



Workshop:

Cognitive Corpus Linguistics: Current Issues in Theory and Methodology

Freiburg Institute for Advanced Studies

Freiburg, October 12, 2008

More on
what corpora can tell us about cognition:
the case of entrenchment

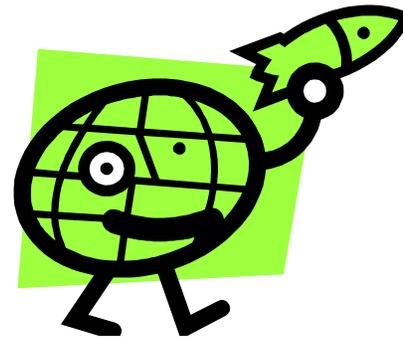


Daniel Wiechmann

Friedrich Schiller University, Jena



*The world of cognitive linguistics
according to me*



Some things that I think we should sit down and talk about...



- 1 Many key **constructs** are **far from being up to date** and are in strong need of serious revision (MEMORY (LOAD), PROCESSING DIFFICULTY)
- 2 Some constructs in (cognitive/functional) linguistics are **ill-defined** (e.g. PREDICTABILITY in Goldbergian CxG)
- 3 Sometimes we have many (**ontologically inconsistent**) definitions of a construct (e.g. PROTOTYPICALITY).
- 4 **Directionality** of definitions: we have concept C and then think about how we can express C (e.g. we assume COMPLEXITY, then count words or nodes and then say something about the role of COMPLEXITY for phenomenon X)
- 5 **Methodological issues are misunderstood** (“all X does is number crunching & data mongering“)
 - ▶ theory building on the basis of nil-hypothesis testing
$$P(\text{data} | \text{hypothesis}) \neq P(\text{hypothesis} | \text{data})$$



A stronger commitment to
operationalism
might help us out here





What exactly is our position on the relation between **measurement – (scientific) meaning ?**

* _____

Bridgman, P. W. “The Operational Character of Scientific Concepts,” in Boyd, Gasper, and Trout, *The Philosophy of Science*, pp. 57–69.

Hempel, C. “A Logical Appraisal of Operationism,” in Brody and Grandy, *Readings in the Philosophy of Science*, pp. 12–20.

Hempel, C. “Empiricist Criteria of Cognitive Significance: Problems and Changes,” in Boyd, Gasper, and Trout, *The Philosophy of Science*, pp. 71–84.





Percy W. Bridgman

“Never again are concepts to prevent us from seeing what nature tries to show us. The way to prevent this is to be sure that something in nature answers to each of our concepts. And the way to do that is to define each scientific concept solely in terms of the operations required to detect or measure instances of the concept. Thus, length is to be identified, not with some property, such as taking up space, but with the procedures for using a meter stick. This is all that length means.”

(Bridgman 1999)

 operationalism





Percy W. Bridgman

“Never again are concepts to prevent us from seeing what nature tries to show us. The way to prevent this is to be sure that something in nature answers to each of our concepts. **And the way to do that is to define each scientific concept solely in terms of the operations required to detect or measure instances of the concept.** Thus, length is to be identified, not with some property, such as taking up space, but with the procedures for using a meter stick. This is all that length means.”

(Bridgman 1999)

➔ operationalism





Maybe we should **not entertain definitions/heterogeneous concepts** such as **PROTOTYPICAL(x)**, if they are conceived as something that is ...



Percy W. Bridgman

- cognitively salient
- most frequent
- central
- oldest
- first acquired
- easily learned for L2
- ...

Each operational procedure should be identified with a distinct concept





If meaning is identified with measurement operations, then we don't know what we mean when we don't know how to measure it...





And now something completely different...



WHAT CAN CORPUS DATA TELL US ABOUT COGNITION ?



