Bernd Kortmann*

Reflecting on the quantitative turn in linguistics

https://doi.org/10.1515/ling-2019-0046
Received December 29, 2019; accepted January 1, 2021; published online July 21, 2021

Abstract: Linguistics, English linguistics in particular, has witnessed a remarkable quantitative turn since the 1990s and the early 2000s. It was a turn both in scale and in quality, a turn concerning the degree (including the degree of sophistication) to which quantitative empirical studies, statistical techniques, and statistical modelling have come to be used and determine linguistic research. Which role have corpus linguistics and probabilistic linguistics, including usage-based approaches, played in this development? Has this turn been to the detriment of qualitative methods, or even of linguistic theorizing in general? Has linguistics reached the point of a "quantitative crisis", or is it still a discipline characterized by a healthy equilibrium, if not mutual reinforcement, of quantitative and qualitative approaches? What are, or should be, major repercussions of the strong quantitative turn for the publication system of (English) linguistics? These are the major overarching questions underlying the reflections offered in this opinion paper.

Keywords: corpuslinguistics; entrenchment; multi-method design; probabilistic linguistics; processing complexity

1 Introduction

Linguistics, English linguistics in particular,[1] has witnessed a remarkable quantitative turn over the last two decades. Has this turn been to the detriment of qualitative methods, or even of linguistic theorizing in general? Has linguistics reached the point of a "quantitative crisis", or is it still a discipline characterized by

1 This article emerged from a paper given at the workshop on the "quantitative crisis", cumulative science, and English linguistics", which was part of the ISLE 5 conference (International Society for the Linguistics of English) at University College, London, in July 2018. This explains the focus of this article on English, even though the rise of quantitative methods and, especially, the issues discussed here are equally important in linguistics, in general (cf. also, for example, Janda 2013, 2017).

***Corresponding author: Bernd Kortmann**, University of Freiburg, Freiburg, Germany, E-mail: bernd.kortmann@anglistik.uni-freiburg.de

a healthy equilibrium, if not mutual reinforcement, of quantitative and qualitative approaches? What are, or should be, major repercussions of the strong quantitative turn for the publication system of (English) linguistics? These are the major overarching questions underlying the reflections offered here, in what may perhaps most appropriately be characterized as an opinion paper. One general conclusion will be that, apart from the need for an increased level of sophistication and problem-awareness in choosing, applying and interpreting statistical methods in linguistic research, the natural next step for a strongly quantitatively oriented linguistics needs to be the increasing adoption of multi-method approaches. Overall, this paper is fueled by the optimism that linguistics stands a real chance to retain all its traditional strengths and, at the same time, to develop into an even more respected showcase of the Digital Humanities, capable of bridging the disciplinary boundaries especially to the behavioral and neurosciences. The all-important precondition for this is that linguistics continues to force itself onwards on the thorny path of standing up to the rigorous standards of quantification, statistical analysis and modelling of these highly developed quantitative sciences.

The quantitative turn addressed in this paper is the one we have been witnessing on a broad scale since the 1990s and the early 2000s. It was a turn both in scale and in quality, a turn concerning the degree (including the 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 like sociolinguistics, interactional linguistics, multimodal studies, or even in semantics and pragmatics, all of which typically, or even dominantly, work with qualitative methods,[2] some degree of quantification is increasingly expected (especially for journal publications). Overall then, the inclusion of quantitative methods has changed from a "nice to have" to a "better to have" status.

2 The quantitative turn

In Brian Joseph's final editorial as the editor of *Language*, one of the long-time flagship journals of the discipline, he comments on recent developments in the field, among them the following:[3]

2 The following are taken to be among the major characteristics of qualitative approaches and methods: the contextualization and context-embeddedness of their object of study; targeting the insider's point of view; aiming at interpretation, empathic understanding, and understanding the actors' perspectives, in general; a strong role of interviews and narratives; applying inductive, natural, non-interventionist methods, such as participant observation; the guiding idea that case studies reveal (subsets of) the complexities of life affecting individuals.

3 The trend Joseph identifies is roughly supported by the somewhat superficial longitudinal studies by Sampson (2005, 2013) on the proportion of empirical, i.e., "evidence-based" versus

> 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)

Joseph also identifies what I consider as the two major aspects, or even drivers, of this development:

> […] research papers are more experimentally based than ever before ('experimental' in the sense of pertaining to any sort of controlled investigation). Also, they are more corpus-based, with many studies using as primary data (not just as corroborating data) examples that have been gleaned from available corpora, including the Internet (2008: 687)

Whereas the experimental turn is increasingly, however much more slowly, making its inroads into linguistics (but see Sections 3 and 4), the rise of corpus linguistics has been vastly influential and has left a lasting imprint on the field. Thus Gries (2015: 93, 113), one of the first and still major protagonists of a statistics-heavy corpus linguistics, is right when stating that "corpus linguistics has been among the fastest-growing methodological disciplines in linguistics", "has become mainstream […]", and bears witness to the trend" that linguistics in general has become much more quantitative/statistical in nature […]: 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".

