

Measuring productivity and fixedness in lexico-syntactic constructions

Prof. Dr. Stephanie Evert

Lehrstuhl für Korpus- und Computerlinguistik | www.linguistik.fau.de



fixed
arbitrary
lexicon
specific

productive
compositional
rules
schematic

Corpus Linguistics

Morphology

Construction Grammar (CxG)

Lexicography

Phraseology

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 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 crowded 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)



– 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

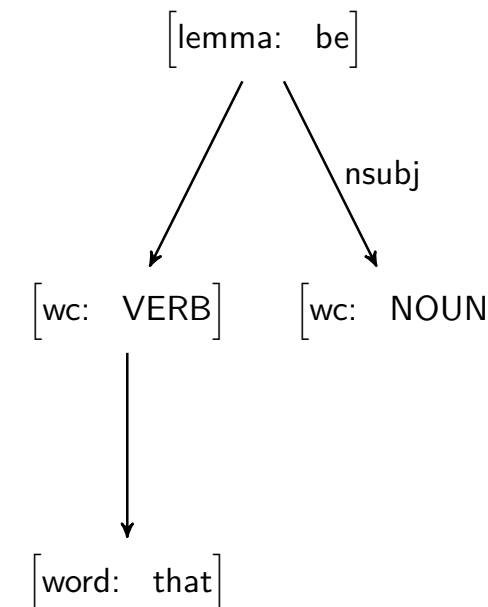
– many thanks to Sascha Diwersy for the hard work!

– data table for quantitative analysis

– noun lemma, realization of copula, main verb of *that*-clause, pre- and post-modification of noun, full sentence context, ...

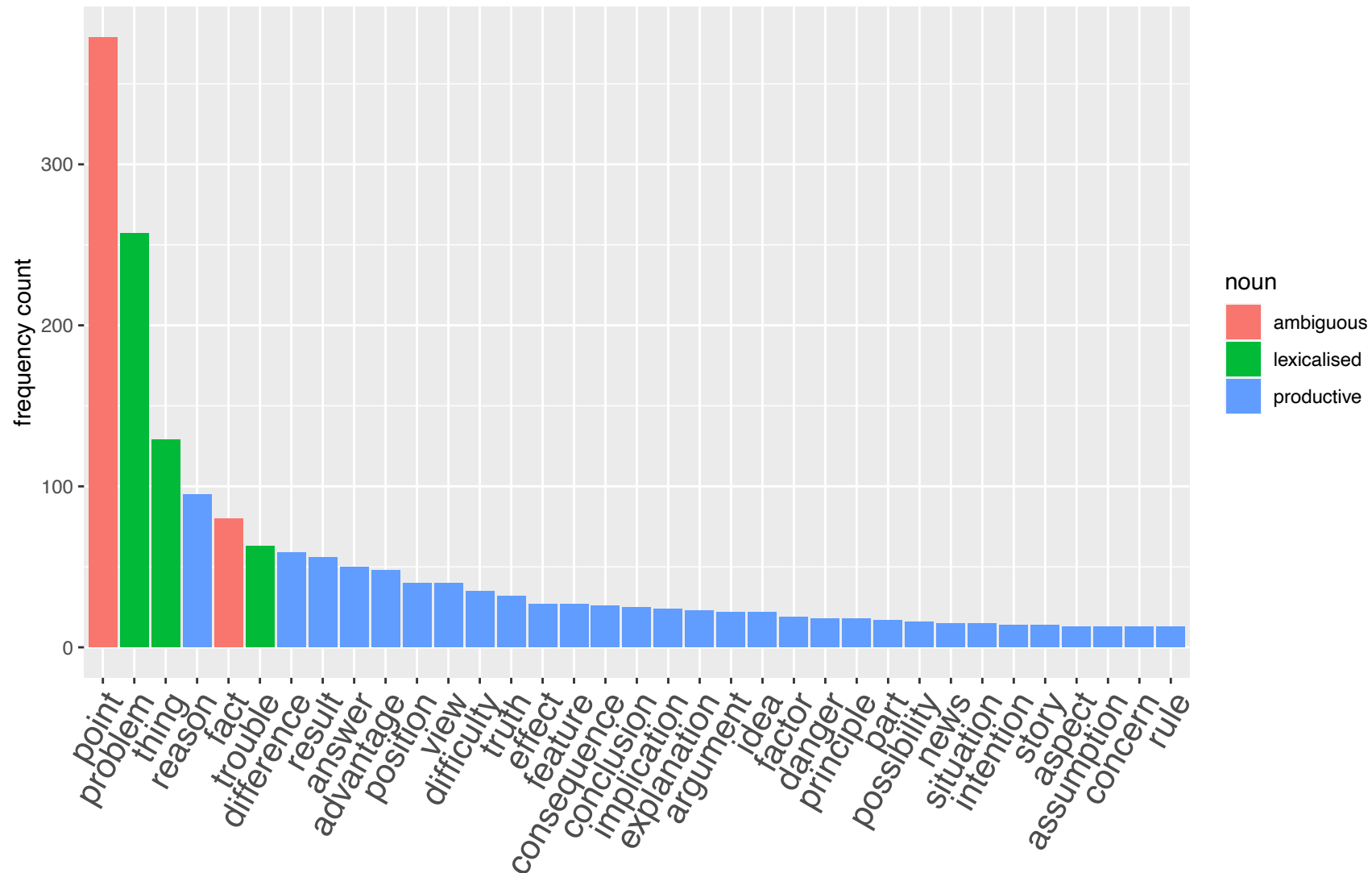
– manual classification of nouns into shell noun classes (NB: a few nouns are ambiguous)

