Lukas Sönning* and Valentin Werner

The replication crisis, scientific revolutions, and linguistics

https://doi.org/10.1515/ling-2019-0045
Received December 28, 2019; accepted July 8, 2021; published online September 16, 2021

Keywords: methodology; paradigm shift; philosophy of science; quantitative linguistics; replication crisis; scientific revolution

1 Introduction

This introductory article sets the scene for the special issue of *Linguistics* entitled *The replication crisis: Implications for linguistics*. We start out by sketching key issues surrounding the replication crisis, a state of methodological unrest that has caused turbulence across quantitative disciplines since the early 2000s. Given the trend throughout the language sciences towards a greater reliance on empirical methods, we argue that the linguistic community should be perceptive to this broader discourse and foster an open and lively discussion culture. To this end, the contributions to this special issue engage with methodological aspects that are at the core of this debate, taking an explicitly linguistic point of view. After an overview of the individual articles, we offer a perspective of the replication crisis from an angle inspired by the philosophy of science. We turn to Kuhn ([1962] 1996) to understand today's methodological controversies in the context of his cyclical model of scientific progress. This allows us to interpret the replication crisis as a transitory stage, with revolutionary forces propagating a paradigm shift. We attempt to capture and contrast essential features of the status quo and what may eventually emerge as the new methodological paradigm in quantitative linguistics.

We would like to start with the basic observation that over the period of roughly two decades, many branches of linguistics have seen a shift towards an increased use of empirical methods. While this certainly does not imply that

***Corresponding author: Lukas Sönning**, Institut für Anglistik und Amerikanistik, Otto-Friedrich-Universität Bamberg, An der Universität 9, 96045 Bamberg, Germany,
E-mail: lukas.soenning@uni-bamberg.de. https://orcid.org/0000-0002-2705-395X
Valentin Werner, Institut für Anglistik und Amerikanistik, Otto-Friedrich-Universität Bamberg, An der Universität 9, 96045, Bamberg, Germany, E-mail: valentin.werner@uni-bamberg.de.
https://orcid.org/0000-0003-2669-3557

theoretical and qualitative work is obsolete, we find traces of this development in every corner of linguistic study. Thus, statistics feature prominently in conference talks, and the same is true for book contributions and journal articles, whose claims about language are often grounded in quantitative data (Brinton et al. 2019; Palacios Martínez 2020; Sampson 2005, 2013). The methodological literature, which has kept up with this expansion, now offers a wide range of textbooks on the statistical analysis of language data (Baayen 2008; Brezina 2018; Cantos Gómez 2013; Desagulier 2017; Eddington 2015; Grant et al. 2017; Gries 2017, 2021; Johnson 2008; Levshina 2015; Loerts et al. 2020; Oakes 1998; Rasinger 2013; Rietveld and van Hout 2005; Schneider and Lauber 2019; Sonderegger et al. 2018; Wallis 2021; Winter 2019). Relevant issues also feature prominently in methodological handbooks (see, e.g., part 3 of Krug and Schlüter 2013; part II of Podesva and Sharma 2014), in focus and tutorial articles in journals (see, e.g., Nicenboim and Vasishth 2016; Vasishth and Nicenboim 2016; Vasishth et al. 2018a), and in several dedicated blogs, such as *corp.ling.stats*,[1] *Experimental Linguistics in the Field*,[2] *Shravan Vasishth's Slog*,[3] or *Doing Linguistics with a Corpus*.[4] In addition, the *Journal of Research Design and Statistics in Linguistics and Communication Science* has emerged as a specialized outlet that highlights the increasing relevance of empirical know-how in the language sciences and related areas.

Given that linguists nowadays often take a quantitative approach to investigating language, the language sciences should be attentive to issues that cause turbulence and stimulate discussion in other, arguably more mature quantitative disciplines. The present special issue is a step in this direction. It aims to sensitize the linguistic community to certain controversies surrounding the "replication crisis", which have been unsettling empirical branches of science since the early 2000s. This cover term designates a state of pronounced methodological introspection across the quantitative sciences. The current replication crisis was triggered by serious indications of data-based claims being less reliable than researchers had believed. Thus, it turned out that statistical conclusions could in many cases not be reproduced upon confrontation with new data. Large-scale replication projects have lent credence and publicity to this concern (e.g., Open Science Collaboration 2015) and prompted debate both inside and outside academia (see, e.g., Blech 2019; Carroll 2017; Yong 2018). The ensuing reflective state of mind bears great potential, both for individual scholars and the community at large, to work towards better ways of learning from data.

1 https://corplingstats.wordpress.com/.
2 https://experimentalfieldlinguistics.wordpress.com/.
3 https://vasishth-statistics.blogspot.com/.
4 https://linguisticswithacorpus.wordpress.com/.

It is beyond doubt that the issues raised in the context of this debate are relevant for our discipline. In fact, with linguistics being a relative newcomer to the quantitative scene (Köhler 2005: 2–4), methodological expertise is still developing. While certain specialized areas have a strong empirical tradition (see Section 2), data-based work in many fields of linguistics is in a state of relative infancy. It is therefore likely that human factors that are seen as important causes of the widespread misapplication and misinterpretation (and even misuse) of statistical methods are, on average, even more pronounced in our community. The language sciences should therefore take the opportunity to learn from the methodological advances that other sciences have put on the agenda. The overarching purpose of this special issue, then, is to raise awareness among linguists and engage them in the debate.

In this introductory part, we pursue three broader aims. For one, we will provide some context on the replication crisis and touch upon certain issues permeating current debates. We are only able to offer a glance at the vast body of contributions, and will therefore point to some helpful references for further reading. Second, we give an overview of the contributions to this special issue and how they connect to the literature on the replication crisis, the research process, and to each other. Finally, we take a step back from the heated debates surrounding methodological reform movements to consider current developments from a broader philosophical perspective. Most prominently, in this regard, Thomas Kuhn ([1962] 1996: 88) has noted that it is "particularly in periods of acknowledged crisis that scientists have turned to philosophical analysis as a device for unlocking the riddles of their field". We take up this inspiration and consider the replication crisis from the viewpoint of Kuhn's scheme of scientific revolutions. A small dose of philosophical musing should not only constitute a refreshing change from our usual form of research work, but we believe that it may bring forth some additional insights beyond the more concrete engagements offered by the individual contributions to this special issue of *Linguistics*.

2 The replication crisis

Concerns about the use and interpretation of statistical methods in scientific contexts have a long tradition (e.g., Berkson 1938, 1942; Cohen 1994; Meehl 1967, 1990; Nickerson 2000; for book-length treatises see, e.g., Cumming 2012; Kline 2013 [2004]; Ziliak and McCloskey 2008). One of the central concerns in these works is the tendency of scholars to misinterpret, or misuse, *p*-values. More recently, the debate about the utilization of quantitative methodologies in applied research was

quite forcefully re-kindled by Ioannidis (2005), in a paper provocatively titled "Why most published research findings are false".

Relevant publications have explored major shortcomings of current research and publication practice, to make researchers aware of issues relating, for instance, to study design, inadequate use of statistical methods, overreliance on *p*-values, selective reporting of results, etc. Among the focal problems identified within this broader discourse are

- a lack of transparency in methodology and data analysis,
- the non-reproducibility of scholarly work, as, for example, original data and analysis procedures are not accessible,
- reluctance to undertake replication studies as purportedly "unoriginal" (and unprestigious) despite their potential to put previous findings in perspective, and
- concerns about high rates of false-positive findings in the published scientific literature.

Efforts have been made to identify potential causes of the replication crisis, and several explanations have been put forward, including the following:

- structural factors and institutionalized incentives, for instance a strong link between success in scholarly careers and publication records (Nosek et al. 2012; Smaldino and McElreath 2016; Stark and Saltelli 2018),
- overreliance on significance testing and incorrect interpretations of *p*-values (e.g., Greenland et al. 2016),
- poor research design and quality that lead to underpowered studies with a fragile basis for statistical inference (Ioannidis 2005; Loken and Gelman 2017),
- cognitive biases that interfere with a neutral and reasoned interpretation of empirical data (Greenland 2017; Munafò et al. 2017),
- unintentional, data-contingent analysis decisions that alter the error-rates of statistical procedures (Gelman and Loken 2014; Wicherts et al. 2016), and
- weak theory, which fails to inform, guide, and constrain data-based work (Meehl 1967, 1990; Muthukrishna and Henrich 2019; Smaldino 2019).

