Means of Productivity

On the Statistical Modelling of the Restrictedness of Lexico-Grammatical Patterns

Sascha Diwersy[1], **Stefan Evert**[2], **Philipp Heinrich**[2], and **Thomas Proisl**[2]

[2] *Praxiling*, **Université Paul-Valéry** (Montpellier 3)
[2] *Chair of CCL*, **Friedrich-Alexander-Universität Erlangen-Nürnberg**

Aims of the talk

- tackle the question of fixed vs. free combinatorics from a predominantly distributional point of view
- need for viable (lexico-statistical) methodology
- starting point: work on morphological and syntactic productivity and adequate measures

Lexique-Grammaire

- Gross (1996); Mejri (1998): work on fixedness (French *figement*)
- several criteria:
  - restrictions of syntactic transformations
  - restrictions on syntactic extensions (insertion of modifiers)
  - restrictions on the use of determiners
  - restrictions on paradigmatic commutation
- degrees of fixedness: continuum reaching from totally fixed to more or less free expressions (see amongst others Gross (1996: 16–17); Le Pesant (2003: 106))

Cognitive Linguistics

- syntax-lexicon continuum (Croft and Cruse, 2004: 255) ranging from
  - atomic and substantive units (e.g. monomorphemic words) to
  - complex and schematic units (e.g. syntactic patterns)

Construction type	Traditional name	Examples
Complex and (mostly) schematic	**syntax**	[SBJ *be*- TNS VERB *-en by* OBL]
Complex, substantive verb	**subcategorization frame**	[SBJ *consume* OBJ]
Complex and (mostly) substantive	**idiom**	[*kick*-TNS *the bucket*]
Complex but bound	**morphology**	[NOUN-*s*], [VERB-TNS]
Atomic and schematic	**syntactic category**	[DEM], [ADJ]
Atomic and substantive	**word/lexicon**	[*this*], [*green*]

- our focus is on complex units with different degrees of schematicity or substantiveness (depending on the perspective one might take)

Syntactic Productivity

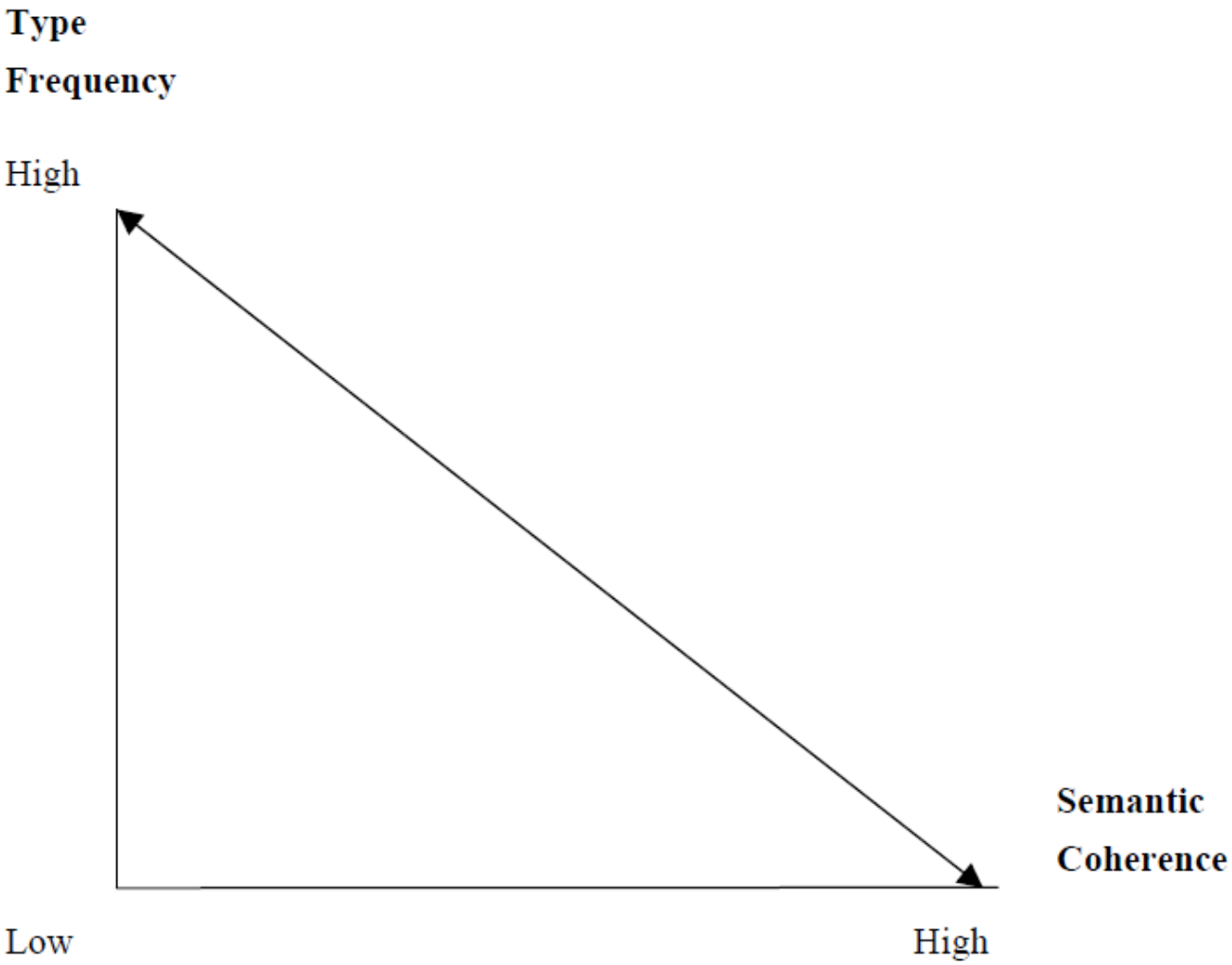
- Baroda (2008): productivity cline ranging from schematicity to specificity

- inverse correlation of type frequency and semantic coherence:
  - schematicity:
    high type frequency + low semantic coherence
  - specificity:
    low type frequency + high degree of semantic coherence

- full productivity by schema extension vs. productivity by analogic extension

Syntactic Productivity

type frequency and semantic coherence (Barðdal, 2008)



Productivity of the N+*be*+*that* pattern

- the use of so called "shell nouns" (Schmid, 2000) as subject of copula clauses involving the linking verb BE and a THAT-clause functioning as subject complement
- shell nouns serve specific semantic, cognitive and textual functions (Schmid, 2000: 14):
  - semantic: characterizing and perspectivizing complex chunks of information expressed in textual segments of various length
  - cognitive: encapsulation of complex chunks of information in temporary nominal concepts with apparently rigid and clear-cut conceptual boundaries
  - textual: linking these nominal concepts with clauses or other pieces of text which contain the actual details of information

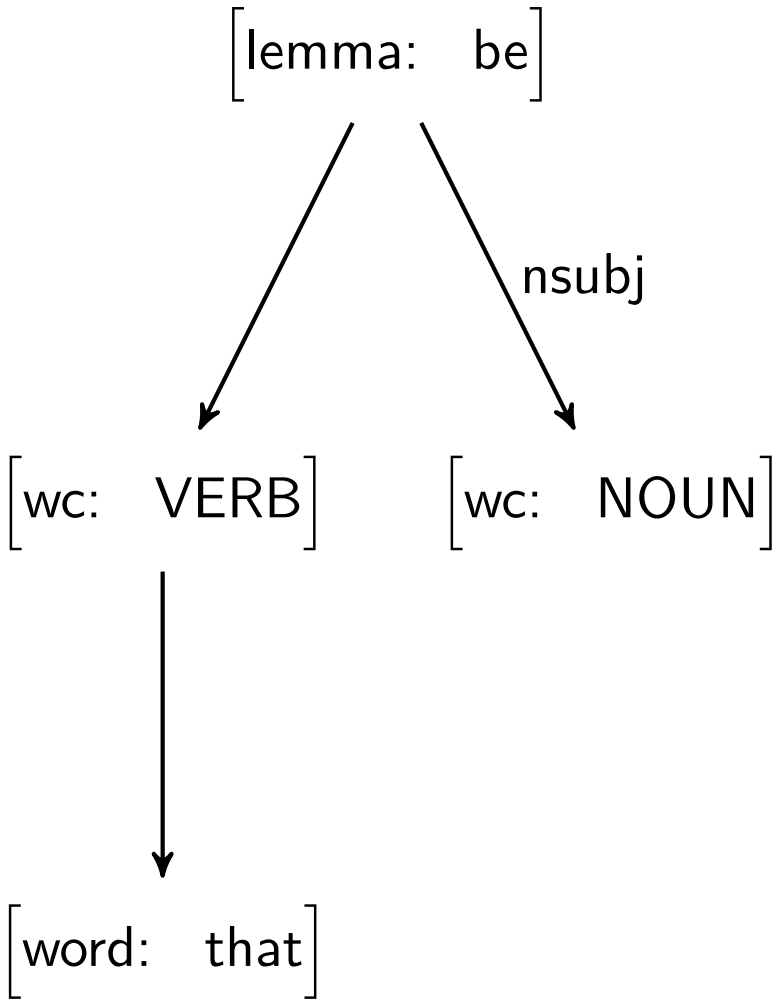
Productivity of the N+*be*+*that* pattern

examples of shell noun uses

- Our main concern as a group is that we do not waste the money. [BNC AT1: 2091]
- The problem here is that having so easy access and the largest concentration of easy routes, it is very crowded at holiday time. [BNC A15: 876]
- But the fact is that the very lack of evidence seems to fan the flames of suspicion. [BNC CB8: 298]
- The point is, that for the first time in decades, the environmentalists have a powerful voice – and a Government which claims to listen. [BNC AAG: 68]

Data extraction

- Treebank.info (Proisl and Uhrig, 2012): `http://treebank.info`
- British National Corpus XML Edition
  - original tokenization
  - Stanford Parser 1.6.9
  - Erlangen lemmatizer



Data extraction

- number of matches extracted via treebank.info: 32,907
- random sample of 10%
- manual validation of 3,290 matches
- elimination of 765 occurrences
  - parsing errors
  - sentence duplicates in corpus
- final sample size: 2,525 items

Data set

- data table with 2,525 instances of the construction:
  - noun lemma (`N_lemma`)
  - realization of copula (`to_be`)
  - verb of embedded clause (`that_V`)
  - pre/postmodification of noun (`PreMod`, `PostMod`, `hasPreMod`, `hasPostMod`)
  - as well as BNC text/sentence ID (`BNCTextID`, `SentenceID`) and BNC metadata for the respective text

- the full sentence (`Sentence`) is included with noun, copula and embedded verb marked

Semantic classification of shell noun uses

shell noun uses are classified into 6 categories (Schmid, 2000):

`Factual`	*thing*, *problem*
`Linguistic`	*promise*, *story*
`Mental`	*idea*, *worry*
`Modal`	*possibility*, *truth*
`Eventive`	*mistake*
`Circumstantial`	*place*, *way*

Semantic classification of shell noun uses

examples

- `Factual`: The main thing is that we're bubbling again and the lads know we can do much better. [BNC K32: 2446]
- `Linguistic`: The most popular story concerning her conception was that a golden egg tumbled out of Chaos in the beginning of the world . [BNC CAC: 1107]
- `Mental`: In the ancient world, the belief was that each person was represented by a star. [BNC CEJ: 656]
- `Modal`: Their 31-year-old marriage has been described as unconventional but the reality is that they live entirely separate lives. [BNC HAE: 4911]

Semantic classification of shell noun uses

ambiguous, or rather vague shell noun uses

- `Linguistic` | `Mental` (see Schmid, 2000: 137f.):
  admission, *assumption*, *claim*, *forecast*, *guess*, *prediction*
- *fact*: `Factual` | `Modal` (see Schmid, 2000: 97):

  > "[T]his noun is used by speakers in the focusing pattern N-be-cl, i.e. in the collocation *the fact is* + *that*-clause, as an emphatic gesture. With the noun *fact*, however, the emphasis is not so much on the relevance of the shell content but on the claim that what is expressed in the *that*-clause is true. Such uses are therefore emphatics for epistemic necessity and will be looked at again in the section on epistemic uses (...)."

