DOI 10.1515/zgl-2012-0004 — ZGL 2012; 40(1): 38–67 DE GRUYTER

Stefan Th. Gries

Statistische Modellierung

Abstract: This paper provides an overview of central aspects of statistical modeling of linguistic data. Starting from a general definition of *model*, the paper discusses the goals of modeling as well as a variety of issues bearing upon the formulation/definition of statistical models. It then surveys model selection, the choice of 'the best model' and three fundamental notions affecting the interpretation of models. Finally, the paper turns to validation and replicability and addresses a variety of challenges researchers face during modeling processes.

Prof. Dr. Stefan Th. Gries: Department of Linguistics, University of California, Santa Barbara, CA 93106-3100, USA, E-Mail: stgries@linguistics.ucsb.edu

1 Einleitung

Die Sprachwissenschaft erfährt zurzeit einen massiven Wandel hin zu empirischeren Studien. In den fünfziger und sechziger Jahren des 20. Jahrhunderts führte der Aufstieg der generativen Grammatik dazu, dass empirische Studien aus der europäischen diachronen Sprachwissenschaft und dem amerikanischen Strukturalismus stark an Bedeutung verloren und von introspektiven und wenig nachprüfbaren Akzeptabilitätsurteilen abgelöst wurden. Mit Ausnahme eines harten Kerns in der generativen Grammatik hat sich die methodologische Landschaft in der Sprachwissenschaft wieder in Richtung methodologisch ausgereifterer Verfahren entwickelt, was zu einem großen Teil dem Fortschritt in den Teildisziplinen der Korpus- und Psycholinguistik sowie anderen technologischen Entwicklungen geschuldet ist. Zum einen sind mittlerweile immer mehr Korpora verfügbar, die es gestatten, auch vergleichsweise seltene und/oder abstrakte Phänomene zu untersuchen; zum anderen haben experimentelle Ansätze, die anfänglich nur in der Psycholinguistik zu finden waren, ihren Weg in den sprachwissenschaftlichen methodologischen Mainstream gefunden.

Diese Hinwendung zu datenintensiveren Ansätzen ging mit einem gewachsenen Bewusstsein dafür einher, dass derartige Daten auch einer entsprechenden Analysemethodik bedürfen. Die Menge und Komplexität der Daten, die sich aus Korpora und Experimenten ergeben, macht es üblicherweise erforderlich, die Daten mit statistischen Methoden auszuwerten, insbesondere sog. *multifaktoriellen Verfahren*, bei denen zwei und mehr Prädiktoren und ihre Effekte auf eine abhängige Variable analysiert werden. Während komplexere statistische Methoden in der Psychologie und Psycholinguistik schon lange üblich waren, wächst diese Erkenntnis und Akzeptanz in vielen anderen Feldern der Sprachwissenschaft noch. Zumindest oberflächlich betrachtet ist die Soziolinguistik hier eine Ausnahme, da in diesem Feld bereits in den siebziger Jahren des letzten Jahrhunderts die sog. *variable rule analysis* (VARBRUL, vgl. Paolillo 2002) verwendet wurde, ein multifaktorielles Verfahren zur statistischen Analyse soziolinguistischer Variation; in Abschnitt 2.2 werde ich allerdings argumentieren, dass VARBRUL eine Reihe von Problemen aufweist, die ihren Vorreiteranspruch unterminieren. Trotz VARBRULs Schwächen ist jedoch nicht zu leugnen, dass multifaktorielle Ansätze erst deutlich später Eingang in die Korpus- und kognitive Linguistik gefunden haben (Leech, Francis & Xu 1994 und Gries 1999 sind die ersten solchen Studien).

Aus der Tatsache, dass komplexere statistische Verfahren in vielen Teilbereichen der Sprachwissenschaft noch nicht lange Fuß gefasst haben, folgt leider auch, dass viele Anwender noch mit Anfangsschwierigkeiten zu kämpfen haben. Zwar gibt es inzwischen einige Einführungen in die Statistik für Sprach-

wissenschaftler (Baayen 2008, Gries 2008, 2009, Johnson 2008)[1] und viele Überblicksartikel, doch besteht weiterhin ein Bedarf an Ausbildung und Training – insbesondere, da in vielen Institutionen statistisches Wissen für Linguisten nicht oder nicht hinreichend unterrichtet wird. Der vorliegende Artikel diskutiert einen der häufigsten Aspekte statistischer Analysen, nämlich die quantitative Modellierung sprachwissenschaftlicher Daten, mit besonderem Blick auf Probleme und Risiken solcher Analysen. In Abschnitt 2 wird kurz der Begriff des statistischen Modells definiert sowie erläutert, wie solche Modelle formuliert werden und welche Arten und Ziele solcher Modelle üblicherweise unterschieden werden; dabei wird besonders auch auf häufige Fehler hingewiesen, die es bei der Formulierung von Modellen zu vermeiden gilt. Abschnitt 3 ist den Themen der Modellselektion und -interpretation gewidmet: Wie ermittelt man das beste Modell für die vorliegenden Daten, und wie ist das Modell zu verstehen? Abschnitt 4 diskutiert kurz, wie Modelle validiert und repliziert werden können. Abschnitt 5 behandelt dann mehrere Herausforderungen, die komplexe Daten häufig beinhalten, bevor Abschnitt 6 abschließt.

2 Modelle, ihre Formulierung und ihre Ziele

2.1 Die Begriffe *Modell* und *Modellierung*

Die Ziele, die gemeinhin mit empirischen wissenschaftlichen Studien verfolgt werden, sind

- *beschreiben*, i. e., das Beantworten der Frage „was passiert hier?“
- *erklären*, i. e., das Beantworten der Frage „warum passiert …?“
- *vorhersagen*, i. e. das Beantworten der Frage „was wird passieren, wenn …?“

Ein (statistisches) Modell bezieht sich hauptsächlich auf Ziele 1 und 3: Ein *Modell* ist eine formale Repräsentation, die Zusammenhänge zwischen Prädiktoren – unabhängigen Variablen und ihren Interaktionen (oft angenommenen Ursachen) – und abhängigen Variablen (oft angenommenen Effekten) mittels mathematischer Relationen/Gleichungen so beschreibt, dass Gleichungen Vorhersagen für Situationen

1 Ich verweise hier nur auf die Einführungen, die nichtkommerzielle Software verwenden; die statistische Software, die in der Linguistik am meisten an Anwendern gewinnt, ist R (R Development Core Team 2011).

gestatten, die den beobachteten ähnlich (genug) sind.[2] Erklären kann ein statistisches Modell dagegen nichts – warum passiert, was das Modell beschreibt, muss durch den Forscher in Form einer Theorie postuliert und begründet werden.

Der Begriff *Modellierung* bezieht sich auf den Prozess, mittels dessen ein Modell entwickelt und dann geprüft wird, das (i) die empirischen Daten gut (genug) beschreibt und (ii) konservativ ist, d. h. nicht mehr Prädiktoren als unbedingt nötig beinhaltet. Diese zweite Bedingung ist als *Ockhams Rasiermesser* bekannt (vgl. Crawley 2007: 325): ‚*Entia non sunt multiplicanda praeter necessitatem*, frei übersetzt im momentanen Kontext, es ist die Beschreibung (i. e., das Modell) eines Phänomens *P* vorzuziehen, die mit der geringsten Menge an Prädiktoren die Daten gut genug beschreibt'. In der Praxis bedeutet dies, dass wenn der Unterschied darin, wie gut zwei Modelle *P* beschreiben, nicht signifikant ist, dann muss man das einfachere der beiden Modelle annehmen.

Häufig beginnt dieser Prozess mit einer vorläufigen Formulierung eines Modells (siehe Abschnitt 2.2), welches dann auf die beiden obigen Bedingungen hin geprüft wird und, falls nötig, in einem iterativen Prozess, der sog. *Modellselektion* (siehe Abschnitt 3) solange modifiziert wird, bis entweder ein Modell übrig bleibt, das die beiden Bedingungen erfüllt oder eins, das alle Prädiktoren als wirkungslos (i. e., nicht signifikant) eliminiert hat und man schlussfolgern muss, dass das untersuchte Phänomen mit den angenommenen Prädiktoren nicht gut (genug) beschrieben werden kann. In dem Rest dieses Abschnitts und in Abschnitt 3 werden diese Aspekte genauer beschrieben.

2.2 Modellformulierung

Die Formulierung eines vorläufigen Modells als erstem Schritt der Modellierung beinhaltet üblicherweise, aber nicht notwendigerweise, die Entwicklung einer *Modellgleichung*, die versucht, die Verteilung einer (oder mehrerer) abhängigen Variablen durch die lineare Kombination einer (oder mehrerer) unabhängiger Variablen und ggf. ihrer Interaktionen zu beschreiben. Als *Interaktion* bezeichnet man das nicht additive, i. e. das nicht erwartbare, Zusammenwirken zweier oder mehrerer Variablen. Zum Beispiel, wenn hohe Werte von Variable A generell die Reaktionszeit auf einen Stimulus verringern und hohe Werte von Variable B generell auch die Reaktionszeit verringern, jedoch hohe Werte von A *und* B

2 Diese Definition des Begriffs *Modell* ist liberaler als manchmal üblich, da sie beispielsweise auch explorative multivariate Verfahren wie Clusteranalysen oder Faktorenanalysen mit einschließt. Hier werden jedoch nur regressionsanalytische und ähnliche Methoden behandelt.

zusammen die Reaktionszeit auf einen Stimulus erhöhen oder weniger verringern als erwartet, dann liegt eine Interaktion – ein unerwartetes Zusammenwirken – von A und B vor. In (1) wird eine einfache Modellformulierung in Form einer Gleichung exemplifiziert: eine abhängige Variable wird beschrieben als eine Funktion von („~“) der linearen Kombination („+“) dreier Prädiktoren, nämlich einer unabhängigen Variablen$_1$, einer unabhängigen Variablen$_2$ und der Interaktion der beiden unabhängigen Variablen (angezeigt durch den Doppelpunkt).