The core concern of the discourse surrounding the replication crisis in the quantitative sciences is to ensure conscientious *practice*. The substantive conclusions drawn from data should be sound, and uncertainty must be represented and communicated properly. To this end, a set of proposed solutions and best practices has emerged to increase the validity and cumulative value of scientific work. Apart from efforts directed towards better methodological training of scholars, novel strategies include:

- the preregistration of data collection and analysis plans, with the emergence of registered reports as a new publication category (Chambers 2013),
- reporting guidelines, which aim to encourage full disclosure of quantitative results (e.g., Larson-Hall and Plonsky 2015),
- the propagation of open science practices to ensure full transparency of data analysis procedures and encourage – and in some cases require – that data, materials and code be made publicly available (e.g., Paquot and Callies 2020),
- a scientific mentality that has been referred to as 'meta-analytic thinking' (Cumming 2012), a cumulative stance towards knowledge construction and the information value of individual studies (e.g., Brandt et al. 2014; Peels 2019; Schmidt 1996), and
- the endorsement of alternative inferential frameworks, such as estimation (Cumming 2012) and Bayesian inference (Kruschke 2010), to counter valid concerns about null hypothesis significance testing.

In order for a discipline to promote and adopt better quantitative practice, awareness of relevant issues must be raised in the community. While vivid and controversial discussion is cultivated in other scholarly fields, efforts towards this end have arguably been insufficient in the language sciences (Gries 2015). In certain highly specialized, strongly quantitative subdisciplines of linguistics, however, the level of awareness and reflection is more advanced. While the debate has so far not been elevated to a broader, general linguistic plane, the discourse that has ensued in these métiers reveals the relevance of core issues for the study of language. Thus, important insights have already been gained in fields such as:

- psycholinguistics (e.g., Baayen et al. 2008; Jaeger 2008; Vasishth et al. 2018b),
- corpus linguistics and learner corpus research (e.g., Flanagan 2017; Gries 2006, 2018; Gries and Deshors 2021; Paquot and Callies 2020; Paquot and Plonsky 2017),
- applied linguistics, including second language acquisition research (e.g., Larson-Hall and Herrington 2010; Larson-Hall and Plonsky 2015; Marsden et al. 2018; Mulder 2020; Nassaji 2012; Norris and Ortega 2007; Plonsky 2015; Plonsky and Gass 2011; Porte 2015; Porte and McManus 2019),
- sociolinguistics (e.g., Aguilar-Sánchez 2014, 2017; D. E. Johnson 2009, 2014), and
- the phonetic sciences (e.g., Roettger 2019; Roettger et al. 2019; Winter 2011).[5]

5 For the sake of completeness, we would also like to mention the area of computational linguistics/natural language processing. For relevant discussion in this field, see Church and Liberman (2021), Geman and Johnson (2004), M. Johnson (2009), and Wieling et al. (2018), for instance.

With current research in the language sciences being driven by a strong reliance on quantitative methods, there can be little doubt about the relevance of the issues pertaining to the replication crisis for our field. In fact, it is rather disconcerting that as yet this debate has been given little explicit attention among linguists.[6] We argue that it is time to forcefully put this topic on the broader linguistic agenda, to encourage communication across disciplines and thereby to reach out to those areas of language study where quantitative methods are used but a reflected analysis of this practice has not taken place.

While the language sciences have diversified to a considerable degree, there are nevertheless a number of unifying aspects that are particularly relevant to quantitative practice in general. For one, our shared object of study, language, exhibits qualities that foreground certain facets of this debate. These concerns, for instance, the adequacy and validity of statistical models used for analysis and description. Further, there are parallels in terms of methodology. Thus, the language sciences share a strong reliance on naturalistic corpus data and experimental designs for knowledge construction (while participant observation and elicitation, e.g., through questionnaire surveys and interviews, have also featured prominently in certain branches; see, e.g., Krug et al. 2013: 8).

3 Overview of this special issue

The individual contributions pursue two goals: First, the reader will be sensitized to a particular aspect of the debate revolving around the replication crisis and will be offered, at a general level, a transparent and explicit exposition of empirical problems and proposed solutions that have been identified in this discourse. Second, the focal aspect will be discussed from the perspective of linguistics, with a focus on whether and how discipline-specific nuances may qualify or emphasize certain points.

In his article titled "Reflecting on the quantitative turn in linguistics", Bernd Kortmann contextualizes the overall concerns of this special issue through the retrospective lens of someone directly affected by and involved in what could be considered a fundamental paradigm shift in his discipline. He starts from the observation that linguistic research increasingly has turned into a quantitative endeavor (see Section 1) and asks (i) whether this quantitative turn has been to the detriment of qualitative methods, or even of linguistic theorizing in general,

6 For exceptions that discuss individual aspects related to the replication crisis, see, e.g., Arppe et al. (2010), Berez-Kroeker et al. (2018), Gawne and Berez-Kroeker (2018), and Haspelmath and Siegmund (2006).

(ii) whether linguistics has reached the point of crisis yet, and (iii) what repercussions the strong quantitative turn has for the publication system of linguistics. While he concludes that linguistics could become an important part of the digital humanities, and that using quantitative empirical approaches is a welcome development in principle, he postulates that certain preconditions, such as adequate statistical training of those involved, must be met (see also Vasishth and Gelman's article), and pitfalls, such as the "from-corpus-to-cognition fallacy", must be evaded. Regarding its establishment as a serious quantitative discipline, Kortmann's verdict is that linguistics still has "a quite hard and long way to go". To achieve this, he argues for efforts (i) to avoid mistakes that other quantitative sciences have made and (ii) to strengthen open science practices in the community. On a related note, he suggests that gatekeepers in the linguistic publication system should take initiatives to ensure appropriate empirical practice, for instance by regularly relying on statistical consultants and by introducing article types such as registered reports, which help to avoid potential biases towards the "new" and "significant" (see also Roettger's article).

Preregistration in linguistic research is advocated in Timo B. Roettger's article by the same title. He first identifies two weaknesses in current linguistic research and publication practice, namely (i) a reluctance to accept null results or direct replication as work worth sharing with other researchers, and (ii) the presence of "researcher degrees of freedom", that is (post-hoc) flexibility at various stages (data collection, measurement, and analysis) in the research process. While (i) may lead to a pronounced publication bias toward the "new" and "significant" (affecting the types of studies that are conducted and submitted for publication in the first place), (ii) may result in HARKing (Hypothesizing After Results are Known) and selective reporting of findings. As a remedy for these potentially detrimental practices, he recommends and introduces preregistration, generally defined as "a time-stamped document that specifies how data is to be collected, measured, and analyzed prior to data collection". He outlines various forms and principles of preregistration, provides pointers to resources that facilitate the process, and draws attention to registered reports as an emerging article type that explicitly incentivizes preregistration. While Roettger embraces preregistration as an important component of open science practices, he also highlights its limitations and suggests that, given the wide variety of linguistic research concerns, it certainly is not a one-size-fits-all solution.

