

Gabriel Altmann

Die Entstehung diatopischer Varianten

Ein stochastisches Modell

Mit Hilfe eines adaptierten Poissonschen Geburts- und Todesprozesses wird ein Modell der Verteilung diatopischer Varianten in einem Sprachgebiet abgeleitet und am Atlas linguistique de la France überprüft

1. Methodologische Voraussetzungen

Die Tatsache, daß alle Sprachen und deren Einheiten (Phoneme, Morpheme etc.) Variation aufweisen, ist hinreichend bekannt. Variation (Variabilität, Polymorphie etc.) ist eine Eigenschaft, die – neben vielen anderen – Sprache mit allen lebendigen Organismen gemeinsam hat. Dabei kann sich die Variation in der *Synchronie* (gleichzeitige Formenvielfalt), in der *Diatopie* (sprachgeographische Polymorphie), in der *Diastratie* (soziolinguistische Polymorphie) oder in der *Diachronie* (Polymorphie im zeitlichen Nacheinander) äußern. Dies gilt auch für die *Polysemie* einzelner Spracheinheiten (Wörter, Morpheme etc.). Diese können zu einem gegebenen Zeitpunkt (synchronische Perspektive) oder über einen bestimmten Zeitraum hinweg (diachronische Perspektive) mehrere Bedeutungen aufweisen. Sowohl die gleichzeitige Vermehrung der Formen mit derselben Bedeutung (Problem der Onomasiologie: etwa bei sprachgeographischer Variation) oder der Bedeutungen einer bestimmten Form (Problem der Polysemie) sowohl aus synchronischer wie aus diachronischer Sicht stellen jedesmal *Prozesse* dar, aus denen synchronisch beobachtbare Sprachstrukturen entstehen.

Für den Bereich der Sprache muß es eine ganze Reihe von derartigen Prozessen geben, wiewohl bisher nur wenige entdeckt wurden. Es ist möglich, mit ihrer Hilfe sprachliche Erscheinungen nicht nur zu *beschreiben*, sondern auch zu *erklären*. Die Erfassung dieser Prozesse führt zur Aufstellung von *Gesetzhypothesen*, d. h. von wohlgeformten und sinnvollen Aussagen, die *deduktiv abgeleitet* (= theoretisch begründet), *induktiv* relativ gut *bestätigt* (= empirisch begründet), für alle Sprachen gültig und in ein System derartiger Aussagen eingebettet sind (vgl. Bunge 1969: 351). Aussagen dieser Art stehen auf einer höheren erkenntnistheoretischen Stufe als zum Beispiel „Lautgesetze“ oder grammatische Regeln, die lediglich *empirische Generalisierungen* darstellen. Hinzuzufügen ist, daß prinzipiell zwischen *empirischen Generalisierungen* (mit niederem erkenntnistheoretischem Wert) und *deduktiv gewonnenen Gesetzhypothesen* (mit höherem erkenntnistheoretischem Wert) unterschieden werden muß.

1.1. Die Systemisierung einer Hypothese ist im derzeitigen Entwicklungsstand der Linguistik noch etwas schwierig (vgl. Altmann 1978; Köhler/Altmann 1983). Es gibt jedoch Wege, dieser Systemisierung näher zu kommen: wir postulieren, daß die Sprache einheitlich konstruiert ist, d.h. daß die Auswirkungen eines bestimmten Prozesses auf allen Ebenen der Sprache in relativ einheitlicher Weise nachweisbar sind. Läßt sich bei einer theoretisch abgeleiteten Aussage empirisch nachweisen, daß sie für mindestens zwei Sprachebenen gilt, so können wir sie bereits als einen Gesetzes-„Kandidaten“ betrachten. Die Systemisierung einer derartigen Aussage liegt dann vor, wenn sie zusammen mit analogen Aussagen aus anderen Sprachebenen unter einen einzigen Prozeß oder unter ein allgemeineres Gesetz subsumiert wird. In bezug auf dieses allgemeine Gesetz erscheint sie nunmehr als ein *Spezialfall*.

Da wir verlangen, daß ein Gesetz für alle Sprachen gilt, umgekehrt aber den Menschen ein erschöpfender empirischer Zugriff auf die Gesamtheit aller Sprachen verwehrt ist, kann man von seiner endgültigen Verifikation nie sprechen. Man kann jedoch derartigen Allaussagen durch zahlreiche Tests einen hohen Grad an Gültigkeit (Zuverlässigkeit, Kreditibilität) verleihen. Es ist interessant festzustellen, daß einige Gesetzhypothesen der *Linguistik* auch in *anderen Wissenschaften* Gültigkeit haben. So wurde gezeigt, daß das *Zipf-Mandelbrotsche Gesetz* nicht nur auf allen Sprachebenen, sondern auch in der Musik, Malerei, Biologie, Psychologie, Verhaltensforschung, Soziologie u.a. gilt (cf. Orlov/Boroda/Nadarejşvili 1982; Guiter/Arapov 1982). Ein weiteres Gesetz, das *Menzerathsche Gesetz*, gilt nicht nur bei der Konstruktion sprachlicher Formen und Bedeutungen, sondern auch in der Biologie (z. B. für gewisse Aspekte der Strukturierung von DNA-Ketten) und der Soziologie der Primaten (z. B. für die Strukturierung von Primatenpopulationen in Herde, Bande, Clan etc.) (vgl. Altmann/Schwibbe et al. 1985). Das *Piotrowski-Gesetz* schließlich gilt nicht nur bei Prozessen der Sprachentwicklung und Fremdwortentlehnung, sondern auch bei analog verlaufenden Prozessen in der Biologie, Ökonomie, Epidemiologie, Gerücheverbreitung etc. (vgl. Altmann/v. Buttler/Rott/Strauß 1983; Best/Kohlhase 1983).

