UNIVERSITY OF BIRMINGHAM
University of Birmingham Research at Birmingham

Observation, experimentation, and replication in linguistics

Grieve, Jack

DOI:
10.1515/ling-2021-0094

License:
None: All rights reserved

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Grieve, J 2021, 'Observation, experimentation, and replication in linguistics', *Linguistics*, vol. 59, no. 5, pp. 1343-1356. https://doi.org/10.1515/ling-2021-0094

Link to publication on Research at Birmingham portal

General rights
Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

•Users may freely distribute the URL that is used to identify this publication.
•Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.
•User may use extracts from the document in line with the concept of 'fair dealing' under the Copyright, Designs and Patents Act 1988 (?)
•Users may not further distribute the material nor use it for the purposes of commercial gain.

Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.

When citing, please reference the published version.

Take down policy
While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

If you believe that this is the case for this document, please contact UBIRA@lists.bham.ac.uk providing details and we will remove access to the work immediately and investigate.

Download date: 19. Apr. 2024

Jack Grieve*

Observation, experimentation, and replication in linguistics

https://doi.org/10.1515/ling-2021-0094
Received June 26, 2020; accepted May 29, 2021; published online June 16, 2021

Abstract: In this paper, I propose that replication failure in linguistics may be due primarily to inherent issues with the application of experimental methods to analyze an inextricably social phenomenon like language, as opposed to poor research practices. Because language use varies across social contexts, and because social context must vary across independent experimental replications, linguists should not be surprised when experimental results fail to replicate at the expected rate. To address issues with replication failure in linguistics, and to increase methodological rigor in our field more generally, I argue that linguists must use experimental methods carefully, keeping in mind their inherent limitations, while acknowledging the scientific value of observational methods, which are often the only way to pursue basic questions in our field.

Keywords: descriptive linguistics; empirical linguistics; experimental linguistics; observational linguistics; research design in linguistics

1 Replication failure in linguistics

Over the past decade, there has been growing concern across the social sciences, first and foremost in psychology, that published results are not being replicated at the expected rate, calling into question the generalizability of research in these fields (Shrout and Rodgers 2018). Various solutions for this "replication crisis" have been proposed, primarily involving improvement of specific research practices, including the adoption of more rigorous approaches to statistical analysis, larger sample sizes, and greater openness in hypotheses, data, methods, and results (e.g., Asendorpf et al. 2013; Finkel et al. 2017; Maxwell et al. 2015; Shrout and Rodgers 2018; Tackett et al. 2019). Over the last few years, linguists have begun to accept that our field likely suffers from similar issues, leading to calls for the adoption of better research practices in linguistics (e.g., Roettger 2019; Sönning

***Corresponding author: Jack Grieve**, Department of English Language and Linguistics, University of Birmingham, Frankland Building, Edgbaston, Birmingham B15 2TT, UK,
E-mail: J.Grieve@bham.ac.uk

and Werner this issue), including pre-registration (Roettger this issue). Clearly, any effort to encourage methodological rigor in linguistics should be welcome, but what is unclear is whether improving specific research practices is sufficient to address excessive replication failure in linguistics, or if this issue points to a fundamental challenge with the nature of linguistic inquiry.

In this paper, assuming that linguistics has issues with replication like other social sciences, I propose that failed replications in linguistics may result more from an inherent and general limitation with experimentation for the study of language than from issues with specific research practices. After reviewing the main empirical research paradigms in linguistics, I argue that we should expect experimental results will often fail to replicate, even when best and consistent research practices are followed, because language is an inextricably social phenomenon, making it impossible for linguists to fully control social context across independent replications. As opposed to the physical sciences, where the parameters of an experiment can be kept stable across replications, every independent replication of a linguistic experiment is a unique social event, making exact replication impossible. I therefore conclude by calling for linguists to apply and interpret the results of experimental research with care and to recognize the fundamental importance of observational research in linguistics.