Desiderata

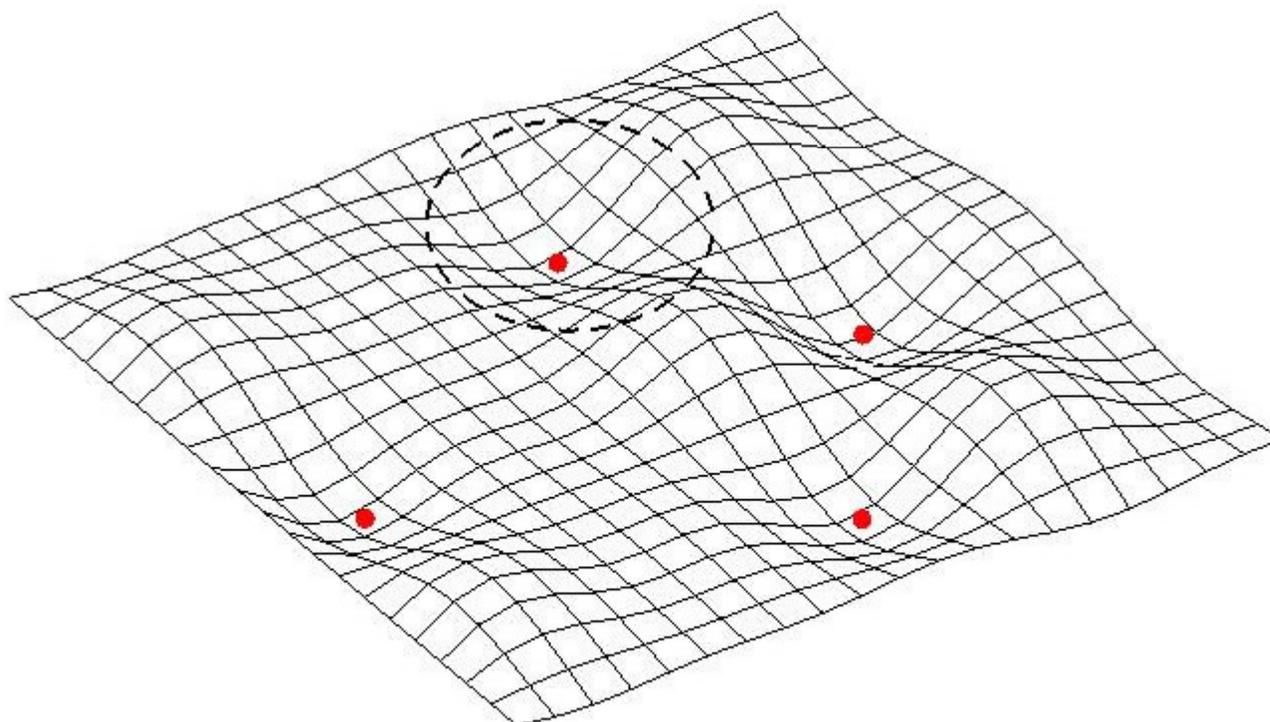
- ecological validity (naturalness)
- representativity; disclose frequencies/distributional properties
- available in amounts that allow for multivariate designs
- allow for grammar induction (analogical modeling)
- easy to come by

So, before I try & make a case for corpora we might ask ourselves:
If we want these things, what else should we turn to?





Entrenchment, language processing & corpora



Entrenchment, language processing & corpora



I. Processing difficulty of an expression E \equiv categorisation difficulty of E

(► Daelemans 1999 *Memory-Based Language Processing*

Tabor and Tanenhaus 1997 *Visitation Set Gravitation Model*)

Ib Processing difficulty is recognition time match cue set against stored (invariant) representation (i.e. spike patterns; memory traces)

II. Categorization difficulty of E is a function of degree of entrenchment of E

(► point attractor strength (dynamical systems theory)

III. Degree of entrenchment of E is a function of frequencies of E in ambient language

IV. Frequencies of E in ambient language can be approximated in corpora

V. Assuming the truth of I-IV, we can assess processing difficulty and entrenchment values of E from corpora



Some caveats...



- Perception (cognition) is all about spatio-**temporal** patterns
- Corpora do not contain information about the time course of state-changes
- Patterns I have talked about are not in any obvious way temporal (in fact, they have been characterized as rather atemporal and stable).
- Suggestion here: we must help ourselves to a simplifying assumption:
 - *A static configuration (type) detected by CFA or similar pattern oriented techniques represent (and label) a spatio-temporal sequence of electro-chemical activity (spikes, activation potentials)*





***AUTOMATIZATION** is the process observed in learning to tie a shoe or recite the alphabet: through repetition or rehearsal, a complex structure is thoroughly mastered to the point that using it is virtually automatic and requires little conscious monitoring. In CG parlance, a structure undergoes progressive **ENTRENCHMENT** and eventually becomes established as a unit“*

(Langacker 2008:16)





***AUTOMATIZATION** is the process observed in learning to tie a shoe or recite the alphabet: through repetition or rehearsal, a complex structure is thoroughly mastered to the point that using it is virtually automatic and requires little conscious monitoring. In CG parlance, a structure undergoes progressive **ENTRENCHMENT** and eventually becomes established as a unit“*

(Langacker 2008:16)





Preferred data:

Are there

Corpus-data that are best suited for research question
(here: spoken, natural/spontaneous, parsed)

Approach:

Detection of higher order n-grams ($n > 2$)

Possible solutions:

CONFIGURAL FREQUENCY ANALYSIS

(focus on patterns; does away with independents -> dependent VAR)

Alternatives (that I am aware of)

Association rule mining

CONTRAST SET LEARNING

WEIGHTED CLASS LEARNING

K-OPTIMAL PATTERN DISCOVERY



http://michael.hahsler.net/research/bib/association_rules/

Entrenchment, language processing & corpora



Property A
(e.g. *CONSTRUCTION(x)*)

Property B
(e.g. *CONSTRUCTION(Y)*)