(1) $\text{AbhVar} \sim \text{UnabhVar}_1 + \text{UnabhVar}_2 + \text{UnabhVar}_1 : \text{UnabhVar}_2$

(2) $\text{AbhVar}_{\text{Vorhersage}} = b_0 + b_1 \times \text{UnabhVar}_1 + b_2 \times \text{UnabhVar}_2 + b_3 \times \text{UnabhVar}_1 : \text{UnabhVar}_2$

Mit anderen Worten, die in (1) abgebildete Modellgleichung ist der Versuch, die Relation zwischen einer abhängigen Variablen und mehreren Prädiktoren abzubilden – im Idealfall so gut, dass die Werte/Ausprägungen der abhängigen Variable auf der Basis der Prädiktoren vorhergesagt werden können. Bei der vorhergesagten abhängigen Variable kann es sich um direkt relevante Variablen handeln (z. B. Reaktionszeiten oder Akzeptabilitätsurteile), aber auch um Wahrscheinlichkeiten kategorialer Variablenausprägungen (z. B. konstruktionale Alternativen, richtige-vs.-falsche Kategorisierungen) oder um Häufigkeiten etc. Diese Modellgleichung wird dann mit einer Form von Regressionsanalyse untersucht, und als Resultat erhält man für jeden Prädiktor

- einen Koeffizienten (repräsentiert in (2) als b_1, $b_2$und b_3), der anzeigt, wie sehr und in welche Richtung der Prädiktor die abhängige Variable beeinflusst; ein positiver Koeffizient bedeutet, dass der entsprechende Prädiktor die Werte der vorhergesagten abhängigen Variable erhöht, da die Vorhersage der abhängigen Variablen darauf beruht, dass der Wert / die Ausprägung des entsprechenden Prädiktors mit dem Koeffizienten multipliziert wird;
- eine Prüfstatistik (oft ein *t*-Wert) und ihre Streuung (oft ein Standardfehler), die herangezogen werden, um einen Signifikanzwert/*p*-Wert für den entsprechenden Prädiktor zu berechnen;
- einen *p*-Wert für den entsprechenden Prädiktor, um anzuzeigen, ob dieser Prädiktor einen signifikanten (d. h., höchstwahrscheinlich nicht zufälligen) ‚Einfluss‘ auf die abhängige Variable hat.

Dieser Analyseschritt wird in Abschnitt 3 ausführlicher diskutiert.

Während dieser erste Schritt und diese Art der Repräsentation unkontrovers erscheinen mögen, darf nicht übersehen werden, dass bereits an dieser Stelle erste und ernsthafte Risiken bestehen, nämlich dass schon die erste Formulierung eines Modellkandidaten (i) fehlerhaft und/oder (ii) unvollständig ist, was im Folgenden kurz kommentiert werden soll.

Eine Modellgleichung kann z. B. fehlerhaft sein, weil sie Prädiktoren auf eine Weise beinhaltet, die deren ‚wahrer Natur' nicht gerecht wird. Frequenzeffekte sind hier ein gutes Beispiel. Oftmals werden Frequenzeffekte am besten auf einer logarithmierten Skala gemessen, d. h. auch wenn $Wort_1$ 10 Mal so häufig ist wie $Wort_2$, dann ist der Effekt von $Wort_1$ auf die abhängige Variable – z. B. Reaktionszeit – nicht 10 Mal so stark wie der von $Wort_2$, sondern vielleicht nur *ln*(10) so stark (vgl. Tryk 1986). Das bedeutet, dass nicht Frequenz, sondern *ln*(Frequenz) in der Modellgleichung stehen sollte, was einen massiven Einfluss auf die statistische Auswertung haben kann. Besonders mit psycholinguistisch relevanten Variablen können derartige Effekte häufiger als erwartet auftauchen, da z. B. Lerneffekte, Vergessenseffekte, Habituationseffekte etc. oft besser logarithmisch skaliert werden sollten. Es kann daher nicht genug betont werden, wie wichtig eine sorgfältige (und häufig visuelle) Inspektion der Daten ist, um derartige Entscheidungen im Vorwege treffen zu können.

Ein weiterer verwandter und wichtiger Aspekt ist, dass die Prädiktoren idealerweise auf dem höchstmöglichen Informationsniveau gemessen werden und *nicht* faktorisiert werden, ohne dass die Daten eine solche Transformation ohnehin suggerieren oder wenigstens zulassen würden. Solches Faktorisieren von Intervalldaten – häufig in VARBRUL-Analysen oder wenn Forscher unbedingt eine ANOVA verwenden wollen – führt nicht nur potentiell zu einem großem Informationsverlust, sondern kann auch die nachfolgende Modellselektion beeinträchtigen: Baayen (2010a) zeigt, wie Regressionen eine bessere Trennschärfe als ANOVAs aufweisen; vgl. auch Harrell (2001: 6).

Neben der o. g. potentiellen Fehlerhaftigkeit kann eine Modellgleichung auch auf zwei Arten unvollständig sein. Zum einen ist es möglich, dass wichtige unabhängige Variablen übersehen wurden – dies ist vergleichsweise trivial und hat wenig mit Modellierung *per se* zu tun. Zum anderen muss im Vorwege genau erwogen werden, ob und wie Interaktionen zwischen unabhängigen Variablen in den Modellierungsprozess einbezogen werden sollen – dies kann einschneidende Konsequenzen haben (und wird trotzdem in zahllosen Studien oft übersehen und/oder unterschätzt). Drei Positionen können hier unterschieden werden.

Einer Position zufolge ist es sinnvoll, Interaktionen von Anfang an mit einzubeziehen, um unerwartete Effekte aufspüren zu können: (i) Wenn eine Interaktion nicht in die Modellgleichung eingebaut wird, wird sie nicht auf Signifikanz getestet und man kann nicht klären, ob man mit der Interaktion das untersuchte Phänomen nicht vielleicht besser beschreiben könnte oder sogar müsste (siehe Abschnitt 2.2.2). (ii) Wenn man nur Interaktionen mit einbezieht, deren Existenz man theoretisch motivieren kann, dann wird es schwerer als nötig, jemals unerwartete und daher nicht *a priori* motivierte Effekte zu finden, bevor sie theoretisch motiviert werden können.

Die zweite Position wurde gerade eben schon impliziert: Prädiktoren werden nur eingeschlossen, wenn es theoretisch motivierte Gründe dafür gibt. Dieser Ansatz hat den Vorteil, dass er ‚exzessivem und unmotiviertem Fischen in den Daten' vorbeugt, aber eben auch den Nachteil, dass es schwierig wird, Unerwartetes zu finden.

Die dritte ‚Position' ist bedauerlicherweise wahrscheinlich noch die häufigste. In dieser Art Studien werden Interaktionen nicht als Prädiktoren miteinbezogen – entweder weil die Relevanz von Interaktionen nicht bekannt ist oder weil die verwendete Software die Analyse von Interaktionen erschwert – und leider ist dies wieder besonders häufig in soziolinguistischen VARBRUL-Analysen der Fall. Dies hat zwei wichtige unerwünschte Konsequenzen, die in den beiden folgenden Abschnitten erläutert werden.

2.2.1 Der potentielle Einfluss fehlender Interaktionen auf Modellgleichungen

Die erste unerwünschte Konsequenz folgt aus dem obigen: Wenn Interaktionen nicht in der Modellgleichung enthalten sind, können sie *per definitionem* nicht erkannt werden. Das hat zur Folge, dass

- die Analyse ein Phänomen weniger gut beschreibt als möglich: die Variabilität in den Daten, die durch die Interaktion beschrieben würde, bleibt unbeschrieben;
- schlimmer noch, diese unbeschriebene Varianz – die sog. *Residuen* oder *Residualvarianz* – beeinträchtigt die Resultate für *alle* Prädiktoren, die in der Modellgleichung enthalten sind. Damit sind in der Regel nicht nur die Koeffizienten der Prädiktoren falsch, sondern auch ihre *p*-Werte.

Hier ein stark vereinfachtes Beispiel, um diesen letzten Punkt zu illustrieren. Nehmen wir an, 80 Schüler (Muttersprachler des Deutschen) – 40 aus je einer von zwei Schulklassen – hätten an einem Diktat in Deutsch und einem Diktat in Englisch teilgenommen. Um zu prüfen, ob es einen Zusammenhang zwischen der Fehlerzahl im Englischdiktat (die abhängige Variable) und der Fehlerzahl im Deutschdiktat (eine unabhängige Variable) sowie der Klassenzugehörigkeit gibt (eine zweite unabhängige Variable), können zwei multifaktorielle Modellgleichungen formuliert werden: eine mit der Interaktion (vgl. (3)), eine ohne (vgl. (4)):

(3) FehlerEngl ~ FehlerDeut + Schulklasse + FehlerDeut:Schulklasse
(4) FehlerEngl ~ FehlerDeut + Schulklasse

Werden beide Modellgleichungen in ein lineares Modell eingegeben, erhält man für diesen fiktiven Datensatz die Resultate in Tabelle 1 (für (3)) und Tabelle 2 (für (4)).

(Diese Tabellen werden in Abschnitt 3.2 genauer diskutiert; das entsprechend besser geeignete gemischte Modell wird hier nicht behandelt.)

	Quadrat-summe	**Koeffizient**	**Standard-fehler**	***t***	***p***
Achsenabschnitt	24,9	2,82208	1,15488	2,444	0,0169
FehlerDeut	2461,42	**1,64459**	0,06769	**24,294**	‹ 0,0001
Schulklasse	0,25	**–0,28249**	1,15488	–0,245	0,8074
FehlerDeut: Schulklasse	241,73	–0,51538	0,06769	–7,613	‹ 0,0001
Residualvarianz	**316,95**				
Korrelation/ Signifikanztest	Mult. R^2 = 0,985	Korr. R^2 = 0,984		$F_{3,\ 76}$ = 1661	p ‹ 0,0001

Tabelle 1: Resultate des Modells in (3) mit der Interaktion FehlerDeut:Schulklasse

	Sum Sq	**Koeffizient**	**Standard-fehler**	***t***	***p***
Achsenabschnitt	23,61	2,74795	1,52324	1,804	0,0751
FehlerDeut	2931,69	**1,75395**	0,08726	**20,101**	‹ 0,0001
Schulklasse	3010,30	**–8,72058**	0,42813	–20,369	‹ 0,0001
Residualvarianz	**558,68**				
Korrelation/ Signifikanztest	Mult. R^2 = 0,974	Korr. R^2 = 0,973		$F_{2,\ 77}$ = 1416	p ‹ 0,0001

Tabelle 2: Resultate des Modells in (4) ohne die Interaktion FehlerDeut:Schulklasse

Drei wichtige und miteinander verbundene Punkte sind anzumerken. Erstens, das lineare Modell in (4) – das ohne die Interaktion – beschreibt die Daten sehr viel schlechter (vgl. die beiden fettgedruckten Zahlen in der Spalte „Quadratsumme“): die unbeschriebene Residualvarianz in Tabelle 2, i. e. die Variabilität in FehlerEngl, die das Modell in (4) nicht beschreiben kann, ist viel höher als die unbeschriebene Residualvarianz in Tabelle 1, i. e. die Variabilität in FehlerEngl, die das Modell in (3) nicht beschreiben kann. In der Tat würde ein entsprechender Test zeigen, dass das lineare Modell in (4) nicht nur schlechter, sondern sogar signifikant schlechter ist als das Modell in (3): $p < 10^{-10}$.