– here: **focus on noun lemma slot** + **shell noun classes**



Frequency distribution of shell nouns

in N-BE-THAT construction

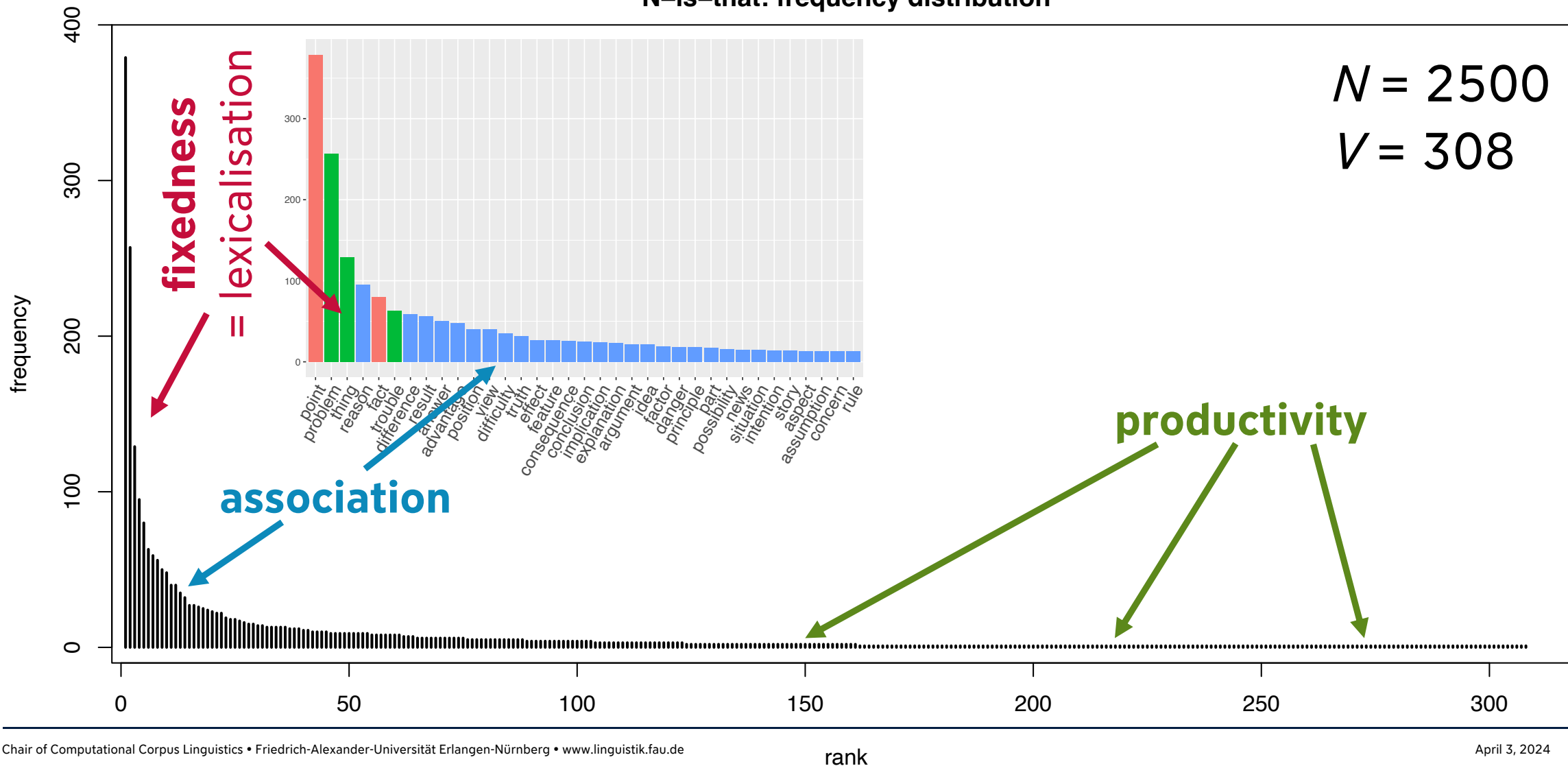


Frequency distribution of shell nouns

in N-BE-THAT construction



N-is-that: frequency distribution



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

- relative frequency (proportion of instances)
 - *point, problem, thing* each account for > 5% of Cx use → likely to be lexically fixed
 - but entirely unclear 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** $f_r \approx \frac{C}{(r + b)^a}$

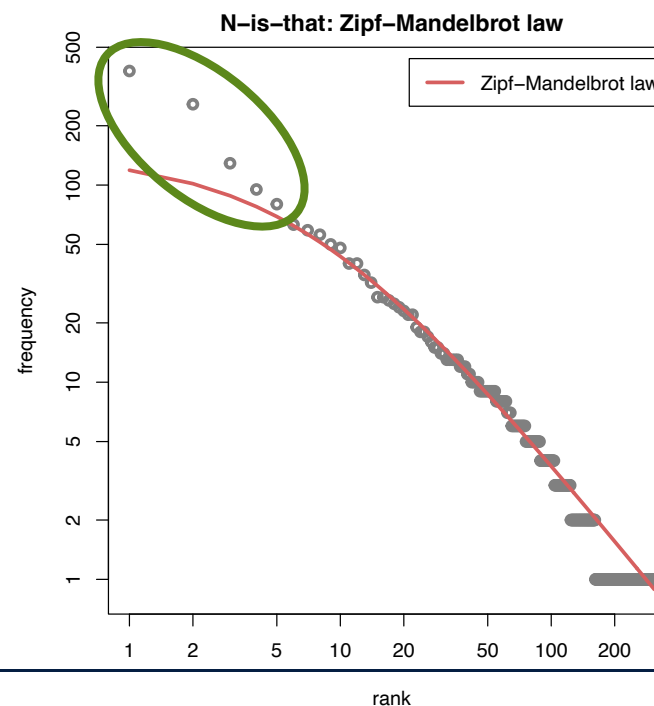
– (Zipf 1949; Mandelbrot 1962)

– 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



– term clustering / dispersion

- (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}	$E[f_{rep}]$
<i>point</i>	379	286	157	100.3
<i>problem</i>	257	216	75	50.8
<i>thing</i>	129	118	21	14.5
reason	95	90	9	8.2
<i>fact</i>	80	68	17	5.9
<i>trouble</i>	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

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)

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

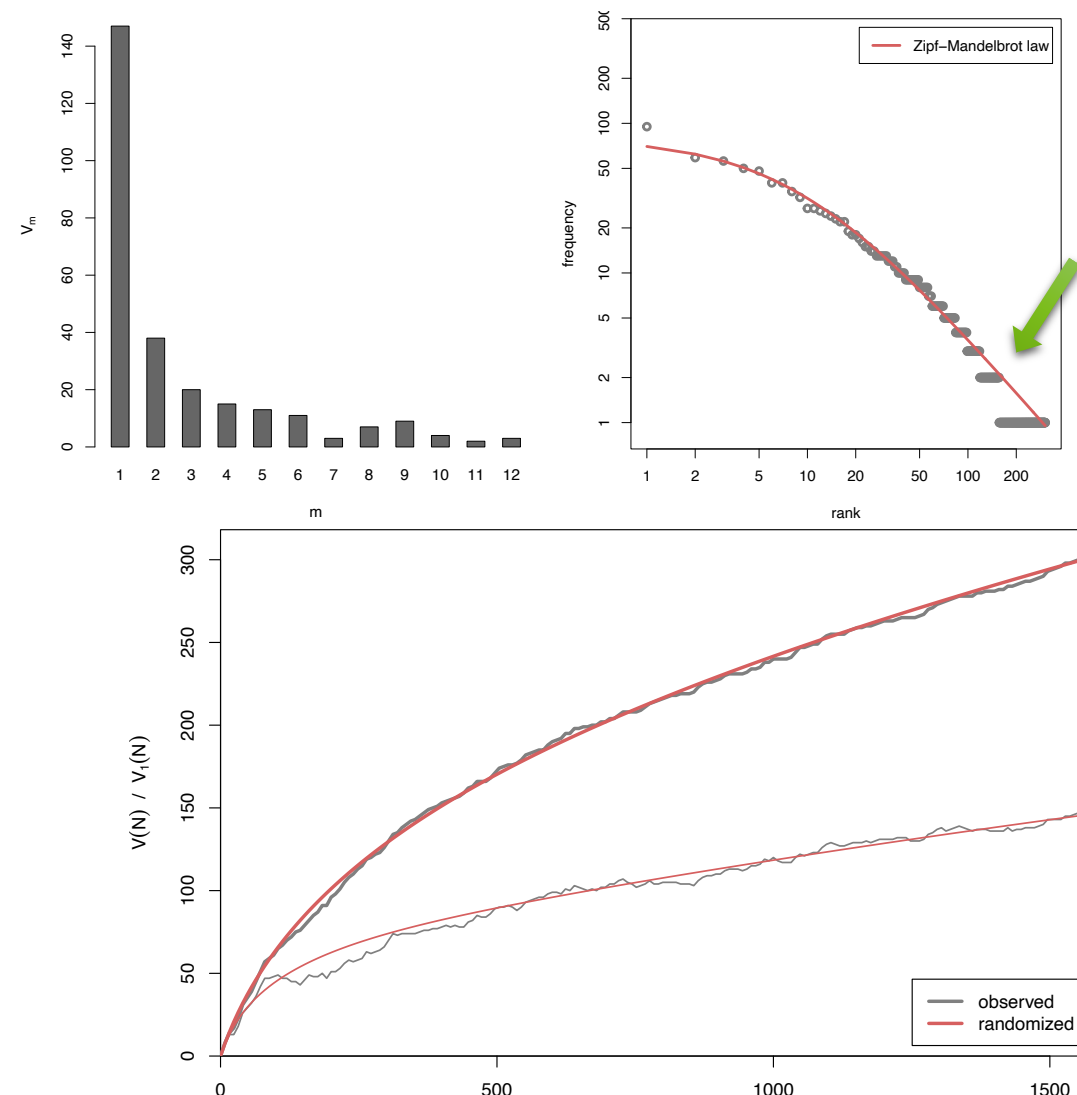
Indicators of productivity

(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)

