QUANTITATIVE LINGUISTICS

Vol. 17

Mathematische Methoden in der historischen Linguistik

von

M. V. Arapov

M. M. Cherc

Übersetzt von

R. Köhler

P. Schmidt

Studienverlag Dr. N. Brockmeyer

Bochum 1983

QUANTITATIVE LINGUISTICS

CIP-Kurztitelaufnahme der Deutschen Bibliothek

Arapov, Michail V.:
Mathematische Methoden in der historischen Linguistik / von M. V. Arapov; M. M. Cherc. Übers. von R. Köhler; P. Schmidt. - Bochum: Studienverlag Brockmeyer, 1983.
(Quantitative Linguistics; Vol. 17)
Einheitssacht.: Matematiceskie metody v istoriceskoj lingvistike ‹dt.›
ISBN 3-88339-300-2
NE: Cherc, Maja M.:; GT

ISBN 3-88339-300-2

Querenburger Höhe 281, 4630 Bochum 1
Druck Thiebes GmbH & Co Kommanditgesellschaft Hagen

We would like to express our gratitude
to the

STIFTUNG VOLKSWAGENWERK

a generous grant from which made possible the
German edition of this book

I N H A L T

VORWORT

Dieser Arbeit ist folgendes einfaches Faktum zugrunde gelegt: Es existiert ein Zusammenhang zwischen der Häufigkeit eines Wortes und der Zeit seiner Entstehung in der Sprache. Es zeigt sich, daß die Mehrheit der Wörter mit großer Auftretenshäufigkeit von den sehr alten Wörtern gebildet wird; umgekehrt ist die Chance dafür, daß es sich bei einem Wort um einen Neologismus handelt, umso größer, je geringer seine Häufigkeit ist. Ordnet man die Wörter nach ihrer Häufigkeit in Texten der Gegenwartssprache, so nimmt der Anteil von Wörtern, die zu einem Zeitpunkt t in der Vergangenheit bereits in der Sprache existierten, je Häufigkeitsklasse mit sinkender Häufigkeit ab, und zwar umso stärker, je weiter t zurückliegt.

Hier ist zunächst anzumerken, daß das Bestehen eines statistischen Zusammenhangs zwischen Alter und Häufigkeit von Wörtern spätestens seit 1947 bekannt ist. Wohl zuerst wurde er von G. Zipf in seiner Arbeit (1) erwähnt, der darüberhinaus auch seine Bedeutung für die quantitative Analyse der Sprachgeschichte und -vorgeschichte unterstrich. Die Arbeit von Zipf fand, obwohl in mehreren Publikationen dargelegt, keine große Beachtung, vielleicht deshalb, weil er den gefundenen Zusammenhang nicht quantitativ ausdrücken konnte und seine praktische Anwendung nur sehr allgemein umriß.

Unsere Arbeit hat folgende zweifache Aufgabenstellung. Zum einen soll ein mathematisches Modell für die Veränderung des Lexikons aufgestellt und auf dessen Grundlage der Zusammenhang zwischen dem Alter eines Wortes und seinem Häufigkeitsrang theoretisch abgeleitet werden. Die Lösung dieser Aufgabe ist Inhalt von Kapitel 2.

Das theoretisch gefundene Gesetz muß noch empirisch überprüft werden, weil die von Zipf angeführten Daten weder quantitativ noch qualitativ befriedigen können. Dazu trägt Kapitel 3 bei, wo Material aus sechs Sprachen untersucht wird: Russisch und Tschechisch (Slavisch), Deutsch und Englisch (Germanisch), Französisch

und Rumänisch (Romanisch). Um nicht nur verschiedene Sprachen, sondern auch mehrere Häufigkeitswörterbücher für jede Sprache zu betrachten, wurden für das Russische, Französische, Deutsche und Englische je zwei Häufigkeitswörterbücher, für das Tschechische eines und für das Rumänische drei untersucht, wobei die letzteren allerdings nicht auf der Gegenwartssprache, sondern auf Texten des 16., 17. und 18. Jh. basieren.

In Kapitel 3 wird versucht, die Frage zu beantworten, wie man, vorausgesetzt man kennt die Verteilung der Wörter im Häufigkeitswörterbuch, die bis zu einem Zeitpunkt t in der Vergangenheit entstanden sind (z.B. bis zum Zerfall der Grundsprache), den absoluten Wert von t (in unserem Beispiel das Jahrhundert des Zerfalls der Grundsprache) bestimmen kann.