The joint contribution "Independence and generalizability in linguistics" by Bodo Winter and Martine Grice addresses the persistent issue that data points in linguistics regularly are treated as independent, even though often they actually lack this property. The authors argue that this potential source of variation has regularly been ignored and that the uncritical use of statistical measures

that rely on the independence assumption can lead to inaccurate results and interpretations. They base their further discussion on the claim that "for any reasonably complex data set, it is practically impossible not to violate the independence assumption with respect to some dimension of non-independence when using standard significance tests" (see also the contribution by Vasishth and Gelman). While non-independence of data points is now regularly accounted for in linguistic studies, they show that the topic is not consistently reflected in the linguistic textbook literature. With a view to addressing this lacuna, their contribution first offers a general introduction to mixed-effects models, independence in linguistic data, and the connected broader issue of generalizability (see also the contribution by Grieve). To raise awareness of the ubiquity of non-independent data structures in linguistics, Winter and Grice then present an overview of such structures in various linguistic fields, extending beyond the regularly acknowledged factors of speaker/text and lexical item. Specifically, they discuss spatial dependence, language and language family dependence, temporal and sequence dependence, talker effects, dyad effects, exact repetitions, and the nested hierarchical structure of corpora. Overall, while they acknowledge that mixed-effects models are a potential methodological remedy in many cases, they suggest that non-independence of data points should already be tackled at the design stage of a study.

In their paper "Variables are valuable: Making a case for deductive modeling", David Tizón-Couto and David Lorenz draw attention to the differential status of predictor variables in the context of statistical modeling. Taking the implications of the "quantitative turn" in linguistics as their starting point (see also Kortmann's contribution), they clearly welcome quantitative approaches that use increasingly sophisticated statistical machinery, but underscore the ancillary role of statistical analysis in the field (see also Section 4.2). They argue that linguists – potentially due to the absence of "established" guidelines in the field – may sometimes feel overwhelmed by the amount of techniques on offer. This may tempt researchers to choose one (allegedly more sophisticated) procedure over another, irrespective of the realities of the linguistic phenomenon or area under investigation (see Section 4.4). The authors then focus on regression modeling as a widely used multifactorial approach and discuss principles of deductive ("intuitive", i.e., informed by theoretical pre-data knowledge) and predictive ("minimal", i.e., parsimonious, significance-based) modeling, mainly in terms of variable selection (and exclusion). Based on a dataset illustrating variation in phonetic realizations of *have to*, they contrast the results and ensuing interpretation of both modeling approaches. They conclude that the deductive approach may be preferable if groundedness in linguistic (rather than statistical) theory and comparability of effects of variables across models (e.g. for replications) is aimed for. They also suggest that such an

approach, which hinges on the conscious pre-selection of variables, may lead to higher accuracy rates for individual coefficients, and thus may serve to develop a clearer picture of linguistic realities.

In a collaborative effort, Shravan Vasishth and Andrew Gelman set out to show "How to embrace variation and accept uncertainty in linguistic and psycholinguistic data analysis". They argue that traditional dichotomous approaches to statistical inference fail for language data for various reasons or may lead to overconfident, deterministic statements of results and their interpretations. They propose that a principal reason for this state is a lack of statistical literacy in the linguistic community at large as regards the scope of answers that statistical analysis can offer (see Section 4.6). After briefly revisiting the logic and potential weaknesses of null hypothesis significance testing, they argue that "conclusions based on data are almost always uncertain, and this is regardless of whether the outcome of the statistical test is statistically significant or not", which makes a strong case for an estimation approach (see also the contribution by Winter and Grice). In the following, they present an extended case study on agreement attraction effects that highlights the affordances of an estimation approach and additionally presents uncertainty estimates within the scope of a "region of practical equivalence approach" (ROPE). As an alternative to power analysis for sample size calculation, the authors suggest that planning of future studies could extensively rely on information gathered either through a ROPE approach or a Bayes factor design analysis. At the same time, they argue that a strong theoretical foundation of the analysis is important (see Section 4.6). Eventually, Vasishth and Gelman address several commonly encountered objections to the approaches they present, and submit that theory can be better assessed through estimation, which crucially involves an open mindset towards variation and uncertainty.

In his position paper "Observation, experimentation, and replication in linguistics" Jack Grieve frames language as a unique, complex, and inherently social phenomenon, a fact that may pose some limits to the application of experimental methods and an obstacle to replication in the language sciences at large. He acknowledges that linguistics, like other empirical disciplines (see Section 1), currently suffers from a "replication crisis", but raises the question of whether this crisis is due to inadequate scientific practice or to problems inherent in language (use) as an object of study. He further draws a fundamental distinction between experimental and observational approaches to language study, which has direct implications for replication, given the various degrees of naturalness of and researcher control on the data. Grieve highlights the possibility that both observational and experimental studies may fail due to poor research design or unknown variation in the study populations. Yet, his main line of argument is that replication failures in experimental linguistics are mainly caused by language use

varying strongly across social contexts, and by social contexts varying across independent replications, a situation that cannot be fully controlled by linguists even in carefully designed research settings. Therefore, he eventually suggests that "replication failure is an inevitable product of using experimentation to probe a highly social phenomenon like language". With a broader view to valid and generalizable findings, for Grieve this situation implies that experimental methods should only be used very carefully, while the inherent value of observation should be held in (equally) high esteem whenever core questions of linguistic inquiry are addressed.

In a brief reply to Grieve, entitled "Context sensitivity and failed replications in linguistics", Timo B. Roettger emphasizes that experimental linguists are aware of the inherent limitations of their approaches and that the study of linguistic phenomena always has to recognize their context-dependence. As a partial remedy, he proposes that, rather than accepting context-sensitivity as an "inevitable fate", more efforts should be invested to exert control over nuisance variables and avoid biases arising from sampling procedures. Roettger suggests three concrete strategies to this end: (i) the specification of the target population with as much detail as possible, (ii) a description of factors that may constrain the generalizability of specific findings, and (iii) the application of more conservative statistical models for data analysis.

As can be gauged from the foregoing summaries, it is clear that the replication crisis is a multifaceted phenomenon (or rather multi-headed hydra?) and that attempts at addressing and overcoming it come in many shapes. While we are confident that the present special issue covers a fair range of central topics, there certainly are additional aspects to be treated and connections to be made in the future, for instance to the more general concerns of open science (see, e.g., Berez-Kroeker et al. forthcoming; Garellek et al. 2020; Heise and Pearce 2020).

4 Scientific revolutions

Let us now take a step back and consider the replication crisis from the viewpoint of Thomas Kuhn's theory of scientific revolutions. His perspective on the progress of science deviates from earlier accounts such as Karl Popper's falsificationism in that it is based on socio-historical observation rather than logical arguments (see, e.g., Chalmers 2013: 97–119). First published in 1962, his landmark work, *The Theory of Scientific Revolutions*, has earned him a place in the pantheon of twentieth-century philosophers of science. Even though Kuhn's work concentrates on theoretical paradigms, we submit that it can be fruitfully applied to methodology. By taking an evolutionary perspective on research methodology, we can

locate the current state of crisis in Kuhn's cyclical description of scientific progress. This framing allows us to understand it as a transitory stage, on the cusp, perhaps, of potentially far-reaching changes in methodological standards. Further, it highlights the important role played by sociological factors and the scientific community – aspects that are often not the primary focus in the literature prompted by the replication crisis.

4.1 Kuhn's structure of scientific revolutions

Kuhn's description of the cyclical evolution of scientific theory discerns three recurrent stages. A period of normal science is characterized by little controversy over the theoretical underpinnings of scholarly work, with researchers routinely working on small problems within the confines of a theory. Research activity may observe anomalies, which point to shortcomings of the abstracted state of knowledge but can be mitigated by theoretical refinements. A crisis develops if mismatches between theory and nature shake the very foundations of a paradigm, which leads practitioners to question the shared set of norms. The state of crisis culminates in a revolutionary overthrow of established codes, which are replaced by a new paradigm. A new period of normal science follows, which will eventually give rise to the next crisis, and so on (see Figure 1).

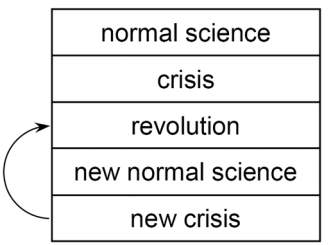

Figure 1: The cyclical evolution of scientific theory according to Kuhn ([1962] 1996).

