Keyword analysis: Progress through regression **(?)**

Lukas Sönning
University of Bamberg

lsoenning.github.io

Objectives

osf.io/**mc26t**

- Explore use of regression for keyness analysis
- Illustration: Key verbs in academic writing
- Advantages + disadvantages
- Conclusion
- Manuscript: psyarxiv.com/**25mwj**

Keyness

osf.io/**mc26t**

Keyness analysis

- Identification of items that are typical of a particular text variety
- Target corpus vs. reference corpus
- Keyness metrics: Ranking of items

Keyness as a multidimensional construct

- Most keyness metrics: Frequency comparisons (target vs. reference corpus)
- Different aspects of keyness: **Distinctiveness** vs. **generality** Egbert & Biber 2019, Gries 2021
- **Frequency**-oriented vs. **dispersion**-oriented approaches

Four dimensions

	Frequency-oriented	Dispersion-oriented
Target corpus in isolation	Discernibility • Occurrence rate (e.g. pmw)	Generality • Range [l] • *TD* [m] • D_{KL} [g] • D, S_{adj}, D_{2}, D_{P}, D_{A} [b]
Comparison to reference corpus	Distinctiveness • Rate ratio [a] • Rate difference [b] • *PS* [b] • Log ratio [c] • Difference coefficient [d] • %DIFF [e] • Odds ratio [f] • Signed D_{KL} [g] • Chi-squared test [d] • Likelihood-ratio test [h] • Wilcoxon test [j] • t-test [j] • BIC [k]	Comparative generality • *TD* ratio [m] • *TD* difference[b] • D_{KL} difference [g] • D, S_{adj}, D_{2}, D_{P}, D_{A} difference [b] • *TD*-based LR test [m]

Descriptive
Inferential

[a] Kilgarriff 2009
[b] Sönning 2022a
[c] Hardie 2014
[d] Hofland & Johansson 1982
[e] Gabrielatos & Marchi 2011
[f] Pojanapunya & Watson Todd 2016
[g] Gries 2021
[h] Dunning 1993
[j] Kilgarriff 1996
[k] Wilson 2013
[l] Rayson 2003
[m] Egbert & Biber 2019

Count regression models

- Data type: Non-negative integers
- Large family e.g. Cameron & Trivedi 2013
- Basic version: Poisson regression
- Count regression underused in linguistics Winter & Bürkner 2021

