Friedrich-Alexander-Universität
Erlangen-Nürnberg

FAU

Measuring productivity and fixedness in lexico-syntactic constructions

Prof. Dr. Stephanie Evert

Chair of Computational Corpus Linguistics | www.linguistik.fau.de

FAU MoD Lecture | 13 Nov 2024

The fixedness–productivity continuum

fixed
arbitrary
lexicon
specific

productive
compositional
rules
schematic

Corpus Linguistics

Morphology

Construction Grammar (CxG)
DFG GRK 2839

Phraseology

Lexicography

Theoret. Ling.
ca. 1957–1990

Fixedness and productivity in constructions (Cx)

meaningful only for individual slots, not entire Cx

– {*emotional*, *ideological*, *spiritual*, *intellectual, mental*, ...} *baggage*

X *baggage*

– *Telefonitis*, *Fresseritis*, *Subventionitis*, *Festivalitis*, *Inselitis*, ... (non-medical use, Lüdeling & Evert 2005)

X + *-itis*

– {*white*, *orange*, *denim*, *being Irish*, ...} *is the new black* | *comedy is the new rock'n'roll*

X *is the new* Y

– *this success earned him world-wide recognition* (Herbst 2018)

X *earns* Y Z

– here: *the fact is that you'll have to listen to the entire talk* (Schmid 2000)

N + BE + *that*

Shell noun constructions

(Schmid 2000, 2018)

Shell nouns are abstract nouns that can be used as conceptual shells for complex pieces of information.

– four main constructions: N + *that*-clause, N + *to*-infinitive, N + BE + *that*-clause , N + BE + *to*-infinitive
  – *The problem here is that having so easy access [...], it is very croweded at holiday times.* [BNC A15: 876]
  – *But the fact is that the very lack of evidence seems to fan the flames of suspicion.* [BNC CB8: 298]
  – *Our main concern as a group is that we do not waste the money.* [BNC AT1: 2091]

Goal: explore fixedness-productivity continuum for the N slot of the construction

  – could also focus on other elements: determiner, modifiers of N, main V of *that*-clause, subject of *that*-clause, ...

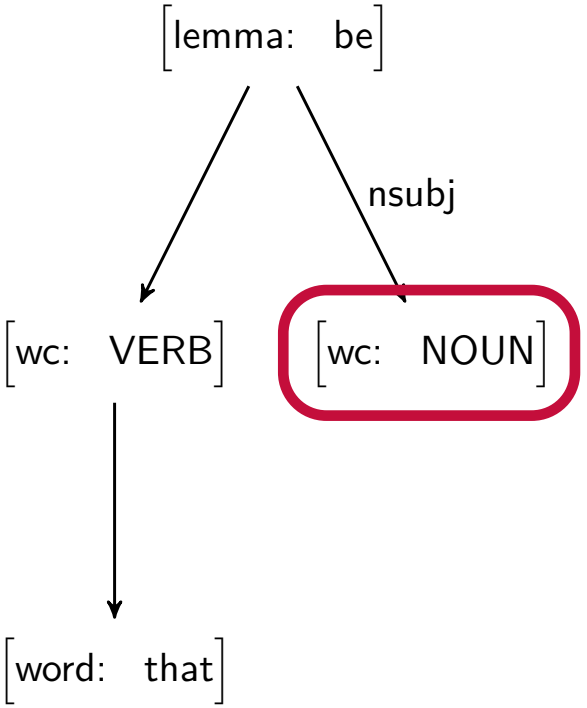
Schmid (2000) identifies six shell noun classes with different functions

– FACTUAL (*fact*, *thing*, *problem*), MENTAL (*idea*, *worry*), LINGUISTIC (*promise*, *story*), MODAL (*possibility*, *truth*), EVENTIVE (*mistake*), CIRCUMSTANTIAL (*place*, *way*)

– secondary research question: Is there evidence that these classes form individual sub-Cx?

Case study: N-BE-THAT Cx in the BNC

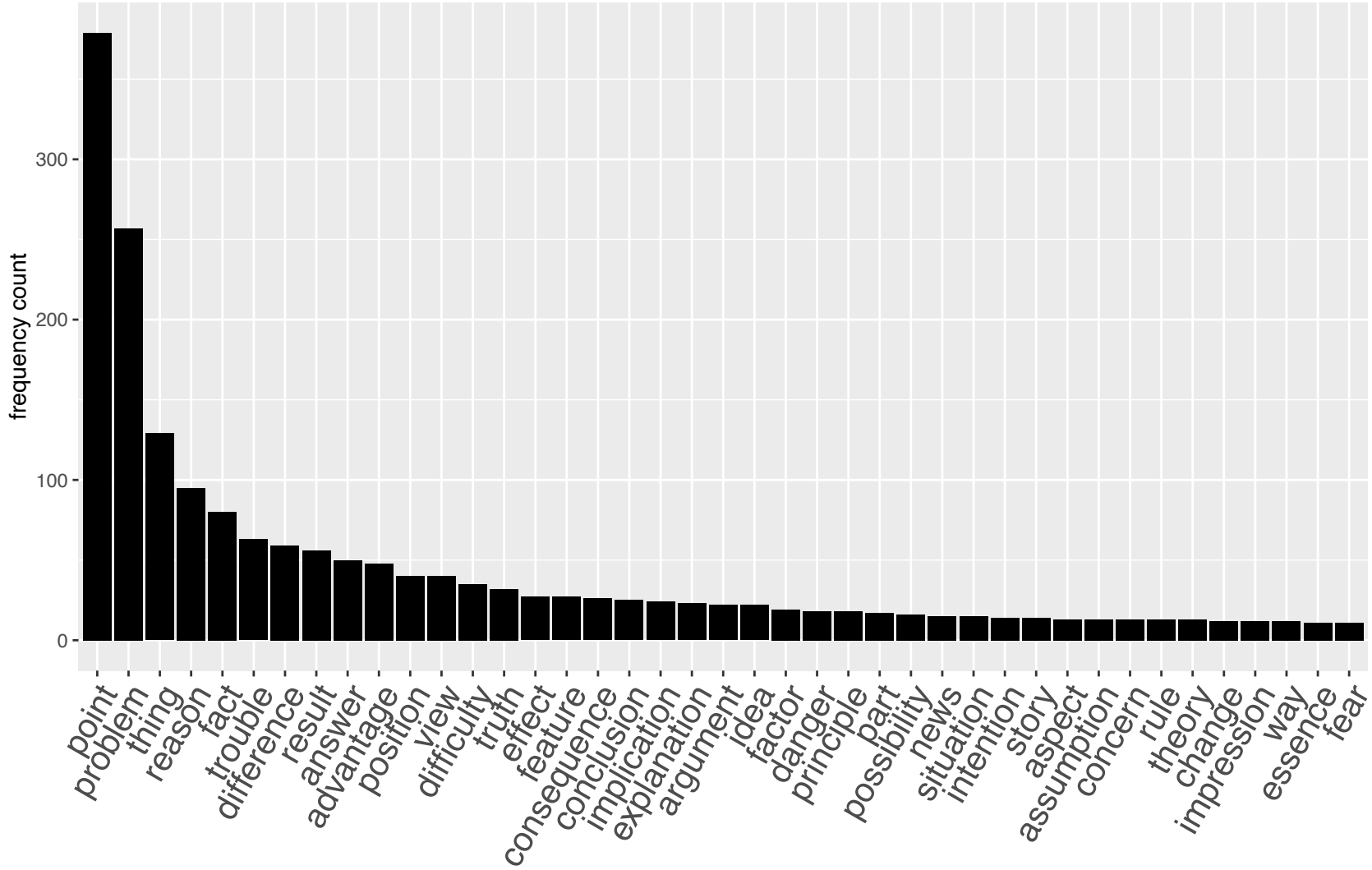
(Diwersy et al. 2019)



- lexico-syntactic pattern extracted from parsed version of British National Corpus
  - https://www.treebank.info/ (using Stanford parser v1.6.9)
  - 32,097 matches in entire BNC
- manual validation of 10% sample → 2,500 instances of Cx
  - eliminate parsing error and sentence duplicates
  - verify actual shell noun Cx rather than just syntactic pattern
  - many thanks to Sascha Diwersy for the hard work!
- data table for quantitative analysis
  - noun lemma, realization of copula, main verb of *that*-clause, pre- and post-modification of noun, full sentence context, ...
  - manual classification of nouns into shell noun classes (NB: a few nouns are ambiguous)
  - here: focus on noun lemma slot + shell noun classes

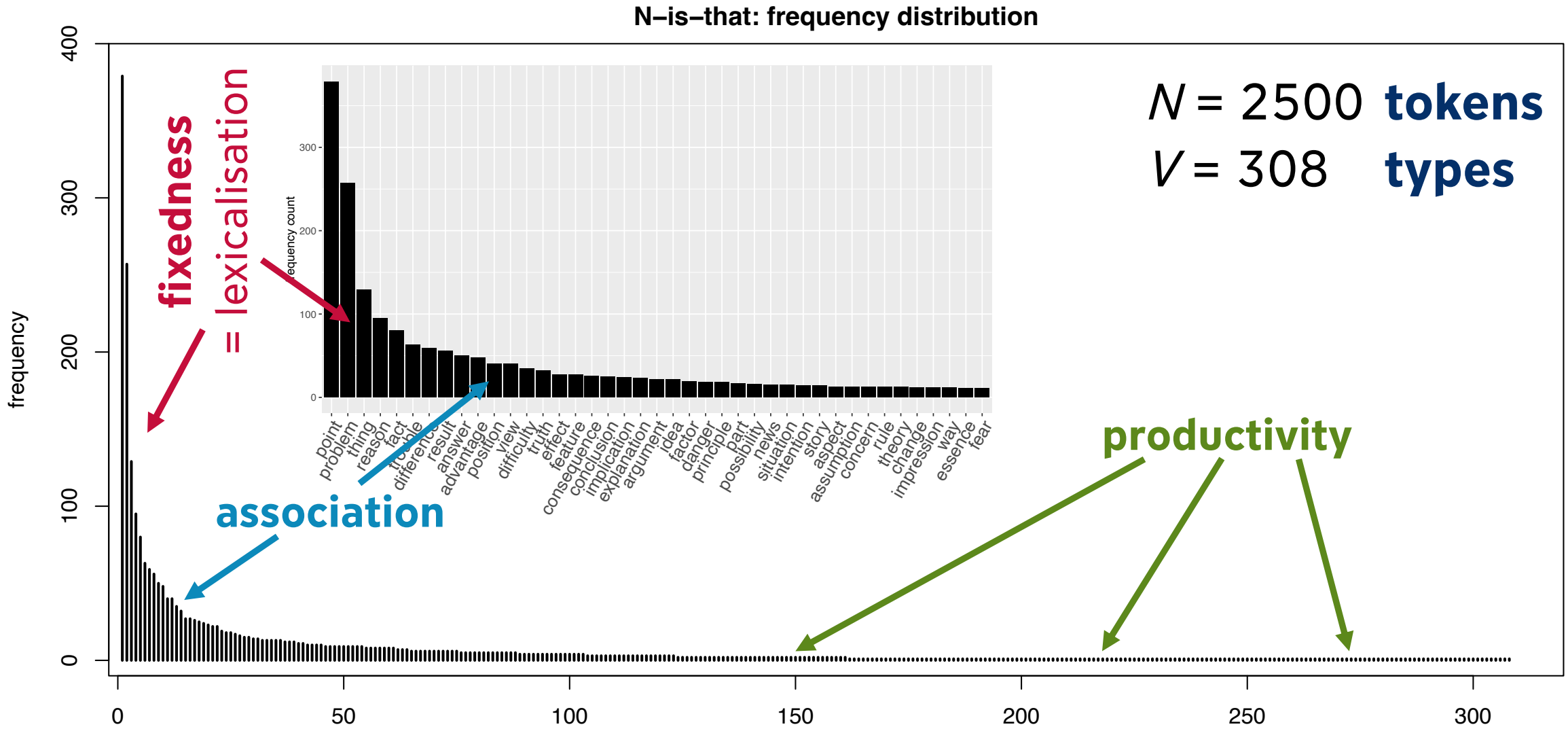
Frequency distribution of shell nouns

in N-BE-THAT construction



Frequency distribution of shell nouns

in N-BE-THAT construction



Ranges in the fixedness-productivity continuum

Distribution of shell nouns in N-BE-THAT Cx suggests three ranges in the fixedness-productivity continuum:

– fixedness: very frequent nouns likely form lexicalised sub-Cx
  – *point* (f = 379 = 15.2%), *problem* (f = 257 = 10.3%), *thing* (f = 129 = 5.2%), *fact* (f = 80 = 3.2%), *trouble* (f = 63 = 2.5%), …
– association: larger group of nouns with higher-than-expected frequency
  – as habitual fillers that sound familiar, such nouns appear to be attracted to the Cx in the sense of CollCxG (Herbst 2020)
– productivity: very many nouns with very low frequency indicate a productive pattern
  – V_1 = 147 hapax legomena (achievement, crux, deal, …), V_2 = 38 dis legomena (basis, flaw, proposal, …), V_3 = 20, V_4 = 15, …

Quantitative description of continuum faces two **key challenges**:

– no clear separation between overlapping ranges of fixedness, association and productivity
– finding appropriate statistical measures for each range

1. Fixedness

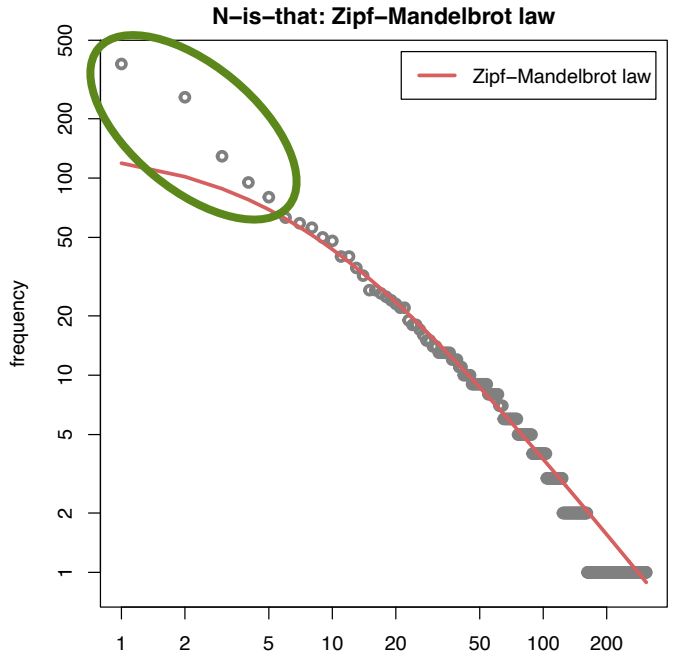
Indicators of fixedness

– relative frequency (proportion of instances)
  – *point*, *problem*, *thing* each account for > 5% of Cx use ➞ likely to be lexically fixed
  – but no mathematical or linguistic justification where to set a cutoff point