Kuhn's scheme may be applied to research *methodology* as follows. The period of normal science is characterized by the adoption of a shared set of methodological procedures. Practicing scientists adhere to this paradigm, which defines the standards for empirical conduct. During the period of normal science, anomalies may occur and challenge the capability and effectiveness of these prescriptions. While minor modifications to existing methodology may silence some of these concerns, some dissonances may resist a resolution by means of elaborations of

the current framework. If anomalous phenomena gain momentum and are widely recognized among the members of a scientific community, research practice enters a state of crisis. This crisis is resolved when a new paradigm emerges and practitioners adhere to a new set of methodological standards. Eventually, this new paradigm will also encounter anomalies, followed by a crisis, a revolution, and the next paradigm shift, and so on.

In a similar way to theory, we posit, methodology may evolve in a cyclical fashion. Before we look at the individual states in more detail and explicate them in reference to our discipline, we will highlight some ways in which theoretical and methodological paradigms differ, with a specific view to linguistics where appropriate.

4.2 Methodological versus theoretical paradigms

In general, we may consider a methodological paradigm as a set of shared norms that define the legitimate methods for empirical work and (statistical) analysis. There are some differences between theoretical and methodological paradigms, to which we would like to turn now.

In contrast to theoretical paradigms, standards relating to empirical conduct tend to be normative. This surfaces in the discipline-specific methodological literature (see Section 1), which sets down standard procedures, best practices, and hard-and-fast rules for data-based work. The symptomatic cookbook style that is typical of many textbooks gives the impression that there is a correct "test" or analysis strategy for each problem (see also Section 4.4). Prescriptive features of a methodological paradigm also materialize in peer review processes and editorial decisions, where normative constraints function as a bottleneck for determining the quality, reliability, and, eventually (and arguably most vitally), the publishability of linguistic claims and interpretations. Theoretical paradigms do not necessarily exert this pressure for purported righteousness, perhaps because the verbal nature of most frameworks leaves too much leeway to enforce the foundations of a paradigm.

A second difference concerns the role of theory and methodology in the research process. Methodology plays an ancillary role in two regards. First, it is a means to an end, as it serves to increase our knowledge and understanding of the subject-matter, in our case language. Contrasted with theory, it therefore assumes an inferior position. Second, this imbalance is reflected in the relative level of competence that members of a scientific community have. As linguists, our key domain of expertise is the study of language. We are trained to work on questions of linguistic interest, and it is our job to learn about and understand how humans

acquire and use language and how it may be represented in the mind. To us, methodological know-how is useful, perhaps essential to achieve our aims. However, we often lack the insight and background knowledge that is necessary to confidently adopt and defend a position in methodological debates. This is usually in stark contrast with theoretical concerns, where we navigate more confidently.

These differences between methodological and theoretical paradigms are relevant in the subsequent discussion. A discordance that is not difficult to discern is that between the normative character and the ancillary role of methodological standards. On the one hand, rigorous methodological prescriptions may be advocated, if not enforced, by agents whose expertise actually is in other matters. On the other hand, linguistic and theoretical arguments should easily, but rarely do, override methodological norms. We will elaborate on this tension further below.

4.3 Normal science

During a state of normal science, most research consists of "puzzle-solving" (Kuhn [1962] 1996: 35). Operating within the confines of a paradigm, scientists address problems or phenomena at a high degree of specialization. Given the comfort and security of a shared and accepted theoretical foundation, research specialists can allocate resources to the boundaries of their knowledge, and scholarly work proceeds in small, cumulative steps. New generations of normal scientists acquire the accepted standards and techniques through supervision by skilled seniors and engagement in the scientific community. This professionalization grants membership in the community. However, as Kuhn ([1962] 1996: 64) notes, it leads "to an immense restriction of the scientist's vision", a kind of acquired academic narrow-mindedness. A key role in the transfer and conservation of norms is played by textbooks, which function as a source of authority, a "pedagogical vehicle for the perpetuation of normal science" (Kuhn [1962] 1996: 137). Remaining uncritical of the prevailing paradigm, normal scientists confidently work on problems under the presupposition that the extant paradigm offers the means for solving these. Their knowledge about the underlying framework will be tacit, however; that is, they will not be able to give a precise description or definition of the paradigm. In general, then, normal science can be understood as a relatively quiescent period of productive work and agreement on theoretical essence.

From the viewpoint of methodology, a period of normal science features methodological consensus in a scholarly community. Similar to a theoretical paradigm, an agreed-upon set of techniques creates a safe haven for empirical conduct, allowing scholars to focus on the puzzles in their specific fields of study. Only in periods of methodological tranquility can empirical researchers devote

their full resources to their primary object of study. If practitioners were instead engaged in methodological reflection and debate, progress on the subject-matter frontier (in our case, the language frontier) would be slowed. States of consensus about empirical practice are therefore highly desirable, as overt disagreement over elementary questions of data-based work would hamper scientific progress.

By the same token, in a period of normal methodology, quantitatively oriented linguists direct their attention and resources to their object of study and confidently rely on a set of techniques that constitute the current methodological norm. The adequacy of this accepted ensemble of procedures is assumed as given. It is passed on to new generations, either by their supervisors or by textbooks documenting and introducing a virtually axiomatic inventory of methods. Being concerned primarily with their substantive domain of research, normal scientists adhere to this paradigm, which permeates empirical work carried out in their research community.

4.4 The current methodological paradigm in quantitative linguistics

Kuhn concedes that it is difficult to give a precise definition of a particular paradigm. He invokes Wittgenstein's notion of family resemblance and argues that it is impossible to establish features common to all work within a paradigm. Among practitioners, knowledge at this meta-level is also tacit. Due to the implicit and usage-based acquisition, attributes of a paradigm are often difficult to formulate explicitly. Bearing in mind the inherent difficulty of the task, it is nevertheless a worthwhile reflective exercise to try to articulate some features of the current quantitative paradigm in linguistics. We should stress, however, that this attempt is bound to do injustice to many individuals, if not whole subdisciplines (see Section 2).

Judging from the quantitative stance that is embodied in relevant textbooks (e.g., Baayen 2008; Grant et al. 2017; Gries 2021; Levshina 2015), the current paradigm appears to revolve around the notion of testing. As such, the conception of data-based learning as a mechanistic testing procedure has its roots in the statistical literature, specifically the null hypothesis significance testing approach devised by Ronald A. Fisher at the beginning of the twentieth century (see, e.g., Fisher 1925).[7] Testing usually entails discrete decisions, and binary choices therefore permeate our current paradigm. Most prominently, we encounter them in

7 It should be noted that Fisher himself argued against strict adherence to *p*-value thresholds, stating that "no scientific worker has a fixed level of significance at which from year to year, and in

the interpretation of statistical inferences about patterns or comparisons in a set of data. Additionally, testing produces categorical decisions in other analytic settings, either actively by the user (as in variable selection tasks in regression modeling) or in a fully automated fashion (as in certain machine learning tasks).

The prominence of testing may have given rise to a second central feature of the current methodological "code of conduct". In general, both the analysis and interpretation of language data typically rely quite strongly on computational and statistical aids. Despite the truism that "[l]inguistics is done by linguists, not by computers" (Egbert et al. 2020: 69), which could be seen as a strong indicator that some rethinking is already taking place, it appears that background knowledge, linguistic theory, and the research questions that motivate an analysis are granted less authority in data-based work. Thus, the illustration of techniques in textbooks rarely takes explicit account of the scientific context and research objectives, but rather portrays quantitative research problems as solvable by means of objective tests and algorithms. These, however, ignore epistemological concerns, causal structure, and the intricacies of the underlying linguistic phenomenon. The rote advice given for the statistical analysis of language data therefore consists mostly of technological rather than scientific strategies for problem solving. This, of course, conflicts with the ancillary role that statistics, and methodology in general, have in scientific work. A similar imbalance features in a shift towards methodological/statistical prose at the cost of linguistic description, as recently surveyed by Larsson et al. (2021) for the domain of corpus linguistics, for instance (see also Section 4.6). While an increase in methodological/statistical explicitness is a welcome development in principle, adverse effects on the engagement with the core subject-matter are clearly undesirable.