2 Empirical approaches to linguistics

Linguists seek to understand language through diverse methodologies, ranging from rationalist introspection into the nature of general linguistic knowledge to empirical investigation into specific forms of language use (Krug et al. 2013; Podesva and Sharma 2013). Research in empirical linguistics is often comparative – looking for relationships between dependent linguistic variables and an almost endless array of independent linguistic and extra-linguistic variables, including independent variables related to time, language, setting, mode, task, and speaker. Making meaningful comparisons, however, is challenging when studying a social phenomenon like language, any instance of which is unique, produced at a specific point in time and in a specific social context. To make sense of such a complex situation, linguists attempt to compare language knowledge and use in many different ways. At the most basic level, a distinction can be drawn between *experimental* and *observational* approaches to linguistics (Klavan and Divjak 2016).

In general, experimental linguists collect data by directly manipulating the context in which data is produced, exposing subjects to specific stimuli so as to allow for the causal effect of that intervention on a measure of language production

or perception to be isolated. Manipulation, however, does not ensure confounding factors are controlled, as there may be unknown characteristics that distinguish between the groups under comparison that could be responsible for the observed differences in the dependent linguistic variable under analysis. *Randomized experiments* therefore assign subjects at random to groups for comparison before the experimental intervention, where each group is exposed to different levels of the independent variable under analysis, so as to balance any unknown characteristics that might otherwise distinguish between those groups (Sadish et al. 2002). By combining manipulation and random assignment, randomized experiments have the potential to directly test for causality, at least within the context of that experiment. For example, Gibson et al. (2017) found that ungrammatical utterances are more likely to be interpreted as meaningful when produced in a non-native accent, based on a series of three randomized experiments, where randomized groups of native speakers of English were asked to interpret utterances produced in native and non-native accents.

Not all experiments, however, are randomized. First, a distinction can be drawn between randomized *between-subject designs*, of the type just described, and non-randomized *within-subject designs*, also known as *repeated measures designs,* where each subject is exposed to all levels of the independent variable being manipulated, thereby controlling for differences across groups and allowing for causation to be tested directly (Charness et al. 2012). For example, Han et al. (2020) investigated the syntactic status of null objects in Korean based on two within-subjects experiments, where subjects were asked to judge if a range of sentences, both with and without null objects, accurately represented a specific situation, as described in a longer text, finding evidence for an ellipsis-based analysis. In general, there are advantages and disadvantages to both designs, and the choice should depend primarily on the nature of the hypotheses being investigated (Charness et al. 2012). Crucially, within-subjects designs are generally more economical, allowing for increased statistical power with fewer subjects, but there is a risk that subjects may adjust their behavior, consciously or subconsciously, across levels, for example, based on their perception of the goals of the experiment. Although within-subjects designs appear to be far more common in linguistics, between-subjects designs with randomization are therefore often preferred in experimental research more generally, all things being equal (Charness et al. 2012), because they naturally control for more confounds, despite what is sometimes claimed in discussions of experimental methods in linguistics (e.g., Arunachalam 2013).

Second, *quasi-experiments* involve manipulation with pre-existing but comparable groups, where the researcher only controls the experimental intervention (Sadish et al. 2002). For example, Alrabai (2014) compared the effect of

motivational teaching strategies on a measure of language learner achievement by varying instruction across otherwise comparable classrooms. Crucially, because they work with pre-existing groups, the relationships identified in quasi-experiments are generally less likely to be causal than in other experimental designs: even when best research practices are followed, there may be unknown characteristics that distinguish between the groups under comparison that could explain any detected effects. For example, variation in achievement in a classroom experiment could be due to other differences between the groups of subjects being compared, including factors that would be difficult to measure or control such as evolving classroom dynamics. For this reason, quasi-experiments are often used to test hypotheses where other experimental designs would be difficult to implement, for example, due to limited resources or ethical concerns (Remler and Van Ryzin 2015). There are, however, more principled reasons to prefer a quasi-experimental approach, at least in linguistics, where we are often interested in *contextualized* research questions that would be impossible to answer through more controlled experiments, such as the effect of different language teaching strategies *in the classroom*.