Wenn wir nun nach einem Gesetz für die *Verteilung diatopischer Varianten* suchen, so können wir bereits auf eine vorhandene Gesetzhypothese zurückgreifen, nämlich auf das „Diversifikationsgesetz“ (vgl. Beöthy/Altmann 1984a, b; Altmann 1985; Rothe 1984; Best/Altmann/Kind 1985), dessen Gültigkeit im Bereich der Semantik durch zahlreiche Anwendungsfälle bestätigt wurde. Da es sich bei diatopischen Varianten um die *geographische Diversifikation dialektaler Formen* handelt, müssen hier entsprechend andere *Begriffe* eingesetzt werden. Daher finden wir es zweckmäßig, den gesamten Gedankengang, der zum Modell der diatopischen (sprachgeographischen) Variation führt, in der Folge ausführlich zu besprechen.

2. Das Modell

Goebl (1984) hat für das Gebiet der Normandie (eigentlich Nordwestfrankreichs unter vollem Einschluß der Normandie) eine zufällige Stichprobe von Atlaskarten aus dem „Atlas linguistique de la France“ (ALF) erhoben. Dabei zählte er ab, wieviele lexikalische (d.h. onomasiologische) Varianten es auf den einzelnen ALF-Karten gibt. So erhielt er beispielsweise für ein Korpus von 100 aleatorisch aus dem ALF erhobenen Atlaskarten die folgende Verteilung (cf. Tabelle 1).

x	f _x
1	32
2	22
3	10
4	9
5	8
6	7
7	4
8	2
9	1
10	3
11	1
12	1

Dabei bedeuten:

x Anzahl der lexikalischen Varianten auf einer ALF-Karte

f_x Anzahl der ALF-Karten mit genau x Varianten

Tabelle 1. Verteilung der Variantenanzahl anhand eines aleatorischen Teilkorpus („Staffelkorpus“) des Arbeitskorpus LEXNORMANDIE (Goebl 1984: 72)

2.1. Ansatz des Modells

Die ersten beiden Zeilen der Tabelle 1 sind wie folgt zu lesen: es gibt 32 Karten, auf denen nur 1 Variante aufscheint (d.h., daß der betreffende Begriff keine Varianten kennt); es gibt 22 Karten, auf denen es jeweils 2 Varianten gibt, usw. Da diese Zahlen den Eindruck erwecken, daß hinter ihnen irgendeine Gesetzmäßigkeit steckt, wollen wir nach ihr suchen. Dies umsomehr, als wir – Bunges 8. . .

ontologischem Prinzip der wissenschaftlichen Forschung folgend – glauben, daß „every thing abides by laws“ (Bunge 1977: 16f.). Wäre es nämlich nicht so, dann gäbe es nur Chaos, und die Wissenschaft als Suche nach Ordnung würde sich crübrigen. Bei unseren Überlegungen gehen wir davon aus, daß eine beliebige Spracheinheit (Wörter, Laute, Morphe etc.) innerhalb eines bestimmten Zeitintervalls variiert; so wird z. B. ein Wort (type) bei jeder erneuten Verwendung (in der Form eines tokens) etwas anders ausgesprochen. Dabei bleibt die neue Variante (token) entweder bestehen und setzt sich durch oder verschwindet wieder. Oder: ein Wort bekommt innerhalb eines Zeitintervalles eine neue Nebenbedeutung dazu bzw. verliert eine alte. Oder: in einem bestimmten Sprachgebiet entsteht eine neue onomasiologische Variante des Begriffes 'Pferd' und so fort. All dies beruht auf dem Zusammenwirken zwischen Sprachgemeinschaft und Sprachvariation.

Betrachten wir den Variabilitätszustand einer Spracheinheit, d. h. ihre formale oder semantische Diversifikation zum Zeitpunkt t . Wir stellen fest, wie viele Varianten sie im Zeitintervall $(0, t)$ bekommen hat. Der Zeitpunkt t ist der Zeitpunkt der Untersuchung, der Zeitpunkt 0 ist irrelevant. Wenn die Diversifikation gesetzesartig verläuft, dann reicht es, die Variabilitätsveränderungen in einem kleinen Zeitintervall $(t, t + \Delta t)$ zu beobachten, um aus diesen Veränderungen den ganzen Prozeß zu entschlüsseln. Zu diesem Zweck stellen wir einige Annahmen auf, die für diatopische Variation spezifisch sind. Sie spielen die Rolle von Postulaten, aus denen ein Modell abgeleitet wird.

Postulat I:

Die Wahrscheinlichkeit, daß im Zeitabschnitt Δt genau 1 neue diatopische Variante entsteht („geboren wird“), ist der Größe dieses Intervalles proportional, d. h.

$$a \Delta t,$$

wobei a als die „Geburtsrate“ bezeichnet werden kann.

Neue Varianten werden ständig geboren, unabhängig davon, wie viele es schon gibt. Dies sind jedoch nur potentielle Varianten bzw. potentielle Rivalen der schon existierenden Varianten. Die „Neugeburten“ in der Sprache verhalten sich nämlich anders als die in der biologischen Realität. Die Geburt einer Variante bedeutet noch keineswegs einen Zuwachs der Population, da die neue Variante gleich nach ihrer ersten lautlichen Realisierung („Geburt“) – der Natur von Sprache entsprechend – verschwindet und damit eine vorderhand nur potentielle Existenz hat. Wenn man diese Variante als eine reale Entität betrachtet, so muß man davon ausgehen, daß sie sich gegen die existierenden Varianten durchzusetzen hat. D. h., die neue Variante sollte nicht nur zum Zeitpunkt ihrer

ersten Realisierung durch einen Sprecher („Geburt“), sondern auch zum Zeitpunkt ihrer Erhebung durch einen Linguisten im Sprachschatz der betreffenden Sprachgemeinschaft nachweisbar sein. Um diesen Sachverhalt wiederzugeben, müssen wir noch eine weitere Annahme treffen:

Postulat II:

Die Wahrscheinlichkeit, daß sich 1 Variante im Zeitabschnitt Δt gegen x Rivalen durchsetzt, ist diesen beiden Größen proportional und kann als

$$bx\Delta t$$

bezeichnet werden. Dabei ist b die „Durchsetzungsrate“ und x die Zahl der konkurrierenden Varianten.

