FRIAS

Lunch Lecture 2017/18:
Quantitative vs. qualitative approaches across sciences

11 January 2018

Reflecting on the quantitative turn in linguistics

Bernd Kortmann

UNI FREIBURG

Research Colloquium in English Linguistics

Luke Bradley, "Statistical creep in linguistics“

today 4:15 to 5 p.m.

this building (KG I), this floor, room 1034

1. Introduction

Inspired by general guiding questions of this LL. Central today:

- What does ***quantitative*** mean concerning the methodologies currently used in linguistics?
- Which **consequences** has it had on linguistics as a discipline that quantitative methods have made such dramatic inroads? Which consequences should it (not) have?
- In a number of disciplines we see increasingly critical responses towards the strong quantitative turns they have taken and warnings of a **“quantitative crisis”**. Is this the case for linguistics, too?

Quantitative / Statistical / Mathematical Linguistics

quantification and statistics nothing new in linguistics (going back especially to first half of 20th century)

- QL investigates languages using statistical methods
- most important objective: formulation of language laws and, ultimately, a set of interrelated language laws, all of which abide by universal laws which can be formulated strictly mathematically
- nature of these laws: **stochastic**, i.e. they are not observed in every single case, but rather determine the probabilities of the events or proportions under study, incl. statistical regularities in language, such as Zipf's Laws and Principles
  -> frequency of use

Zipf, George Kingley. 1949. *Human Behavior and the Principle of Least Effort.* Cambridge, Mass.: Addison-Wesley Press:

(i) Zipf's Law: There is a constant relationship between the rank of a word in a frequency list and the frequency with which it is used in a text: “The frequency of words in a given text is inversely proportional to their rank in frequency lists.”

(ii) Zipf's Principle of Economical Abbreviation: formal complexity correlates inversely with frequency of use;

(iii) Zipf's Law of Diminishing Returns: "... *there is a tendency for old age, small size, versatility of meaning, and a multiplicity of permutational associations - all to be directly correlated with high frequency of usage*" (1949: 121) -> 3 principles

- information theory (Claude E. Shannon, *A Mathematical Theory of Communication*, 1948):

  entropy as a measure of the unpredictability of a given sign, sign sequence, or message; alternatively: a measure of their average information content;

  compression techniques for measuring the information content of messages (-> Kolmogorov complexity)

- stylistics -> stylostatistics or, better known, stylometry (interface with literary studies)

- 1960s: the new paradigm of Quantitative sociolinguistics introduced by William Labov

- 1960s: compilation of the first corpora of English

- **Quantitative turn in linguistics since the 1990s and the early 2000s** (spearheaded by English linguistics):

  a turn in scale and quality, a turn concerning the degree (including: degree of sophistication) to which quantitative empirical studies, statistical techniques and statistical modelling have come to be used and determine linguistic research

  even in fields of linguistics dominantly working with qualitative methods some degree of quantification is increasingly expected (especially for journal publications):
  “nice to have > better to have”

Structure

1. Introduction
2. The Quantitative Turn

   2.1 Corpus linguistics
   2.2 Two applications: Aggregation and metricization
   2.3 Probabilistic linguistics

3. From Corpus to Cognition? (or: Do corpora mirror psychological reality?)
4. Major Caveats
5. A Quantitative Crisis in Linguistics?
6. Conclusion

2. The Quantitative Turn since the 1990s & early 2000s

- In Brian Joseph's final editorial as the editor of what many see as the flagship journal of the discipline, *Language,* he commented on recent developments in the field. One of the recent developments he has seen happening is the following:

> Linguistics has always had a numerical and mathematical side ... but the use of quantitative methods, and, relatedly, formalizations and modeling, seems to be ever on the increase; rare is the paper that does not report on some statistical analysis of relevant data or offer some model of the problem at hand. (Joseph 2008: 687)

Gries in Krug/Schlüter 2013: 361

For several reasons, this appears to be a development for the better:

- First, it situates the field of linguistics more firmly in the domains of social sciences and cognitive science… Other fields in the social sciences and in cognitive science - psychology, sociology, computer science, to name but a few - have long recognized the power of quantitative methods for their respective fields of study, and since linguists deal with phenomena just as multifactorial and interrelated as scholars in these disciplines, it was time we also began to use the tools that have been so useful in neighboring disciplines.

Gries in Krug/Schlüter 2013: 361

- Second, the quantitative study of phenomena affords us with a higher degree of comparability, objectivity, and replicability.

- Third, there is increasing evidence that much of the cognitive and/or linguistic system is statistical or probabilistic in nature.

- […] and if one adopts a probabilistic theoretical perspective, then the choice of probabilistic – i.e. statistical – tools is only natural; […]

Gries in Krug/Schlüter 2013: 361f.

The major driver: corpus linguistics

- […] corpus linguistics has been among the fastest-growing methodological disciplines in linguistics […]

- […] corpus linguistics has become mainstream […] (p. 113)

- „… that linguistics in general has become much more quantitative/statistical in nature is a trend we also witness in corpus linguistics: For example, 10 or 15 years ago it would have been quite difficult to find papers with multifactorial statistical techniques in corpus-linguistics papers – now, monofactorial statistical tests at least are much more frequent, and multifactorial statistical methods are on the rise. (p. 93)

Gries 2016: 93, 113

2.1 Corpus linguistics

as a method,

- is empirical, analyzing actual patterns of use in natural texts
- utilizes a large and principled collection of natural texts, known as a “corpus”, as the basis for analysis
- makes extensive use of computers for analysis, using both automatic and interactive techniques
- depends on both quantitative and qualitative analytical techniques

Biber 2010: 160f.

Two major research approaches

- corpus-based research: attempts to describe the systematic patterns of variation and use for linguistic features and constructs that have been previously identified by linguistic theory

- corpus-driven research makes minimal a priori assumptions regarding the linguistic features that should be employed for the corpus analysis; it rather attempts to uncover new linguistic constructs through inductive analysis of very large, representative corpora combined with computational tools for analysis

Biber 2010: 169

Corpus-driven research: Lexical Bundles

- In its most basic form, corpus-driven analysis assumes only the existence of words, while concepts like “phrase” and “clause” have no a priori status. Rather, **cooccurrence patterns** among words, **discovered** from the corpus analysis, are the basis for subsequent linguistic descriptions.

- Lexical bundles [or: *fixed-word sequences*, *n-grams*] are defined as the multi-word sequences that recur most frequently and are distributed widely across different texts. Lexical bundles in English conversation are word sequences like *I don't know if* or *I just wanted to.* They are usually neither structurally complete nor idiomatic in meaning.

