

Diversität besser erfassen

Methodologische Entwicklungen in der experimentellen Linguistik

Renate Delucchi Danhier

Ausgehend von einigen philosophischen Überlegungen zur methodologischen Entwicklung der experimentellen Linguistik werden im vorliegenden Artikel zentrale Aspekte experimentellen Arbeitens in der Linguistik kritisch reflektiert. Der Fokus wird dabei auf das Spannungsfeld zwischen dem Streben nach generalisierbaren Erkenntnissen und der Berücksichtigung sprachlicher Vielfalt und Diversität gelegt. Es wird aufgezeigt, wie Diversität besser berücksichtigt und verstanden werden kann, während gleichzeitig die Methoden sauberer angewandt werden. Hierbei wird der Fokus auf die ersten drei Schritte in der Durchführung eines linguistischen Experiments gemäß dem NHTS-Modell gelegt: (1) die Formulierung von Hypothesen, (2) das Experimentdesign und (3) die Datensammlung. Es werden praktische Empfehlungen für die Durchführung experimenteller Studien unter Einhaltung von Mindeststandards in der aktuellen (Psycho)-Linguistik gegeben. Darüber hinaus dient dieser Beitrag der kritischen Reflektion und Diskussion des Experiments als effektive quantitative Methode zur Erkenntnisgewinnung in der Linguistik.

Einführung

Dieser Beitrag setzt sich zum Ziel, experimentelle Anwendungen in der Linguistik und ihre inhärenten Beschränkungen, insbesondere im Zusammenhang mit Vielfalt und Diversität, kritisch zu reflektieren. Hierfür wird zunächst die Entwicklung und Stellung der Linguistik als Wissenschaft im Laufe der letzten Jahrzehnte betrachtet. Im Anschluss daran werden die derzeit in der Psycholinguistik verbreiteten empirischen Methoden erörtert und der Frage nachgegangen, inwiefern sie dazu geeignet sind, Diversitätsaspekte zu berücksichtigen. Die hier diskutierten formalen Herausforderungen der experimentellen Methodik sind bereits mehrmals angesprochen worden (z.B. Baayen et al.; Jaeger; Vasisht et al.). Der Mehrwert die-

ses Beitrags liegt in der kombinierten Betrachtung von *Best Practices*-Empfehlungen mit den Chancen der Diversitätsforschung.

Die Stellung der Linguistik als Wissenschaft

Die Linguistik hat in den letzten Jahrzehnten nach dem Vorbild der Psychologie einen Paradigmenwechsel von den am Konstruktivismus orientierten, eher qualitativ-beobachtend arbeitenden Geisteswissenschaften hin zu einer stärker am positivistischen Paradigma der Naturwissenschaften orientierten quantitativ arbeitenden Sozialwissenschaft vollzogen. Ziel war, eine größere Überprüfbarkeit und Quantifizierbarkeit der Forschungsergebnisse zu erzielen und eine größere methodische Stringenz zu erreichen. Auch wenn diese Entwicklung zu zahlreichen neuen Erkenntnissen geführt hat, birgt sie vielfältige Herausforderungen.

Die Wertung, dass sich verschiedene Wissenschaftszweige in der Stringenz der angewandten Methoden unterscheiden, reicht mindestens bis zu Auguste Comtes »Hierarchie der Wissenschaften« im 19. Jahrhundert zurück. Abbildung 1 veranschaulicht humorvoll diese Hierarchie, wobei die traditionelle Linguistik am äußersten linken Rand der Disziplinen mit einem niedrigen Grad an methodischer Reinheit eingestuft wird.

Der Comic bezieht sich auf zwei reduktionistische Hypothesen, die besagen, dass zum einen alle Dinge denselben grundlegenden naturwissenschaftlichen Gesetzen folgen und zum anderen die in den Naturwissenschaften verbreiteten Vorgehensweisen eins zu eins auf die Geistes- und Sozialwissenschaften übertragbar wären. Die Naturwissenschaften, die auf dem Kontinuum weiter rechts angeordnet sind, werden häufig als »wissenschaftlicher« und prestigeträchtiger angesehen, weil sie auf diesen Gesetzen fußen.

Diese Hierarchie der Wissenschaften zeigt sich auch in der alltagsgebräuchlichen binären Unterscheidung zwischen den sogenannten »harten« und »weichen« Wissenschaften. Harte Wissenschaften (rechts in Abbildung 1) beschäftigen sich mit Naturphänomenen und versuchen objektive Gesetze und Prinzipien mittels quantitativer Methoden und experimenteller Ansätze zu beschreiben. Weiche Wissenschaften wie die Sozial- und Geisteswissenschaften (links), zu denen auch die Linguistik gezählt wird, untersuchen hingegen menschliches Verhalten, soziale Strukturen und kulturelle Aspekte. Sie verwenden traditionell qualitative Methoden wie Beobachtung, Interviews und Textinterpretation, um komplexe soziale Phänomene zu verstehen. Interessanterweise ist die Trennung der Disziplinen in einigen Sprachen durch unterschiedliche Begriffe ersichtlich (z. B. auf Englisch: *Sciences and Humanities*), während im Deutschen die Geisteswissenschaften ebenfalls als Wissenschaften bezeichnet werden.

Hierarchie der Wissenschaften

weiche Wissenschaften — stringendere Anwendung der Methodik → harte Wissenschaften

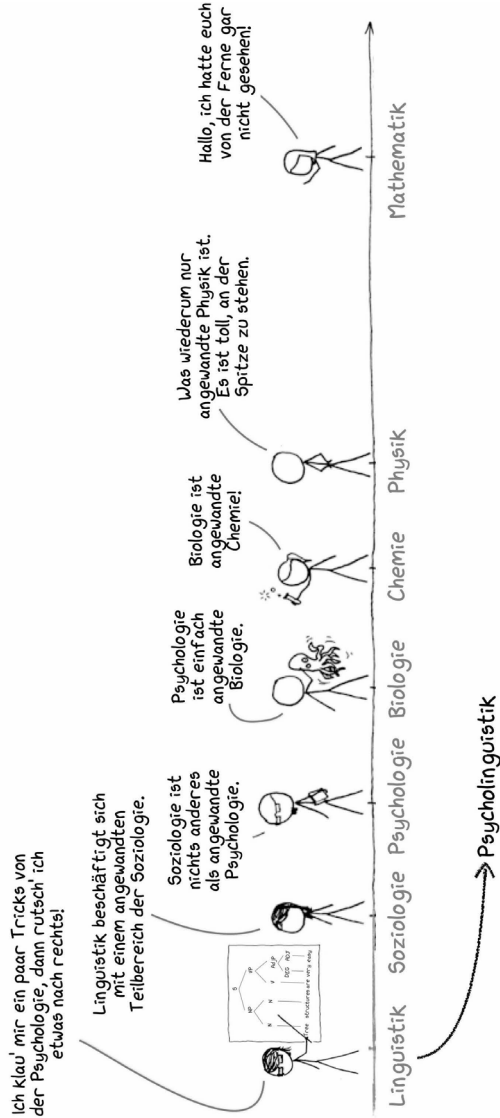


Abbildung 1: Renate Delucchi Danhier: Disziplinen geordnet nach methodologischer Stringenz (adaptiert von Munroe 2008), eigene Darstellung, lizenzfrei.

Aus diesem Grund gibt es seit einigen Jahrzehnten in der Linguistik, der Psychologie und ihren Nachbarwissenschaften die Bestrebung, Methoden anzuwenden,

die dem positivistischen Paradigma der Naturwissenschaften des 17. Jahrhunderts folgen. Dieser quantitative Turn hat zu einem großen Erkenntnisgewinn geführt, birgt jedoch Herausforderungen, da menschliche Forschungssubjekte und die damit einhergehende Komplexität andere Anforderungen an die Methodik stellen als die Erforschung naturwissenschaftlicher Gesetzmäßigkeiten und daher eine eigene Methodik erfordern (Anderson 177).

Mit diesem Paradigmenwechsel in der Linguistik und den damit einhergehenden neuen methodischen Möglichkeiten, kamen auch neue Forschungsfelder dazu, wie z.B. die kognitiven Prozesse der Produktion, Rezeption und Entwicklung sprachlicher Strukturen, was zur Entwicklung der neuen interdisziplinären Disziplin der Psycholinguistik geführt hat. Mithilfe von quantitativen Daten aus Beobachtungsstudien, Experimenten und Pseudoexperimenten und deren Auswertung können theoretisch fundierte Hypothesen über das Verhalten von Sprechenden anhand empirischer Beobachtungen bestätigt oder widerlegt werden. Die experimentelle Forschung hat sich für viele (Psycho-)Linguist*innen als optimaler Ansatz etabliert, linguistische Fragestellungen anzugehen, da sie direkte Tests kausaler Zusammenhänge ermöglichen soll (für eine kritische Diskussion vgl. Roettger et al.). Nach dieser Auffassung würden sich Forschungsbestrebungen nicht-experimenteller Natur lediglich auf »Briefmarkensammeln« beschränken, d.h. Beobachtungen zu katalogisieren und zu systematisieren, wie es das berühmte Zitat »*all science is either physics or stamp collecting*«, das Ernest Rutherford zugeschrieben wird, impliziert. Chomsky scheint zum Beispiel die Meinung zu teilen, dass das Experiment als wissenschaftlicher *golden standard* in der Linguistik gelten kann (Andor 97).