A final aspect we wish to mention is, in our view, also connected to the dominant testing approach to data analysis. On the whole, the current paradigm consists of a patchwork of methods borrowed from other, not necessarily neighboring, scientific and technological disciplines. Traditional techniques adopted from experimental sciences co-exist with recently popular tools copied from the domain of machine learning. Little effort has been devoted to establishing a set of unifying, language-specific principles for empirical work. Linguistically grounded axioms could provide signposts towards a consensus on core techniques or defaults for language data analysis and thereby constrain methodological choices in a principled way. Current practice, however, appears to reflect a state of methodological proliferation, where the field is open to and

---

all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas" (Fisher 1956: 42).

readily adopts an increasingly diverse set of procedures from other fields of study. We are lacking linguistic imperatives to tell the chaff from the wheat.

In summary, language data analysis, as it is currently taught and practiced, embraces the notion of testing and bears a strong resemblance to technology rather than theoretically-grounded science. Responsibility is often handed to statistical procedures and algorithms, with substantive, subject-specific (i.e., language-related) considerations only playing a minor role. The predominant quantitative mentality shows empiricist features and may be described by the belief that the application of objective methods and tests allows us to extract from data meaningful insights about language.

4.5 Crisis

The "puzzle-solving" activity in periods of normal science produces experiences and observations that are incongruent with the tenets of a paradigm. Kuhn refers to clashes between theory and nature as anomalies. As mentioned above, such mismatches may prompt modifications to the theoretical apparatus in the form of elaborations that allow for the discrepancies noted. A symptom of crisis, as Kuhn ([1962] 1996: 75) notes, is therefore the "proliferation of theories". The rapid increase of modifications to and elaborations of theories reflects the effort of agents working in the old paradigm to amend the existing framework to account for anomalies. A state of crisis develops if anomalous experiences gain critical mass and cannot be offset by alterations to the theoretical status quo. Irregularities are particularly severe if they forcefully point to limits and/or shortcomings at the fundamentals of a paradigm. In the domain of theory, these would be empirical observations that contradict or falsify central tenets of a framework. If anomalies gain prominence and attention throughout the community, the inadequacy of the prevailing paradigm is recognized by the profession. What follows, according to Kuhn, is a transition from normal to extraordinary research. A deconstruction of the paradigm loosens the rules and conventions for research. Normal puzzle-solving activities give way to debates about the theoretical core of a research enterprise and a state of scientific unrest follows. Paradigm-induced constraints are relaxed, which leads to overt expression of discontent, philosophical debate, and disorder in scientific conduct.

Anomalies in research methodology also describe mismatches. Here, inconsistencies occur between nature and inferences about nature. These inferences are statistical rather than scientific, which is to say that we are inferring, from patterns in our data, the presence of patterns in the real world. A mismatch occurs when the statistical inferences we make based on a sample do

not in fact generalize to the wider set of contexts in which we are interested. Of course, procedures for extending the scope of quantitative statements allow for a margin of error, and anomalous experiences are therefore to be expected at a certain (pre-specifiable) rate. If the frequency of mismatches exceeds that rate, however, this will strike at the very foundation of our empirical apparatus. Recognition of severe methodological anomalies would lead to discomfort, quarrels, and discontent in much the same way as described above. A state of methodological proliferation, or anarchy, would follow, with researchers promoting and turning to alternative ways of distilling insights from data. This produces an overabundance of techniques for the analysis of data.

4.6 Methodological anomalies and the replication crisis

The methodological framework currently underlying most scientific activity has given rise to a critical state. The "replication crisis" has shown that methods fail to yield warranted claims about nature at the assumed level of inferential error (see Sections 1 and, among others, the arguments developed in the individual contributions summarized in Section 3). Thus, in the present methodological paradigm, the probability that a quantitative statement about nature will materialize when double-checked using new data, is not as high as the nominal error-rates would lead us to expect. In fact, mismatches between statistical inference and nature abound in the quantitative sciences, which has given rise to today's crisis.

Anomalies produced by current methodological conventions go further, however. Thus, a mismatch can be found between the actual meaning of inferential probabilities and the way researchers interpret and understand them. This discrepancy is widely known and difficult to avoid without conscious effort. This is perhaps due to the fact that the type of probability that is signaled by an inferential *p*-value is at odds with human cognition. This misfit presumably arises from the fact that (i) the researcher is required to formulate a null hypothesis that may not be of direct substantive interest; (ii) *p* denotes a conditional rather than an "ordinary" probability; (iii) *p* refers to the long-run error rate of a statistical procedure (such as a null hypothesis significance test or the method for constructing a confidence interval) instead of a degree of plausibility, as suggested by the everyday sense of the term "probability"; and (iv) *p* is the probability of observing a more extreme test statistic, and warrants no statements about the likelihood of a research hypothesis or a value of interest. The fact that *p*-values do not align well with the way our brains are wired is a distorting factor when it comes to the statistical and substantive interpretation of research findings.

A further anomaly may be found in the interplay between, and relative authority of, methodology and theory in the research process. Thus, as suggested above, the subsidiary role of methodology in scientific work clashes with its normative force. In some contexts, it appears to be elevated to an exclusive criterion for the quality of a linguistic research contribution; in other contexts, it seems to be an end in itself and we sometimes get the impression that statistical methods, rather than linguistic insights, are taking over as selling points of scholarly work (see also Larsson et al. 2021). Another imbalance on the theory-method scale occurs between scientific and statistical considerations and arguments in data analysis. Quantitative work is often driven by crude statistical defaults and cut-offs rather than domain expertise, and data-analytic decisions and claims about language often hinge on statistical rather than linguistic arguments (see Section 3), involving an increased danger of over-abstraction from the data and reduced linguistic validity of interpretations (see also Egbert et al. 2020: 51).

As we argued earlier, the language sciences are in a state of methodological proliferation, which, according to Kuhn's scheme, is symptomatic of a state of crisis. Over the past decade, we have seen an enormous influx of data-analytic techniques from other fields, in particular from the machine learning community. Some of these tools trick us into believing that meaningful insights about language are only a click away. We see different methods being applied to the same problem, which, as Mayo (2018: 27) notes, "is all to the good, […] provided one understands how to interpret competing answers". This, however, rarely appears to be the case. Rather, it seems that, due to the enormous diversification of statistical methods, we are in a period of "pronounced professional insecurity" (Kuhn [1962] 1996: 67), which, in Kuhn's view, typically precedes a paradigm change, which is brought about by a scientific revolution.

4.7 Revolution

If an existing paradigm ceases to be supportive of the scientific goals of a discipline, crisis may lead to a scientific revolution and the emergence of a new paradigm. At this stage of evolution, a key role is played by revolutionaries, who tend to share certain traits. Thus, Kuhn ([1962] 1996: 90) notes that those "who achieve these fundamental inventions of a new paradigm have been either very young or very new to the field whose paradigm they change. […] [B]eing little committed by prior practice to the traditional rules of normal science, [they] are particularly likely to see that those rules no longer define a playable game and to conceive another set that can replace them". However, these pioneers are likely to confront

resistance to novelties and change. Antagonists tend to have high stakes in the old paradigm, perhaps because a successful career has committed them to the old norms. Paradigm shifts are therefore likely to require a generational change.

Kuhn notes that a paradigm shift usually brings about a gestalt switch in the perception of the members of a scientific community. Thus, research problems come to be seen in a very different light and phenomena are conceptualized and understood in a very different way. Moving back and forth between, say, a generativist and functional-cognitive account of the same linguistic phenomenon illustrate such a shift of vision. Simplifying somewhat, this example represents a further feature of Kuhn's framework: Theoretical paradigms are incommensurable. They tend to have different epistemological priorities, attach different senses to the same technical terms, and work "in different worlds" (Kuhn [1962] 1996: 150), which is to say that they tend to see different things when confronted with the same observations.

Methodological revolutions may likewise be obstructed by stakeholders. These could, on the one hand, be members of the community whose expertise in and contributions to the old paradigm have earned them prestige, visibility, and influence. These actors may resist change since adhering to a new set of standards would undermine their work and achievement, if not their reputation. On the other hand, resistance may also be offered by scholars who, though aware of anomalies and able to move forward, feel disincentivized by systemic forces from doing so.

