QCL – Domains of application Complex Statistical Associations

---

## Corpus-derived predictions about human language processing II

### I. Complex associations (ngrams with $n > 2$ )

Hypothesis:

Processing difficulty of complex construction is a function of degree of entrenchment (cognitive routinization) of construction type



# QCL – Domains of application Complex Statistical Associations

---

Corpus-derived predictions about  
human language processing II

## I. Complex associations (ngrams with $n > 2$ )

**TASK:**  
Assessing complex  
associative relationships



seit 1558

# Fundamentals in QCL

## Data frames

no	add	medium	synR.int	fin.RC.type.bi	synR.int.bi	text.type	formality
1179	<ICE.GB:W2A.002#026:1>	written	do	fin.RC	Nsubj	academic.writing	formal
206	<ICE.GB:S1A.087#073:1:A>	spoken	do	fin.RC	Nsubj	direct.conv	informal
292	<ICE.GB:S1B.019#038:1:C>	spoken	do	Nfin.RC	Nsubj	formal	formal
1930	<ICE.GB:W2E.006#004:1>	written	obl	fin.RC	Nsubj	reportage	informal
1482	<ICE.GB:W2B.018#007:1>	written	obl	fin.RC	Nsubj	non.academic.tech. writing	formal
1151	<ICE.GB:W1A.015#061:3>	written	s	fin.RC	subj	non.prof.writing	informal
1499	<ICE.GB:W2B.020#058:1>	written	a	fin.RC	subj	non.academic.tech. writing	formal
1198	<ICE.GB:W2A.006#066:1>	written	a	fin.RC	subj	academic.writing	formal
239	<ICE.GB:S1B.002#100:1:A>	spoken	a	fin.RC	subj	formal	formal
1640	<ICE.GB:W2B.039#003:1>	written	a	fin.RC	subj	non.academic.tech. writing	formal
1121	<ICE.GB:W1A.009#068:1>	written	a	fin.RC	subj	non.prof.writing	informal

Let's have a look



Daniel Wiechmann – FSU Jena



seit 1558

# Fundamentals in QCL Statistical Analysis

**Case:**

The man<sub>RC</sub> [ that John hates ] is actually quite nice.

*external syntax*

- **NP [Det N RC] VP**
- NPVP [V NP [Det N RC]]
- ...

*internal syntax*

- Subject relative (SRC)
- **Object<sub>DIRECT</sub> relative (ORC)**
- Object<sub>INDIRECT</sub> relatives
- ...



seit 1558

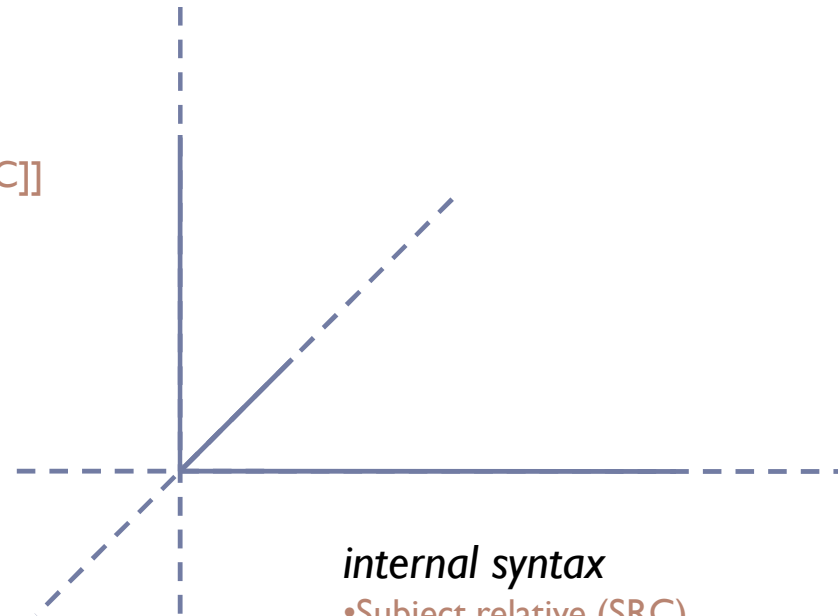
# Fundamentals in QCL Statistical Analysis

Case:

The man<sub>RC</sub> [ that John hates ] is actually quite nice.

*external syntax*

- NP [Det N RC] VP
- NPVP [V NP [Det N RC]]
- ...



*finiteness*

- + finite
- - finite (reduced passive)  
*the horse raced past the barn*
- - to-infinitival

*internal syntax*

- Subject relative (SRC)
- Object<sub>DIRECT</sub> relative (ORC)
- Object<sub>INDIRECT</sub> relatives
- ...

