2 Vorplanung einer empirischen Untersuchung

Planung

Hat man sich dafür entschieden, ein empirisches Forschungsprojekt durchzuführen – und hier ist es unwesentlich, ob das Projekt ein sehr kleines ist (wie zum Beispiel eine Studie im Rahmen einer Seminararbeit, die einen Fragebogen einsetzt) oder ein größeres (wie für ein Dissertationsprojekt, das im Rahmen des Untersuchungsdesigns eventuell mehrere Erhebungsinstrumente verwendet) –, ist der wichtigste Schritt die Planung. Wenn man versucht, in der Planungsphase Zeit zu sparen, zum Beispiel weil man schnell Daten erheben möchte, verliert man erheblich mehr Zeit bei der Auswertung und Interpretation, weil die falschen Versuchspersonen ausgewählt wurden, weil die Daten nicht das Phänomen abbilden, das sie abbilden sollten, oder weil man nicht weiß, was man mit den gesammelten Daten anfangen soll. In solchen Fällen – und das passiert häufiger, als man denkt – muss man völlig neu beginnen.

Man kann viele Fallen vermeiden, indem man sich vorab grundlegende Gedanken macht. Bevor man ein passendes Untersuchungsdesign (Versuchspersonen, Erhebungsinstrumente, Vorgehensweise, Datenerhebung) auswählt, sollten der Untersuchungsgegenstand, die Fragestellung und entsprechende Hypothesen genau festgelegt sein, wobei man die Gütekriterien empirischer Forschung und mögliche Störfaktoren im Auge behalten muss. Auf jeden dieser Schritte gehen wir im Folgenden ein.

2.1 Auswahl eines Untersuchungsgegenstandes

Untersuchungsfragestellung

In der Sprachlehrforschung hat man den Vorteil, dass viele interessante Fragestellungen sich direkt aus der Praxis ergeben. Als Lehrperson wird man täglich mit Fragen und Problemen konfrontiert, die sich hervorragend für kleinere oder größere Projekte eignen: Fällt es Schülern in bestimmten sprachlichen Kontexten leichter, französische Possessivpronomen korrekt zu verwenden? Hilft es, wenn Schüler anstatt des regulären Sprachunterrichts jede Woche eine Stunde lang Grammatik mit Hilfe einer Computerlernsoftware üben? In welchen Situationen schafft es Renate, ein Adverbial ins Vorfeld zu setzen, ohne vor dem Verb noch das Subjekt einzufügen („Heute lerne ich Deutsch“ anstatt „Heute, ich lerne Deutsch“)? Und hat Peter, der aus Dänemark kommt, mit dieser Struktur mehr Probleme als Madeline aus den Niederlanden?

Aber auch wenn man nicht in der beruflichen Praxis steht, trifft man auf erforschenswerte Bereiche: durch das eigene persönliche Umfeld, durch Seminarthemen oder durch Literaturrecherche. Wichtig dabei ist, dass man weiß, was **genau** untersucht werden soll. Der Untersuchungsgegenstand muss exakt beschrieben werden. Gerade in der Sprachlehrforschung, bei der

es um recht verschiedene Dinge gehen kann, die alle ihren Einfluss auf den Unterrichtsprozess haben, muss man seinen Interessensgegenstand sehr klar eingegrenzt haben, was keineswegs einfach ist. Ein Beispiel dazu: Nehmen wir an, es soll untersucht werden, ob man das Hörverständnis von Fremdsprachenlernern besser mit der Methode X oder mit der Methode Y fördern kann. Dann wird eine ganz entscheidende Frage für den Wert der Untersuchung sein, ob es gelingt, den Faktor „Hörverständnis" hinreichend von anderen Faktoren zu isolieren (wir operationalisieren das Konstrukt). Denn ob die Lerner richtig reagieren, hängt nicht nur vom eigentlichen Hörverständnis ab, sondern sie nehmen z.B. ihr Weltwissen und Informationen aus der Situation zu Hilfe, und in der Planung der Untersuchung muss man das berücksichtigen.

Auch bei linguistischen Untersuchungen kommt es vor, dass der Untersuchungsgegenstand nicht genau genug bestimmt wird, selbst in vermeintlich klaren Fällen wie etwa einer Auszählung, in welchen Satzarten bestimmte Modalpartikeln vorkommen. Wenn z.B. die Beschreibung der Fälle, in denen *eben* als Modalpartikel angesehen wird (*Männer sind **eben** so*), nicht exakt genug ist, werden Adverbien mitgezählt (***Eben** war Fritz noch nüchtern).*

Hat man dann einen interessanten Untersuchungsgegenstand festgelegt, sind noch weitere Planungen notwendig, bevor eine Forschungsfrage gestellt wird: Das Projekt muss machbar sein und es muss auf der Basis des relevanten Informationsstandes der Forschung aufgebaut werden.