Die stärkere Orientierung der Psycholinguistik hin zu quantitativen und experimentellen Methoden und somit in Richtung auf »härteren« Wissenschaften ist, wie oben dargestellt u.a. auf den Einfluss der Psychologie zurückzuführen. In der psychologischen Forschung reicht die Tradition, linguistische Fragestellungen experimentell anzugehen, mehr als ein Jahrhundert zurück (Binet und Henri). Linguisten haben jedoch erst in den 1970er Jahren begonnen, Experimente und empirische Daten zu verwenden, um linguistische Hypothesen zu untersuchen (Köhler 2). Das könnte die großen Diskrepanzen in der Qualität und Tiefe der Methodenausbildung unter den in der Psycholinguistik tätigen Forschenden teilweise erklären.

Eine grundlegende Frage ist, ob die methodischen Forschungspraktiken und Standards der Naturwissenschaften überhaupt auf die Geistes- und Sozialwissenschaften angewendet werden können:

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. (Grieve 1344)

Experimentieren im eigentlichen Sinne ist auch in anderen Wissenschaftszweigen aufgrund der Natur der Phänomene und der Variablen unter Analyse unmöglich, beispielsweise in der Geologie oder der Astronomie. Solche Disziplinen, wie auch die Linguistik, haben eine lange Tradition von Beobachtungsstudien und Korrelationsstudien, die unser Verständnis der Welt immens weitergebracht haben.

Seit den frühen 2000er Jahren hat sich die Forschungsgemeinschaft in den Sozialwissenschaften und der Psychologie, aber auch in der Linguistik, zunehmend über die mangelnde Reproduzierbarkeit veröffentlichter Ergebnisse Sorgen gemacht. Die Wiederholung von Experimenten führt zu selten zu vergleichbaren Ergebnissen. Beispielsweise wurden in einer großen Metastudie hundert psychologische Experimente und korrelative Studien versucht zu replizieren. Lediglich 36 % bis 47 % (abhängig vom Replikationskriterium) der Studien konnten repliziert werden (Open Science Collaboration). Diese sogenannte Replikationskrise wirft generelle Fragen zur Verallgemeinerbarkeit von Forschung auf (Shrout und Rodgers; Roettger).

Entwicklung der Methodik in der aktuellen linguistischen Praxis

Obwohl es in der Natur der wissenschaftlichen Praxis liegt, dass manche Forschungsergebnisse letztlich im Rahmen des wissenschaftlichen Erkenntnisgewinns durch neue Beweise widerlegt werden, sollte dies ein langsamer Prozess sein, der den wissenschaftlichen Fortschritt im Laufe der Zeit widerspiegelt. Stattdessen scheitert eine alarmierende Anzahl von zeitnahen psychologischen und linguistischen Studien an der Replikation, was auf methodologische Probleme hinweist, da bereits etwas mit den ersten berichteten Ergebnissen nicht zu stimmen scheint. Bewertungen des modernen wissenschaftlichen Status von Disziplinen zeigen, dass signifikante statistische Ergebnisse in den Geistes- und Sozialwissenschaften nach Kontrolle der Anzahl der getesteten Hypothesen 2,3-mal häufiger sind als in den Naturwissenschaften (Fanelli). Zudem sind die Geisteswissenschaften anfälliger für einen raschen Wandel von *Buzzwords* (Bentley), und der wissenschaftliche Konsens ist in den Geistes- und Sozialwissenschaften niedriger als in den Naturwissenschaften (Fanelli und Glänzel).

Mögliche Gründe hierfür könnten (1) das fehlende Bewusstsein der Forschenden für Bias, (2) geringe Teststärke aufgrund von kleinen Stichprobengrößen in (psycho-)linguistischen Studien und (3) in einzelnen Fällen das Fehlen einer guten wissenschaftlichen Praxis sein, entweder aufgrund von Unkenntnis oder fragwürdiger Forschungspraktiken, wie z.B. die selektive Berichterstattung und unzureichende Spezifizierung der Rahmenbedingungen, was eine Replikation unmöglich macht.

Eine niedrige Replikationsquote könnte als der linguistischen Forschung inhärent angesehen werden, da Sprache als soziales Phänomen eine große Dynamik aufweist, oder wie es Grieve zusammenfasst,

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. (Grieve 1344)

Jedoch würde dies die methodologische Wende der Linguistik zum quantifizierten Experimentieren im Allgemeinen und insbesondere zur Hypothesentestung in Frage stellen und den großen Erkenntnisgewinn der letzten Jahrzehnte invalidieren. Dies führt zu der Frage, ob es nicht besser wäre zur qualitativen Forschung zurückzukehren, wenn es unmöglich ist, im Rahmen des experimentellen Arbeitens in der Linguistik die Hauptgütekriterien von Objektivität, Reliabilität und Validität sicherzustellen.

Experimentieren in der Linguistik

In der Linguistik besteht das Ziel darin, Sprache zu verstehen: Wie sie verwendet wird, wie sie erworben wird, wie sich ihre Nutzung durch verschiedene Gesellschaftsgruppen unterscheidet oder wie sie im Gehirn repräsentiert wird. Dies kann durch rein theoretische Überlegungen erfolgen, was jedoch Grenzen hat und sich nicht verifizieren lässt. Das empirische Arbeiten stellt einen systematischen Ansatz zur Erkenntnisgewinnung dar, der darauf abzielt, Wissen zu generieren, das durch Beobachtung oder Erfahrung verifiziert werden kann, anstatt nur durch Theorie oder Logik. Beobachtungsstudien können durch Korrelationen untermauert werden, die jedoch nicht dazu geeignet sind, Kausalzusammenhänge zu erklären. Experimente hingegen, können direkte Tests für kausale Zusammenhänge ermöglichen. Die Forschung in der empirischen Linguistik sucht oft nach Beziehungen zwischen abhängigen linguistischen oder behavioristischen Variablen und unabhängigen Variablen, die sprachlich oder außersprachlich sein können, wie z.B. Erstsprache, Aufgabe, Sprecher*in, Alter usw.

Die Nutzung von Experimenten ist nicht zwingend notwendig, viele sprachliche Theorien und Modelle wurden auf Grundlage von Beobachtungen entwickelt und bewertet. Die Beobachtung birgt jedoch die Herausforderung, dass die Anwesenheit der beobachtenden Forschenden das Verhalten der beobachteten Forschungssubjekte verändern und damit die Ergebnisse beeinflussen kann. In der Soziolinguistik wird dieses Phänomen sogar als Beobachterparadoxon bezeichnet, und es wird allgemein akzeptiert, dass dieses Phänomen innerhalb des Kontextes soziolinguistischer Interviews und Umfragen nie vollständig überwunden werden kann (Labov).

Auch in experimentellen Studien spielt das Beobachterparadoxon eine Rolle, jedoch ist der Vorteil, dass man den Kontext der Sprachproduktion kontrollieren kann.

In einem psycholinguistischen Experiment wird eine (oder werden mehrere) Bedingung(en) systematisch variiert, wobei diese unabhängige Variable die zu untersuchende Variable ist, und die (kausale) Wirkung der Variation auf die gemessenen Ereignisse (abhängige Variable) beobachtet wird. Die unabhängige Variable ist häufig ein Merkmal, das Gruppen unterscheidet, wie zum Beispiel das Beherrschen einer bestimmten Sprache. Hierbei werden Proband*innen der verschiedenen Gruppen einem spezifischen Stimulus ausgesetzt, der es erlaubt, den kausalen Effekt der Intervention auf die Sprachproduktion oder -wahrnehmung zu messen und zu vergleichen. Die Manipulation ist nicht perfekt, da man die Proband*innen aufgrund ihrer Charakteristika den Gruppen zuordnet, anstatt eine zufällige Zuordnung zu treffen. Dadurch kann nicht sichergestellt werden, dass Störfaktoren kontrolliert werden, da es unbekannte Merkmale geben kann, die sich zwischen den zu vergleichenden Gruppen unterscheiden und die der eigentliche Grund für die beobachteten Unterschiede in der gemessenen abhängigen Variable sein könnten. Aus diesem Grund handelt es sich in der Linguistik häufig um Quasi-Experimente, da bestimmte Faktoren nicht manipulierbar sind. In einem echten Experiment wird die unabhängige Variable aktiv manipuliert, und die Gruppierung der Proband*innen erfolgt zufällig. In einem Quasi-Experiment dagegen werden die Proband*innen aufgrund natürlicher Gegebenheiten gruppiert (monolingual und bilingual; Sprecher*in einer oder einer anderen Sprache).