– polysemy = ambiguity between shell noun classes = polyvalent function
  – *point* (FACTUAL + LINGUISTIC + MENTAL) and *fact* (FACTUAL + MODAL)

– quantitative evidence: Zipf's law $f_r \approx \frac{C}{(r+b)^a}$
  – (Zipf 1949; Mandelbrot 1962)
  – many linguistic distributions follow Zipf's law

  – but the most frequent nouns don't fit!
  – not part of regular distribution ➞ lexicalised



noun lemma	f	%
point	379	15.2
problem	257	10.3
thing	129	5.2
reason	95	3.8
fact	80	3.2
trouble	63	2.5
difference	59	2.4
result	56	2.2
answer	50	2.0
advantage	48	1.9
position	40	1.6

Indicators of fixedness

– term clustering / dispersion
  – "The chance of two Trumps is closer to $p/2$ than to p^2"
    (cf. Katz 1996; Church 2000; Gries 2008)
  – lexicalised items (e.g. domain terminology) tend to repeat in same text
  – quantitative: *f* (# tokens) vs. *df* (# texts)
  – suggested criterion: f_{rep} = number of tokens in texts with repetition
    (compared to expectation E[f_{rep}] for random distribution)
  – indicates term clustering for *trouble*, but not for *reason*

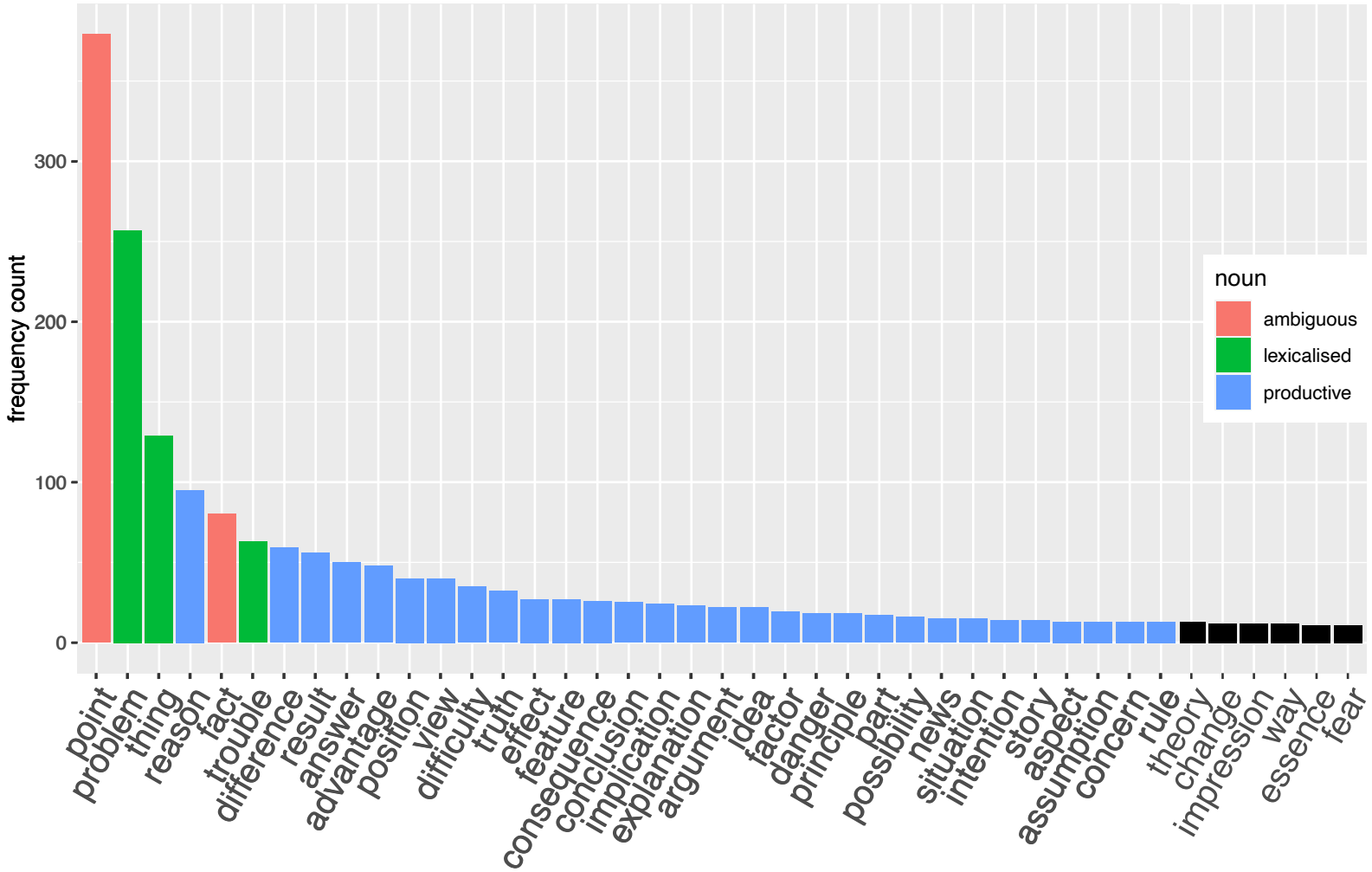
– final list of fixed nouns: *point*, *problem*, *thing*, *fact*, *trouble*

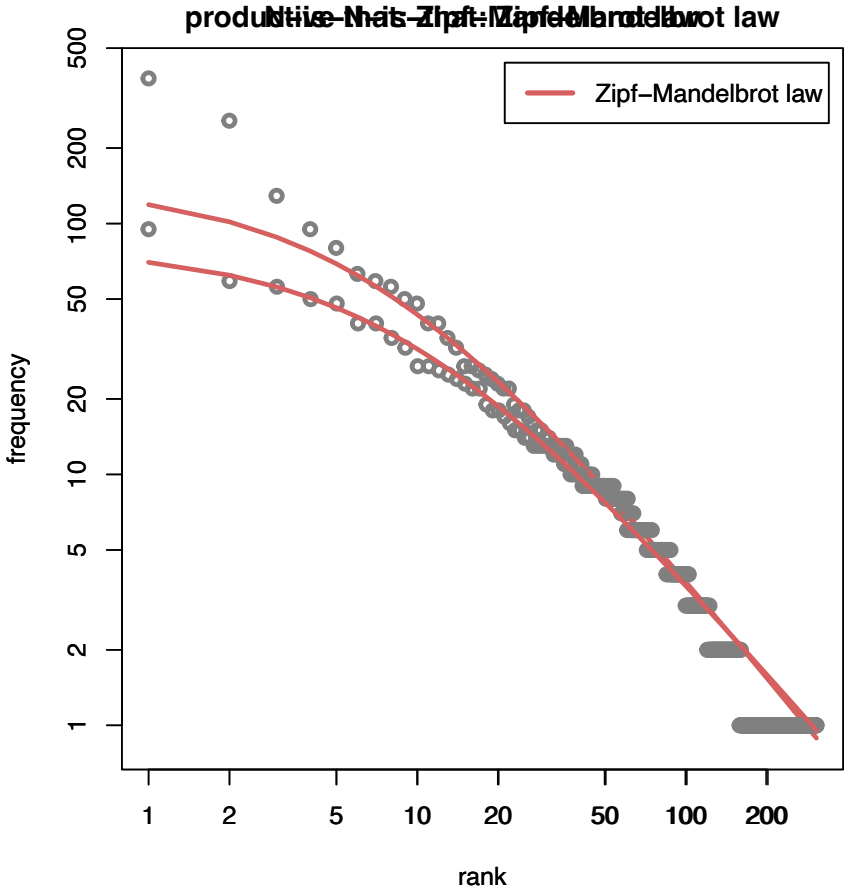
– analyse as lexicalised sub-Cx ➞ e.g. adjectival modifiers of noun
  – excluded from data set for further analysis of "productive" N-BE-THAT Cx

noun lemma	f	df	f_{rep}	E[f_{rep}]
point	379	286	157	100.3
problem	257	216	75	50.8
thing	129	118	21	14.5
reason	95	90	9	8.2
fact	80	68	17	5.9
trouble	63	53	18	3.7
difference	59	56	6	3.3
result	56	53	6	3.0
answer	50	48	4	2.4
advantage	48	43	8	2.2
position	40	36	8	1.5

Consequence: exclude fixed shell nouns

from productive N-BE-THAT construction





POINT-BE-THAT and FACT-BE-THAT Cx

Further analysis of (potential) lexicalised sub-Cx focuses on other elements and on semantics / connotation

e.g. determiner of shell noun

– based on manual annotation of random sample (400 items)

– other: *one*, *another*, possessive (*my*, *his*, *the government's*), ∅

X-BE-THAT Cx	definite	indefinite	other
point	75%	10%	15%
fact	94%	3%	4%
other N	65%	8%	27%

e.g. adjectival modifiers of shell noun

– measure association between modifier and construction (→ part 3)

– POINT-BE-THAT Cx: *important*, *essential*, *whole*, *crucial*, *final*, *key*

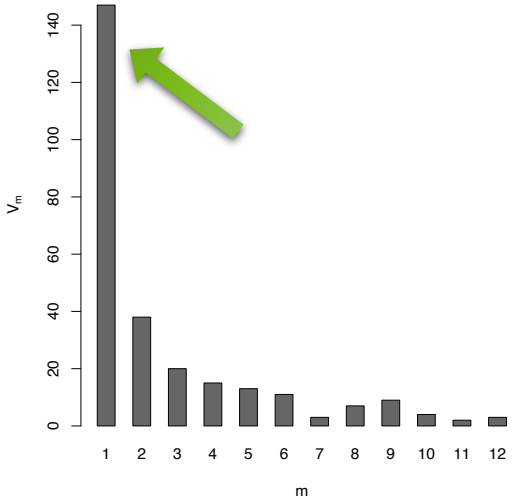
– FACT-BE-THAT Cx: *simple*, *plain*, *sad*, —

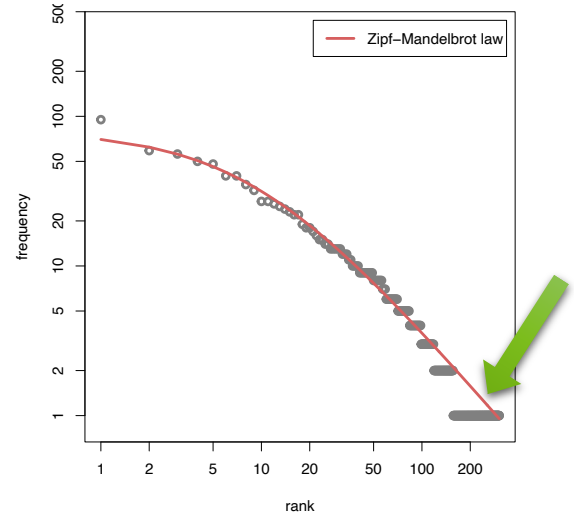
2. Productivity

Indicators of productivity

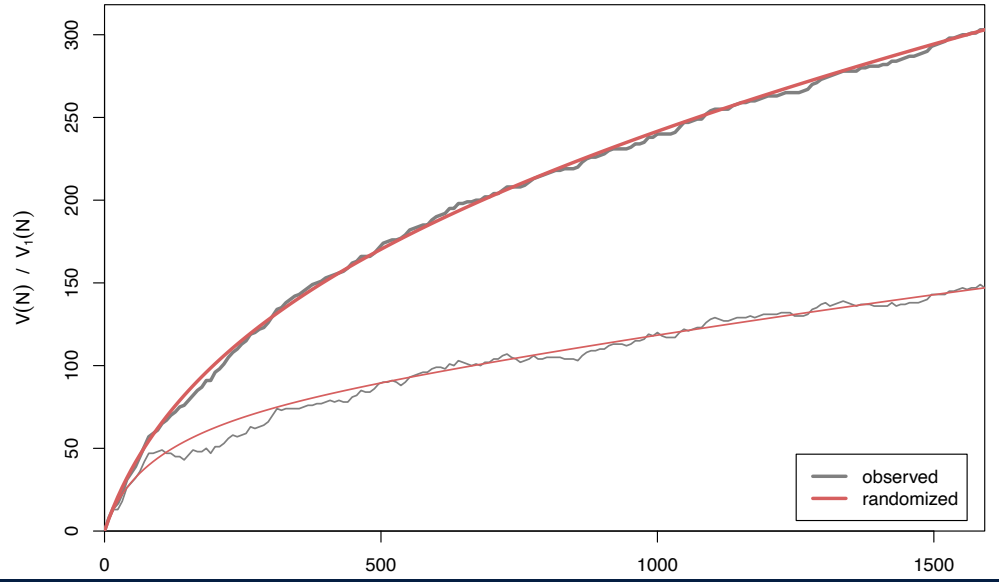
(e.g. Tweedie & Baayen 1998; Baayen 2001; Zeldes 2012)

- frequency spectrum: large number of rare types
  - V_1 = number of hapax legomena with $f = 1$
  - V_2 = number of dis legomena with $f = 2$
  - V_m = number of types with $f = m$





- vocabulary growth: spontaneous creation of new types
  - vocabulary growth curve (VGC) plots V (# types) against N (# tokens)
  - often randomised for smoothing



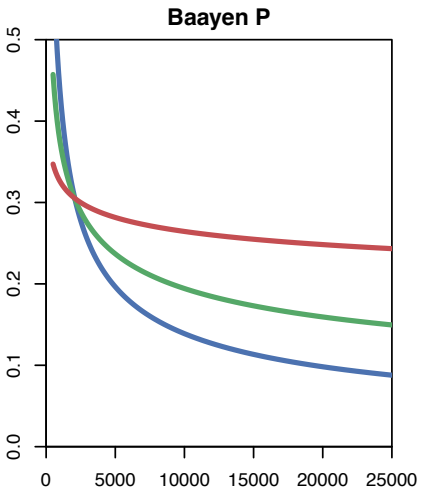
- mathematics: slope of VGC = productivity index
  - (Baayen 1992)

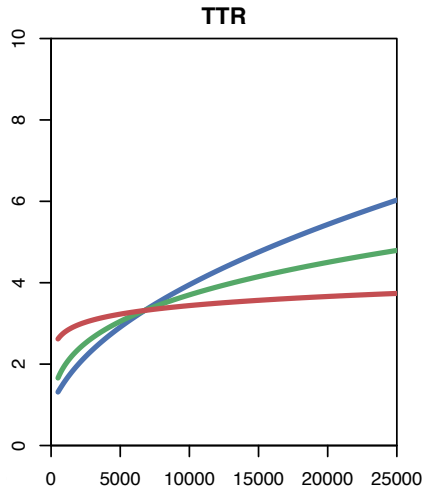
$$\mathscr{P} = \frac{V_1}{N}$$

Productivity measures

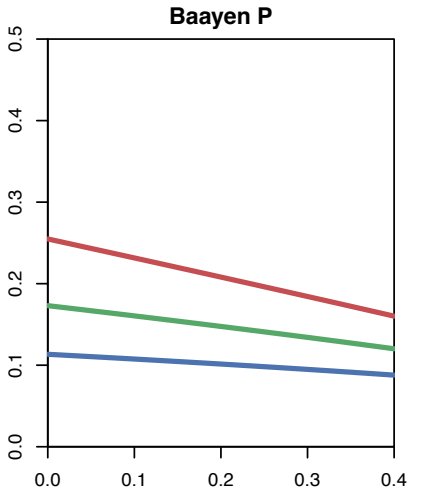
https://zipfr.r-forge.r-project.org/lrec2018.html

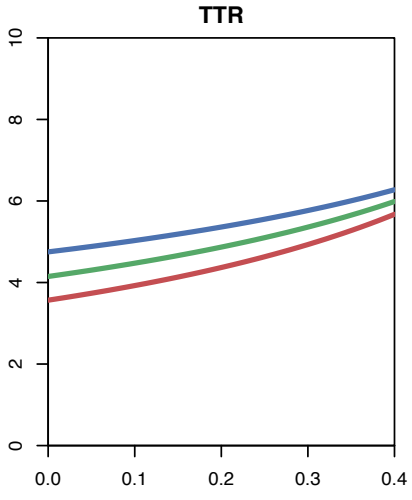
- Baayen's productivity index $\mathscr{P}$
- type-token ratio TTR = V / N
- Herdan's law C = log V / log N
- Sichel S = V_2 / V
  - and many, many more …





- many measures affected by sample size N
  - "normalised" measures used for comparison (e.g. MATTR, MTLD)
  - average over fixed-length bins ➞ dependency on bin size





- most affected by probability mass of frequent types
  - but we want to focus on productive range of continuum!