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

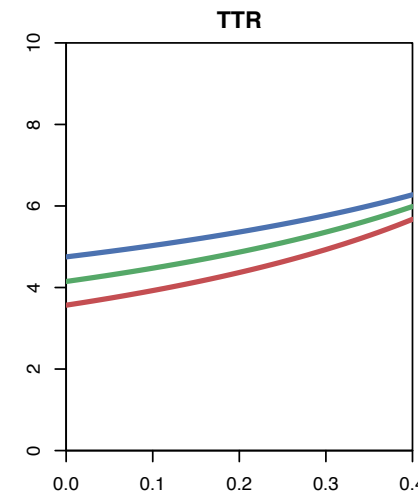
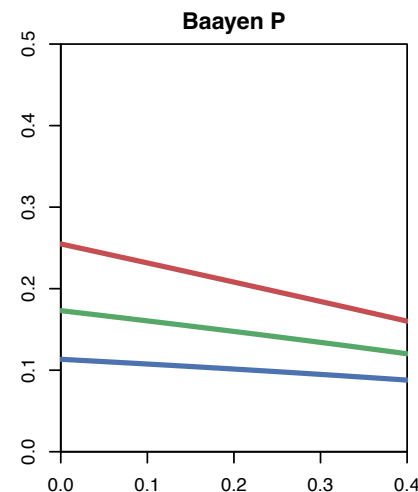
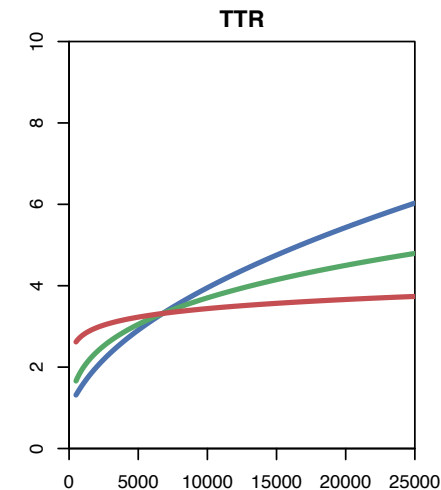
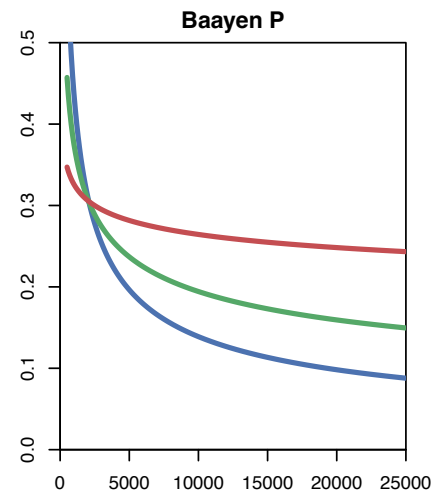


Productivity measures

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



- Baayen's productivity index \mathcal{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, MTLT)
 - average over fixed-length bins \rightarrow dependency on bin size
- still affected by probability mass of frequent types
 - but we want to focus on **productive range** of continuum!



Statistical LNRE models

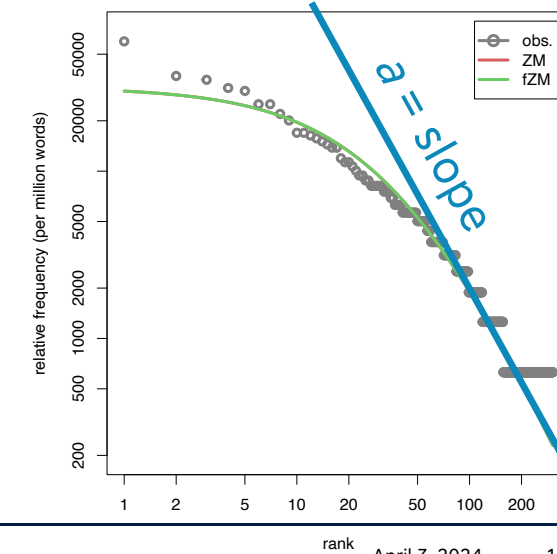
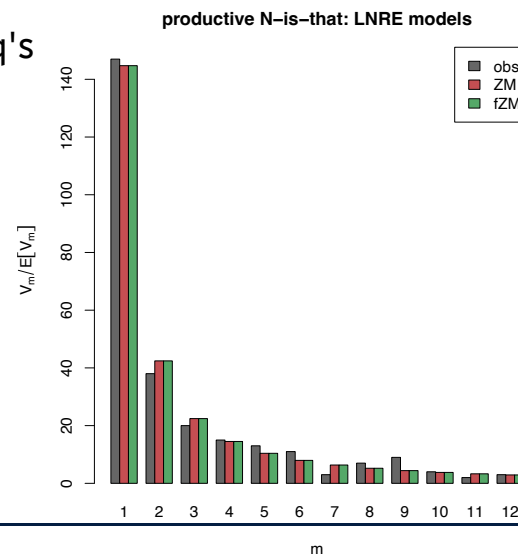
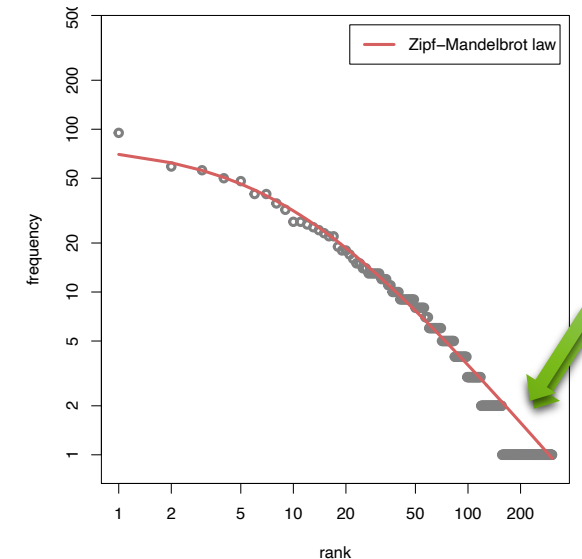
(Baayen 2001; Evert 2004; Evert & Baroni 2007)