Zweitens, die Koeffizienten unterscheiden sich in den beiden Modellen, besonders die für Schulklasse (vgl. die fettgedruckten Zahlen in der Spalte „Koeffizient“). Um einen beliebigen Beispielwert zu betrachten: aus den Roh-

daten geht z. B. hervor, dass der Mittelwert von FehlerDeut ca. 17 Fehler und der Mittelwert von FehlerEngl für Schüler aus Klasse A, die 17 Fehler im Deutschdiktat machten, 20 beträgt. Das Modell in (3) sagt für FehlerDeut = 17 in Klasse A 21,736 Fehler voraus und ist damit 8,68 % zu hoch.[3] Das Modell in (4) dagegen sagt für FehlerDeut = 17 in Klasse A 23,845 Fehler voraus und liegt damit im Mittel um 19,23 % zu hoch.[4] Das heißt, die Schätzung, die auf dem Modell in (4) basiert, liegt mehr als 2,2 Mal mehr falsch, was ausschließlich darauf basiert, dass dieses Modell die Interaktion nicht berücksichtigt. Vergleicht man alle Schätzungen beider Modelle miteinander, so zeigt sich, dass das Modell in (4) mit seinen Schätzungen im Mittel mehr als ein Drittel mehr danebenliegt als das Modell in (3).

Drittens, die Signifikanzwerte für die Prädiktoren sind unterschiedlich. In diesem Fall erklärt das Modell in (4) den Prädiktor Schulklasse für einen signifikanten Haupteffekt, was bedeutet, dass dieses Modell fälschlicherweise suggeriert, dass der Unterschied der beiden Klassen ein Mittelwertunterschied ist: die Schüler in Klasse A machen im Mittel 8,7 weniger Fehler. Das Modell in (3) erkennt dagegen korrekt, dass Schulklasse nur in der Interaktion signifikant ist, was bedeutet, dass dieses Modell korrekt suggeriert, dass der Unterschied der beiden Klassen ein Steigungsunterschied ist: mit steigender Fehlerzahl in Deutsch machen die Schüler in Klasse A weniger Fehler in Englisch (vgl. auch nächster Abschnitt).

2.2.2 Der fehlende Signifikanztest von Interaktionen

Neben den im letzten Abschnitt genannten schwerwiegenden Nachteilen gibt es noch eine zweite unerwünschte Konsequenz, die aus der Vernachlässigung von Interaktionen resultieren kann. Zur Veranschaulichung dieses Punktes betrachten wir die Diktatdaten, die in Abbildung 1 dargestellt sind: Auf der *x*-Achse sind die Fehlerzahlen im Deutschdiktat, auf der *y*-Achse die Fehlerzahlen im Englischdiktat eingetragen, die Buchstaben „A“ und „B“ symbolisieren Schüler aus den entsprechenden Klassen und die Korrelation in jeder Klasse ist mit einer Regressionslinie (und ihrem Konfidenzintervall) dargestellt.

3 Achsenabschnitt + FehlerDeut + Schulklasse:A + Interaktion = vorhergesagter Wert 2,82208 + 17 * 1,64459 + −0,28249 + 17 * −0,51538 = 21,73616

4 Achsenabschnitt + FehlerDeut + Schulklasse:A = vorhergesagter Wert 2,74795 + 17 * 1,75395 + −8,72058 = 23,84452

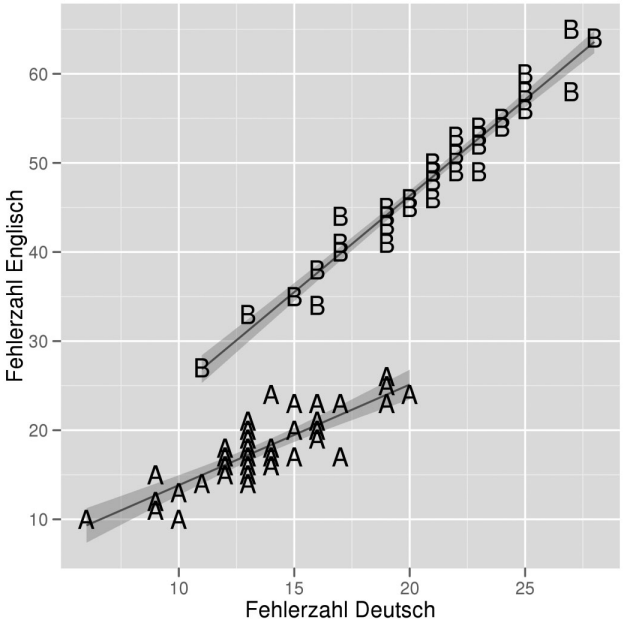

Abbildung 1: Die Korrelation zwischen der Fehlerzahl in Deutsch- und Englischdiktaten

Gelegentlich werden solche Daten analysiert, indem je ein lineares Modell zwischen FehlerEngl und FehlerDeut für jede Schulklasse separat gerechnet wie in (5) und (6) gezeigt:

(5) $\text{FehlerEngl}_{\text{Klasse A}} \sim \text{FehlerDeut}_{\text{Klasse A}}$

(6) $\text{FehlerEngl}_{\text{Klasse B}} \sim \text{FehlerDeut}_{\text{Klasse B}}$

Die fehlerhafte Logik dahinter ist vermutlich, dass man damit die in Abschnitt 2.2.1 genannten Probleme zu umgehen glaubt, denn man ‚berücksichtigt' ja die Interaktion FehlerDeut:Schulklasse indem man explizit zulässt, dass die Korrelation zwischen FehlerEngl und FehlerDeut in jeder Schulklasse unterschiedlich sein kann, was Abbildung 1 ja auch zeigt und was in den entsprechenden statistischen Resultaten in Tabelle 3 ebenfalls deutlich wird.

	Koeffizient/Steigung	***p***
Schulklasse: A	1,13	‹ 0,0001
Schulklasse: B	2,16	‹ 0,0001

Tabelle 3: Resultate der linearen Modelle in (5) und (6) (d. h., separat für jede Schulklasse)

Ganz offensichtlich ist die Korrelation zwischen den Fehlerzahlen in Klasse B stärker als in Klasse A: für jeden weiteren Fehler, den ein Schüler in Klasse A in Deutsch macht, macht er im Schnitt 1,13 Fehler mehr im Englischen, aber für jeden weiteren Fehler, den ein Schüler in Klasse B in Deutsch macht, macht er im Schnitt 2,16 Fehler mehr im Englischen. Nichtsdestotrotz ist diese Analyse streng genommen aus mindestens zwei Gründen problematisch. Erstens läuft diese Analyse Gefahr, aufgrund multipler Tests des gleichen Datensatzes signifikante Resultate zu liefern, die nicht wirklich signifikant sind. Der zweite Grund ist jedoch der hier wichtigere. Die vorliegende Form der Auswertung erlaubt es dem Forscher nicht zu klären, ob der in Abbildung 1 und Tabelle 3 dargestellte Unterschied in den Koeffizienten signifikant ist, denn die Interaktion Fehler-Deut:Schulklasse taucht in den Modellen in (5) und (6) ja nicht auf, was zur Folge hat, dass man keinen Signifikanztest und *p*-Wert für diese Interaktion bekommt. Nur das Modell in (3)/Tabelle 1 beinhaltet diese Interaktion und damit auch ihren *p*-Wert: Um den Koeffizienten für Klasse A zu erhalten, zieht man von der Steigung in Tabelle 1 (1,64459) den Wert der Interaktion ab (–0,51538) und erhält den Koeffizenten Tabelle 3 für Klasse A (ca. 1,13) – und der *p*-Wert für diese Interaktion in Tabelle 1 zeigt, dass dieser Unterschied bereits signifikant ist.

All dies führt zur generellen wichtigen Erkenntnis, dass Studien, in denen die Daten in Teile aufgesplittet werden, für die dann separate Modelle gerechnet werden, nicht klären können, ob die Resultate der Teile sich signifikant voneinander unterscheiden … Bedauerlicherweise ist dies wieder eine Art Analyseform, die besonders in der Soziolinguistik nicht selten ist: Wenn z. B. separate VARBRUL-Analysen für verschiedene Zeitperioden durchgeführt werden, ist es mit diesem Design nur schwer möglich zu testen, ob sich Zeitperioden überhaupt signifikant voneinander unterscheiden (vgl. z. B. Jankowski 2004).

Zusammenfassend lässt sich sagen, dass das Nichteinbeziehen von Interaktionen also nicht nur eine Unterlassungssünde darstellt, wegen der man vielleicht Interaktionen nicht findet bzw. nicht auf Signifikanz prüfen kann – der potentielle Schaden ist viel höher, weil das Nichteinbeziehen von Interaktionen eben auch die Resultate selbst derjenigen Prädiktoren beeinträchtigen kann, die man einbezogen hat. Vor diesem Hintergrund sollte deutlich werden, dass die Vernachlässigung von Interaktionen eine Kardinalsünde darstellt, und dies gilt für alle Modellierungsverfahren. Außerdem sollte auch sehr deutlich geworden sein, dass es gerade die Einbeziehung von mehreren Prädiktoren und Interaktionen ist, die es quantitativen Studien in der Sprachwissenschaft ermöglicht, gänzlich neue Fragen zu adressieren sowie komplexe (Interaktions)Strukturen in Datensätzen zu finden, deren Größe/Komplexität eine eher informelle oder intuitive Betrachtung ausschließen; siehe auch Harrell (2001: Abschnitt 1.4) zu Modellformulierung.

2.3 Ziele von Modellierung und Arten des Modellierungsprozesses

Obgleich der letzte Abschnitt vergleichsweise lang war, hat er hoffentlich verdeutlicht, wie wichtig es ist, gleich zu Beginn ein adäquates Modell zu formulieren: wenn die Prädiktoren nicht richtig skaliert sind und Interaktionen nicht berücksichtigt werden, ist jede Modellierung zum Scheitern verurteilt. In diesem Abschnitt soll noch einmal genauer auf die Ziele des Modellierungsprozesses eingegangen werden.

Zu Beginn des Abschnitts 2.1 wurde bereits erwähnt, dass die typischen Ziele eines Modellierungsprozesses Beschreiben und Vorhersagen sind. Hier soll kurz verdeutlicht werden, dass diese beiden Ziele wiederum zwei Hauptanwendungen haben. Wie oben erwähnt, wird Modellierung meist *hypothesenprüfend* eingesetzt, wie im obigen Beispiel, wenn es darum geht zu testen, ob Prädiktoren einen signifikanten Effekt auf eine oder mehrere abhängige Variablen haben. In solchen Fällen kommen typischerweise verschiedene Arten von Regressionen oder ähnlichen Methoden zum Einsatz, und die exakte Auswahl einer Methode hängt ab (i) von der zu testenden Art des angenommenen Zusammenhangs sowie (ii) der Art der abhängigen Variablen.