Postulat III:

Varianten unterliegen nicht nur dem Phänomen der Entstehung (d. h. der Geburt und der Durchsetzung), sondern auch dem Phänomen des Vergehens (d. h. dem Tod). Dies geschieht nicht nur durch den Druck, den die Hochsprache auf die Dialekte ausübt, sondern auch durch die Unifikationskraft, mit der der passive Sprachbenützer (d. h. der Hörer) die formale Diversifikation, die ihm vom aktiven Sprachbenützer (d. h. vom Sprecher) angeboten wird, abbremst oder rückgängig macht. Je mehr Varianten vorhanden sind, desto stärker muß die Unifikationskraft werden, um ein Ausufern der Variabilität – wodurch es zum Zusammenbruch der Kommunikation kommen könnte – abzufangen. Deswegen nehmen wir an, daß die Wahrscheinlichkeit, mit der im Zeitabschnitt Δt von den vorhandenen Varianten 1 verschwindet („stirbt“), sowohl der Größe dieses Intervalles als auch der Zahl der Konkurrenten proportional ist, d. h.

$$cx\Delta t.$$

Dabei ist c die „Sterberate“.

Postulat IV:

Bisher haben wir nur die Wahrscheinlichkeit berücksichtigt, daß in einem Zeitintervall nur 1 Variante geboren wird, sich durchsetzt oder stirbt. Wir müssen aber noch zwei weitere Annahmen machen. Die erste dieser beiden Annahmen betrifft die Wahrscheinlichkeit, mit der in dem Zeitintervall Δt 2 oder mehr Varian-

ten geboren werden, sich durchsetzen oder sterben. Die Größe dieser Wahrscheinlichkeit entspricht der Ordnung des gewählten Intervalles, $0(\Delta t)$, und ist so klein, daß sie vernachlässigt werden kann. Man nimmt an, daß $0(\Delta t)/\Delta t \rightarrow 0$, wenn $\Delta t \rightarrow 0$.

Postulat V:

Die zweite der eben erwähnten zwei Annahmen berücksichtigt den Fall, daß in Δt 0 Veränderungen entstehen, d. h., es wird keine Variante geboren, keine setzt sich daher durch, und keine stirbt. Die Wahrscheinlichkeit dafür berechnet sich aus dem Produkt der Komplemente der obigen Wahrscheinlichkeiten, d. h. aus

$$(1 - a\Delta t)(1 - bx\Delta t)(1 - cx\Delta t).$$

Da $(\Delta t)^2$ noch kleiner als $0(\Delta t)$ ist, kann man es vernachlässigen und erhält somit die Wahrscheinlichkeit von 0 Veränderungen in der folgenden Form:

$$1 - [a + (b + c)x] \Delta t.$$

Postulat VI:

Wie bei Poisson-Prozessen üblich, verlangt man, daß die Intervalle disjunkt und die in ihnen ablaufenden Ereignisse von den Ereignissen in den anderen Intervallen unabhängig sind. Dieses Axiom muß aus wahrscheinlichkeitstheoretischen Gründen angesetzt werden.

Von diesen Annahmen ausgehend können wir nunmehr den gesamten Diversifikationsprozeß erfassen. Wir bezeichnen die Wahrscheinlichkeit, daß im Intervall $(0, t)$ genau x Varianten (z. B. eines Begriffes) vorhanden sind, als $P_x(t)$ und die Wahrscheinlichkeit, daß im Intervall $(t, t + \Delta t)$ genau x Varianten vorhanden sind, als $P_x(t + \Delta t)$. Wir untersuchen nunmehr die letztere Wahrscheinlichkeit. Diese setzt sich aus *fünf Teilen* (s. unten) zusammen. Vorab überlegen wir aber:

- entweder gab es bis zur Zeit t genau $x - 1$ Varianten und 1 wurde geboren oder es gab bis zur Zeit t genau $x - 1$ Varianten und 1 potentielle neue Variante setzte sich durch;
- oder es gab $x + 1$ Varianten und 1 starb;
- oder es gab x Varianten und nichts veränderte sich;
- oder es gab $x \pm r$ Varianten und es gab r Veränderungen ($r > 1$).

Diese vier Überlegungen kann man formal unter Rückgriff auf die Postulate I bis VI leicht dazu verwenden, die oben genannte Wahrscheinlichkeit $P_x(t + \Delta t)$ zu erhalten.

Teil (i):

Bis zum Zeitpunkt t gab es $x - 1$ Varianten [mit der Wahrscheinlichkeit $P_{x-1}(t)$], und in Δt wurde 1 Variante geboren [mit der Wahrscheinlichkeit $a \Delta t$ laut Postulat I]. Da laut Postulat VI die Ereignisse voneinander unabhängig sind, können wir die Wahrscheinlichkeiten ausmultiplizieren und erhalten:

$$(1) \quad a \Delta t \cdot P_{x-1}(t)$$

Teil (ii):

Bis zum Zeitpunkt t gab es $x - 1$ Varianten, und von den potentiellen Varianten hat sich genau 1 gegen die $x - 1$ Rivalen durchgesetzt [mit der Wahrscheinlichkeit $b(x - 1) \Delta t$ laut Postulat II]. Das Produkt ergibt also:

$$(2) \quad b(x - 1) \Delta t \cdot P_{x-1}(t).$$

Teil (iii):