▶ *the right thing to do*

• ...



seit 1558

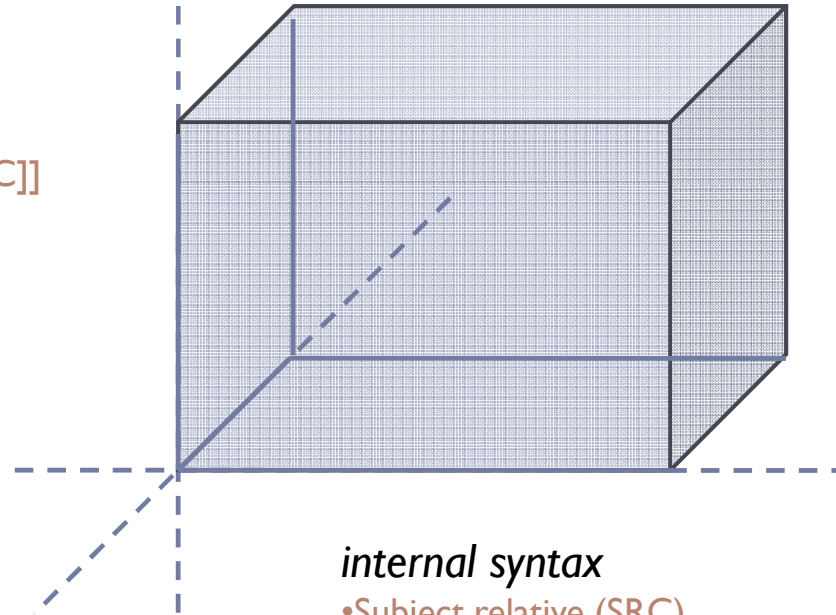
# Fundamentals in QCL Statistical Analysis

**Case:**

The man<sub>RC</sub> [ that John hates ] is actually quite nice.

*external syntax*

- NP [Det N RC] VP
- NPVP [V NP [Det N RC]]
- ...



*finiteness*

- + finite
- - finite (reduced passive)  
*the horse raced past the barn*
- - to-infinitival

*internal syntax*

- Subject relative (SRC)
- Object<sub>DIRECT</sub> relative (ORC)
- Object<sub>INDIRECT</sub> relatives
- ...

▶ *the right thing to do*

• ...