Durchführbarkeit

Machbar zu sein bedeutet mehreres. So muss das Projekt so weit eingegrenzt werden, dass es auch mit den zur Verfügung stehenden Mitteln und in der zur Verfügung stehenden Zeit durchgeführt werden kann. Ist es dagegen zu weit eingegrenzt, läuft man Gefahr, dass die Ergebnisse nicht mehr valide sind (s. Kapitel 2.4.3). Man wird also z.B. Überlegungen zur Größe der untersuchten Gruppe anstellen. Möchte man herausfinden, ob es einen Unterschied zwischen Chinesen und Russen in der benötigten Unterrichtszeit zum Erreichen des GER A1-Niveaus in Deutsch gibt, kann man unmöglich sämtliche chinesischen und russischen Lernenden testen, aber auch nicht einen chinesischen und einen russischen Lerner, die man zufällig kennt. Ebenso wenig ist es für die meisten Untersuchungen sinnvoll, „Sprachkenntnisse" generell überprüfen zu wollen: Man wählt einen bestimmten, eingegrenzten Bereich aus und versucht, diesen genau zu untersuchen. So könnte man sich dafür entscheiden, Genuskongruenz in der Nominalphrase **oder** Erfolg beim Verstehen von Richtungsanweisungen **oder** das Ergebnis im Sismik-Test bei Kindergartenkindern einer bestimmten Herkunft zu untersuchen.

Machbar zu sein bedeutet aber auch, dass alles, was man für die Untersuchung braucht, auch vorhanden oder zu beschaffen ist. Möchte man lexikalische Entscheidungstests durchführen (dies wird im Kapitel „Experimente" noch genauer erklärt), dann kann die Studie nur dann durchgeführt werden, wenn die Forscherin auch eine zuverlässige Möglichkeit hat, Reaktionszeiten zu messen. Möchte man Fehler analysieren, die deutschsprachige Lernende

beim Schreiben arabischer Texte machen, muss man diese Fehler auch kompetent erkennen können (d.h. man muss entweder selbst fundierte Arabischkenntnisse oder Zugriff auf jemanden mit diesen Kenntnissen haben). Und ebenso braucht man eine Gruppe von Menschen (Versuchspersonen), die zur Forschungsfrage passen – möchte man den bilingualen Spracherwerb untersuchen, hilft es wenig, wenn man nur erwachsene Lernende kennt.

Kenntnis der bisherigen Forschung

Zudem muss die Studie informiert sein. Das bedeutet nichts anderes, als dass man sich – genau wie für eine Seminararbeit – vor der endgültigen Formulierung einer Forschungsfrage durch vertiefte Literaturrecherche über die vorliegenden Erkenntnisse zum Thema informiert, und zwar sowohl über den Gegenstand als auch über die mögliche Forschungsmethodik. Das hilft sowohl, das Projekt zu planen, als auch Fallen beim Untersuchungsaufbau zu vermeiden: Man lernt aus den Überlegungen anderer, vor allem, wenn dabei die einzelnen Schritte der Planung und Durchführung besprochen werden.

Auch muss man wissen, zu welchem Zweck die Daten erfasst werden sollen. Bereits vor der Datenerhebung muss man überlegen, welche Schlussfolgerungen man hinterher tatsächlich aus den Ergebnissen ziehen kann. Soll z.B. eine bestehende Hypothese oder Theorie überprüft werden, so wird man sich fragen, welche Vorhersagen diese Hypothese für bestimmte Situationen trifft. Dann kann man versuchen, diese Situationen zu beobachten oder sie künstlich zu schaffen, um zu überprüfen, ob sich die Wirklichkeit den Vorhersagen dieser Theorie entsprechend verhält.

Viele laienhaft angelegte empirische Untersuchungen sammeln Daten, die dann nicht interpretierbar sind. Man beobachtet z.B. bei Befragungen oft, dass Studierende Fragen stellen, ohne sich vorher genug überlegt zu haben, was die möglichen Antworten für ihre Untersuchung überhaupt bedeuten können („Ich sehe mal, was dabei herauskommt"). Den daraus resultierenden Problemen kann man entgehen, indem man, nachdem der Fragebogen oder das Interview konzipiert ist, systematisch untersucht, ob die zu erwartenden Antworten überhaupt für die Argumentation zu gebrauchen sind. Wenn man z.B. für eine Untersuchung zum Gebrauch der Vergangenheitstempora bei Katholiken und Nicht-Katholiken eine Operationalisierung (d.h. wie man das zu Untersuchende im Rahmen einer Untersuchung definiert, auf etwas Beobachtbares/Messbares hin konkretisiert und dieses Konstrukt dann misst) von „katholisch" als „Religionszugehörigkeit nach der Lohnsteuerkarte" bestimmt, dann sind Fragen wie „Gehen Sie regelmäßig in die Kirche?" oder „Was halten Sie vom Papst?" nicht relevant für die Untersuchung. Derartige Fragen hätten nur einen Sinn, wenn anzunehmen wäre, es spiele beim Gebrauch der Tempora eine Rolle, ob jemand das „Katholisch-Sein" mehr oder weniger intensiv praktiziert.