Bis zum Zeitpunkt t gab es $x + 1$ Varianten, und 1 davon wurde in Δt von ihren Rivalen verdrängt, was laut Postulat III folgendes ergibt:

$$(3) \quad c(x + 1) \Delta t \cdot P_{x+1}(t).$$

Teil (iv):

Bis zum Zeitpunkt t gab es x Varianten, und in Δt entstand keine Veränderung. Das Produkt dieser beiden Wahrscheinlichkeiten ergibt laut Postulat V:

$$(4) \quad \{1 - [a + (b + c)x] \Delta t \cdot P_x(t)\}.$$

Teil (v):

Bis zum Zeitpunkt t gab es genau $x \pm r$ Varianten, und in Δt kamen r Veränderungen zustande (bei $x - r$ wurden r Varianten geboren oder setzten sich durch; bei $x + r$ starben r Varianten). Die entsprechende Wahrscheinlichkeit ist laut Postulat IV gleich

$$(5) \quad 0(\Delta t).$$

Diesen Veränderungsprozeß, der als *Geburts- und Todesprozeß* bekannt ist, kann man in seiner Gesamtheit übersichtlich durch Tabelle 2 darstellen.

Zahl der Varianten (types) bis zum Zeitpunkt t	Veränderung in $t + \Delta t$	Wahrscheinlichkeit
$x - 1$	1	$a \Delta t P_{x-1}(t) + b(x-1) \Delta t P_{x-1}(t)$
$x + 1$	1	$c(x+1) \Delta t P_{x+1}(t)$
x	0	$\{1 - [a + (b+c)x] \Delta t\} P_x(t)$
$x \pm r$	r	$0(\Delta t)$

Tabelle 2. Komponenten des Prozesses der Variantenbildung

2.2. Aufstellung des Modells

Man bedenke, daß Geburt und Durchsetzung keine gleichzeitigen, sondern alternative Prozesse sind. Daher kann man die betreffenden Wahrscheinlichkeiten addieren (siehe Zeile 1 der Tabelle 2). Dies gilt auch für alle in der dritten Spalte der Tabelle 2 aufgeführten Ereignisse, da diese alle alternativ sind. Aus diesem Grund bekommen wir die Wahrscheinlichkeit, daß es im Zeitintervall $(t, t + \Delta t)$ genau x akzeptierte Varianten geben wird, in der Form der Summe der in der Tabelle 2 angeschriebenen Wahrscheinlichkeiten, d. h.

$$(6) \quad P_x(t + \Delta t) = [a + b(x-1)] \Delta t P_{x-1}(t) + c(x+1) \Delta t P_{x+1}(t) + \{1 - [a + (b+c)x] \Delta t\} P_x(t) + 0(\Delta t).$$

Subtrahiert man von den beiden Seiten der Formel (6) $P_x(t)$ und dividiert die beiden Seiten sodann durch Δt , so erhält man:

$$(7) \quad \frac{P_x(t + \Delta t) - P_x(t)}{\Delta t} = [a + b(x-1)] P_{x-1}(t) + c(x+1) P_{x+1}(t) - [a + (b+c)x] P_x(t) + 0(\Delta t)/\Delta t$$

Wählt man nun Δt so klein, daß es gegen Null strebt, d. h. geht man im Ausdruck (7) zur Grenze über, wenn $\Delta t \rightarrow 0$, dann erhält man:

$$(8) \quad \frac{dP_x(t)}{dt} = [a + b(x-1)] P_{x-1}(t) + c(x+1) P_{x+1}(t) - [a + (b+c)x] P_x(t)$$

für $x \geq 1$

weil $0(\Delta t)/\Delta t$ wegen Postulat IV zu Null wird.

Bevor wir (8) lösen, machen wir eine *Vereinfachung*, die die theoretische Weiterentwicklung unseres Modells etwas erleichtert, jedoch seine Korrektheit und die Resultate nicht beeinträchtigt. Wenn man sagt, daß eine Spracheinheit keine Variabilität aufweist, dann ist nur eine Grundform (type), aber keine Variante von ihr (token) vorhanden. Das heißt, daß in der Tabelle von Goebel (siehe hier in Tabelle 1) an jener Stelle, wo $x = 1$ steht, eigentlich $x = 0$ stehen müßte. Denn auf diesen (mononymen) Karten gibt es keine Variation. Analog muß man für $x = 2$ in Tabelle 1 nunmehr $x = 1$ setzen und so fort. Praktisch bedeutet dies nichts anderes als eine einfache Verschiebung unserer Variablen auf der Zahlenachse um 1 nach links, so daß die Verteilungen bei $x = 0$ anfangen. Dies ist in der Tat nur eine formale Veränderung, obwohl sie auch sachlich vertretbar wäre. Hingegen ist bei der semantischen Diversifikation, wo die Variable x die Ränge darstellt, diese Verschiebung sachlich nicht gerechtfertigt. Durch die eben erwähnte Vereinfachung bekommen wir für $P_0(t)$

$$(9) \quad \frac{dP_0(t)}{dt} = cP_1(t) - aP_0(t)$$

Um die Gleichungen (8) und (9) lösen zu können, machen wir noch eine rein linguistische Annahme, die gut begründbar ist.

Postulat VII:

Es ist bekannt, daß Sprache, geographisch gesehen, inhomogen ist, d. h., es gab höchstwahrscheinlich schon immer Variabilität, und diese Variabilität war recht stabil. Es wäre ziemlich unplausibel anzunehmen, daß es im Althochdeutschen weniger oder mehr Varianten gab als im heutigen Deutschen. Ja sogar bei der Rekonstruktion der indogermanischen „Ursprache“ benutzt man implizit diese Annahme, andernfalls wäre es kaum gerechtfertigt, für eine angeblich „homogene“ Ursprache bis zu 10 Laryngale anzusetzen.