Mit Bezug auf die Art des Zusammenhangs lassen sich nonlineare und lineare Methoden unterscheiden, wobei die letzteren noch unterteilt werden können in Modelle, die in Regressionsgeraden resultieren (vgl. Abbildung 1 sowie Abbildung 2, links), und Modelle, die in Regressionskurven resultieren (vgl. Abbildung 2, rechts); in der Sprachwissenschaft sind nonlineare Modelle, die sich nicht

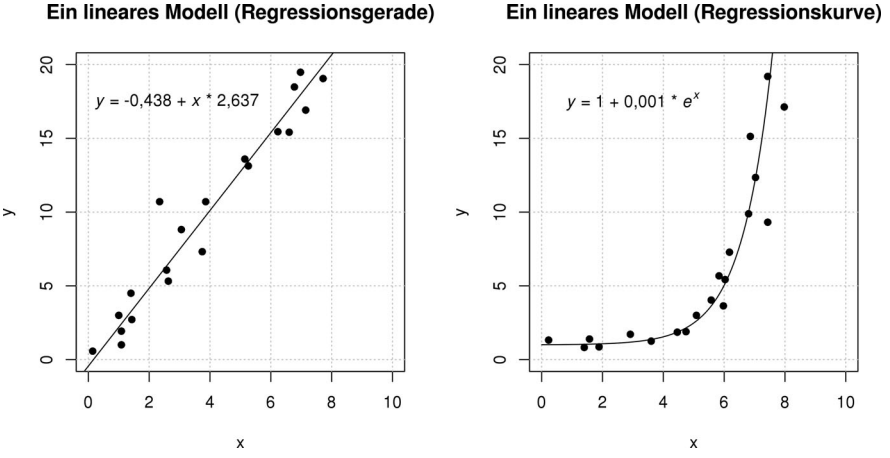

Abbildung 2: Zwei lineare Modelle

direkt in lineare Modelle überführen lassen, eher selten, aber beide in Abbildung 2 dargestellten lineare Modelle sind weit verbreitet.

Die Notwendigkeit einer gründlichen visuellen Inspektion der zu modellierenden Daten kann nicht genug betont werden – ein Forscher, der Abbildung 1 generiert hätte, würde nie auf die Idee kommen, die Interaktion FehlerDeut:Schulklasse nicht zu berücksichtigen. Dies soll durch zwei weitere Beispiele verdeutlicht werden. Das erste behandelt die Entwicklung der Häufigkeit der *keep* V-*ing* Konstruktion im Englischen (z. B. *They keep on using too few graphs in linguistics*); die Daten und Diskussion basieren auf Hilpert (2011) und Gries (2011). Das linke Panel in Abbildung 3 zeigt, dass die Häufigkeit der Konstruktion zunimmt. Würde man eine lineare Regression für das Modell in (7) rechnen, erhielte man die angegebene sehr signifikante Korrelation, die einen signifikanten und kontinuierlichen Anstieg suggeriert.

(7) Häufigkeit ~ Jahr
(8) Häufigkeit ~ Jahr + BisEinschliesslich1996 + Jahr: BisEinschliesslich1996

Der zentrale Aspekt ist hier jedoch, dass, auch wenn lineare Regressionsgeraden gemäß Ockhams Rasiermesser die Voreinstellung bei Modellierungen sind, das nicht heißt, dass sie notwendigerweise die beste Modellform darstellen. Im linken Panel wird neben der Regressionsgerade auch ein sog. *Smoother* dargestellt, der einen kurvilinearen Zusammenhang suggeriert. Im rechten Panel wird daher gezeigt, dass eine bessere Analyse – *regression with breakpoints* – zeigt, dass die Daten in zwei Zeitperioden fallen: ein steiler Häufigkeitsanstieg bis 1996 ($r = 0{,}91$), gefolgt von einem Plateau ab 1997 ($r = 0{,}1$). Ein entsprechender Vergleich der

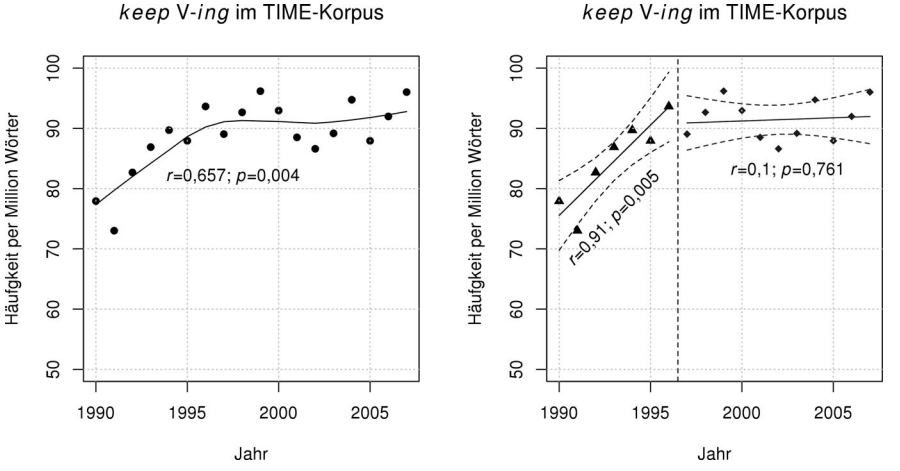

Abbildung 3: Die Verwendung von *keep* V-*ing* im TIME-Korpus: zwei Modelle

Modelle in (7) und (8) zeigt, dass, obwohl das Modell in (8) mehr Prädiktoren beinhaltet, es trotzdem die Daten in der Tat soviel besser beschreibt, dass es dem in (7) vorzuziehen ist, wie in Tabelle 4 zusammengefasst wird.

	mult. R^2/ korr. mult. R^2	*AIC* (s. u.)	ANOVA-Vergleich
Modell in (7)	0,4315 / 0,396	57,64	$F_{2,\ 24} = 7{,}864$
Modell in (8)	0,7323 / 0,6749	48,08	$p = 0{,}005$

Tabelle 4: Vergleich der Modelle in (7) und (8)

Ein ähnliches Beispiel ist in Tuttle & Lovicks (2007) Regression zu finden, die in Abbildung 4 dargestellt ist. Diese Daten basieren auf der Analyse von zwei gesprochenen narrativen Texten; auf der *x*-Achse ist die Länge von Pausen eingetragen, auf der *y*-Achse die Länge der Reime (in ms). Auch wenn es möglich ist, eine lineare Regressionsgerade durch diese Punktewolke zu zwingen, heißt das nicht, dass ein anderer Ansatz nicht adäquater wäre, zumal Abbildung 4 überdeutlich zeigt, dass die Verteilungsannahmen linearer Modelle hier verletzt werden. Beispielsweise sind die Residuen höchstwahrscheinlich nicht normalverteilt, und es erscheint sogar möglich, dass der Zusammenhang der beiden Variablen eher kurvilinear als linear ist.

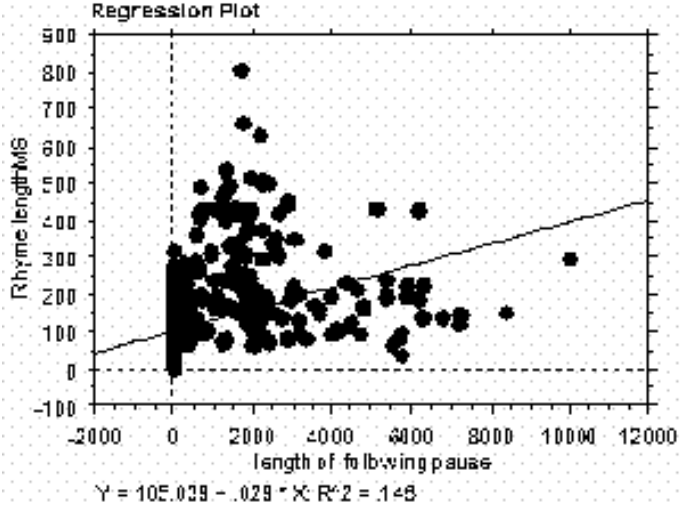

Abbildung 4: Der Zusammenhang von Reimlänge und Pausenlänge in Tuttle & Lovick (2007: 213)

Hinsichtlich der abhängigen Variablen lassen sich verschiedene lineare Modelle unterscheiden. Der konzeptuell/mathematisch einfachste ‚Standardfall' sind die oben exemplifizierten (multiplen) linearen Regressionen für abhängige Variablen, die intervallskaliert sind. Derartige lineare Modelle führen zur Berechnung einer Geraden von vorhergesagten Werten, was auch bedeutet, dass sie ab einem bestimmten Punkt unrealistische Werte vorhersagen: Jede Regressionsgerade, die nicht parallel zur *x*-Achse verläuft, sagt Werte vorher zwischen $-\infty$ und $+\infty$ und wird entsprechend irgendwann negative Werte vorhersagen, obwohl z. B. Auftretenswahrscheinlichkeiten und Häufigkeiten nicht negativ sein können. Daher werden andere Arten von abhängigen Variablen mit anderen Arten von Regressionen untersucht:

- binäre logistische Regressionen für binäre abhängige Variablen, die Auftretenswahrscheinlichkeiten der abhängigen Variablen vorhersagen;
- multinomiale logistische Regressionen für kategoriale abhängige Variablen, die Auftretenswahrscheinlichkeiten der abhängige Variablen vorhersagen;
- Poissonregressionen für abhängige Variablen, die Häufigkeiten darstellen und selbige vorhersagen.[5]

Um den Regressionsansatz auf solche abhängige Variablen anwenden zu können, ohne unzulässige Werte vorherzusagen, werden diese abhängigen Variablen mit einer sog. *Verbindungsfunktion* vom vorhergesagten Wertebereich zwischen $-\infty$ und $+\infty$ in den angemessenen Wertebereich transformiert. Für die binäre logistische Regression z. B. werden die von der Regressionsgleichung zwischen $-\infty$ und $+\infty$ liegenden vorhergesagten Werte mit der inversen logit-Funktion in (9) in den Wertebereich zwischen 0 und 1 umgewandelt; für die Poissonregression werden die zwischen $-\infty$ und $+\infty$ liegenden vorhergesagten Werte mit einer Exponential-

5 Bei den o. g. Regressionsmodellen handelt es sich um ‚Standardansätze' – aus Platzgründen kann auf Alternativen nur am Rande eingegangen werden. Eine Alternative zu linearen Regressionen sind die sog. *Generalisierten additiven Modelle*, die sich besonders gut eignen, kurvilineare Zusammenhänge zu entdecken (vgl. Hastie & Tibshirani 1990). Eine interessante Alternative besonders zu logistischen Regressionen sind *Klassifikations- und Regressionsbäume sowie Random Forests*. Diese sind oft flexibler, da sie weniger Annahmen hinsichtlich der Verteilung der Daten machen, aber erstere können Schwierigkeiten bei der Identifikation von Interaktionen haben. Eine weitere Alternative sind lineare oder quadratische *Diskriminanzanalysen*, aber mittlerweile sind viele Verfahren verfügbar; neuronale Netzwerke und andere Lernalgorithmen, *support vector machines*, clusteranalytische Verfahren, etc. Im Zweifelsfall sollten Resultate verschiedener Methoden auf Konvergenz geprüft werden. Überblicke bieten Faraway (2006) und Hastie, Tibshirani & Friedmann (2009); Beispiele für Anwendungen der o. g. Methoden sind Baayen (2010b, 2011), Teich & Frankhauser (2010) sowie Jarvis (2011) für einen Vergleich verschiedener Klassifikationsalgorithmen.

funktion in (10) in den Wertebereich ≥ 0 umgewandelt, etc., wie in Abbildung 5 verdeutlicht wird.