Statistical LNRE models

(Baayen 2001; Evert 2004; Evert & Baroni 2007)

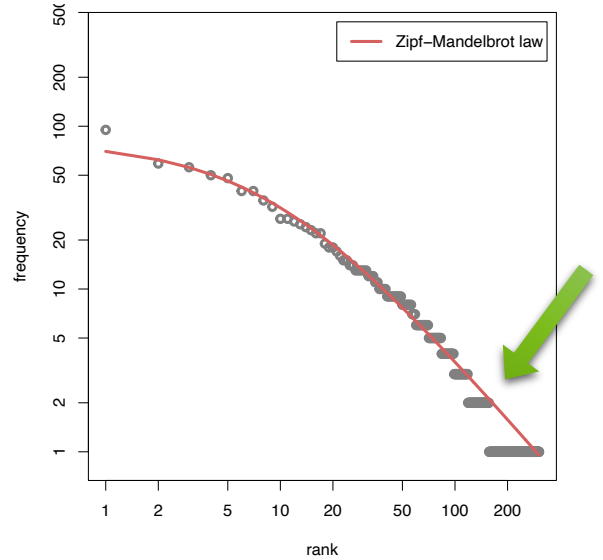
LNRE (*large number of rare events*) are statistical models for sampling from a distribution characterised by Zipf's law (or a similar type-token distribution)

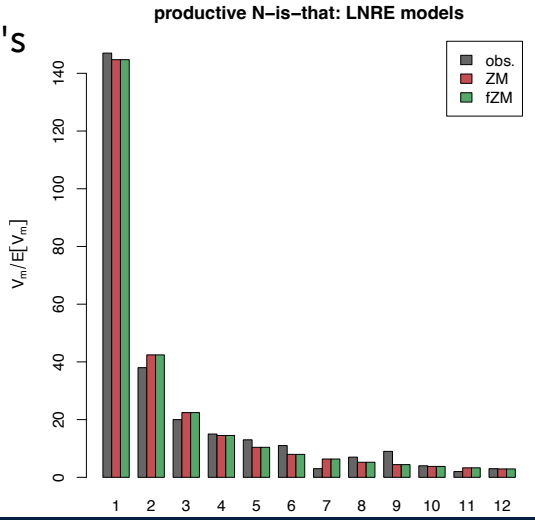
- intuition: fit Zipf's law to observed Zipf ranking
  - but doesn't account very well for hapax legomena and dis legomena

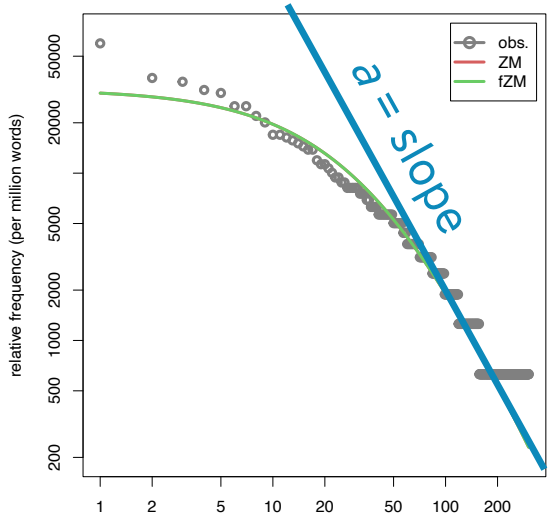
- LNRE version: ZM and fZM models (Evert 2004)
  - fitted to frequency spectrum ➞ focus on lowest frequency classes V_m
  - better hapax & dis legomena, less accurate for medium to high freq's

- quantitative measure: Zipf slope $a > 1$

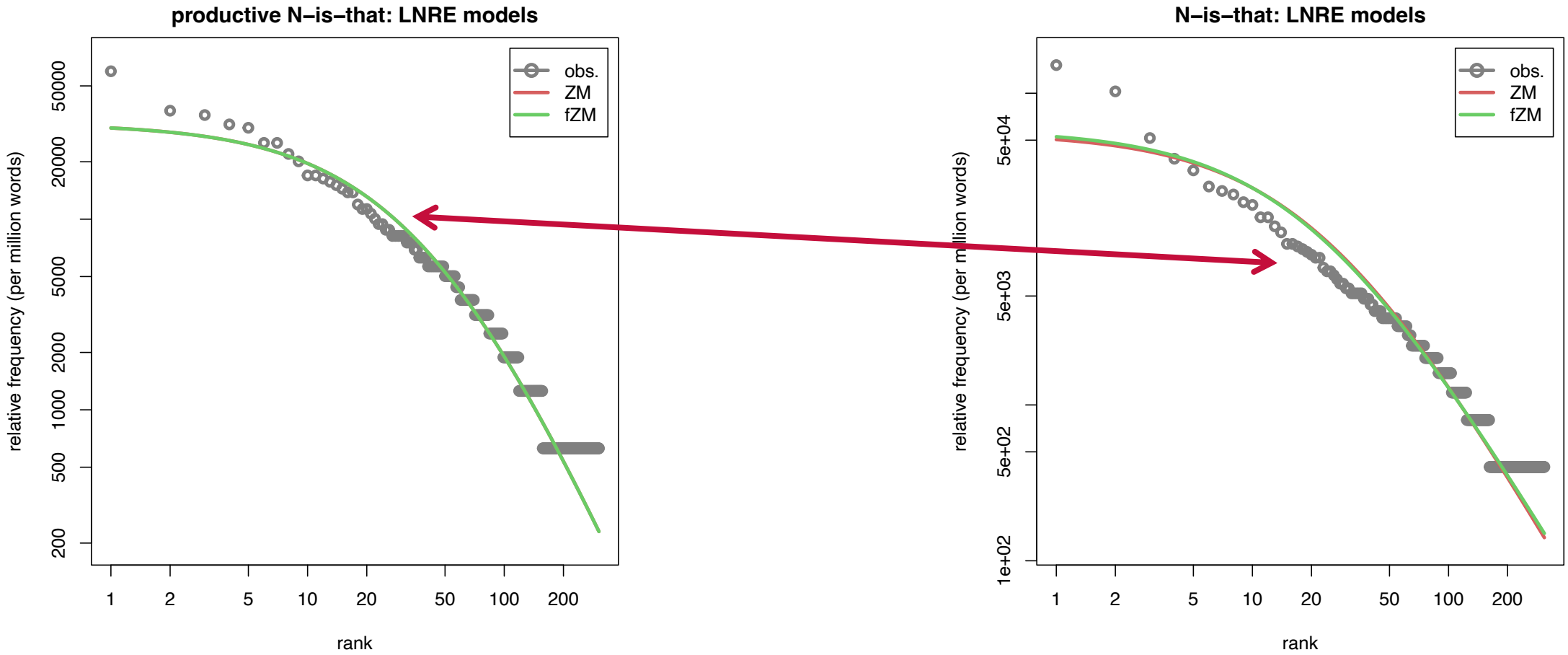
$$\pi_i = \frac{C}{(i+b)^a} \qquad (i = 1, \ldots, S)$$







Why we needed to exclude lexicalised types



The mathematics behind LNRE models

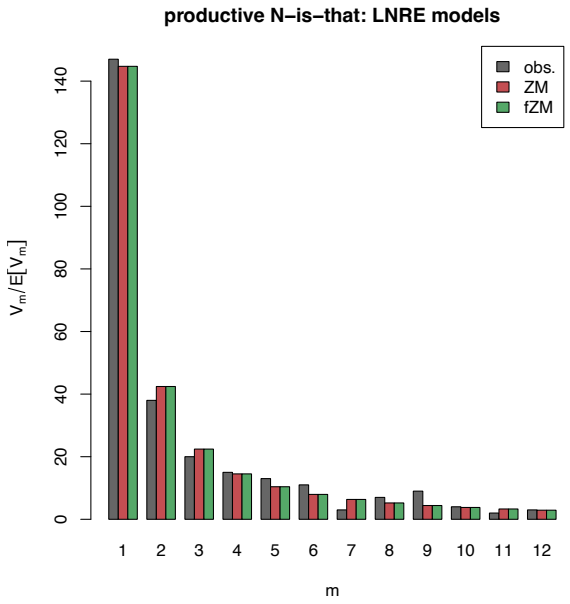
https://zipfr.r-forge.r-project.org/lrec2018.html

- parameter estimation fits E[V_m] to V_m and E[V] to V
  - simplification: Poisson sampling instead of binomial (conditioned on N) ➞ good approximation for large N

$$E[V_m] = \sum_{i=1}^{S} e^{-N\pi_i} \frac{(N\pi_i)^m}{m!} \qquad E[V] = \sum_{i=1}^{S} \left(1 - e^{-N\pi_i}\right)$$

- hypothesis testing & extrapolation based on Var[V_m] and Var[V]
  - goodness-of-fit test (multivariate chi-squared)
  - expected vocabulary growth curves with prediction intervals

- problem: discrete finite sums are inconvenient
  - both mathematically and numerically
  - especially for large samples ($N > 10^6$) and populations ($S > 10^5$)



The mathematics behind LNRE models

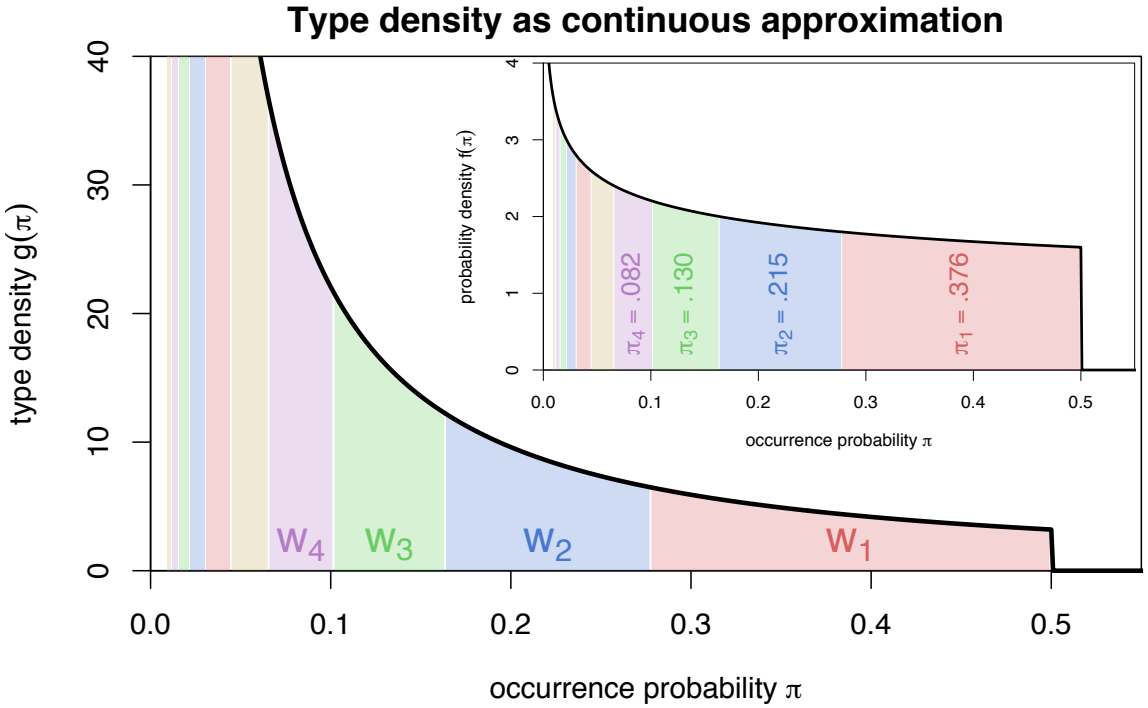
https://zipfr.r-forge.r-project.org/lrec2018.html

– simplification: approximate discrete types by continuous type density function $g(\pi)$

$$\pi_i = \frac{C}{(i+b)^a}$$

$$\Updownarrow$$

$$g(\pi) = \begin{cases} C \cdot \pi^{-\alpha-1} & A \leq \pi \leq B \\ 0 & \text{otherwise} \end{cases}$$