Operationalisierung

Replikationsstudien

Im Übrigen sind, vor allem für erste empirische Versuche, z.B. im Rahmen von Seminar-, Master- oder Examensarbeiten, Replikationsstudien sehr zu empfehlen (vgl. Porte 2002, 35). Replikationsstudien versuchen, die Ergebnisse aus anderen Studien zu bestätigen. Sie können sehr eng an die Vor-

gaben der ersten Studie angelehnt werden, indem sie z.B. dieselbe Untersuchungsfrage und dasselbe Untersuchungsdesign einsetzen, oder leicht unterschiedlich sein, indem sie z.B. mit einer unterschiedlichen Probandengruppe oder einer unterschiedlichen Zielsprache die gleiche Untersuchung durchführen. Wenn Sie eine Replikationsstudie durchführen, haben Sie die Gelegenheit, erstens den Forschungsprozess stark geleitet zu erleben und zu reflektieren, zweitens aber auch, die Ergebnisse aus anderen Untersuchungen zu hinterfragen (und zu bestätigen bzw. in Frage zu stellen).

Wir resümieren kurz, worüber man sich im Klaren sein muss, bevor man sich an den Aufbau einer Fragestellung begibt:

1. Was genau soll untersucht werden?
2. Ist es durch die Auswahl einer geeigneten Methode möglich, den Untersuchungsgegenstand tatsächlich zu erforschen?
3. Bin ich über die vorhergehende Forschung zum gleichen Untersuchungsgegenstand und über mögliche Methoden informiert?
4. Zu welchem Zweck wird die Studie durchgeführt?
5. Wie sollen die zu untersuchenden Variablen operationalisiert werden?

Hat man diese Vorfragen zufriedenstellend geklärt, ist der nächste Schritt, eine angemessene Fragestellung zu formulieren.

2.2 Was ist eine Forschungsfrage?

Eine quantitativ angelegte empirische Studie bestimmt eine (oder mehrere) Forschungsfrage(n) und stellt Hypothesen auf, die sich auf diese Fragestellung beziehen. Die formulierte Forschungsfrage verdeutlicht möglichst präzise, worum es in der Studie geht; sie entwickelt sich aus dem Forschungsinteresse und der Literaturrecherche und spiegelt häufig den theoretischen Rahmen wider, in dem die Studie eingebettet ist.

Nehmen wir an, wir interessieren uns für den Einfluss einer Sprachlernsoftware auf das Lernen des Unterschieds zwischen dem *present perfect tense* und dem *simple past tense* im Englischen. Angemessen ist eine Fragestellung, die das Forschungsinteresse möglichst klar darstellt, z.B. „Lernen Schüler, die – nach derselben Einführung in den Unterschied zwischen den beiden Tempusformen – dazu zwei Stunden Übungen mit dem Softwarelernprogramm X machen, besser als Schüler, die im gleichen Zeitraum dieselben oder sehr ähnliche Übungen im Arbeitsbuch lösen, und zwar gemessen an der Leistung bei einem Entscheidungstest mit diesen beiden Tempusformen?“ Diese Frage sagt uns, (1) was der Forschungsgegenstand ist (Vergleich zwischen dem Lernen am Rechner und dem Lernen mit einem Arbeitsbuch in einer Situation, in der möglichst nur das Lernmedium verschieden ist), (2) wie der Forschungsgegenstand operationalisiert wird (Lernen des Unterschieds zwischen *present perfect* und *simple past*), und (3) wie „Lernen“ operationalisiert wird (z.B. Ergebnis bei einem Test, in dem die

Schüler entscheiden müssen, ob die richtige Tempusform eingesetzt worden ist).[1] Problematisch dagegen wäre eine Fragestellung wie die folgende: „Lernen Schüler besser mit Hilfe eines Sprachlernprogramms?", denn diese Frage sagt uns weder, was unter „Lernen" verstanden wird, noch was die Schüler lernen sollen (man kann z.B. relativ sicher sein, dass ein Sprachlernprogramm weniger hilfreich ist, wenn man in der Fremdsprache streiten lernen möchte), noch wie das Lernen gemessen werden sollte.