Unlike successive theoretical research programs, however, methodological paradigms need not be incommensurable with each other and may not lead to the type of gestalt switch described by Kuhn. In the domain of quantitative practice, proposed reforms rather bring about smaller shifts of vision in certain parts of empirical conduct. Examples are (i) the complementary role and synthesis of significance testing and estimation, (ii) the unification of a large arsenal of hypothesis tests under the generalized linear model framework, (iii) the appreciation of frequentist (classical) methods as a special case of Bayesian inference, and (iv) a stronger commitment to as well as formalization and acknowledgment of open science practices.

5 Towards a new methodological paradigm in quantitative linguistics

Against the backdrop of Kuhn's evolutionary model, quantitative methodology in the language sciences may be facing a period of unrest and transition. Let us speculate on some of the changes a methodological paradigm shift might produce.

Table 1 gives a summary of this list, which, we must concede, is necessarily selective.

The testing mentality that appears to be so deeply entrenched in current empirical practice will give way to estimation as the predominant mode of expressing the statistical uncertainty of indications in the data. The *p*-value will celebrate its 100-year anniversary (Fisher 1925) in good company, with effect sizes and their uncertainty limits offering a more transparent paraphrase of its inferential information (see Vasishth and Gelman's article). The shift in attention to point and uncertainty estimates will be accompanied by a de-emphasis on discrete significance cut-offs. Instead, more care will be taken to locate linguistic claims on the continuum between exploration and confirmation, with insights based on preregistered analysis plans enjoying a special status (see Roettger's contribution). To readers, this will clarify what evidential force to assign to statistical conclusions.

Table 1: The old and new paradigm in contrast.

Old paradigm	New paradigm
Null hypothesis significance testing, *p*-values	Estimation: Point and uncertainty estimates for substantively meaningful quantities
P-values as publication thresholds	Linguistic substance as a key criterion; claims are located on the exploratory-confirmatory continuum
Bottom-up, data-driven analysis	Top-down, theory-driven analysis
Language-specific data features not taken into account (or considered as a nuisance) during analysis	Linguistically informed analysis; efforts to establish a set of 'language data universals', i.e., typical and recurrent features of (natural) language data
Statistical modeling: Reliance on algorithms and fit indices	Deductive modeling: Guidance by scientific objectives, context, and domain knowledge
Methodological proliferation	Mixed-effects regression as default
Communication of quantitative results at a technical level	Audience design: Empathy and minimal standards for the communication of results
Private/proprietary science: Data not shared, concerns about data breaches or potential criticism	Open science: Open data and code; data seen as a public commodity; culture of mutual respect; constructive atmosphere
Overconfidence in findings of a single study	Cumulative thinking; findings seen as preliminary indications and part of a larger empirical context
Focus on novel findings, discoveries	Focus on replicable, severely tested claims
Confident attitude, trust in data and statistics	Cautious attitude, acceptance of uncertainty and variation

Overall, data analysis will proceed in close synchrony with theoretical and background knowledge, and fewer decisions will be based on purely statistical criteria. Where appropriate, data-driven, technological aids will be accompanied by linguistically motivated, theory-driven decisions. Statistical modeling, which relies on technical indices developed by statisticians with no interest in or knowledge of the linguistic background, will co-exist with scientific, linguistically informed modeling, where attention to context and subject-matter considerations are driving the analysis of language data (see Tizón-Couto and Lorenz's article).

The abundance of procedures and techniques will be winnowed to a principled selection of defaults. This then will constitute the core agenda for methodological training and textbooks, and data work that is guided by clear objectives can proceed in an informed manner. Given certain unifying features of language data, multilevel (mixed-effects) regression may eventually be considered a useful default for addressing focused questions. The navigation between different analysis strategies will rely on guidance provided by "language data universals", a heuristic set of features that permeate many language data settings. Such features could include (i) aspects of data structure, for instance the hierarchical (clustered) grouping of observations (see Winter and Grice's contribution); (ii) distributional patterns, such as Zipfian frequency profiles; and (iii) linguistically meaningful correlations between (predictor) variables, for instance the association between the frequency and length of a word. These "universals" are by no means absolute, but describe tendencies: recurrent features that need to be given thought to and, where necessary, accounted for.

When it comes to the communication of data-based claims, researchers will aim for a conceptual exposition of findings, avoid incomprehensible jargon and keep in focus the primary linguistic concern of their investigation. Data analysis and presentation will be responsive to the underlying subject-matter questions, and the mode and style of presentation will respect the level of statistical literacy of the (intended) audience. For core analysis techniques, analysts and their linguistic audience will be able to communicate at eye level, based on a consensus on minimal standards for both productive and receptive communication skills.

Data and analysis code will be made available during peer review and after publication. This will foster an open data culture characterized by respectful interaction. Less experienced scholars must not fear methodological attacks on their analyses, which are instead seen as informing interim interpretations that may require future modification. The fact that data are published along with the narrative distilled from it will disburden the researcher from having to offer a definitive analysis of the empirical evidence at hand. This may also serve to weaken normative constraints during the review process, and may generate fruitful discussions on methodological issues within the community.

Overall, researchers will assume a more cautious attitude towards their empirical contributions and consider them as additions to a cumulative whole. It follows that claims about language, especially if unanticipated, will require independent confirmation, ideally based on multiple lines of evidence. This goes hand in hand with routine efforts to replicate data-based insights and a more critical evaluation of novel, isolated discoveries. Overall, then, with an increased awareness of the limitations of statistical assessments and a recalibration of the information value of detached claims, empirical work in the new paradigm will be carried out much more cautiously.

6 Concluding remarks

We certainly owe an apology to our readers for advancing a set of rather terse, sometimes vague, and surely overgeneralizing observations and statements. Some points we touched upon would require considerable elaboration to have any persuasive force, and admittedly, we have dealt with most issues at a highly abstract level. The partly polemic tone, however, was intentional: Our critical reading of Kuhn's theory, and our sketch of the next methodological paradigm in quantitative linguistics merely serve to set the scene for a more involved and detailed discussion of individual aspects. This part we leave to the individual contributions that constitute this special issue of *Linguistics*, and to future engagements with the issue. If we have left some readers in an argumentative mood, then we would consider this introduction to have been worthwhile. We can skip small talk during the next coffee break.

We have taken the liberty granted by the genre "Introduction to the special issue" to attempt to frame the replication crisis from a broader philosophy of science perspective. While this may have overstretched the epistemological means of our profession, we believe that it has allowed us to see the current methodological debate in a somewhat different light. On the assumption that Kuhn's *Structure of Scientific Revolutions* readily extends to the methodological realms of science, a hypothetico-deductive engagement with his theory would lead us to predict that a methodological revolution, and subsequent paradigm shift, is upon us. We could sit back and wait for the empirical evidence to settle the case. But let us not forget that, at this level of empirical observation, we are the data points, the participants – each of us a potential revolutionary. This twist of fate makes it perfectly legitimate for us to exploit researcher degrees of freedom to manipulate the outcome – at last. We should embrace this role and think about and discuss what the future of language data analysis could look like.

Acknowledgments: First ideas surrounding the larger project on the replication crisis in linguistics were discussed at the workshop "The 'quantitative crisis', cumulative science, and English linguistics", hosted at the 5th meeting of the *International Society for the Linguistics of English* in July 2018 at University College London. We would like to thank the conference organizers and the scientific advisory board for providing a venue where many important discussions were sparked, as well as the audience and other presenters not part of the current publication for direct input at an early stage. In addition, we would like to thank Julia Schlüter and the *Linguistics* reviewers for helpful feedback on this introductory article. *Linguistics* editor-in-chief Volker Gast deserves a special mention for repeatedly sharing his time to provide constructive criticism and encouragement throughout the publication process of the overall issue.

References

Aguilar-Sánchez, Jorge. 2014. Replicability of (socio)linguistics studies. *Journal of Research Design and Statistics in Linguistics and Communication Science* 1(1). 5–25.