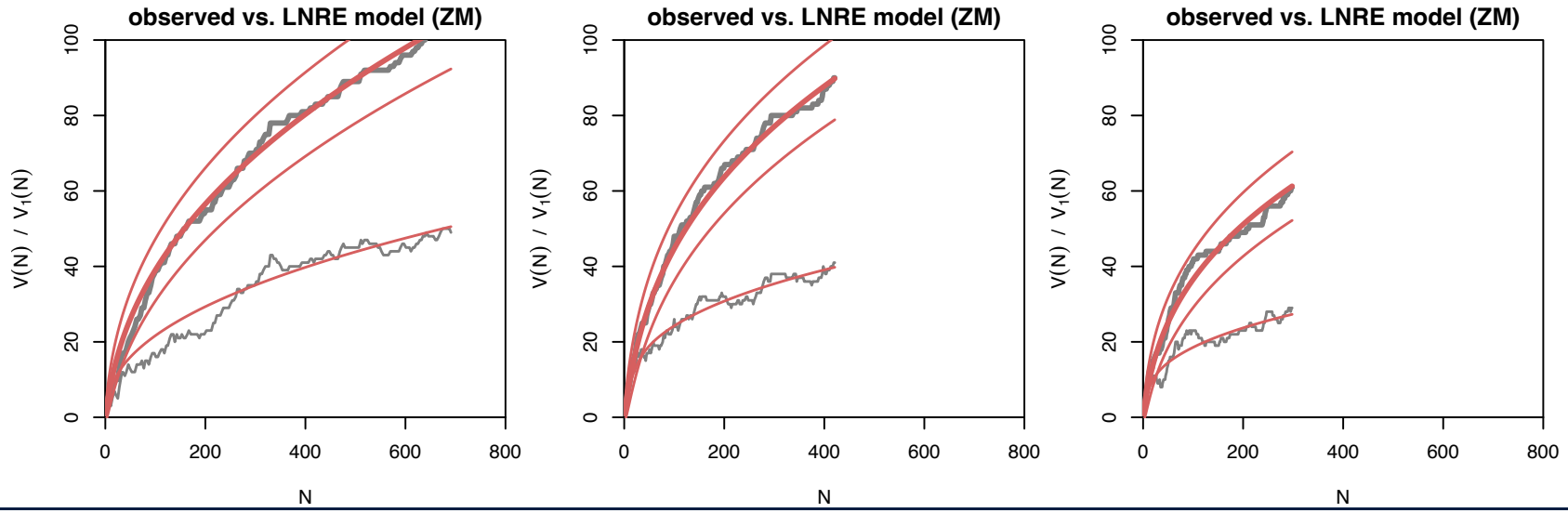


– discrete sums replaced by continuous integrals

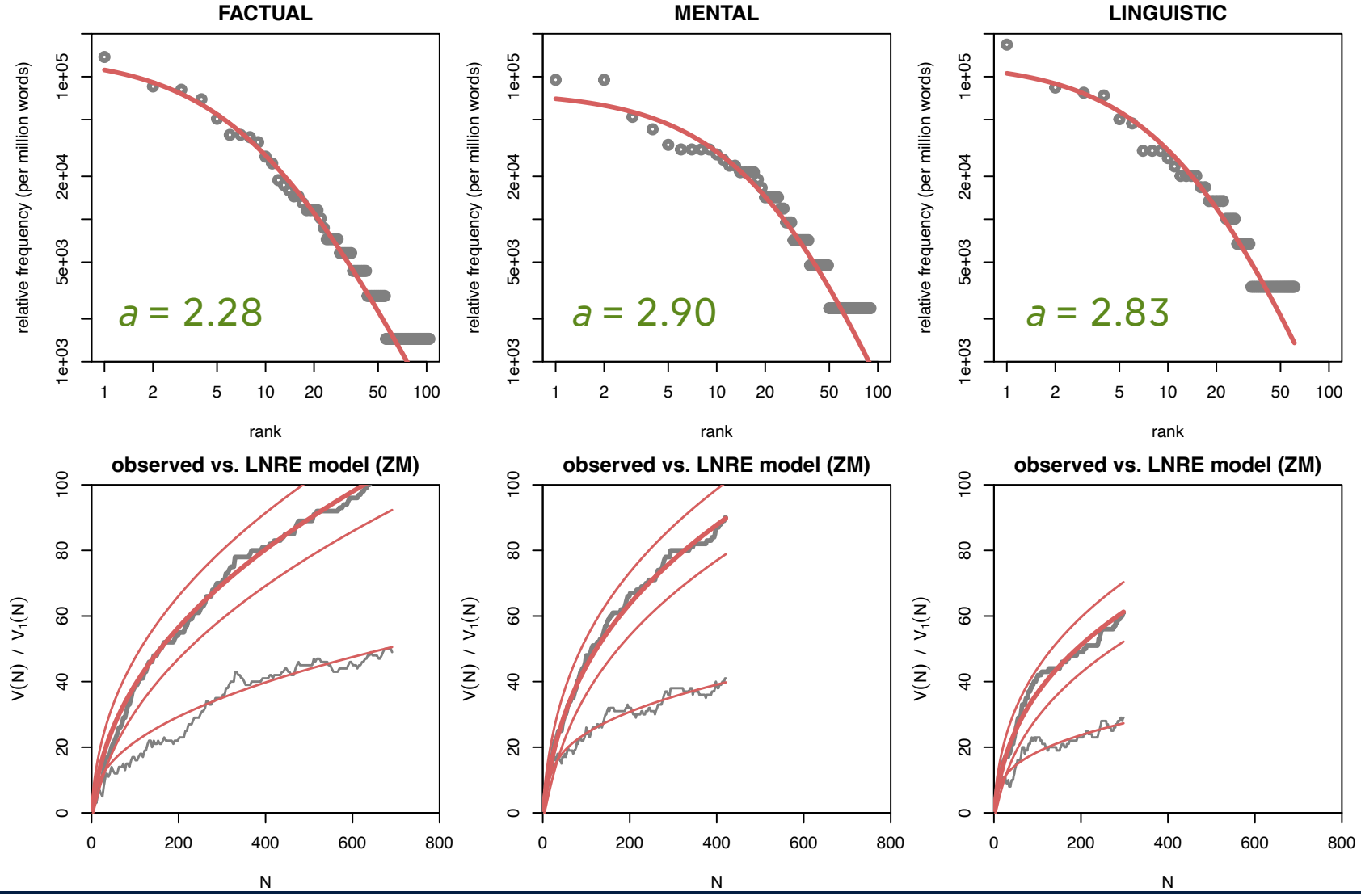
$$E[V_m] = \int_0^\infty \frac{(N\pi)^m}{m!} e^{-N\pi} g(\pi) d\pi \qquad E[V] = \int_0^\infty \left(1 - e^{-N\pi}\right) g(\pi) d\pi$$

Comparison of shell noun classes

FACTUAL	MENTAL	LINGUISTIC

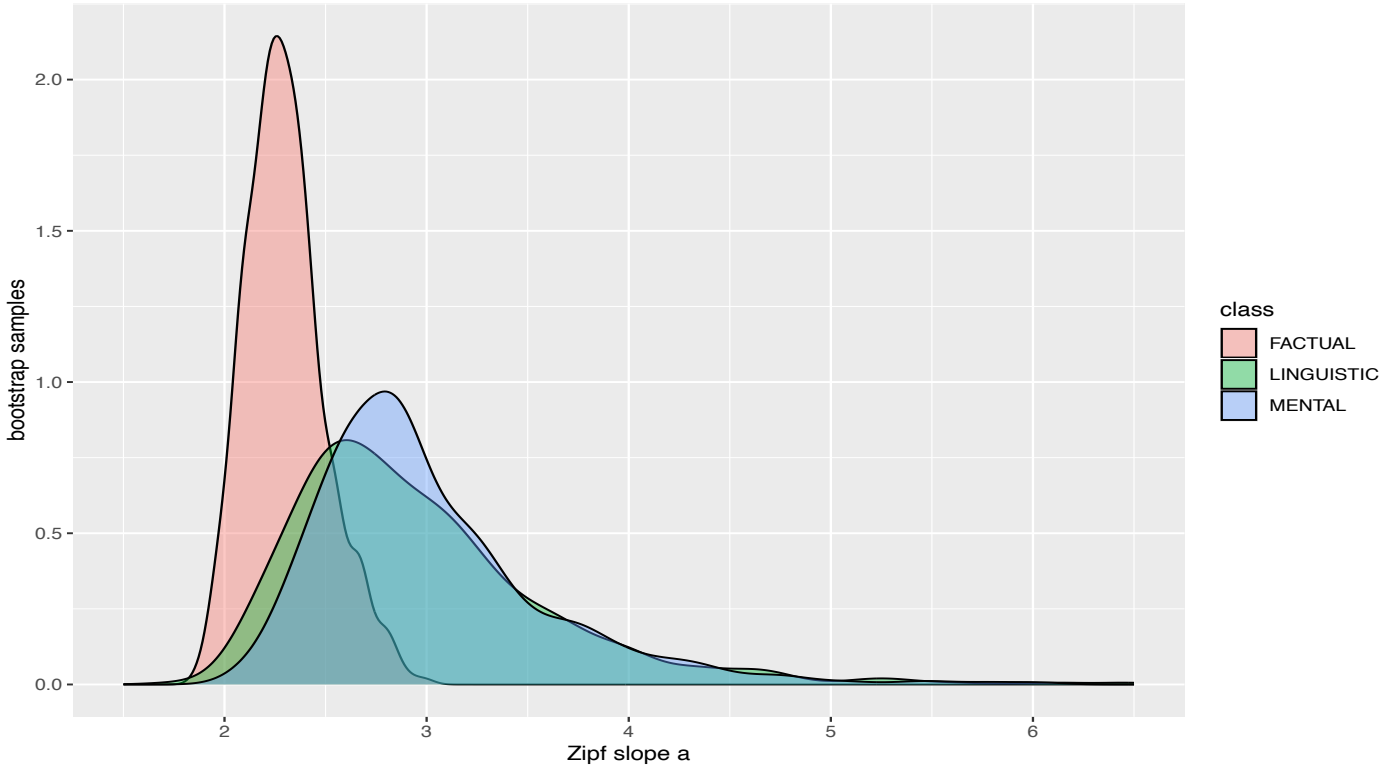


Comparison of shell noun classes



A caveat: sampling variation

- estimation of LNRE models difficult: sampling variation for small data sets
  - a different random sample from the same population might result in substantially different estimated LNRE parameters
- simulation experiment: parametric bootstrapping from LNRE model
  - shows how much estimated Zipf slope differs from true Zipf slope of LNRE model across samples
  - good measure of statistical uncertainty
  - assess (non-)significance of differences
- reliable LNRE estimation is open problem!



3. Association

Indicators of association

keyword & collocation analysis

Grey area between fixedness and productivity: neither completely spontaneous nor fully lexicalised

– hypothesis: collocational patterns, statistical association of fillers with Cx

Quantitative approach:

– compare frequency of N in Cx slot with its frequency outside Cx

– can be quantified by a wide range of association measures or keyness measures

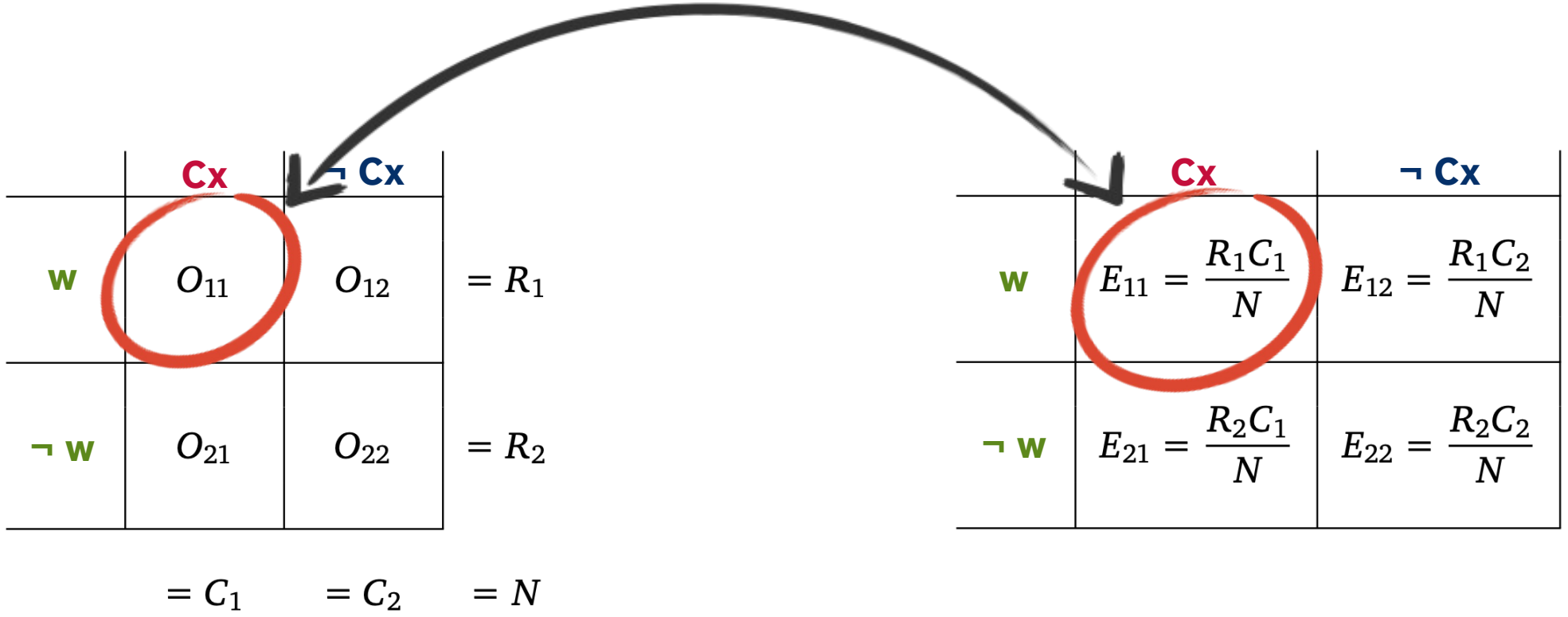
– statistical association is one of the central quantitative concepts in corpus linguistics!