Forschungsfragen können prinzipiell in drei Kategorien geteilt werden: deskriptiv, korrelativ oder kausal. Eine deskriptive Forschungsfrage interessiert sich für die Beschreibung einer Begebenheit, z.B.: „Wie häufig erhalten Grundschüler mit türkischem Migrationshintergrund eine Gymnasialempfehlung?" Eine korrelative Fragestellung fragt, welche Variablen häufig zusammenkommen, z.B.: „Erhalten Grundschüler mit türkischem, russischem und italienischem Migrationshintergrund mit unterschiedlicher Häufigkeit eine Gymnasialempfehlung im Vergleich zu autochthonen Kindern?" Eine kausale Fragestellung versucht dagegen, Gründe oder Auslöser für bestimmte Variablen nachzuweisen, z.B.: „Erhalten Grundschüler mit türkischem Migrationshintergrund häufiger eine Gymnasialempfehlung, wenn ihre Schulnoten und das Ergebnis in der DESI-Studie anonymisiert an eine externe Kommission gegeben werden?" Welche Art von Frage Sie stellen, hängt vom Erkenntnisinteresse ab und beeinflusst die gestellten Hypothesen sowie das Untersuchungsdesign.

deskriptiv

korrelativ

kausal

2.3 Wie formuliere ich eine Hypothese?

Eine Hypothese ist ein Satz, der empirisch falsifizierbar ist. Empirische Forschung untersucht, ob eine bestimmte Hypothese der Überprüfung in der Realsituation standhält – es wird gefragt, ob die Ergebnisse die formulierte(n) Hypothese(n) unterstützen oder nicht. Somit kann eine Hypothese (zumindest vorerst) bestätigt oder verworfen werden.

falsifizierbar

Plant man, eine empirische Studie durchzuführen, hat man meistens eine Idee, was dabei als Ergebnis herauskommen könnte. Deswegen sind Hypothesen normalerweise direktional – das bedeutet, dass eine bestimmte Richtung vermutet wird. Hypothesen können auch nicht-direktional sein; in diesem Falle besagen sie einfach, dass eine Beziehung zu finden sein wird – aber nicht, was für eine. (Die Nullhypothese, die in anderen Wissenschaften wie z.B. der Psychologie häufig angewendet wird, besagt, dass es keine Beziehung zwischen den einzelnen untersuchten Faktoren in der Studie gibt. Da die Aufstellung einer Nullhypothese in der Sprachlehrforschung eher untypisch ist, wird sie hier nicht weiter behandelt.)

direktional

Nullhypothese

[1] Weitere Probleme – die Gruppen können vorher unterschiedlich viel gewusst haben, wir wissen nicht, was wirklich der Lernzuwachs ist, wenn der Test nicht vorher schon gemacht wurde, usw. – behandeln wir später in diesem Kapitel bei den Stör- und Kontrollfaktoren.

Die wichtigsten Kriterien für eine falsifizierbare Hypothese sind:[2]

1. Eine Hypothese ist eine Aussage, die Allgemeingültigkeit anstrebt – das heißt, sie geht über den Einzelfall hinaus. Bei der Fragestellung in 2.2 könnte die Hypothese lauten: „Schüler, die mit der Sprachlernsoftware üben, erzielen bessere Ergebnisse als Schüler, die nur mit dem Arbeitsbuch üben." Hier sehen wir auch gleich die vermutete Richtung, wir sagen also für eine der beiden Möglichkeiten des Übens voraus, dass sie bessere Ergebnisse haben wird. In unserer Untersuchung werden wir natürlich nicht sämtliche Schüler, die es gibt, untersuchen; trotzdem soll die Hypothese nicht nur für die Schüler gelten, die an unserer Untersuchung teilgenommen haben.
2. Die Konstrukte werden (wenn auch meist implizit) durch den logischen Operator „wenn-dann" (bzw. „je-desto") verbunden. Also in unserem Fall, **wenn** ein Schüler am Computer übt, **dann** lernt er besser (als wenn er mit einem Buch übt).
3. Die Aussage ist potenziell falsifizierbar – es muss möglich sein, zu beweisen, dass die Hypothese nicht gilt. Es ist durchaus denkbar, dass unsere Untersuchung entweder keinen Vorteil für Schüler, die mit der Lernsoftware geübt haben, ergibt (also beide Gruppen im Test gleich gut sind) oder dass es einen Vorteil für die zweite Gruppe (Kontrollgruppe) gibt. In beiden Fällen wäre die Hypothese zu verwerfen.