Aguilar-Sánchez, Jorge. 2017. Copula + Adjective: An a-posteriori power analysis for the generalizability of results. *Journal of Research Design and Statistics in Linguistics and Communication Science* 4(2). 91–123.

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.

Baayen, R. Harald. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baayen, R. Harald, Douglas J. Davidson & Douglas M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59(4). 390–412.

Berez-Kroeker, Andrea L., Bradley McDonnell, Eve Koller & Lauren B. Collister (eds.). Forthcoming. *The open handbook of linguistic data management*. Cambridge, MA: MIT Press.

Berez-Kroeker, Andrea L., Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David I. Beaver, Shobhana Chelliah, Stanley Dubinsky, Richard P. Meier, Nick Thieberger, Karen Rice & Anthony C. Woodbury. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56(1). 1–18.

Berkson, Joseph. 1938. Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association* 33. 526–536.

Berkson, Joseph. 1942. Tests of significance considered as evidence. *Journal of the American Statistical Association* 37. 325–335.

Blech, Jörg. 2019. Professor Zufall [professor coincidence]. *Der Spiegel*. 20 April 2019. Available at: https://magazin.spiegel.de/SP/2019/17/163511563/index.html.

Brandt, Mark J., Hans Ijzerman, Ap Dijksterhuis, Frank J. Farach, Jason Geller, Roger Giner-Sorolla, James A. Grange, Marco Perugini, Jeffrey R. Spies & Anna van't Veer. 2014. The replication

recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology* 50. 217–224.
Brezina, Vaclav. 2018. *Statistics in corpus linguistics: A practical guide*. Cambridge: Cambridge University Press.
Brinton, Laurel J., Patrick Honeybone, Bernd Kortmann & Elena Seoane. 2019. Editorial. *English Language and Linguistics* 23(1). i–ii.
Cantos Gómez, Pascual. 2013. *Statistical methods in language and linguistic research*. Sheffield: Equinox.
Carroll, Aaron E. 2017. Science needs a solution for the temptation of positive results. *The New York Times*. 29 May 2017. Available at: https://www.nytimes.com/2017/05/29/upshot/science-needs-a-solution-for-the-temptation-of-positive-results.html.
Chalmers, Alan. 2013. *What is this thing called science?* Queensland: University of Queensland Press.
Chambers, Christopher D. 2013. Registered reports: A new publishing initiative at Cortex. *Cortex* 49. 609–610.
Church, Kenneth & Mark Liberman. 2021. The future of computational linguistics: Beyond alchemy. *Frontiers in Artificial Intelligence* 4. 625341.
Cohen, Jacob. 1994. The earth is round ($p < 0.05$). *American Psychologist* 49. 997–1003.
Cumming, Geoff. 2012. *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.
Desagulier, Guillaume. 2017. *Corpus linguistics and statistics with R*. Cham: Springer.
Eddington, David. 2015. *Statistics for linguists: A step-by-step guide for novices*. Newcastle: Cambridge Scholars.
Egbert, Jesse, Tove Larsson & Douglas Biber. 2020. *Doing linguistics with a corpus*. Cambridge: Cambridge University Press.
Fisher, Ronald A. 1925. *Statistical methods for research workers*. Edinburgh: Oliver & Boyd.
Fisher, Ronald A. 1956. *Statistical methods and scientific inference*. Edinburgh: Oliver & Boyd.
Flanagan, Joseph. 2017. Reproducible research: Strategies, tools, and workflows. In Turo Hiltunen, Joe McVeigh & Tanja Säily (eds.), *Big and rich data in English corpus linguistics: Methods and explorations*. Helsinki: VARIENG. http://www.helsinki.fi/varieng/series/volumes/19/flanagan/.
Garellek, Marc, Matthew Gordon, James Kirby, Wai-Sum Lee, Alexis Michaud, Christine Mooshammer, Oliver Niebuhr, Daniel Recasens, Timo B. Roettger, Adrian Simpson & Kristine Yu. 2020. Toward open data policies in phonetics: What we can gain and how we can avoid pitfalls. *Journal of Speech Science* 9(1). 3–16.
Gawne, Lauren & Andrea L. Berez-Kroeker. 2018. Reflections on reproducible research. In Bradley McDonnell, Andrea L. Berez-Kroeker & Gary Holton (eds.), *Reflections on language documentation 20 years after Himmelmann 1998*, 22–32. Honolulu: University of Hawai'i Press.
Gelman, Andrew & Erik Loken. 2014. The statistical crisis in science. *American Scientist* 102(6). 460–465.
Geman, Stuart & Mark Johnson. 2004. Probability and statistics in computational linguistics. A brief review. In Mark Johnson, Sanjeev P. Khudanpur Mari Ostendorf & Roni Rosenfeld (eds.), *Mathematical foundations of speech and language processing*, 1–26. New York: Springer.
Grant, Tim, Urszula Clark, Gertrud Reershemius, Dave Pollard, Sarah Hayes & Garry Plappert. 2017. *Quantitative research methods for linguists: A questions and answers approach for students*. London: Routledge.

Greenland, Sander. 2017. Invited commentary: The need for cognitive science in methodology. *American Journal of Epidemiology* 186(6). 639–645.
Greenland, Sander, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman & Douglas G. Altman. 2016. Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology* 31. 337–350.
Gries, Stefan Th. 2006. Some proposals towards a more rigorous corpus linguistics. *Zeitschrift für Anglistik und Amerikanistik* 54(2). 191–202.
Gries, Stefan Th. 2015. Some current quantitative problems in corpus linguistics and a sketch of some solutions. *Language and Linguistics* 16(1). 93–117.
Gries, Stefan Th. 2017. *Ten lectures on quantitative approaches in cognitive linguistics: Corpus-linguistic, experimental, and statistical applications*. Leiden: Brill.
Gries, Stefan Th. 2018. On over- and underuse in learner corpus research and multifactoriality in corpus linguistics more generally. *Journal of Second Language Studies* 1(2). 276–308.
Gries, Stefan Th. 2021. *Statistics for linguistics with R*. Berlin: Mouton de Gruyter.
Gries, Stefan Th. & Sandra C. Deshors. 2021. Statistical analyses of learner corpus data. In Nicole Tracy-Ventura & Magali Paquot (eds.), *The Routledge handbook of second language acquisition and corpora*, 119–132. London: Routledge.
Haspelmath, Martin & Sven Siegmund. 2006. Simulating the replication of some of Greenberg's word order generalizations. *Linguistic Typology* 10(1). 74–82.
Heise, Christian & Joshua M. Pearce. 2020. From open access to open science: The path from scientific reality to open scientific communication. *SAGE Open* 10(2). 2158244020915900.
Ioannidis, John P. A. 2005. Why most published research findings are false. *PLoS Medicine* 2(8). e124.
Jaeger, T. Florian. 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language* 59(4). 434–446.
Johnson, Keith. 2008. *Quantitative methods in linguistics*. Malden, MA: Blackwell.
Johnson, Daniel E. 2009. Getting off the GoldVarb standard: Introducing Rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass* 3(1). 350–383.
Johnson, Daniel E. 2014. Progress in regression: Why sociolinguistic data calls for mixed-effects models. Available at: http://www.danielezrajohnson.com/johnson_2014.pdf.
Johnson, Mark. 2009. How the statistical revolution changes (computational) linguistics. In Timothy Baldwin & Vaila Kordoni (eds.), *Proceedings of the EACL 2009 workshop on the interaction between linguistics and computational linguistics: Virtuous, vicious or vacuous?* 3–11. Athens: ACL. Available at: https://www.aclweb.org/anthology/W09-0103.pdf.
Kline, Rex B. 2013 [2004]. *Beyond significance testing: Statistics reform in the behavioral sciences*, 2nd edn. Washington, DC: American Psychological Association.
Köhler, Reinhard. 2005. Aims and methods of quantitative linguistics. In Reinhard Köhler, Gabriel Altmann & Rajmund G. Piotrowski (eds.), *Quantitative linguistics: An international handbook*, 1–16. Berlin & New York: Mouton de Gruyter.
Krug, Manfred & Julia Schlüter (eds.). 2013. *Research methods in language variation and change*. Cambridge: Cambridge University Press.
Krug, Manfred, Julia Schlüter & Anette Rosenbach. 2013. Introduction: Investigating language variation and change. In Manfred Krug & Julia Schlüter (eds.), *Research methods in language variation and change*, 1–14. Cambridge: Cambridge University Press.
Kruschke, John K. 2010. What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences* 14. 293–300.