Putting Gries' statements to test, we took the journal *English Language and Linguistics* (short: ELL) as a case study. For one thing, English linguistics was the front-runner in developing corpora, annotation techniques, and appropriate corpus-linguistic methodology. For another thing, ELL was founded in 1997, i.e., just at a time which can be characterized as the beginning and formative years of the quantitative turn in linguistics. All 380 articles published since 1997 (the journal's inception) until 2019 were analyzed for their methodologies.[4] The articles were categorized into four overarching levels:

"intuition-based" articles (or articles neutral to this distinction) per volume in a sample of *Language* volumes between 1960 and 2011. For Sampson, articles "… which quoted observational support for at least two separate data items counted as evidence-based" (2005: 25), regardless whether these data items were observed in standard corpora, self-designed corpora or other sources, such as overheard conversations. From a low in 1970 (just below 30%), the percentage of "empirical" articles in *Language* has steadily increased, with almost all sampled volumes in the 2000s exceeding 70% (2005: 31, 2013: 286–287).

4 Squibs and reviews were excluded, as these are qualitative formats. It would be strange to find quantitative methods in a book review.

–	"no"	– The article uses no quantitative methods. It is purely qualitative (or uses non-statistical phonological methods such as measuring formant frequencies).
–	"freq"	– The article reports descriptive statistics such as mean, standard deviation, or relative frequency. The vast majority of these articles are corpus-based and report frequency counts.
–	"simple"	– The article compares tables or means without modeling. The majority of these articles use Chi-squared tests, but *t*-tests, Mann–Whitney tests, and similar methods are also included in this category.
–	"advanced"	– The article uses more complex statistical modeling. Most of these articles use either linear or logistic regression, though many recent articles also use mixed-effects models. Other methods included in this category include random forests, variable rules analysis (aka Varbrul or Goldvarb analyses), as well as a few unique modeling methods.

To account for the differing amounts of articles per year, all graphs show the percentage of each article type by year, proportional to the total articles in that year. Figure 1 shows a linear regression line fitted to the four categories listed above plotted against the percentage of the total articles by year. Note that there is a lot of variance by year, which cannot be seen in such a strictly linear model. Thus, Figure 2 shows the same categorization as a bar graph, again proportional to the total number of articles by year. Both figures are telling the same story, though: the proportion of qualitative articles per volume is strongly decreasing (from some 75% to less than 25%), the proportions of articles making use of simple and, even more pronounced, advanced statistics are steadily increasing (ending up between 25 and 30% each), while the proportion of "freq"-articles remains fairly stable (at a level slightly below the 25% mark).

In Figure 3, we collapse simple and advanced models into one category ("statistical"), and plot it against "frequency" and "qualitative" to see statistical measures compared to articles reporting frequency counts and qualitative works.

Finally, combining frequency together with the other quantitative methodologies provides an even broader overview. Figure 4 shows a linear regression line for qualitative articles compared to quantitative articles in general (i.e., including frequency counts) as a percentage of the total articles for that year. This figure says it all: the quantitative turn in linguistics as reflected by the longitudinal study of one of the leading journals in the field could not be more pronounced.[5]

5 A parallel trend Janda (2019: 8) identified for the journal *Cognitive Linguistics* between 1990 and 2017, with journal volumes between 1990 and 2007 exhibiting between 20 and 40% quantitative articles vis-á-vis 50–80% in the volumes from 2008 onwards.

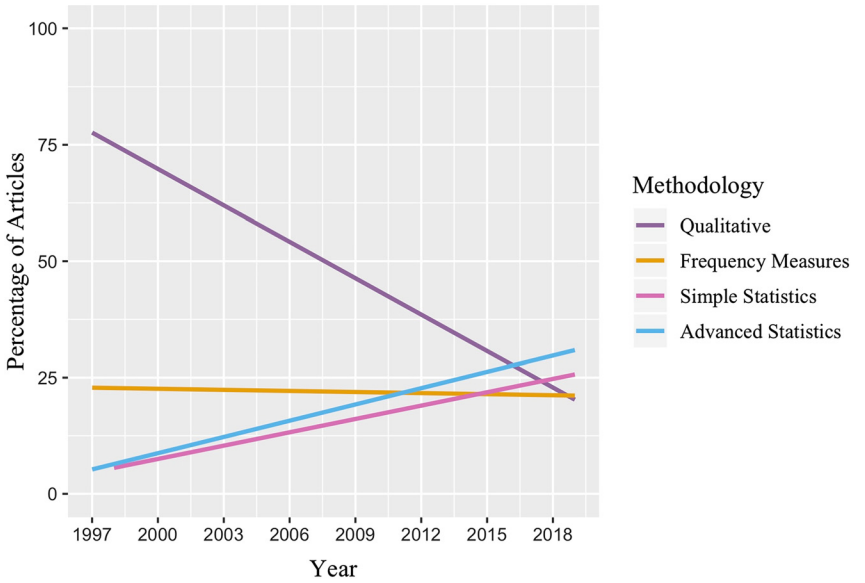

Figure 1: Linear regression lines for all four methodological categories as a percentage of the total articles by year in *English Language and Linguistics* from 1997 to 2019.