Es gibt jedoch Möglichkeiten, die Formen zumindest im Hinblick auf die äußeren Rahmenbedingungen, unter denen die Untersuchung durchgeführt wird, zu kontrollieren. (Psycho-)Linguistische (Quasi-)Experimente finden oft in einem Labor statt, in dem Störvariablen sorgfältig kontrolliert werden können. Die gesammelten Daten werden anschließend mit dem statistischen Rahmen des Nullhypothese-Statistiktests (NHST) analysiert. Die Verwendung der frequentistischen statistischen Inferenz in der Datenanalyse, beispielsweise durch Anwendung eines t-Tests, beinhaltet typischerweise eine binäre Entscheidung: Eine zuvor gestellte Nullhypothese wird unter Verwendung statistischer Signifikanztests entweder abgelehnt oder akzeptiert (Greenland). Das Signifikanzniveau ist im Grunde genommen eine willkürlich vereinbarte Schwelle, die in der Linguistik normalerweise bei $p = .05$ liegt. Linguist*innen nehmen demzufolge in Kauf, 5 % der Ergebnisse, die eigentlich nur auf Zufall beruhen, fälschlicherweise als signifikant anzusehen. Dies wird als Fehlertyp I bezeichnet. Ein signifikantes Ergebnis bedeutet, dass das gemessene Ergebnis als zu extrem angesehen wird, um zufällig entstanden zu sein. Nicht signifikante Ergebnisse sind hingegen schwer zu interpretieren, weil das Nichtfinden eines signifikanten Effekts nicht automatisch bedeutet, es liege keine Wirkung vor (Vasishth and Gelman).

Jedoch sind die identifizierten kausalen Zusammenhänge nur innerhalb des spezifischen Experimentkontextes garantiert gültig und nicht generalisierbar, da die Experimente unter kontrollierten Laborbedingungen durchgeführt werden und die Aufgaben in gewisser Weise künstlich sind, um kausale Zusammenhänge deutlich aufzeigen zu können. Es kann auf dieser Grundlage keine grundsätzliche Aussage über linguistische Phänomene in der komplexen realen Welt getroffen werden.

In jedem Schritt des Experiments muss der/die Forschende zahlreiche Entscheidungen treffen, die häufig aus methodologischen Gesichtspunkten relativ willkürlich sind, so dass man eine relativ große Entscheidungsfreiheit hat, die als Freiheitsgrade von Forschenden (*researcher degrees of freedom*) eingestuft werden (Simmons et al.; Wicherts et al.). Diese Entscheidungen beeinflussen das Ergebnis der durchgeführten Signifikanztests, also welche Ergebnisse als signifikant gesehen werden und welche nicht, und damit in entscheidender Weise auch die Schlussfolgerungen aus der Forschung. Wenn diese Freiheitsgrade absichtlich ausgenutzt werden, um die Wahrscheinlichkeit falsch positiver Ergebnisse (Fehlertyp 1) zu erhöhen, kann dies zur Nicht-Reproduzierbarkeit von Ergebnissen beitragen.

Letztendlich können anhand der Inferenzstatistik i.d.R. nur ungewisse Schlussfolgerungen aus den Daten gezogen werden, sogar bei signifikanten Ergebnissen. Die einzige Möglichkeit, im Rahmen eines experimentellen Forschungsdesigns mit absoluter Sicherheit eine absolute Antwort zu erhalten ist, jeden in der zu untersuchenden Population erschöpfend zu testen. Dies ist in den meisten Fällen in der Linguistik absolut impraktikabel. Jedoch kann man durch die korrekte Anwendung frequentistischer Statistik anhand NHST und p -Werten zuverlässig und recht einfach zu vernünftigen Schlussfolgerungen gelangen. Durch eine etwas anspruchsvollere Statistik kann man sogar die Unsicherheit der Schlussfolgerungen statistisch quantifizieren.

In letzter Zeit haben gemischte Modelle (*Mixed Models*) an Popularität gewonnen, was dazu führt, dass frequentistische Statistiken zunehmend als unzureichend angesehen werden. Die Probleme bei der Interpretation von frequentistischen Statistiken sind jedoch größtenteils auf die falsche Anwendung der Methode zurückzuführen, die von Unwissenheit und Naivität bis hin zur absichtlichen Manipulation durch p -Phishing oder HARKing reichen.

Im Folgenden wird der Fokus auf drei Ansätze gelegt, die sicherstellen können, dass die frequentistischen Statistiken tatsächlich richtig funktionieren: 1) Die Verpflichtung zu einer klar definierten falsifizierbaren Hypothese, 2) ausreichende Teststärke durch eine ausreichend große Stichprobe und 3) eine Kontrolle sowie angemessene und transparente Beschreibung des Kontexts der Datenerhebung und ein Verzicht auf die Verallgemeinerung nicht allgemein gültiger Forschungsergebnisse, da praktisch jedes linguistische Phänomen kontextsensibel ist. Wenn

diese Bedingungen nicht gegeben sind, besteht ein hohes Risiko, das Feld mit nicht replizierbaren Ergebnissen von zweifelhafter Glaubwürdigkeit zu überfluten.

Hier wird die Meinung vertreten, dass die experimentelle Methodik für die Linguistik zielführend ist, wenn diese richtig angewendet wird. Die modernen Standards im Experimentdesign und in der statistischen Auswertung, die sich aus der Replikationskrise herauskristallisiert haben, bieten sogar Möglichkeiten, mehr Diversität in den untersuchten Daten im experimentellen Forschungsprozess zu berücksichtigen als mit den alten frequentalistischen Standards. Gleichzeitig können die Repräsentativität und die Qualität der linguistischen Analysen gesteigert werden, was eine Chance für die Diversitätsforschung darstellt. Im Folgenden wird erörtert, wie diese doppelten Ziele erreicht werden könnten.

Der experimentelle Prozess besteht aus einer Reihe von aufeinander aufbauenden Schritten. Aspekte, die in den ersten Schritten nicht berücksichtigt werden, können auch später nicht untersucht werden. Entsprechend kann man zwar signifikante, jedoch nicht-repräsentative Ergebnisse erwarten, wenn eine nicht-repräsentative Stichprobe schlecht analysiert wird (auch bekannt als das *garbage in/garbage out*-Prinzip). Aufgrund der Bedeutung der ersten drei Schritte für die experimentelle Methodik, wird der Fokus hierauf gelegt: (1) Die Formulierung der Hypothese, (2) der Entwurf eines Experiments und (3) die Datensammlung.

Die natürliche Variabilität im Sprachgebrauch

Da Kommunikation stets Selbstdarstellung bedeutet, fließen Kommunikation und Identität ineinander über. Als identitätsstiftendes Mittel kann Sprache Menschen verbinden (Gruppenzugehörigkeit) oder trennen (Markierung einer Abgrenzung). Sprachgebrauch unterscheidet sich naturgemäß zwischen Sprachen, Dialekten, Registern und einzelne Sprechenden. Sprache ist zugleich Mittel der Kommunikation und der Identität, da sich Menschen in der Kommunikation ausdrücken, sich positionieren und sich damit als Akteure ihrer eigenen Identität verhalten. Identitätsrelevante Merkmale beeinflussen die Sprachverwendung und können deshalb nie gänzlich von Experimenten ausgeblendet werden: Faktoren wie das Bildungsniveau, das Alter oder die lokale Herkunft sowie die Zugehörigkeit zu einer kulturellen Gruppe, das Geschlecht des Sprechers, das Geschlecht der Gesprächspartner*in, die Geburtsreihenfolge, das Einkommen der Eltern und die Zuhörendenschaft (Johnstone 50) beeinflussen die Färbung des Sprachgebrauchs. Diese Aspekte entfalten sich besonders in der Gemeinschaft mit anderen (Joseph 17), weil es nur in dieser Situation darum geht, jemand zu sein und einen Teil der Identität zu offenbaren. Denn zu Hause oder in einem psycholinguistischen Labor ohne Kontakt zu anderen Individuen ist es nicht sozial entscheidend, sich darzustellen.

Diese Variabilität wird manchmal als zu komplex und variabel betrachtet, um von linguistischen Experimenten berücksichtigt zu werden. Beobachtende Studien können die Variation dieser Variablen untersuchen, sogar ohne jegliches Eingreifen der Forschenden. Im Gegenteil, der Haupteinwand gegen ein experimentelles Vorgehen in der Linguistik war ursprünglich, dass man die Komplexität der menschlichen Realität nicht durch eine Hypothesenprüfung, durch experimentelle Bedingungen und Messungen erfassen könne. Eine hohe ökologische Validität, also ein Experimentdesign, was mit den Merkmalen der natürlichen Umgebung des untersuchten Phänomens übereinstimmt, und die Stringenz der Methodenanwendung im naturwissenschaftlichen Sinne sind meistens entgegengesetzte Ziele, die in einem Experiment ausbalanciert werden müssen.