- We will use: **Negative binomial regression**

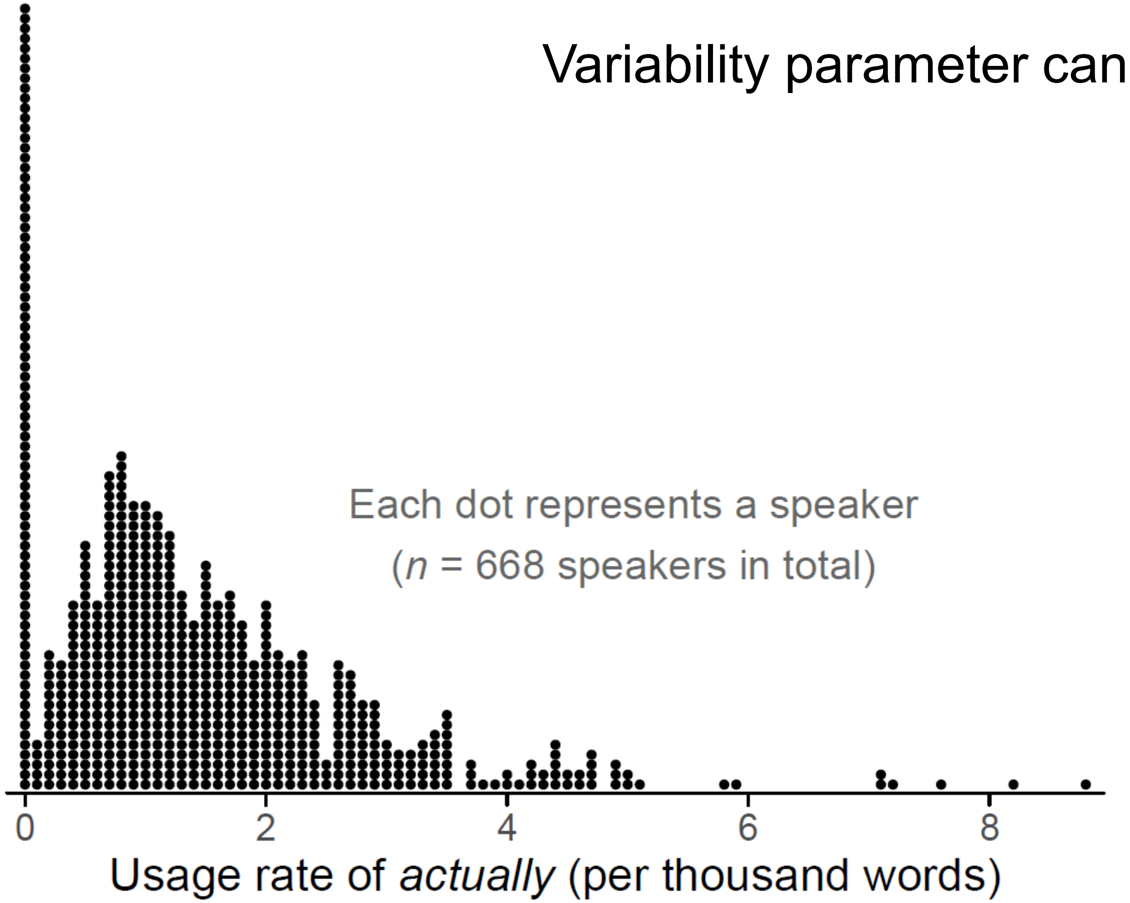
Keyness regression

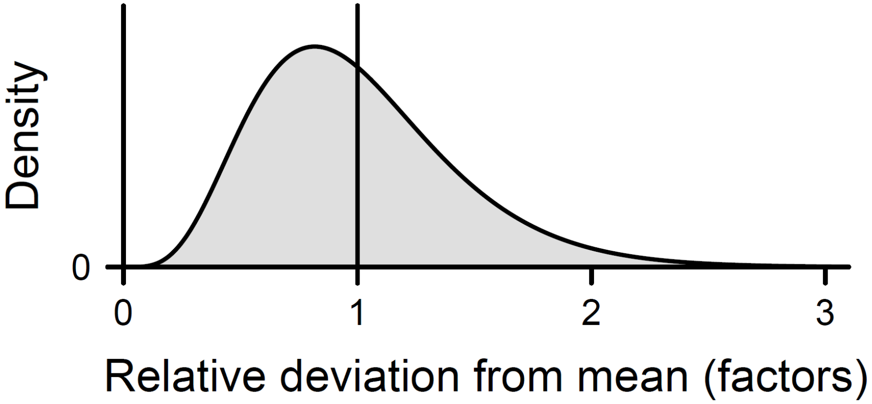
Negative binomial regression

- Poisson too restrictive: Assumes probability of *actually* is constant across speakers/situations
- Negative binomial: Allows for variation among speakers/texts

  Additional parameter describing variability

  Variability parameter can be converted into a dispersion measure (D_{NB})





Data: SpokenBNC2014 Love et al. 2017

Keyness regression

Keyness metrics

	Frequency-oriented	Dispersion-oriented
Target corpus in isolation	Discernibility **Occurrence rate** (normalized frequency) **+ 95% CI**	Generality **Dispersion D_{NB}** **+ 95% CI**
Comparison to reference corpus	Distinctiveness **Rate ratio** **+ 95% CI**	Comparative generality **Difference in dispersion D_{NB}** **+ 95% CI**

Case study

Key verbs in published academic writing

- COCA Davies 2008-
  - Year 2019 only
  - Target corpus: ACAD
  - Reference corpus: NEWS

- Focus on **578 verb lemmas** (data: Sönning 2022b)
  - Higher normalized frequency in ACAD
  - Normalized frequency > 10 pmw in ACAD