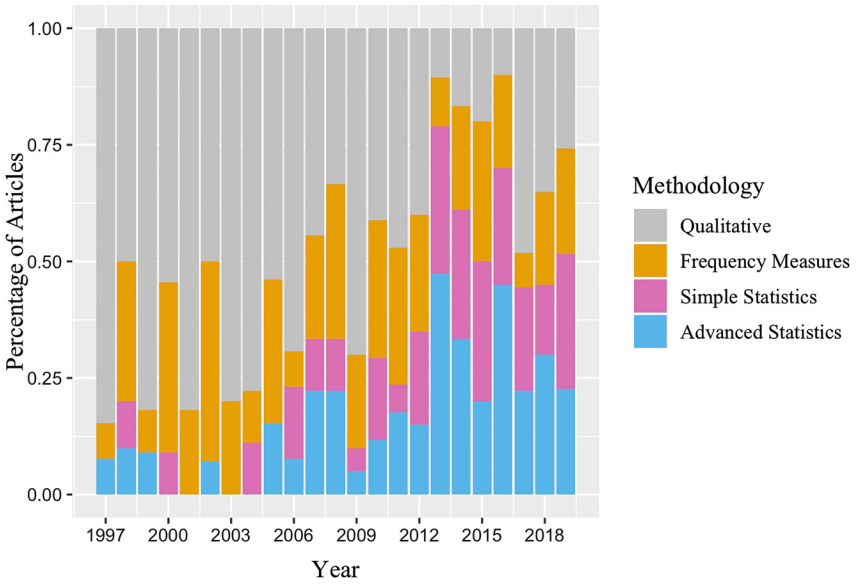

Figure 2: Bar graph for all four methodological categories as a percentage of the total articles by year in *English Language and Linguistics* from 1997 to 2019.

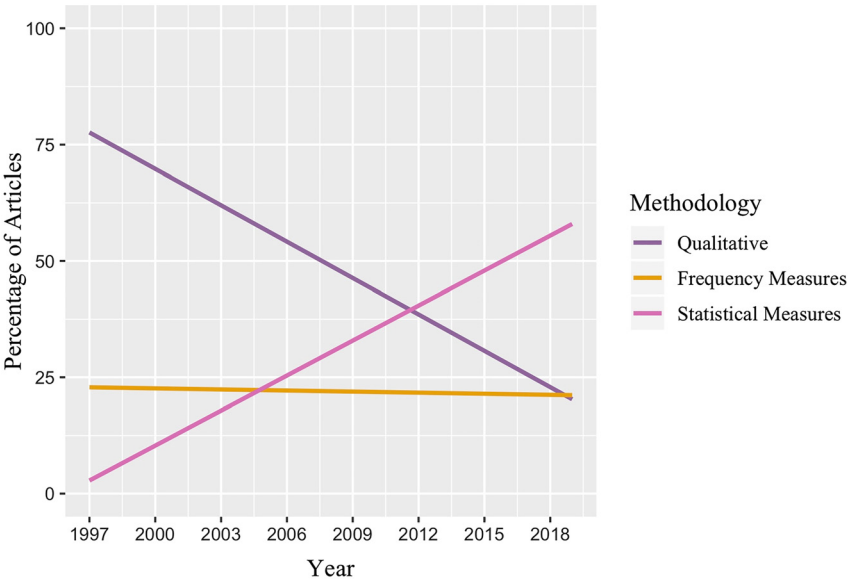

Figure 3: Linear regression lines for qualitative methodologies compared to frequency-based and statistical methodologies as a percentage of the total articles by year in *English Language and Linguistics* from 1997 to 2019.

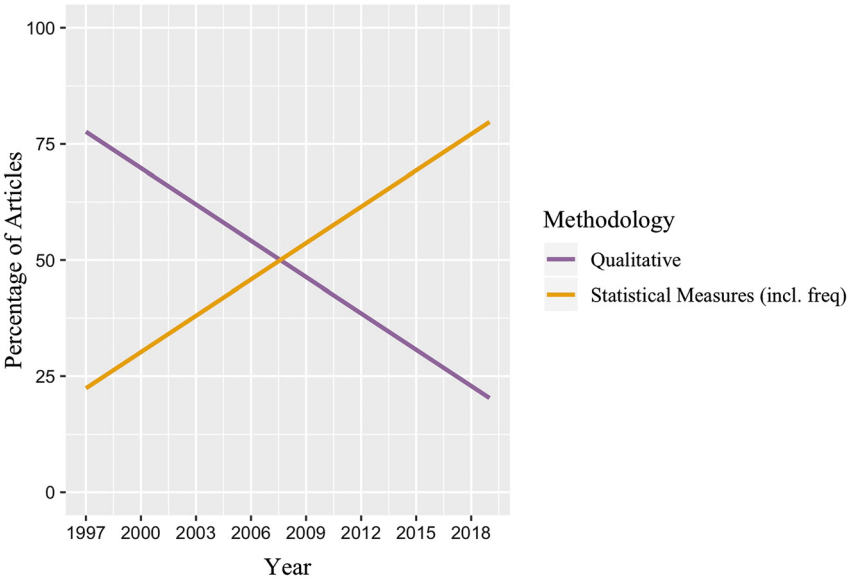

Figure 4: Linear regression lines for qualitative methodologies compared to quantitative methodologies as a percentage of the total articles by year in *English Language and Linguistics* from 1997 to 2019.

The rise, power, and pervasiveness of corpus linguistics can be read off easily from the following figures as well. At the time of writing this article (December 2019), the number of currently publicly available corpora just for English has passed the threshold of 100 (excluding subcorpora of the International Corpus of English (ICE) or the International Corpus of Learner English (ICLE), comprising some 20 billion words. To a substantial part (some 12 billion words), this includes web-based corpora (e.g., NOW – News on the Web, GloWbE – Corpus of Global Web-based English, the Hansard Corpus, the Wikipedia Corpus), which are growing on a daily basis. Add to this, for example, about 2,000 billion words in Google Books or data that can constantly be compiled from the internet, notably from social networks (e.g., Twitter). Further striking indicators of the corpus-linguistic and quantitative turn in linguistics are the following (see Appendix): 16 introductions to Corpus Linguistics have been published since the very first one (by McEnery and Wilson) in 1996, five relevant handbooks since 2009, five book series since 1998, six journals since 2002, and nine introductions to statistics for linguist(ic)s since 1998, four of them since 2015 (see Appendix). Moreover, methodological textbooks have standardly come to include sections on corpus use (e.g., Krug and Schlüter 2013).