- *point*: `Factual` | `Linguistic` | `Mental` (see Schmid, 2000: 96)

Quantitative analysis: fixedness and productivity

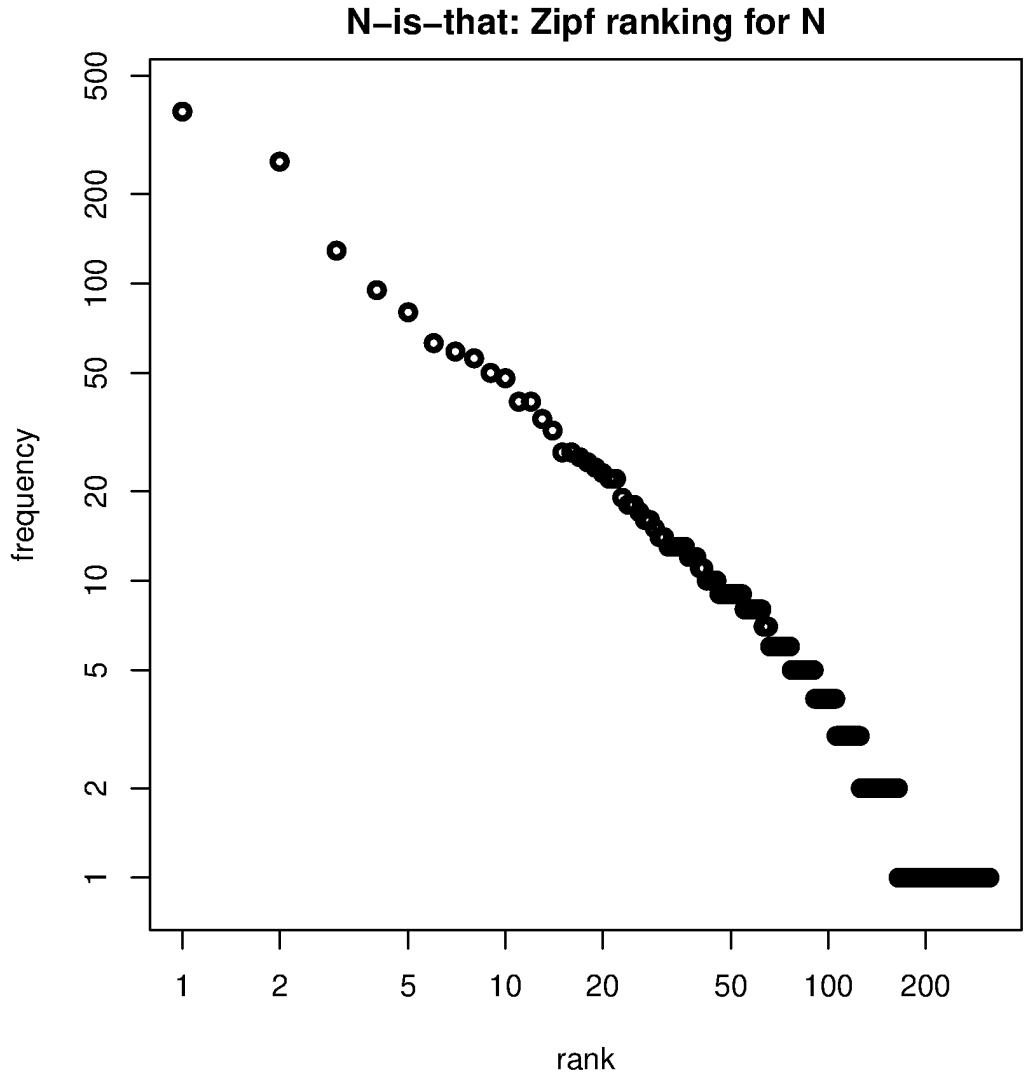
for the full sample of 2,525 instances

rank	f	type	
1.	379	point	result, answer,
2.	257	problem	advantage, position,
3.	129	thing	view, difficulty, truth,
4.	95	reason	effect, feature,
5.	80	fact	consequence,
6.	63	trouble	conclusion, implication,
7.	59	difference	explanation, argument
		...	
41.	11	fear	theory, change,
42.	10	feeling	impression, way,
43.	10	finding	essence, snag,
44.	10	significance	drawback, hope,
46.	9	belief	justification, message,
		...	objection, reality
167.	1	achievement	algorithm, rumour,
247.	1	objective	attitude, figure,
286.	1	satisfaction	subject, development,
301.	1	target	favour, practice, driver

Quantitative analysis: fixedness and productivity

for the full sample of 2,525 instances

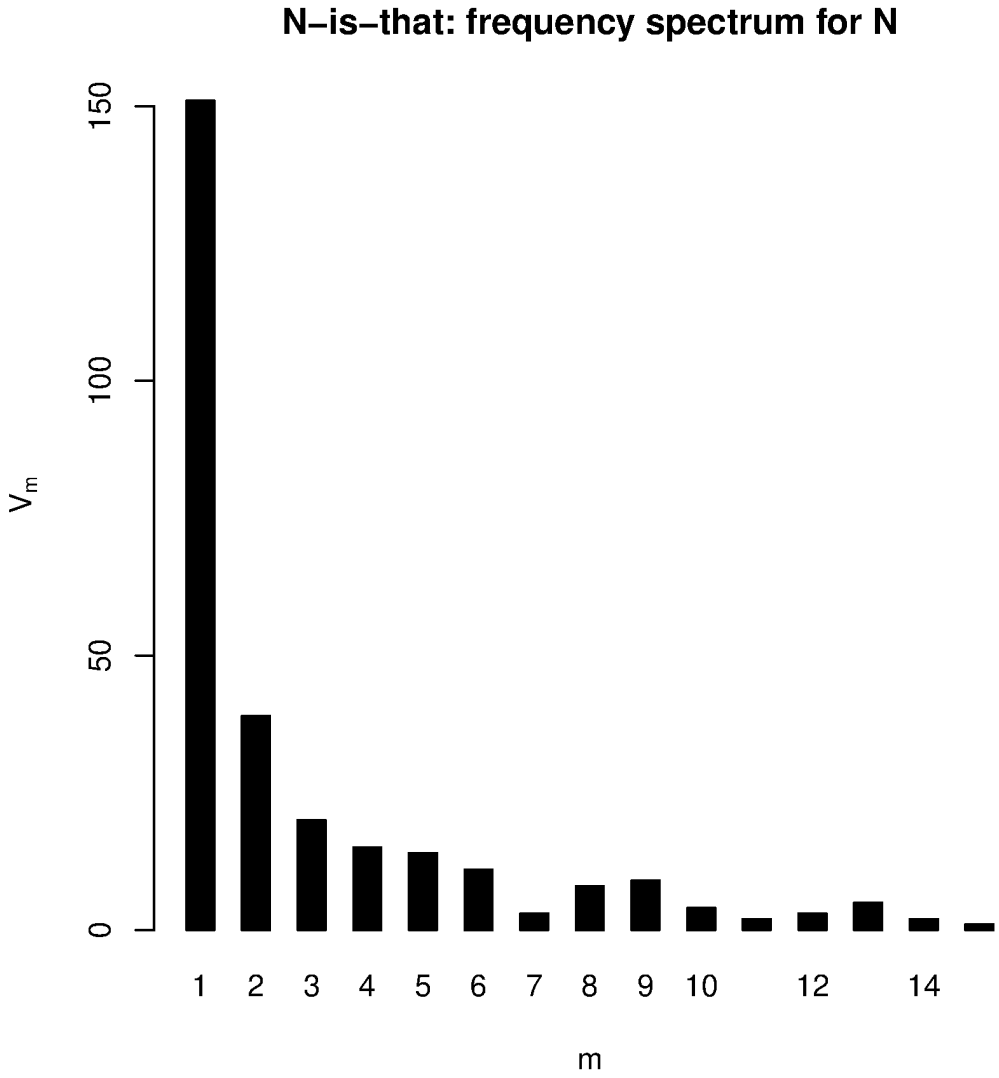
- relevant quantitative data: **type-token distribution** (Baayen, 2001)
- $N = 2525$ tokens
- $V = 315$ types
- $V_1 = 151$ hapaxes