(9) inverse logit von $x = \frac{1}{1e^x}$

(10) Exponentialfunktion von $x = e^X$

Auf der *x*-Achse sind die von der Regressionsgleichung vorhergesagten Werte eingetragen, auf der *y*-Achse die gemäß (9) und (10) transformierten Werte. Das linke Panel z. B. zeigt deutlich, dass, selbst wenn die Regressionsgleichung einer binären logistischen Regression negative Werte vorhersagt (z. B. wenn $x = -2$), die Regressionsvorhersagen durch die inversen logit-Funktion in den Wertebereich zwischen 0 und 1 gezwungen werden, der für Auftretenswahrscheinlichkeiten erforderlich ist. Auf diese Weise kann eine Familie von Modellen auf sehr unterschiedliche Datenarten angewendet werden.

Während diese Arten von hypothesenprüfenden Verfahren die Standardanwendung von Modellierung sind, so wird Modellierung auch *explorativ* verwendet, d. h. in Situationen, wo noch keine präzise formulierten Null- und Alternativhypothesen vorliegen, sondern wo es darum geht, mit Hilfe von Modellen Strukturen in Daten zu finden. In derartigen Kontexten wird vielleicht noch keine abhängige Variable berücksichtigt, sondern es kann z. B. darum gehen, Strukturen und Interdependenzen zwischen vielen unabhängigen Variablen zu finden; dabei kommen oft Verfahren wie Clusteranalysen, Faktorenanalysen,

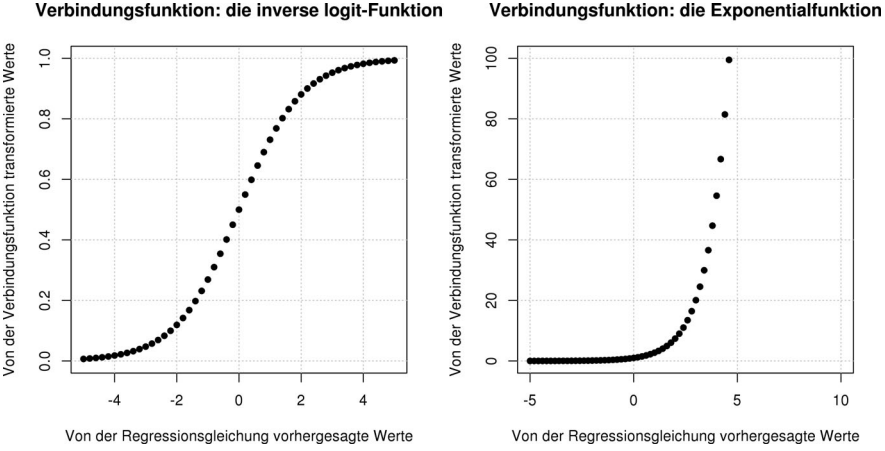

Abbildung 5: Die Transformation vorhergesagter Werte (auf der *x*-Achse) durch zwei Verbindungsfunktionen (repräsentiert auf der *y*-Achse)

Korrespondenzanalysen, multidimensionale Skalierung etc. zum Einsatz, deren Resultate dann in hypothesenprüfende Verfahren wie z. B. Regressionen eingehen.

3 Modellselektion und -interpretation

Nachdem eine erste Modellformulierung unter Berücksichtigung der obigen Aspekte erfolgt ist, folgt oft ein schrittweiser Modellselektionsprozess, in dem das erste Modell wie oben erwähnt schrittweise angepasst wird, bis es die Daten gut genug beschreibt und nur Prädiktoren enthält, deren Verwendung gegenüber Ockhams Rasiermesser gerechtfertigt werden kann. Wenn man solch ein endgültiges Modell erhalten hat, ist der dann folgende Schritt die Interpretation der Ergebnisse. Diese Selektion des besten Modells für die Daten und seine Interpretation werden in diesem Abschnitt behandelt.

3.1 Modellselektion

Der Prozess der Modellselektion wird hauptsächlich durch zwei Parameter bestimmt, der *Richtung*, in der Modellselektion betrieben wird, und dem *Kriterium*, das verwendet wird, um Prädiktoren in ein Modell aufzunehmen oder aus einem Modell auszuschließen.

3.1.1 Die Richtung der Modellselektion

Drei verschiedene Ansätze können unterschieden werden:

- *vorwärts*: in diesem Fall beginnt man mit einem kleinen Modell und testet sukzessive, ob man weitere Prädiktoren hinzufügen kann; der Selektionsprozess bricht ab entweder, wenn das Hinzufügen von Prädiktoren das Modell nicht mehr deutlich verbessert, oder wenn alle verfügbaren Prädiktoren bereits im Modell enthalten sind. Der Extremfall des ersten, kleinsten Modells ist, dass man für alle Daten nur deren Mittelwert vorhersagt. Die Addition von Prädiktoren folgt dabei oft der Regel, dass man sich vom niedrigsten Grad an Interaktivität (Haupteffekte, i. e. Variablen in Isolation) in Richtung höherer Grade an Interaktivität bewegt.
- *rückwärts*: in diesem Fall beginnt man mit einem umfangreichen Modell und testet sukzessive, ob man Prädiktoren weglassen kann; der Selektionsprozess bricht ab entweder, wenn kein Prädiktor mehr weggelassen werden kann,

ohne das Modell deutlich zu verschlechtern, oder wenn keine Prädiktoren mehr im Modell enthalten sind. Der Extremfall des ersten, maximalen Modells ist, dass man alle Haupteffekte und alle ihre Interaktionen einschließt. Die Subtraktion von Prädiktoren folgt dabei der Regel, dass man sich vom höchsten Grad an Interaktivität in Richtung niedrigerer Grade an Interaktivität bewegt und keinen Prädiktor subtrahieren darf, der an einer höheren deutlichen Interaktion beteiligt ist.
- *bidirektional*: in diesem Fall beginnt man mit einem Startmodell und gestattet dem Algorithmus, Prädiktoren hinzuzufügen oder abzuziehen, je nachdem, was das Modell am deutlichsten verbessert.

Es scheint keinen Konsens darüber zu geben, welcher Ansatz der beste ist; mir persönlich scheint der zweite Ansatz am meisten verbreitet zu sein. Es soll jedoch nicht unterschlagen werden, dass es auch gute Argumente dafür gibt, überhaupt keine derartige Modellselektion durchzuführen; der interessierte Leser sei auf Harrell (2001: Abschnitt 4.3) oder Faraway (2005: Abschnitt 8.2) verwiesen.

3.1.2 Das Kriterium für die Modellselektion

Die obigen Ausführungen waren beabsichtigterweise vage, als es darum ging, wann Prädiktoren zu einem Modell hinzugefügt bzw. aus einem Modell entfernt werden können – der oben verwendete Ausdruck war, dass eine Prädiktor in einem Modell einen „deutlichen" Beitrag leisten muss. Diese Vagheit rührt daher, dass es verschiedene Kriterien gibt, mit denen ‚Deutlichkeit' gemessen werden kann; die zwei häufigsten sind die folgenden:
- *Signifikanzwerte/*p*-Werte*: gemäß diesem Ansatz wird ein Prädiktor in ein Modell aufgenommen, wenn seine Aufnahme das Modell signifikant verbessert (wobei es dem Forscher obliegt, welches Signifikanzniveau gewählt wird, der Standard von 5 % oder ein anderer Wert), und ein Prädiktor wird aus einem Modell entfernt, wenn seine Entfernung zu keiner signifikanten Verschlechterung des Modells führt.
- *AIC (Akaike Information Criterion)*: *AIC* ist ein Maß, das die Güte eines Modells in Beziehung setzt zu seiner Menge an Parametern/Prädiktoren; es ist damit ebenfalls eine Operationalisierung von Ockhams Rasiermesser: wenn zwei Modelle Daten gleich gut beschreiben, dann wird *AIC* für das Modell mit weniger Prädiktoren niedriger sein und dieses Modell bevorzugen. Gemäß diesem Ansatz wird ein Prädiktor in ein Modell aufgenommen oder aus einem Modell entfernt, wenn dies *AIC* verringert.

Andere Maße werden z. B. bei Harrell (2001: Abschnitt 9.8) diskutiert; vgl. auch Faraway (2005: Kap. 8).

3.2 Modellinterpretation

Am Ende einer Modellselektion steht ein Modell, das einen oder mehrere Prädiktoren – Haupteffekte und oft Interaktionen – beinhaltet und für jeden Prädiktor Koeffizienten, Prüfstatistiken, Freiheitsgrade und *p*-Werte enthält; vgl. Tabelle 1. Dieser Output wird normalerweise anhand von drei Fragen interpretiert:

- gibt es einen signifikanten Zusammenhang zwischen dem/den Prädiktor(en) und der/den abhängigen Variable(n)?
- was ist die Art/Natur des (signifikanten) Zusammenhangs?
- wie gut erklärt das Modell die beobachtete(n) abhängige(n) Variable(n)?

3.2.1 Gibt es einen signifikanten Zusammenhang?

Diese Frage ist normalerweise einfach zu beantworten, indem man den *p*-Wert des Modells im Allgemeinen und die *p*-Werte der interessierenden Prädiktoren betrachtet. Letztere werden auf der Basis der *t*-Werte berechnet, welche wiederum die Quotienten der Koeffizienten und ihrer Standardfehler sind, und sie quantifizieren die Unsicherheit, die mit Koeffizienten assoziiert ist: je größer der Standardfehler ist in Relation zum Koeffizienten, desto kleiner wird der *t*-Wert und desto größer wird der *p*-Wert. In den meisten Fällen, besonders wenn ein durchdachtes hypothesenprüfendes Design vorliegt, wird es mindestens einen signifikanten Befund geben. Im Beispielmodell aus (3) ist das gesamte Modell hoch signifikant ($F_{3,\ 76} = 1661$; $p < 0{,}001$), und die entsprechenden *p*-Werte der Koeffizienten sind in Tabelle 1 zusammengefasst.