This trade-off between control and naturalness is closely related to the distinction between *internal validity* and *external validity* in experimental design. On the one hand, internal validity is the degree to which a cause-and-effect relationship has been directly established given the design of an experiment (Sadish et al. 2002). Randomization and manipulation help maximize the internal validity of experiments, as do other aspects of research design (e.g., sufficient sample size, appropriate statistical analysis). Because they analyze pre-existing groups, the internal validity of quasi-experiments is generally lower than for other types of experiments (Remler and Van Ryzin 2015). On the other hand, external validity is the degree to which the results of an experiment generalize to target populations in the real world. Because they are conducted in highly controlled contexts, there are always questions about the external validity of experiments (Sadish et al. 2002). For example, a causal relationship identified in an experiment may have relatively little importance in the real world, where other factors may be far more important. These types of concerns will often be reduced for quasi-experiments, as they are conducted in a more natural context with pre-existing groups. Nevertheless, quasi-experiments are still far from natural, as they involve researchers creating variation that would not otherwise exist so as to isolate specific cause-and-effect relationships.

In contrast to experiments, an observational approach to linguistics involves describing how language varies naturally – for example, across languages, dialects, registers, and speakers. The fundamental difference between these two approaches is that experimental linguists directly vary the independent variable or

variables across the groups under comparison to see how the dependent linguistic variable changes, whereas observational linguists describe how the value of the dependent linguistic variable changes across groups that vary naturally in terms of the independent variable or variables of interest, without any intervention from the linguist. As a consequence, an observational approach is often the only choice in linguistics because the independent variables of interest cannot be assigned to subjects directly via intervention. This includes the comparative analysis of languages and dialects, as well as the analysis of many other social variables, precluding the direct use of experimentation for pursuing many important lines of research in linguistics. For example, a linguist can bring speakers of different languages or dialects into the lab to compare some aspect of language production or perception, but such analyses will necessarily be observational, as the linguist cannot directly control the social background of speakers.

For this reason, linguistics has traditionally been an observational field of study. Observation has long been the basis of building descriptive grammars and dictionaries, which are some of the oldest and most basic endeavors that can be described as linguistic research. Observation was also the basis of traditional philology, from which modern linguistics emerged, as well as its many important discoveries, including, for example, the existence and structure of the Indo-European language family. Similarly, observation is the basis of many modern fields of linguistics, including corpus linguistics, discourse analysis, historical linguistics, sociolinguistics, dialectology, and typology – fields that have made major contributions to linguistics, allowing for important theories and models of language to be developed and assessed, generally without any reliance on experimentation.

In addition to the types of questions that tend to drive linguistic research, observational studies are so common in our field because language is so easy to observe. Even the strongest form of observational research, which involves absolutely no intervention from the researcher on the act of data production, essentially maximizing external validity (Rosenbaum 2002), is often possible in linguistics, because language is often naturally recorded at the time of production and made publicly available as part of the social context in which it was produced. This situation allows linguists to study language by directly sampling instances of real language use from these social contexts (e.g., texts) without effecting the form of the language being observed, much like how an astronomer can study the universe. In particular, this type of true observational linguistics underlies research in corpus linguistics (Biber 1993; McEnery and Wilson 1996) and discourse analysis (Coulthard 2014), as well as considerable research across the breadth of linguistics that draws on these methodologies.

Of course, not all varieties of language are naturally recorded at the time of production, including most forms of speech. In such cases, the linguist must intervene in the social context in which language is produced to collect or even directly elicit data from informants – for example, through surveys, interviews, and ethnography. Although such approaches to data collection are less natural than the approaches adopted in corpus linguistics and discourse analysis, they generally qualify as forms of observational research, because the linguist does not and often cannot directly manipulate the independent variables of interest (e.g., dialect, language). The inherent limitations of elicitation-based approaches to observation are generally acknowledged by linguists working with these methods, who often go to considerable lengths to mitigate the effect of researcher intervention during the act of data collection. For example, field linguists embed themselves in remote communities for years so as to accurately describe isolated languages (e.g., Everett 2009), while sociolinguists conduct carefully designed interviews so as to elicit naturalistic language from informants by limiting their attention paid to speech (Labov 1984). Sociolinguists have even labelled this effect as the *Observer's Paradox*, and it is generally accepted that this effect can never be completely overcome within the context of sociolinguistic interviews and surveys (Labov 1972). Despite this limitation, there are clearly advantages to elicitation compared to simple observation in linguistics, depending on one's research goals, especially in terms of having greater control on the context in which language is produced.