Quantitative analysis: fixedness and productivity

for the full sample of 2,525 instances

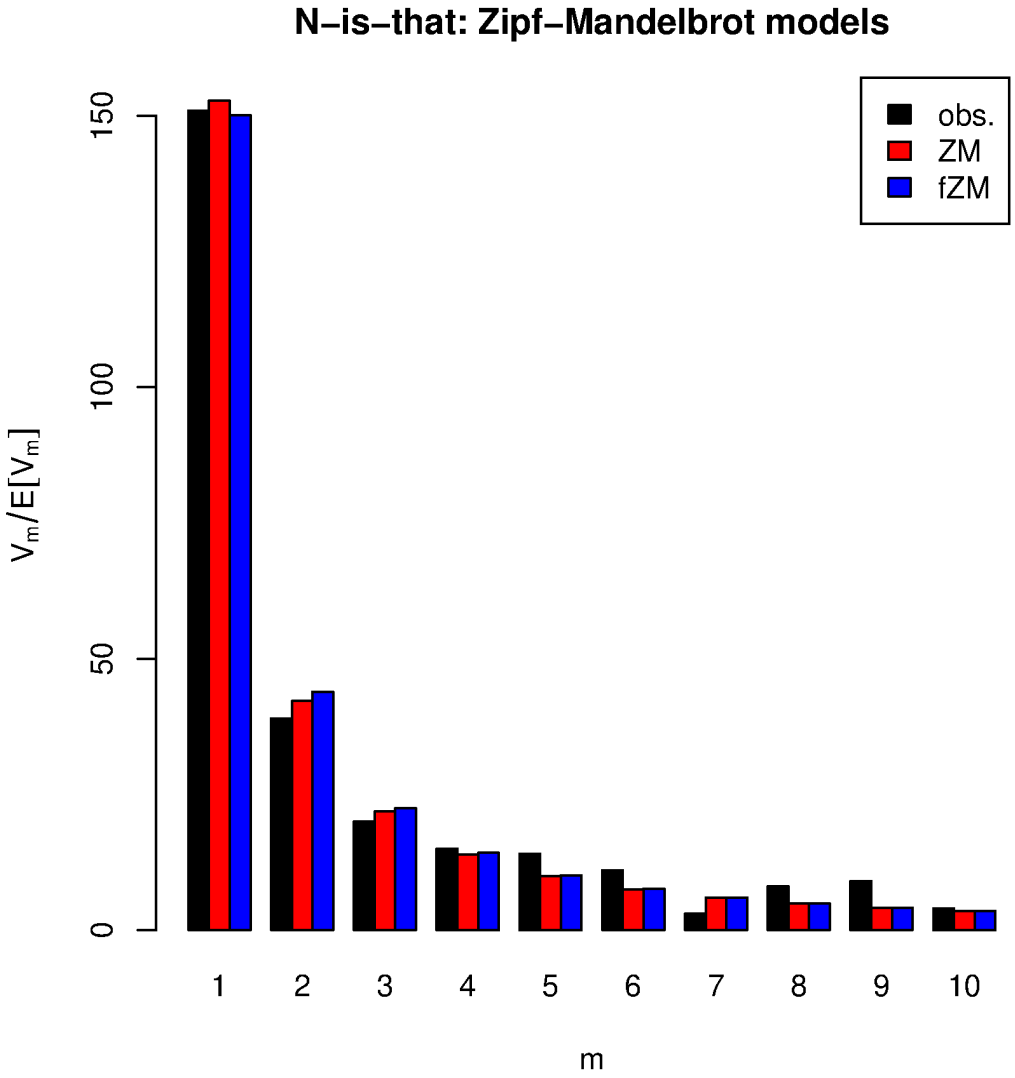
- relevant quantitative data: **type-token distribution** (Baayen, 2001)
- $N = 2525$ tokens
- $V = 315$ types
- $V_1 = 151$ hapaxes
- **frequency spectrum** $V_m \rightarrow$ productivity



Quantitative analysis: fixedness and productivity

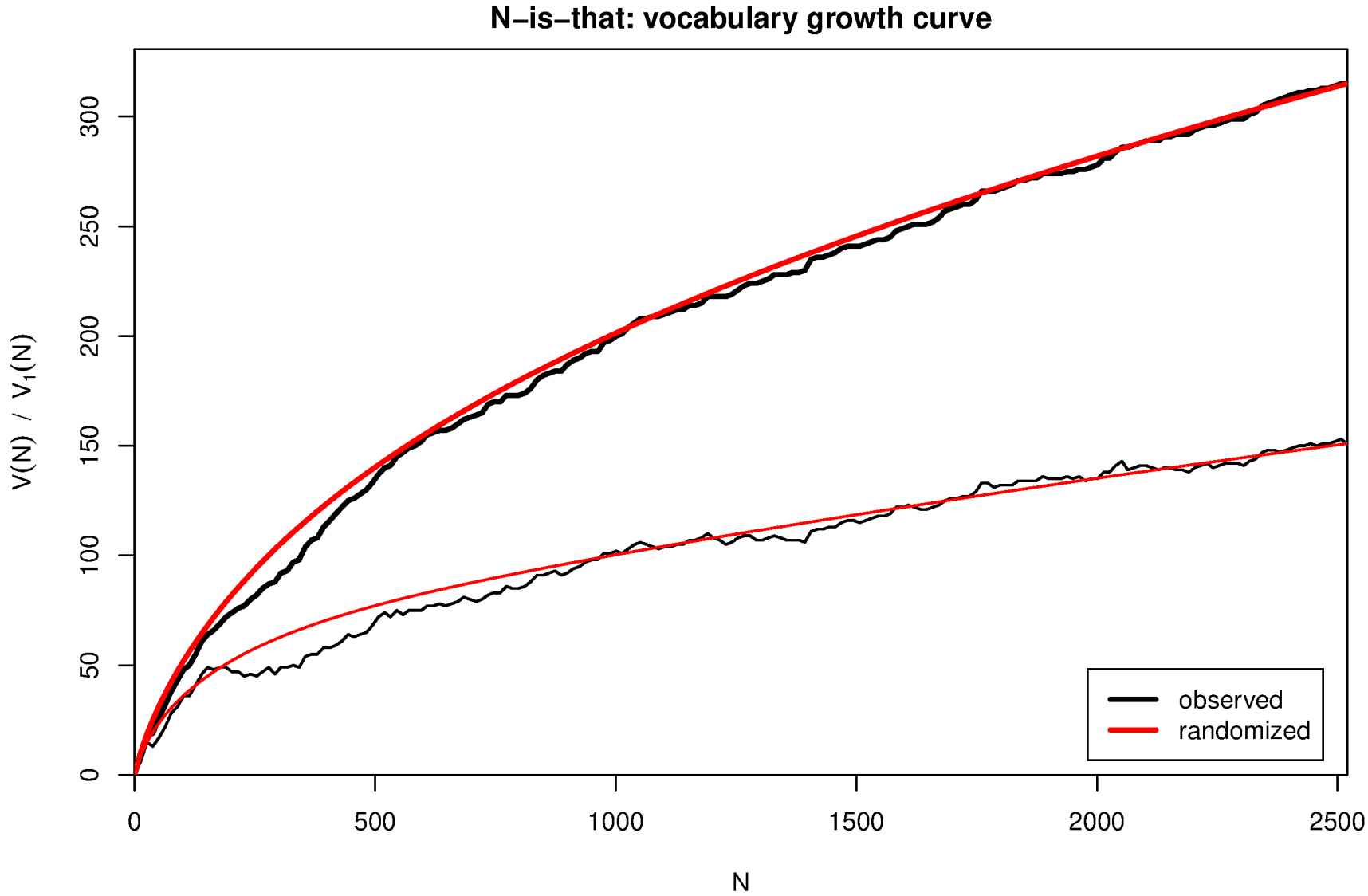
for the full sample of 2,525 instances

- relevant quantitative data: **type-token distribution** (Baayen, 2001)
- $N = 2525$ tokens
- $V = 315$ types
- $V_1 = 151$ hapaxes
- **frequency spectrum** $V_m \rightarrow$ productivity
- statistical analysis with **LNRE / ZM** (Baayen, 2001; Evert, 2004)



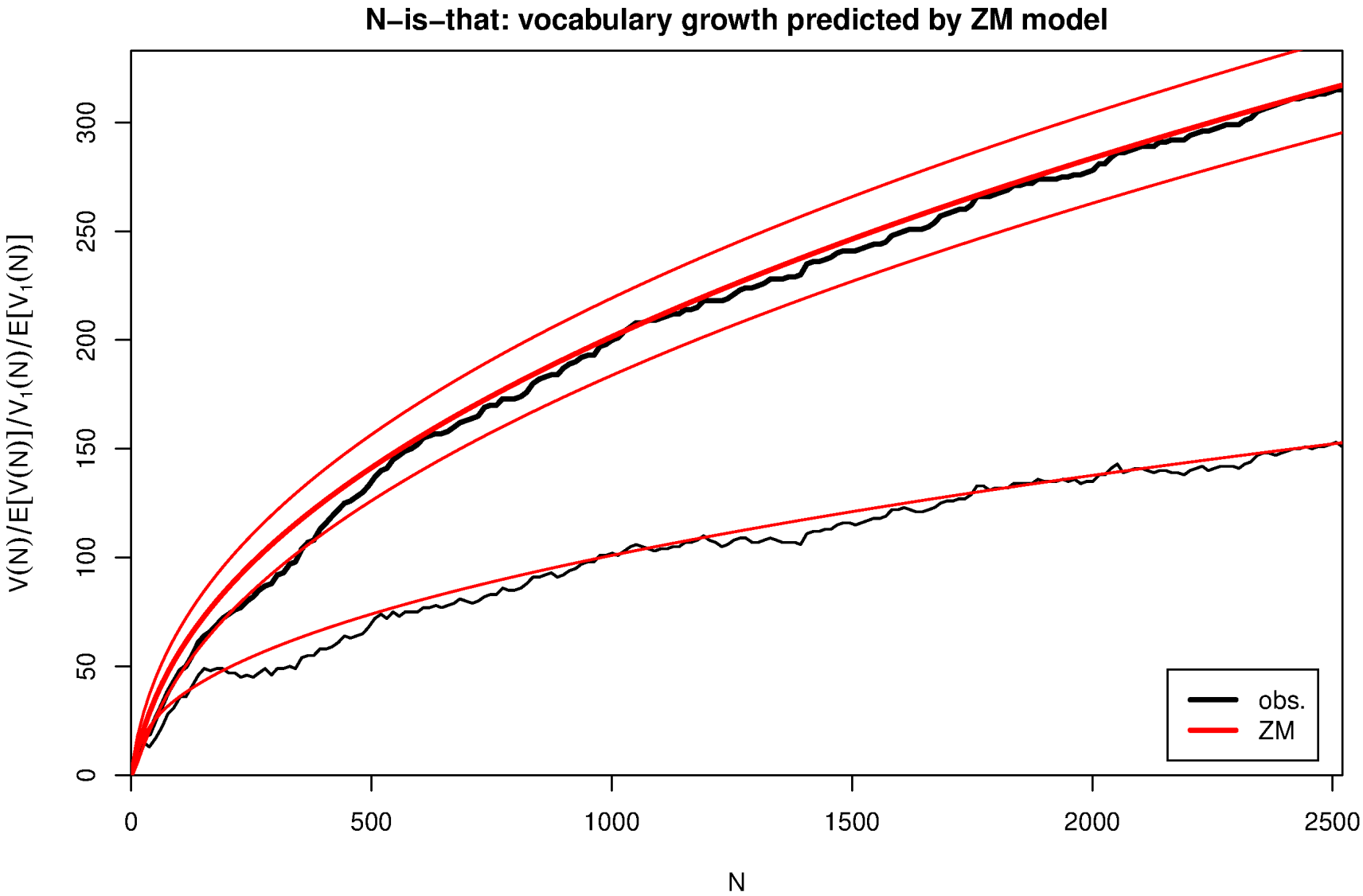
Vocabulary growth curves & non-randomness