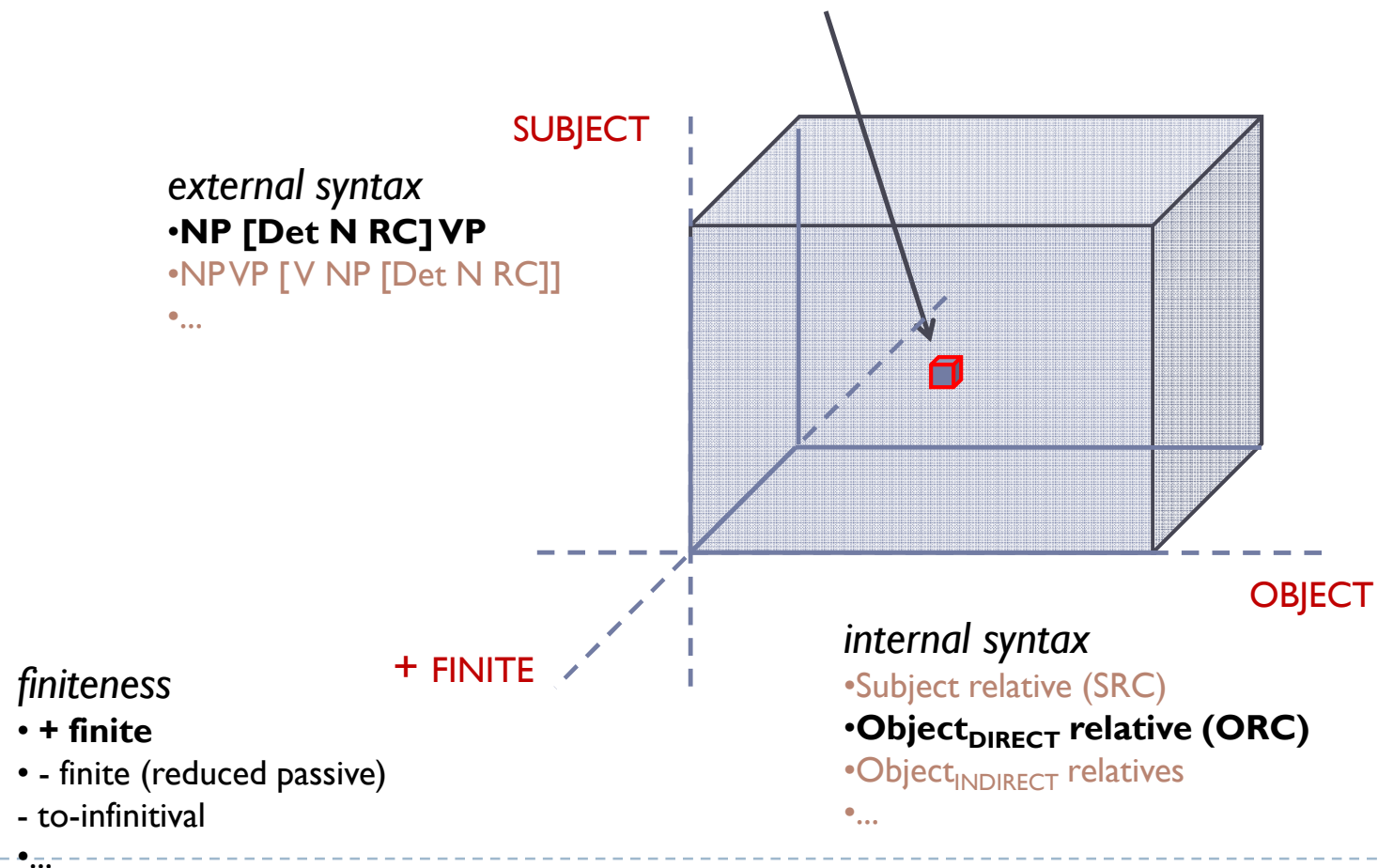


seit 1558

# Fundamentals in QCL Statistical Analysis

Case:

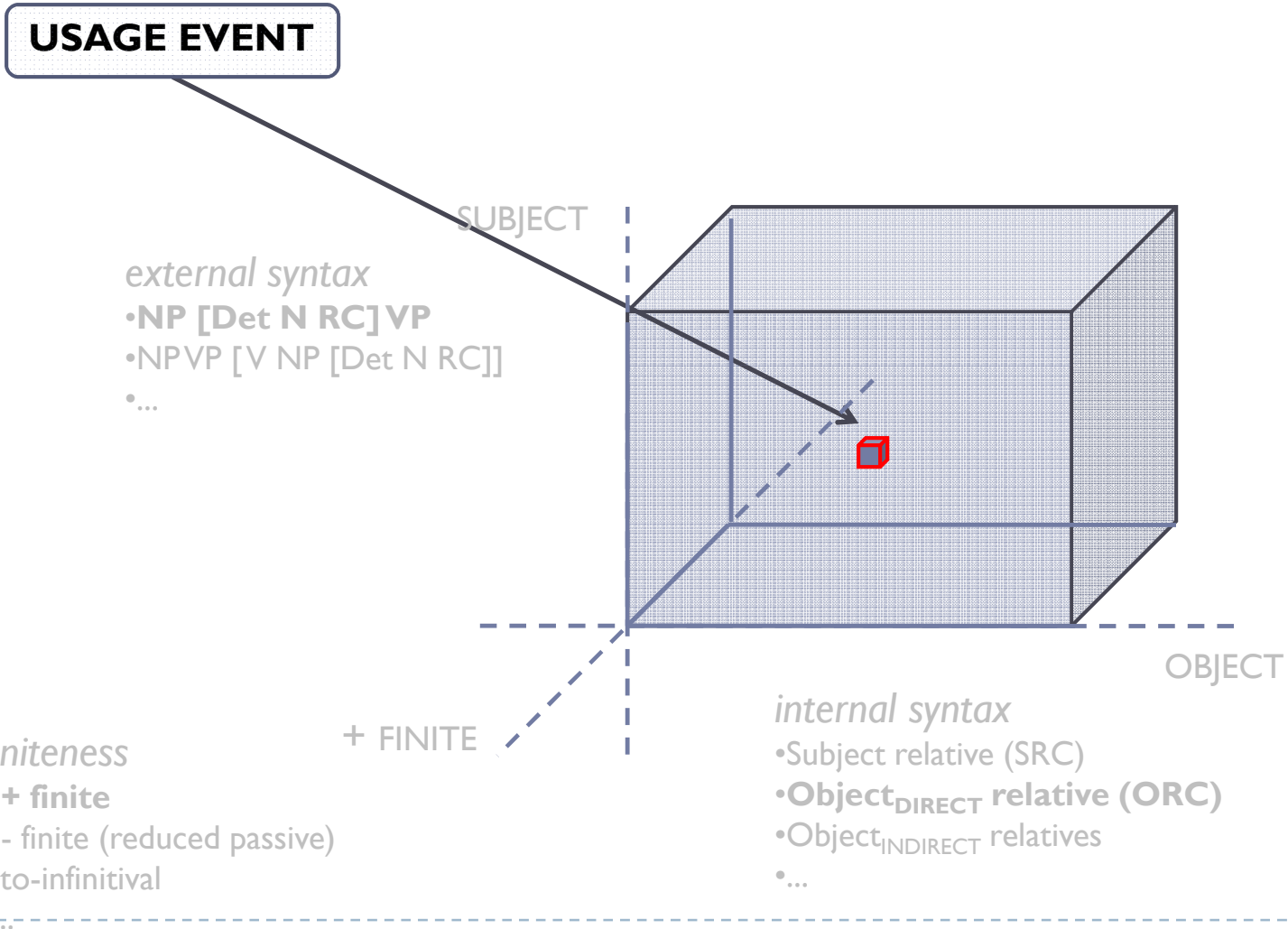
The man<sub>RC</sub> [ that John hates ] is actually quite nice.