Biber 2010: 162, 170

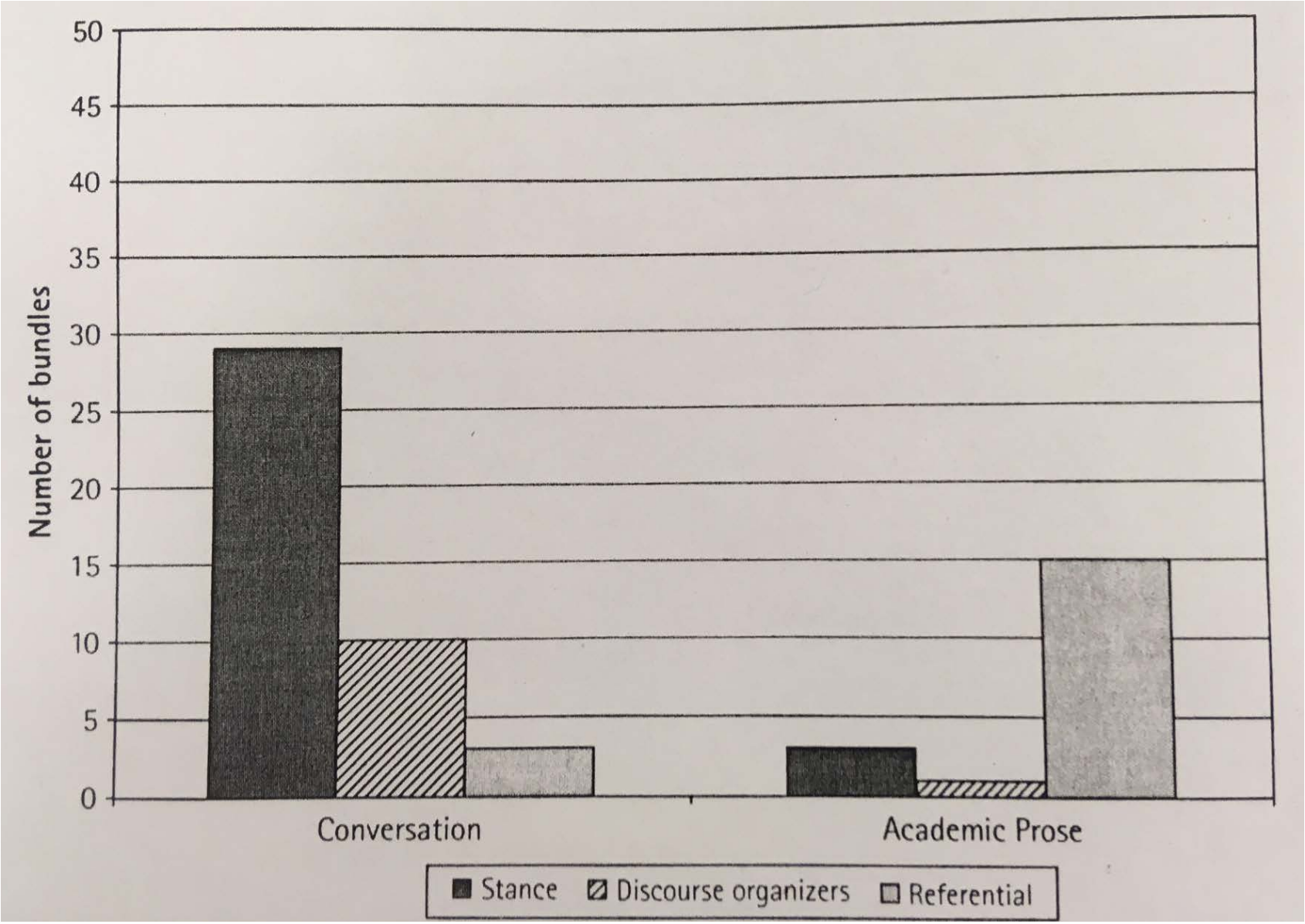


Figure 8.6. Distribution of lexical bundles across functional types (4-word bundles occurring more than 40 times per million words)

Biber 2010: 174

Currently available corpora just for English

- **103 corpora** (excluding sub-corpora of e.g. ICE or ICLE)

- **19,041,436,907 words** **-> BIG DATA** indeed

  out of these: 5.5 billion words in NOW (News on the Web Corpus), 2 billion words each in GloWbE (Corpus of Global Web-Based English), Hansard Corpus and Wikipedia

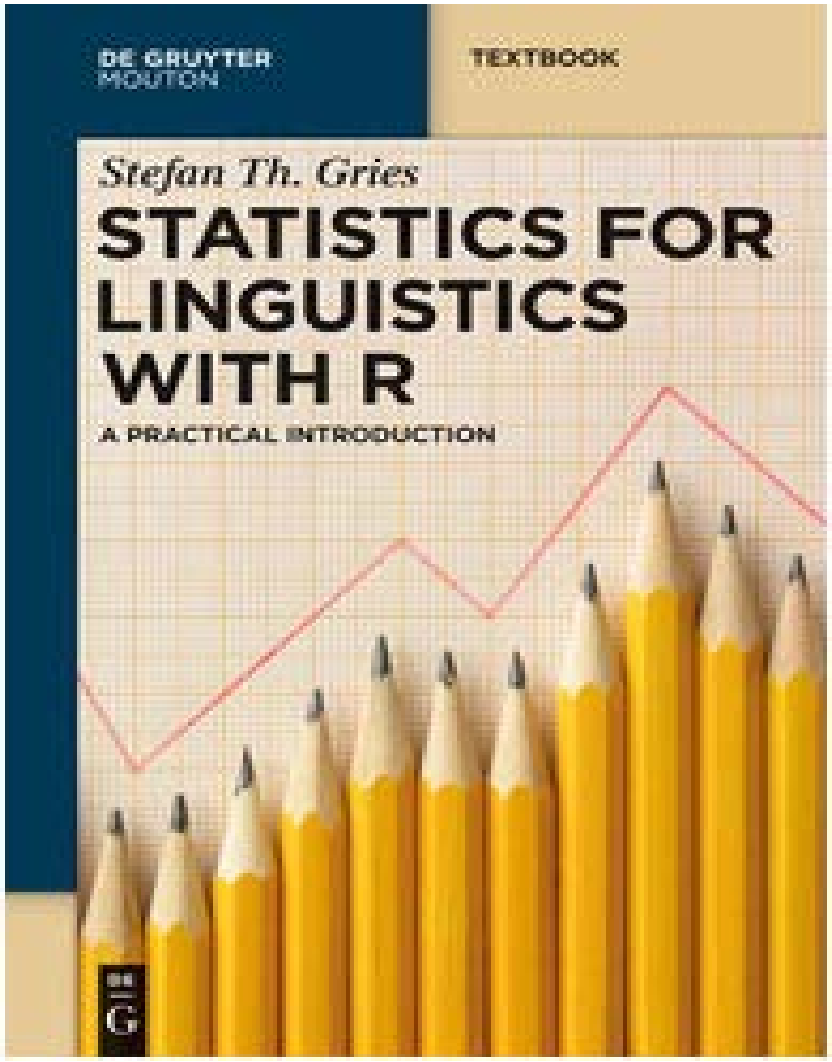
  -> web-based corpora grow on a daily basis
  (e.g. NOW: by 5-6 million words every day)