3.2.2 Was ist die Natur des (signifikanten) Zusammenhangs?

Der Natur des Zusammenhangs zwischen dem/den Prädiktor(en) und der abhängigen Variable(n) kann sich auf mehrere Arten genähert werden. Die erste basiert auf der Inspektion des Korrelationswertes für das gesamte Modell, welche in Tabelle 1 in der letzten Zeile dargestellt sind: je näher die R^2-Werte an 1 sind, desto stärker ist die Korrelation zwischen allen Prädiktoren und der abhängigen Variablen.

Die zweite betrifft die Effekte der einzelnen Prädiktoren und basiert auf den (signifikanten) Koeffizienten im Modell (i. e., den Werten in Tabelle 1). Die dort gelisteten Werte haben folgende Bedeutungen:

- der Achsenabschnitt ist der ungewichtete Mittelwert der zwei vorhergesagten Fehlerzahlen im Englischdiktat von Schülern in beiden Klassen, die keinen Fehler im Deutschdiktat machten. Mit anderen Worten, wenn Schüler keine Fehler im Deutschdiktat machten, sagt die Regressionsgleichung vorher, dass sie im Mittel beider Klassen ca. 3 Fehler (2.82) im Englischdiktat machen werden;
- der Achsenabschnitt plus der erste Koeffizient (eine Steigung) ist der ungewichtete Mittelwert der zwei vorhergesagten Fehlerzahlen im Englischdiktat von Schülern in beiden Klassen, die einen Fehler im Deutschdiktat machten. Mit anderen Worten, wenn Schüler einen Fehler im Deutschdiktat machten, sagt die Regressionsgleichung vorher, dass sie im Mittel beider Klassen ca. $2.82 + 1.64 \approx 4.5$ Fehler im Englischdiktat machen werden;
- der Achsenabschnitt plus der zweite Koeffizient (ein Mittelwertsunterschied) ist die vorhergesagte Fehlerzahl eines Schülers aus Klasse A, der keinen Fehler im Deutschdiktat machte: $2.82 + -0.28 \approx 2.54$;
- der Achsenabschnitt plus alle drei Koeffizienten ist die vorhergesagte Fehlerzahl eines Schülers aus Klasse A, der einen Fehler im Deutschdiktat machte: $2.82 + 1.64 + -0.28 + -0.52 \approx 3.67$.

Diese Ausführungen zeigen zwei Dinge sehr deutlich: Erstens, die Ausgabe eines Regressionsmodells beinhaltet eine Menge an Informationen, die sehr genau Auskunft geben über die Art, wie Prädiktoren die abhängige Variable beeinflussen, und die ggf. sehr genaue Vorhersagen ermöglichen. Zweitens, obwohl es sich bei diesem Beispiel nur um ein sehr einfaches Modell handelt – ein einfacheres Modell mit Interaktionen ist kaum möglich –, ist die textuelle Zusammenfassung schon schwer verdaulich: Modellinterpretation auf der Basis der Koeffizienten alleine ist besonders für statistisch untrainierte Leser intuitiv kaum möglich. In den allermeisten Fällen ist es daher unabdingbar, die Resultate des Modells auf eine zweite Art zu verdeutlichen, nämlich sie in grafischer Form zusammenzufassen – verglichen mit der obigen Zusammenfassung des Modells aus (3) in Tabelle 1 ist Abbildung 1 leicht verständlich.

Für derartige grafische Zusammenfassungen müssen zwei Arten von Entscheidungen getroffen werden. Die erste wird nicht immer explizit gerechtfertigt und bezieht sich darauf, ob die Visualisierung auf der Basis der tatsächlich *beobachteten Daten* (z. B. beobachtete Mittelwerte für eine binäre unabhängige Variable) oder der von signifikanten Prädiktoren *vorhergesagten Werte* (vgl. die vorhergesagten Werte der obigen Modellzusammenfassungen) passieren soll. Der Vorteil der tatsächlich

beobachteten Daten ist, dass sie – intuitiv betrachtet – das sind, was man beschreiben will. Ihr Nachteil ist, dass sie insbesondere in komplexen Datensätzen oft so viel Variabilität aufweisen, dass sie schwieriger zu verstehen sind als Zusammenfassungen auf der Basis der Daten, die von einem Modell vorhergesagt wurden, mit dem sie ja signifikant korrelieren, und daher weniger variabel sind. Zusätzlich muss oft auch noch die am besten verständliche Skalierung der zu visualisierenden Daten gewählt werden. Die Ergebnisse logistischer Regressionen beispielsweise können als Odds, logarithmierte Odds, oder vorhergesagte Wahrscheinlichkeiten visualisiert werden,[6] und jede dieser Formen hat Vor- und Nachteile.

Die zweite Entscheidung ist die für eine bestimmte grafische Darstellungsform. Man kann folgende Faustregeln angeben:

- für kategoriale Variablen stellen oft Punkte- oder Balkendiagramme (ggf. mit *error bars*) Mittelwerte gut dar, da die Koeffizienten kategorialer Variablen Mittelwertsunterschiede repräsentieren;
- für Intervallvariablen sind Streudiagramme mit einer Regressionsgeraden/kurven (ggf. mit Konfidenzintervallen) oft nützlich, da die Koeffizienten von Intervallvariablen Steigungen repräsentieren;
- für Interaktionen werden diese kombiniert, so dass unterschiedliche Kombinationen von Balken oder verschiedene Regressionsgeraden/kurven die Kombinationen der Interaktion darstellen (s. o. Abbildung 1).

Die folgenden Quellen diskutieren viele dieser Aspekte: Unwin, Theus & Hofmann (2006), Cook & Swayne (2007), Sarkar (2008), Wickham (2009), Keen (2010), Murrell (2011).

3.2.3 Wie gut erklärt das Modell die Daten?

Einer der letzten Schritte ist zu beurteilen, wie gut das Modell die Daten erklärt. In Abhängigkeit von der Datenlage und dem verwendeten Modellalgorithmus kann man hauptsächlich zwei Quantifizierungsarten unterscheiden:

- wenn die abhängige Variable intervallskaliert ist (und besonders in entsprechenden regressionsanalytischen Verfahren), kann die Modellgüte durch den Prozentsatz an Varianz, den das Modell beschreiben kann, wiedergegeben werden. Im obigen Modell (3) ist die Varianzaufklärung außergewöhnlich hoch: das sog. *Bestimmtheitsmaß*, multiples R^2, beträgt 0,985, i. e. 98,5 %.

6 Die Odds eines Ereignisses *E* ist der Quotient der Wahrscheinlichkeit, dass *E* eintritt, geteilt durch die Wahrscheinlichkeit, dass *E* eintritt: p_E / p_{nichtE}.

In den allermeisten Fällen wird jedoch nicht einfach nur multiples R^2angegeben, sondern ein R^2-Wert, der mit Hinblick auf die Menge an Prädiktoren nach unten korrigiert wurde (vergleichbar mit dem *AIC*); in diesem Fall enthält das Modell nur wenige Prädiktoren und das korrigierte R^2ist daher fast genauso hoch: 0,9844.

– wenn die abhängige Variable binär oder kategorial ist – das Modell also ein Klassifikationsproblem zu lösen hatte –, dann kann die Modellgüte u. a. durch den Prozentsatz an korrekt klassifizierten Ausprägungen der abhängigen Variablen wiedergegeben werden. Hierzu wird für jeden Datenpunkt die Wahrscheinlichkeit aller möglichen Klassifikationen ermittelt und dann die Kategorie mit der höchsten Wahrscheinlichkeit gewählt und mit der tatsächlich beobachteten Variablenausprägung verglichen. Eine etwas genauere Alternative hierzu ist die Berechnung einer Korrelation zwischen den beobachteten Datenpunkten und den vorhergesagten Wahrscheinlichkeiten der entsprechenden Variablenausprägung (vgl. der *C*-Wert oder D_{xy} in der logistischen Regression). Eine m. E. unterschätzte letzte hier zu nennende Alternative sind sog. PRE-Maße, für *proportional reduction of error*. PRE-Maße geben an, um wie viel Prozent sich die Genauigkeit der Klassifikation der abhängigen Variable verbessert hat, wenn man, statt einfach immer die häufigste Ausprägung zu raten, die unabhängigen Variablen berücksichtigt; PRE-Maße sind also intuitiv sehr leicht verständlich.

4 Validierung und Replizierbarkeit

Die letzten beiden Schritte, die der Modellselektion und -interpretation folgen, beziehen sich auf die Validierung des endgültigen Modells und seine Replizierbarkeit.

4.1 Validierung

Während *Validität* die Eigenschaft bezeichnet, dass die erhobenen Variablen tatsächlich das messen, was sie messen sollen, versteht man unter *Validierung* im Kontext von Modellierung die Prüfung, ob ein Modell, das auf der Basis eines bestimmten Datensatzes entwickelt wurde und dort eine bestimmte Vorhersage-/Klassifikationsgenauigkeit erreichte (siehe den vorigen Abschnitt), auch für andere aber vergleichbare Datensätze verwendet werden kann. Man testet sozusagen gegen den schlimmstmöglichen Fall, dass das Modell nur auf Besonderheiten im ersten Originaldatensatz beruhte, für andere Daten aber unbrauchbar ist.

Eine nützliche und weit verbreitete Art der Validierung ist die sog. *Kreuzvalidierung*. Bei diesem Verfahren werden die Daten in zwei Teile aufgeteilt, einen (größeren) Teil, der zur Modellformulierung und -selektion verwendet wird, und einen (kleineren) Teil, auf den das Modell angewendet wird, um dort die abhängige Variable vorherzusagen. Der Vorteil dieses Ansatzes ist, dass das Modell nicht auf dieselben Daten angewendet wurde, von denen es abgeleitet wurde. Um jedoch die Validierung nicht zu abhängig von nur einer einzigen Aufteilung der Daten zu machen, wird dieser Prozess mehrfach durchgeführt; die häufigste und oft effizienteste Anwendungsform teilt die Daten in 10 Teile auf, und in 10 Schritten wird jeweils ein Zehntel der Daten auf der Basis der neun anderen Zehntel modelliert (vgl. Molinaro et al. 2005: 3306). Ein kleinteiligerer Ansatz ist als *leave-one-out* Methode bekannt und teilt einen Datensatz mit n Datenpunkten n Mal in zwei verschiedene Stichproben auf: die erste Validierung versucht Datenpunkt 1 auf der Basis eines Modells für Datenpunkte 2 bis n vorherzusagen, die zweite Validierung versucht Datenpunkt 2 auf der Basis eines Modells für Datenpunkte 1 und 3 bis n vorherzusagen etc.