Sprache weist eine große Variabilität auf. Jedoch erfordert die Hypothesenprüfung im Sinne von Fisher ein niedriges Signal-Rausch-Verhältnis, um signifikante Ergebnisse zu liefern. Es gibt zwei Wege, trotz der großen Variabilität die Hypothesenprüfung sinnvoll anzuwenden: (1) Der Effekt der Variabilität wird entweder durch eine möglichst homogene Stichprobe reduziert (z.B. nur männliche Rechtshänder zwischen 20 und 30 Jahren mit Hochschulabschluss in beiden Vergleichsgruppen) oder (2) durch ein Ausbalancieren der Variablen kontrolliert (z.B. 50 % Frauen, 50 % Männer in beiden Gruppen). Experimentelle Designs erproben bewusst die Variation, die in einer natürlichen Umgebung vorhanden wäre. Diversität wird also ausgeschlossen, um Rauschen in den Daten zu minimieren. Dadurch sind die Ergebnisse jedoch weniger repräsentativ und verallgemeinerbar. Dies ist insbesondere vor dem Hintergrund zu hinterfragen, dass die Replikationskrise aufzeigt, dass Methoden häufig nicht sauber angewendet werden, was dazu führt, dass die Ergebnisse ohnehin nicht generalisierbar sind.

Praktische und ethische Überlegungen setzen zudem Grenzen für experimentelle Untersuchungen an menschlichen Subjekten, was eine saubere Anwendung der Methoden weiter erschwert. Die Fähigkeit einer Person, die Methodik korrekt anzuwenden, ist also oft durch praktische Grenzen sowie durch Bias eingeschränkt. Ein Ausweg aus diesem Dilemma ist die Anwendung von anspruchsvolleren statistischen Modellen wie gemischten Modellen (*Mixed Models*), die auf strukturierte Weise mit viel mehr Variation umgehen können. Damit kann mehr Vielfalt in den Studien berücksichtigt werden.

Das Untersuchungsobjekt der Linguistik ist die Sprache, nicht die Sprechenden der Sprache, sondern die von diesen Menschen produzierte Sprache selbst. Eine genaue Definition, was alles Sprache umfasst, ist schwierig (so wie auch in der Biologie eine genaue Definition von Leben schwierig ist). Diese zu untersuchende *Sprache* ist jedoch eine Idealisierung, sogar für die Linguisten*innen, die deskriptiv und nicht normativ arbeiten. Linguistische Studien untersuchen meistens nur die Standardsprache, in der Germanistik also meistens das Hochdeutsche. Dialekte oder Soziolekte (Jugendsprache, Ausdrucksweisen verschiedener Subkulturen oder

sozioökonomischer Milieus) werden manchmal in speziellen Untersuchungen betrachtet. Wenn diese sprachliche Diversität nicht das primäre Forschungsthema der Studie ist, wird versucht, diese Quellen von Variabilität aus den Daten rauszuhalten, indem z.B. Dialektsprechende gar nicht rekrutiert werden. Dies spiegelt die außerhalb der Wissenschaft verbreiteten Vorurteile gegenüber diesen dialektalen Formen wider und verstärkt diese eventuell sogar. Zum Beispiel wird das doppelte Perfekt (»Ich habe es gehört gehabt.«) pejorativ als »Hausfrauen-Perfekt« bezeichnet (Sick 180).

Die Hypothesenstellung

Einstellungen und Werte der Forschenden haben einen Einfluss auf die Formulierung von Forschungsfragen und entsprechenden Hypothesen sowie bei der Sammlung und Analyse von Daten, die zu Verzerrungen (*bias*) führen können. Welche Forschungsfragen überhaupt gestellt werden, hängt auch von politischen Faktoren wie z.B. der vorherrschenden Förderpolitik und der Hochschulpolitik, aber auch hochschulinternen Politiken ab, die die Machtverhältnisse in der Gesellschaft widerspiegeln. Zugunsten der »Verallgemeinerung von Ergebnissen« wird oft in linguistischen Studien nicht die tatsächliche komplexe Realität betrachtet, sondern ein vereinfachter Auszug der Realität. Die Ergebnisse einer Studie können keine Aussagen über Gruppen von Menschen treffen, die vorher systematisch bei der Erhebung der Daten ausgeschlossen wurden. Dies führt dazu, dass bestimmte Minderheiten für die Forschung »unsichtbar« werden. In manchen Fällen werden sogar Forschungsergebnisse undifferenziert verallgemeinert, so dass unklar ist ob die Ergebnisse auf bestimmte Gruppen (z.B. Frauen) zutreffen (Criado Pérez).

Mit der Hilfe von gemischten Modellen kann man mit erheblich diverseren Datensätzen umgehen, sodass man schon bei der Experimentplanung mehr Diversität einbeziehen kann. Bei der Formulierung von Hypothesen bietet sich also die erste Gelegenheit, Diversität in die Forschung einzubeziehen, indem nicht nur Hypothesen über die Sprechergruppen aufgestellt werden, die der linguistischen Norm entsprechen, sondern auch über Minderheitsgruppen. Es gibt Studien zur Produktion und Rezeption von Sprache durch unterschiedliche soziale Gruppen, inklusive Nutzer von Dialekten (Montrul et al.; Walker), Bilingualen (Werkmann Horvat et al.) und Zweitsprachnutzenden, sowie über Nutzende von Jugendsprache und wenig bekannten und selten gesprochenen Sprachen (Fast).

Um den NHST-Ansatz sauber durchzuführen, muss die Hypothese klar formuliert werden und sich aus der Theorie ableiten. In der Hypothesenformulierung sollten die unabhängige Variable und die abhängige Variable vor der Durchführung der Studie spezifiziert werden, ebenso wie die erwartete Richtung des Effekts, idealerweise mit einem Vorregistrierungsprotokoll (Wagenmakers et al.). Wenn diese ge-

naue Operationalisierung fehlt, besteht später zu viel Flexibilität bei der Analyse und Interpretation der Ergebnisse (Schaller).

Explorative Studien ohne theoretische Begründung für erwartete Effekte, die darauf abzielen, »interessante« (d.h. in der Regel statistisch signifikante) Muster in den Daten zu finden, liefern praktisch mathematisch garantiert irgendwelche statistisch signifikanten Ergebnisse (Wagenmakers et al.), sofern keine angemessenen statistischen Korrekturen für Mehrfachtests vorgenommen werden. Eine Hypothese erst dann aufzustellen, wenn man weiß, dass die Ergebnisse auf diese Weise signifikant werden, widerspricht dem grundlegenden Prinzip des Hypothesentests, und ist als HARKing (*Hypothesizing After Results are Known*) bekannt (Kerr). Statistisch signifikante Ergebnisse sind jedoch nur in *Confirmatory Studies* interpretierbar, d.h. wenn eine spezifische und klar formulierte Hypothese abgelehnt oder angenommen werden kann.

Experimentelles Design

Experimentelle Designs unterschätzen oft die Variation, die in natürlicher Sprache und in menschlichen Gesellschaften vorhanden ist. Innerhalb des positivistischen Paradigmas sollen alle möglichen Subjektivitäten, einschließlich Identitätsmerkmale, vermieden werden, aber es scheint offensichtlich, dass der Sprachgebrauch einer Person oder Gruppe von mehr als einer relevanten Variable abhängig sein wird. Ein klares Beispiel, wo diese Probleme aufeinandertreffen, ist die Forschung zu Mehrsprachigkeit in Regionen, in denen die Zugehörigkeit zu einer oder mehreren Sprachen eine politische Dimension hat, wie in Katalonien oder im Süden der Vereinigten Staaten. In solchen Fällen können Forschende selbst persönliche Vorurteile haben, die ihre Suche nach bestimmten Ergebnissen beeinflussen. Es ist wichtig, diese persönlichen Vorurteile offenzulegen, wie es in der qualitativen Forschung Usus ist. Selbst wenn der/die Forschende eigentlich nur am Aspekt der Mehrsprachigkeit interessiert ist und nicht persönlich von der politischen Seite des Themas betroffen ist, kann es schwierig sein, bei der Rekrutierung von Proband*innen sicherzustellen, dass sie die Identitätskriterien und sprachlichen Kriterien sauber voneinander trennen können.

Es stellt sich die Frage, ob manche etablierten experimentellen Designs tatsächlich alle relevanten Quellen für die Variation der Daten erfassen. Aufgrund der Komplexität von Sprache und Gesellschaft ist es schwierig, relevante Variablen zu identifizieren und zu messen. Als Veranschaulichung dient anekdotisch ein berichteter Zwist zwischen Fitzgerald und Hemingway in Paris im Jahr 1920. Die beiden finanziell erfolgreichen Schriftsteller hatten so unterschiedliche Weltanschauungen und literarische Stile:

Fitzgerald: »Die Reichen sind anders als wir.«

Hemingway: »Ja, sie haben mehr Geld.«

Fitzgeralds Äußerung deutet auf einen fundamentalen Unterschied zwischen den Reichen und anderen hin, möglicherweise in Bezug auf ihre Erfahrungen, Prioritäten oder Werte. Hemingways witzige Antwort verengt den Fokus lediglich auf den materiellen Aspekt des Reichtums und verharmlost andere Unterschiede. Dieser Austausch spiegelt die gegensätzlichen Perspektiven und Persönlichkeiten von Fitzgerald (bekannt für seine Faszination für die Elite) und Hemingway (ein Mann, der sich als unprätentiös betrachtet) wider. Insgesamt zeigt aber auch dieser kurze Austausch, wie quantitative Maßnahmen dazu neigen können, komplexe soziale Phänomene auf eine einzige »essentielle« Eigenschaft oder ein einziges Merkmal zu reduzieren.