- **Ranking** Discernibility

  Distinctiveness

  Generality

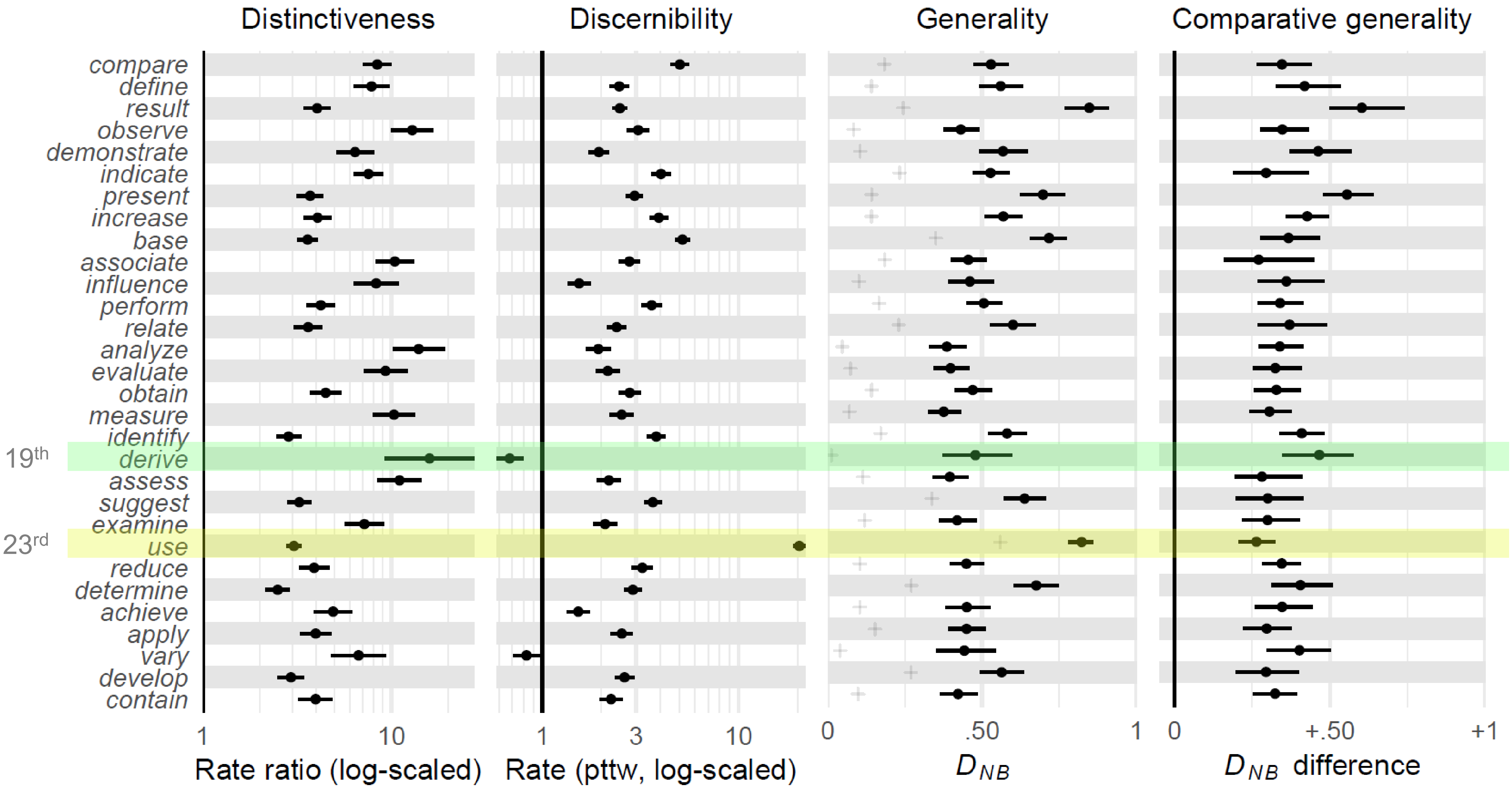
  Comparative generality

  Equal weight

Keyness regression

Error bars: 95% CIs

osf.io/**mc26t**



Keyness regression

Advantages

- Text-level analysis e.g. Baroni & Evert 2009; Lijffijt et al. 2014
- Confidence intervals for all keyness metrics cf. Gries 2022
- Descriptive and inferential information
- Interpretable metrics cf. Sönning 2022a

Keyness regression

Disadvantages

- Specialized statistical software needed
  - R package gamlss Rigby & Stasinopoulos 2005
- Data-hungry
- Computationally expensive

Conclusion

osf.io/**mc26t**

Keyness regression (?)

- Unsuited as a routine tool
- Implementation in corpus analysis software not possible

- Generally: Use of simpler metrics as screening devices

- Feasibility depends on analysis task
    - Unattractive for supplementary keyness analyses
    - Attractive for first-rate keyness analyses (e.g. for producing vocabulary lists)

Thanks for listening.

References

Baroni, Marco & Stefan Evert. 2009. Statistical methods for corpus exploitation. In Anke Lüdeling & Merjy Kytö (eds.), *Corpus linguistics: An international handbook*. Berlin: Mouton de Gruyter, 777–803.