Andere Arten der Validierung beruhen auf *Sampling- oder Permutationsmethoden*. Für Samplingmethoden wie den *Bootstrap* kann man z. B. s (Hunderte, Tausende, ...) unterschiedlich große Stichproben mit Zurücklegen aus dem eigentlichen Datensatz generieren, für jede dieser s Stichproben das beste Modell ermitteln und dann die s Modelle aggregieren; das auf diese Weise erhaltene aggregierte Modell ist wesentlich weniger von idiosynkratischen Datenpunkten oder Mustern beeinflusst; eine derartige Implementierung für clusteranalytische Verfahren z. B. ist in Suzuki & Shimodaira (2011) zu finden. Eine Permutationsmethode zur Modellprüfung dagegen wurde in Gries (2006) eingesetzt. Die Resultate einer logistischen Regression auf Korpusdaten wurden verglichen mit den Resultaten aller 2^{13}-1 = 8191 Korpusteile, die man generieren kann, wenn man 13 Korpusteile auf alle möglichen Arten kombiniert; auf diese Weise erhält man ein genaues Bild der Variabilität des Phänomens, die die (hier, Korpus-) Daten enthalten. Siehe Fox (1997: Kap. 16), Harrell (2011: Abschnitte 5.1–5.2) oder Fox (2008: Kap. 21–22) für Diskussion.

4.2 Replizierbarkeit

Ein weiterer Gütetest neben der o. g. Validierung ist die Replikation, um die Ergebnisse für einen neuen Datensatz mit dem ersten zu vergleichen; dies kann entweder durch den gleichen Forscher oder andere Autoren geschehen. In beiden Fällen ist es natürlich essentiell, dass die Vergleichbarkeit der Studien durch so ähnlich wie mögliche Operationalisierungen und Auswertungsverfahren gewähr-

leistet ist. Dies wiederum erfordert, dass derartige methodologische Entscheidungen in Forschungsarbeiten maximal explizit dokumentiert werden, was leider noch nicht immer der Fall ist: selbst, wenn Modellselektion erwähnt wird, bleibt nicht selten unklar, welche Richtung und welches Kriterium (und ggf. welches Signifikanzniveau) verwendet wurde, etc. Es gibt jedoch erste Bestrebungen (vgl. das *Journal of Experimental Linguistics* <http://www.elanguage.net/journals/index.php/jel/index>), derartige Standards umzusetzen.

Was in der Sprachwissenschaft noch sehr selten ist, ist, dass neben der Explikation aller methodologischen Entscheidungen auch die Urdaten selbst für eine einfache oder erweiternde Reanalyse zur Verfügung gestellt werden. Auch dies wird mittlerweile mehr diskutiert, aber legale und andere Hürden sowie Ressourcenprobleme stehen dem noch im Weg, wie z. B. Korpora, die nicht voll verfügbar sind und/oder wachsen oder anderweitig verändert werden, Tonaufnahmen und Transkripte von Sprachgemeinschaften, die nicht ohne deren Zustimmung weitergegeben werden dürfen, etc. Hinzu kommen Probleme, die aus der zunehmenden *publish-or-perish* Kultur in der Akademie erwachsen. So befürchten viele Forscher, die über Jahre in Einzelarbeit Daten gesammelt haben, dass die vollständige Publikation dieser Daten dazu führt, dass selbige sofort wenig karriereförderndnd von anderen großen Forschergruppen ausgeschlachtet werden. In diesem Bereich hat die Sprachwissenschaft noch einiges an Entwicklung vor sich.

5 Herausforderungen

Aufgrund der Vielfalt an Anwendungsbereichen und Methoden ist Modellierung ein Themengebiet, dem mehr durch eine Enzyklopädie als durch einen Überblicksartikel Rechnung getragen werden kann, und die daher notwendige Verkürzung muss viele potentielle Klippen und Risiken daher unerwähnt lassen. In diesem letzten Abschnitt soll besonders für Leser mit Vorkenntnissen jedoch kurz auf einige potentielle Probleme und Herausforderungen eingegangen werden, die in Modellierungsprozessen regelmäßig entstehen. Einige der größten und konzeptuell wichtigsten Herausforderungen wurden in den Abschnitten und 2.3 bereits angesprochen: das Formulieren der (ersten) Modellgleichung und die Fragen, (i) auf welchen Skalen Prädiktoren berücksichtigt werden sollen, und (ii) welche und wie viele Interaktionen berücksichtigt werden sollen. In den folgenden kurzen Abschnitten soll es hingegen um Herausforderungen eher technischer Natur gehen.

5.1 Verteilungsannahmen von, und Korrelationen in, den Daten

Ein Punkt, der bisher übergangen wurde, ist, dass die Signifikanztests in Modellierungsverfahren und die Interpretation von Modellen bestimmte Anforderungen an die Daten stellen. Viele Signifikanztests in regressionsanalytischen Ansätzen erfordern beispielsweise, dass Varianzen in den untersuchten Teilen der Daten homogen sind, dass Residuen normal verteilt sind, dass andere Dispersionsparameter bestimmte Werte aufweisen, etc., und idealerweise sollten alle regressionsanalytischen Studien ihre Daten auf diese Anforderungen hin prüfen. Ein weiterer Punkt, der in vielen Modellierungsprozessen zentral sein kann, ist die sog. *Kollinearität* (oder *Multikollinearität*), das Phänomen, dass Prädiktoren stark miteinander korrelieren. Solche Interkorrelationen können dazu führen, dass die Schätzungen der Regressionskoeffizienten so instabil werden, dass sich sogar ihr Vorzeichen (i. e., die angenommene Richtung ihres Einflusses auf die abhängige Variable) umkehren kann, was sowohl ihre Signifikanztestung als auch Interpretation sehr erschwert. Kollinearität kann auf mehrere Arten ermittelt werden, wie z. B. paarweise Korrelationen zwischen Prädiktoren, einen hohen Varianzinflationsfaktor *VIF* oder, heuristisch, wenn z. B. eine Regression eine signifikante Korrelation ergibt, aber keiner oder kaum einer der Prädiktoren signifikant ist. Als Gegenmaßnahmen kann man hoch interkorrelierte Prädiktoren aus dem Modell entfernen oder mit explorativen Methoden wie einer Faktorenanalyse zusammenfassen; manchmal hilft schon eine Zentrierung oder *z*-Standardisierung der Daten oder die Wahl eines alternativen Regressionsverfahren, welches nicht so sehr von Kollinearität beeinflusst wird; vgl. Harrell (2001: Abschnitte 4.6–4.7) und Faraway (2005: Kap. 5).[7] Die Prüfung dieser Voraussetzungen und Modellcharakteristika (und häufig anderer) sollte ein fester Bestandteil des Modellierungsprozesses sein, da ansonsten das Risiko inkorrekter Resultate zunimmt, ein Punkt, der auch im folgenden Abschnitt aufgenommen wird.

7 Die Werte einer Intervallvariablen werden zentriert, indem man von jedem Variablenwert den Mittelwert der Variablen abzieht. Die Werte einer Intervallvariablen werden *z*-standardisiert, indem man von jedem Variablenwert den Mittelwert der Variablen abzieht und alle diese Differenzen durch die Standardabweichung der Variablen dividiert.

5.2 Spezielle Datenpunkte

Einhergehend mit der Prüfung dieser Voraussetzungen sollten die Daten des Weiteren *post hoc* analysiert werden mit Hinblick auf

- *Ausreißer*, i. e. Datenpunkte, die sich stark von der Punktewolke des ganzen Datensatzes unterscheiden und durch ein Modell schlecht beschrieben oder klassifiziert werden;
- einflussreiche Datenpunkte mit Hebelwirkung (*leverage points*), i. e. Datenpunkte, die einen großen Einfluss auf die Struktur eines Modells haben (weil sie sich stark vom Mittelwert eines Prädiktors unterscheiden).

Zur Identifikation dieser Datenpunkte gibt es eine Vielzahl von Diagnostiken wie Ausreißertests, Hebelwirkungen, Residuen, *Cook's distances* etc. Modelldiagnostik dieser Art ist extrem wichtig, um die Genauigkeit und Interpretierbarkeit der Resultate sicherzustellen, da schon ein einziger Datenpunkt mit sehr großer Hebelwirkung die Steigung einer Regressionsgeraden massiv verändern kann; Faraway (2005: Kap. 4–5) bietet einen guten Überblick über dieses in der praktischen Forschung oft vernachlässigte Thema.

Eine weitere große Herausforderung können fehlende Daten (*missing data*) sein, z. B. Datenpunkte, für die Versuchspersonen keine Antwort gaben, Korpusbeispiele, die nicht hinsichtlich eines Prädiktors kodiert werden konnten, etc. In der bisherigen Praxis scheinen fehlende Daten meist einfach ignoriert zu werden, d. h. sie werden nicht in die Modellierung einbezogen. Dies ist bedauerlich, da zum einen potentiell wichtige Information verloren geht – warum fehlen diese Datenpunkte, und nicht andere? – und da zum anderen das Weglassen dieser Daten die Stichprobengröße verkleinert, was das Identifizieren signifikanter Effekte erschwert. Daher sind die beiden folgenden Strategien oft nützlicher:

- Exploration der fehlenden Daten: treten die fehlenden Daten gehäuft in einer bestimmten Experimentalbedingung, Versuchsperson, oder Interaktion auf? Zu diesem Zweck kann es in der Tat nützlich sein, die Ab-/Anwesenheit von fehlenden Daten zu modellieren: gibt es Kombinationen von Prädiktoren, die das Auftreten von fehlenden Daten gut vorhersagen, und was suggeriert das für die gegenwärtigen Daten und nachfolgende Studien?
- *Imputation* der fehlenden Daten: solange der Prozentsatz an fehlenden Daten nicht zu hoch ist, kann versucht werden, die fehlenden Datenpunke aus den vorhandenen zu schätzen. Eine der einfachsten Methoden ist, jeden fehlenden Datenpunkt durch den Mittelwert des entsprechenden Prädiktors zu ersetzen. Eine bessere Methode ist, die fehlenden Daten durch ein Modell/eine Regression über alle vorhandenen Prädiktoren vorherzusagen oder über einen clusteranalytischen Ansatz diejenigen *x* Fälle in den Daten zu finden,

die dem fehlenden Datenpunkt am ähnlichsten sind, und dann aus diesen eine Vorhersage für das fehlende Datum zu generieren, etc.; vgl. Harrell (2001: Kap. 3), Faraway (2005: Kap. 12) oder Torgo (2011: Abschnitt 2.5).