Dieses Problem der Essentialisierung kann zu einer Verzerrung der Realität führen, da es die Vielfalt und Komplexität sozialer Phänomene nicht angemessen berücksichtigt und wiederum zu Stereotypen, Vorurteilen und unangemessenen Verallgemeinerungen führen kann. Die Überbetonung eines einzelnen Merkmals (z.B. Geschlecht, Muttersprache, sozioökonomischer Status) und die Reduzierung einer Person oder Gruppe auf dieses eine Merkmal blendet andere Identitätsmerkmale aus. Essentialisierung beinhaltet außerdem die Annahme, dass Menschen aufgrund einzelner Merkmale objektiv und eindeutig bestimmten Gruppen zugeordnet werden können. Wenn dies von den Betroffenen als Fremdzuschreibung empfunden wird, geht die Essentialisierung mit Ausschlusspraktiken einher. Die Lösung im Rahmen quantitativer Methoden könnte sein, nicht nur anhand einzelner Variablen zu operieren, sondern ein Konstrukt auf Basis mehrerer Variablen herzustellen, die dank einer Faktorenanalyse unterschiedliche Gewichtungen erhalten.

Häufig werden in der Linguistik binäre Kategorien gebildet und miteinander verglichen, um die statistische Datenanalyse zu erleichtern. Jedoch passt kaum eine menschliche Kategorie sauber in eine binäre Unterteilung. Als Beispiel kann die Kategorie Geschlecht hier genannt werden. Das abstrakte Konzept von Gruppen, die sich gegenseitig ausschließen, scheint in der Theorie überholt zu sein. Die meisten Kategorisierungen erfolgen in einem Spektrum. Auch die monolingual/bilingual-Unterteilung hat in den letzten Jahren eine spektrale Behandlung erfahren, als ein Kontinuum mit Monolingualen an einem Extrem und Mehrsprachigen am anderen. Das gleiche passiert mit der Dichotomie der »alltäglichen Erzählung« und der literarischen Erzählung, die jetzt auch als zwei Pole in einem Kontinuum gesehen werden.

Problematisch im Sinne der Freiheitsgrade von Forschenden ist, wenn eine Variable in ihrer Funktion in der Studie oder in der gemessenen Skala nachjustiert wird. Zum Beispiel könnte die Variable *Alter*, wenn gemessen in Jahren, nachträglich bei der Datenanalyse als kontinuierliche Variable beibehalten werden, oder in

eine dreistufige Skala (jung, mittleres Alter und alt) oder in eine zweistufige Skala (alt und jung) umgewandelt werden. Dies kann entweder durch Neudefinition der Kategoriengrenzen oder durch das Ausschließen einer Teilnehmendengruppe erfolgen. Darüber hinaus kann die gemessene Altersvariable in späteren Analysen verschiedene Rollen übernehmen: als unabhängige Variable zur Untersuchung von Altersunterschieden oder als Kontrollvariable zur Erklärung eines Teils der innerhalb der Bedingungen auftretenden Variation in der abhängigen Variable. Diese Flexibilität betrifft alle demografischen Variablen (z.B. Einkommen, sozioökonomischer Status, Bildungsniveau, sogar die Muttersprache). Idealerweise sollten bei der Experimentplanung die Operationalisierung aller Variablen spezifiziert werden.

Ein weiteres beständiges Problem im Bereich des Experimentdesigns ist die Abhängigkeit von Datenstrukturen: Oft wird bei statistischen Auswertungen in der Linguistik angenommen, dass alle Datenpunkte aus statistischer Sicht unabhängig voneinander sind, obwohl ihnen diese Eigenschaft ganz offensichtlich fehlt, zum Beispiel weil mehrere Messungen auf eine einzige Sprecherin zurückzuführen sind. In einem Experiment würde diese Annahme erfüllt sein, wenn jeder der Datenpunkte von einem anderen Teilnehmenden stammen würde. Falls jede*r Teilnehmende mehr als einen Datenpunkt für jede Bedingung liefert, sollte ein Durchschnitt dieser mehreren Punkte genommen werden, damit das, was in den statistischen Test einfließt, ein Datenpunkt pro Teilnehmer*in pro Bedingung ist. Das Nichtberücksichtigen von Informationen über einzelne Teilnehmende im Modell führt zu einer künstlichen Erhöhung der Stichprobengröße (basierend auf *trials* und nicht auf Proband*innen), was fälschlicherweise zu einer Überschätzung der Teststärke führt (Hurlbert).

Die unkritische Verwendung von Statistiktest, die auf der Unabhängigkeitsannahme beruhen, kann zu falschen Ergebnissen und Interpretationen führen (Winter und Grice) und ist eine fragwürdige Forschungspraxis, über die inzwischen ein größeres Bewusstsein herrscht. Als Lösungsansatz für diese nicht unabhängigen Datenstrukturen in Studiendesign ist der Gebrauch von gemischten Modellen in der Linguistik vorgeschlagen worden, sodass die wichtigen Quellen von Abhängigkeit in den Daten als Zufallseffekte in das statistische Modell einbezogen werden, z.B. Datenpunkte, die von denselben Teilnehmenden stammen. Auch andere Quellen der Variation können im Modell berücksichtigt werden: Proband*innen, die sich näher sind (weil sie sich kennen, denselben Dialekt sprechen, dieselbe berufliche Ausbildung haben, auf die gleiche Weise sozialisiert wurden usw.), werden eher linguistische Merkmale teilen, was eine Form der Nicht-Unabhängigkeit einführt, wenn eine Studie mehrere Sprecher*innen derselben Gruppe einschließt. Durch Verwendung von gemischten Modellen kann ein Zufallseffekt für den Rekrutierungsort, den Dialekt, die Ausbildung oder das Geschlecht dem Modell hinzugefügt werden. Somit kann mithilfe von gemischten Modellen nicht nur das Problem der Verletzung der Unabhängigkeitsannahme durch nicht unabhängige Datenstrukturen angegangen

werden, sondern es können auch vielfältigere Gruppen von Proband*innen für die Studie zugelassen werden, ohne das Risiko zu erhöhen, dass dadurch das Verhältnis von Daten zu Rauschen zu hoch wird.

Ein Beispiel, wo konfundierende Variablen die Interpretation der Ergebnisse dramatisch verändern können, ist der viel berichtete sprachliche Relativitätseffekt, wonach angeblich Sprecher*innen von Sprachen ohne grammatisierte Formen für Präsens und Zukunft mehr Geld sparen als Sprecher*innen von Sprachen, mit Präsens/Zukunft Unterscheidung (Chen). Bei Berücksichtigung genealogischer Abhängigkeiten und Einbeziehung von Zufallseffekten für Sprachfamilie und geographischer Ort zeigte sich jedoch, dass der Effekt zwischen Grammatik und Sparverhalten viel schwächer und schwerer zu interpretieren war als ursprünglich angenommen (Roberts et al.). Ein weiteres Beispiel im Bereich der sprachlichen Relativität stellt das berühmte Turning-Tables-Experiment dar (Brown und Levinson), welches einen Zusammenhang zwischen dem Referenzrahmen einer Sprache (basierend auf Himmelsrichtungen oder basierend auf Links/Rechts-Unterscheidungen) und dem Verhalten bei Rotationsproblemen zeigte. Die Ergebnisse wurden kritisiert, weil außer der manipulierten Variable der ersten Sprache (Niederländisch vs. Tzeltal) weitere Variablen wie Schulbildung, Urbanisierung, und gesellschaftliche Insularität für die gemessenen Unterschiede verantwortlich sein könnten (Li und Gleitman). Diese Kritik wurde aufgegriffen (Levinson et al.), indem gezeigt wurde, dass tatsächlich das Alter, das Geschlecht und der berufliche Hintergrund der Teilnehmenden kontrolliert wurden, sowie die Umgebung, in der die Datenerhebung stattfand (draußen/drinnen). Darüber hinaus wurde die Effekte von Lese- und Schreibfähigkeit (*literacy*), Alter und Geschlecht statistisch überprüft (Levinson).

In dieser Replik argumentieren Levinson und Kollegen, dass die Testumgebung und die Protokolle zur Datenerhebung sorgfältig kontrolliert werden müssen (Gedächtnisbelastung, Verbergen der eigentlichen Aufgabe usw.), um sicherzustellen, dass die Aufgabe für alle Teilnehmenden ausreichend undurchsichtig ist. Wichtig ist auch, dass die Ergebnisse über verschiedene Sprachen hinweg (Longgu, Tzeltal und Arrernte) stabil sind.