Regardless of the specific approach to observational research that is adopted, observational studies in linguistics often employ some form of random sampling, so as to allow for valid generalizations to be made about populations of speakers and texts. Much like random assignment in experimental research, selecting speakers or texts at random from a population balances unknown sources of variation in the sample. Assuming sufficiently large random samples are obtained, the results of comparative observational research should hold on average for the populations under analysis, allowing for meaningful comparisons to be made. Random sampling, however, does not allow for causality to be tested directly, as unknown factors that are distributed unevenly across the populations would be preserved in the random samples being compared. Observational studies can therefore identify real differences between populations, but they cannot directly determine why those differences exist, because the independent variables under analysis cannot be manipulated.[1] Even if large numbers of independent variables

1 A natural experiment is a special type of observational study where an argument can be made for natural randomization (Dunning 2012). Specifically, natural experiments are characterized by randomization without manipulation, as opposed to real experiments, which always involve

are factored into the analysis, we can never be certain that other important but unknown variables are not responsible for the observed variation. Instead, causative links must be considered in light of the results of the study and the possible existence of confounding factors that have gone uncontrolled. This is the primary limitation of adopting an observational approach to linguistics.

To illustrate the difference between experiment and observation in linguistics, consider two hypothetical studies involving reading times. In the first study, researchers are interested in understanding the relationship between reading time and font size. They recruit subjects from a population, divide the subjects into two random groups, and then give both groups the same text to read, varying the font size. If they observe a difference, they can be reasonably confident that font size directly affects reading times, at least within the context of that study. This is an experiment. In the second study, researchers are interested in understanding the relationship between reading time and sex. They recruit subjects from a population, divide the subjects into two groups based on their self-declared sex, and then give both groups the same text to read. However, if they observe a difference, although they can be reasonably confident the difference is real, at least within the context of that study, they cannot be reasonably confident, based on these results, that variation is sex *causes* variation in reading times, because unknown factors that differentiate between these groups may have caused the observed variation in reading times. Perhaps differences in cognitive abilities across sexes were observed, but this variation may also be due to other factors, for example, cultural differences in the way children of different sexes are educated. This is an observational study.

Although uptake of experimental and observational methods across linguistics largely and rightfully reflects the specific questions that researchers pursue, experimental research is sometimes presented as the optimal approach to linguistics because it allows for causal links to be tested directly (for a critical discussion, see Roettger et al. 2019). For example, Chomsky, who is the most influential proponent of this view of linguistic inquiry, has repeatedly argued that observational research, especially corpus-based research, does not qualify as scientific because it is non-experimental. Most notably, in an interview published in the journal *Intercultural Pragmatics*, Andor (2004) quotes Chomsky as saying

---

manipulation of the independent variables under analysis, and standard observational studies, which involve neither manipulation nor randomization. Natural experiments, however, are rare in general and appear to be especially uncommon in linguistics, although the term is sometimes used incorrectly to refer to observational studies more generally (e.g., Auer et al. 2000).

> Corpus linguistics doesn't mean anything. It's like saying suppose a physicist decides, suppose physics and chemistry decide that instead of relying on experiments, what they're going to do is take videotapes of things happening in the world and they'll collect huge videotapes of everything that's happening and from that maybe they'll come up with some generalizations or insights. Well, you know, sciences don't do this (Andor 2014: 97).

Despite such claims, this is exactly what many scientists do (for further discussion, see Divjak et al. 2017). Observation and exploration of natural phenomena is generally recognized as an indispensable part of scientific inquiry (Nilsen and Bowler 2020; Tukey 1977), not only as the basis of hypothesis generation across the physical and social sciences, but as the primary source of data for investigating causative relationships in several fields, including archaeology, astronomy, geology, and geography, where experimentation is often impossible because of the nature of the variables under analysis. Claims such as Chomsky's therefore not only erroneously limit the application of scientific methodologies in linguistics but the domains in which scientific discovery can be pursued.