In den Kapiteln 2 und 3 wird also der erste Teil der Aufgabe gelöst und gewissermaßen ein Minimalprogramm realisiert, welches die mathematische Analyse der von Zipf entdeckten Gesetzmäßigkeit einschließt. Zur Realisierung dieses Programms gehört weiter die Auffindung einer Methode zur absoluten Datierung, mit deren Hilfe sich dieselben Probleme lösen lassen wie mit der glottochronologischen Methode von M. Swadesh. Dabei enthält das hier dargestellte, von uns Rangmethode genannte Verfahren in einem gewissen Sinn die Glottochronologie als Spezialfall.

Zum anderen stellten wir uns aber eine umfassendere Aufgabe. Die Anwendung einer quantitativen Methode, sei es der unten vorgeschlagenen, sei es der von Swadesh, auf die Sprachgeschichte erfordert eine klarere und konsequentere Abgrenzung zwischen Erb- und Lehnwörtern, Sprachverwandtschaft und Sprachinterferenz. Die Rangmethode wie auch die glottochronologische Methode sind ausschließlich auf die Datierung von älteren oder jüngeren Schichten des Erbwortschatzes anwendbar.

Allerdings kann, wenn durch etymologische Untersuchungen für ein Paar (oder eine Gruppe) von Sprachen ein gemeinsamer lexikalischer Fundus festgestellt wird, mit den existierenden Methoden der historischen Sprachwissenschaft nicht entschieden werden, ob dieser sich durch genetische Verwandtschaft erklärt oder das Resultat von Entlehnungen darstellt. (Das gilt besonders, wenn nur

sehr entfernte Verwandtschaft angenommen werden kann und die sprachwissenschaftlichen Argumente nicht durch historische Belege gestützt werden).

In dieser Arbeit wird ein eindeutiges Kriterium für Sprachverwandtschaft formuliert. Es zeigt sich nämlich, daß sich für die Begriffe der Verwandtschaft und der Entlehnung solche völlig natürlichen Explikationen angeben lassen, daß sich als Verwandtschaftskriterium eine spezielle Häufigkeitsverteilung der "gemeinsamen Wörter" (d.h. der Wörter einer Sprache, die eine Entsprechung in der anderen besitzen) ergibt: Wenn in beiden Lexika der Anteil der "gemeinsamen Wörter" mit sinkender Häufigkeit monoton abnimmt, so sind die Sprachen verwandt. Der Lehnwortschatz ist nach einem komplizierteren Gesetz verteilt; zwar wächst der Anteil der Lehnwörter mit abnehmender Häufigkeit, jedoch nur in einzelnen Abschnitten des Häufigkeitswörterbuchs.

In Kapitel 1 wird eine Reihe von Begriffen zu Sprachverwandtschaft und Lehnbeziehungen zwischen Sprachen eingeführt; außerdem werden zwei grundlegende sprachgeschichtliche Prozesse vorgestellt: die Divergenz, d.h. der Zerfall einer Grundsprache in mehrere Tochtersprachen, und die Konvergenz, d.h. die Annäherung von Sprachen. In diesem Kapitel versuchen wir auch, das Wesen der glottochronologischen Methode von Swadesh zu klären und ihre enge Beziehung zur Konzeption einer Gruppe verwandter Sprachen als Resultat des Zerfalls (Divergenz) ihres Vorläufers, der Grundsprache zu zeigen.

Die Kapitel 2 und 3, die bereits unter dem Gesichtspunkt des allgemeinen Forschungsprogramms angesprochen wurden, enthalten den Abriß einer Theorie des Zerfalls des Protolexikons, von dem unterschiedliche Fragmente in den Tochterlexika erhalten sind.

Kapitel 4 beschäftigt sich mit einer mathematischen Theorie des lexikalischen Zuwachses aufgrund von Entlehnungen und mit den Beziehungen, die bei der Entlehnung zwischen den beteiligten Sprachen bestehen.

In Kapitel 5 werden empirische Daten zur Stützung einiger Thesen aus Kapitel 4 aufgeführt.