Cameron, A. Colin & Pravin K. Trivedi. 2013. *Regression analysis of count data*. Cambridge: CUP.

Davies, Mark. 2008-. *The Corpus of Contemporary American English*. www.english-corpora.org/coca.

Dunning, Ted. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics* 19(1), 61–74. doi:10.5555/972450.972454

Egbert, Jesse & Douglas Biber. 2019. Incorporating text dispersion into keyword analysis. *Corpora* 14(1), 77–104. doi:10.3366/cor.2019.0162

Gabrielatos, Costas & Anna Marchi. 2011. Keyness: Matching metrics to definitions. *Corpus Linguistics in the South* 1, University of Portsmouth, 5 November 2011. Available online at: http://eprints.lancs.ac.uk/51449

Gries, Stefan Th. 2021. A new approach to (key) keywords analysis: Using frequency, and now also dispersion. *Research in Corpus Linguistics* 9(2): 1–33. doi:10.32714/ricl.09.02.02

Gries, Stefan Th. 2022. Toward more careful corpus statistics: uncertainty estimates for frequencies, dispersions, association measures, and more. *Research Methods in Applied Linguistics* 1(1), 100002. doi:10.1016/j.rmal.2021.100002

Hardie, Andrew. 2014. Log ratio – An informal introduction. Post on the website of the ESRC Centre for Corpus Approaches to Social Science CASS. Available online at: http://cass.lancs.ac.uk/?p=1133

Hofland, Knut & Stig Johansson. 1982. *Word frequencies in British and American English*. London: Longman.

Kilgarriff, Aadam. 1996. Which words are particularly characteristic of a text? A survey of statistical approaches. In L. J. Evett & T. G. Rose (eds.) *Language Engineering for Document Analysis and Recognition* (*LEDAR*). AISB96 Workshop proceedings, Brighton, England. Faculty of Engineering and Computing, Nottingham Trent University, UK, 33–40.

Kilgarriff, Adam. 2009. Simple maths for keywords. In M. Mahlberg, V. González-Díaz & C. Smith (eds.) *Proceedings of the Corpus Linguistics Conference*, *CL2009*. Liverpool: University of Liverpool. Available online at: http://ucrel.lancs.ac.uk/publications/CL2009/171_FullPaper.doc

Lijffijt, Jefrey, Terttu Nevalainen, Tanja Säily, Panagiotis Papapetrou, Kai Puolamaki & Heikki Mannila. 2014. Significance testing of word frequencies in corpora. *Digital Scholarship in the Humanities*. doi:10.1093/llc/fqu064

Pojanapunya, Punjaporn & Richard Watson Todd. 2018. Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. *Corpus Linguistics and Linguistic Theory* 14(1), 133–167. doi:10.1515/cllt-2015-0030.

Rayson, Paul. 2003. *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison*. PhD thesis, Lancaster University.

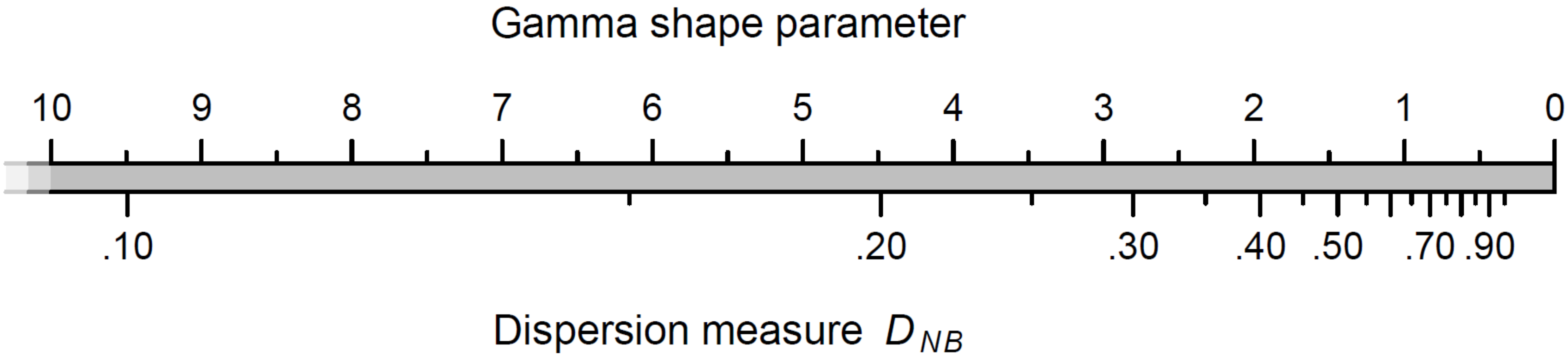
Rigby, Robert A. & Mikis D. Stasinopoulos. 2005. Generalized additive models for location, scale and shape (with discussion), *Applied Statistics* 54(3): 507–554.