In part, because of this rhetoric, there is considerable confusion surrounding the application of basic research paradigms in linguistics. For example, it is easy to find recent studies published in major linguistics journals that present observational research as experimental. In general, these studies compare performance on various production and perception tasks by subjects from different social groups, including comparisons based on first language (e.g., Vanek and Mertins 2020), dialect (e.g., Montrul et al. 2015; Walker 2019; Yun and Kang 2019), and multilingualism (e.g., Werkmann Horvat et al. 2021). These studies are valid and informative, but they are not experiments: the researchers did not manipulate, and could not have manipulated, the social variables under analysis, precluding the direct isolation of cause and effect relationships. These types of misunderstandings have even found their way into research handbooks. For example, the introduction to experimental methods in linguistics presented in Abbuhl et al. (2013) opens with a hypothetical example where acceptability judgments are collected from L1 and L2 speakers. Although acceptability judgments are regularly embedded within experimental research designs (see Schutze 2016),[2] this example would not qualify because of the nature of the independent extra-linguistic variable under analysis. If linguistics as a field is truly dedicated to conducting rigorous and replicable scientific research, we must use standard scientific terminology accurately, acknowledging the strengths

2 It should be stressed that the nature of the dependent linguistic variable under analysis does not determine whether or not a given study is experimental: in general, any measure of language production or perception can be embedded into an observational or an experimental study. Rather, the status of the study is determined by how the levels of the independent variable are defined (i.e., naturally or via researcher manipulation).

and weaknesses of the methodologies we apply, while taking advantage of the full range of research paradigms at our disposal, motivated wholly by the nature of the specific research questions we pursue.

3 Replication failure across research paradigms

The choice between observational and experimental research paradigms represents a trade-off between naturalness and control, with clear advantages and disadvantages associated with both approaches. An observational approach allows for differences between real groups to be described accurately, but it does not guarantee these differences exist due to membership in these groups, whereas an experimental approach allows for cause-and-effect relationships to be identified, but it does not guarantee these relationships matter in the real world. These differences also have important ramifications for understanding issues around replication and generalizability in our field. Specifically, there are at least three basic reasons why a linguistic analysis may fail to replicate as expected, depending crucially on the research paradigm adopted.

First, both observational and experimental studies can fail to replicate at the expected rate because the design or implementation of the study (or the replication) was technically flawed. In such cases, we should not expect that the results of the study will generally replicate. These issues can include small sample size, measurement error, problematic statistical analysis, and dishonest research practices. Such threats to the validity of a study's design should be minimized regardless of the research paradigm adopted. These types of issues have been the focus of most discussions on how to improve replication rates in linguistics (Roettger 2019, this issue; Sönning and Werner this issue), including through preregistration, but it is important to acknowledge that there are other reasons why studies may fail to replicate.

Second, both observational and experimental studies can fail to replicate due to unknown variation in the population or populations being analyzed across replications. Random sampling allows for generalizations to be made within the context of a given study, but it does not ensure that the samples analyzed across independent studies come from the same population or populations. Two studies that follow an otherwise identical research design may produce different results if they sample people or texts from different populations, even when these populations appear to be the same superficially. No matter how much care is taken to match populations across replications, the comparability of independent samples is always up for debate, especially as language varies systematically and substantially across social contexts. Consequently, linguists cannot, and generally do not, assume that results

obtained for one language, era, dialect, register, or person hold more generally. Whenever a study fails to replicate, it is possible that this is because the replication focused on a different population. Crucially, in such cases, replication may have failed, but the results of each study can still generally be independently valid and insightful, identifying potentially unappreciated sources of variation in the phenomena of interest across the populations under analysis.