for the full sample of 2,525 instances



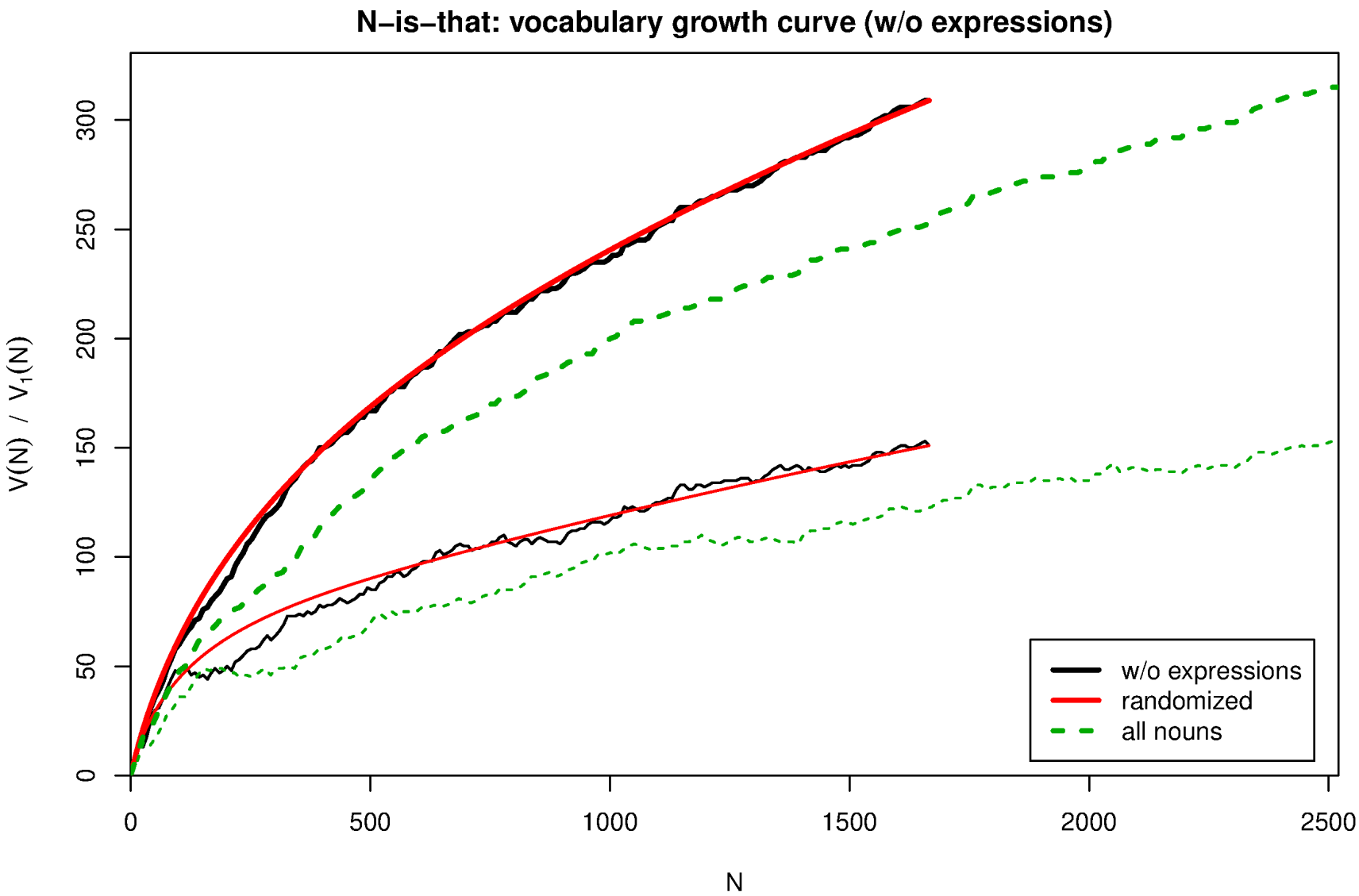
Vocabulary growth curves & non-randomness

for the full sample of 2,525 instances



Vocabulary growth curves & non-randomness

1,666 instances w/o expressions *the point/problem/fact/trouble/position/difficulty is that*

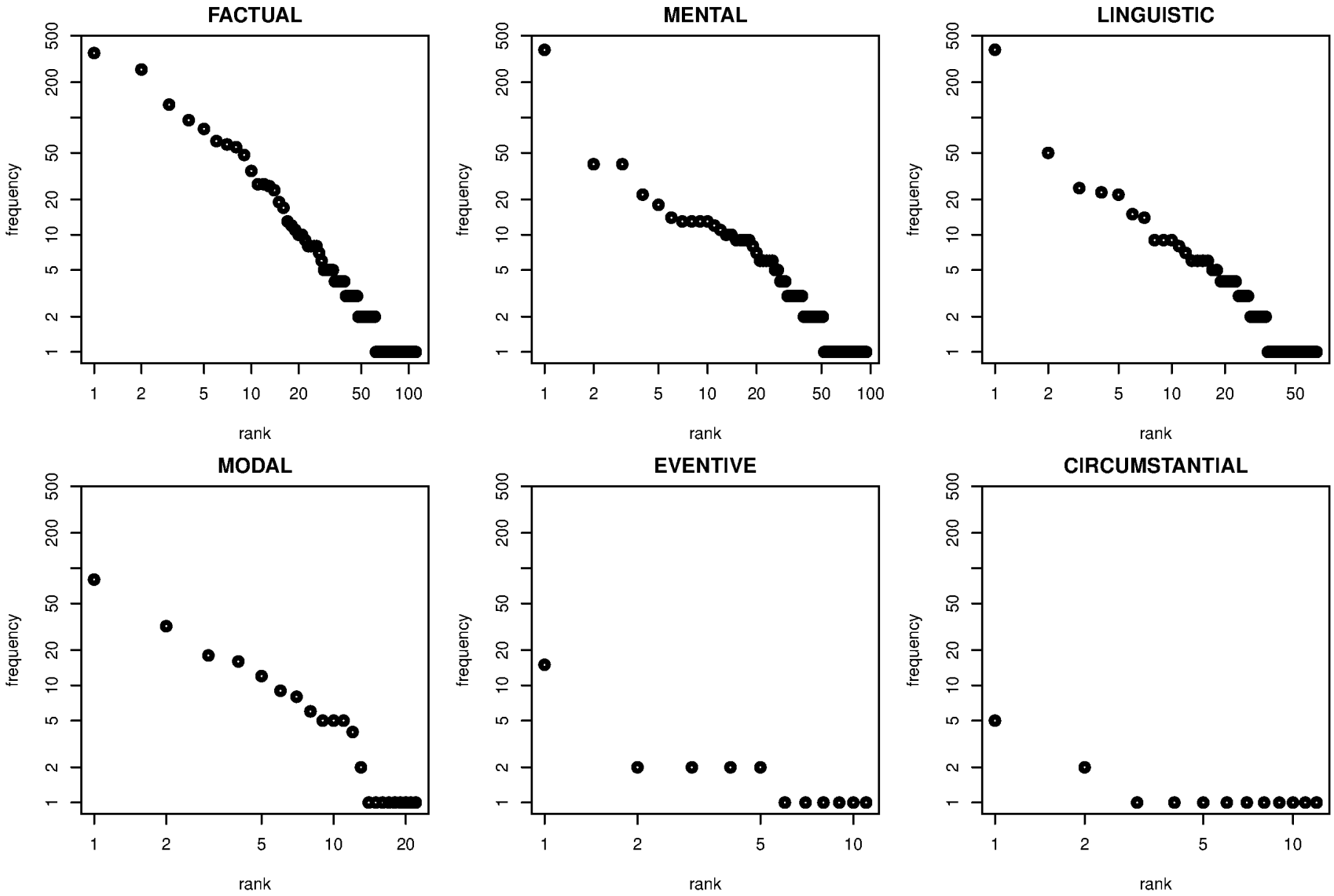


Comparison of shell noun categories

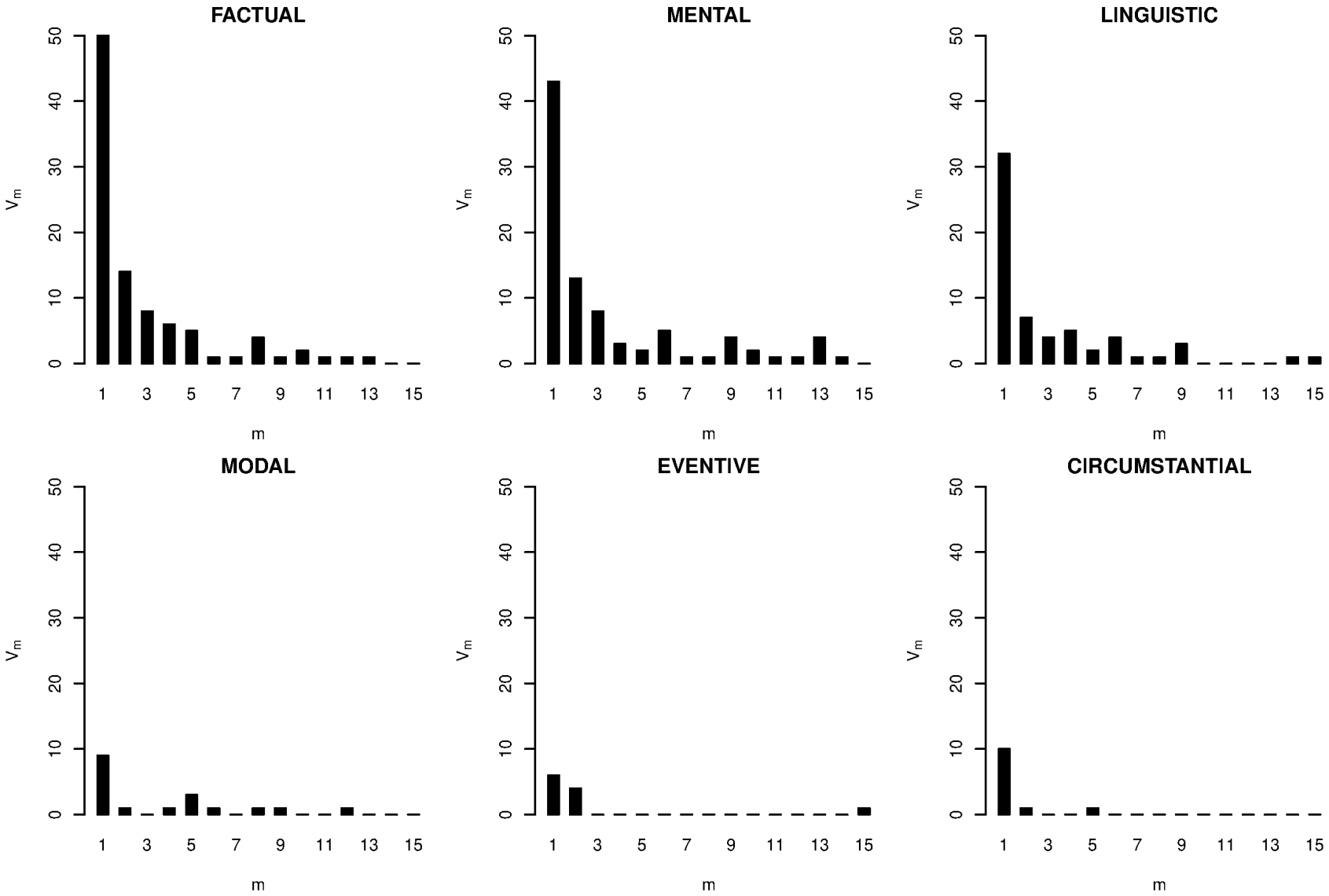
	all		w/o expr	
category	*V*	*N*	*V*	*N*
Circumstantial	12	17	12	17
Eventive	11	29	11	29
Factual	111	1578	106	788
Linguistic	66	682	65	303
Mental	94	803	92	385
Modal	22	211	21	131