	Value 1	Value 0	
Value 1	State 1	State 2	
Value 0	State 3	State 4	



Entrenchment, language processing & corpora



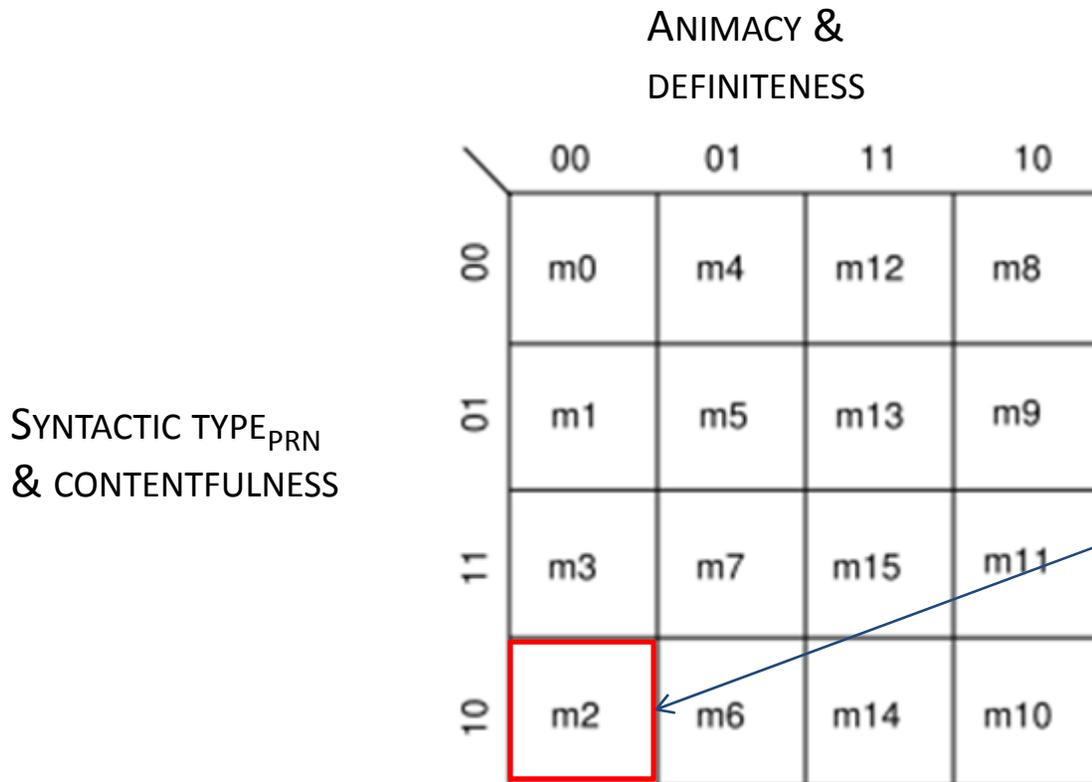
Property B
(e.g. DEFINITE(x))

Property A
(e.g. ANIMATE(x))

	Value 1	Value 0	
Value 1	State 1	State 2	
Value 0	State 3	State 4	



Entrenchment, language processing & corpora



Each cell represents
a state (=configuration)

m2 = 10 00
+ pronominal
- contentful
- definite
- animate

e.g. *something*



Representing *state space of four binary variables*
via **Karnaugh maps**

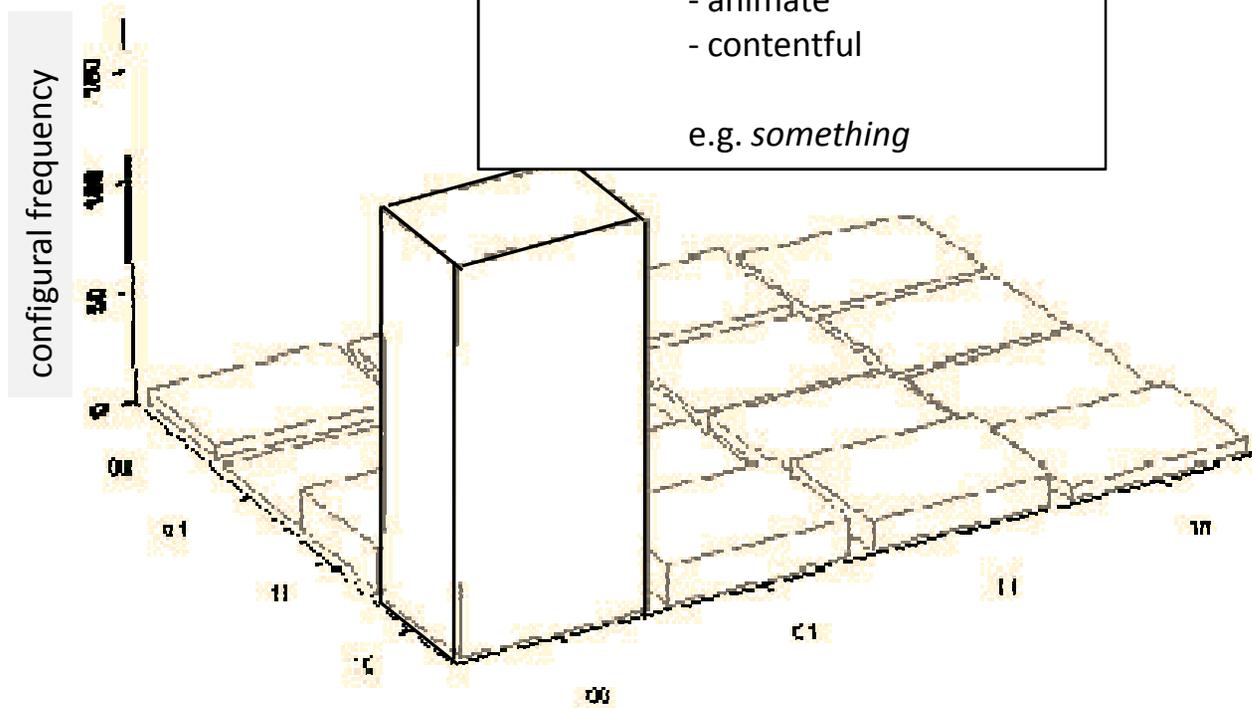
Entrenchment, language processing & corpora