Finally, experimental studies (as well as observational studies that rely on elicitation) can fail to replicate because the context in which data was collected differs across replications. For example, variation in the researchers conducting the experiment, the setting of the experiment, and the time of the experiment, all represent variation in the experimental context and by extension in the social context in which language data was collected. This variation can then lead to variation in results. Such threats to replicability are distinct from threats to internal validity, as each experiment can be independently valid, and vary only in ways that are necessary for *independent* replications to be conducted. An experiment can therefore fail to replicate, even when best research practices are followed consistently, because social context inevitably changes across independent replications. Recent research has shown that contextual factors can affect experimental results in psychology (Van Bavel et al. 2016), but such considerations seem especially important in linguistics (e.g., see Hay et al. 2009), where we know variation in communicative context substantially and systematically affects linguistic form (Biber and Conrad 2019). Alternatively, in true observational studies, variation in context is just as important, but it amounts to variation in the population under analysis (e.g., texts sampled from two different registers represent two different populations of texts), as the researcher has had no effect on the context in which the language under analysis was produced.

Taken together, these final two sources of replication failure effectively undermine the interpretability of experimental research in linguistics, as well as elicitation-based research more generally. Even after the design of a study has been validated and applied consistently across replications, the source of lower-than-expected replication rates is always uncertain: it may be due to real but unknown variation in the populations being analyzed, pointing to new and important insights, but it may also be due to unintended variation in the context in which data is collected, an artefact of researcher intervention. This is why there is concern about a replication crisis in linguistics: if failed replications only pointed to real but unknown variation in the populations under analysis, as they do in true observational research, this would be considered scientific discovery. Any instance of language is situated in the real world, even the instantiation of a thought, and its form is linked to the social context in which it was produced. The essence of experimental research, however, is to manipulate this social context,

which will necessarily have an effect on language production and perception. Because independent experimental replications *must* involve variation in social context, replication rates will generally be lower than expected, depending broadly on the social sensitivity of the linguistic phenomena under analysis. This is why replication failure in linguistics cannot simply be addressed by adopting better experimental research practices: replication failure is an inevitable product of using experimentation to probe a highly social phenomenon like language.

4 In defense of observational linguistics

Observational methods have long been the basis of linguistics and are necessary to pursue many of the core research questions that have driven scientific research on language for centuries. Although observational studies cannot directly test for causal links, as unknown factors may vary across the groups being compared, the relationships identified through observation can still be causal, and their causal status can still be inferred through the careful interpretation of observational results, as the history of our own field clearly attests. The status of causal links in observational research is always up for debate, but so too is the status of causal links in experimental research: the question only shifts from whether the study has identified a causal link to whether the causal link matters in the real world. Even spurious relationships identified in observational research can point to causative relationships in the real world,[3] which may otherwise have gone unnoticed. Ultimately, spurious relationships found in nature can be more informative – and more likely to replicate – than causative relationships found in the lab.

If we are to truly address issues around replication and generalizability in linguistics, we must embrace the value of observational research for extending our understanding of language, taking advantage of large corpora and modern methods in data science and causal inference. When combined with robust and open research practices, an observational approach to linguistics provides us with a pathway to make meaningful comparisons, where we can be confident that our findings represent valid descriptions of natural phenomena, as opposed to the by-products of researchers intervening in a social domain, thereby addressing

3 Despite how the term is often used, it is important to acknowledge that spurious relationships are not necessarily untrue: these correlations can be robust and replicate without issue. They are spurious because there is no causative link between the variables: in particular, a correlation may exist because both variables under analysis are in a causative relationship with a third unobserved variable. The analysis of spurious correlations can therefore lead to the identification of causative relationships, for example, by considering what other factors might cause the spurious relationships that have been observed.

replication failure in its most pernicious form. Of course, experimentation is an extremely important part of linguistic inquiry, used to test specific claims of cause and effect across much of linguistics. However, as a field, we must stop elevating experimental research above observational research (Klavan and Divjak 2016; Roettger et al. 2019). The choice between experiment and observation should be driven entirely by the research questions we ask: one approach is not more scientific than the other, and, in many cases, observation is the only option, not because of practical concerns, but because the variables under analysis, like the first language or dialect of a speaker, cannot be manipulated experimentally. Furthermore, we must remember that causal links identified in rigorously conducted experiments are only necessarily valid within the context of those experiments. Those same links may be of very little relevance for explaining phenomena as observed in the real world, which is ultimately our goal as scientists.