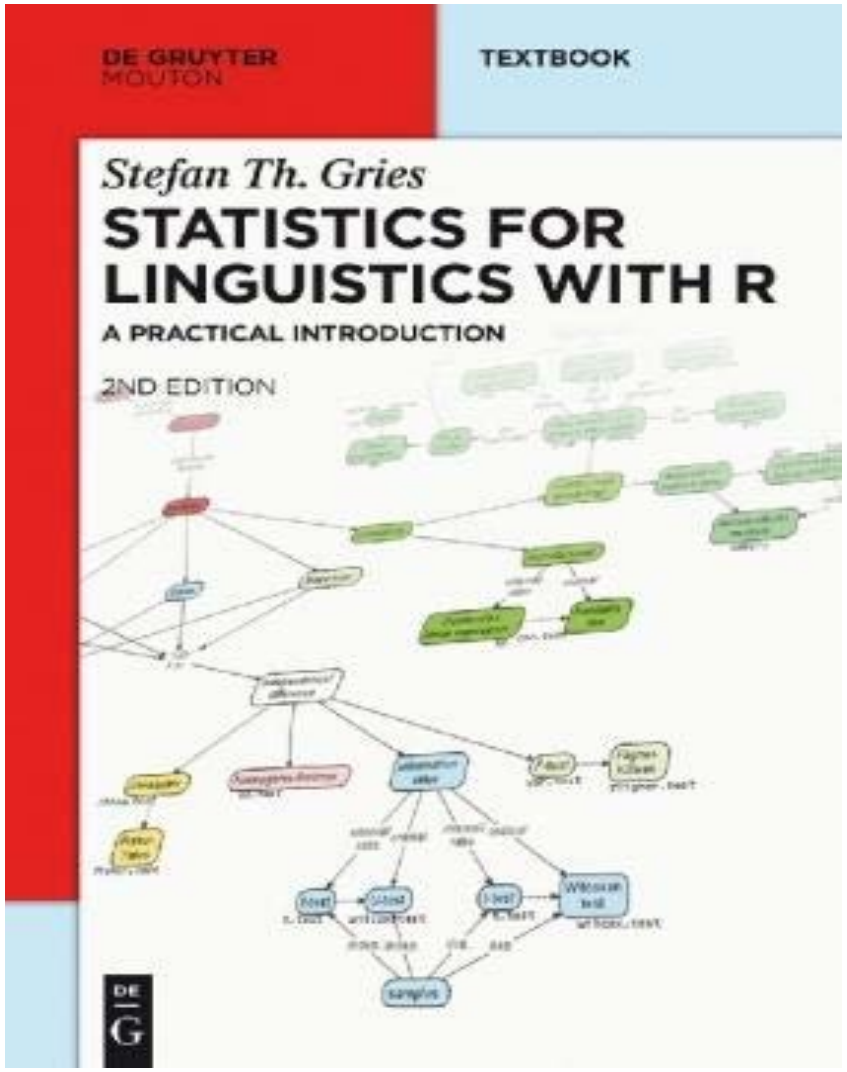
- plus: 2,000 billion words in Google Books
- plus: constantly compiled data from social networks
  (e.g. Twitter)

Some indicators of the corpus-linguistic and quantitative turn

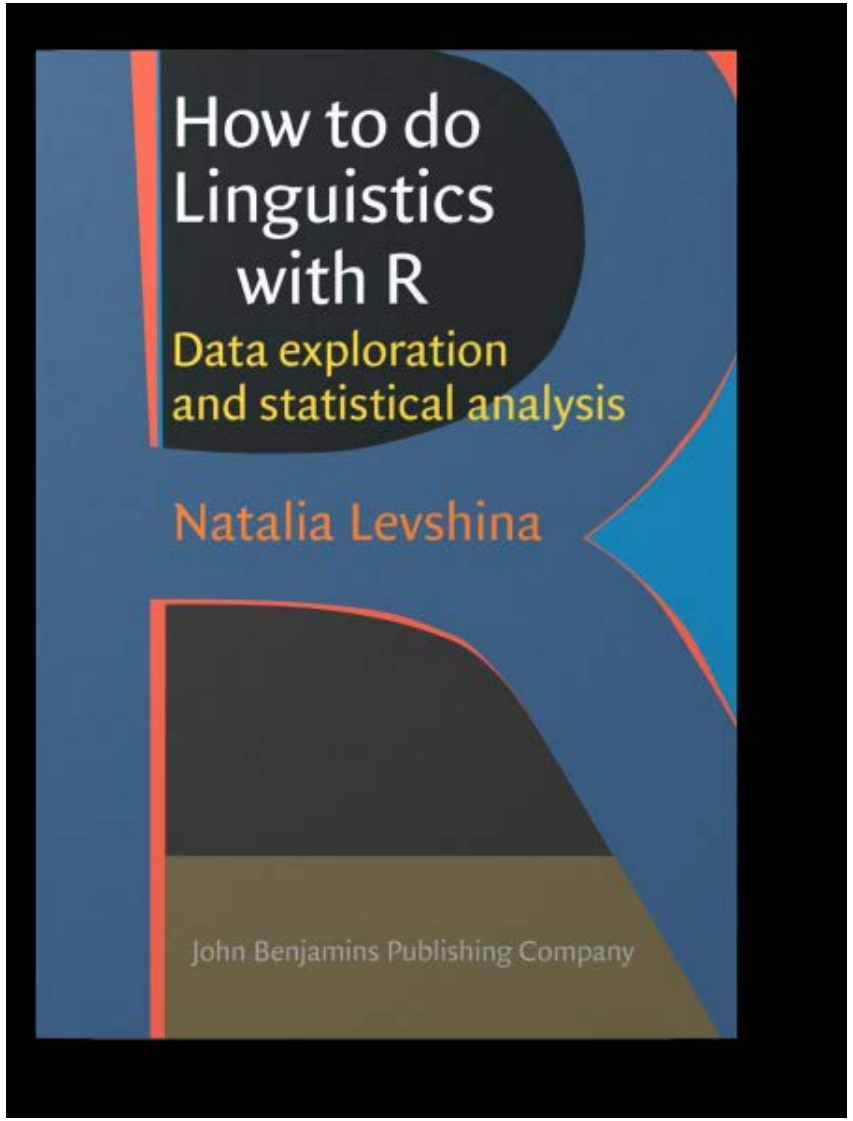
- **13 Introductions to Corpus Linguistics** (since 1998)
- **5 Handbooks** (2009; 2012, 2015, 2015, 2017)
- **6 Book series** (since 1988)
- **4 Journals** (since 2002)
- **3 Introductions to Statistics for Linguist(ic)s** (since 2008)

- ***Statistics for Linguistics with R: A Practical Introduction***
  By: Gries, Stefan Th. Berlin, Germany: de Gruyter Mouton; 2009, 2013 2nd rev. ed., xiii, 359 pp.





- ***How to Do Linguistics with R: Data Exploration and Statistical Analysis***
  By: Levshina, Natalia. Amsterdam, Netherlands: Benjamins; 2015. xi, 443 pp.



2.2 Two applications: Aggregation and metricization

Metrics

e.g. for measuring

- structural (dis)similarity

  of languages or linguistic varieties
- structural complexity

Computing linguistic distances between varieties

- Focus on accent differences (based on a comparison of 110 words in 59 accents of English, i.e. a total of 6490 sound files)

- Rhoticity:
  rhotic accents (/r/ pronounced in all positions) vs
  non-rhotic accents (only pre-vocalic /r/ is pronounced)

- […] the phylogenetic software NeighborNet… is widely used in interdisciplinary research, having been developed for population biology, but implemented for data from anthropology, archaeology and linguistics (…). Clearly, using the same visualization software improves our options for comparison with data from other disciplines.