Most frequent type:

m2 = 01 00
+ pronominal
- definite
- animate
- contentful

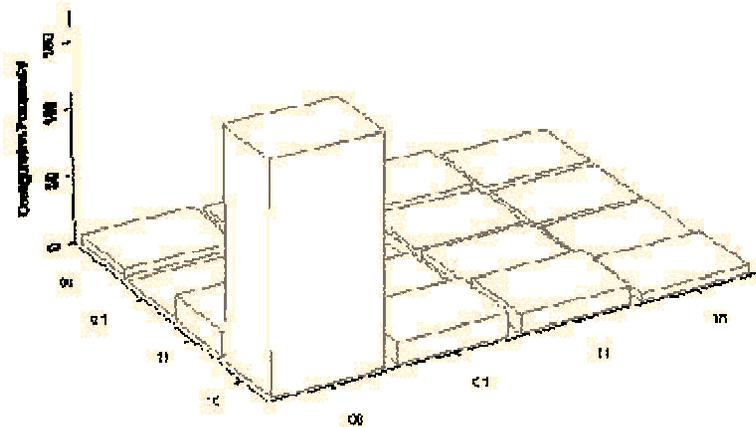
e.g. *something*



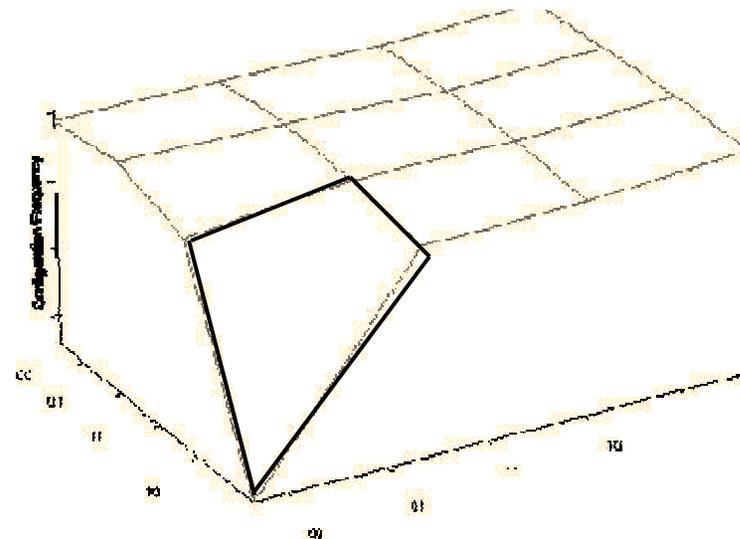
Entrenchment, language processing & corpora