Die Variabilität ist also zeitunabhängig und hält sich ständig in einem Gleichgewicht, das im Lichte unserer linguistischen Argumentation durch die *Balance* zwischen *Geburt/Entstehung* und *Vergehen/Tod* versinnbildlicht werden kann. Diese Annahme ermöglicht uns, in (8) und (9) die Zeit t zu eliminieren und die beiden ersten Ableitungen gleich Null zu setzen. Wir bekommen also:

$$(10) \quad \begin{aligned} -aP_0 + cP_1 &= 0 \\ -[a + (b + c)x]P_x + [a + b(x - 1)]P_{x-1} + c(x + 1)P_{x+1} &= 0, \quad x \geq 1 \end{aligned}$$

Diese Differenzengleichungen kann man auf mehrere Weisen lösen. Hier ver-

wenden wir die Methode der schrittweisen Eliminierung. Bezeichnen wir die Größen

$$\lambda_x = a + bx$$

$$\mu_x = cx$$

so ergibt sich:

$$(11) \quad P_x = P_0 \prod_{i=1}^x \frac{\lambda_{i-1}}{\mu_i} = P_0 \frac{\lambda_0 \lambda_1 \dots \lambda_{x-1}}{\mu_1 \mu_2 \dots \mu_x}.$$

Setzen wir unsere Größen in (11) ein, so erhalten wir:

$$(12) \quad P_x = P_0 \frac{a(a+b)(a+2b) \dots a + (x-1)b}{x!c^x}$$

Hier klammern wir b aus und erhalten:

$$(13) \quad P_x = P_0 \frac{b^x}{c^x x!} \frac{a}{b} \left(\frac{a}{b} + 1 \right) \left(\frac{a}{b} + 2 \right) \dots \left[\frac{a}{b} + (x-1) \right].$$

Es muß $c > b$ sein, sonst gäbe es kein Gleichgewicht. Denn, wenn die Durchsetzungsrate größer wäre als die Sterberate, würde das zu einem Ausufern der Variantenzahlen und schließlich zum Zerfall der Kommunikation führen. Wir setzen

$$(14) \quad \frac{b}{c} = q, \quad \text{wo} \quad 0 < q < 1 \quad \text{und}$$

$$\frac{a}{b} = r.$$

Dann wird (13) zu:

$$(15) \quad P_x = P_0 \frac{r(r+1) \dots (r+x-1)}{x!} q^x = P_0 \binom{r+x-1}{x} q^x.$$

P_0 berechnen wir aus der Bedingung, daß $\sum P_x = 1$ sein muß, d. h.,

$$1 = \sum P_x = P_0 \sum \binom{r+x-1}{x} q^x = P_0 (1-q)^{-r},$$

woraus $P_0 = (1-q)^r$ folgt.

Da $q < 1$, bezeichnen wir $1 - q = p$, woraus sich dann $P_0 = p^r$ ergibt. Schließlich erhalten wir:

$$(16) \quad P_x = \binom{r+x-1}{x} p^r q^x, \quad x = 0, 1, \dots$$

Das ist die *negative Binomialverteilung*, die ein *Modell* für *Goebels Daten* darstellen sollte. Die theoretische Zahl der Karten mit x Varianten ergibt sich dann als NP_x , wobei N die Zahl der untersuchten Karten ist (Stichprobenumfang).

Hätten wir oben die Verschiebung der Variablen nicht durchgeführt, so hätten wir statt (16) die *1-verschobene negative Binomialverteilung* $P_x = \binom{r+x-2}{x-1} p^r q^{x-1}$, für $x = 1, 2, \dots$, – wie auch bei anderen Diversifikationsproblemen – erhalten.

3. Überprüfung des Modells anhand empirischer Daten

Die Parameter der Verteilung, p und r , sind nicht bekannt und müssen anhand der Daten geschätzt werden. Von den zahlreichen Schätzmethoden für diese Verteilung (vgl. Johnson/Kotz 1969: 131–135) liefert hier eigentlich jede gute Resultate. Wir benutzten die *Maximum-likelihood-Methode*. Die zwölf empirischen Verteilungen von Goebel (1984: 71–73), die nach steigender Zahl der Karten (also nach wachsendem Stichprobenumfang) angeordnet sind, findet man zusammen mit den theoretischen Häufigkeiten und dem Chiquadrat-Test in der Tabelle 3. Die Verschiebung der Variablen in den Tabellen ist – wie oben erwähnt – nur ein Konventionsproblem.

Wie man sieht, ist die Anpassung des Modells an die empirischen Daten sehr gut, so daß man vorläufig annehmen kann, daß die diatopische Diversifikation einem Poissonschen Geburts- und Todesprozeß folgt. Es bedarf jedoch noch zahlreicher Überprüfungen anhand anderer diatopischer Daten aus anderen Sprachgebieten, bis man die gegebene Form des Modells als „gut unterstützt“ betrachten kann. Eine starke Unterstützung dieses Modells ist jedoch durch seine Gültigkeit im Bereich der Semantik und zahlreicher nichtlinguistischer Disziplinen (vgl. Haight 1967; Parzen 1967 etc.) gegeben.