Übrigens sollen Hypothesen natürlich auch eine sinnvolle Fragestellung betreffen und theoretisch hergeleitet (d.h. nicht nur auf persönlicher Erfahrung beruhend) sein. Eine Hypothese wie „Studenten, die Deutsch als Erstsprache sprechen, machen im Deutschen weniger Genusfehler als Studenten, die Englisch als Erstsprache sprechen" wird höchstwahrscheinlich bestätigt – sie ist aber ziemlich uninteressant.

Beispiel

Versuchen wir es jetzt mit einem etwas problematischeren Beispiel. Man könnte die Hypothese aufstellen: „Katholische Deutschsprachige gebrauchen bei der Bezeichnung von Vergangenem das Perfekt, andere Deutschsprachige nicht." Die Hypothese versucht, eine Aussage über katholische Deutschsprachige im Allgemeinen zu machen – also nicht nur solche, die z.B. in Mainz leben, – sowie über alle weiteren Deutschsprachigen. In unserer Hypothese gilt aber implizit auch eine Universal-Aussage, d.h. ein einziger Katholik, der das Präteritum oder das Plusquamperfekt statt des Perfekts gebrauchen würde, würde unsere Hypothese schon widerlegen. Ebenfalls wäre die Hypothese bereits mit dem Nachweis falsifiziert, dass eine einzige nicht-katholische Person einmal das Perfekt benutzt hat.

Eine realistischere Hypothese wäre dagegen: „Katholische Deutschsprachige gebrauchen das Perfekt häufiger als nicht-katholische." Die Hypothese besagt also, dass wenn ein Sprecher katholisch ist, dann gebraucht er das Perfekt bei der Bezeichnung von Vergangenem häufiger, als wenn er nicht ka-

[2] Die folgende Darstellung lehnt sich an die Ausführungen in Bortz/Döring 1995, 7 an.

tholisch ist. Die Hypothese könnte jetzt durch einen statistischen Befund bestätigt werden, nämlich durch den, dass man, wenn man eine hinreichend große Gruppe von katholischen und von nicht-katholischen Sprechern des Deutschen untersucht, bei den katholischen Sprechern, bezogen auf die Gesamtverteilung der Vergangenheitstempora, prozentual mehr Perfekt findet als bei den nicht-katholischen. Die Hypothese wäre falsifiziert bzw. widerlegt, wenn sich kein im statistischen Sinne signifikanter Unterschied (d.h. kein Unterschied, der groß genug ist, um den Zufall mit ausreichender Sicherheit als Ursache auszuschließen) in der Häufigkeit des Perfektgebrauchs aufzeigen lässt. Eine Falsifizierung der aufgestellten Hypothese heißt noch nicht, dass das Gegenteil bewiesen wäre (also dass Katholiken das Perfekt weniger gebrauchen als andere Sprecher des Deutschen), es heißt nur, dass unsere Daten keine Unterstützung für die Annahme liefern, Katholiken gebrauchten mehr Perfekt als Nicht-Katholiken.

Statistische Signifikanz

Eine gute Hypothese ist also eine Aussage, die sich direkt auf die Forschungsfrage bezieht, die falsifizierbar ist, die Beziehungen zwischen den untersuchten Faktoren darstellt, die Konstrukte verwendet, die man operationalisieren (bestimmen und beobachten) kann, und die durch die gesichtete Literatur unterstützt wird (oder für die es zumindest eine gute Erklärung gibt, warum sie bestimmte Ergebnisse vorhersagt).

2.4 Gütekriterien für empirische Untersuchungen

Bei jeder empirischen Untersuchung entstehen einige naheliegende grundsätzliche Fragen, die bei der Planung beachtet werden müssen und die wir anhand von drei Beispielen erläutern wollen.[3]

Angenommen, wir wollen die Übersetzungsfertigkeit von Studenten messen und haben dazu einen Text von 200 Wörtern ausgesucht, der in die Fremdsprache übersetzt werden sollte. Manche Studenten haben eine gute Übersetzung geschrieben, manche eine mittelmäßige oder schlechte. Wissen wir auf Grund dieser Übersetzungen dann, ob diese Studenten gut oder schlecht übersetzen können? Oder könnte es sein, dass wir ganz andere Resultate bekommen hätten, wenn wir einen anderen Text ausgesucht hätten, einen Text über ein anderes Thema, einen viel längeren Text oder einfach eine Liste von Wörtern?

Und angenommen, wir hätten diesen Text von einer von drei Gruppen Erstsemester-Studenten der Anglistik übersetzen lassen, können wir dann anhand der Resultate dieser Übersetzungen etwas über das zu erwartende Übersetzungsfertigkeitsniveau der anderen beiden Gruppen aussagen? Oder über die Fähigkeiten von Erstsemestern allgemein im Übersetzen?