Im letzten Kapitel (Kap. 6) wird versucht, das entwickelte Instrumentarium zur Messung der Produktivität verschiedener Ausdrucksmittel der Gegenwartssprache zu verwenden.

In den Schlußbemerkungen haben wir uns bemüht, ohne die Ergebnisse der einzelnen Kapitel zu wiederholen, die Hauptprobleme hervorzuheben, die in dieser Arbeit berührt werden.

VORWORT ZUR DEUTSCHEN AUSGABE

Seit dem Erscheinen der Originalausgabe dieses Buches sind nunmehr zehn Jahre vergangen. In der Zwischenzeit sind den Autoren einige Beschränkungen des vorgelegten Ansatzes und Unzulänglichkeiten in seiner Darstellung deutlich geworden, so daß bereits einige Verallgemeinerungen des in der Arbeit verwendeten formalen Apparats und präzisere numerische Resultate möglich wären. Die Autoren haben es jedoch vorgezogen, den Text nicht wesentlich zu verändern, sondern sich auf die Korrektur von Fehlern und einige zusätzliche Erläuterungen zu beschränken.

Der Grund ist darin zu sehen, daß mögliche Weiterentwicklungen ohne radikale Änderung der Gesamtkonzeption kaum zu einem wesentlich tieferen Verständnis des lexikalischen Wandels geführt hätten, da zur Formulierung einer umfassenderen Theorie zunächst wenigstens zwei grundsätzliche Probleme zu lösen sind.

Zum einen sind die existierenden Modelle in gewisser Hinsicht stationär; sie beschreiben nur die Veränderung der signifiants, d.h. des materiellen Aspekts des Lexikons bei gleichbleibender Struktur. Im Modell von M. Swadesh ist die Konstante Struktur die universale Liste der signifié's. In der vorliegenden Arbeit konnten die fundamentalen Gesetze des lexikalischen Wandels unabhängig von dieser Struktur formuliert werden. Die heute bekannten Zusammenhänge zwischen Bedeutung und Häufigkeit von Wörtern zeigen dagegen, daß unser Modell eine implizite Konstanzannahme in bezug auf die semantische Struktur enthält.

In Wirklichkeit ändert sich aber auch die semantische Struktur. Wenn man die Ergebnisse unserer Arbeit extrapoliert, zeigt sich, daß sich keine Sprache mehr als 7000 bis 8000 Jahre zurückverfolgen läßt, wenn man die Entstehungszeit der ältesten Texte als Ausgangspunkt nimmt. Innerhalb von 7000 bis 8000 Jahren wird das lexikalische Material vollständig ausgetauscht. Wenn sich aber parallel zum materiellen Wandel auch die semantische Struktur ändert, ergibt sich u.U. ein völlig anderer Zeitraum.

Zum anderen wird der Prozeß der Sprachspaltung in diesen Modellen des lexikalischen Wandels nur in stark vereinfachter Form berücksichtigt; sowohl dialektale und subdialektale Gliederung von Einzelsprachen als auch Parallelentwicklung verwandter Spra-

chen werden vernachlässigt. Insbesondere wäre es zur Weiterentwicklung der Theorie wichtig zu klären, inwieweit die Ränge (Häufigkeiten) einander entsprechender Wörter in den Lexika gleichzeitig existierender Dialekte einer Sprache und in den Lexika verschiedener Sprachen übereinstimmen. Erste Resultate hierzu zeigen bereits, daß die innere Differenzierung der Sprache in direktem Zusammenhang mit ihrer Entwicklung steht.

So wurde $\sqrt{i}$ - die Quadratwurzel des Rangs eines Wortes im Häufigkeitswörterbuch -, die in die Ausgangsformeln unserer Theorie eingeht und die Wahrscheinlichkeit angibt, mit der ein Wort des Rangs i beim lexikalischen Wandel erhalten bleibt, zunächst rein empirisch bestimmt. Bei der Untersuchung der Übereinstimmung von Lexika tritt diese Größe dagegen schon aufgrund rein theoretischer Überlegungen als Maß für die Streuung (quadratische Abweichung) der Ränge eines Wortes auf, dessen mittlerer Rang für eine Gruppe verwandter Lexika i beträgt.