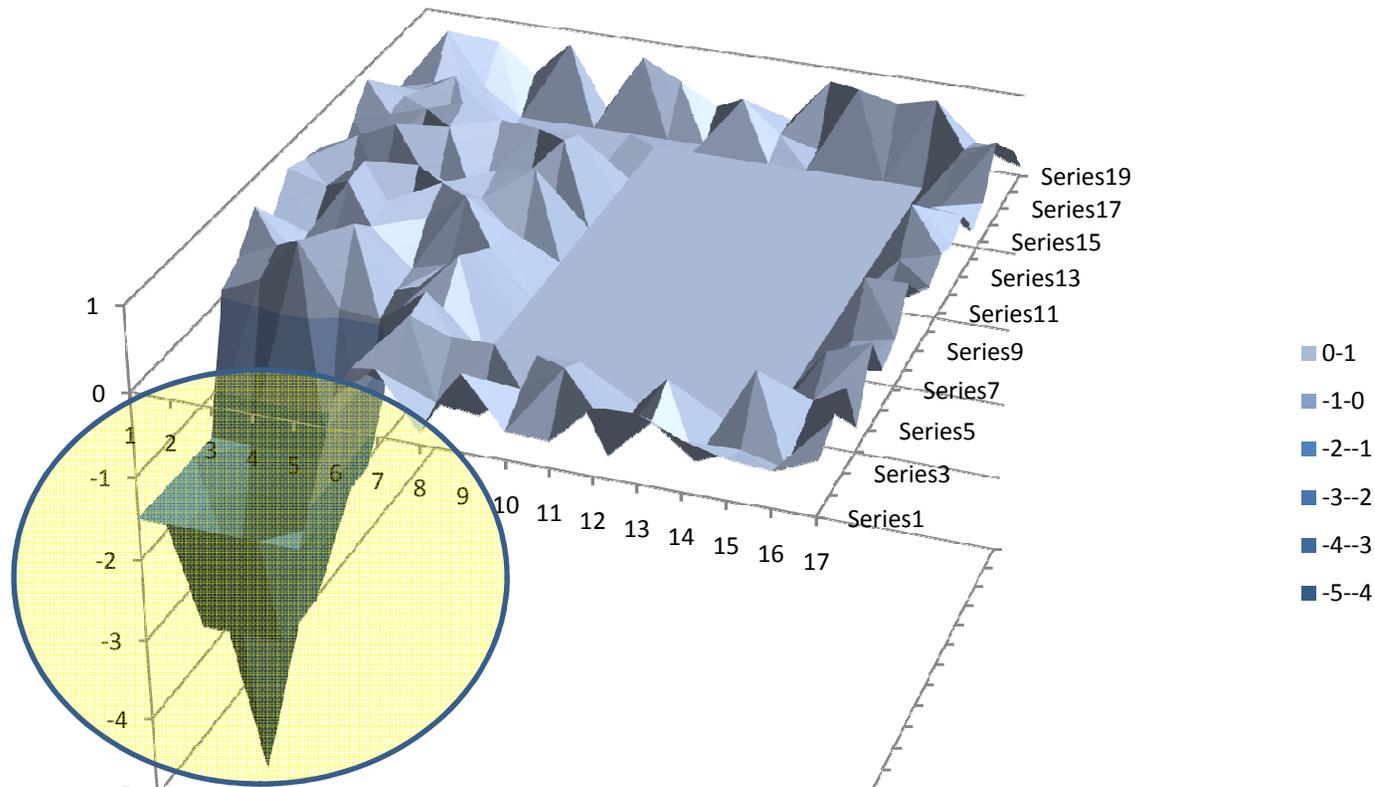
Most frequent type (m4)
is **highest block**



Most frequent type (m4)
is **deepest valley**



Entrenchment, language processing & corpora



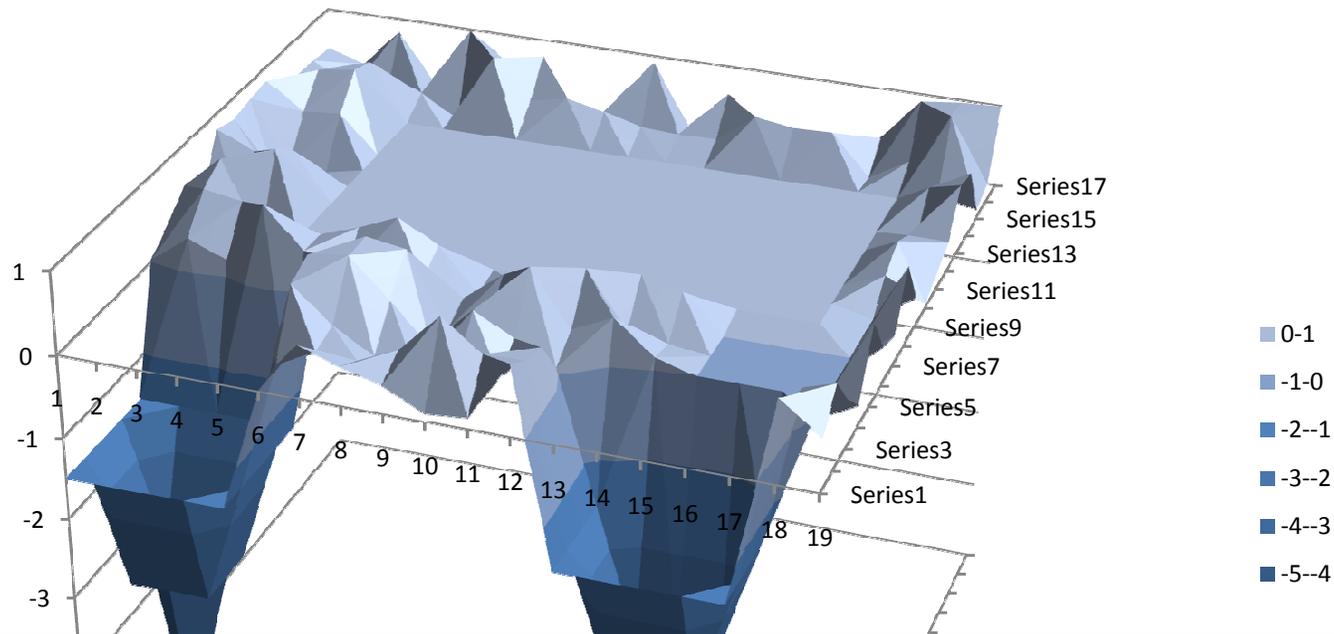
► Deep cone corresponds to deeply **entrenched pattern** which can be expressed as a **type** in **CFA** or a **significant rule** in **k-optimal pattern discovery** technique, ... (?)