– mathematical approach: contingency table

	N-BE-THAT pattern	N outside pattern
reason	f_1	f_2
¬ reason	$n_1 - f_1$	$n_2 - f_2$

Contingency tables & association measures

keyword & collocation analysis



	Cx	¬ Cx	
w	O_{11}	O_{12}	$= R_1$
¬ w	O_{21}	O_{22}	$= R_2$
	$= C_1$	$= C_2$	$= N$

	Cx	¬ Cx
w	$E_{11} = \frac{R_1 C_1}{N}$	$E_{12} = \frac{R_1 C_2}{N}$
¬ w	$E_{21} = \frac{R_2 C_1}{N}$	$E_{22} = \frac{R_2 C_2}{N}$

H_0: statistical independence

$p_{cooc} = p_1 \times p_2$

observed

expected

Contingency tables & association measures

keyword & collocation analysis

recommended measures (Evert 2008)

$$\text{MI} = \log_2 \frac{O}{E}$$

$$\text{MI}^k = \log_2 \frac{O^k}{E}$$

$$\text{local-MI} = O \cdot \log_2 \frac{O}{E}$$

$$\text{z-score} = \frac{O - E}{\sqrt{E}}$$

$$\text{t-score} = \frac{O - E}{\sqrt{O}}$$

$$\text{simple-ll} = 2 \left(O \cdot \log \frac{O}{E} - (O - E) \right)$$

$$\text{chi-squared} = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$\text{chi-squared}_{\text{corr}} = \frac{N\left(|O_{11}O_{22} - O_{12}O_{21}| - N/2\right)^2}{R_1 R_2 C_1 C_2}$$

$$\text{log-likelihood} = 2 \sum_{ij} O_{ij} \log \frac{O_{ij}}{E_{ij}}$$

$$\text{average-MI} = \sum_{ij} O_{ij} \cdot \log_2 \frac{O_{ij}}{E_{ij}}$$

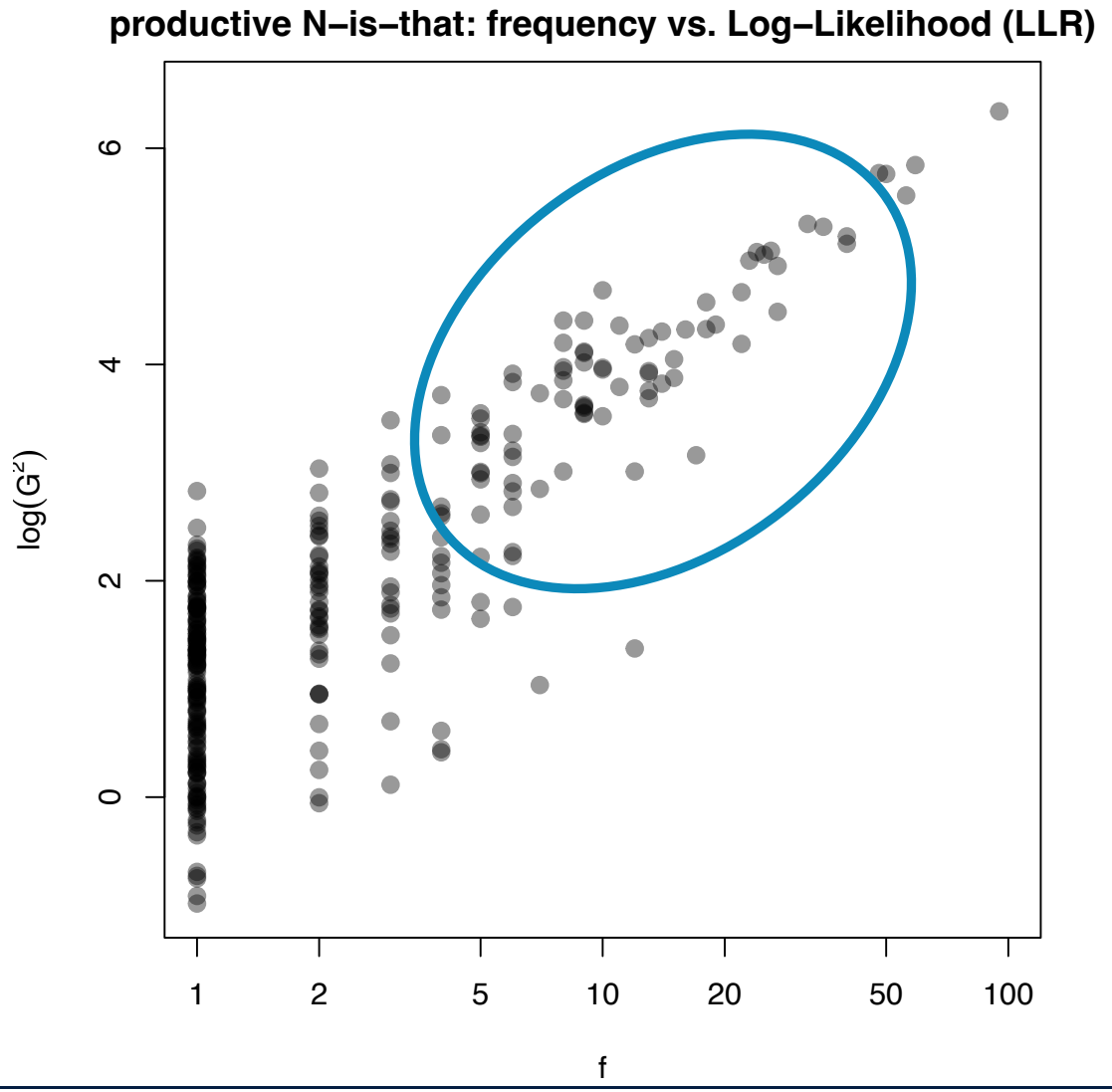
$$\text{Dice} = \frac{2O_{11}}{R_1 + C_1}$$

$$\text{odds-ratio} = \log \frac{\left(O_{11} + \frac{1}{2}\right)\left(O_{22} + \frac{1}{2}\right)}{\left(O_{12} + \frac{1}{2}\right)\left(O_{21} + \frac{1}{2}\right)}$$

Measures of significance

http://www.collocations.de/AM/

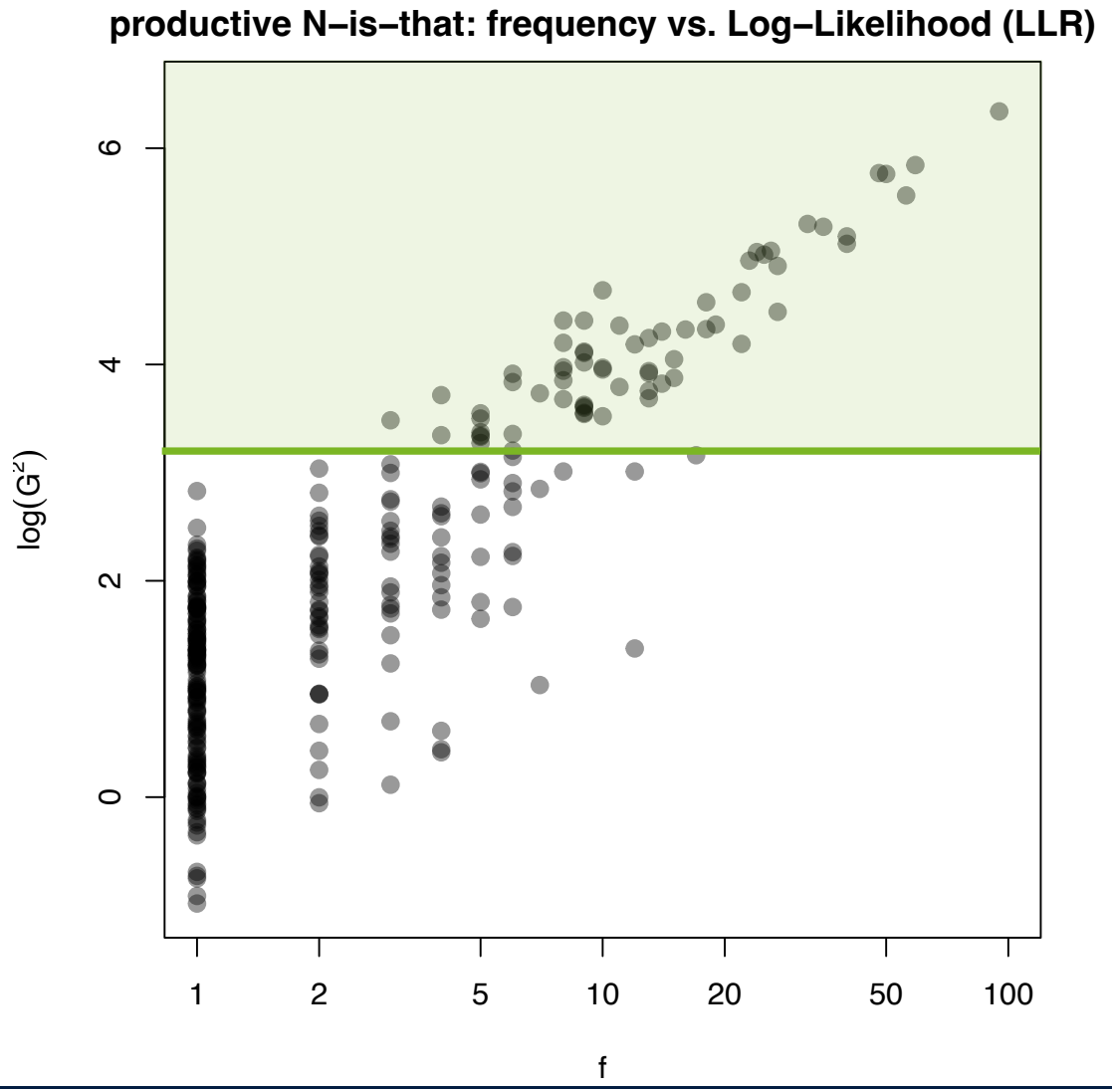
- recommended measure: log-likelihood G^2
  - goal: differentiation between medium-frequency noun types
- known for strong correlation with frequency
  - provides little additional information on collocability of nouns



Measures of significance

http://www.collocations.de/AM/

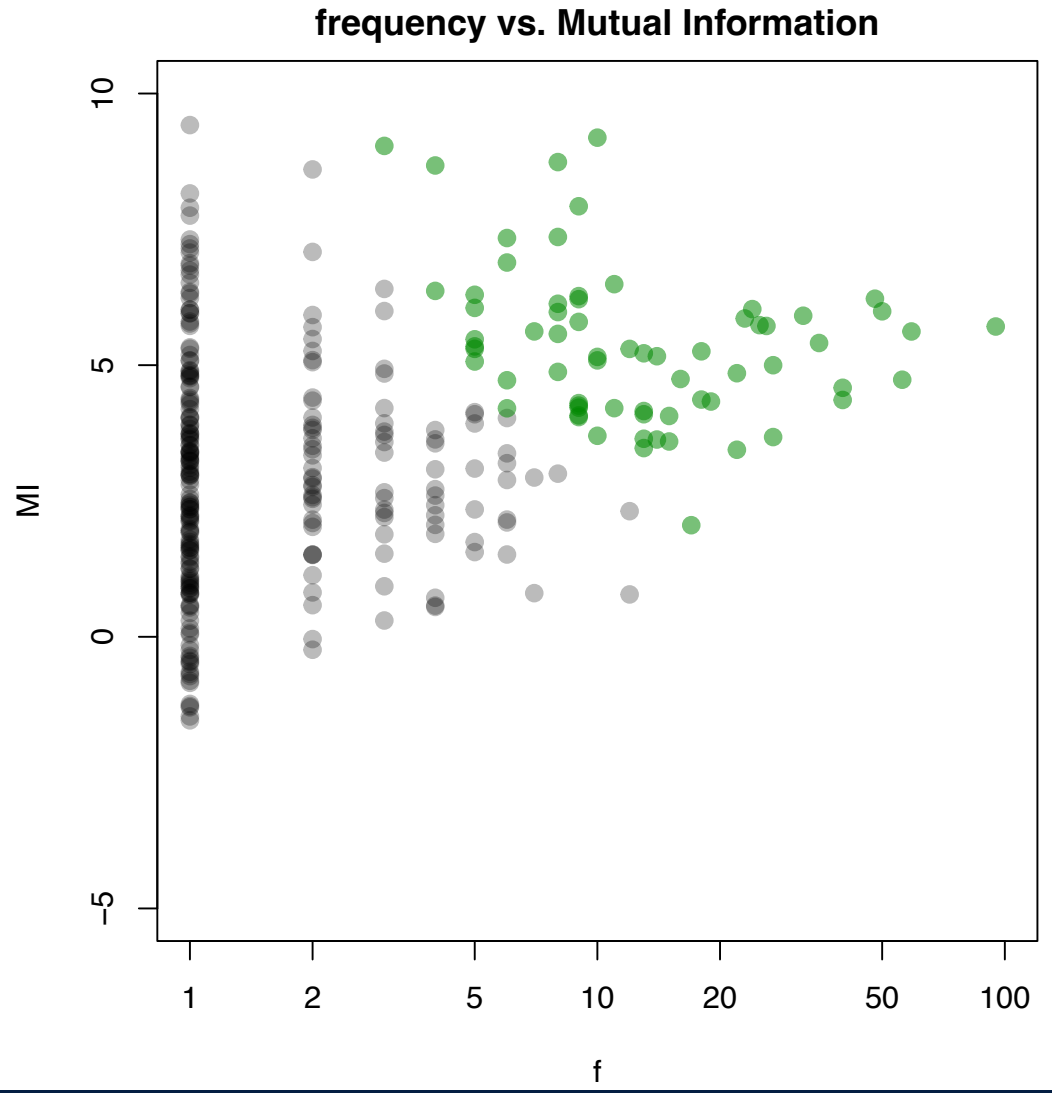
- recommended measure: log-likelihood G^2
  - goal: differentiation between medium-frequency noun types
- known for strong correlation with frequency
  - provides little additional information on collocability of nouns
- but useful as significance test
  - need to correct for multiple tests by controlling family-wise error rate (FWER) with Bonferroni adjustment (Hardie 2014)
  - resulting significance cutoff $G^2 \geq 15.6$
- nouns with reliable evidence for a positive association
  - NB: other nouns might collocate with Cx, too



Measures of effect size

http://www.collocations.de/AM/

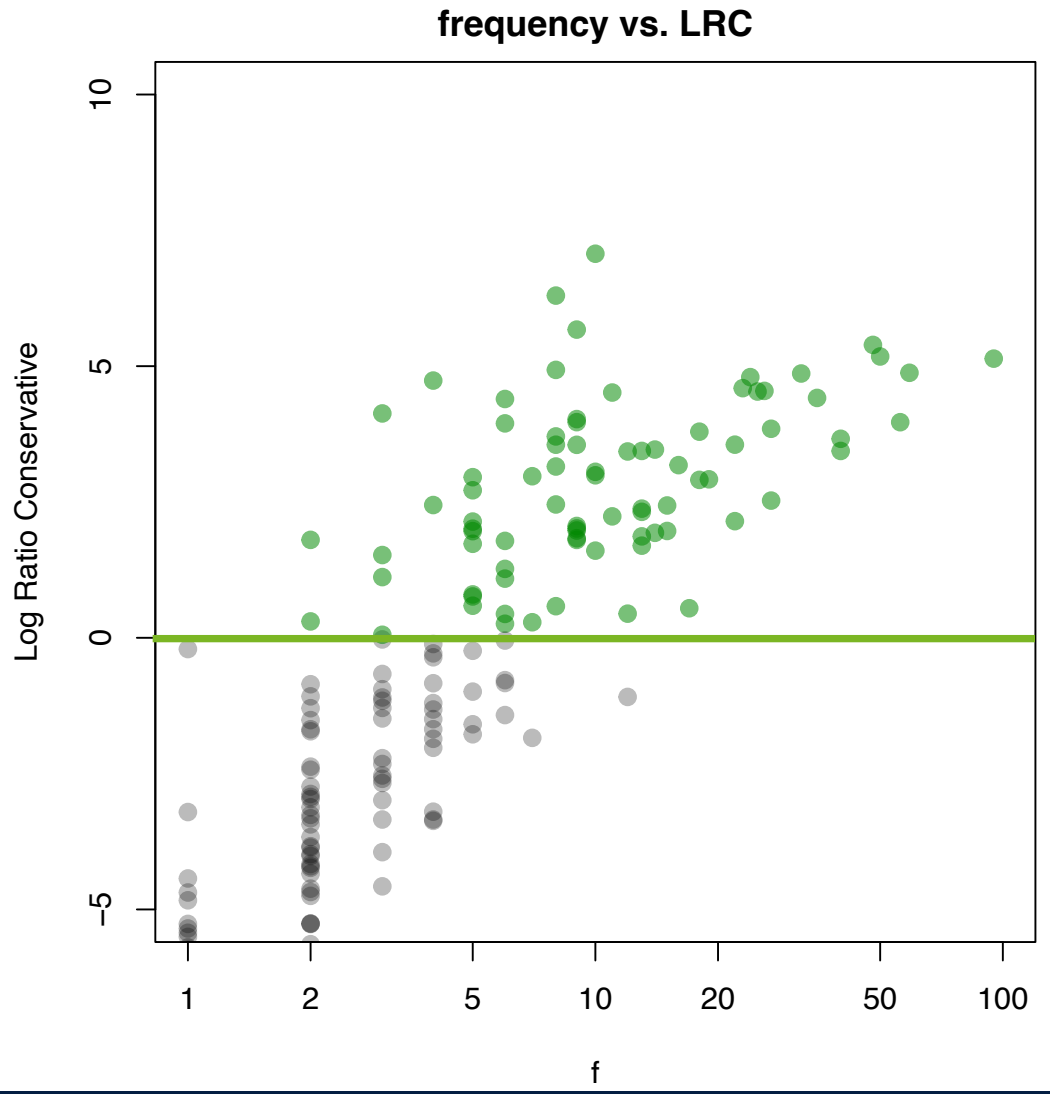
- highly popular measure: mutual information MI
  - especially in lexicography (Church & Hanks 1990)
- well-known bias towards low-frequency nouns
  - productive hapax and dis legomena among highest MI values
  - not very useful without additional filters
- combine with G^2 significance test (green points)
  - MI measures degree of association for significant collocates