McMahon/Maguire in Krug/Schlüter 2013: 421

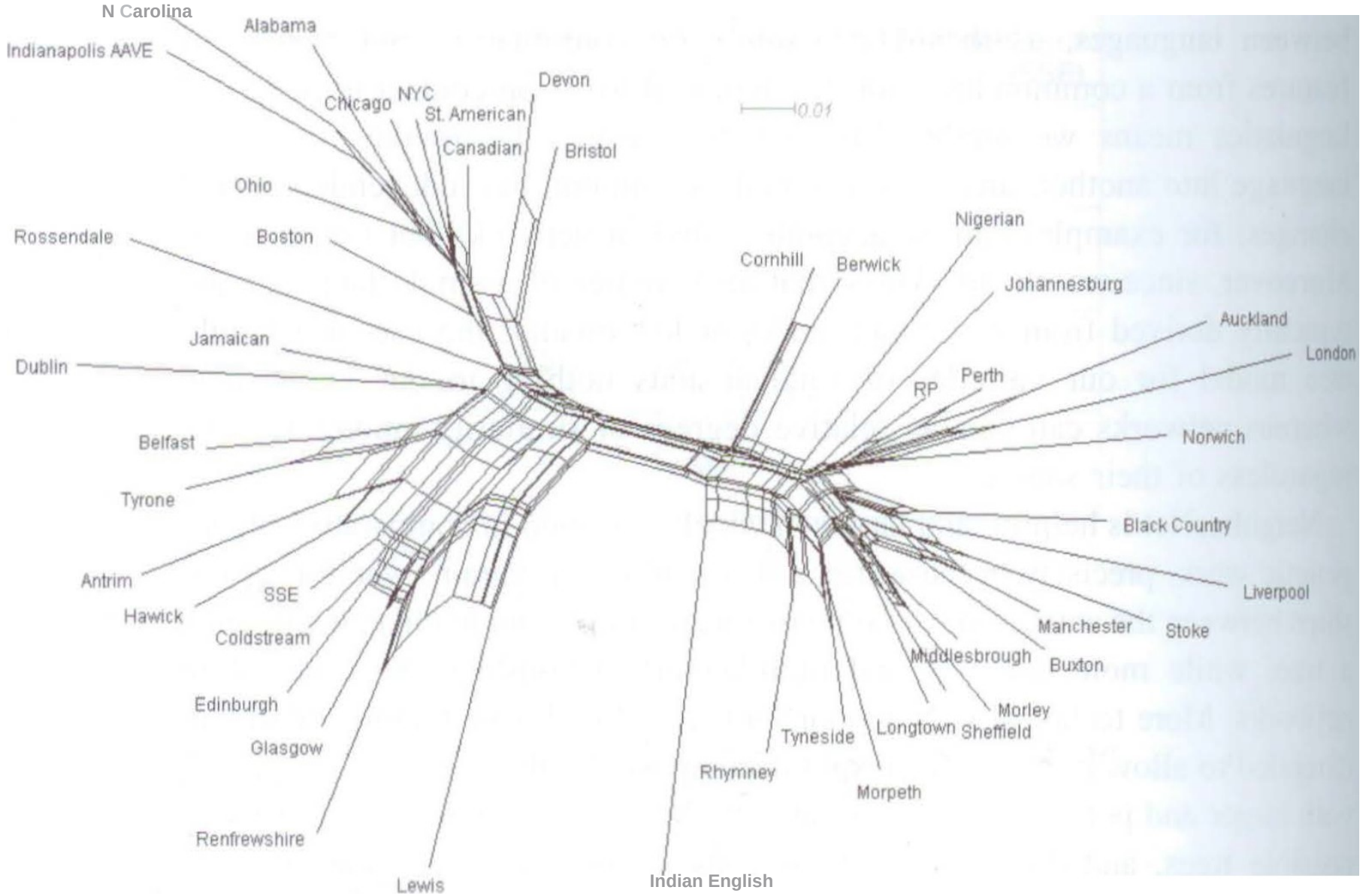


Figure 22.2. *NeighborNet of all modern English Typical varieties*

McMahon/Maguire in Krug/Schlüter 2013: 426

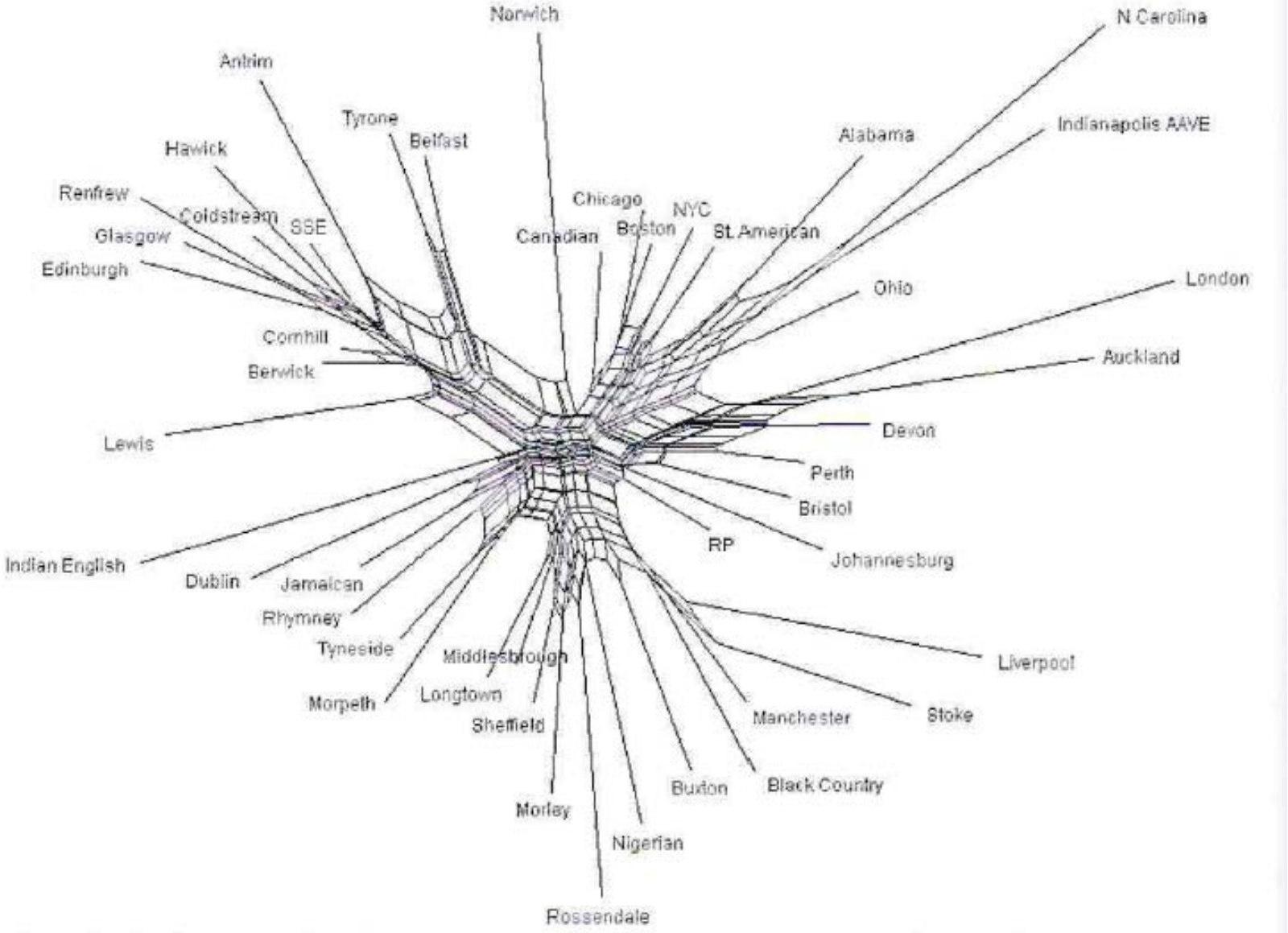


Figure 22.4. *All modern English Typical varieties; items with postvocalic /r/ excluded*