[3] Für die qualitative Forschung gelten andere Gütekriterien, auf die wir hier nicht eingehen; wenn Sie sich weiter informieren möchten, verweisen wir auf Steinke 1999.

Oder nehmen wir an, wir lassen eine Dozentin Essays beurteilen, die Studenten in einem Kurs „Schriftlicher Ausdruck" geschrieben haben. Wie sicher können wir sein, dass die Kriterien, die die Dozentin benutzt, gut und konsistent sind? Würde eine andere Dozentin dieselben Noten vergeben? Und können wir, wenn ein Student eine gute Note für diesen Essay bekommen hat, annehmen, dass er sich gut schriftlich ausdrücken kann?

Beispiel

Um ein ausführliches Beispiel zu geben: Wir wollen herausfinden, wie Wörter in unserem mentalen Lexikon zusammenhängen, ob Wörter, die in bestimmten Kontexten häufig zusammen vorkommen, auch im Gehirn so repräsentiert sind, dass sie einander aktivieren. Wir betrachten Wörter aus demselben Script wie *Arzt*, *Krankenschwester*, *Krankenhaus* oder wie *Schlüssel* und *abschließen*, oder wir betrachten Wörter aus demselben Wortfeld wie *hell* und *dunkel* oder *Vogel* und *Spatz*. Um das zu untersuchen, haben wir ein sog. „lexikalisches Entscheidungsexperiment" entwickelt, in dem Wortpaare auf einem Computerbildschirm präsentiert werden, wobei das zweite „Wort" manchmal nur eine Buchstabenfolge ohne Bedeutung (ein Pseudowort) ist. Unter den Wörtern gibt es dann solche, die eine Beziehung zum ersten Wort des Wortpaares haben, und solche, die keine von den o.a. Beziehungen aufweisen. Das erste Wort wird kurz auf dem Computerbildschirm gezeigt und direkt danach das zweite. Die Versuchsteilnehmer (Studenten) müssen auf einen roten Knopf drücken, wenn das zweite kein Wort ist, und auf einen grünen, wenn es ein Wort ist. Dabei interessiert uns nur, wie sie auf tatsächliche Wörter reagieren. Wir messen die Reaktionszeit von der Präsentation des zweiten Wortes bis zum Drücken des Knopfes. Wir erwarten, dass die Reaktionszeit kürzer ist, wenn das präsentierte Wortpaar *Schlüssel – Tür* ist, als wenn das präsentierte Wortpaar *Schlüssel – Zug* ist.

praktische Fragen

Dabei kommen Fragen auf wie:

- Wie viele Wortpaare brauchen wir eigentlich, um ein einigermaßen zuverlässiges Ergebnis zu bekommen? Reichen fünf? Oder zehn? Oder brauchen wir erheblich mehr?
- Wenn wir Unterschiede in der Reaktionszeit finden, was können wir dadurch genau über unsere Fragestellung aussagen?
- Angenommen, die Reaktionszeit war tatsächlich kürzer, wenn es eine – wie auch immer geartete – Beziehung zwischen den beiden Wörtern des Wortpaares gab. Können wir dann sicher sein, dass die daraus gezogenen Schlüsse nicht nur für unsere kleine Gruppe von Versuchsteilnehmern gelten, sondern für alle Deutschsprachigen?
- Können wir sicher sein, dass die verwendeten Paare von Wörtern hinreichend ähnlich sind oder müssen wir befürchten, dass wir ganz verschiedene Arten/Grade von Beziehungen vermischen?
- Können wir sicher sein, dass die Auswertung der Ergebnisse nicht durch andere Faktoren (wie z.B. das Verhalten der Versuchsleiterin) beeinflusst wurde, weder während des Experiments noch bei der Interpretation der Daten?

Die oben gestellten Fragen beziehen sich unter anderem auf die Reliabilität, die Validität und die Objektivität der jeweiligen Untersuchung. Diese Gütekriterien behandeln wir zwar wie üblich getrennt, sie greifen aber häufig ineinander.

2.4.1 Zuverlässigkeit (Verlässlichkeit, Reliabilität)

Die Begriffe „Zuverlässigkeit", „Verlässlichkeit" und „Reliabilität" werden synonym gebraucht; sie bezeichnen dasselbe, und zwar, ob das Messverfahren das, was gemessen werden soll, exakt erfasst und ob die Daten, die damit gewonnen wurden, zuverlässig ausgewertet sind. Als verlässlich gilt eine Erhebung (und das bei dieser Erhebung benutzte Instrument) also dann, wenn die Messung genau ist.