This corpus linguistic turn, spearheaded by English linguistics, is largely considered a development for the better. Three powerful reasons are given, again, by Gries (2013: 361–362):

> 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.
>
> 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; […].

In the following, I will first look a bit more closely into exactly this probabilistic theoretical perspective which is so strongly fueling the quantitative turn in linguistics (Section 3) and the related question whether corpora do indeed mirror psychological reality (Section 4). This will be followed by some major caveats, formulated in the form of *do*s and *don't*s, concerning the quantitative turn and the

use of heavy quantitative machinery (Section 5) and, in conclusion, by brief answers to the three overarching questions raised at the outset (Section 6).

3 Probabilistic linguistics

Bod (2010: 634) aptly summarizes the crucial underlying assumption of probabilistic linguistics as follows:

> 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.

One of the strongest arguments for using probabilities comes from the study of frequency effects in language. Frequent words and constructions have been shown to be learned faster than infrequent ones (e.g., Ellis 2002; Goodman et al. 2008), just as frequent combinations of phonemes, morphemes and structures could be shown to be perceived as more grammatical, or well-formed, than infrequent combinations (cf. Coleman and Pierrehumbert 1997; Manning 2003). Frequency effects have also been shown to pervade gradience in language (Bybee and Hopper 2001).[6]

Now what is important, not only when exploring frequency effects, is that probabilistic linguistics makes a strong cognitive claim, namely that our language faculty is probabilistic, probabilities being "operative in acquisition, perception, production, language change, language variation, language universals, and more" (Bod 2010: 634; cf. similarly Bresnan and Ford 2010: 205).

For language acquisition this cognitive claim is worked out in more detail by Gries and Ellis (2015: 241):

> [l]anguage learners do not consciously tally […] corpus-based statistics. The frequency tuning under consideration here is computed by the learner's system automatically during language usage. The statistics are implicitly learned and implicitly stored […]; learners do not have conscious access to them. Nevertheless, every moment of language cognition is informed by these data […].

6 In a nine-year concerted effort, frequency effects in language variation, language change, and language acquisition have been explored from a variety of perspectives in the Freiburg Graduate Training College between 2009 and 2018 (GRK 1624; http://frequenz.uni-freiburg.de/index.php?id=320&language=en).

Indeed, since the early 2000s, linguistics has seen "a growing number of experimental studies confirming principled correlations between statistical generalizations over corpus data and subjects' experimental behaviors at various levels of language description" (Blumenthal-Dramé 2016). And yet there is one fundamental question, and enormous challenge, that the cognitive claim of probabilistic linguistics (and usage-based linguistics, in general) needs to answer: Is their model of language processing and language storage realistic? According to this model, called *exemplar store model* by Blumenthal-Dramé (2016), the mind is conceived as a multi-dimensional memory space that is permanently switched on like a multi-channel recording system. Through this system, the entirety of experience (down to the most fine-grained linguistic and non-linguistic facets of the input) is channeled, indexed and stored away in memory, and language users compute statistical generalizations over usage data. Thereby the system is continuously updated (cf. Docherty and Foulkes 2014: 51). The question is whether the "exemplar store" really models what goes on in people's minds. Moreover, can this question, which is asked by an increasing number of people, be answered by using off-line linguistic, i.e., corpus, data? Do corpora mirror psychological reality, at all? This set of questions will be addressed, however selective and briefly, in the following section.

4 From corpus to cognition?

An early statement concerning the strong links between cognition and recurrent usage events was formulated by one of the pioneers of cognitive linguistics, Ronald Langacker (1987: 100): "[a]n event […] becomes more and more deeply entrenched through continued repetition". Corpora enter the picture as, according to Arppe et al. (2010: 8–9), Langacker's "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". Such tests have, however, increasingly been conducted in recent years,[7] and thus two relevant case studies from English linguistics will briefly be presented in this section, one from word-formation, the other from inflectional morphology and morphosyntax. Both book-length studies, this is important to note, adopt a multi-method design, and both studies show that corpora offer *no* shortcut to cognition.

7 For an early collection of articles arguing for multi-method approaches in cognitive and corpus-linguistic research, see Schönefeld (2011).

4.1 Entrenchment

In the following, entrenchment will be understood as characterized by Zeschel in his account of the usage-based hypothesis, which "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)" (in Arppe et al. 2010: 10). It is this hypothesis that Blumenthal-Dramé puts to the test in her 2012 monograph, operationalizing entrenchment in gestalt psychological terms and conducting a series of behavioral and neuroimaging experiments on the processing and storage of complex words in English. By adopting a multimethod design, she wants to reach a better understanding of entrenchment and explore the crucial methodological question whether cognitively realistic insights into speakers' linguistic knowledge can be reached via the indirect method of making quantitative generalizations over offline linguistic data, i.e., corpus data. In a nutshell, the main results and conclusions to be drawn from them can be summarized as follows. First, it is at most a weak version of the corpus-to-cognition principle Blumenthal-Dramé (2012: 205) finds evidence for (my emphases):

> "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." And even then this "[…] *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 […]."