McMahon/Maguire in Krug/Schlüter 2013: 428

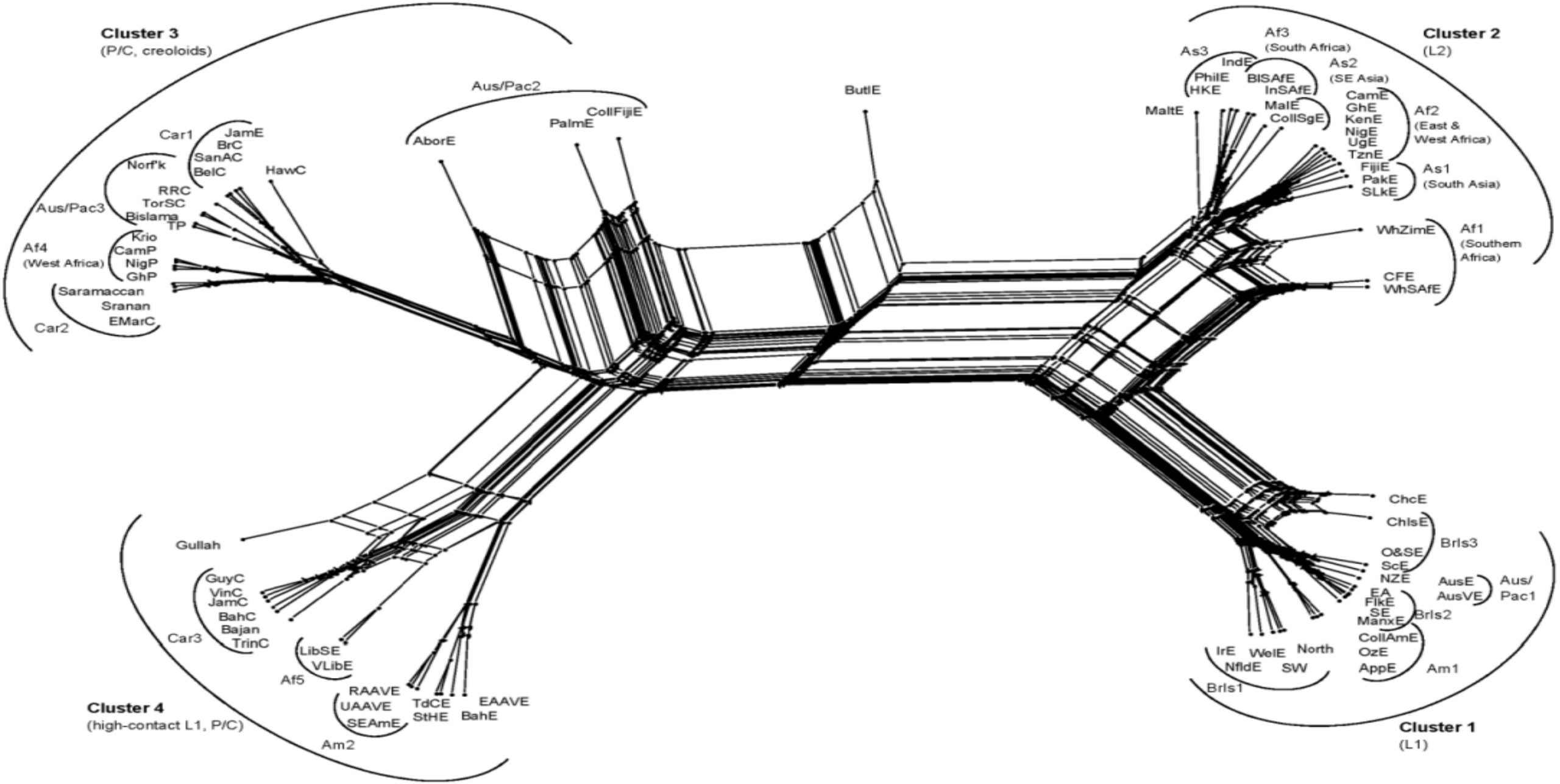


Figure 1: WAVE_Global network for the entire feature set (235)

Analysing aggregated linguistic data

- Aggregate Data Analysis (AGA), also known as *Data Synthesis*, *Mass Data Analysis* or, especially in biology, as (*Numerical*) *Taxonomy*: concerned with the joint analysis of multiple characteristics

- AGA is a methodical cornerstone in e.g.

-> biology: species categorization

-> economics: assessing macroeconomic changes

-> marketing research

-> consumer creditworthiness modeling

Szmrecsanyi in Krug/Schlüter 2013: 433

Aggregating part-of-speech frequencies:
analyticity vs. syntheticity in British English text types
(or: registers, genres)

1. Spoken texts are significantly more analytic than written texts. The average spoken text exhibits 50 more analytic markers per 1,000 words of running text than the typical written text.

2. Written texts are significantly more synthetic than spoken texts, in that the former exhibit, on average, approximately 30 more synthetic markers per 1,000 words of running text than the latter.

3. Variability among written texts is more sizable than variability among spoken texts: in Figure 23.1, the cloud embedding spoken genres is substantially more compact than its written counterpart.

Szmrecsanyi in Krug/Schlüter 2013: 439

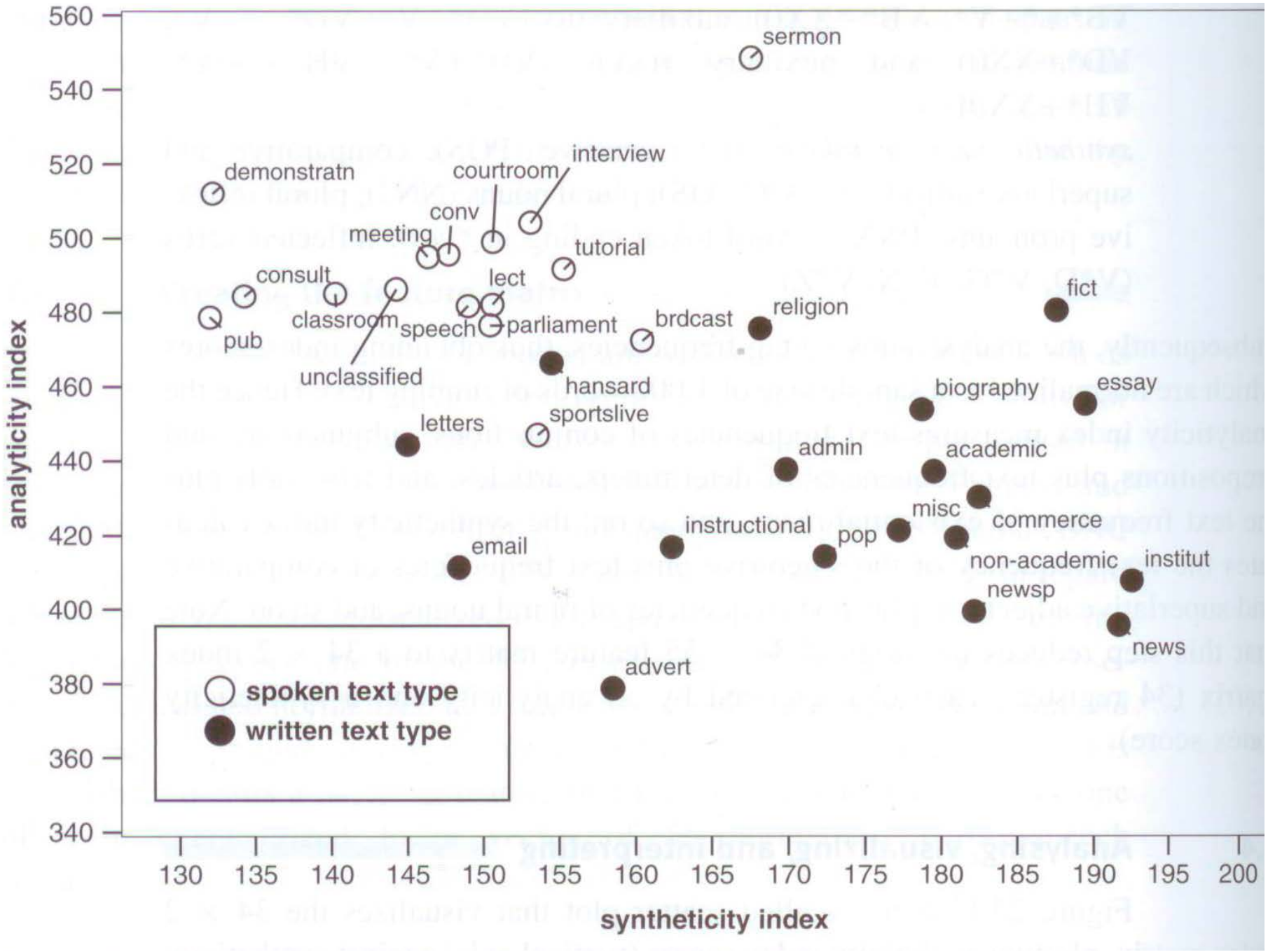


Figure 23. 1. *Visualization of the 34* x *2 index matrix: BNC macro registers - analyticity by syntheticity (in index points,* ptw*). Black dots indicate written registers, white dots indicate spoken registers*