Conservative estimates

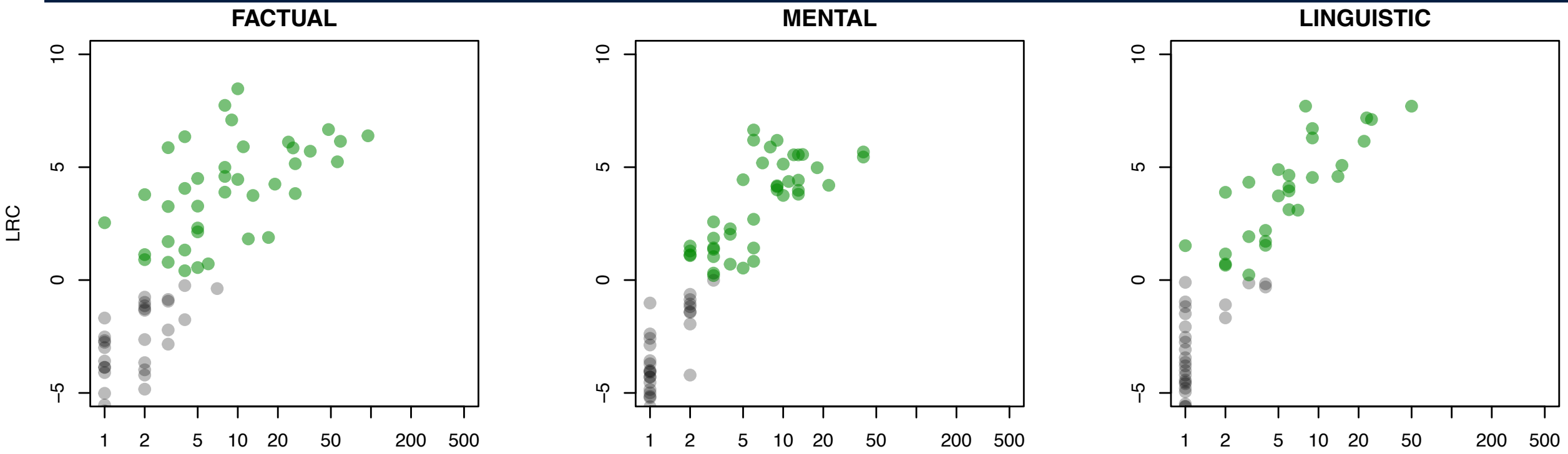
http://www.collocations.de/AM/ | https://osf.io/cy6mw/

- conservative estimates combine both aspects: Conservative LogRatio LRC (Evert 2022)
  - based on confidence interval for relative risk
  - recommended as first-line assocation and keyness measure
- LRC shows LogRatio supported by significant evidence
  - true LogRatio value will often be higher (but can't be confident)
  - significance cutoff LRC ≥ 0
- observation: overlap of ranges, with significant association for productive nouns as rare as $f = 2$ or 3
- observation: strongest associations found in medium frequency range



Comparison of shell noun classes

top LRC collocates with f < 20



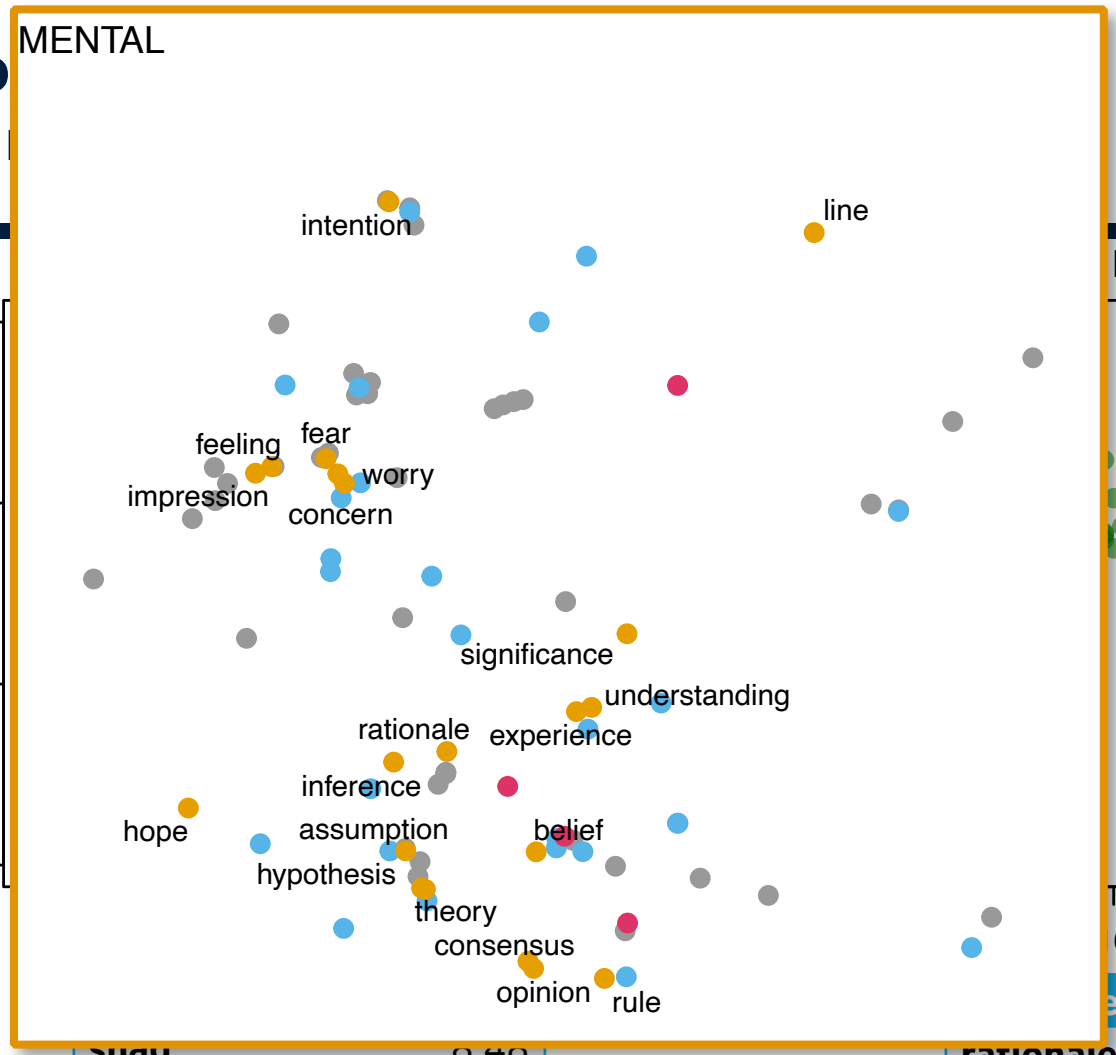
collocate	LRC
snag	8.48
corollary	7.74
drawback	7.09
upshot	6.35
essence	5.90
downside	5.86

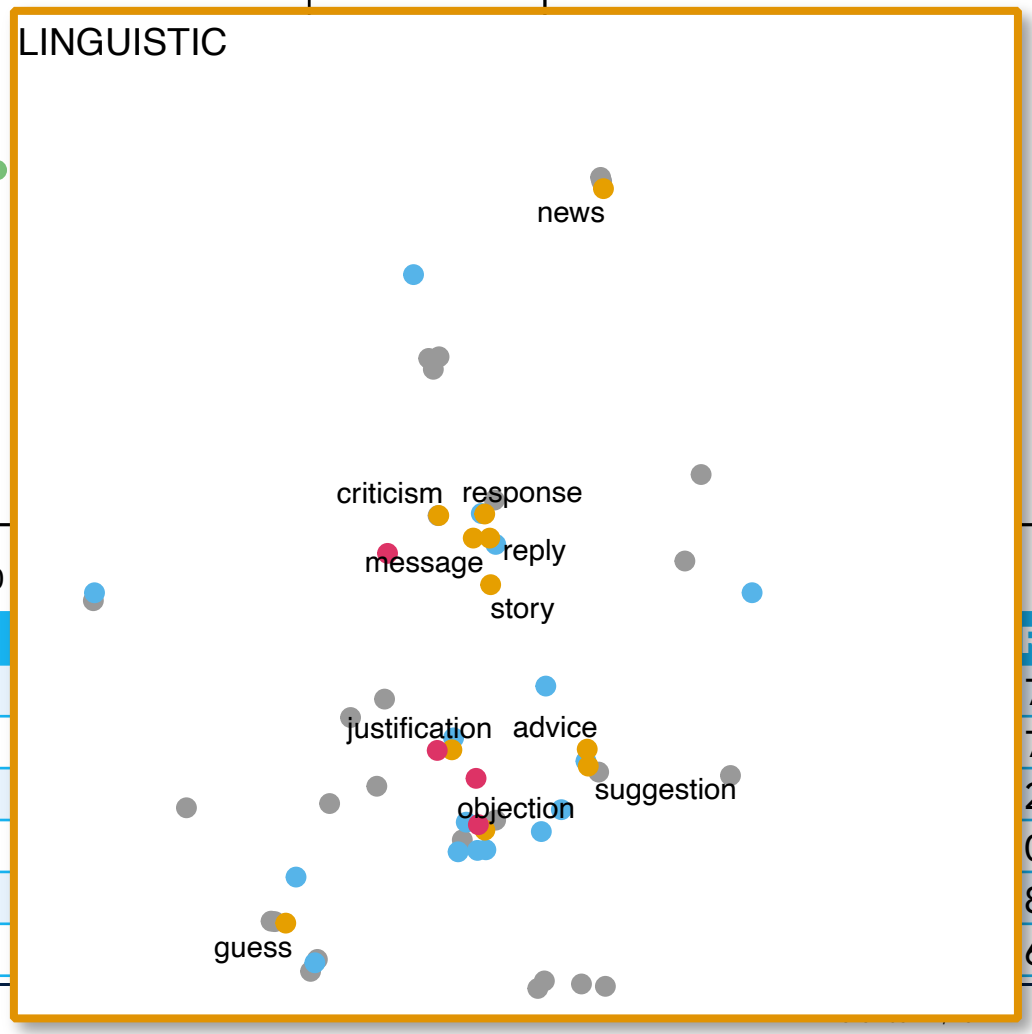
collocate	LRC
rationale	6.65
inference	6.20
worry	6.19
consensus	5.89
intention	5.56
impression	5.55

collocate	LRC
guess	7.70
justification	6.71
objection	6.29
news	5.08
proposition	4.89
reply	4.64

Co

top





corollary	7.74
drawback	7.09
upshot	6.35
essence	5.90
downside	5.86

rationale
inference
worry
consensus
intention
impression

RC
70
71
29
08
89
64

Conclusion

Conclusions & next steps

Fixedness — productivity continuum for variable slots in lexico-syntactic patterns

– indicators of fixedness: frequency, dispersion, polysemy, …

– indicators of productivity: type-token statistics, LNRE models, semantic regularity, …

– indiactors of association: statistical association measures, semantic clustering, …

Bigger picture:

– correlations between multiple slots (association, productivity)

– application in CxG: **comparison of candidate Cx**

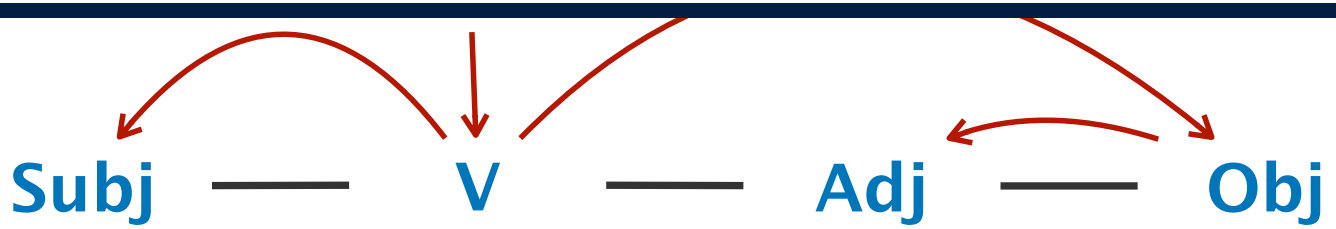
Research training group GRK 2839: Dimensions of Constructional Space

– PhD project *Corpus evidence for delineating constructions* (Malin Patel)

– see https://www.cxg.phil.fau.eu/project-1/

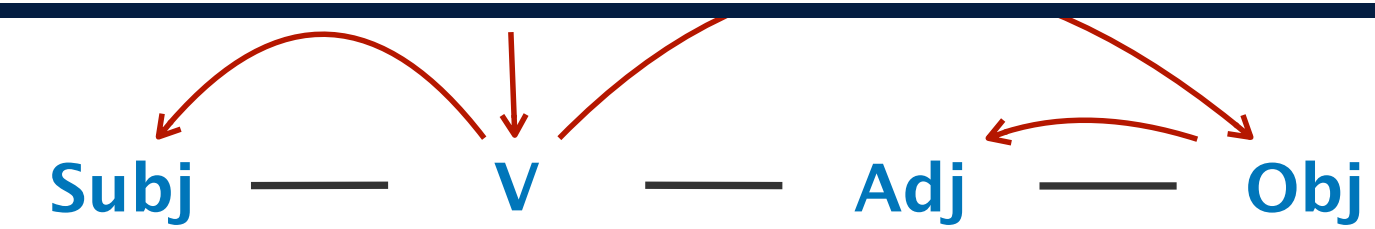
Delineating Cx: The “slot” model

corpus evidence on Cx as tabular data



Delineating Cx: The “slot” model

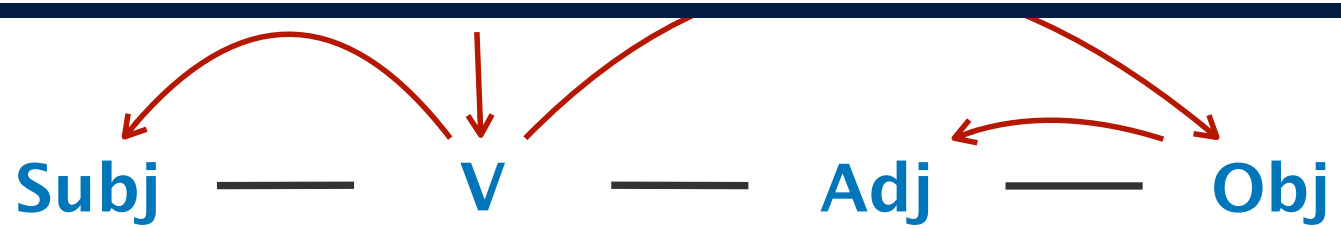
corpus evidence on Cx as tabular data



Subj	V	Adj	Obj
they	earn	—	money
he	earn	—	chance
manager	earn	—	—
I	earn	first	salary
labourer	earn	—	more
—	earn	more	money
Jane	earn	much	sympathy
—	earn	—	salary
Doris	buy	fresh	food
—	buy	—	something
you	buy	—	it
they	buy	nationalist	support
…	…	…	…