Sönning, Lukas. 2022a. Evaluation of keyness metrics: Reliability and interpretability. *PsyArXiv preprint*. https://psyarxiv.com/eb2n9/

Sönning, Lukas. 2022b. Key verbs in academic writing: Dataset for "Evaluation of keyness metrics: Reliability and interpretability", https://doi.org/10.18710/EUXSMW, DataverseNO, DRAFT VERSION

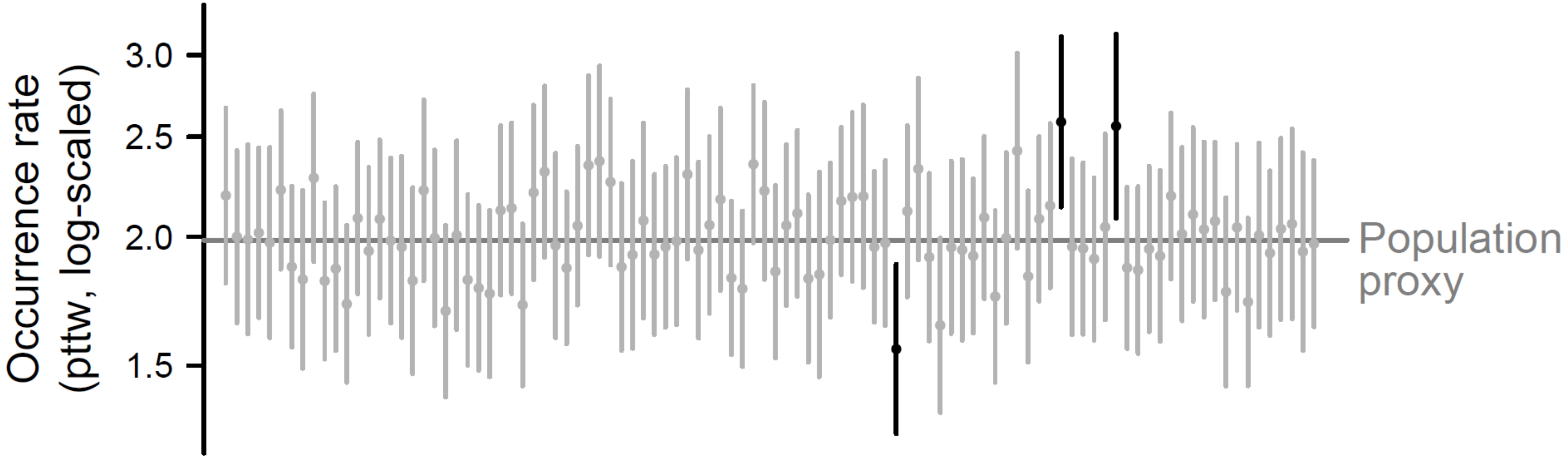
Wilson, Andrew. 2013. Embracing Bayes Factors for key item analysis in corpus linguistics. In Markus Bieswanger & Amei Koll-Stobbe (eds.), *New approaches to the study of linguistic variability*, 3-11. Frankfurt: Peter Lang.