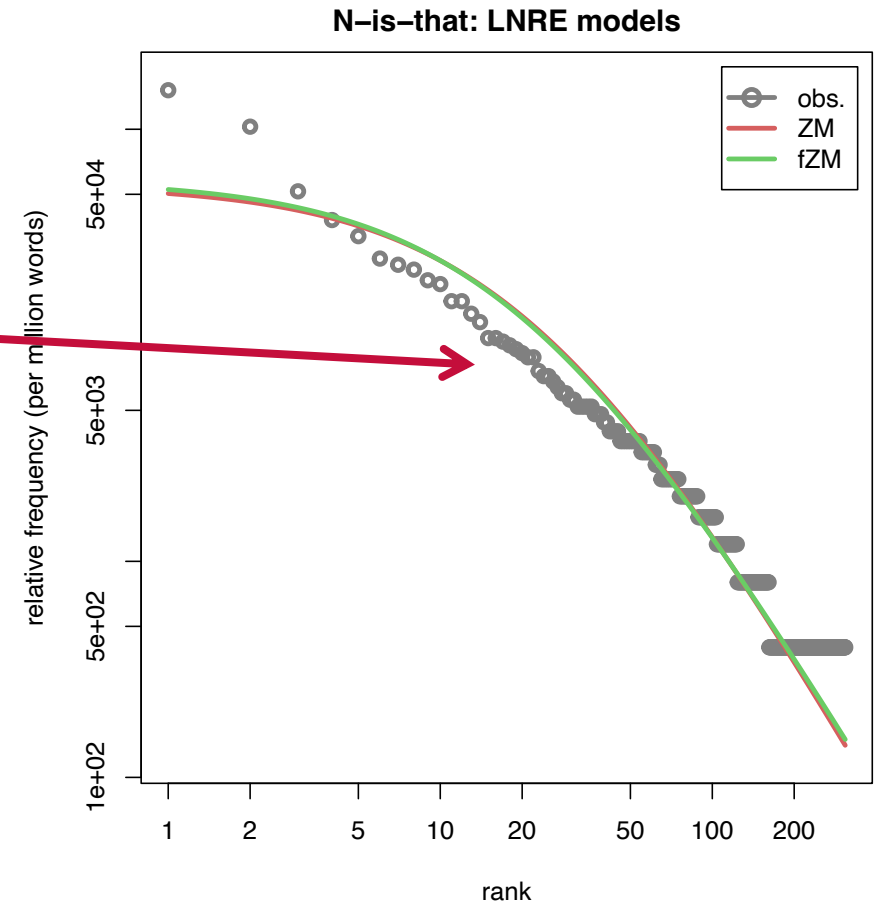
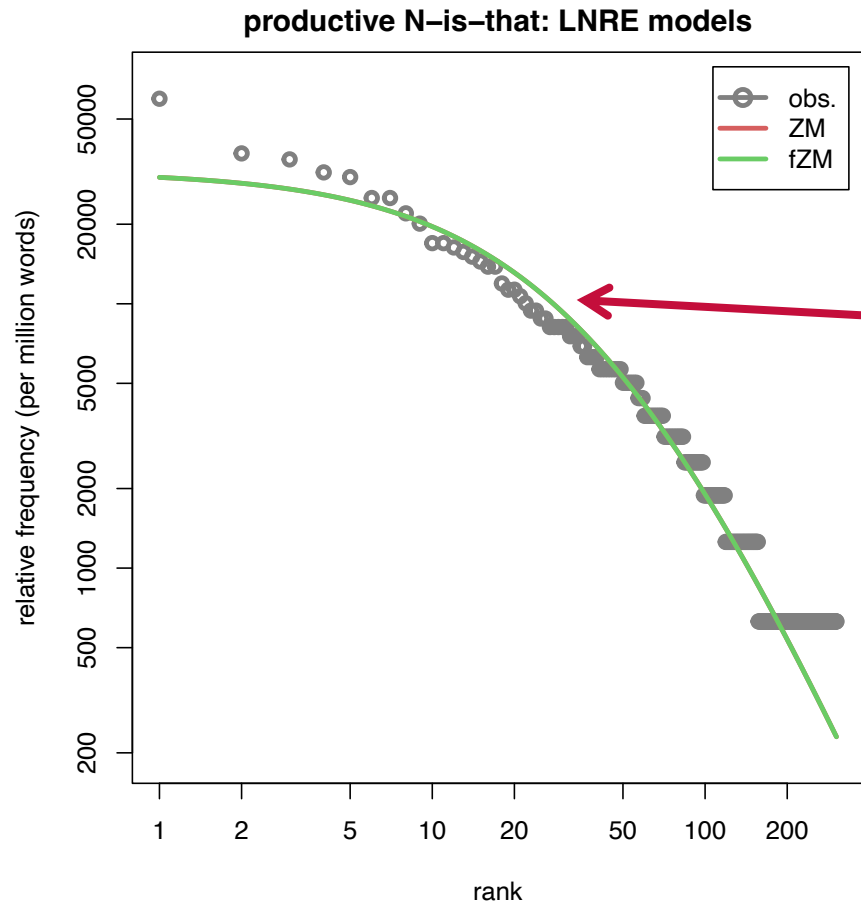
LNRE (*large number of rare events*) are statistical models for sampling from a distribution characterised by Zipf's law (or a similar type-token distribution)

- intuition: fit Zipf's law to observed data
 - but doesn't account very well for hapax legomena and dis legomena
- LNRE version: ZM and fZM models (Evert 2004)
 - fitted to frequency spectrum with focus on lowest frequency classes V_m
 - better hapax & dis legomena, less accurate for medium to high freq's
- quantitative measure: **Zipf slope** $a > 1$
 - or parameter $0 < \alpha < 1$ of type density function

$$\pi_i = \frac{C}{(i + b)^a} \iff g(\pi) = C \cdot \pi^{-\alpha-1}$$



Why we needed to exclude lexicalised types

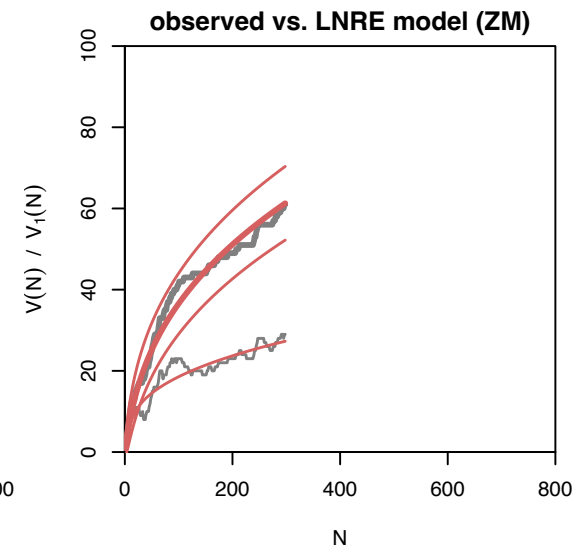
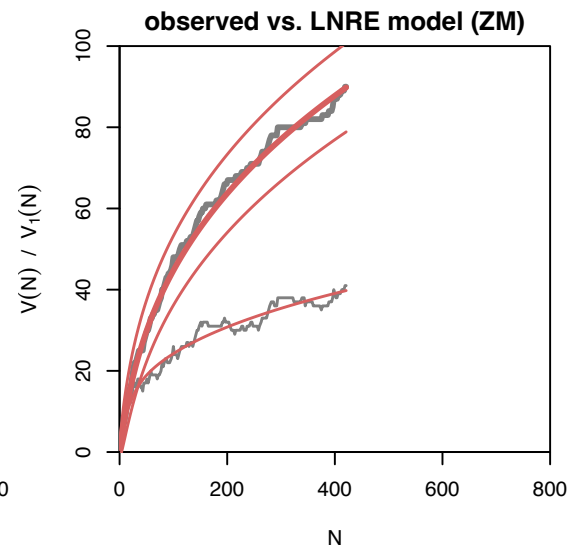
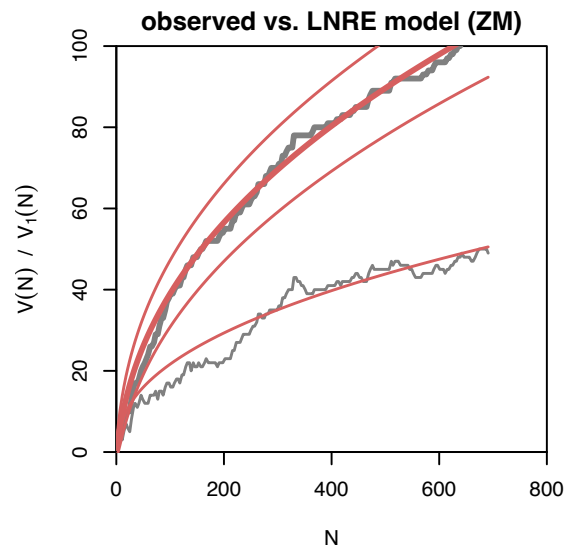