For some analyses, the following expressions are excluded:
the point/problem/fact/trouble/position/difficulty is that

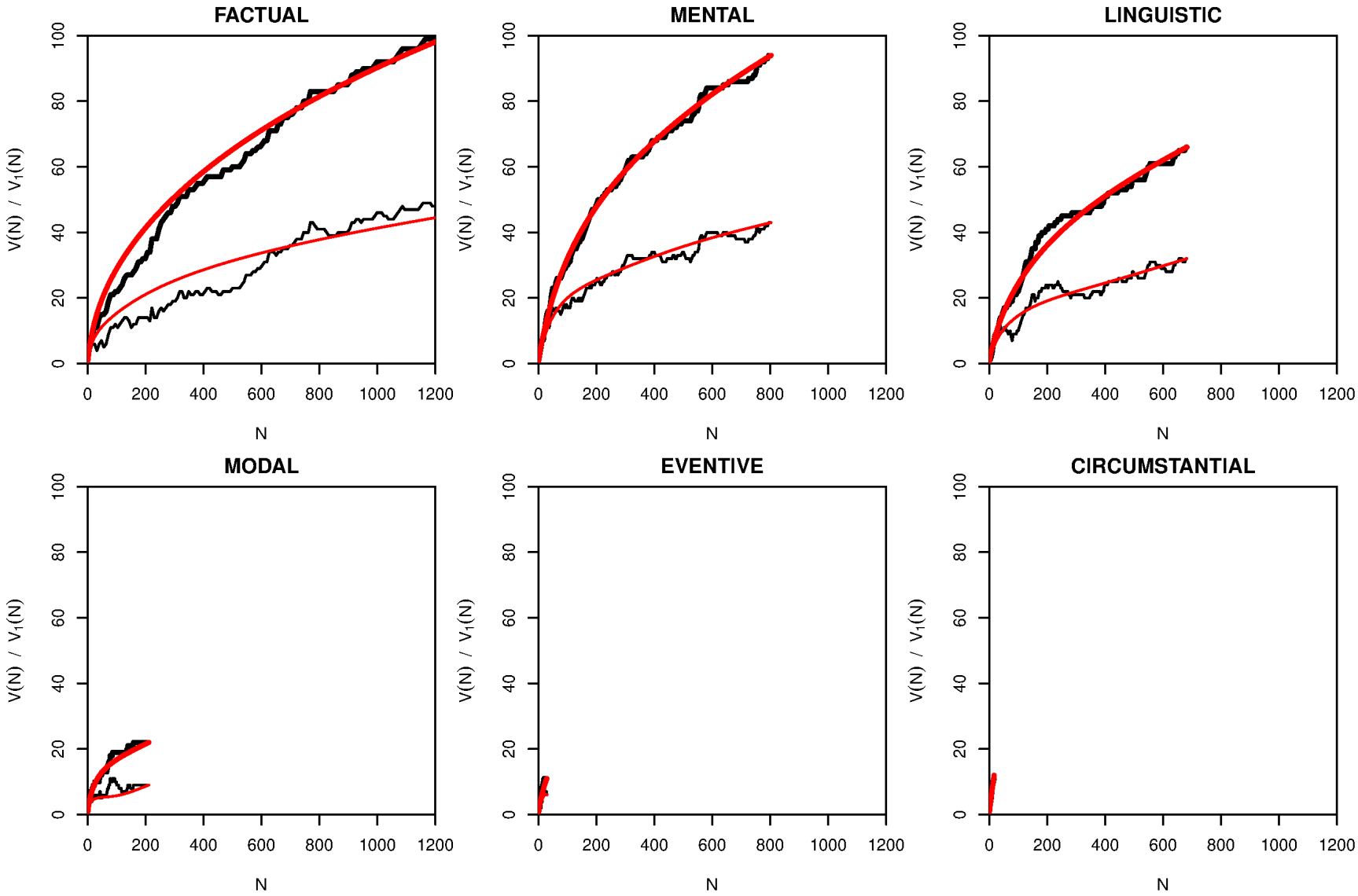
Comparison of shell noun categories



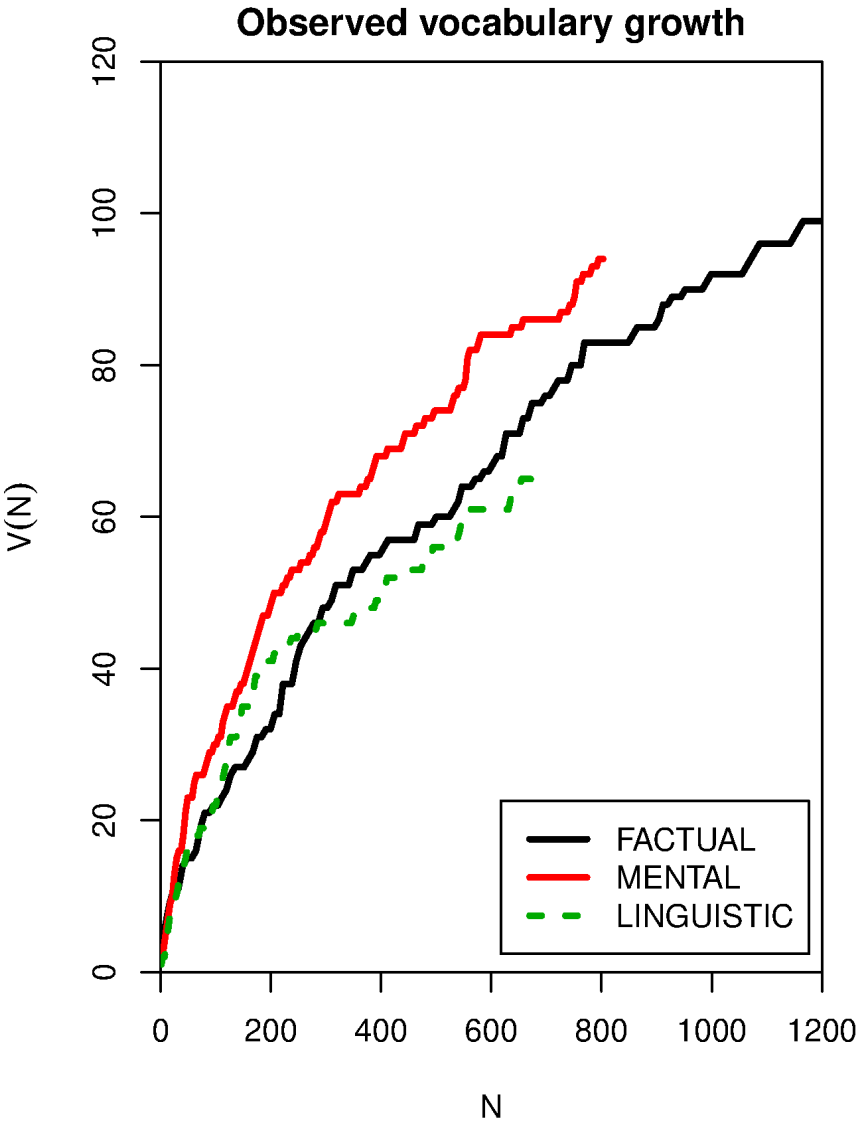
Comparison of shell noun categories

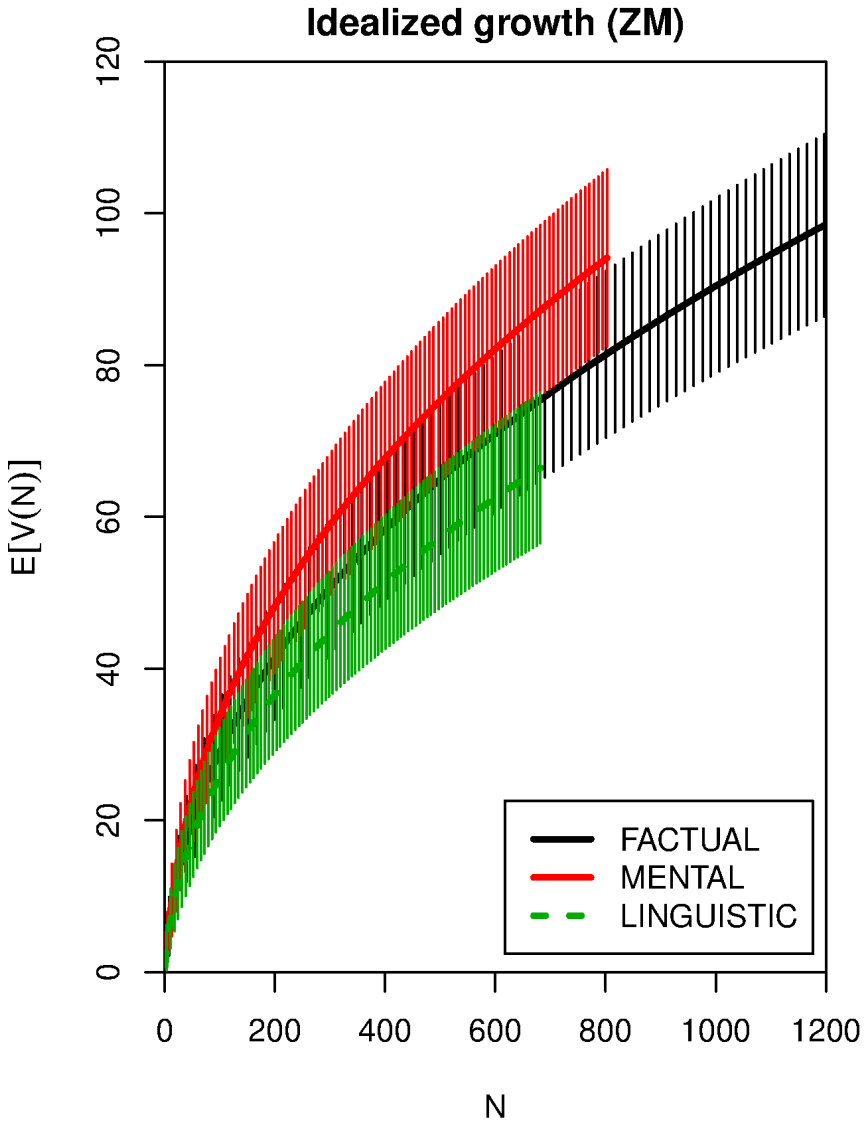


Comparison of shell noun categories

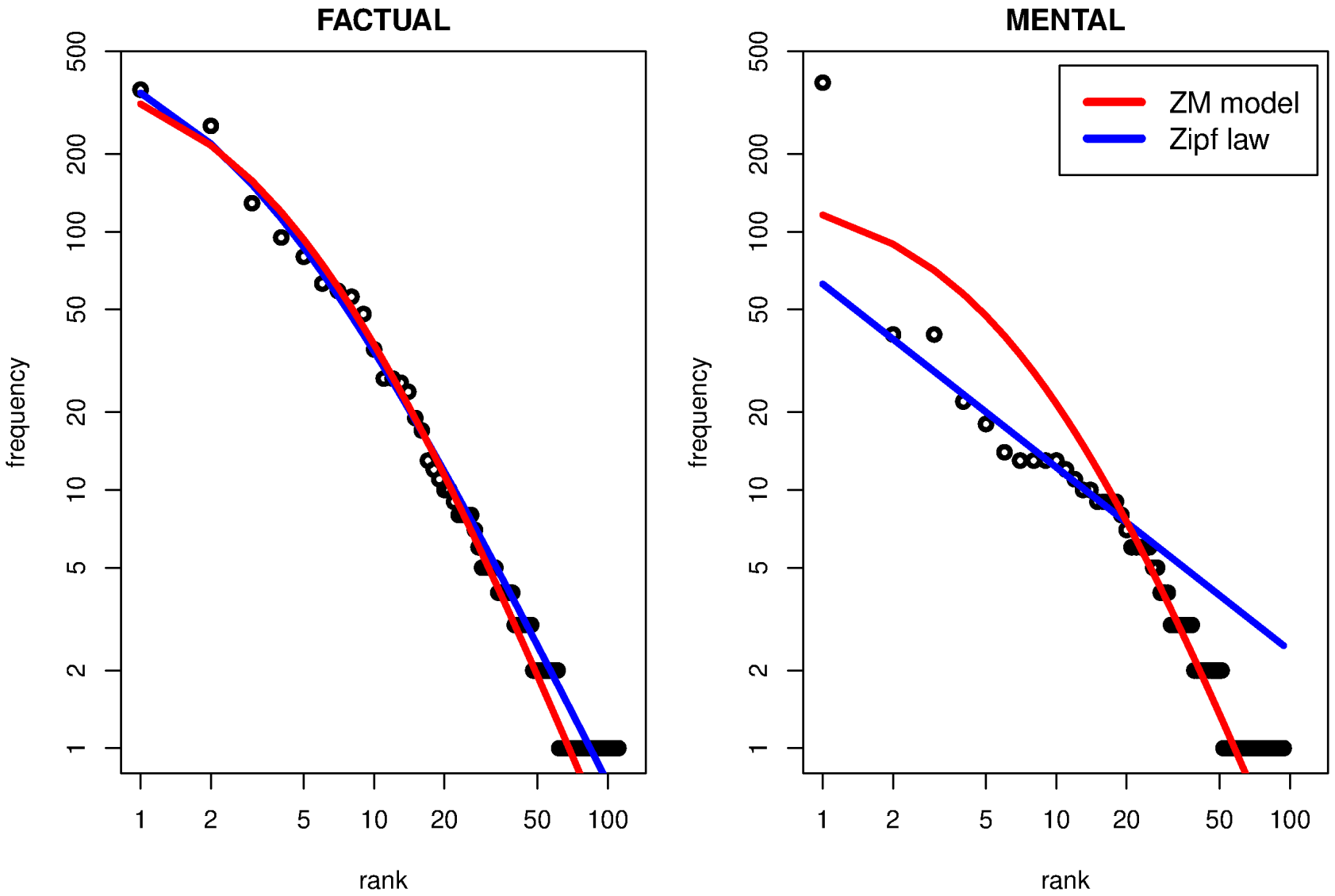


Statistical analysis with LNRE models

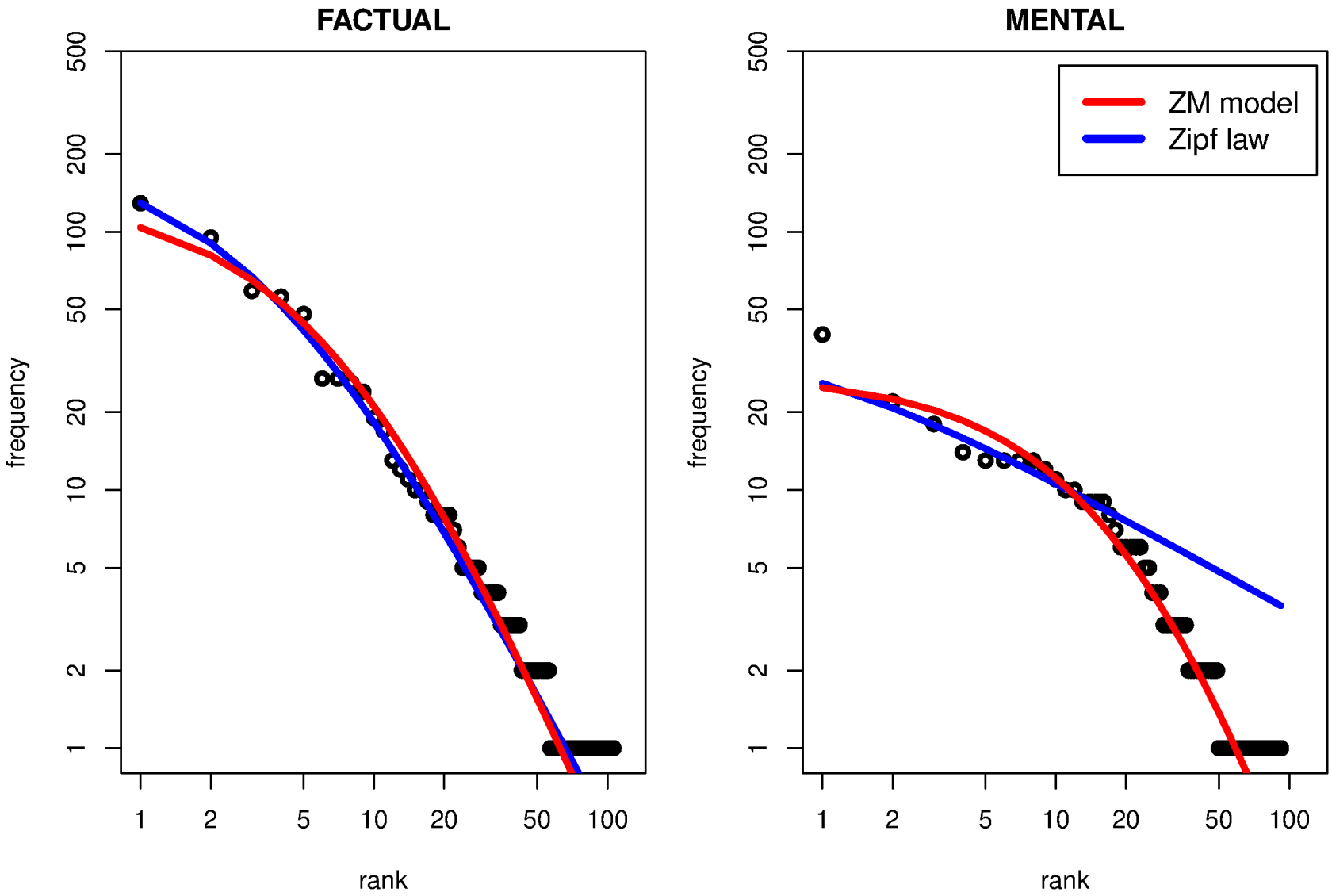




Limitations of current LNRE models



Limitations of current LNRE models (w/o expressions)

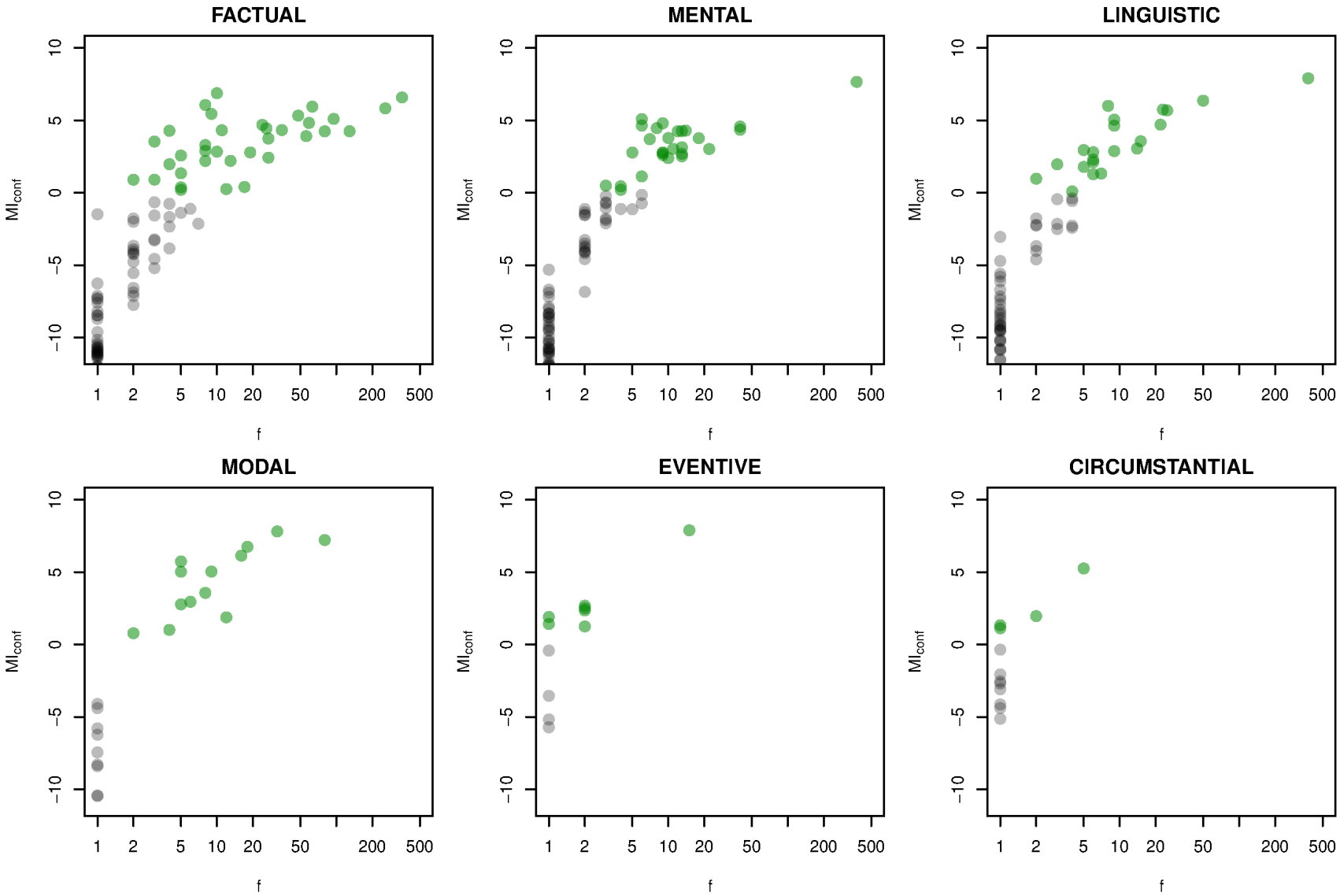


The middle ground: Statistical association

for the full sample of 2,525 instances

rang	f	type	E[f]	log G^2	MI_{conf}
1	379	point	4.787	7.850	5.986
2	257	problem	5.600	7.290	5.127
3	129	thing	7.510	6.198	3.535
4	95	reason	2.870	6.177	4.380
5	80	fact	4.198	5.771	3.517
6	63	trouble	0.947	6.006	5.217
7	59	difference	1.899	5.678	4.088
		...			
41	11	fear	0.942	3.554	1.189
42	10	feeling	1.218	3.240	0.529
43	10	finding	0.445	3.787	1.981
44	10	significance	0.465	3.768	1.917
46	9	belief	0.746	3.378	0.904
		...			
167	1	achievement	0.458	0.391	-14.817
247	1	objective	0.729	0.086	-15.487
286	1	satisfaction	0.287	0.728	-14.143
301	1	target	0.902	0.010	-15.794

The middle ground: Statistical association



Conclusion

- combination of quantitative approaches to capture the three sides of the syntax-lexicon continuum
  1. fixedness — frequency + concordance
  2. preference — association strength + semantics
  3. productivity — type-token distribution (LNRE models)

- methodological improvements needed
  - more flexible & robust LNRE models
  - integration of type-token statistics with association measures

Conclusion

- aspects of productivity and fixedness in terms of functional and structural parameters pertaining to the N+*is*+*that* pattern
  - differences between semantic classes have to do with the central role of the *that* clause from a functional point of view: characterizing propositions (`Factual`, `Linguistic`, `Mental`) vs. characterizing state of affairs
  - highly frequent (as well as ambiguous or vague) nouns, e.g. *point*: loss of (semantic) characterizing function in favour of the (textual) linking function → *the point is that* as emphatic focus marking connector
  - variation of the internal structure of the subject NP will need to be taken account of: *the* + N vs. DET:poss | POSS + N vs. DET:indef + PREMOD + N

Thanks for listening.
Questions?