Szmrecsanyi in Krug/Schlüter 2013: 439

Aggregation even in language typology

- Szmrecsanyi, Benedikt & Bernhard Wälchli (eds.) 2014. ***Aggregating dialectology, typology, and register analysis: Linguistic variation in text and speech***. Berlin: de Gruyter.
- builds on a workshop held at (FRIAS) in February 2011
- the volume deals with the quantitative analysis of linguistic variation from text via features to aggregation – what we take the liberty to call the text-feature-aggregation pipeline in variation studies. [….]
- some typological studies based on the analysis of parallel corpora
- “doculect”: A doculect is any documented language variety, be it as raw data (e.g., a sound file), primary data (e.g., a transcribed text or wordlist), or secondary data (e.g., a glossed text or a grammatical description) of whatever size. (Wälchli/Cysouw 2012)

(Wälchli/Szmrecsanyi 2014: 6)

2.3 Probabilistic linguistics

- underlying assumption:
  Knowledge of language is sensitive to distributions of previous language experiences. Whenever an expression is processed, it is seen as a piece of evidence that affects the probability distribution of language experiences. New expressions are constructed by probabilistically generalizing over previous expressions.

- Probabilistic linguistics *enriches* linguistic theory with statistics by defining probabilities over complex linguistic entities, from phonological to semantic representations. Probabilistic linguistics does therefore not abandon all the progress made by linguistics thus far; on the contrary, it integrates this knowledge with a probabilistic perspective.

Bod 2010: 634, 636

Frequency effects (note: GRK 1624 Freiburg)

- One of the strongest arguments in favor of using probabilities comes from the wealth of **frequency effects** that pervade gradience in language (Bybee and Hopper 2001,…). Frequent words and constructions are learned faster than infrequent ones (Goodman et al. 2008). Frequent combinations of phonemes, morphemes and structures are perceived as more grammatical, or well-formed, than infrequent combinations (Coleman and Pierrehumbert 1997; Manning 2003).

- We can best model these effects by making explicit the link between frequency and probability: probability theory not only provides tools to working with the frequency of events but also with the frequency of combinations of events.
  -> often-used probability model: logistic regression

Bod 2010: 634

Cognitive claim

Probabilistic linguistics makes a cognitive claim:

- Probabilities are operative in acquisition, perception, production, language change, language variation, language universals, and more. **All evidence points to a probabilistic language faculty.**

Bod 2010: 634

3. From Corpus to Cognition?

(or: Do corpora mirror psychological reality?)

Cognitive Corpus Linguistics

Abstract (FRIAS Conference 2008)

- Within cognitive linguistics, there is an increasing awareness that the study of linguistic phenomena needs to be grounded in usage. Ideally, research in cognitive linguistics should be based on authentic language use, its results should be replicable, and its claims falsifiable. Consequently, more and more studies now turn to corpora as a source of data. While corpus-based methodologies have increased in sophistication, the use of corpus data is also associated with a number of unresolved problems. **The study of cognition through off-line linguistic data is, arguably, indirect**, even if such data fulfils desirable qualities such as being natural, representative and plentiful. Several topics in this context stand out as particularly pressing issues.

Arppe/Gilquin/Glynn/Hilpert/Zeschel 2010: 1

Corpus frequencies and psychological reality

[….] Cognitive linguistics posits strong links between cognition and actual usage events, so that '[a]n event [. . .] becomes more and more deeply entrenched through continued repetition' (Langacker, 1987: 100). This assumption entails that corpora, which contain information about what is likely to be repeated or not in language, should make it possible to identify those items that have a special status in the mind. However, this assumption is mainly that – an assumption, and linguists have made relatively few efforts hitherto to test the cognitive reality of corpora.

- **Corpora are no shortcut to cognition: 2 case studies**

Arppe/Gilquin/Glynn/Hilpert/Zeschel 2010: 8f

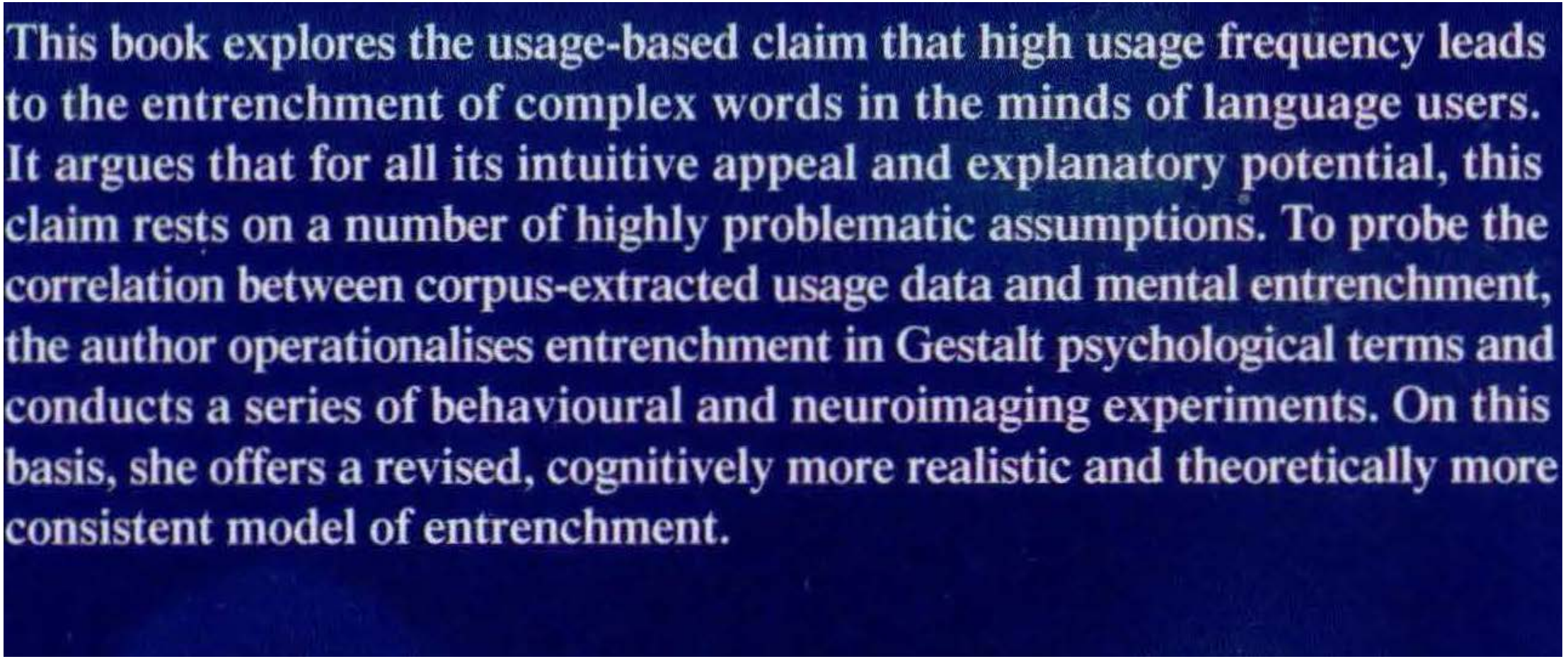
3.1 Entrenchment

- The usage-based hypothesis assumes that there is a connection between the usage frequency of linguistic structures and their degree of cognitive routinisation, or likelihood to be memorised/stored (entrenchment). (…) (Langacker, 2000).