Besides, science is not just the search for cause-and-effect relationships through hypothesis testing: science is about understanding the world around us (Glass and Hall 2008). And linguistics is about understanding language: how it is structured, how it is acquired, how it is used, how it varies over time and across society, and how it is represented in the mind. Big questions like these cannot be answered simply by testing causal links involving variables we can manipulate. Language is a complex social system: if we take one part of language out of this system and analyze it in isolation, we cannot hope to fully understand how it operates within the larger system, or how the properties of the system emerge from the interaction of its constituents (Kretzschmar 2015). Language is a history of utterances, every one unique, every one produced by a unique person, with a unique purpose, at a unique moment in time. It belies the social reality of language to attempt to reduce something so intricate, intentional, and inexorable down to a set of rules.

Acknowledgments: I would like to thank Dagmar Divjak, Jason Grafmiller, Nick Groom, Susan Hunston, Gary Lupyan, Akira Murakami, Elliot Murphy, Lukas Sönning, Timo Roettger, Rory Turnbull, Emily Waibel, Valentin Werner, and Bodo Winter for discussing this paper and its contents with me.

References

Abbuhl, Rebekha, Susan Gass & Alison Mackey. 2013. Experimental research design. In Robert J. Podesva & Devyani Sharma (eds.), *Research methods in linguistics*, 116–134. Cambridge: Cambridge University Press.

Alrabai, Faikieh. 2014. The effects of teachers' in-class motivational intervention on learners' EFL achievement. *Applied Linguistics* 37. 307–333.

Andor, József. 2014. The master and his performance: An interview with Noam Chomsky. *Intercultural Pragmatics* 1. 93–111.

Arunachalam, S. 2013. Experimental methods for linguists. *Language and Linguistics Compass* 7(4). 221–232.

Asendorpf, Jens B., Mark Conner, Filip De Fruyt, Jan De Houwer, Jaap J. A. Denissen, Klaus Fiedler, Susann Fiedler, David C. Funder, Reinhold Kliegl, Brian A. Nosek, Marco Perugini, Brent W. Roberts, Manfred Schmitt, Marcel A. G. van Aken, Hannelore Weber & Jelte M. Wicherst. 2013. Recommendations for increasing replicability in psychology. *European Journal of Personality* 27. 108–119.

Auer, Edward T., Lynn E. Bernstein & Paula E. Tucker. 2000. Is subjective word familiarity a meter of ambient language? A natural experiment on effects of perceptual experience. *Memory & Cognition* 28(5). 789–797.

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8. 243–257.

Biber, Douglas & Susan Conrad. 2019. *Genre, register and style*, 2nd edn. Cambridge: Cambridge University Press.

Charness, Gary, Uri Gneezy & Michael Kuhn. 2012. Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization* 81(1). 1–8.

Coulthard, Malcolm. 2014. *An introduction to discourse analysis*. London: Routledge.

Divjak, Dagmar, Tomaž Erjavec & Serge Sharoff. 2017. Slavic computational and corpus linguistics. *Journal of Slavic Linguistics* 25(2). 171–199.

Dunning, Thad. 2012. *Natural experiments in the social sciences: A design-based approach*. Cambridge: Cambridge University Press.

Everett, Daniel L. 2009. *Don't sleep, there are snakes: Life and language in the Amazonian jungle*. London: Profile Books.

Finkel, Eli J., Paul W. Eastwick & Harry T. Reis. 2017. Replicability and other features of a high-quality science: Toward a balanced and empirical approach. *Journal of Personality and Social Psychology* 113. 244–253.

Gibson, Edward A., Caitllin M. Tan, Richard Futrell, Kyle Mahowald, Lars Konieczny, Barbara Hemforth & Evelina Fedorenko. 2017. Don't underestimate the benefits of being misunderstood. *Psychological Science* 28(6). 703–712.

Glass, David J. & Ned Hall. 2008. A brief history of the hypothesis. *Cell* 134. 378–381.