References

R. Harald Baayen. *Word Frequency Distributions*. Kluwer Academic Publishers, Dordrecht, 2001.

Jóhanna Barðdal. *Productivity: Evidence from Case and Argument Structure in Icelandic*, volume 8 of *Constructional Approaches to Language*. John Benjamins Publishing Company, Amsterdam, December 2008.

William Croft and D. Alan Cruse. *Cognitive Linguistics*. Cambridge University Press, Cambridge, 2004.

Stefan Evert. 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, 2004.

Gaston Gross. *Les expressions figées en français, noms composés et autres locutions*. Ophrys, Paris, 1996.

Denis Le Pesant. Quelques schèmes productifs de noms composés de forme N de N. *Cahiers de lexicologie*, (82):105–115, 2003.

Salah Mejri. Le figement lexical : descriptions linguistiques et structuration sémantique. *L'information grammaticale*, 76(1):50–51, 1998.

Thomas Proisl and Peter Uhrig. Efficient dependency graph matching with the IMS Open Corpus Workbench. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2750–2756, Istanbul, 2012. European Language Resources Association.

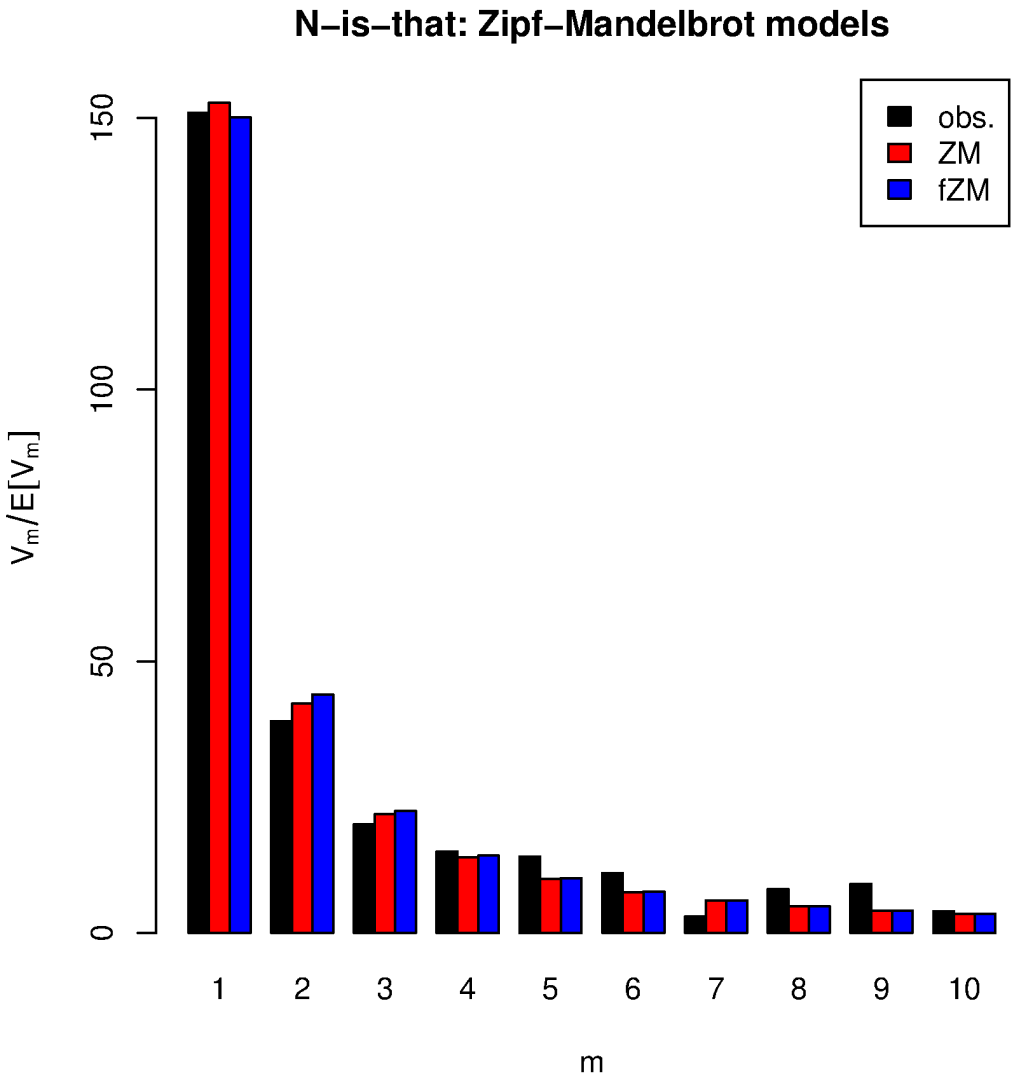
Hans-Jörg Schmid. *English Abstract Nouns as Conceptual Shells. From Corpus to Cognition*. De Gruyter Mouton, Berlin, Boston, 2000.

Statistical analysis with LNRE models

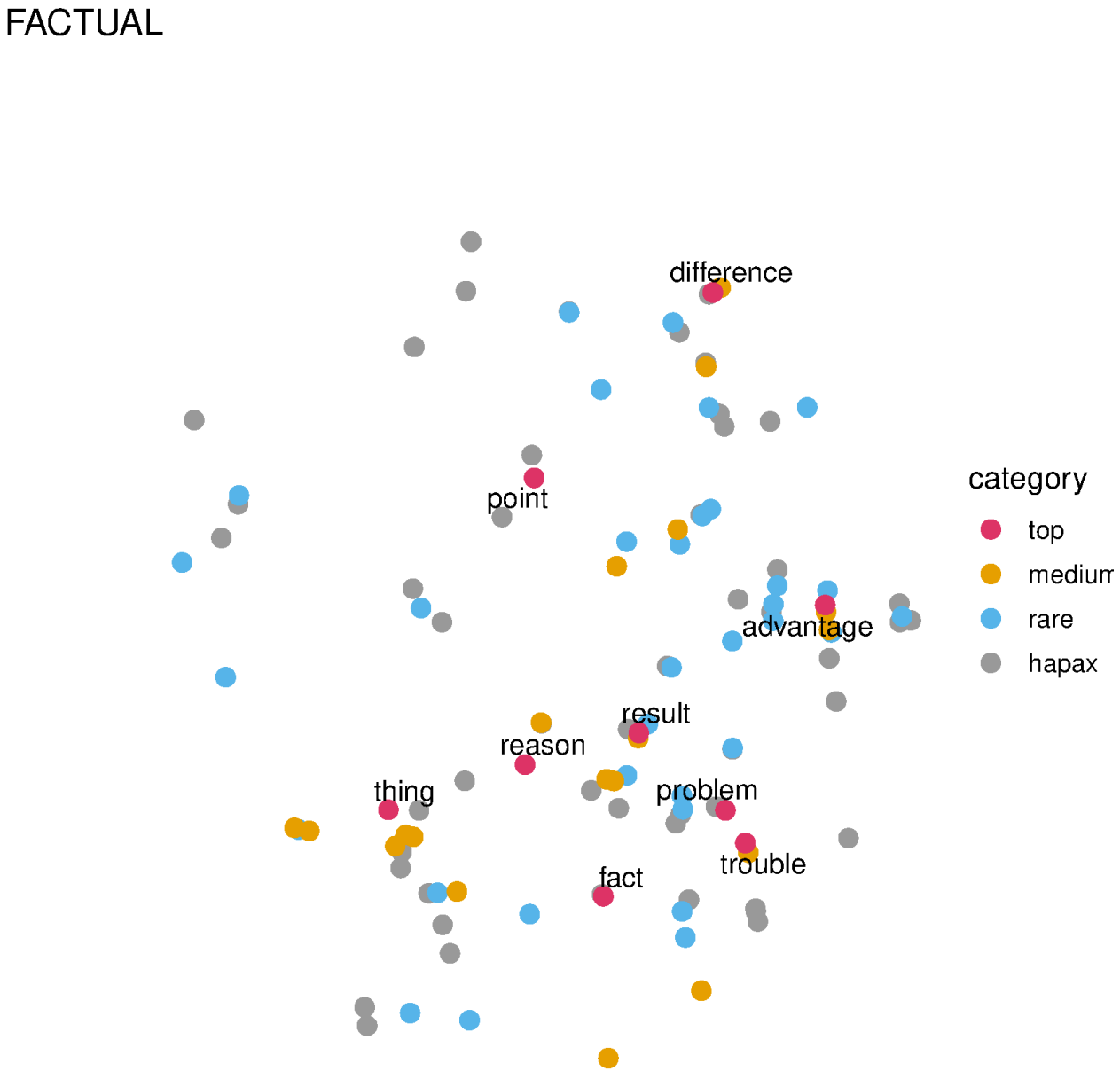
- LNRE model (Baayen, 2001) assumes Zipfian population
- parameters estimated from comparison of observed and expected frequency spectrum
- here: Zipf-Mandelbrot