Drei Umstände sind hier noch zu bemerken:

- (1) Mit der Vergrößerung der Stichprobe scheinen sich die Parameter p und r zu stabilisieren. Anscheinend konvergiert p gegen $\frac{1}{3}$ und r gegen 1. Dies kann man natürlich als eventuell nur für die vorliegenden empirischen Daten gültig

	x	f _x	NP _x	x	f _x	NP _x	x	f _x	NP _x
	0	9	7.11	0	15	15.13	0	23	22.84
	1	2	6.27	1	11	10.82	1	17	15.66
	2	6	4.43	2	8	7.54	2	8	10.90
	3	2	2.87	3	3	5.20	3	5	7.62
	4	4	1.78	4	5	3.58	4	7	5.34
	5	1	1.07	5	2	2.45	5	5	3.75
	≥6	1	1.48	6	3	1.68	6	4	2.64
				7	0	1.41	7	2	1.85
				8	1	0.78	8	1	1.31
				9	1	0.53	9	2	0.92
				≥10	1	1.14	≥10	1	2.19
N	25			50			75		
p	0.4689			0.3221			0.2939		
r	1.6600			1.0550			0.9710		
X ²	7.18			4.29			5.43		
FG	4			7			7		
P	0.13			0.75			0.61		

	x	f _x	NP _x	x	f _x	NP _x	x	f _x	NP _x
	0	32	32.02	0	48	49.25	0	55	59.02
	1	22	20.63	1	36	31.36	1	50	42.24
	2	10	14.07	2	16	21.15	2	27	29.76
	3	9	9.77	3	13	14.54	3	21	20.86
	4	8	6.85	4	10	10.08	4	12	14.58
	5	7	4.83	5	9	7.03	5	9	10.58
	6	4	3.42	6	6	4.92	6	8	7.10
	7	2	2.42	7	2	3.45	7	2	4.94
	8	1	1.72	8	4	2.42	8	6	3.44
	9	3	1.22	9	3	1.71	9	5	2.40
	10	1	0.87	10	2	1.20	10	3	1.67
	≥11	1	2.18	≥11	1	2.89	≥11	2	3.81
N	100			150			200		
P	0.2801			0.2877			0.3065		
r	0.8950			0.8940			1.0320		
X ²	6.21			7.31			11.07		
FG	8			9			9		
P	0.62			0.60			0.27		

	x	f _x	NP _x	x	f _x	NP _x	x	f _x	NP _x
	0	71	75.20	0	88	91.96	0	103	109.06
	1	60	51.99	1	71	63.50	1	88	76.59
	2	33	36.36	2	42	44.04	2	51	52.66
	3	25	25.53	3	30	30.59	3	35	35.94
	4	17	17.96	4	20	21.27	4	23	24.44
	5	11	12.65	5	13	14.79	5	13	16.59
	6	10	8.92	6	11	10.29	6	12	11.24
	7	3	6.29	7	4	7.16	7	4	7.61
	8	7	4.44	8	7	4.98	8	7	5.15
	9	5	3.13	9	5	3.47	9	5	3.48
	10	4	2.21	10	4	2.41	10	4	2.35
	11	3	1.56	11	3	1.68	11	3	1.59
	≥12	1	3.76	12	1	1.17	12	1	1.07
				≥13	1	2.68	≥13	1	2.22
N	250			300			350		
P	0.2924			0.3033			0.3273		
r	0.9770			0.9910			1.0440		
X ²	11.30			7.55			9.16		
FG	10			11			11		
P	0.33			0.75			0.61		

	x	f _x	NP _x	x	f _x	NP _x	x	f _x	NP _x
	0	117	123.37	0	133	138.89	0	148	154.65
	1	100	86.71	1	111	98.86	1	125	110.35
	2	58	59.91	2	68	68.23	2	72	76.06
	3	37	41.16	3	40	46.59	3	47	51.81
	4	29	28.19	4	33	31.65	4	38	35.08
	5	15	19.28	5	18	21.43	5	19	23.67
	6	13	13.17	6	15	14.48	6	16	15.93
	7	6	8.99	7	7	9.77	7	8	10.71
	8	9	6.13	8	9	6.58	8	10	7.18
	9	5	4.18	9	5	4.43	9	5	4.82
	10	5	2.85	10	5	2.98	10	6	3.23
	11	3	1.94	11	3	2.00	11	3	2.16
	12	1	1.32	12	1	1.35	12	1	1.44
	≥13	2	2.82	≥13	2	2.75	≥13	2	2.91
N	400			450			500		
P	0.3209			0.3316			0.3350		
r	1.0350			1.0650			1.0730		
X ²	8.84			7.20			8.99		
FG	11			11			11		
P	0.64			0.78			0.62		

Tabelle 3. Goebels Verteilungen der Variantenzahlen im Dialektatlas der Normandie

betrachten. Ist aber $r = 1$, so ist $a = b$ (vgl. Formel 14). Wenn $r = 1$, dann reduziert sich die *negative Binomialverteilung* auf die *geometrische Verteilung*, was man direkt am Modell zeigen kann. Wenn

$$\lambda_x = a + bx = a + ax = a(x + 1)$$

$$\mu_x = cx,$$

dann ergibt sich die Lösung von (10) als:

$$(17) \quad x = P_0 \prod_{i=1}^x \frac{\lambda_{i-1}}{\mu_i} = P_0 \prod_{i=1}^x \frac{ai}{ci} = P_0 \left(\frac{a}{c} \right)^x.$$

Da laut (14) $b/c = a/c = q$, so haben wir jetzt:

$$(18) \quad P_x = P_0 q^x, \quad x = 0, 1, \dots$$

$$\text{Aus } 1 = \sum P_x = P_0 \sum_{x=0}^{\infty} q^x = P_0 \frac{1}{1-q} = P_0 \frac{1}{p}, \text{ d. h. } P_0 = p \text{ und}$$

$$(19) \quad P_x = pq^x, \text{ für } x = 0, 1, \dots, \dots$$

was die Wahrscheinlichkeitsfunktion der geometrischen Verteilung ist.