In der Linguistik und Sprachlehrforschung sind v.a. zwei Arten von Zuverlässigkeit von Interesse: Bewerterzuverlässigkeit und Testzuverlässigkeit.

Bewerterzuverlässigkeit

In unserem Beispiel von oben zur Beurteilung des schriftlichen Ausdrucks in Essays kann es schwierig sein, die Bewerterzuverlässigkeit (engl. *rater reliability*) zu garantieren. Um Inter-Bewerterzuverlässigkeit zu bestimmen, müssten wir mindestens zwei Dozentinnen – unabhängig voneinander – die geschriebenen Texte bewerten lassen und die Ergebnisse dann miteinander vergleichen. Andererseits ist es auch wichtig, zu bestimmen, ob die Dozentinnen selbst die Ergebnisse konsistent bewerten (dass sie z.B. nicht je nach Müdigkeit ähnliche Texte sehr unterschiedlich bewerten); das ist dann Intra-Bewerterzuverlässigkeit. Bei Testverfahren, bei denen das Messinstrument sehr wenig Spielraum lässt, also strikte Vorgaben macht, wird die Bewerterzuverlässigkeit tendenziell höher sein (so z.B. bei unserem lexikalischen Entscheidungsexperiment). Bei Testverfahren, die subjektive Entscheidungen zulassen (wenn z.B. die Qualität bei „Schriftlicher Ausdruck" nicht ganz genau definiert wird), wird die Bewerterzuverlässigkeit niedriger sein. Deswegen ist es bei solchen Verfahren sehr wichtig, mindestens zwei unabhängige, kompetente Auswertungen der Ergebnisse machen zu lassen. Ebenso wichtig ist es sicherzustellen, dass Erwartungshaltungen seitens der Versuchsleiterin nicht dazu beitragen, dass Daten unterschiedlich bewertet werden. Daher ist es auch häufig sinnvoll, dass die Bewerterinnen die erwarteten Ergebnisse nicht vor der Datenauswertung kennen.

Testzuverlässigkeit

Die Testzuverlässigkeit dagegen versichert, dass das Testverfahren konsistent ist. Zur Ermittlung dieser gibt es im Prinzip drei Verfahren:

Testwiederholung: Unter gleichen Bedingungen sollten dieselben Ergebnisse erzielt werden. Bei einfachen Messverfahren ist dies einfach: Die Länge meines Tisches sollte dieselbe sein, wenn ich sie zweimal nacheinander mit demselben oder einem anderen Zollstock messe. Bei Untersuchungen in der Sprachlehrforschung ist es nicht mehr so einfach. So kann man – nach einer gewissen Zeit, damit sich die Versuchsteilnehmer nicht mehr daran erinnern – einfach denselben Test mit denselben Teilnehmern noch einmal machen.

Dabei sieht man, ob das Ergebnis der zweiten Durchführung des Tests mit der ersten übereinstimmt. Dieses Verfahren kann man bei Grammatikalitätsurteilen durchaus anwenden, bei den meisten Instrumenten (u.a. Sprachtests, Lese- oder Schreibaufgaben etc.) funktioniert es aber nicht, weil die Lerner in der Zwischenzeit – oder durch den ersten Test selbst – Lernfortschritte gemacht haben können, weil sie sich an den Test erinnern (das gilt insbesondere für Kinder, die sich manchmal überraschend lange z.B. an gelesene oder erzählte Geschichten oder an sonstige Testaufgaben erinnern können), oder weil sie wenig motiviert sind, denselben Test noch einmal durchzuführen.

Paralleltest: Man untersucht dieselben Versuchsteilnehmer ohne nennenswerten zeitlichen Abstand mit einer zweiten Version des Tests, den man eingesetzt hat. Das erfordert allerdings, dass die beiden eingesetzten Tests wirklich äquivalent sind. Zudem lässt sich das Verfahren nicht anwenden, wenn ein Lerneffekt durch das Bearbeiten des ersten Tests eintritt.

Interne Konsistenzprüfung: Wenn man die Versuchsteilnehmer nicht zweimal testen kann, kann man die Konsistenz innerhalb eines Tests überprüfen. Das Einfachste ist, man macht eine Testhalbierung (engl. *split-half*). Man unterteilt dazu die Ergebnisse des Tests in zwei Hälften – zum Beispiel alle geraden und alle ungeraden Fragen – und überprüft mit statistischen Verfahren (durch eine Berechnung der Korrelation zwischen den zwei Hälften), ob diese zwei Hälften wesentlich verschiedene Ergebnisse haben. Wenn das so ist, gilt der Test als inkonsistent und damit wenig verlässlich.