Moreover, Blumenthal-Dramé (2012: 205) alerts us to the much-overlooked fact that the assumption of an *average* language user may distort a far more complex reality. There may well exist systematic intersubjective differences in processing style, some language users employing more holistic, others less holistic cognitive styles, for example. Such differences must not be handled (or rather dismissed) as mere noise when interpreting the statistical models employed in the relevant studies.

4.2 Analyticity versus syntheticity

The second case study putting to test the cognitive reality of corpora (Kunter 2017)[8] is concerned with English inflectional morphology and morphosyntax, more exactly with the alternation between analytic and synthetic forms in comparison

8 A revised version of the postdoctoral thesis cited here is scheduled to appear in the De Gruyter Mouton series Topics in English Linguistics [TiEL] in 2022.

and possessive marking. The author's primary target of investigation is processing complexity, but similar to Blumenthal-Dramé in her study on entrenchment he starts out from the observation that frequency distributions and observations on structural complexity in corpora are often taken as indirect evidence for cognitive processing complexity. Kunter (2017: 4) states:

> 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-occurring 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.

In other words, corpus studies are in essence observational and perfect for identifying correlations between variable features, on the basis of which 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. In short: corpus data reveal no more than indirect evidence for cognitive processes; only experimental data offer the chance to reveal direct evidence. Applied to Kunter's processing complexity study of analytic versus synthetic coding in English comparison and possessive marking, he therefore finds the following:

> Consequently, the main claim of the analytic support hypothesis is supported almost exclusively by indirect evidence, the claim that speakers use the analytic variants as an immediate reaction to increases in the cognitive complexity of the environment, and not merely together with increases in cognitive complexity. In other words, the hypothesis states a causal relation between high processing effort and the choice of analytic forms, but the bulk of supporting evidence comes from observational data that is not ideal for identifying causal links. (Kunter 2017: 221)

This is why Kunter, like Blumenthal-Dramé in her study presented in Section 4.1, conducts production experiments alongside corpus studies, arriving at the following overall conclusion of this dual approach design (my emphases):

> 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)

Moreover, Kunter's findings on the alternation between synthetic and analytic forms also remind us of the often overlooked (or at least underrated) fact that speakers and hearers may well, and often clearly do, have competing motivations as regards language production and processing (recall Zipf's 1949 monograph on *Human behavior and the principle of least effort*) – again something that corpus analyses alone tell us little, if anything, about: "Synthetic comparatives have been found to be relatively easy to process by listeners, but analytic comparatives may be preferred by speakers" (Kunter 2017: 226).

What the above two case studies show very clearly is that only experimental data offer the chance to reveal direct evidence, whereas corpus data are purely observational, show us how people use language but reveal no more than indirect evidence for cognitive processes, and may thus at most give rise to the formulation of research hypotheses concerning cognitive processes. A multi-method design exploring both (offline) corpus and (online) experimental data is thus indispensable. This is what any corpus linguist embarking on a research question relating to language cognition should keep in mind. All this is also convincingly argued in Arppe et al.'s (2010) thought-provoking debate on cognitive corpus linguistics, which ends in three simple pieces of advice for corpus-based cognitive linguists. Of these, I will quote only the first and, in my view, most important one: "First, a certain degree of humility could not hurt" (Arppe et al. 2010: 21).

5 Major caveats

In this section the perspective taken on quantitative approaches in linguistics will be significantly broadened. Especially, but certainly not exclusively, for an (early) early-career readership, some major caveats concerning the quantitative turn and the use of heavy quantitative machinery will be offered, formulated in the form of *do*s and *don't*s. Some of these are linguistics-specific, some relate to principles of good (quantitative) science, in general. All of them have been formulated from the perspective of an experienced journal and book series editor who, additionally, has supervised and/or reviewed the theses and studies of numerous generations of Master students as well as doctoral and postdoctoral researchers.

– *In general*, *do* everything that is necessary for achieving a maximum of methodological transparency, rigor, statistical validity, robustness, reproducibility, falsifiability and, ultimately, for a maximum of mileage when putting to test existing, or even formulating novel, linguistic hypotheses. *Do first* formulate intelligent research questions and a solid research- and theory-grounded set of (potentially competing) hypotheses, which can in a next step be statistically tested/falsified.

- *Do*, in a next step, identify the most appropriate (corpus) data sets and (quantitative or qualitative) methods, notably statistical methods, for the purposes of your research question(s).
- *Don't* take statistical compatibility with a given hypothesis immediately as (sufficient) proof, nor as automatically implying incompatibility with a competing hypothesis or theory.
- *Do* distinguish between confirmatory and exploratory analyses (as done, for example, by Granlund et al. 2019). Confirmatory analyses would be ones of the type in the previous two points. But your findings may lead you to test novel hypotheses, to which you should be absolutely open, which may then serve as the basis for confirmatory statistical testing in follow-up studies, but
- *Don't* take exploratory analyses as a starting point, even less take them as an escape route or, in the worst case, as pseudo-confirmatory analyses of hypotheses formulated post hoc (i.e., when interpreting the results of your statistical analyses).
- *In general*, *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. More specifically, *don't* use a statistical model that is more complex than needed for your data, *don't* multiply statistical testing beyond necessity, and *don't* engage in statistics-driven research! Statistical machinery must not determine the research question.
- *Do* interpret your quantitative findings against previous linguistic research and theory building. Moreover:
- *Don't* confuse correlations with causes as this may, for example, lead you to fall in the circularity trap. Take frequency effects: the strong correlation between frequency of use and semantic bleaching often described in grammaticalization studies, for example, does not tell you anything *per se* about whether it is semantic bleaching that is the cause for a high usage frequency. It may just as well be the other way round or, as a third option, simply no more than an association, with no unidirectionality in either direction.
- *Don't* see language or some variety of a given language *exclusively* through the lens of (available relevant) corpora, as this may artificially narrow the object of study. For example, features salient for a particular variety may well be low in frequency. Theoretically interesting structural patterns or structural patterns currently undergoing change as observed by native speaker linguists, for example, may not figure at all (yet) in especially smaller-sized corpora like the sub-corpora of the International Corpus of English. Likewise the study of