Han, Chung-hye, Kyeong-min Kim, Keir Moulton & Jeffrey Lidz. 2020. Null objects in Korean: Experimental evidence for the argument ellipsis analysis. *Linguistic Inquiry* 51 2. 319–340.

Hay, Jennifer, Katie Drager & Paul Warren. 2009. Careful who you talk to: An effect of experimenter identity on the production of the NEAR/SQUARE merger in New Zealand English. *Australian Journal of Linguistics* 29(2). 269–285.

Klavan, Jane & Dagmar Divjak. 2016. The cognitive plausibility of statistical classification models: Comparing textual and behavioral evidence. *Folia Linguistica* 50(2). 355–384.

Krug, Manfred, Julia Schüler & Anette Rosenbach. 2013. Introduction: Investigating language variation and change. In Manfred Krug & Julia Schlüter (eds.), *Research methods in language variation and change*, 1–13. Cambridge: Cambridge University Press.

Kretzschmar, William A. Jr. 2015. *Language and complex systems*. Cambridge University Press.

Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia, PA: University of Pennsylvania Press.

Labov, William. 1984. Field methods of the project on linguistic change and language in use. In John Baugh & Joel Sherzer (eds.), *Readings in sociolinguistics*, 28–54. Englewood Cliffs, NJ: Prentice-Hall.

McEnery, Tony & Andrew Wilson. 1996. *Corpus linguistics*. Edinburgh: Edinburgh University Press.
Maxwell, Scott E., Michael Y. Lau & George S Howard. 2015. Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist* 70. 487–498.
Montrul, Silvina, Rakesh Bhatt & Roxana Girju. 2015. Differential object marking in Spanish, Hindi, and Romanian as heritage languages. *Language* 564–610. https://doi.org/10.1353/lan.2015.0035.
Nilsen, Erlend B., Diana E. Bowler & John D. C. Linnell. 2020. Exploratory and confirmatory research in the open science era. *Journal of Applied Ecology* 57(4). 842–847.
Remler, Dahlia K. & Gregg G. Van Ryzin. 2015. *Research methods in practice*. London: Sage.
Robert J. Podesva & Devyani Sharma (eds.). 2013. *Research methods in linguistics*. Cambridge: Cambridge University Press.
Roettger, Timo B. 2019. Researcher degrees of freedom in phonetic sciences. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 10. 1–27.
Roettger, Timo B., Bodo Winter & Harald Baayen. 2019. Emergent data analysis in phonetic sciences: Towards pluralism and reproducibility. *Journal of Phonetics* 73. 1–7.
Rosenbaum, Paul R. 2002. *Observational studies*, 2nd edn. New York: Springer.
Sadish, William R., Thomas D. Cook & Donald T. Campbell. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.
Schütze, Carson T. 2016. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Berlin: Language Science Press.
Shrout, Patrick E. & Joseph L. Rodgers. 2018. Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology* 69. 487–510.
Tackett, Jennifer L., Cassandra M. Brandes, Kevin M. King & Kristian E Markon. 2019. Psychology's replication crisis and clinical psychological science. *Annual Review of Clinical Psychology* 15. 579–604.
Tukey, John Wilder. 1977. *Exploratory data analysis*. Reading, MA: Addison-Wesley.
Van Bavel, Jay J., Peter Mende-Siedleckia, William J. Bradya & Diego A. Reineroa. 2016. Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences* 113. 6454–6459.
Vanek, Norbert & Barbara Mertins. 2020. Defying chronology: Crosslinguistic variation in reverse order reports. *Linguistics* 58(2). 569–603.
Walker, Abby. 2019. The role of dialect experience in topic-based shifts in speech production. *Language Variation and Change* 31(2). 135–163.
Werkmann Horvat, Anna, Marianna Bolognesi & Katrin Kohl. 2021. Creativity is a toaster: Experimental evidence on how multilinguals process novel metaphors. *Applied Linguistics*.
Yun, Suyeon & Yoonjung Kang. 2019. Variation of the word-initial liquid in North and South Korean dialects under contact. *Journal of Phonetics* 77. 100918.