Wenn die Zuverlässigkeit überprüft wird, werden diese Ergebnisse meist in dem Kapitel, in dem man seine verwendete Methodik darstellt, anhand eines Korrelationskoeffizienten präsentiert (wie man den berechnet, behandeln wir in Kapitel 9 genauer).

2.4.2 Objektivität

Objektivität

Die Objektivität bezieht sich darauf, ob die Erhebung, Auswertung und Interpretation der Ergebnisse durch die Forscherin beeinflusst wurden. Es soll möglichst gesichert werden, dass Daten, die von der Forscherin notiert und ausgewertet werden, auch richtig erhoben wurden, ohne dass eine (subjektive) Interpretation einfließt. Je stärker die Auswertung nach einem fest vorgeschriebenen Schema verläuft, umso geringer ist die Gefahr von subjektiven Einflüssen. Das Messen von Reaktionszeiten durch den Computer beim o.a. lexikalischen Entscheidungsexperiment ist objektiver als die Beurteilung der Qualität von Schulaufsätzen durch Lehrpersonen. Objektivität und Zuverlässigkeit sind beide für die Herstellung von Gültigkeit notwendig, reichen aber hierfür nicht aus. Daher gehen wir jetzt zur Problematik der Gültigkeit über.

2.4.3 Gültigkeit (Validität)

Gültigkeit/Validität

Mit „Gültigkeit“ oder „Validität“ bezeichnet man, inwiefern das Messverfahren das misst, was es zu messen vorgibt. Wir kennen alle ironische Sprüche wie „Ich weiß zwar nicht genau, was ich messe, aber das messe ich ganz genau“. Die Validität einer Untersuchung ergibt sich also daraus, ob tatsächlich das erhoben, erfragt oder beobachtet und gemessen wird, was untersucht werden soll. Nicht nur das Messinstrument selbst (also z.B. der benutzte Fragebogen oder der benutzte Test), sondern das gesamte Untersuchungsdesign muss bei einer Prüfung der Gültigkeit kontrolliert werden, denn Fehler können auch in anderen Punkten liegen, etwa einer fehlerhaften Auswahl der Befragten oder der Verwendung von Begriffen, die von verschiedenen Personenkreisen unterschiedlich gebraucht werden, u.a.m.

Es gibt unterschiedliche Arten der Gültigkeit, wobei in der Sprachlehr- und -lernforschung vor allem interne und externe Validität von Belang sind. Deswegen gehen wir jetzt auf beide ein.

interne Validität

Die interne Validität bezieht sich darauf, inwiefern die Ergebnisse das abbilden, was sie abbilden sollen – und ob sie von weiteren Faktoren (Störfaktoren) beeinflusst worden sind. So ist es z.B. wichtig zu wissen, dass die Versuchspersonen, die wir testen, zur anvisierten Gruppe gehören. Will man untersuchen, unter welchen Bedingungen Sätze wie *Ich gehe gern ins Freibad, weil da sind die Leute so nett* für Muttersprachler akzeptabel sind, dann ist es schlecht, wenn sich unter den befragten Personen auch Nicht-Erstsprachler oder Bilinguale befinden (oder wenn viele Dialektsprecher dabei sind usw.). Ebenso kann bei wiederholten Messverfahren (s. u.a. Kapitel 10.1.2) die interne Validität beeinträchtigt werden, wenn viele Versuchspersonen bei den weiteren Messungen nicht mehr dabei sind – z.B. weil sie weggezogen sind, oder weil sie nicht mehr an der Studie teilnehmen wollen. Außerdem können Probleme entstehen, wenn ein Testverfahren sehr lang ist (Ermüdung führt meist zu schlechteren Ergebnissen, ebenso Langeweile) oder wenn die Versuchspersonen wissen, was die Forscherin von ihnen erwartet, und versuchen, sich so zu verhalten (das kann vor allem bei Befragungen problematisch werden). Schließlich kann interne Validität durch äußere Faktoren gestört werden, z.B. durch Lärm im Versuchsraum, der die Konzentration der Versuchsteilnehmer beeinträchtigt.

externe Validität

Mit externer Validität oder Geltungsbereich ist gemeint, unter welchen Gegebenheiten die Ergebnisse einer Untersuchung für bestimmte Untersuchungsobjekte gelten – über die Studie hinaus. Der Geltungsbereich bei sprachwissenschaftlichen Untersuchungen ist häufig einer der am ehesten angreifbaren Punkte. Sehr häufig wird recht naiv davon ausgegangen, dass Daten, die vor zwanzig Jahren oder vor noch längerer Zeit erhoben wurden, immer noch Aussagen über die heutige Sprache erlauben. Ebenso naiv wird oft angenommen, dass Daten, die in einer bestimmten Region erhoben worden sind, Aussagen über die Landessprache insgesamt erlauben.