subtle semantic phenomena and pragmatic phenomena involving knowledge about the entire sociological context (incl. language attitudes) and situational communicative (incl. multimodal) context, the interlocutors' intentions, etc., requires a broader take on data than many corpora can currently offer. Sampson (2005: 26) is therefore right when stating, "Intersubjectively observable evidence is evidence, wherever it is found. Corpus linguists compile and work with standard corpora because they are especially convenient data sources, but there is no reason to suggest that evidence has to occur in a recognized corpus to count as evidence".

- *Don't* commit the "from-corpus-to-cognition fallacy": rather conduct, as is appropriate for the research question, experimental studies alongside corpus studies (i.e., use a dual-approach or multi-method design).[9]
- *Do* make your data and statistical analyses accessible in the true open science spirit.[10]
- Finally, 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.

6 Conclusion: a quantitative crisis in (English) linguistics?

Overall, in my view at least, the quantitative turn in linguistics has been a largely positive development. It has many strengths and great potential *always provided*

9 Interestingly, the survey by Palacios Martínez (2020) on methods of data collection in 1,143 papers in 32 international and high impact linguistics journals published in 2017 reveals that only about 5% of the studies presented in these journal articles used multi-method approaches. The most widely used data collection methods were experimental studies (33%) and corpus-based studies (18%).

10 For overviews and brief descriptions of open access repositories see, for example, http://www.open-science-repository.com/ or https://open-access.net/en/information-on-open-access/repositories/. Specific repositories include the Open Science Foundation (https://osf.io), and for linguistics The Tromsø Repository of Language and Linguistics (TROLLing; https://dataverse.no/dataverse/trolling). It is also becoming increasingly common for linguistics journals (e.g., *Corpus Linguistics and Linguistic Theory*) to upload statistical code and output as supplementary materials to the article.

that corpus analyses and statistical techniques are selected and applied in a cautious, reflected manner, thus heeding constraints, challenges and dangers such as the limits of what corpora can tell us about cognition and, the risks of simplistic or naive statistical analyses (e.g., cherry-picking or confusing correlations with causes). The crucial point and task for linguists committed to the quantitative turn is "to boldly go where the others already are", as it were, without however repeating the mistakes the others have made along the way. Painful as it may be to realize, in the concert of the quantitative sciences, linguistics still is a (somewhat 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. What is needed *besides* basic and advanced statistical training as part of degree and doctoral training programs, and *besides* statistics-savvy linguists, is that the members of every linguistics department should also have the possibility of consulting with professional (ideally linguistics-savvy) statisticians! If all these conditions are fulfilled, linguistics will increasingly (co)operate on eye level with the behavioral and neuro-sciences. Another piece of good news that needs to be stressed by way of conclusion is that the quantitative turn in linguistics has *not* been to the detriment of qualitative approaches. Rather I see a productive relationship and interplay characterized by mutual respect, reinforcement, and benefit.

What then about the focal problems addressed in the ISLE 5 workshop on "The 'quantitative crisis', cumulative science, and English linguistics": Do quantitative approaches in linguistics suffer from methodological problems such as the non-replicability of studies, high rates of false-positive findings in published research, a lack of transparency as regards methodology and decisions in their analyses, or a negligence of replication studies as "unoriginal" (and nonprestigious)? There is no denying that linguistics in times of the quantitative turn may well suffer from these problems, too. Yet at the same time there is reason for optimism: the quantitative turn in linguistics is still fairly recent or, so I believe at least, recent enough for the research community to develop an awareness of these problems at a fairly early stage and to avoid the mistakes the established quantitative sciences have made along *their* way. This, in turn, together with the recent developments in formulating standards of linguistics publishing in line with open science policies and best practices (notably with regard to the accessibility of data and analyses, the transparency of data coding, methods and statistical tests, the reproducibility of studies and their findings, and, perhaps least pronounced at the current stage, the courage – both on the side of the authors and the journal editors – to publish

"negative" results) offers the realistic opportunity for solving these problems, or avoiding them altogether, within the next few years.

This does not mean, on a final note, that there is no room for improvement with regard to linguistics publishing. A cursory check of some thirty high-ranking linguistics journals regularly publishing state-of-the-art quantitative studies reveals that only a fraction have made a statistical consultant or consultant editor part of the editorial team, and only just about half of them draw at least on the services of members of the editorial or advisory board working quantitatively. Also only a fraction of journals currently have specific guidelines, or at least sections thereof, dedicated to quality standards for the statistics in manuscript submissions. Therefore, even if journal editors regularly involve independent peer reviewers with statistical expertise as one important element for judging the quality of manuscript submissions, in terms of quality assurance and meeting the highest scientific standards of the quantitative sciences most linguistics journals (and, correspondingly, book series) can, and before long need to, do better. At the same time, there are linguistics journals that have strongly changed their policies in accordance with best practice open science policies (see also Footnote 9). Perhaps most ambitious among these, but necessary in light of everything which has been said here on the strong quantitative turn in linguistics (mind the caveats in Section 5), is pre-registration. Thus the journal *Language Learning* has implemented Registered Reports as a new article type as recently as in 2018 (Marsden et al. 2018). This means that, instead of submitting their finished (draft) papers, authors register with journals (or book series, for that matter) the planned design, hypotheses, methods, data preparation and analyses *even before* actually conducting the relevant studies intended to be published with these journals. The journal editors, in turn, pass on the pre-registration information to peer reviewers as the basis for their (first) reports and subsequently make conditional acceptance offers. This new article type, or rather publication policy, offers a range of advantages, among them the following: It increases transparency, makes reviewing more effective in that the peer reviewers can identify methodological problems at a stage early enough to be remedied by the authors before embarking on the actual study, and not least it helps avoid publication bias towards the "new" and "significant".