Stichproben

Die Teststärke ist die Wahrscheinlichkeit, ein signifikantes Ergebnis zu erhalten, wenn die Nullhypothese tatsächlich falsch ist. Wenn die Teststärke niedrig ist, führt das NHST-Verfahren nicht zu aussagekräftigen Ergebnissen, sondern entweder zu schwer interpretierbaren nicht-signifikanten Ergebnissen oder zu signifikanten Effekten, deren wahre Effektstärke aber überschätzt ist (Gelman und Carlin). Es ist tatsächlich bei Replikationsstudien mit größeren Stichproben zu beobachten, dass entweder die signifikanten Ergebnisse der Originalstudie nicht repliziert werden

oder signifikante Effekte mit einer schwächeren Effektgröße gefunden werden (Jäger et al.; Nicenboim et al.; Vasisht et al.).

Studien in der Psycholinguistik haben oft zu geringe Teststärke da Forscherintuitionen meist überoptimistisch sind (Bakker et al.). Psycholinguistische Experimente und Pseudoexperimente arbeiten in der Regel mit 20 bis 40 Proband*innen, die als repräsentativ für eine Population von bis zu mehreren Millionen Menschen angesehen werden. Der Stichprobenumfang wird oft als Vielfaches der zu vergleichenden Bedingungen festgelegt (Friedrichs), was jedoch keine Repräsentativität, sondern nur die Vergleichbarkeit der Stichproben garantiert. Studien mit geringer Teststärke sind anfälliger für Bias aufgrund größerer Stichprobenvariabilität und proportional größerer Auswirkungen von Analyseentscheidungen bei kleineren Stichprobengrößen (Bakker et al.).

Es ist empfehlenswert, in linguistischen NHST-Studien eine formale Power-Analyse durchzuführen, um die Teilnehmendenzahl im Voraus zu bestimmen. Neben der Größe der Stichprobe sollte im Stichprobenplan auch die Population spezifiziert werden, aus der gezogen wird, sowie das Prozedere der Stichprobenziehung. Es ist obligatorisch, erst nach Abschluss der Datenerhebung die Datenanalyse durchzuführen. Das Beenden der Datensammlung nach Berechnung eines signifikanten Zwischenergebnisses oder das Sammeln weiterer Daten bei einem nicht signifikanten Ergebnis führt zu falsch positiven Ergebnissen (John et al.).

Eine niedrige Teststärke aufgrund einer kleinen Stichprobe zusammen mit einer nicht zufälligen Auswahl der Probanden (Bortz und Döring), wie oft in der Psycholinguistik der Fall ist, führt zu irreführenden, nicht replizierbaren Ergebnissen (Gelman und Carlin; Jäger et al.; Nicenboim et al.; Vasisht et al.). Die Idee, dass niedrige Proband*innenzahlen durch mehr Trials pro Proband*in kompensiert werden können, ist falsch, da dies zu nicht unabhängigen Datenpunkten führt. Die wenigen Fälle, in denen es zu rechtfertigen ist, mit kleineren Stichproben zu arbeiten, sind Fragestellungen, bei denen es schwierig ist, Proband*innen zu rekrutieren (Aphasieforschung, Ethnolinguistik). Jedoch sind große Stichproben und direkte Replikationen mit verschiedenen Stichproben unerlässlich in experimentell orientierten Forschungslinien, wenn das Ziel darin besteht, robuste und verallgemeinerbare Ergebnisse zu erzielen (Vasisht und Gelman 1317).

Datensammlung

Die Auswahl von Proband*innengruppen in linguistischen Studien ist eine komplexe Aufgabe. Die Kriterien, nach denen Teilnehmende bestimmten Gruppen zugeordnet werden, variieren zwischen Forschungsgruppen und manchmal sogar zwischen Projekten einer Gruppe. Dies kann auf theoretische Überlegungen zurückzuführen sein, zum Beispiel gilt das Erwerbsalter als Hauptkriterium zur Definition

bilingualer Kinder, bei mehrsprachigen Erwachsenen jedoch nicht unbedingt. Auch praktische Erwägungen spielen eine Rolle, da es oft schwierig ist, ideale Teilnehmende zu finden, und stattdessen Personen rekrutiert werden müssen, die innerhalb eines vernünftigen Zeitrahmens verfügbar sind. Zum Beispiel gestaltet sich die Definition von Monolingualität in kontrastiven Studien oft unterschiedlich, je nachdem, in welchem Land die Studie durchgeführt wird. In vielen Ländern mit einem hohen Bildungsniveau und einer Schulpflicht ist es schwer, tatsächlich monolinguale Sprecher*innen zu finden, da die meisten Menschen Fremdsprachenkenntnisse haben. Bis zu welchem Punkt jemand als monolingual bzw. bilingual betrachtet wird, variiert von Studie zu Studie, sowie bis wann außerhalb ihres Heimatlandes getestete Sprecher*innen noch als monolingual gelten. Es ist auch wichtig, im Voraus zu entscheiden, unter welchen Bedingungen bestimmte Teilnehmende von den Analysen ausgeschlossen werden sollen, um die Manipulation der Daten zugunsten bestimmter Ergebnisse zu vermeiden.

Um valide statistische Untersuchungen durchführen zu können, ist ein größerer Datensatz erforderlich. Dennoch ist eine perfekt ausgewogene Verteilung der Proband*innen (z.B. nach Geschlecht, Händigkeit oder Altersgruppen) schwer zu erreichen. Im Gegensatz zu Laborexperimenten in der Chemie oder Physik, bei denen Variationen einer einzigen Variable unter Beibehaltung aller anderen Variablen möglich ist, ist es in Experimenten mit menschlichen Subjekten unmöglich, alle anderen Variablen konstant zu halten. Dies hat auch Auswirkungen auf die statistische Analyse, insbesondere bei der Durchführung von Regressionsanalysen.

Wenn eine randomisierte Zuweisung der Proband*innen zu Gruppen nicht möglich ist, kann die Verteilung der Kovariaten (ein Variablentyp, der die Ergebnisse beeinflusst, aber nicht von direktem Interesse für die Studie ist) in beiden Gruppen durch eine paarweise Zuordnung (engl. *Matching*, ein Verfahren, mit denen ähnliche Beobachtungen in mindestens zwei Datensätzen verbunden werden) besser kontrolliert werden, indem Proband*innen beider Gruppen anhand bestimmter Kovariaten, wie Geschlecht und Alter, angepasst werden (Stadler et al.). Dennoch können eine Vielzahl anderer Kovariaten, die ebenfalls einen Einfluss auf die abhängige Variable haben könnten, unkontrolliert variieren. Trotz der Bemühungen, die Realität genauer abzubilden, z.B. indem man binäre Variablen durch in einem Spektrum variierende Größen ersetzt, bleibt eine erhebliche Variation innerhalb und zwischen den Proband*innengruppen bestehen. Dies stellt z.B. in der Bilingualismus-Forschung eine Herausforderung dar.

Nach Festlegung des Profils der zu untersuchenden Teilnehmenden steht die Herausforderung der Rekrutierung im Fokus. In linguistischen und sozialwissenschaftlichen Studien stammen die Teilnehmenden hauptsächlich aus dem spezifischen demografischen Kontext der sogenannten »westlich, gebildet, industrialisiert, reich und demokratisch« (W.E.I.R.D.) Personen. Für linguistische Studien werden typischerweise Universitätsstudierende rekrutiert, da diese an

den Universitäten, wo die Studien durchgeführt werden, reichlich verfügbar sind. Studierendengruppen unterscheiden sich in ihren kulturellen, Bildungs-, wirtschaftlichen oder politischen Hintergründen von anderen Gruppen, da sie nur einen begrenzten Ausschnitt der Bevölkerung repräsentieren. Trotzdem werden häufig die so gewonnenen Ergebnisse auf die Gesamtbevölkerung übertragen.

Die Rekrutierung von Proband*innen erfolgt häufig durch einen Aufruf zur Teilnahme, sei es auf dem Campus, im Internet oder über soziale Medien. Diese Methode führt jedoch zu einer Selbstselektion der Teilnehmenden, da sich in der Regel nur Personen melden, die bereits Interesse am Thema haben oder bestimmte Eigenschaften aufweisen. Diese Selbstselektion führt wieder zu Verzerrungen, auch weil sich manche Personen nicht mit bestimmten Gruppen identifizieren, obwohl sie die definierten Eigenschaften aufweisen (z.B. mehrsprachige Menschen, die sich selbst nicht als bilingual ansehen). Eine monetäre Entschädigung ist eine nur begrenzte Lösung, da sie zwar mehr Menschen zur Teilnahme motivieren kann, jedoch dazu führen kann, dass einige Teilnehmende nur am Geld interessiert sind und möglicherweise falsche Angaben machen, um zur Studie zugelassen zu werden. Finanziell übermotivierte Studienteilnehmende, die an so vielen bezahlten Studien wie möglich teilnehmen, verzerren auch die Daten in derselben Art und Weise wie *trap-happy animals* in Bevölkerungsschätzungsstudien in der Biologie. Einige Universitäten haben eine verpflichtende Studienteilnahme gegen *credits* eingeführt, um eine Proband*innenversorgung sicherzustellen; auf diese Weise werden jedoch meistens W.E.I.R.D.-Teilnehmende rekrutiert. Es wird empfohlen, bei der Berichterstattung nicht nur auf die Gruppeneigenschaften einzugehen, sondern auch transparent darzustellen, auf welche Weise die Probanden rekrutiert und zur Teilnahme motiviert wurden.