Delineating Cx: The “slot” model

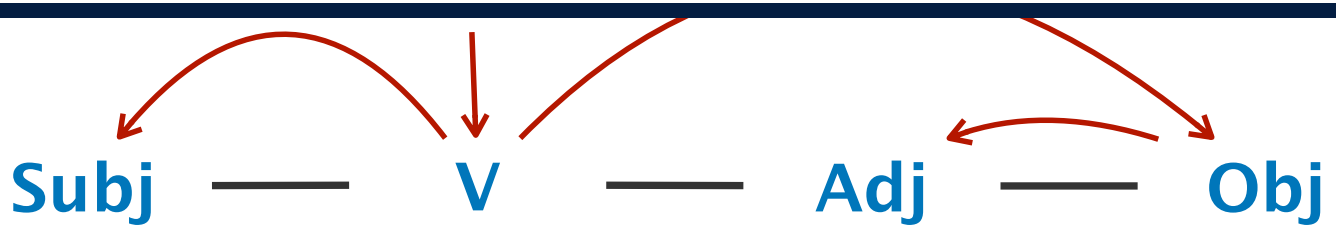
lexico-syntactic patterns = Cx hypotheses



Different fixed & variable-slot LSP within this frame:

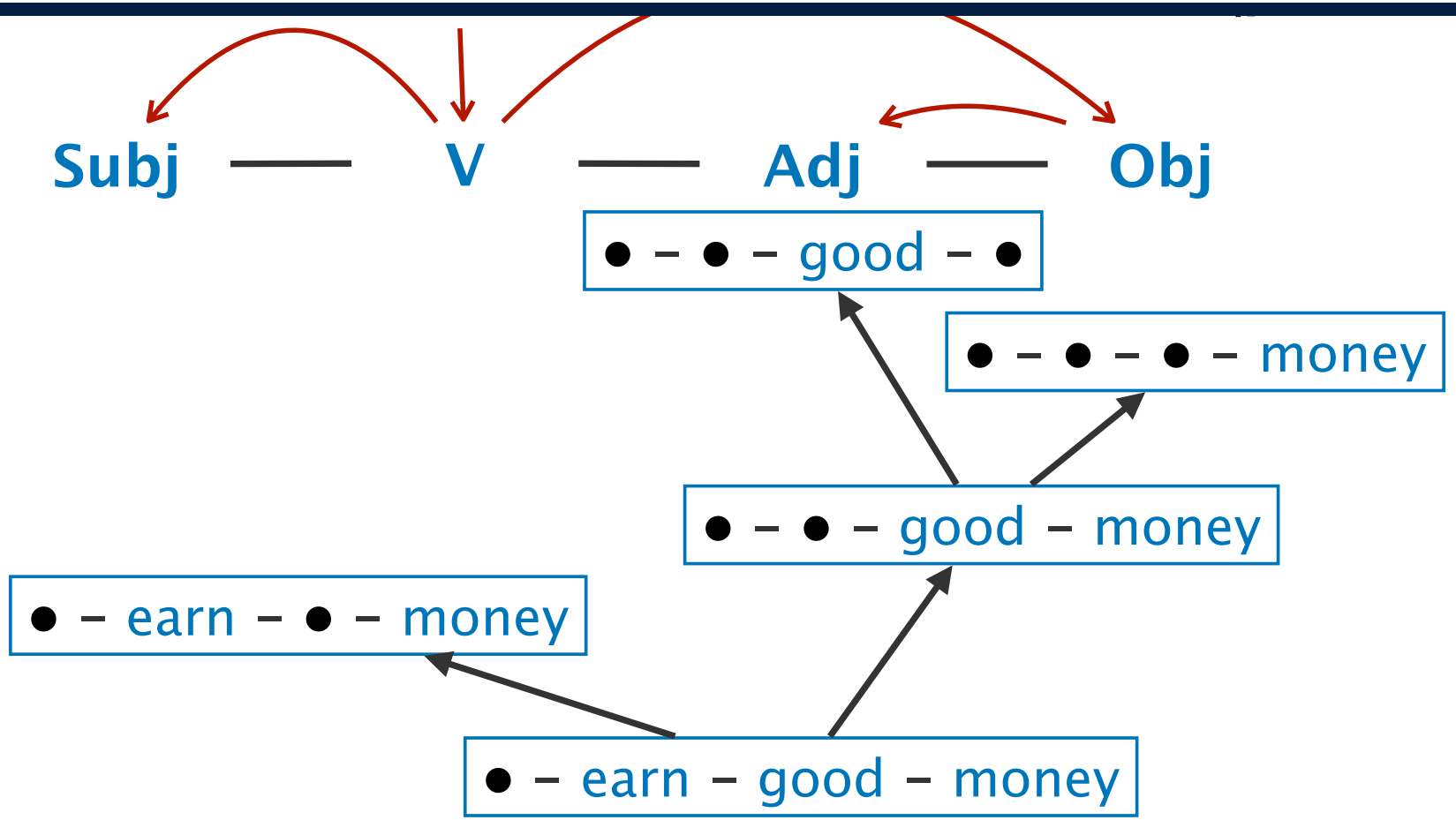
- ● – earn – ● – money
- worker – earn – ● – ●
- ● – earn – good – money
- ● – earn – × – keep
- company – earn – huge – profit
- **Pron** – earn – **A** – support
- worker – earn – ● – [MONEY]