5.3 Abhängige Datenpunkte und gemischte Modelle

Als letzte Herausforderung sei hier die momentan stark an Interesse gewinnende Modellierungsmethode der gemischten Modelle (*mixed effects* oder *multilevel models*) erwähnt. Bei dieser Methode handelt es sich um Regressionsmodelle, die besonders gut geeignet sind, Messwiederholungen/abhängige Stichproben, hierarchisch geschachtelte Prädiktoren sowie unbalancierte Stichproben, wie sie besonders in naturalistischen Daten vorkommen, zu analysieren. Während abhängige Daten in der Psycholinguistik lange durch F_1/F_2/(min)F-Statistiken analysiert wurden (siehe Clark 1973 sowie Forster & Dickinson 1976 zur Entwicklung dieses Standards), werden derartige Daten inzwischen mehr und mehr mit gemischten Modellen analysiert. Diese Modelle erlauben es, die herkömmlichen Statistiken zu berechnen – Koeffizienten, ihre Standardfehler und ihre p-Werte – gemischte Modelle berechnen aber darüber hinaus auch Korrekturen zu Achsenabschnitten und Steigungen für zufällige Effekte wie Versuchspersonen (vgl. *by-subject* Effekte), Stimuli (vgl. *by-item* Effekte) und andere. Dies macht es möglich, versuchspersonen- und stimulusspezifische Effekte zu berücksichtigen, was auch die Schätzungen der Koeffizienten der Prädiktoren präzisiert; siehe Faraway (2006: Kap. 8, 10) und Gelman & Hill (2007) für generelle Einführungen sowie Baayen (2008: Kap. 8) und Johnson (2008: Kap. 7) für linguistische Anwendungen.

Trotz all dieser Vorteile und des stark wachsenden Interesses sind gemischte Modelle hier (noch!) als Herausforderung gelistet. Der Hauptgrund dafür ist, dass – zumindest in sprachwissenschaftlichen Kontexten – diese Modelle m. E. noch in der Entwicklung und zentrale Fragen der Modellierung noch nicht beantwortet scheinen. Beispielsweise scheint die Frage, wie p-Werte für feste und zufällige Effekte berechnet werden, noch nicht für alle Arten von Regressionsmodellen geklärt und/oder implementiert zu sein. Außerdem scheint bei rückwärts-eliminierender Modellselektion (vgl. Abschnitt 3.1.1) noch unklar, wie ein maximales Modell aussieht: sollen alle Prädiktoren – alle Haupteffekte und alle Interaktionen – Korrekturen für alle zufälligen Effekte erhalten bzw. wie entscheidet man, welche Teilmenge von Korrekturen man zulässt? Und in welcher Reihenfolge eliminiert man feste und zufällige Effekte? Vor diesem Hintergrund bedürfen gemischte Modelle in der Linguistik m. E. noch weiterer Erprobung und Erklärung; sobald jedoch die o. g. Fragen klarer beantwortet wurden und robuste

Implementationen vorliegen, werden gemischte Modelle unzweifelhaft und umgehend zu einem der wichtigsten Werkzeuge statistischer Modellierung werden.

6 Schlusswort

Wie oben bereits erwähnt, handelt es sich bei Modellierung um ein Gebiet, das besonders in den letzten Jahren aufgrund mathematisch-theoretischer und technologischer Entwicklungen ungemein an Größe und Diversität gewonnen hat. Die Vielzahl an Entwicklungen findet daher auch nur langsam ihren Weg in das Gebiet der Sprachwissenschaft, das erst seit neuerem wieder so empirisch und quantitativ ist, dass es von derartigen Entwicklungen profitieren kann. Ebenfalls deutlich geworden ist hoffentlich, dass mit zunehmender Komplexität der Daten, Fragestellungen und Methoden eine klare Trennung von hypothesenprüfenden und explorativen Verfahren eigentlich nicht mehr sinnvoll ist. Selbst eine theoretisch streng hypothesenprüfende Studie wird – wenn sie denn mit aller Gründlichkeit durchgeführt wird – zahlreiche explorative Elemente enthalten:

- eine erste Inspektion der Daten, um Skalenniveaus, erforderliche Transformationen, generelle Verteilungseigenschaften sowie potentielle Ausreißer der beteiligten Variablen zu bestimmen; zu diesem Zeitpunkt werden ggf. auch explorative Methoden verwendet, um fehlende Daten zu berechnen;
- einen Modellselektionsprozess, indem Prädiktoren und ihre Ausprägungen durch eine motivierte Mischung aus empirischen Daten und analytischen Erwägungen addiert und eliminiert werden – ist Belebtheit eine signifikante Variable, die in die Regression eingehen muss, und selbst wenn, müssen wirklich z. B. fünf Belebtheitsstufen unterschieden werden oder untermauern die Daten vielleicht nur drei Stufen? Die Antworten auf solche Fragen können in einer Feedbackschleife dazu führen, dass z. B. aufgrund von Kollinearität Prädiktoren durch explorative Verfahren neu definiert werden und der ganze Vorgang von vorne begonnen wird;
- einen Validierungsprozess und eine *post hoc* Exploration auf spezielle Datenpunkte, die ebenfalls wieder die Modellformulierung und -selektion beeinflussen kann etc.; vgl. insbesondere Harrell (2001: Kap. 5) und Crawley (2007: Kap. 9).

Vor diesem Hintergrund sollte deutlich werden, dass neben Anforderungen an Daten und generellen statistischen Kenntnissen die wichtigste Komponente des Modellierungsprozesses immer noch der statistisch *und* theoretisch fachkundige Forscher ist.

Danksagung: Ich danke Stefanie Wulff, Anke Lüdeling und zwei Reviewern für ihren Input zu früheren Stadien dieses Artikels.

Literatur

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

Baayen, R. Harald. 2010a. A real experiment is a factorial experiment? *The Mental Lexicon* 5(1). 149–157.

Baayen, R. Harald. 2010b. Demythologizing the word frequency effect: a discriminative learning perspective. *The Mental Lexicon* 5(3). 436–461.

Baayen, R. Harald. 2011. Corpus linguistics and naive discriminative learning. *Brazilian Journal of Applied Linguistics* 11(2). 295–328.

Clark, Herbert H. 1973. The Language-as-Fixed-Effect Fallacy: a critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior* 12(4). 335–359.

Crawley, Michael J. 2007. *The R book*. Chichester: John Wiley.

Cook, Dianne & Deborah F. Swayne. 2007. *Interactive and dynamic graphics for data analysis. With R and GGobi*. New York: Springer.

Faraway, Julian J. 2005. *Linear models with R*. Boca Raton, FL: Chapman & Hall/CRC.

Faraway, Julian J. 2006. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. Boca Raton, FL: Chapman & Hall/CRC.

Forster, Kenneth I. & R. G. Dickinson. 1976. More on the language-as-fixed-effect fallacy: Monte Carlo estimates of error rates for F_1, F_2, F, and min F. *Journal of Verbal Learning and Verbal Behavior* 15(2). 135–142.

Fox, John. 1997. *Applied regression analysis, linear models, and related methods*. Thousand Oaks, CA: Sage Publications.

Fox, John. 2008. *Applied regression analysis and generalized linear models*. 2nd ed. Thousand Oaks, CA: Sage Publications

Gelman, Andrew & Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Gries, Stefan Th. 1999. Particle movement: a cognitive and functional approach. *Cognitive Linguistics* 10(2). 105–145.

Gries, Stefan Th. 2006. Exploring variability within and between corpora: some methodological considerations. *Corpora* 1(2). 109–151.

Gries, Stefan Th. 2008. *Statistik für Sprachwissenschaftler*. Göttingen: Vandenhoeck & Ruprecht.

Gries, Stefan Th. 2009. *Statistics for linguistics with R: a practical introduction*. Berlin & New York: Mouton de Gruyter.

Gries, Stefan Th. 2011. Commentary. In Kathryn Allan & Justyna Robinson (Hrsg.), *Current methods in historical semantics*, 184–195. Berlin & New York: Mouton de Gruyter.

Harrell, Frank E. Jr. 2001. *Regression modeling strategies. With applications to linear models, logistic regression, and survival analysis*. New York: Springer.

Hastie, Trevor & Robert Tibshirani. 1990. *Generalized additive models*. Boca Raton, FL: Chapman & Hall/CRC.

Hastie, Trevor, Robert Tibshirani & Jerome Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*. 2nd ed. New York: Springer.

Hilpert, Martin. 2011. Diachronic collostructional analysis: how to use it and how to deal with confounding factors. In Kathryn Allan & Justyna Robinson (Hrsg.), *Current methods in historical semantics*, 130–160. Berlin & New York: Mouton de Gruyter.

Jankowski, Bridget. 2004. A transatlantic perspective of variation and change in English deontic modality. *Toronto Working Papers in Linguistics* 23(2). 85–113.

Jarvis, Scott. 2011. Data mining with learner corpora: choosing L1 classifiers for L1 detection. In Fanny Meunier, Sylvie De Cock, Gaëtanelle Gilquin & Magali Paquot (Hrsg.), *A taste for corpora. In honor of Sylviane Granger*, 127–154. Amsterdam & Philadelphia: John Benjamins.

Johnson, Keith. 2008. *Quantitative methods in linguistics*. Malden, MA: Blackwell.

Keen, Kevin J. 2010. *Graphics for statistics and data analysis with R*. Boca Raton, FL: CRC Press, Taylor & Francis Group.

Leech, Geoffrey, Brian Francis & Xfueng Xu. 1994. The use of computer corpora in the textual demonstrability of gradience in linguistic categories. In Catherine Fuchs & Bernard Victorri (Hrsg.), *Continuity in linguistic semantics*, 57–76. Amsterdam & Philadelphia: John Benjamins.

Molinaro, Anette M., Richard Simon & Ruth M. Pfeiffer. 2005. Prediction error estimation: a comparison of resampling methods. *Bioinformatics* 21(15). 3301–3307.

Murrell, Paul. 2011. *R graphics*. 2nd ed. Boca Raton, FL: CRC Press, Taylor & Francis Group.

Paolillo, John C. 2002. *Analyzing linguistic variation: statistical models and methods*. Stanford, CA: CSLI Publications.

R Development Core Team. 2011. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. URL ‹http://www.R-project.org/›.

Sarkar, Deepayan. 2008. *Lattice: multivariate data visualization with R*. New York: Springer.

Suzuki, Ryota & Hidetoshi Shimodaira. 2011. pvclust: Hierarchical clustering with *p*-values via multiscale bootstrap resampling. R package version 1.2–2. ‹http://CRAN.R-project.org/package=pvclust›.

Teich, Elke & Peter Fankhauser. 2010. Exploring a corpus of scientific texts using data mining. In Stefan Th. Gries, Stefanie Wulff & Mark Davies (Hrsg.), *Corpus-linguistic applications: current studies, new directions*, 233–247. Amsterdam: Rodopi.

Torgo, Luís. 2011. *Data mining with R: learning with case studies*. Boca Raton, FL: Chapman & Hall/CRC.

Tryk, H. Edward. 1986. Subjective scaling of word frequency. *The American Journal of Psychology* 81(2). 170–177.

Tuttle, Siri G. & Olga Lovick. 2007. Intonational marking of discourse units in two Dena'ina narratives. *Nouveaux Cahiers de Linguistique Française* 28. 305–316.

Unwin, Antony, Martin Theus & Heike Hofmann. 2006. *Graphics of large datasets: visualizing a million*. New York: Springer.

Wickham, Hadley. 2009. *ggplot2: elegant graphics for data analysis*. New York: Springer.