Dieser Umstand würde aber dafür sprechen, daß Geburts- und Durchsetzungsrate gleich sind. Es läßt sich leicht zeigen, daß (19) mit $p = 1/3$ ($r = 1$) an die letzte der empirischen Verteilungen Goebels ($N = 500$; vgl. Goebel 1984: 73) mit einem $X^2 = 11.82$ bei 12 Freiheitsgraden und $P = 0.4606$ angepaßt werden kann. Wir nehmen aber an, daß dies kein allgemeiner, sondern ein sehr spezieller Fall ist (s. weiter unten).

(2) Auch wenn man nicht *zufällige* sondern *thematische* Stichproben aus der Datenquelle ALF zieht, liefert unser Modell befriedigende Resultate, soweit die erhobenen Themata (Begriffe) eine homogene Gruppe bilden. Für die thematischen Teilkorpora von Goebel (1984: 61–70) ergaben sich die Resultate wie in Tabelle 4 angegeben. Wir bringen hier nicht die vollständige Verteilungen mit der Anpassung, sondern nur das Thema, die Parameter und das Testresultat (X^2 , P).

Die zwei letzten Gruppen (Teilkorpora 20 und 21) zeigen keine Übereinstimmung mit unserem Modell, wohl wegen der großen Heterogenität der ihnen zugeteilten Begriffe. Es sind die zwei einzigen Verteilungen, die nicht monoton fallend verlaufen.

(3) Trotz dieser recht zufriedenstellenden Resultate im Falle der Variantenbildung im Bereich von *Nordwestfrankreich* (ausgehend von der Datenquelle ALF)

Korpus	$\hat{p}$	$\hat{f}$	X^2	FG	P
1. Zeit, Wetter, Jahreszeiten, Wochentage	0.2268	0.3190	6.26	6	0.39
2. Familie, Verwandtschaft, Eigennamen	0.4949	0.9800	3.59	3	0.31
3. Menschlicher Körper, Bekleidung	0.2859	0.6910	11.16	7	0.13
4. Krankheit, Heilung, Toilette, Schlaf	0.4245	0.9610	1.56	3	0.67
5. Affekte, Gefühle, Intelligenz, Perzeption, Sinne	0.1906	0.2120	3.99	3	0.26
6. Verkehr, Bewegung, religiöses und soziales Leben	0.3373	0.7240	5.84	7	0.56
7. Küche, Speisen, Essen, Trinken, Rauchen	0.1946	0.3539	9.59	7	0.21
8. Haus und Hof in Beschreibung und Tätigkeit	0.2947	0.8570	11.69	7	0.11
9. Handwerk, Erwerb, Beruf, Handel	0.2504	0.5550	3.89	6	0.69
10. Werkzeuge, Instrumente, Geräte	0.3312	0.8290	3.88	4	0.42
11. Boden, Gelände, Mineralien, Metalle	0.2437	0.5840	8.53	4	0.07
12. Feld, Wald und Flur und ihre Bewirtschaftung	0.2995	0.9600	2.90	6	0.82
13. Blumen und Pflanzen und ihre Pflege	0.3196	0.9840	3.89	5	0.57
14. Bäume, Sträucher und Früchte und ihre Pflege	0.3216	1.0380	8.95	6	0.18
15. (Haus-)Tiere und ihr Leben	0.2271	0.6060	8.40	9	0.49
16. Jagd, Fischerei	0.3776	0.9100	0.09	1	0.77
17. Adjektive, Numeralia	0.2951	0.2630	4.04	2	0.13
18. Adverbien, Präpositionen, Konjunktionen	0.5356	1.8310	9.70	5	0.08
19. Verben	0.4783	0.7910	3.33	3	0.34
20. Artikel, Pronomen, Substantiv	keine Übereinstimmung				
21. Syntaktisches	keine Übereinstimmung				

Zu den Originaldaten und den Modalitäten ihrer Erhebung vgl. Goebel 1984: 71–73.

Tabelle 4. Anpassung der negativen Binomialverteilung an die thematischen Teilkorpora für Normandie

kann man das hier präsentierte Modell nicht als die endgültige Lösung betrachten. Wenn man nämlich die von Goebel (1984: 52–54) für *Italien* beigebrachten Daten prüft, so sieht man sofort, daß sie anders verlaufen. Durch die Tatsache, daß in Italien 251 Meßpunkte gewählt wurden (dagegen für Nordwestfrankreich/Normandie: 71 Meßpunkte), ergibt sich automatisch eine größere Variation. Dadurch sinkt die Zahl der Karten, auf denen wenige Varianten aufscheinen, und es wächst die Anzahl der Karten mit vielen Varianten. Es steigt mit anderen Worten die onomasiologische Zersplitterung der einzelnen Atlaskarten an. Dies bedeutet, daß gegenüber früher (Nordwestfrankreich/Normandie) der Anfang der Verteilung niedriger liegt und sich der Schweif der Verteilung beträchtlich verlängert. Die negative Binomialverteilung ist dann nicht mehr imstande, diesen Verlauf der Verteilung zu modellieren. Zwar kann man zeigen, daß sie in einigen Fällen noch eine hinreichende Anpassung liefert, aber für eine allgemeine Lösung muß man die höhere Anzahl der Meßpunkte bereits beim Ansatz des stochastischen Prozesses in Betracht ziehen. Mit diesem Problem werden wir uns an anderer Stelle beschäftigen.