State spaces, configurations & entrenched pattern



Processing difficulties are expected if there are *attractors* that are *close to each other*. (competition ~ classification difficulty) (???)

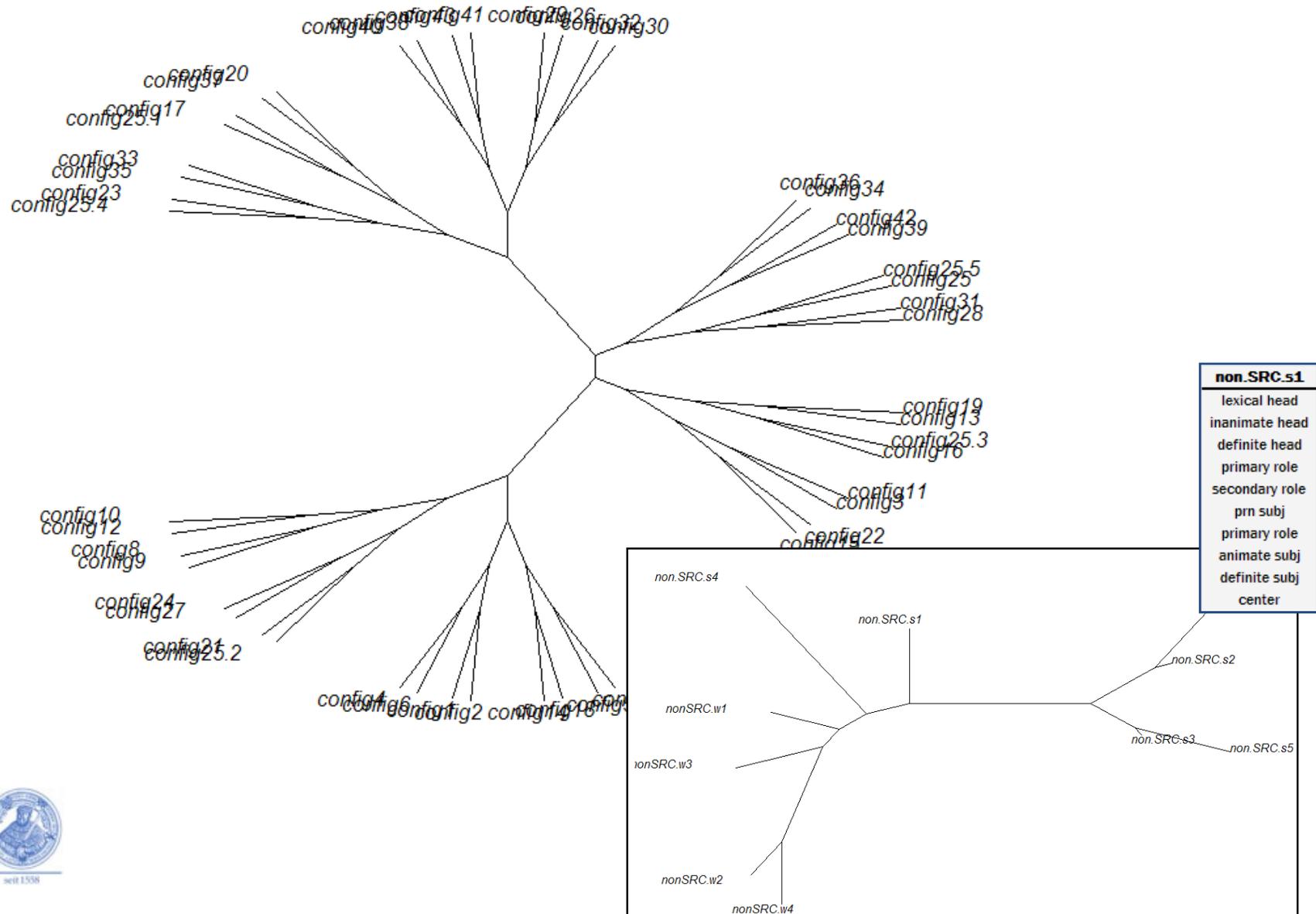


► Visitation Set Gravitation model

Tabor, W., and Tanenhaus, M. 1997. Parsing in a Dynamical System: An Attractor-Based Account of the Interaction of Lexical and Structural Constraints in Sentence Processing. *Language and Cognitive Processes*, 12(2): 211-271.