  Putting this hypothesis to test:

- Alice Blumenthal-Dramé (2012), *Entrenchment in Usage-Based Theories: What Corpus Data Do and Do not Reveal about the Mind*. Boston/Berlin: De Gruyter Mouton.

Blumenthal-Dramé 2012: cover blurb



- Overall, based on the test case of entrenchment, this work has argued for a weak version of the corpus-to-cognition principle:

  If a whole range of caveats is heeded, certain corpus-extracted variables may, to some extent, be used as a yardstick for entrenchment in the brain of an average language user.

- […], which may be rather weakly representative of actual brains. Although this is not in itself objectionable […] wide-scope generalizations of this kind will necessarily miss important generalizations at a higher level of granularity […].

Blumenthal-Dramé 2012: 205

- Moreover,…, there is good reason to expect systematic differences between subjects with more or less holistic cognitive styles, and

statistical models which handle these differences [simply] as noise would be highly distorting with regard to the mind of real language users.

Blumenthal-Dramé 2012: 205

3.2 Analyticity vs. Syntheticity

Gero Kunter (2017), *Processing Complexity and the Alternation Between Analytic and Synthetic Forms in English* (Habilitation thesis, U Düsseldorf, Germany)

- Corpus-based studies on grammatical variation often evoke (or sometimes even make explicit claims concerning) psycho-linguistic processing (e.g. by linking structural complexity/ simplicity to cognitive/processing complexity/simplicity).

- Frequently, these corpus-based studies follow **an indirect line of argumentation**: if a particular grammatical variant is found to co-occur with structures that have relatively high linguistic complexity, this high linguistic complexity is often equated to an increased processing complexity. In a second step, the correlational relation between the grammatical variant and its co-occuring structures is interpreted as a causal relation, thus arguing that the occurrence of a particular grammatical variant does not only co-occur with another structure, but is caused by the higher processing complexity of that structure.

(Kunter 2017: 4)

- In other words: corpus studies are in essence observational and perfect for identifying correlations between variable features; moreover, on the basis of these correlations it is of course possible, and useful to formulate hypotheses on causal relations. HOWEVER, these hypotheses, in turn, must be tested independently, namely by way of experimental (psycholinguistic) studies.

  → corpus data reveal no more than indirect evidence for cognitive processes; only experimental data offer the chance to reveal direct evidence

  → Gerd Antes (U Freiburg, *FAZ*, 2 January, 2018): From the era of causality → the era of correlations; great danger: *spurious correlations* = two phenomena are highly correlated, but deducing from this a causal relation would be utter nonsense

Overall result of Kunter‘s dual-approach study

- “Thus, the two production experiments and the two corpus studies speak against a general compensatory mechanism that can account for both the comparative alternation and of the possessive alternation. There is partial evidence in favour of the *more* support hypothesis in that speakers show a significantly higher tendency to use the analytic comparative with adjectives that are cognitively more complex. However, this effect is one factor alongside other determinants, … The empirical findings provide little reason to assume that processing complexity plays a similar role in the possessive alternation.”

(Kunter 2017: 223)

Competing motivations of speaker and hearer

- Another interesting overall observation AND reminder – speaker and hearer may well have competing motivations:

> “This book adds another facet to the explanation as to why English and other languages show cases of alternation between synthetic and analytic variants: both forms exist because they are preferred either by the speaker or by the addressee.
>
> **Synthetic comparatives have been found to be relatively easy to process by listeners, but analytic comparatives may be preferred by speakers**.”

(Kunter 2017: 226)

3.3 Interim summary & recommendation

- Follow the advice by Arppe et al. (2010: 1): they point to three simple things corpus-based cognitive linguists should heed, the most important of which (to me) is the following one:

  “First, a certain degree of humility could not hurt.”

  Recall **Kunter**: corpus data are purely observational and reveal no more than **indirect** evidence for cognitive processes; only experimental data offer the chance to reveal direct evidence
  -> need for multi-method design

Arppe/Gilquin/Glynn/Hilpert/Zeschel 2010: 1

4. Major Caveats

- **don't** see language or some variety of a given language exclusively through the lens of (available relevant) corpora, as this may artificially narrow (or block) the object of study
- **don’t** take induction (e.g. in corpus-driven research) and statistics to be objective *per se*
- **don’t** take the search for correlations to be an end in itself
- **don’t** confuse correlations with causes
- **do** everything that is necessary (!) for achieving a maximum of methodological transparency, rigour, statistical significance, robustness, reproducibility, falsifiability and, ultimately, explanatory power and mileage for linguistic theory-building (e.g. no statistical cherry-picking), **but**

- **don't** do everything that is statistically possible just because you can do it (heaping minor, or worse: irrelevant, detail on minor/irrelevant detail), even less as a remedy for an imperfect data set or inconclusive data analyses

  → **don’t** multiply statistical testing beyond necessity
  → **no** statistics-driven research! (statistic machinery must not determine the research question), i.e.
  → **don't** let the tail wag the dog

- **do first** formulate intelligent research questions and a solid research- and theory-grounded set of hypotheses, which can **then** be statistically tested/falsified, yet

- **don’t** take statistical compatibility with a given hypothesis immediately as (sufficient) proof

- **don't** commit the “from-corpus-to-cognition fallacy” and conduct, as is appropriate for the research question, experimental studies alongside corpus studies (-> multi-method design)

- however powerful and promising the corpus revolution and quantitative turn may be (or be felt to be): **don’t** forget the rich inventory of theories and **(largely qualitative)** methods which (schools of) linguists have developed and refined over many decades for the analysis of natural language and communication (-> a caveat which applies also vice versa)

- **never** forget the human factor behind everything in communication and language: the intentions, needs, constraints of natural language users in spontaneous verbal interaction

5. A quantitative crisis in linguistics?

- Workshop at ISLE 5 (London, 17-20 July, 2018):
  Sönning/Werner (Bamberg): “The ‘quantitative crisis’, cumulative science, and English linguistics”

-> some focal problems identified in the broader discourse:

- overreliance on a single influential theory determining an entire research paradigm
- non-reproducibility of studies
- high rates of false-positive findings in published research
- lack of transparency as regards methodology and analysis
- negligence of replication studies as “unoriginal” (and unprestigious)

- overreliance on a single influential theory determining an entire research paradigm? **-> NO**

- with regard to the four methodological problems:

  all relevant in linguistics, too, but there is reason for optimism

  → not least due to awareness raising at a fairly early point and

  → the readiness in the publishing of linguistic research to go by the principles of open science, e.g.