Supplement

The dataset for this article can be found on Zenodo: https://doi.org/10.5281/zenodo.5102780.

Acknowledgments: The author wishes to acknowledge and thank the editors, an excellent anonymous reviewer, Natalia Levshina, and the members of his research team for their extremely valuable comments and suggestions.

Appendix

Striking indicators of the corpus-linguistic and quantitative turn in linguistics in the 2000s: Relevant introductions, handbooks, book series, and journals in chronological order[11]

Introductions to Corpus Linguistics

1996 McEnery, Tony & Andrew Wilson. 2001[1996], 2nd edn. *Corpus linguistics*. Edinburgh: Edinburgh University Press.
1998 Biber, Douglas, Susan Conrad & Randi Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.
2002 Meyer, Charles F. 2002. *English corpus linguistics: An introduction*. Cambridge: Cambridge University Press.
2004 Halliday, M. A. K., Wolfgang Teubert, Colin Yallop & Anna Čermáková. 2004. *Lexicology and corpus linguistics: An introduction*. London: Continuum.
2004 Semino, Elena & Mick Short. 2004. *Corpus stylistics: Speech, writing and thought presentation in a corpus of English writing*. New York: Routledge.
2006 McEnery, Tony, Richard Xiao & Yukio Tono. 2006. *Corpus-based language studies: An advanced resource book*. London: Routledge.
2007 Biber, Douglas, Ulla Connor & Thomas A. Upton. 2007. *Discourse on the move: Using corpus analysis to describe discourse structure*. Amsterdam & Philadelphia: John Benjamins.
2009 Mukherjee, Joybrato. 2009. *Anglistische Korpuslinguistik: Eine Einführung*. Berlin: Erich Schmidt.
2010 Reppen, Randi. 2010. *Using corpora in the language classroom*. Cambridge: Cambridge University Press.
2012 McEnery, Tony & Andrew Hardie. 2012. *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.
2014 Gatto, Maristella. 2014. *Web as corpus: Theory and practice*. London: Bloomsbury.
2015 Kübler, Sandra & Heike Zinsmeister. 2015. *Corpus linguistics and linguistically annotated corpora*. London: Bloomsbury.
2016 Crawford, William & Eniko Csomay. 2016. *Doing corpus linguistics*. London: Routledge.
2016 Weisser, Martin. 2016. *Practical corpus linguistics: An introduction to corpus-based language analysis*. Chichester: Wiley Blackwell.
2019 Lindquist, Hans & Magnus Levin. 2019. *Corpus linguistics and the description of English*, 2nd edn. Edinburgh: Edinburgh University Press.
2020 Zufferey, Sandrine. 2020. *Introduction to corpus linguistics*. London: Wiley-ISTE.

11 This overview makes no claim to be exhaustive.

Handbooks

2009 Lüdeling, Anke & Merja Kytö (eds.). 2009. *Corpus linguistics. An international handbook.* Berlin & New York: Mouton de Gruyter.
2012 O'Keeffe, Anne & Michael McCarthy. 2012. *The Routledge handbook of corpus linguistics.* London: Routledge. (2nd edition currently in preparation).
2015 Biber, Douglas & Randi Reppen. 2015. *The Cambridge handbook of English corpus linguistics*. Cambridge: Cambridge University Press.
2015 Granger, Sylviane & Gaëtanelle Gilquin. 2015. *The Cambridge handbook of learner corpus research.* Cambridge: Cambridge University Press.
2017 Durand, Jacques, Ulrike Gut & Gjert Kristoffersen (eds.). 2017. *The Oxford handbook of corpus phonology.* Oxford: Oxford University Press.

Book series

1998 *Studies in Corpus Linguistics* (John Benjamins)
2002 *Routledge Advances in Corpus Linguistics* (Routledge)
2004 *English Corpus Linguistics* (Peter Lang)
2004 *Research in Corpus and Discourse* and *Studies in Corpus and Discourse*: the 2 strands of the series *Corpus and Discourse* (Bloomsbury)
2011 *Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache* (CLIP, Narr)

Introductions to statistics for linguist(ic)s

1998 Oakes, Michael. 1998. *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.
2008 Baayen, Harald R. 2008. *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge: Cambridge University Press.
2008 Gries, Stefan Th. 2008. *Statistik für Sprachwissenschaftler* (Studienbücher zur Linguistik, Band 13). Göttingen: Vandenhoeck & Ruprecht.
2009 Gries, Stefan Th. 2009/2013^2. *Statistics for linguistics with R: A practical introduction*. Berlin: de Gruyter Mouton.
2015 Eddington, David. 2015. *Statistics for linguistics: A step-by-step guide for novices*. Cambridge: Cambridge Scholars.
2015 Levshina, Natalia. 2015. *How to do linguistics with R: Data exploration and statistical analysis*. Amsterdam & Philadelphia: John Benjamins.
2017 Desagulier, Guillaume. 2017. *Corpus linguistics and statistics with R: Introduction to quantitative methods in linguistics*. Cham: Springer.
2018 Brezina, Vaclav. 2018. *Statistics in corpus linguistics: A practical guide*. Cambridge: Cambridge University Press.
2020 Wallis, Sean. 2020. *Statistics in corpus linguistics research: A new approach*. New York: Routledge.