Processing difficulty as distance to entrenched pattern



k -optimal pattern discovery

- Uses highly efficient OPUS search algorithm (Webb 1995, 2000, 2006) to identify association rules that highlight interrelationships among attributes (factor levels)

- **IF** *NAME=Daniel* **THEN** *CONSOLE=yes*

Association rules

- Output coefficients:

- Coverage

- Support

- Strength

- Lift

- **Leverage** (=proportion of additional cases covered by both LHS and RHS above those if LHS and RHS were independent)

Detected rules

Right-hand-side (RHS) restricted to: type of embedding

Rule mining SPECS

Search for rules
Search by leverage
Filter out rules that are insignificant, critical value=0.05

Maximum number of attributes on LHS = 4
Maximum number of rules = 100
Minimum leverage = -1.0
Minimum leverage count = -2147483647
Minimum coverage = 0.0
Minimum coverage count = 1
Minimum support = 0.0
Minimum support count = 0
Minimum lift = 0
Minimum strength = 0.0

All values allowed on LHS

Values allowed on RHS:

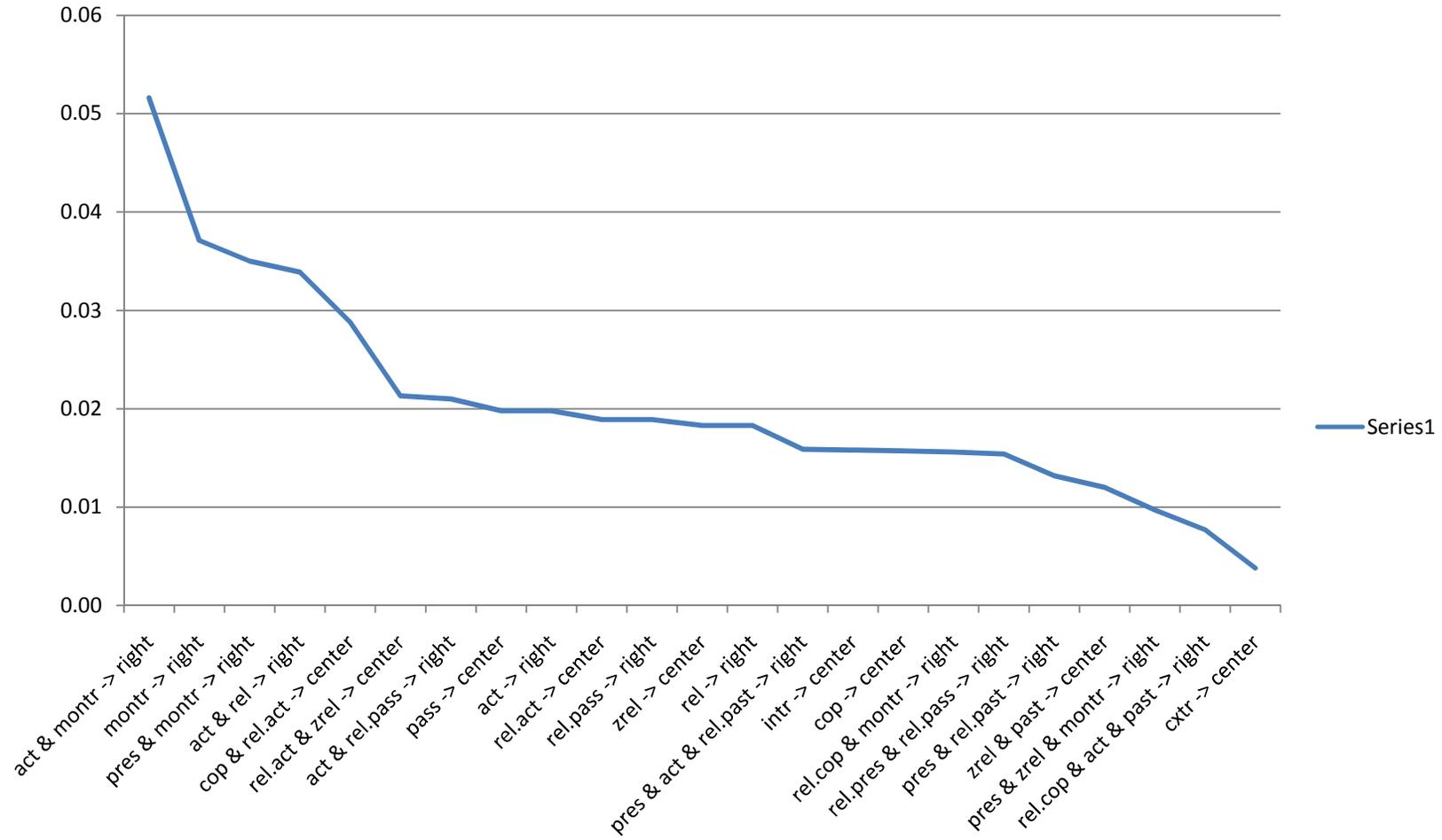
center
right

Only 23 rules satisfy the specified constraints.

act & montr -> right
[Coverage=0.335 (301); Support=0.258 (232); Strength=0.771;
Lift=1.25; Leverage=0.0516 (46.3); p=1.02E-009]