Comparison of shell noun classes

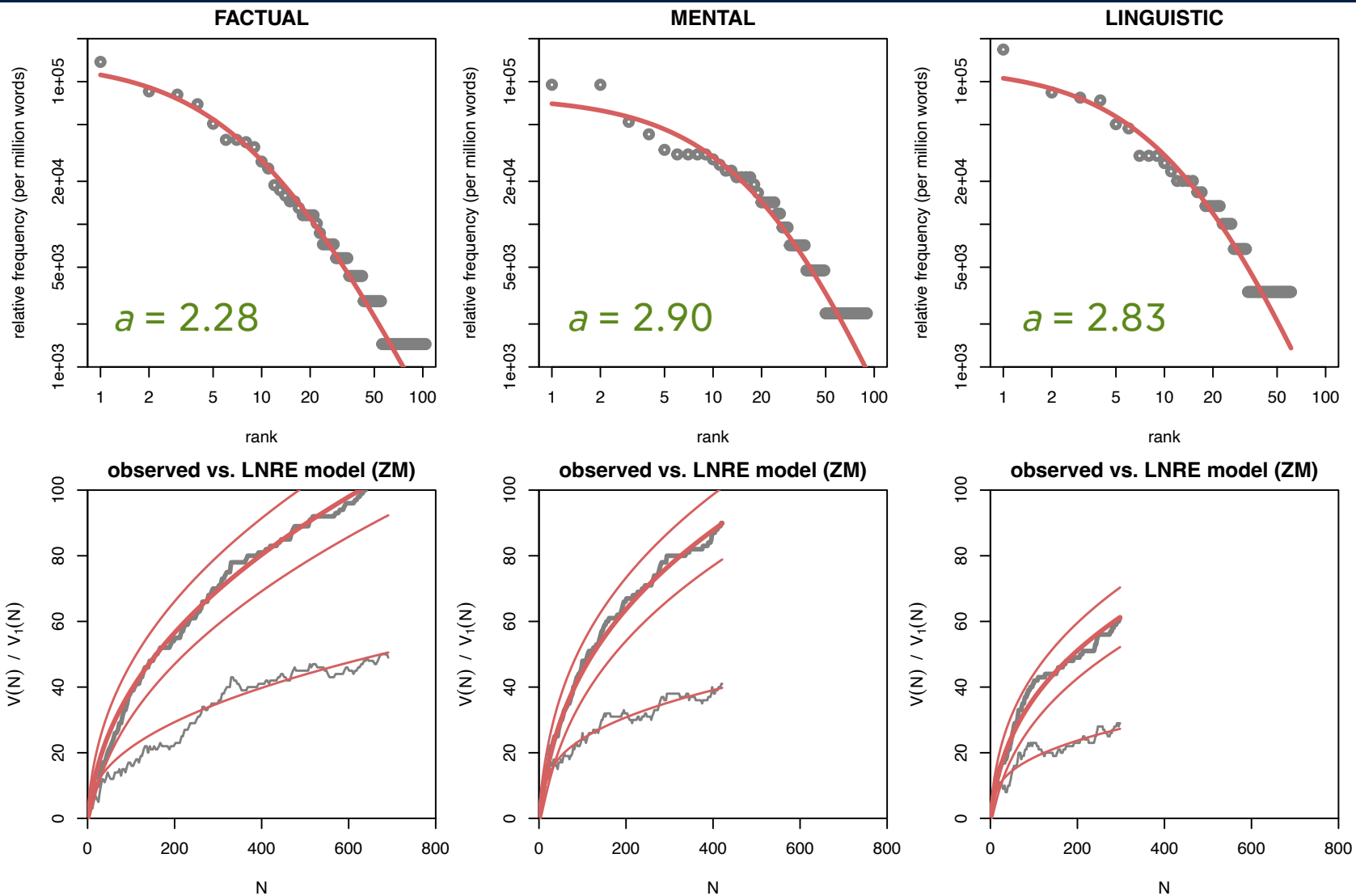
FACTUAL

MENTAL

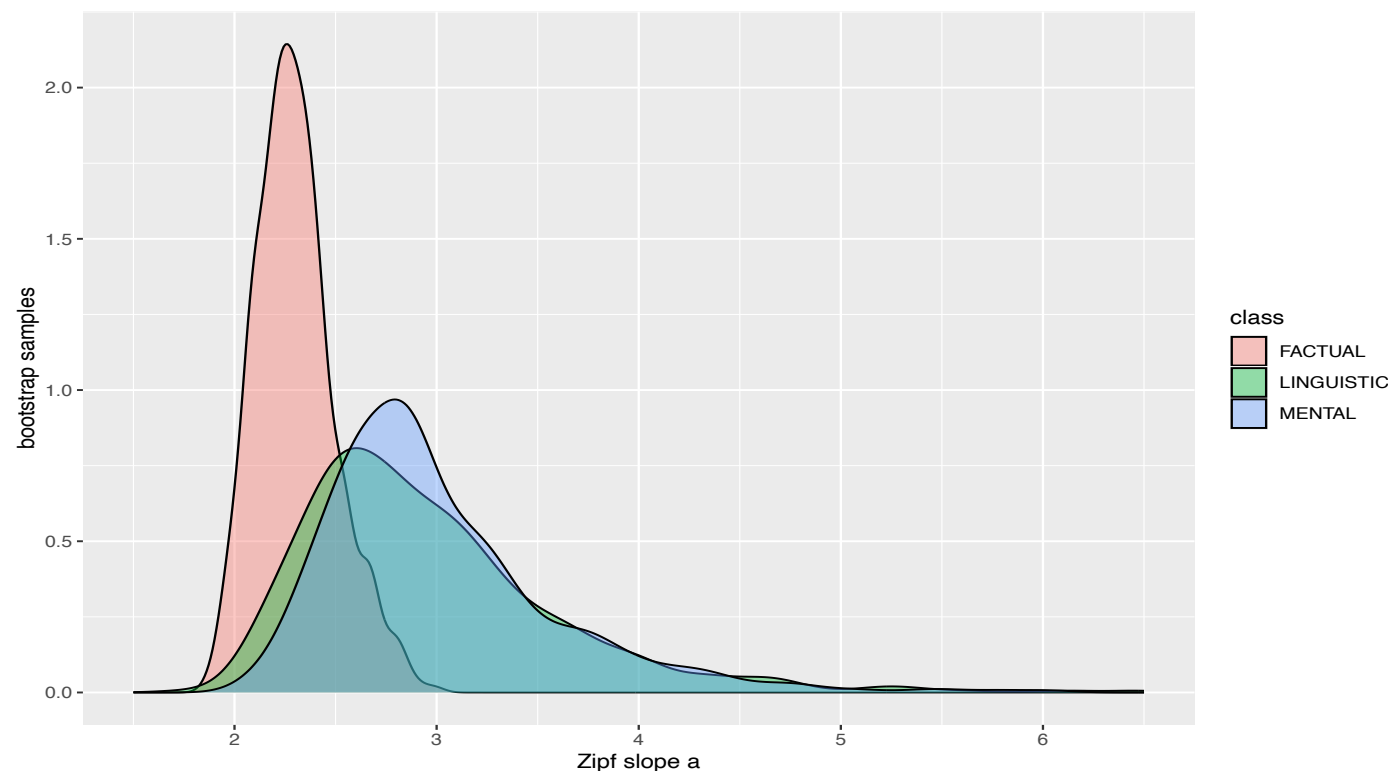
LINGUISTIC



Comparison of shell noun classes



- 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

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

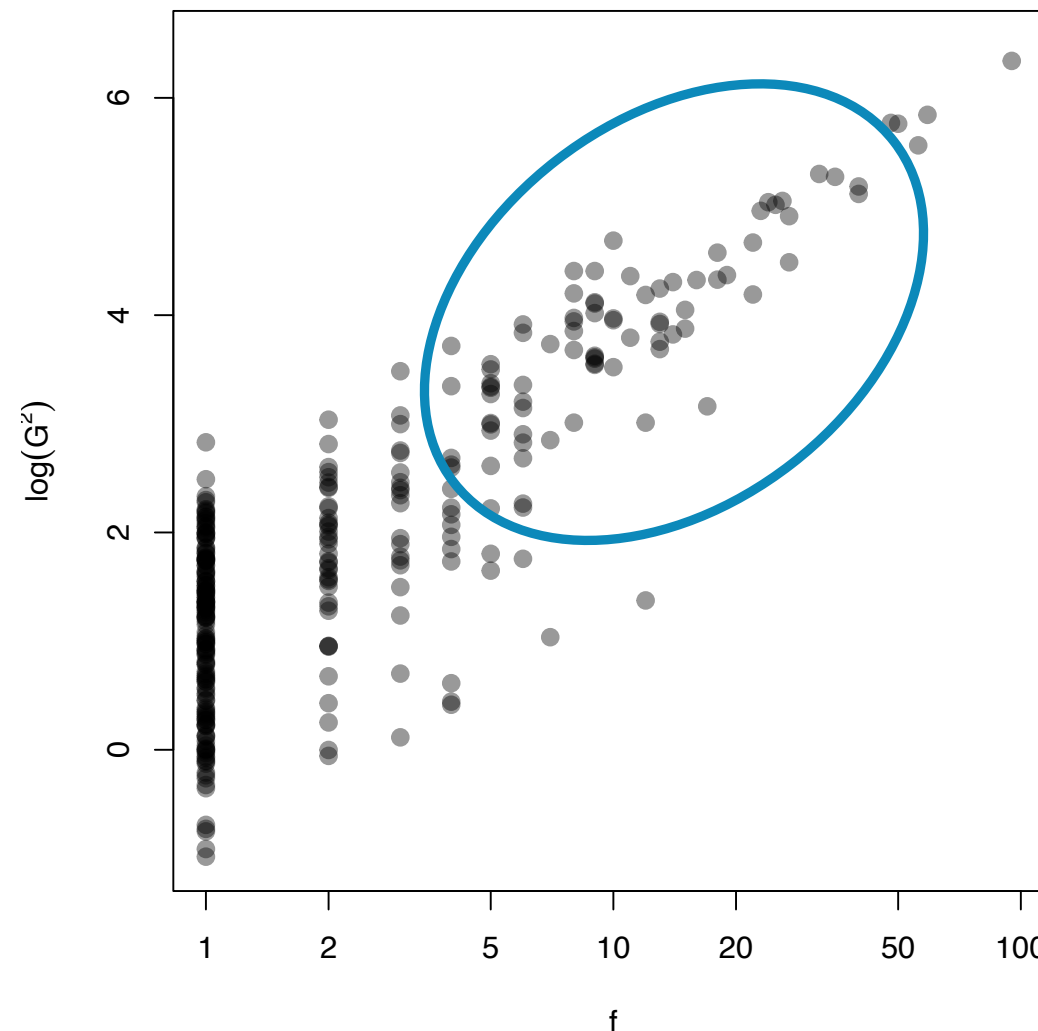
– **hypothesis**: collocational patterns, statistical association of fillers with Cx

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 measures grouped by mathematical properties (e.g. Evert 2008)