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

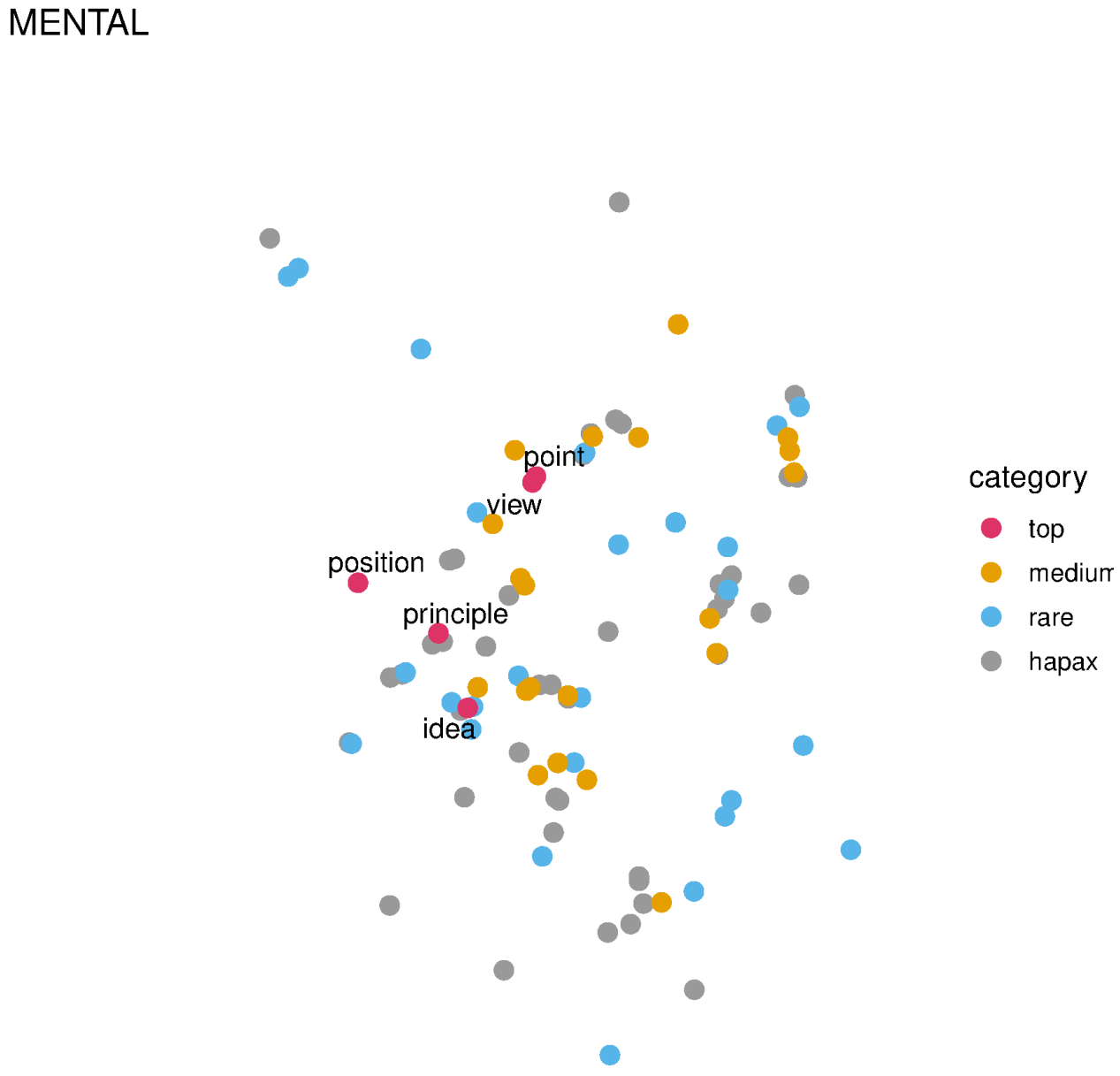
(Evert, 2004)



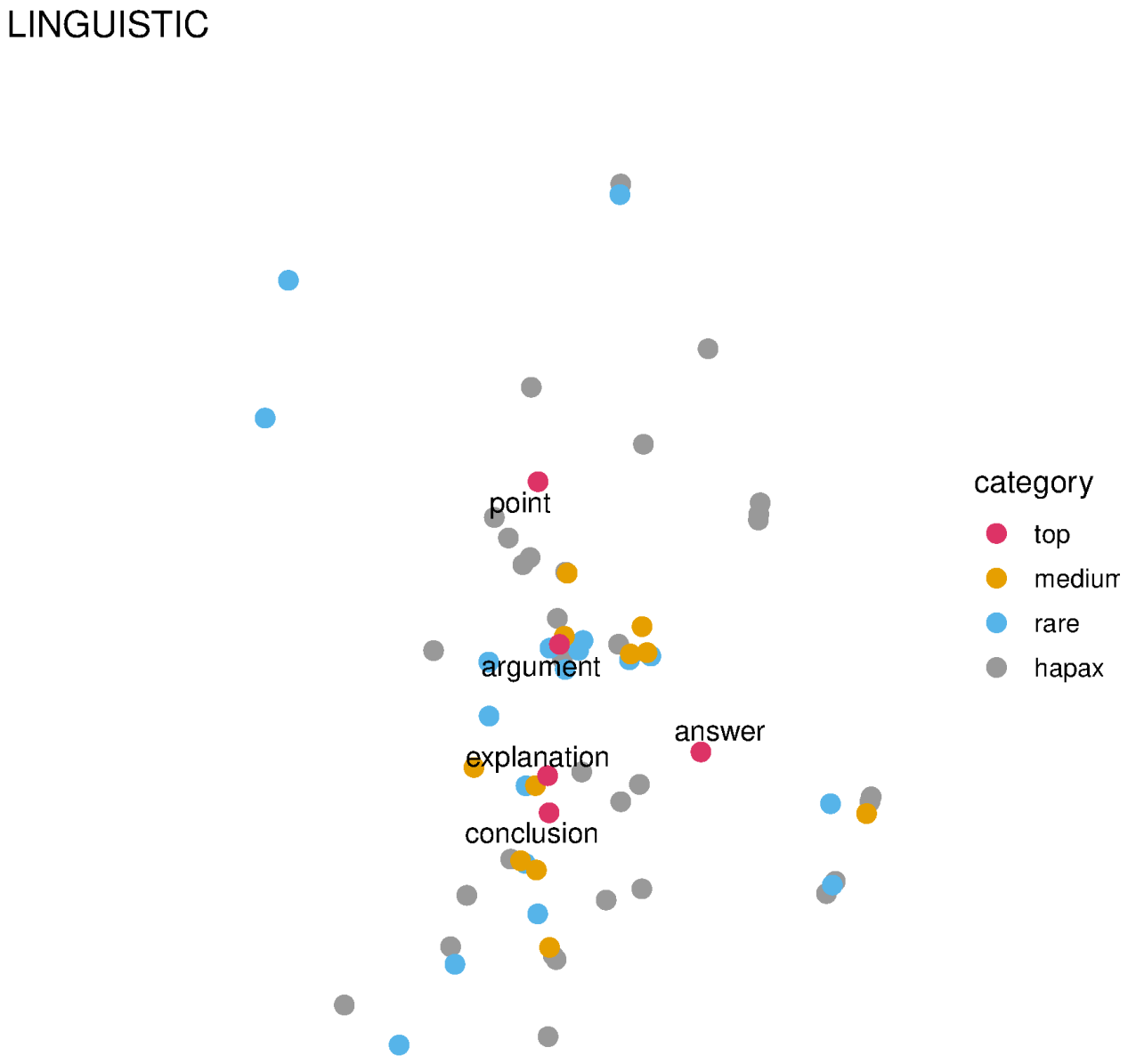
Productivity & semantics: Word embeddings



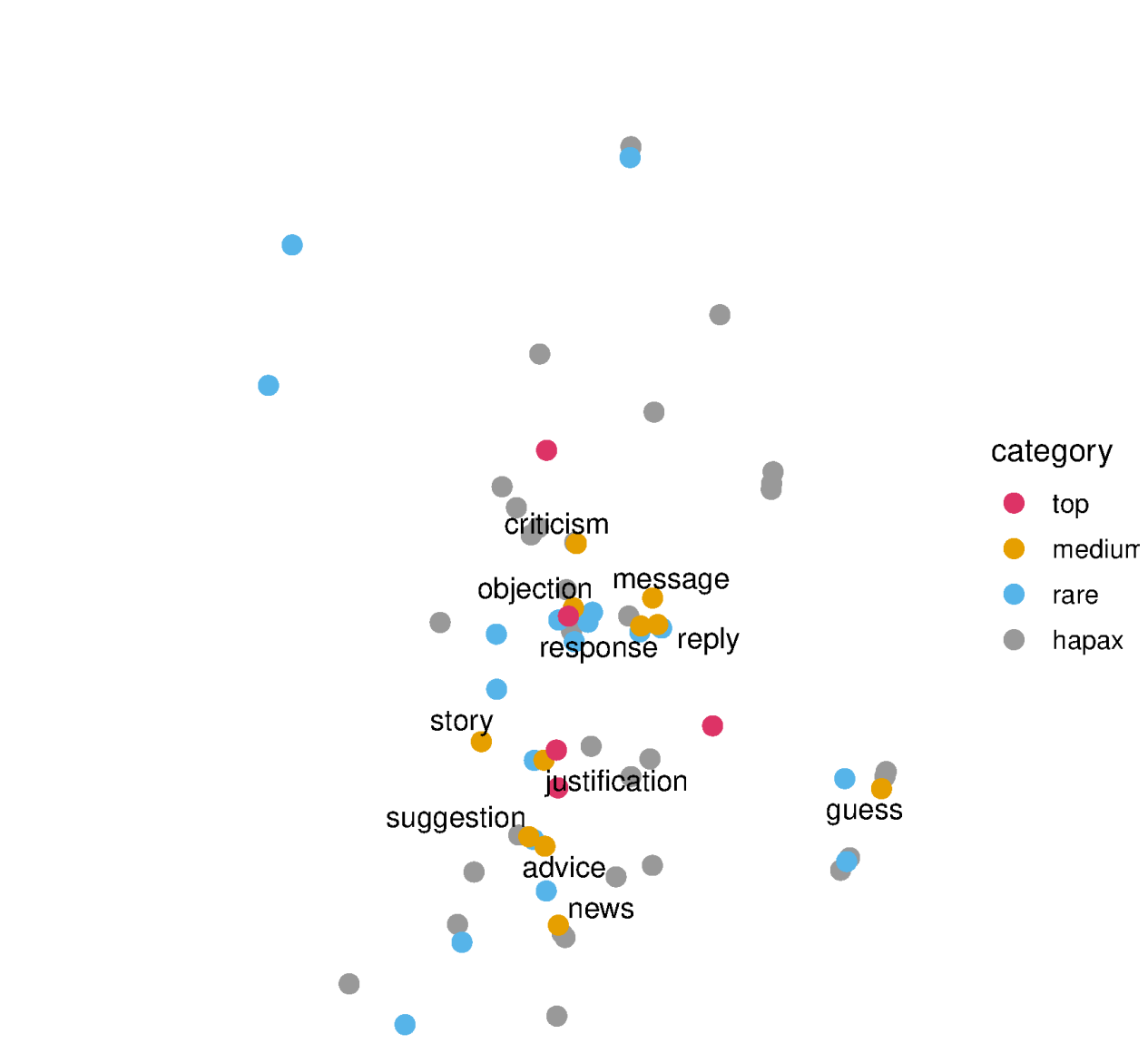
Productivity & semantics: Word embeddings



Productivity & semantics: Word embeddings



Productivity & semantics: Word embeddings



Productivity & semantics: Word embeddings