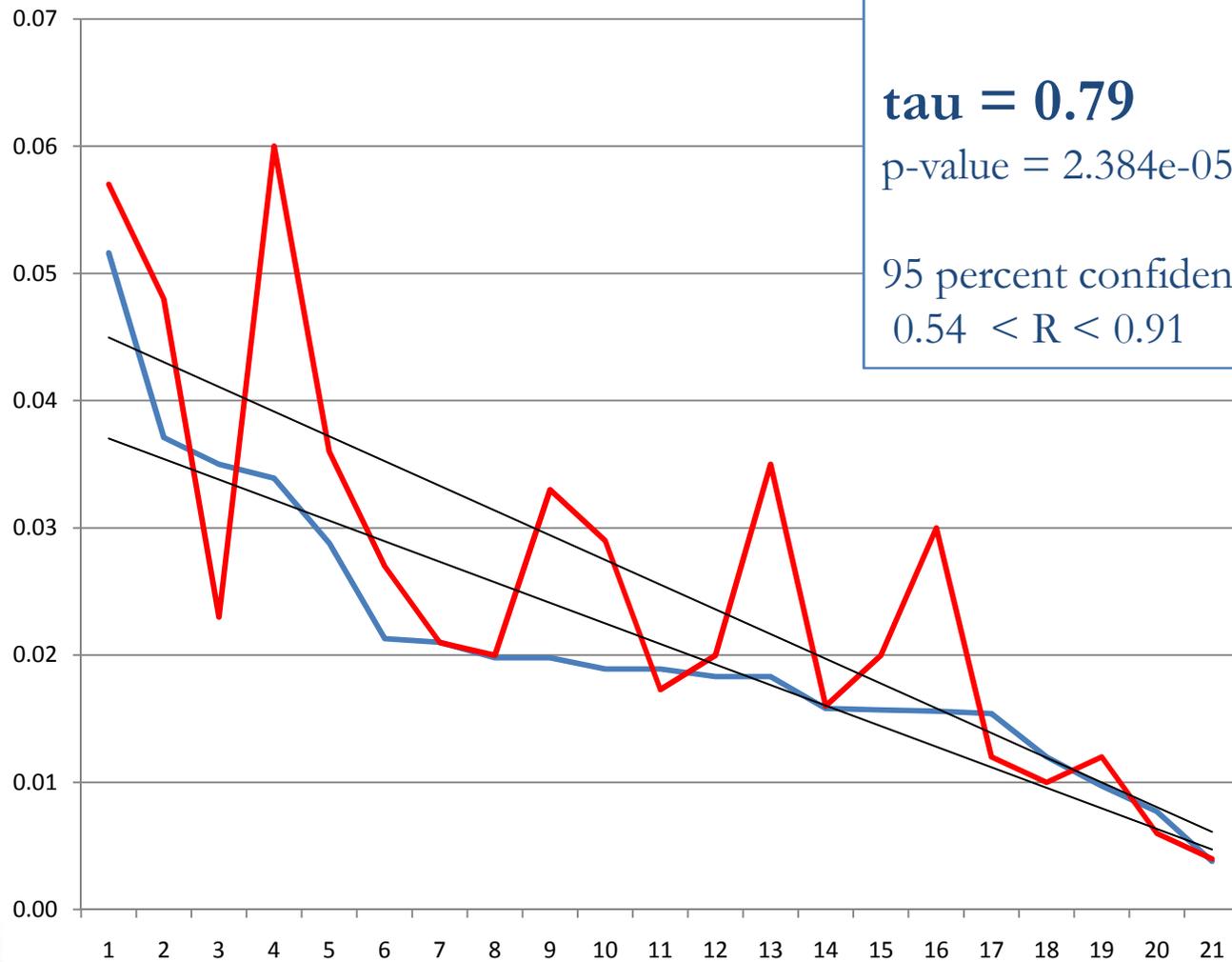
<i>detected rule</i>	<i>leverage</i>
act & montr -> right	0.05
montr -> right	0.04
pres & montr -> right	0.04
act & rel -> right	0.03
cop & rel.act -> center	0.03
rel.act & zrel -> center	0.02
act & rel.pass -> right	0.02
pass -> center	0.02
act -> right	0.02
rel.act -> center	0.02
rel.pass -> right	0.02
zrel -> center	0.02
rel -> right	0.02
pres & act & rel.past -> right	0.02
intr -> center	0.02
cop -> center	0.02
rel.cop & montr -> right	0.02
rel.pres & rel.pass -> right	0.02
pres & rel.past -> right	0.01
zrel & past -> center	0.01
pres & zrel & montr -> right	0.01
rel.cop & act & past -> right	0.01
cxtr -> center	0.00

Results: k-pattern rule discovery rule power (by leverage)





CFA versus k-optimal pattern discovery



Rank order correlation

tau = 0.79

p-value = 2.384e-05

95 percent confidence interval:

0.54 < R < 0.91

- leverage
- Q
- Linear (leverage)
- Linear (Q)



scit 1558