Die Datensammlung stellt ebenfalls eine Herausforderung dar, insbesondere aufgrund der kleinen Forschungsgruppengröße in der Linguistik, die eine Datensammlung nach einem Doppelblind-Protokoll schwierig macht. Idealerweise sollten nicht nur die Teilnehmenden, sondern auch die Experimentleiter*innen über kritische Designentscheidungen im Unklaren sein, um Verzerrungen durch Erwartungen oder unbeabsichtigte Signale zu vermeiden. Dies umfasst Informationen wie das eigentliche Ziel der Studie oder die Zuordnung der Proband*innen zu einer bestimmten Studiengruppe. Diese letzte Hürde ist besonders schwierig zu erreichen, da die Zuweisung oft offensichtlich ist (z.B. wenn Sprecher*innen unterschiedlicher Sprachen getestet werden).

Schlussfolgerungen

Jede Sprachäußerung in der realen Welt ist mit dem sozialen Kontext verbunden, in dem sie produziert wurde. Die Essenz der linguistischen Experimentalforschung

besteht jedoch darin, diesen sozialen Kontext zu manipulieren, was zwangsläufig Auswirkungen auf die Sprachproduktion und -wahrnehmung hat. Linguistische (Quasi-)Experimente stellen eine Möglichkeit dar, das komplexe soziale Phänomen Sprache zu verstehen, indem man den *trade-off* zwischen Natürlichkeit und Kontrolle in gewisser Weise manipuliert. Dieser Ansatz ist jedoch nur sinnvoll, wenn die Methodik richtig angewendet wird.

Angesichts der sogenannten Replikationskrise lässt sich argumentieren, dass die Hinwendung zur quantitativen Forschung in der Linguistik zu Lasten qualitativer Methoden und der linguistischen Theoriebildung im Allgemeinen gegangen ist. Wünschenswert wären ein Gleichgewicht und eine gegenseitige Verstärkung zwischen quantitativen und qualitativen Ansätzen (*Mixed Methods*). Dieser Zustand einer »quantitativen Krise« (Yarkoni) kann jedoch durch hohe Standards bei der Planung der Studie und bei der Sammlung und Analyse der Daten verbessert werden: Entscheidungen sollten im Voraus getroffen werden, um ein klares Protokoll für die Datenerhebung festzulegen, da Variationen im kommunikativen Kontext die sprachliche Form wesentlich und systematisch beeinflussen (Biber und Conrad).

Frequentistische statistische Tests setzen bestimmte Einschränkungen im Design experimenteller Studien voraus. Die Diversität der Teilnehmenden sowie die untersuchten Forschungsgegenstände sind dadurch zugunsten des NHST-Ansatzes eingeschränkt worden. Mit Hilfe von komplexeren statistischen Methoden, welche die Quellen von Variabilität berücksichtigen, ist es möglich, Experimente zu entwerfen, die mehr von der natürlichen Heterogenität der Sprache in die Daten zurückbringen (Yarkoni 19).

Schritte in die richtige Richtung wären klare Hypothesen mit erwarteter Richtung der Ergebnisse, höhere Stichprobengrößen und die Anwendung von statistischen Methoden, welche die Diversität der Proband*innen im statistischen Modell berücksichtigen. Im Modell sollten darüber hinaus variierende Einflussgrößen experimenteller Natur wie zum Beispiel die einzelnen Teilnehmenden, Stimuli oder Kontexte berücksichtigt werden. So könnten Schlussfolgerungen gezogen werden, die über verschiedene Sprechergruppen und Kontexte hinweg besser zu verallgemeinern wären (Yarkoni).

Eine größere Stichprobengröße erlaubt somit mehr Diversität in den Proband*innen, und die gemischten Modelle können mit dieser Diversität umgehen, ohne die Unabhängigkeitsannahme zu verletzen. Dies würde zu einer ökologisch validen Forschung und einer besseren Verallgemeinerbarkeit der Ergebnisse führen, was eine erhebliche Verbesserung im Vergleich zum aktuellen methodologischen Standard darstellen würde. Die aktuelle Methodenpraxis ist oft von (unbewusstem) HARKing geprägt, was statistische Beweise für Muster liefert, die eigentlich schwächer ausgeprägt oder gar nicht vorhanden sind.

Bei der Interpretation der Ergebnisse sollten die Einschränkungen der Studie ehrlich dargelegt werden, und weniger unreflektiert auf Gruppen übertragen wer-

den, die nicht in der Stichprobe vertreten waren. Man sollte nicht davon ausgehen, dass Ergebnisse, die für eine Sprache, eine Epoche, einen Dialekt, ein Register oder eine Aufgabe erhalten wurden, allgemeingültig sind (Grieve 1352).

Es besteht eine große Chance, das weitreichende Potenzial neuer methodischer Ansätze und Analysemethoden zu nutzen, um ökologisch gültige, valide und verallgemeinerbare Ergebnisse zu erhalten: Hierzu zählen zum Beispiel Mixed-Methods-Ansätze, welche quantitative und qualitative Ansätze kombinieren, sowie komplexe statistische Analysemethoden wie gemischte Modelle (*Mixed Models*). Da Sprachnutzer*innen immer im Kontext der Gesellschaft handeln, ist Sprachgebrauch immer im Zusammenhang mit zwischenmenschlicher Interaktion zu verstehen. Linguistische Studien können daher nur Teilbereiche der Komplexitäten der Realität aufdecken. Um komplexe linguistische Fragestellungen besser zu verstehen ist es notwendig, die gesamte Palette der methodischen Ansätze zu nutzen, die uns zur Verfügung steht. Hierbei sollten sich das Forschungsdesign und die gewählten Methoden an der Natur der jeweiligen Fragestellung orientieren. Jedoch werden wir niemals absolut objektiv sein, und wir können uns niemals ganz sicher sein, dass die Antworten, die wir erhalten, die Realität adäquat abbilden. Da in den Geistes- und Sozialwissenschaften Menschen versuchen, Menschen, menschliches Werken und Handeln zu verstehen, spielt Vielfalt eine große Rolle. Die mit Vielfalt einhergehenden Objektivitätsprobleme betreffen die Geisteswissenschaften daher mehr als andere Wissenschaften und sind gleichzeitig zentrale Themen in der geistes- und sozialwissenschaftlichen Forschung.

Bibliografie

- Anderson, Philip W. »More is different. Broken symmetry and the nature of the hierarchical structure of science.« *Science* 177:4047, 1972, 393–396.
- Andor, Józef. »The master and his performance: An interview with Noam Chomsky.« *Intercultural Pragmatics* 1, 2014, 93–111, <https://doi.org/10.1515/IPRG.2004.009>.
- Baayen, R. Harald, Douglas J. Davidson und Douglas M. Bates. »Mixed-effects modeling with crossed random effects for subjects and items.« *Journal of Memory and Language* 59:4, 2008, 390–412, <https://doi.org/10.1016/j.jml.2007.12.005>.
- Bakker, Marjan, Albert Van Dijk und Jelte M. Wicherts. »The rules of the game called psychological science.« *Perspectives on Psychological Science* 7:6, 2012, 543–554, <https://doi.org/10.1177/1745691612459060>.
- Bakker, Marjan, Chris Hartgerink et al. »Researcher's Intuitions about Power in Psychological Research.« *Psychological Science* 27:8, 2016, 1069–1077, <https://doi.org/10.1177/0956797616647519>.