 - measures of **significance**: log-likelihood G^2 , Fisher's exact test (Stefanowitsch & Gries 2003)
 - measures of **effect size**: MI, LogRatio, odds ratio, ΔP , ...
 - **heuristic** coefficients: Dice, MI^2 , MI^3 , Simple Maths, ...
 - **conservative estimates**: LRC (Evert 2022)

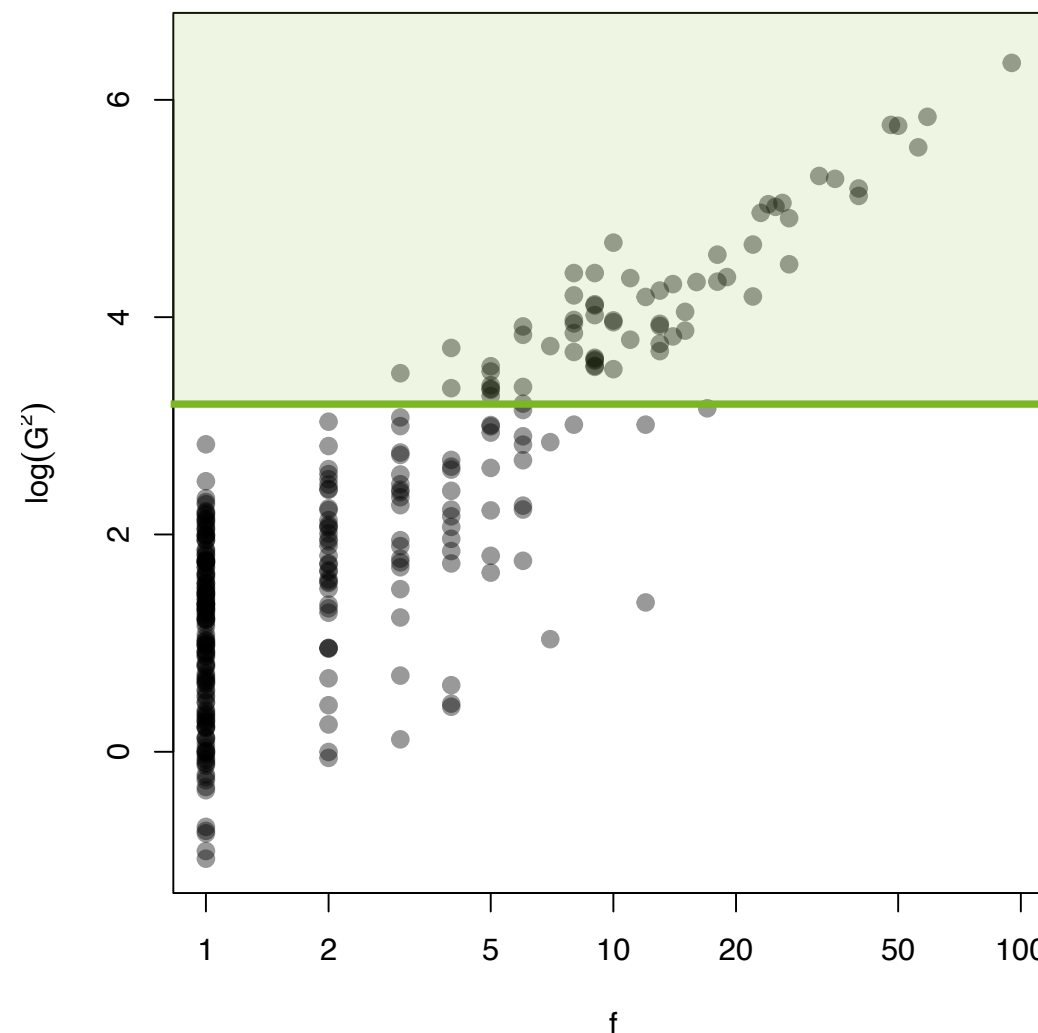
- recommended measure: **log-likelihood G^2**
- goal: differentiation between medium-frequency noun types

productive N-is-that: frequency vs. Log-Likelihood (LLR)