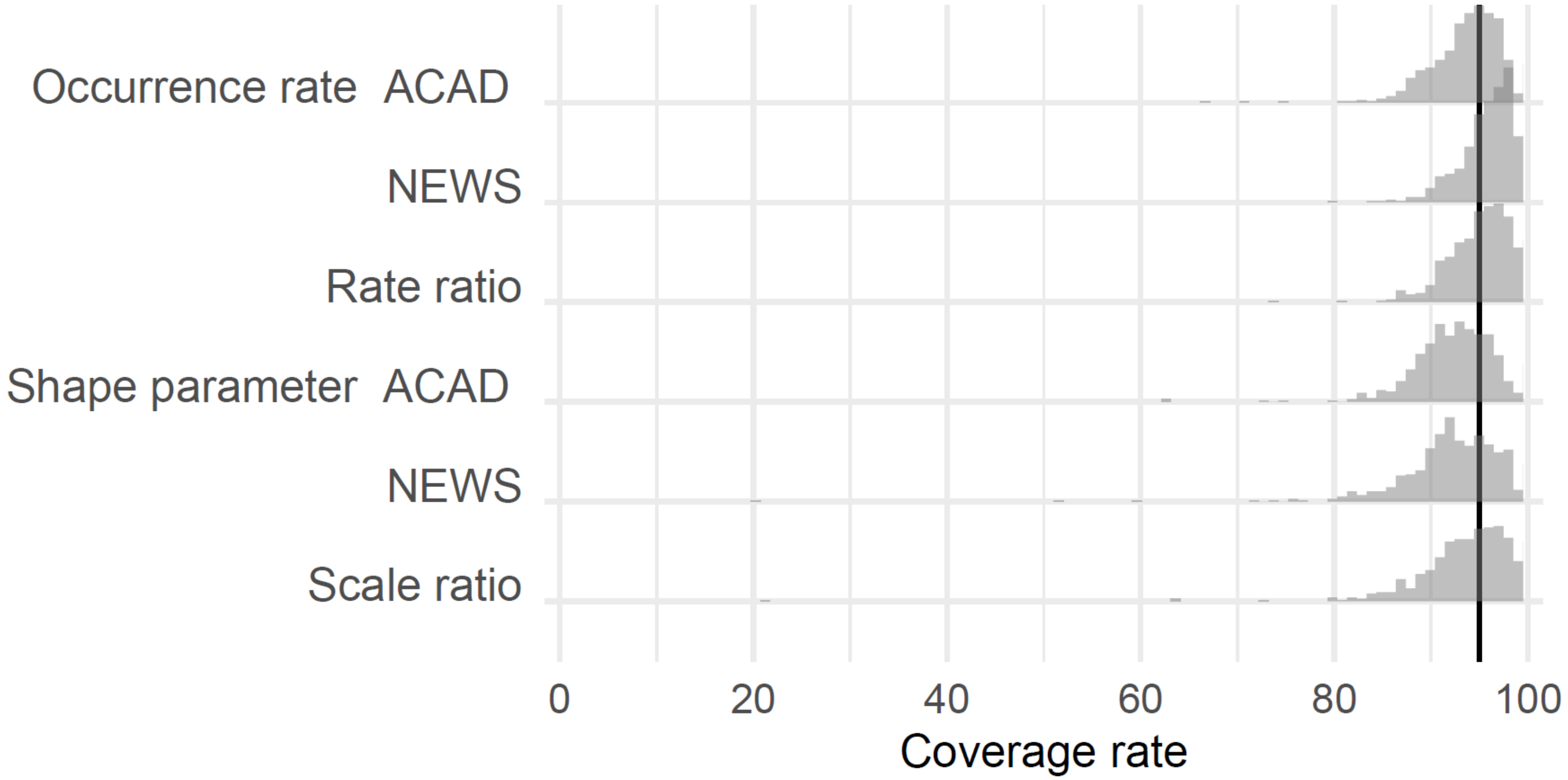
Winter, Bodo & Paul-Christian Bürkner. 2021. Poisson regression for linguists: A tutorial introduction to modelling count data with brms. *Language and Linguistics Compass*, e12439. https://doi.org/10.1111/lnc3.12439

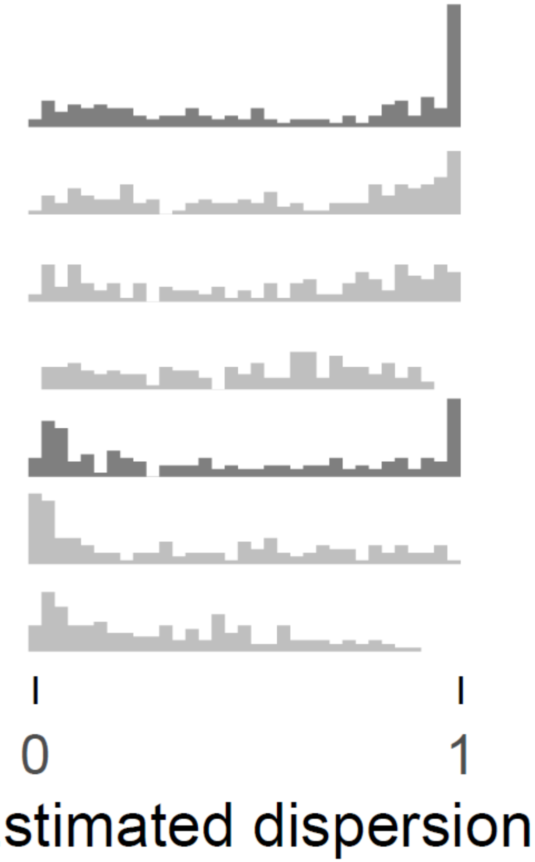

Gamma shape parameter
10
9
8
7
6
5
4
3
2
1
0
.10
.20
.30
.40
.50
.70
.90
Dispersion measure D_{NB}