Delineating Cx: Pairwise comparisons



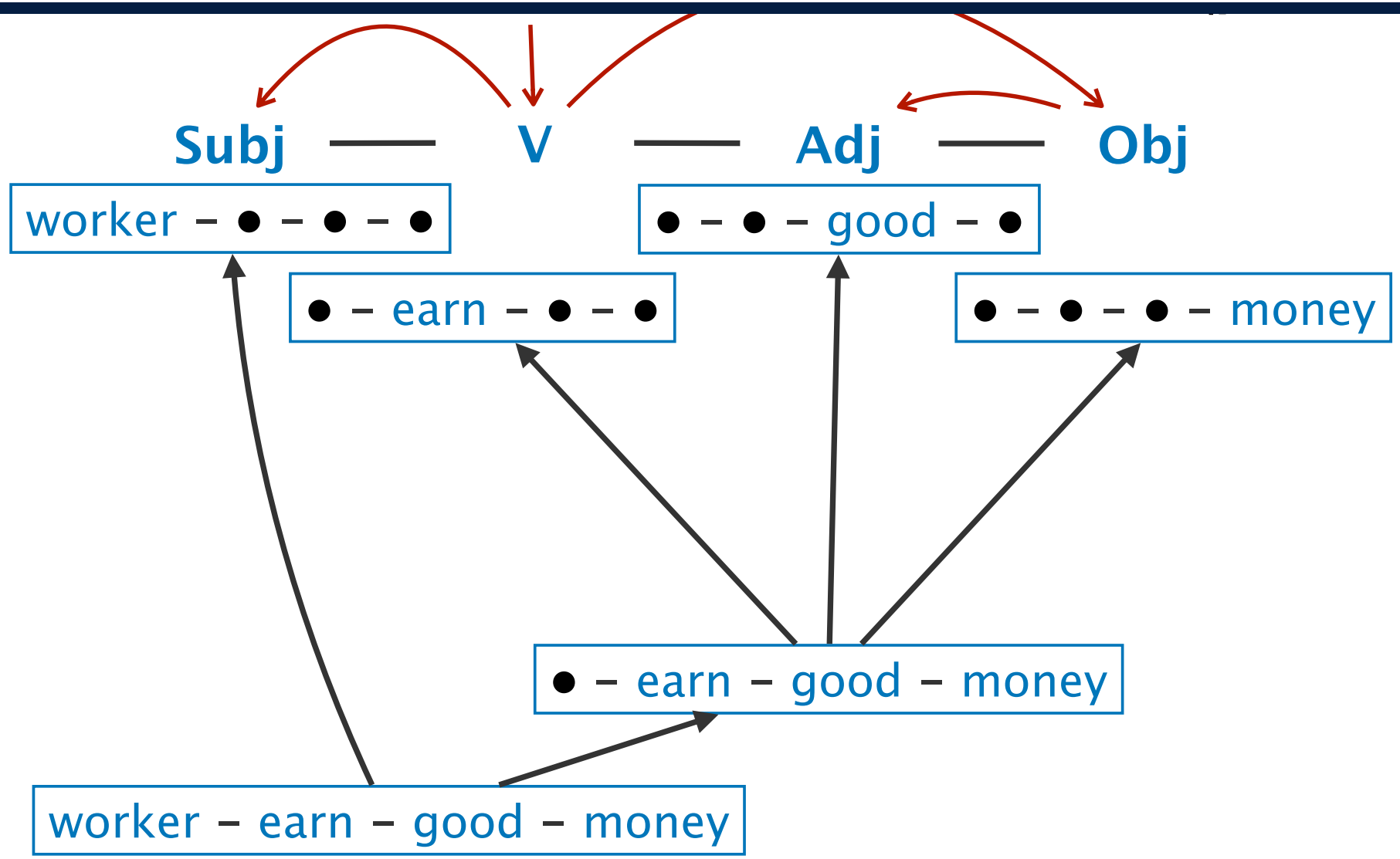
● – earn – good – money

Delineating Cx: Pairwise comparisons



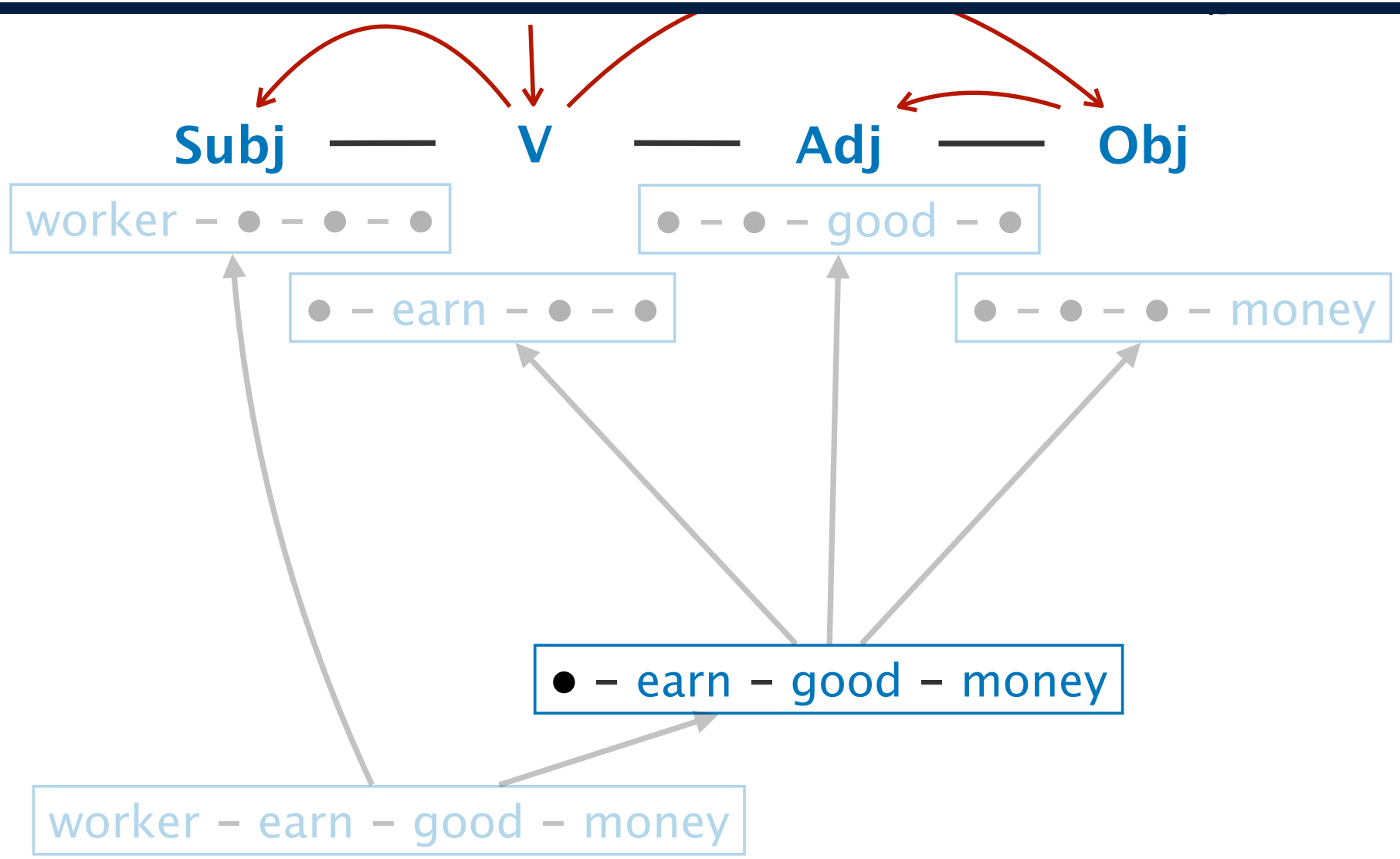
association

Delineating Cx: Pairwise comparisons

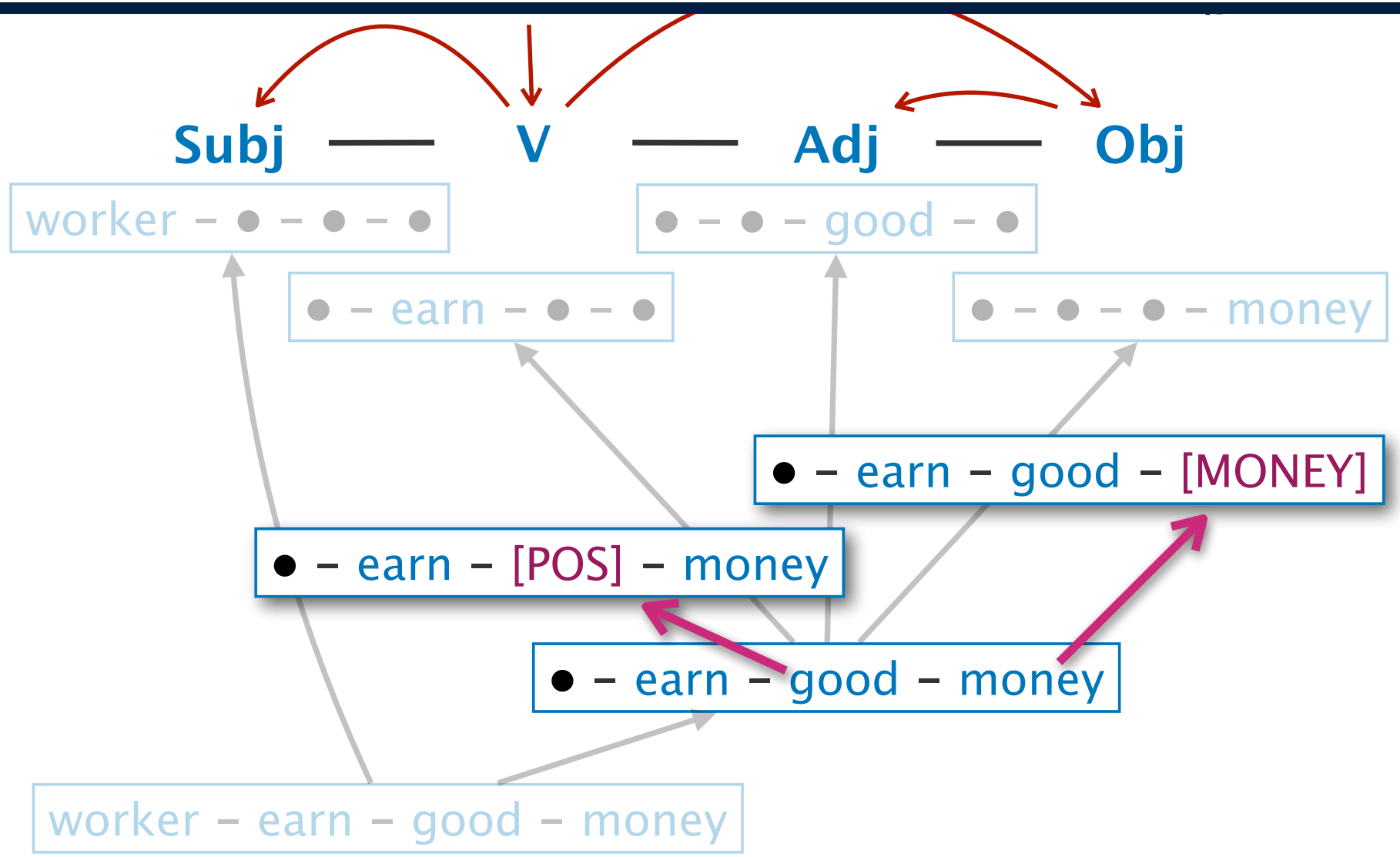


association

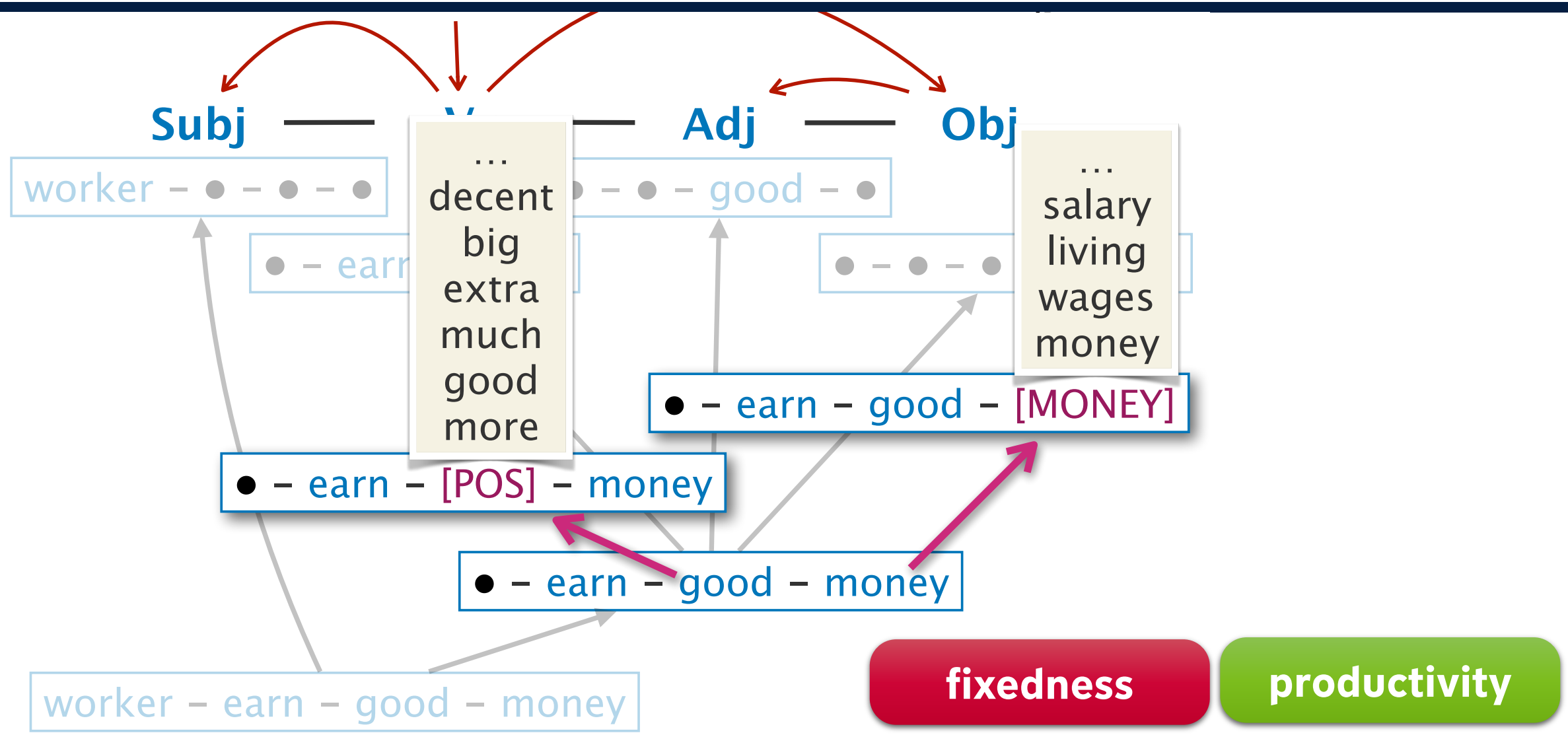
Delineating Cx: Pairwise comparisons



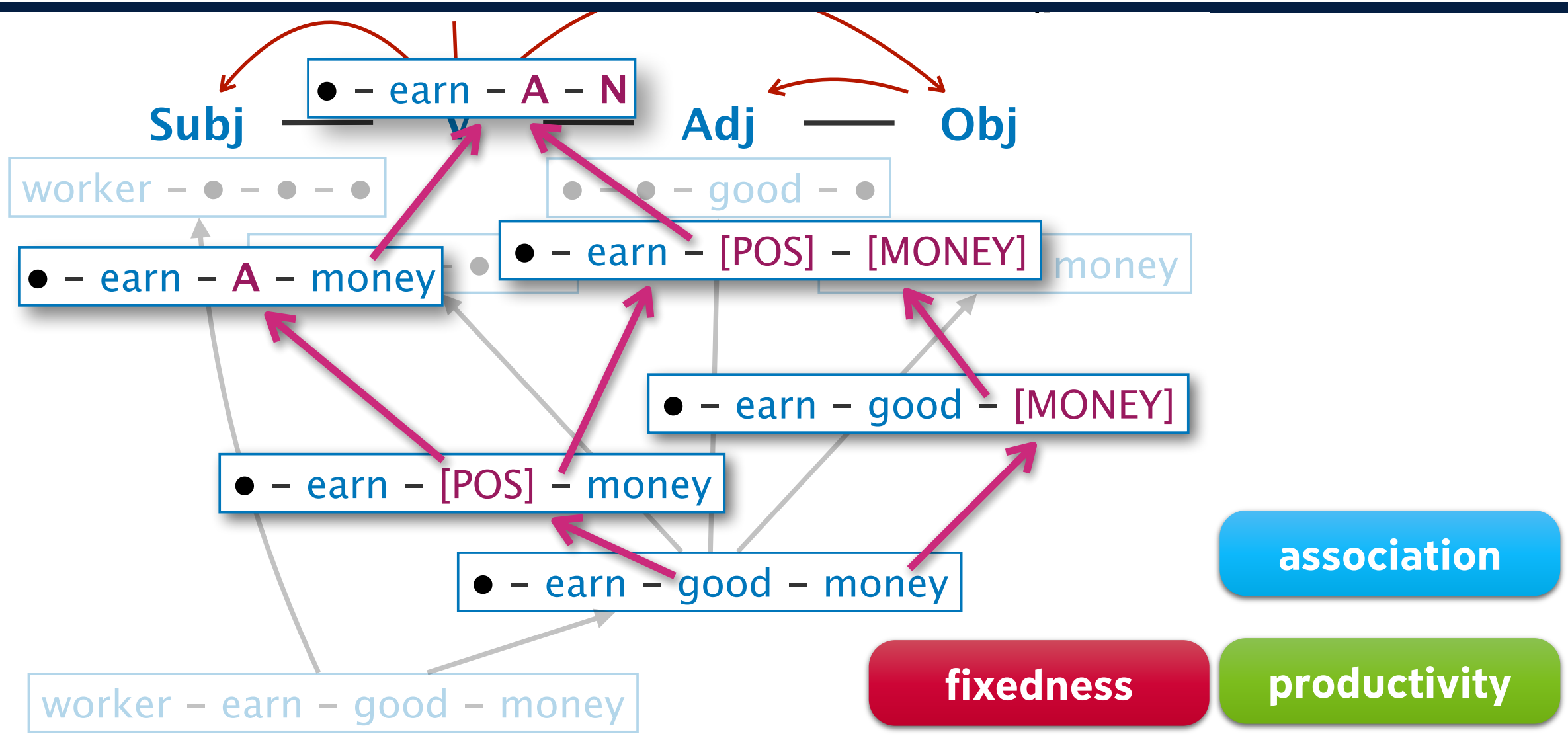
Delineating Cx: Pairwise comparisons



Delineating Cx: Pairwise comparisons



Delineating Cx: Pairwise comparisons



Collaboration in FAU MoD?

What I need help with …

– more sophisticated statistical / quantitative analysis

– e.g. efficient Bayesian inference w/o continuous approximation, accounting for non-randomness

– e.g. better measures of statistical association patterns (including semantics?)

– e.g. association & productivity across multiple slots / between candidate Cx

My overarching research programme: **Digital Hermeneutics**

– goal: integration of human interpretation with quantitative methods and mathematical analysis

– challenges: understanding quantitative results + feedback of human insights into automatic analysis

– other ongoing projects: MMDA, RC21, QuanTOR, DIREGA

– possible connection to model discovery research?

FAU

Thank you for listening!

Questions? Comments?

References

- Baayen, R. H. (2001). *Word Frequency Distributions*. Kluwer Academic Publishers, Dordrecht.
- Baayen, H. (1992). Quantitative aspects of morphological productivity. *Yearbook of Morphology 1991*, pages 109–149.
- Church, K. W. (2000). Empirical estimates of adaptation: The chance of two Noriegas is closer to p/2 than p^2. In *Proceedings of COLING 2000*, pages 173–179, Saarbrücken, Germany.
- Church, K. W. & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, **16**(1):22–29.
- Diwersy, S; Evert, S.; Heinrich, P.; Proisl, T. (2019). Means of productivity – On the statistical modelling of the restrictedness of lexico-grammatical patterns. Presentation at *EUROPHRAS 2019: Productive Patterns in Phraseology*, Santiago de Compostela, Spain.
- Evert, S. (2004). A simple LNRE model for random character sequences. In *Proceedings of the 7èmes Journées Internationales d'Analyse Statistique des Données Textuelles* (JADT 2004), pages 411–422, Louvain-la-Neuve, Belgium.
- Evert, S. (2008). Corpora and collocations. In Lüdeling, A. & Kytö, M., editors, *Corpus Linguistics. An International Handbook*, chapter 58, pages 1212–1248. Mouton de Gruyter, Berlin, New York.
- Evert, S. (2022). Measuring keyness. In *Digital Humanities 2022: Conference Abstracts*, pages 202–205, Tokyo, Japan / online. https://osf.io/cy6mw/.
- Evert, S. & Baroni, M. (2007). zipfR: Word frequency distributions in R. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstrations Sessions*, pages 29–32, Prague, Czech Republic.
- Gries, S. Th. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, **13**(4):403–437.
- Hardie, A. (2014). A single statistical technique for keywords, lockwords, and collocations. Internal CASS working paper no. 1, unpublished.

References

- Herbst, T. (2018). Is language a collostructicon? A proposal for looking at collocations, valency, argument structure and other constructions. In Cantos-Gómez, P. & Almela-Sánchez, M., editors, *Lexical Collocation Analysis: Advances and Applications*, pages 1–22. Springer International Publishing, Cham.
- Katz, S. M. (1996). Distribution of content words and phrases in text and language modelling. *Natural Language Engineering*, **2**(2):15–59.
- Lüdeling, A. & Evert, S. (2005). The emergence of productive non-medical -itis. corpus evidence and qualitative analysis. In Kepser, S. and Reis, M., editors, *Linguistic Evidence. Empirical, Theoretical, and Computational Perspectives*, pages 351–370. Mouton de Gruyter, Berlin.
- Mandelbrot, B. (1962). On the theory of word frequencies and on related Markovian models of discourse. In Jakobson, R., editor, *Structure of Language and its Mathematical Aspects*, pages 190–219. American Mathematical Society, Providence, RI.
- Schmid, H.-J. (2000). *English Abstract Nouns as Conceptual Shells*. Number 34 in Topics in English Linguistics. Mouton de Gruyter, Berlin, New York.
- Schmid, H.-J. (2018). Shell nouns in English: A personal roundup. *Caplletra. Revista Internacional de Filologia*, **64**:109–128.
- Stefanowitsch, A. & Gries, S. T. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, 8(2):209–243.
- Tweedie, F. J. and Baayen, R. H. (1998). How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities*, **32**(5):323–352.
- Zeldes, A. (2012). *Productivity in Argument Selection: From Morphology to Syntax*. Number 260 in Trends in Linguistics: Studies and Monographs (TiLSM). De Gruyter Mouton, Berlin, Boston.
- Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge, MA.