seit 1558

# Fundamentals in QCL Statistical Analysis





seit 1558

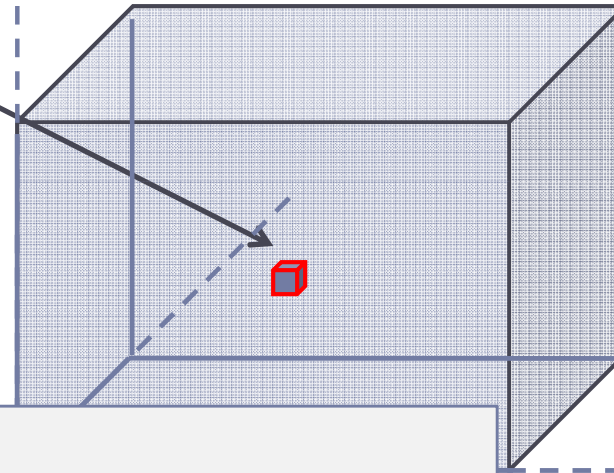
# Fundamentals in QCL Statistical Analysis

## USAGE EVENT

*external syntax*

- NP [Det N RC] VP
- NPVP [V NP [Det N RC]]
- ...

SUBJECT



OBJECT

## Configural frequency analysis

- searches through state space
  - evaluates frequencies of all possible patterns
  - identifies patterns that are statistically special

• finite

- - finite (reduced passive)
- - to-infinitival
- ...

• Object<sub>DIRECT</sub> relative (ORC)

- Object<sub>INDIRECT</sub> relatives
- ...



seit 1558

# Fundamentals in QCL

## Patterns and processing difficulty

---

- *(hierarchical) configural frequency analysis*  
(von Eye and Pena 2004)
    - No functional distinction of variables
    - There are just patterns (configurations of a state space)
- 
- Model type fits the conception of language of **construction grammars**  
(Goldberg 2006)
  - Follows ideas in **exemplar-/memory-based language processing**  
(Daelemans & van den Bosch 2005)





seit 1558

# References

- Baayen, H. 2008. *Analysing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge: CUP.
- Crawley, M. 2005. *The R book*. New York: John Wiley & Sons .
- Goldberg, A. 2006. *Constructions at work: The Nature of Generalisation in Language*. Oxford: OUP.
- Gries, S. 2009. *Quantitative Corpus Linguistics with R*. London: Routledge.
- Stefanowitsch, A. 2005. Quantitative Korpuslinguistik und sprachliche Wirklichkeit. In Christiane Solte-Gresser, Karin Struve, Natascha Ueckmann (eds), *Von der Wirklichkeit zur Wissenschaft. Aktuelle Forschungsmethoden in den Sprach-, Literatur- und Kulturwissenschaften*, 147-161. Hamburg: LIT-Verlag.
- von Eye, A., and Peña, E. 2004. Configural Frequency Analysis: the Search for Extreme Cells. *Journal of Applied Statistics*, 31, 981-997.
- Wiechmann, D. and Fuhs, S. 2006. Concordancing software. *Corpus Linguistics and Linguistic Theory* ,2-1, 109-130
- Wiechmann, D. 2008. On the Computation of Collostruction Strength- Testing measures of associations as expressions of lexical bias. *Corpus Linguistics and Linguistic Theory*, 4-2, 253-290.