- 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

productive N-is-that: frequency vs. Log-Likelihood (LLR)



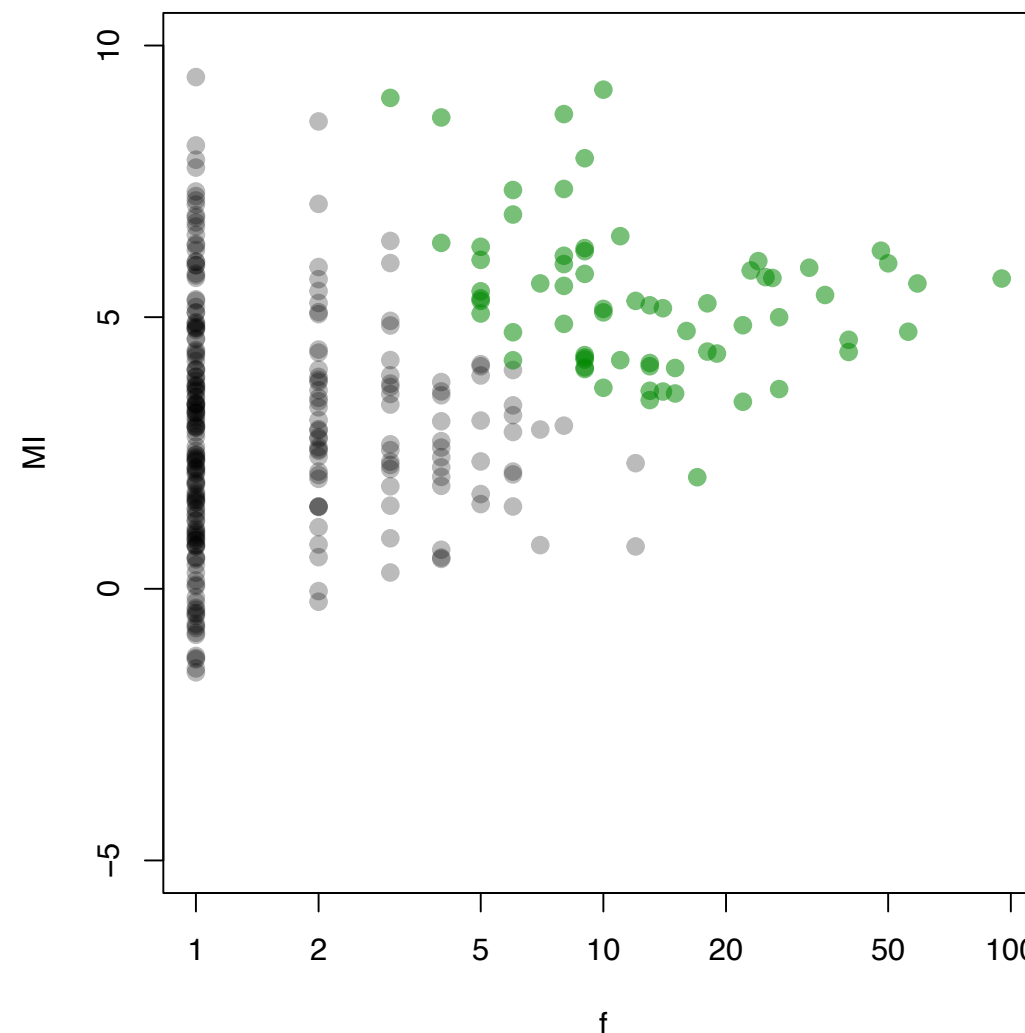
Measures of effect size

<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

frequency vs. Mutual Information

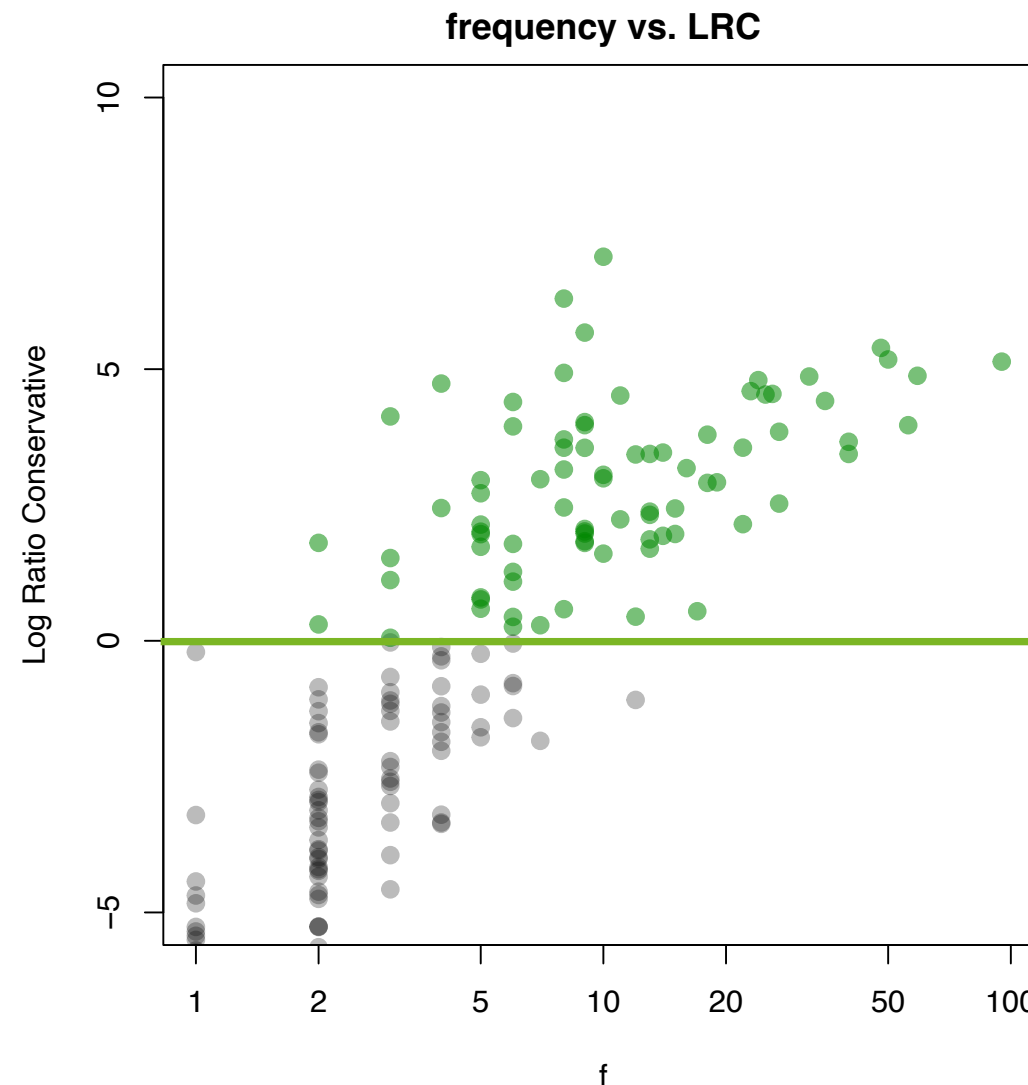


Conservative estimates

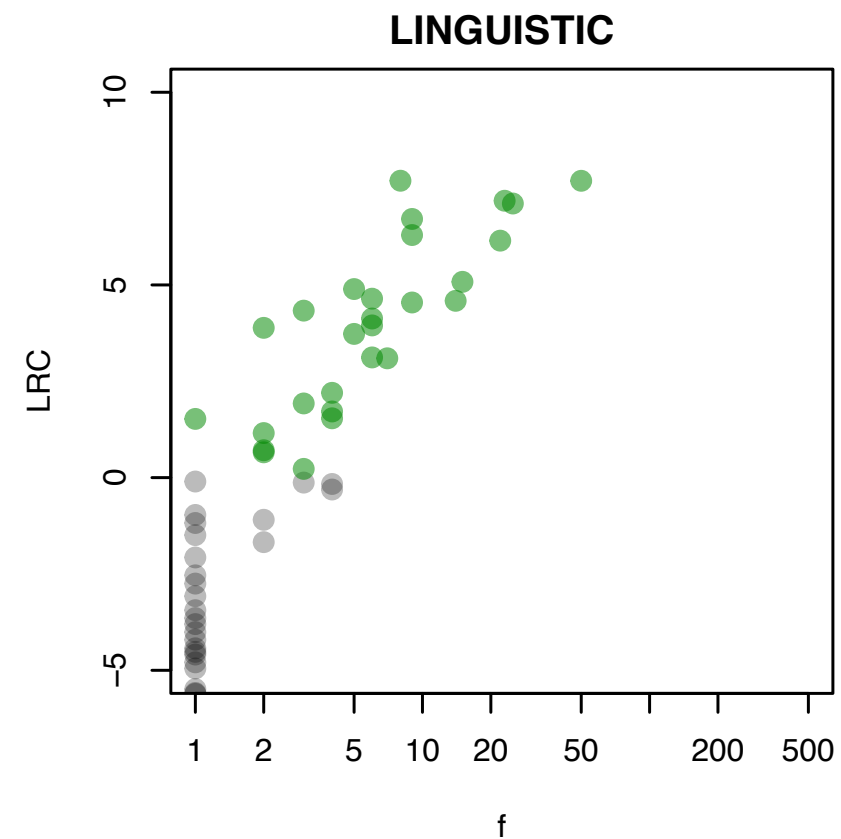
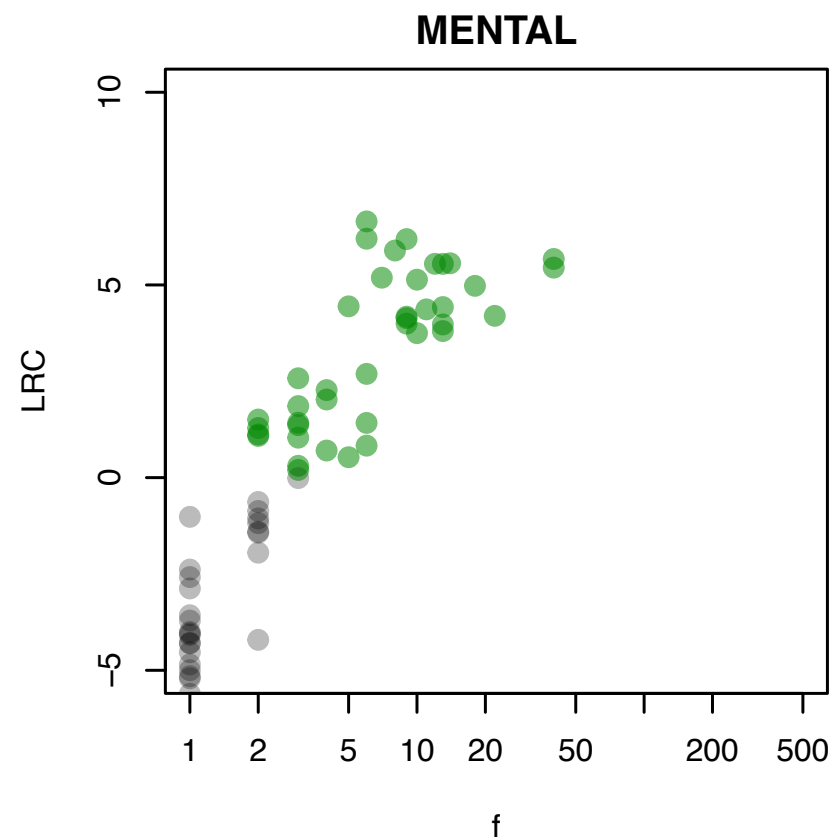
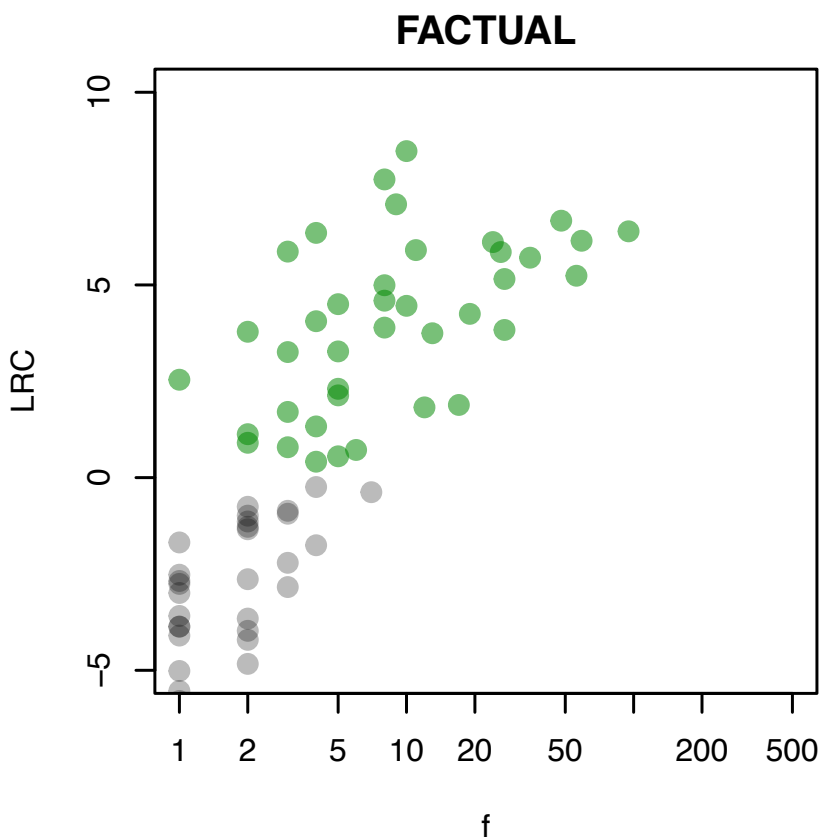
<http://www.collocations.de/AM/> | <https://osf.io/cy6mw/>



- conservative estimates combine both aspects:
Conservative LogRatio LRC (Evert 2022)
 - based on confidence interval for relative risk
 - recommended as first-line association and keyness measure
- LRC shows LogRatio supported by significant evidence
 - true LogRatio value will often be higher (but can't be confident)
 - significance cutoff $LRC \geq 0$
- 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



Thank you for listening!
Questions? Comments?

- 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 lexicogrammatical 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.

- Herbst, T. (2018). Is language a collocation? 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). Collocations: 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.