4. Methodologische Bemerkungen

Abschließend seien uns noch einige methodologische Bemerkungen gestattet. Goebel hat angedeutet (1984: 42), daß die von ihm beigebrachten Verteilungen „entfernt an die von Zipf [...] aufgezeigten Wortschatzgesetze erinnern“. Dieser intuitive Hinweis war korrekt: nicht nur, daß sie daran erinnern, sie sind durch diese Wortschatzgesetze entstanden. Denn das Modell der Diversifikation beruht expressis verbis auf den „Zipfschen Kräften“ (vgl. Altmann 1985). Im vorliegenden Beitrag stellen die Größen $\lambda_x = a + bx$ die Diversifikationskraft und $\mu_x = cx$ die Unifikationskraft in ihrer einfachsten linearen Form dar (siehe hier bei den Formeln 11 und 12). Sie sind allerdings nicht identisch mit physikalischen Kräften, wie auch eine Möglichkeit der Reduktion auf die Physik unserer Meinung nach auszuschließen ist (vgl. Ayala/Dobzhansky 1974). Derartige Kräfte sind aber überall in der Sprache vorhanden, so daß der Erfolg bei der Theoriebildung in der Sprachgeographie und überhaupt in der Linguistik davon abhängt, wie gut es uns gelingt, diese Kräfte in einem bestimmten Bereich zu erfassen und mathematisch auszudrücken. Zwar führte der sogenannte *Reduktionismus* in den Naturwissenschaften zu wichtigen Erkenntnissen. Doch, was den Bereich der Linguistik betrifft, glauben wir eher daran, daß Sprachprozesse *Spezialfälle* von *sehr allgemeinen Prozessen* sind, die sowohl in der Natur als auch in der menschlichen Gesellschaft verlaufen. Dies bedeutet allerdings nicht, daß wir an eine *monokephale* Einheit der Wissenschaften glauben, etwa dergestalt, daß sich alle Disziplinen durch eine einzige Disziplin „erklären“ lassen, wir glauben eher an eine *polykephal* organisierte Wissenschaftshierarchie, in der sich

die Probleme zweier oder mehrerer Disziplinen gemeinschaftlich aus übergeordneten abstrakteren Theorien erklären lassen.

Gesetze werden üblicherweise nach denjenigen Forschern benannt, die die zugrundeliegende empirische Regularität als erste beobachtet haben. Für die diatopische Spielart des im Bereich der Semantik entdeckten Diversifikationsgesetzes hat sich intern bereits die Bezeichnung „Goebelsches Gesetz“ eingebürgert.

Literaturnachweis

- [ALF] ALF: *Atlas linguistique de la France*. Ed. par Jules Gilliéron et Edmond Edmont. – Paris: Champion 1902–1910.
- [Altmann 1978] Altmann, Gabriel: *Towards a theory of language*. – In: Glottometrika 1. Hrsg. von G' A'. – Bochum: Brockmeyer 1978. S. 1–25.
- [Altmann 1985] –: *Semantische Diversifikation*. – In: Folia Linguistica 19 (1985), 177–200.
- [Altmann/Buttlar/Rott/Strauss 1983] Altmann, Gabriel/Buttlar, Haro von/Rott, Walter/Strauss, Udo: *A law of language change*. – In: Historical Linguistics. Hrsg. von Barron Brainerd. – Bochum: Brockmeyer 1983. S. 104–115.
- [Altmann/Schwibbe et al. 1985] Altmann, Gabriel/Schwibbe, Michael/Kaumanns, W./Köhler, Reinhard/Wilde, J.: *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*. [Erscheint 1985].
- [Ayala/Dobzhansky 1974] *Studies in the philosophy of biology. Reduction and related problems*. Ed. by F.J. Ayala and Th. Dobzhansky. – London: Macmillan 1974.
- [Beöthy/Altmann 1984a] Beöthy, Erzsébet/Altmann, Gabriel: *The diversification of meaning of Hungarian verbal prefixes. III. „el-“, „föl-“, „be-“*. – In Glottometrika 7. Hrsg. von U. Rothe. – Bochum: Brockmeyer 1984. S. 45–56.
- [Beöthy/Altmann 1984b] –: *The diversification of meaning of Hungarian verbal prefixes. II. „ki-“*. – In: Finnisch-Ugrische Mitteilungen 8 (1984), 29–37.
- [Best/Altmann/Kind 1985] Best, Karl-Heinz/Altmann, Gabriel/Kind, Bernd: *Eine Verallgemeinerung des Gesetzes der semantischen Diversifikation*. [Manusk.]
- [Best/Kohlhase 1983] *Exakte Sprachwandelforschung. Theoretische Beiträge, statistische Analysen und Arbeitsberichte*. Hrsg.: Karl-Heinz Best u. Jörg Kohlhasse. – Göttingen: edition herodot 1983.
- [Bunge 1969] Bunge, Mario: *Scientific research. I*. – Berlin: Springer 1969.
- [Bunge 1977] –: *Treatise on basic philosophy*. Vol. 3. – Dordrecht: Reidel 1977.
- [Goebel 1984] Goebel, Hans: *Dialektometrische Studien. I*. – Tübingen: Niemeyer 1984.
- [Guiter/Arapov 1982] *Studies on Zipf's law*. Ed. by Henri Guiter and Michail V. Arapov. – Bochum: Brockmeyer 1982.
- [Haight 1967] Haight, Frank A.: *Handbook of the Poisson distribution*. – New York: Wiley 1967.
- [Johnson/Kotz 1969] Johnson, Norman L./Kotz, Samuel: *Discrete distributions*. – Boston: Houghton Mifflin 1969.
- [Köhler/Altmann 1983] Köhler, Reinhard/Altmann, Gabriel: *Systemtheorie und Semiotik*. – In: Zeitschrift für Semiotik 5 (1983), 424–431.
- [Orlov/Boroda/Nadarejvili 1982] Orlov, Jurij K./Boroda, Moisei G./Nadarejvili, Isabella Š.: *Sprache, Text, Kunst. Quantitative Analysen*. – Bochum: Brockmeyer 1982.
- [Parzen 1967] Parzen, Emanuel: *Stochastic processes*. – San Francisco: Holden-Day 1967.
- [Rothe 1984] Rothe, Ursula: *Die Semantik des textuellen et*. – [Manusk.] Bochum 1984.

Eingereicht am 8.1.1985