Journals

1996	*Journal of Quantitative Linguistics*
2002	*International Journal of Corpus Linguistics*
2005	*Corpus Linguistics and Linguistic Theory*
2006	*Corpora*
2013	*Research in Corpus Linguistics*
2014	*Journal of Research Design and Statistics in Linguistics and Communication Science*

References

Arppe, Antti, Gaëtanelle Gilquin, Dylan Glynn, Martin Hilpert & Arne Zeschel. 2010. Cognitive corpus linguistics: Five points of debate on current theory and methodology. *Corpora* 5(1). 1–27.

Blumenthal-Dramé, Alice. 2012. *Entrenchment in usage-based theories: What corpus data do and do not reveal about the mind*. Berlin & Boston: De Gruyter Mouton.

Blumenthal-Dramé, Alice. 2016. What corpus-based cognitive linguistics can and cannot expect from neurolinguistics. *Cognitive Linguistics* 27 (4). 493–505.

Bod, Rens. 2010. Probabilistic linguistics. In Bernd Heine & Heiko Narrog (eds.), *The Oxford handbook of linguistic analysis*, 633–662. Oxford: Oxford University Press.

Bresnan, Joan & Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86(1). 186–213.

Bybee, Joan & Paul Hopper (eds.). 2001. *Frequency and the emergence of linguistic structure*. Amsterdam & Philadelphia: John Benjamins.

Coleman, John & Janet B. Pierrehumbert. 1997. *Stochastic phonological grammars and acceptability*. Association for Computational Linguistics. ArXiv cmp-lg/9707017. https://www.aclweb.org/anthology/W97-1107 .

Docherty, Gerard J. & Paul Foulkes. 2014. An evaluation of usage-based approaches to the modeling of sociophonetic variability. *Lingua* 142. 42–56.

Ellis, Nick. 2002. Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition* 24(2). 143–188.

Goodman, Judith, Philip Dale & Ping Li. 2008. Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language* 35. 515–531.

Granlund, Sonia, Joanna Kolak, Virve Vihman, Felix Engelmann, Elena V. M. Lieven, Julian M. Pine, Anna L. Theakston & Ben Ambridge. 2019. Language-general and language-specific phenomena in the acquisition of inflectional noun morphology: A cross-linguistic elicited-production study of Polish, Finnish and Estonian. *Journal of Memory and Language* 107. 169–194.

Gries, Stefan Th. 2013. Elementary statistical testing with R. In Manfred Krug & Julia Schlüter (eds.), *Research methods in language variation and change*, 361–381. Cambridge: Cambridge University Press.

Gries, Stefan Th. 2015. Some current quantitative problems in corpus linguistics and a sketch of some solutions. *Language and Linguistics* 16. 93–117.
Gries, Stefan Th. & Nick C. Ellis. 2015. Statistical measures for usage-based linguistics. *Language Learning* 65(Suppl. 1). 228–255.
Janda, Laura A. (ed.). 2013. *Cognitive linguistics: The quantitative turn. The essential reader*. Berlin & Boston: De Gruyter Mouton.
Janda, Laura A. 2017. The quantitative turn. In Barbara Dancygier (ed.), *The Cambridge handbook of cognitive linguistics*, 498–514. Cambridge: Cambridge University Press.
Janda, Laura A. 2019. Quantitative perspectives in cognitive linguistics. *Review of Cognitive Linguistics* 17. 7–28.
Joseph, Brian. 2008. The editor's department: Last scene of all… *Language* 84. 686–690.
Krug, Manfred & 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*. Düsseldorf: University of Düsseldorf Postdoctoral Dissertation.
Langacker, Ronald. 1987. *Foundations of cognitive grammar, Vol. 1: Theoretical prerequisites*. Stanford, CA: Stanford University Press.
Manning, Chris. 2003. Probabilistic syntax. In Rens Bod, Jennifer Hay & Stefanie Jannedy (eds.), *Probabilistic linguistics*, 289–341. Cambridge, MA: MIT Press.
Marsden, Emma, Kara Morgan-Short, Pavel Trofimovich & Nick C. Ellis. 2018. Introducing registered reports at *Language Learning*: Promoting transparency, replication, and a synthetic ethic in the language sciences. *Language Learning* 68. 309–320.
Palacios Martínez, Ignacio M. 2020. Methods of data collection in English empirical linguistics research: Results of a recent survey. *Language Sciences* 78. 101263.
Sampson, Geoffrey R. 2005. Quantifying the shift towards empirical methods. *International Journal of Corpus Linguistics* 10. 10–36.
Sampson, Geoffrey R. 2013. The empirical trend: Ten years on. *International Journal of Corpus Linguistics* 18(2). 281–289.
Schönefeld, Doris (ed.) 2011. *Converging evidence: Methodological and theoretical issues for linguistic research*. Amsterdam & Philadelphia: John Benjamins.
Zipf, George Kingsley. 1949. *Human behavior and the principle of least effort*. Cambridge, MA: Addison-Wesley Press.