Kuhn, Thomas S. [1962] 1996. *The structure of scientific revolutions*. Chicago: University of Chicago Press.
Larson-Hall, Jenifer & Richard Herrington. 2010. Improving data analysis in second language acquisition by utilizing modern developments in applied statistics. *Applied Linguistics* 31(3). 368–390.
Larson-Hall, Jenifer & Luke Plonsky. 2015. Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning* 65(s1). 127–159.
Larsson, Tove, Jesse Egbert & Biber Douglas. 2021. On the status of statistical reporting versus linguistic description in corpus linguistics: A ten-year perspective. *Corpora* 17(1).
Levshina, Natalia. 2015. *How to do linguistics with R*. Amsterdam & Philadelphia: John Benjamins.
Loerts, Hanneke, Wander Lowie & Bregtje Seton. 2020. *Essential statistics for applied linguistics: Using R or JASP*. London: Palgrave Macmillan.
Loken, Eric & Andrew Gelman. 2017. Measurement error and the replication crisis. *Science* 355(6325). 584–585.
Marsden, Emma, Kara Morgan-Short, Pavel Trovimovich & Nick C. Ellis. 2018. Introducing registered reports at *Language Learning*: Promoting transparency, replication, and a synthetic ethic in the language sciences. *Language Learning* 68(2). 309–320.
Mayo, Deborah G. 2018. *Statistical inference as severe testing: How to get beyond the statistics wars*. Cambridge: Cambridge University Press.
Meehl, Paul E. 1967. Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science* 34(2). 103–115.
Meehl, Paul E. 1990. Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry* 1(2). 108–141.
Mulder, Gerben. 2020. The new statistics for applied linguistics. *Dutch Journal for Applied Linguistics* 9(1/2). 79–96.
Munafò, Marcus R, Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware & John P. A. Ioannidis. 2017. A manifesto for reproducible science. *Nature Human Behaviour* 1. 0021.
Muthukrishna, Michael & Joseph Henrich. 2019. A problem in theory. *Nature Human Behaviour* 3. 221–229.
Nassaji, Hossein. 2012. Significance tests and generalizability of research results: A case for replication. In Graeme Porte (ed.), *Replication research in applied linguistics*, 92–115. Cambridge: Cambridge University Press.
Nicenboim, Bruno & Shravan Vasishth. 2016. Statistical methods for linguistic research: Foundational ideas – part II. *Language and Linguistics Compass* 10(11). 591–613.
Nickerson, Raymond S. 2000. Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods* 5. 241–301.
Norris, John M. & Lourdes Ortega. 2007. The future of research synthesis in applied linguistics: Beyond art or science. *TESOL Quarterly* 41(4). 805–815.
Nosek, Brian A., Jeffrey R. Spies & Matt Motyl. 2012. Scientific utopia II: Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science* 7(6). 615–631.
Oakes, Michael P. 1998. *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.
Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349(6251). 1–8.

Palacios Martínez, Ignacio M. 2020. Methods of data collection in English empirical linguistics research: Results of a recent survey. *Language Sciences* 78. 101263.
Paquot, Magali & Marcus Callies. 2020. Promoting methodological expertise, transparency, replication, and cumulative learning: Introducing new manuscript types in the *International Journal of Learner Corpus Research*. *International Journal of Learner Corpus Research* 6(2). 121–124.
Paquot, Magali & Luke Plonsky. 2017. Quantitative research methods and study quality in learner corpus research. *International Journal of Learner Corpus Research* 3(1). 61–94.
Peels, Rik. 2019. Replicability and replication in the humanities. *Research Integrity and Peer Review* 4.2. https://doi.org/10.1186/s41073-018-0060-4.
Plonsky, Luke (ed.). 2015. *Advancing quantitative methods in second language research*. New York: Routledge.
Plonsky, Luke & Susan Gass. 2011. Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning* 61(2). 325–366.
Podesva, Robert J. & Devyani Sharma (eds.). 2014. *Research methods in linguistics*. Cambridge: Cambridge University Press.
Porte, Graham. 2015. Replication research in quantitative research. In James Dean Brown & Christine Combe (eds.), *Research in language teaching and learning*, 140–145. Cambridge: Cambridge University Press.
Porte, Graham & Kevin McManus. 2019. *Doing replication research in applied linguistics*. New York: Routledge.
Rasinger, Sebastian M. 2013. *Quantitative research in linguistics: An introduction*. London: Bloomsbury.
Rietveld, Toni & Roeland van Hout. 2005. *Statistics in language research: Analysis of variance*. Berlin & New York: Mouton de Gruyter.
Roettger, Timo B. 2019. Researcher degrees of freedom in phonetic research. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 10(1). 1–27.
Roettger, Timo B., Bodo Winter & R. Harald Baayen (eds.). 2019. *Emerging data analysis in phonetic sciences*. Special issue of the *Journal of Phonetics*. 73. Available at: https://www.sciencedirect.com/journal/journal-of-phonetics/special-issue/10357FT5MD0.
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.
Schmidt, Frank L. 1996. Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods* 1(2). 115–129.
Schneider, Gerold & Max Lauber. 2019. *Statistics for Linguists: A patient, slow-paced introduction to statistics and to the programming language R*. Zurich: University of Zurich.
Smaldino, Paul A. 2019. Better methods can't make up for mediocre theory. *Nature* 575. 9.
Smaldino, Paul A. & Richard McElreath. 2016. The natural selection of bad science. *Royal Society Open Science* 3. 160384.
Sonderegger, Morgan, Michael Wagner & Francisco Torreira. 2018. *Quantitative methods for linguistic data*. Montreal: McGill University. Available at: http://people.linguistics.mcgill.ca/~morgan/book/index.html.
Stark, Philip B. & Andrea Saltelli. 2018. Cargo-cult statistics and scientific crisis. *Significance* 15(4). 40–43.

Vasishth, Shravan & Bruno Nicenboim. 2016. Statistical methods for linguistic research: Foundational ideas – part I. *Language and Linguistics Compass* 10(8). 349–369.

Vasishth, Shravan, Bruno Nicenboim, Mary E. Beckman, Fangfang Li & Eun Jong Kong. 2018a. Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics* 77. 147–161.

Vasishth, Shravan, Daniela Mertzen, Lena A. Jäger & Andrew Gelman. 2018b. The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language* 103. 151–175.

Wallis, Sean. 2021. *Statistics in corpus linguistics research: A new approach*. London: Routledge.

Wicherts, Jelte M., Coosje L. S. Veldkamp, Hilde E. M. Augusteijn, Marjan Bakker, Robbie C. M. van Aert & Marcel A. L. M. van Assen. 2016. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology* 7. 1832.

Wieling, Martijn, Josine Rawee & Gertjan von Noord. 2018. Reproducibility in computational linguistics: Are we willing to share? *Computational Linguistics* 44(4). 641–649.

Winter, Bodo. 2011. Pseudoreplication in phonetic research. In Wai-Sum Lee & Eric Zee (eds.), *Proceedings of the 17th International Congress of the Phonetic Sciences*, 2137–2140. Hong Kong: City University of Hong Kong.

Winter, Bodo. 2019. *Statistics for linguists: An introduction using R*. New York: Routledge.

Yong, Ed. 2018. Psychology's replication crisis is running out of excuses. *The Atlantic*. 19 November 2018. Available at: https://www.theatlantic.com/science/archive/2018/11/psychologys-replication-crisis-real/576223/.

Ziliak, Stephen T. & Deirdre N. McCloskey. 2008. *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. Ann Arbor: University of Michigan Press.