    - accessibility of data & analyses
    - reproducibility of studies and statistical tests
    - no fear of publishing “negative” results

6. Conclusion

- Overall, the QT in linguistics has been a largely positive development. It has many strengths and great potential **always provided** corpus analyses and statistical techniques are selected and conducted/applied cautiously and in a highly reflected manner,

- heeding constraints, challenges and dangers, such as
  - the limits of what corpora can tell us about cognition
  - the risks of simplistic / naïve statistical analysis: cherry-picking, confusing correlations (at worst: spurious correlations) with causes

…but there is still some way to go

- The crucial point and task for linguists committed to the QT:
  „to boldly go where the others already are“

- In the concert of the quantitative sciences, linguistics is still a (somewhat little naive) newcomer, but if it wants to be taken seriously it needs to stand up to the rigorous standards of these sciences -> this is still a quite hard and long way to go

- Besides basic and advanced statistical training as part of degree and doctoral training programmes, besides statistics-savvy linguists, the members of each and every linguistics department should also have the possibility of consulting with professional (ideally linguistics-savvy) statisticians!

- If these conditions are fulfilled, linguistics will become an even more respected showcase of the **Digital Humanities**, and may truly succeed in **bridging the disciplinary boundaries** to the STEM sciences, especially to the behavioral and neuro-sciences.

- Finally: the QT in linguistics has NOT been to the detriment of qualitative approaches; rather: a productive relationship characterized by mutual respect, reinforcement and benefit

THANK YOU!

Lecture on February 1st, 2018

Are the Digital Humanities per se quantitative in nature?

John Nerbonne (U Groningen & U Freiburg)

All lectures are available as a video podcast at
www.frias.uni-freiburg.de/en/media-library

FRIAS

Next lecture on January 18th, 2018

Life in the marine realm – counting microbes… and what else?

Wolfgang Hess (U Freiburg)

All lectures are available as a video podcast at
www.frias.uni-freiburg.de/en/media-library

UNI FREIBURG

References

- Arppe, Antti, Gaëtanelle Gilquin, Dylan Glynn, Martin Hilpert and Arne Zeschel. 2010. Cognitive corpus linguistics: Five points of debate on current theory and methodology. *Corpora* 5(1), 1–27.
- Biber, Douglas. 2010. Corpus-Based and Corpus-Driven Analyses of Language Variation and Use. In: Heine/Narrog, eds. 159-191.
- Blumenthal-Dramé, Alice. 2012. *Entrenchment in Usage-Based Theories: What Corpus Data Do and Do not Reveal about the Mind*. Boston/Berlin: De Gruyter Mouton.
- Bod, Rens. 2010. Probabilistic linguistics. In: Heine/Narrog, eds. 633-662
- Divjak, Dagmar, Natalia Levshina and Jane Klavan. Cognitive Linguistics: Looking back, looking forward. *Cognitive Linguistics* 27(4): 447-463.
- Gries, Stefan Th. 2015. Some Current Quantitative Problems in Corpus Linguistics and a Sketch of Some Solutions. *Language & Linguistics* 16: 93-117.

- Gries, Stefan Th. 2013. Elementary statistical testing with R. In: Krug/Schlüter, eds. 361-381.
- Heine, Bernd and Heiko Narrog, eds. 2010. *The Oxford Handbook of Linguistics Analysis.* Oxford: Oxford University Press.
- Kortmann, Bernd and Kerstin Lunkenheimer, eds. 2013. *The Electronic World Atlas of Varieties of English* 2.0. [eWAVE 2.0]. Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://ewave-atlas.org>
- Kortmann, Bernd and Kerstin Lunkenheimer, eds. 2012. *The Mouton World Atlas of Variation in English.* Berlin/New York: De Gruyter Mouton.
- Krug, Manfred and Julia Schlüter, eds. 2013. *Research Methods in Language Variation and Change*. Cambridge: Cambridge University Press.
- Kunter, Gero. 2017. Processing complexity and the alternation between analytic and synthetic forms in English. Postdoctoral thesis, University of Düsseldorf.

- Mannila, Heikki, Terttu Nevalainen and Helena Raumolin-Brunberg. 2013. Quantifying variation and estimating the effects of sample size on the frequencies of linguistic variables. In: Krug/Schlüter, eds. 337-360.
- McMahon, April and Warren Maguire. 2013. Computing linguistic distances between varieties. In: Krug/Schlüter, eds. 421-432.
- Szmrecsanyi, Benedikt and Bernhard Wälchli, eds. 2014. *Aggregating dialectology, typology, and register analysis: Linguistic variation in text and speech*. Berlin: de Gruyter
- Szmrecsanyi, Benedikt. 2013. Analyzing aggregated linguistic data. In: Krug/Schlüter, eds. 433-455.
- Wälchli, Bernhard and Benedikt Szmrecsanyi. 2014. Introduction: The text-feature-aggregation pipeline in variation studies. In: Szmrecsanyi/Wälchli, eds. 1-25.

References

- Antes, Gerd. Die Medizin im Datenrausch. *FAZ* Nr. 1/2. Januar 2018/S. 9.
- Arppe, Antti, Gaëtanelle Gilquin, Dylan Glynn, Martin Hilpert and Arne Zeschel. 2010. Cognitive corpus linguistics: Five points of debate on current theory and methodology. *Corpora* 5(1), 1–27.
- Biber, Douglas. 2010. Corpus-Based and Corpus-Driven Analyses of Language Variation and Use. In: Heine/Narrog, eds. 159-191.
- Blumenthal-Dramé, Alice. 2012. *Entrenchment in Usage-Based Theories: What Corpus Data Do and Do not Reveal about the Mind*. Boston/Berlin: De Gruyter Mouton.
- Bod, Rens. 2010. Probabilistic linguistics. In: Heine/Narrog, eds. 633-662
- Divjak, Dagmar, Natalia Levshina and Jane Klavan. Cognitive Linguistics: Looking back, looking forward. *Cognitive Linguistics* 27(4): 447-463.
- Gries, Stefan Th. 2015. Some Current Quantitative Problems in Corpus Linguistics and a Sketch of Some Solutions. *Language & Linguistics* 16: 93-117.

- Gries, Stefan Th. 2013. Elementary statistical testing with R. In: Krug/Schlüter, eds. 361-381.
- Heine, Bernd and Heiko Narrog, eds. 2010. *The Oxford Handbook of Linguistics Analysis.* Oxford: Oxford University Press.
- Kortmann, Bernd and Kerstin Lunkenheimer, eds. 2012. *The Mouton World Atlas of Variation in English.* Berlin/New York: De Gruyter Mouton.
- Krug, Manfred and Julia Schlüter, eds. 2013. *Research Methods in Language Variation and Change*. Cambridge: Cambridge University Press.
- Kunter, Gero. 2017. Processing complexity and the alternation between analytic and synthetic forms in English. Postdoctoral thesis, University of Düsseldorf.
- Mannila, Heikki, Terttu Nevalainen and Helena Raumolin-Brunberg. 2013. Quantifying variation and estimating the effects of sample size on the frequencies of linguistic variables. In: Krug/Schlüter, eds. 337-360.

- McMahon, April and Warren Maguire. 2013. Computing linguistic distances between varieties. In: Krug/Schlüter, eds. 421-432.
- Szmrecsanyi, Benedikt and Bernhard Wälchli, eds. 2014. *Aggregating dialectology, typology, and register analysis: Linguistic variation in text and speech*. Berlin: de Gruyter
- Szmrecsanyi, Benedikt. 2013. Analyzing aggregated linguistic data. In: Krug/Schlüter, eds. 433-455.
- Wälchli, Bernhard and Benedikt Szmrecsanyi. 2014. Introduction: The text-feature-aggregation pipeline in variation studies. In: Szmrecsanyi/Wälchli, eds. 1-25.