- Bentley, Robert A. »Random Drift versus Selection in Academic Vocabulary: An Evolutionary Analysis of Published Keywords.« *PLOS ONE* 3:8, 2008, e3057, <https://doi.org/10.1371/journal.pone.0003057>.
- Biber, Douglas, und Susan Conrad. *Genre, Register and Style*. 2. Aufl., Cambridge University Press, 2019, <https://doi.org/10.1017/CBO9780511814358>.
- Binet, Alfred, und Victor Henri. »La mémoire des phrases (mémoire des idées).« *L'Année Psychologique* 1, 1894, 24–59.
- Bortz, Jürgen, und Nicola Döring. *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*. Springer, 2002, <https://doi.org/10.1007/978-3-540-33306-7>.
- Brown, Penelope, und Stephen C. Levinson. »Linguistic and nonlinguistic coding of spatial arrays: explorations in Mayan cognition.« *Working Paper No. 24, Cognitive Anthropology Research Group, Max Planck Institute for Psycholinguistics*, 1993.
- Chen, M. Keith. »The Effect of Language on Economic Behavior: Evidence from Savings Rates, Health Behaviors, and Retirement Assets.« *American Economic Review* 103:2, 2013, 690–731, <http://dx.doi.org/10.1257/aer.103.2.690>.
- Criado Pérez, Caroline. *Invisible Women. Data Bias in a World Designed for Men*. Abrams, 2019.
- Fanelli, Daniele. »Positive Results Increase Down the Hierarchy of the Sciences.« *PLOS ONE* 5:4, 2010, e10068, <https://doi.org/10.1371/journal.pone.0010068>.
- Fast, Karin. *Spatial Language in Tungag*. The Australian National University, 2015.
- Fisher, Ronald A. *Statistical Methods for Research Workers*. Oliver & Boyd, 1925.
- Friedrichs, Jürgen. *Methoden empirischer Sozialforschung*. Westdeutscher Verlag, 1990.
- Gelman, Andrew, und John Carlin. »Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors.« *Perspectives on Psychological Science* 9:6, 2014, 641–651, <https://doi.org/10.1177/1745691614551642>.
- Greenland, Sander. »Invited commentary: The need for cognitive science in methodology.« *American Journal of Epidemiology* 186:6, 2017, 639–645, <https://doi.org/10.1093/aje/kwx259>.
- Grieve, Jack. »Observation, experimentation, and replication in linguistics.« *Linguistics* 59:5, 2021, 1343–1356, <https://doi.org/10.1515/ling-2021-0094>.
- Hurlbert, Stuart H. »Pseudoreplication and the design of ecological field experiments.« *Ecological monographs* 54:2, 1984, 187–211, <https://doi.org/10.2307/1942661>.
- Jaeger, T. Florian. »Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models.« *Journal of Memory and Language* 59:4, 2008, 434–446, <https://doi.org/10.1016/j.jml.2007.11.007>.
- Jäger, Lena A., Daniela Mertzen et al. »Interference patterns in subject-verb agreement and reflexives revisited: A large-sample study.« *Journal of Memory and Language* 111, 2020, 104063, <https://doi.org/10.1016/j.jml.2019.104063>.

- John, Leslie K., George Loewenstein und Drazen Prelec. »Measuring the prevalence of questionable research practices with incentives for truth telling.« *Psychological Science* 23:5, 2012, 524–532, <https://doi.org/10.1177/0956797611430953>.
- Johnstone, Barbara. *The Linguistic Individual. Self-Expression in Language and Linguistics*. Oxford University Press, 1996, <https://doi.org/10.1093/os0/9780195101843.001.0001>.
- Joseph, John E. »Identity.« *Language and Identities*. Hg. von Carmen Llamas und Dominic Watt. Edinburgh University Press, 2010, 9–17.
- Kerr, Norbert L. »HARKing: Hypothesizing after the results are known.« *Personality and Social Psychology Review* 2:3, 1998, 196–217, https://doi.org/10.1207/s15327957pspr0203_4.
- Kobrock, Katja, und Thomas B. Roettger. »Assessing the replication landscape in experimental linguistics.« *Glossa Psycholinguistics* 2:1, 2023, 1–28, <https://doi.org/10.5070/G6011135>.
- Köhler, Reinhard. »Aims and methods of quantitative linguistics.« *Quantitative Linguistik: ein internationales Handbuch/Quantitative Linguistics: An International Handbook*. Hg. von Reinhard Köhler, Gabriel Altmann und Rajmund G. Piotrowski. de Gruyter, 2005, 1–16.
- Labov, William. *Sociolinguistic patterns*. University of Pennsylvania Press, 1972.
- Levinson, Stephen C., Sotaro Kita et al. »Returning the Tables: Language Affects Spatial Reasoning.« *Cognition* 84:2, 2002, 155–188, [https://doi.org/10.1016/S0010-0277\(02\)00045-8](https://doi.org/10.1016/S0010-0277(02)00045-8).
- Levinson, Stephen R. *Space in Language and Cognition: Explorations in Cognitive Diversity*. Cambridge University Press, 2003, <https://doi.org/10.1017/CBO9780511613609>.
- Li, Peggy, und Lila Gleitman. »Turning the Tables: Language and Spatial Reasoning.« *Cognition* 83, 2002, 265–294, [https://doi.org/10.1016/S0010-0277\(02\)00009-4](https://doi.org/10.1016/S0010-0277(02)00009-4).
- Montrul, Silvina, Rakesh Bhatt und Roxana Girju. »Differential object marking in Spanish, Hindi, and Romanian as heritage languages.« *Language* 91:3, 2015, 564–610.
- Munroe, Randall. »Purity.« *xkcd*, <https://xkcd.com/435/> (24.03.2026).
- Nicenboim, Bruno, Shravan Vasishth und Frank Rösler. »Are words pre-activated probabilistically during sentence comprehension? Evidence from new data and a Bayesian random-effects meta-analysis using publicly available data.« *Neuropsychologia* 142, 2020, 107427, <https://doi.org/10.1016/j.neuropsychologia.2020.107427>.
- Open Science Collaboration. »Estimating the reproducibility of psychological science.« *Science* 349:6251, 2015, <https://doi.org/10.1126/science.aac4716>.
- Roberts, Seán G., James Winters und M. Keith Chen. »Future Tense and Economic Decisions: Controlling for Cultural Evolution.« *PLOS ONE* 10:7, 2015, e0132145, <https://doi.org/10.1371/journal.pone.0132145>.

- Roettger, Timo. »Researcher degrees of freedom in phonetic research.« *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 10:1, 2019, 1–27, <https://doi.org/10.5334/labphon.147>.
- Schaller, Markus. »The empirical benefits of conceptual rigor: systematic articulation of conceptual hypotheses can reduce the risk of non-replicable results (and facilitate novel discoveries too).« *Journal of Experimental Social Psychology* 66, 2016, 107–115, <https://doi.org/10.1016/j.jesp.2015.09.006>.
- Shrout, Patrick E., und Joseph L. Rodgers. »Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis.« *Annual Review of Psychology* 69, 2018, 487–510, <https://doi.org/10.1146/annurev-psych-122216-011845>.
- Sick, Bastian. *Der Dativ ist dem Genitiv sein Tod. Neues aus dem Irrgarten der deutschen Sprache*. Folge 2. 15. Aufl., Kiepenheuer & Witsch, 2015.
- Simmons, Joseph P., Leif D. Nelson und Uri Simonsohn. »False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant.« *Psychological Science* 22:11, 2011, 1359–1366, <https://doi.org/10.1177/095679761141763>.
- Stadler, Mara, Philipp Doebler et al. »Statistical modeling of dynamic eye-tracking experiments: Relative importance of visual stimulus elements for gaze behavior in the multi-group case.« *Behavior Research Methods* 53:6, 2021, 2650–2667, <https://doi.org/10.3758/s13428-021-01576-8>.
- Vasishth, Shravan, Daniela Mertzen et al. »The statistical significance filter leads to overoptimistic expectations of replicability.« *Journal of Memory and Language* 103, 2018, 151–175, <https://doi.org/10.1016/j.jml.2018.07.004>.
- Vasishth, Shravan, und Andrew Gelman. »How to embrace variation and accept uncertainty in linguistic and psycholinguistic data analysis.« *Linguistics* 59:5, 2021, 1311–1342.
- Wagenmakers, Eric-Jan, Ruud Wetzels et al. »Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011).« *Journal of Personality and Social Psychology* 100, 2011, 426–432, <https://doi.org/10.1037/a0022790>.
- Wagenmakers, Eric-Jan, Ruud Wetzels et al. »An agenda for purely confirmatory research.« *Perspectives on Psychological Science* 7, 2012, 627–633, <https://doi.org/10.1177/1745691612463078>.
- Walker, Abby. »The role of dialect experience in topic-based shifts in speech production.« *Language Variation and Change* 31:2, 2019, 135–163, <https://doi.org/10.1017/S0954394519000152>.
- Werkmann Horvat, Anna, Marianna Bolognesi und Katrin Kohl. »Creativity is a Toaster: Experimental Evidence on How Multilinguals Process Novel Metaphors.« *Applied Linguistics* 42:5, 2021, 823–847, <https://doi.org/10.1093/applin/amab002>.

- Wicherts, Jelte M., Coos L. Veldkamp et al. »Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking.« *Frontiers in Psychology* 7, 2016, 1832, <https://doi.org/10.3389/fpsyg.2016.01832>.
- Winter, Bodo, und Martin Grice. »Independence and generalizability in linguistics.« *Linguistics* 59:5, 2021, 1251–1277, <https://doi.org/10.1515/ling-2019-0049>.
- Yarkoni, Tal. »The generalizability crisis.« *Behavioral and Brain Sciences* 45, 2020, 1–37, <https://doi.org/10.1017/S0140525X20001685>.