Estimates (+ 95% CIs) from 100 different data subsets

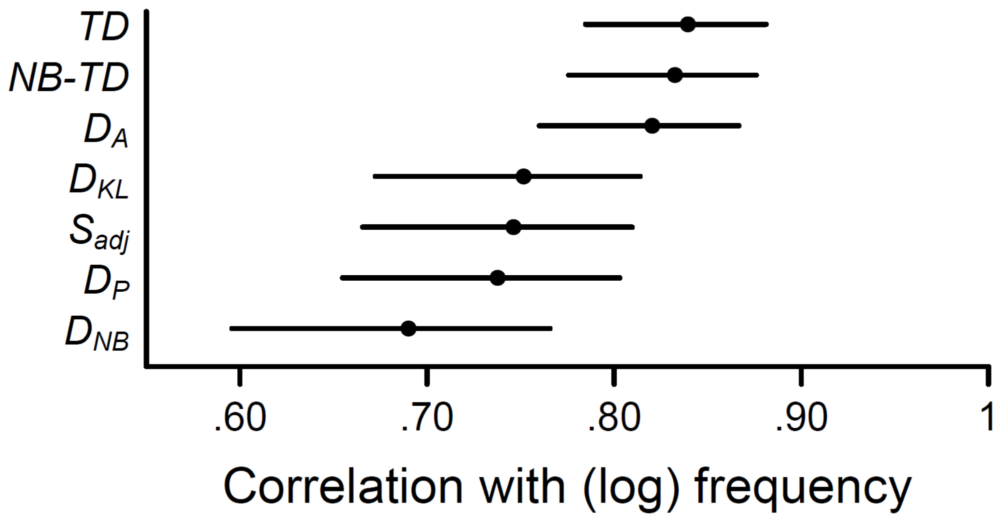
Verb: DEFINE

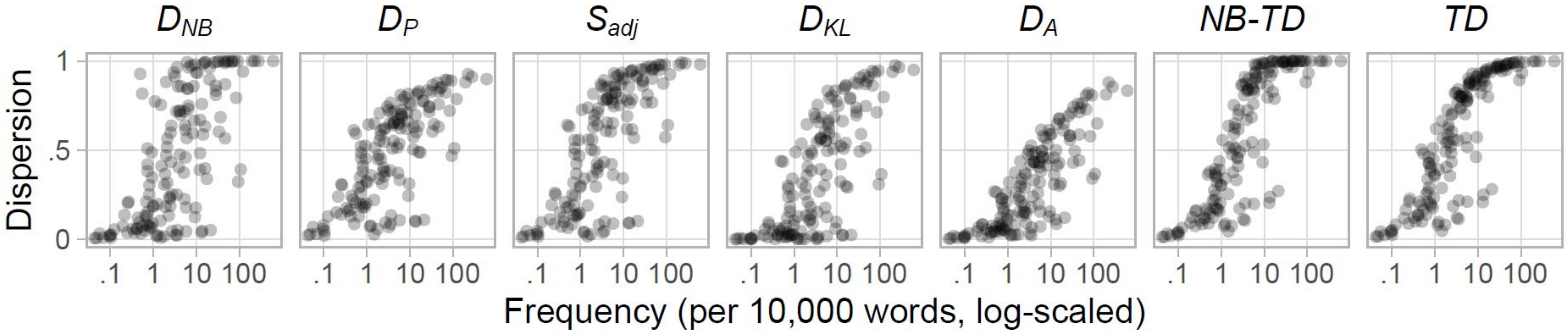


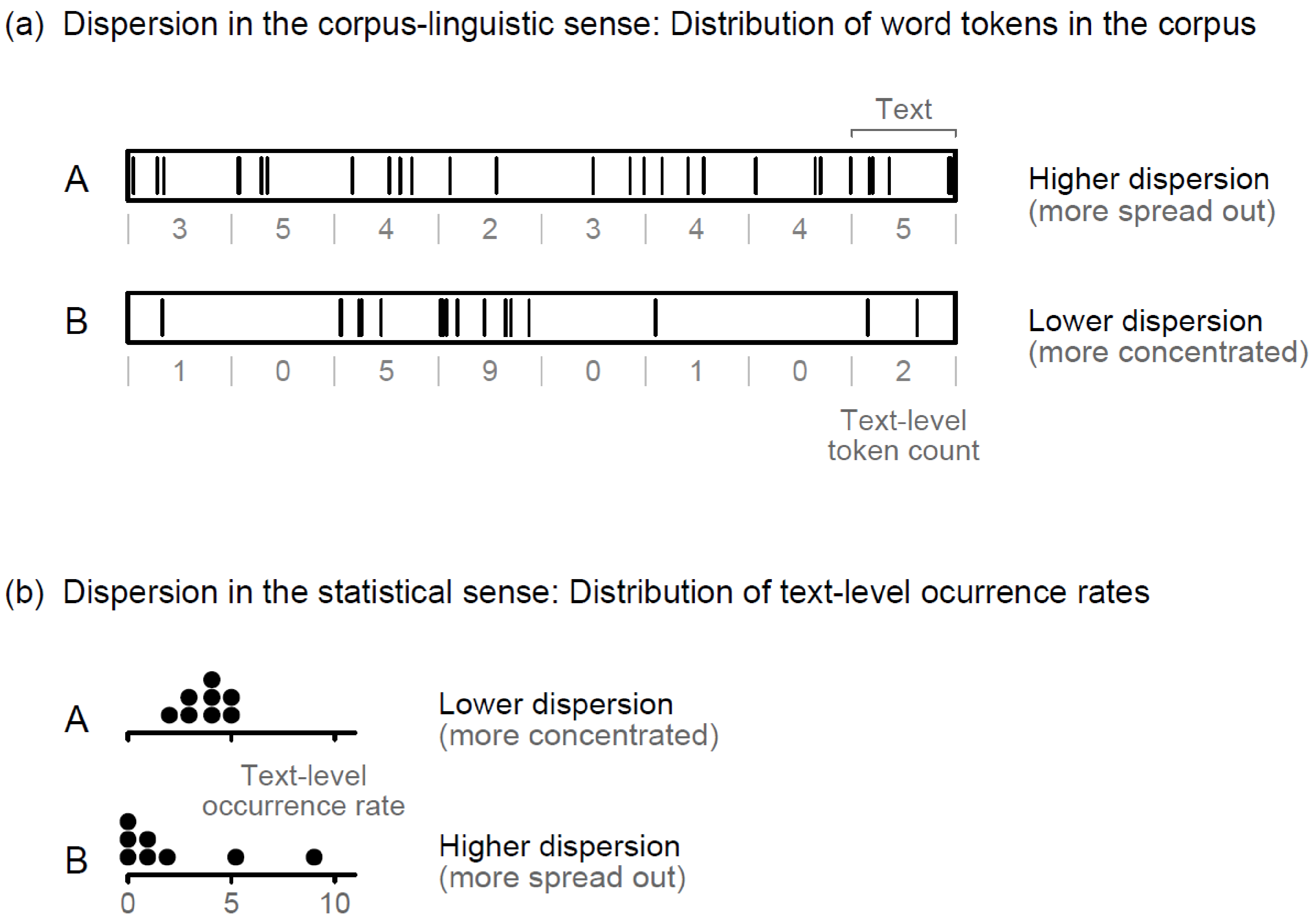

Occurrence rate ACAD
NEWS
Rate ratio
Shape parameter ACAD
NEWS
Scale ratio
0
20
40
60
80
100
Coverage rate



	Mdn	*IQR*	*Range*
NB-TD	.62	.73	[.01;1.0]
TD	.62	.65	[.02;1.0]
S_{adj}	.62	.66	[.01;.99]
D_P	.52	.47	[.02;.93]
D_{NB}	.43	.72	[.01;1.0]
D_{KL}	.33	.61	[.00;.98]
D_A	.29	.41	[.00;.88]


TD
NB-TD
D_A
D_{KL}
S_{adj}
D_P
D_{NB}
.60
.70
.80
.90
1
Correlation with (log) frequency


D_{NB}
D_P
S_{adj}
D_{KL}
D_A
NB-TD
TD
Dispersion
1
.5
0
.1
1
10
100
Frequency (per 10,000 words, log-scaled)



(a) Dispersion in the corpus-linguistic sense: Distribution of word tokens in the corpus

(b) Dispersion in the statistical sense: Distribution of text-level ocurrence rates