An dieser Stelle sei Herrn R. Köhler und Herrn P. Schmidt besonders gedankt, die die deutsche Übersetzung des Buches besorgt haben. Mit ihrer freundlichen Unterstützung konnte eine Reihe von Ungenauigkeiten und Druckfehlern der Originalausgabe korrigiert werden.

26. März 1981

M.V. Arapov
M.M. Cherc

1. DIE LEXIKOSTATISTIK ALS SPRACHHISTORISCHER ANSATZ

1.1. FUNDAMENTALE RELATIONEN ZWISCHEN SPRACHEN

Traditionell spricht man von genealogischer Klassifikation und Verwandtschaft in bezug auf Relationen zwischen einzelnen Sprachen. Der Terminus Sprache wird in der historischen Sprachwissenschaft offensichtlich in mehreren Bedeutungen verwendet. So spricht man von der urslavischen Periode der russischen Sprache und von der urslavischen Sprache, von der russischen Literatursprache des XIX. Jh. und von den Besonderheiten der russischen Literatursprache im XIX. Jh., von der russischen Sprache als Menge von Dialekten und von den dialektalen Eigentümlichkeiten der Sprache eines bestimmten Sprachdenkmals. Es ist hier unmöglich, alle oder auch nur einen Teil der Begriffe zu explizieren, die sich hinter dem Terminus Sprache verbergen. Jedoch enthalten die existierenden Vorstellungen von Sprachverwandtschaft und von den Möglichkeiten der genetischen Klassifikation von Sprachen kaum spezifische Annahmen über die Natur ihrer Objekte, so daß bei der Klassifikation von Sprachen der zugrunde gelegte Sprachbegriff unwesentlich ist, solange es sich bei einer Sprache um ein diskretes Objekt x handelt, das von allen anderen Objekten dieser Art unterscheidbar ist, und ein bestimmtes Zeitintervall $t = [t_1, t_2]$ angegeben werden kann, in dem es existiert. Man kann davon ausgehen, daß die betrachteten Sprachen eine endliche Menge $L = \{x,y,z...\}$ bilden.

Im Folgenden diskutieren wir die Relationen, die auf der Menge L definiert sind. Vorweg sei jedoch bemerkt, daß wir keine Eigenschaft für die gesamte Sprache messen können, und daß sich alle hier dargelegten Ergebnisse nicht auf eine Sprache x, sondern nur auf ihr Lexikon V_x beziehen. Wir werden jedoch Argumente dafür anführen, daß die Beschreibung der lexikalischen Evolution näherungsweise die Evolution der Sprache insgesamt beschreibt, wenn auch rein intuitiv der Charakter des Wandels unterschiedlich ist: Unter Vernachlässigung der Beweglichkeit und Unschärfe zeitlicher und

räumlicher Sprachgrenzen fassen wir Sprachen als diskrete Objekte auf, während die Evolution des Lexikons ein kontinuierlicher Prozeß ist, in dem nur schwer Etappen abgegrenzt oder qualitative Sprünge bestimmt werden könnten. In Kap. 2 wird das Problem der Interpretation der Begriffe Vererbung und Entlehnung, Verwandtschaft und Nachfolge in Bezug auf den lexikalischen Wandel gelöst. Vorher ist jedoch ihre allgemeine Bedeutung zu klären, d.h. ihre Verwendung in Bezug auf Interrelationen zwischen Sprachen.

Wir beginnen mit der Relation, die für die historische Sprachwissenschaft gewissermaßen fundamental ist, obwohl sie dort keine spezielle Bezeichnung besitzt, nämlich {>}: x > y („die Sprache x ist mittelbarer Vorläufer der Sprache y"). Es geht um Fälle, wo eine Sprache x in gewissem Maße den zeitlich unmittelbar folgenden Zustand einer Sprache y bestimmt. Wichtig ist dabei nur, daß ein Einfluß, gleich welcher Stärke, von x auf y bestand, und daß dieser nicht durch eine weitere Sprache z vermittelt war. So ist der Zustand der russischen Gegenwartssprache nicht nur durch die russische Sprache des XIX. Jh. bestimmt, sondern z.B. auch durch den Einfluß der französischen und deutschen Sprache des XIX. Jh.

Die beiden folgenden Relationen - „ist Vorgänger von" und „ist Entlehnungsquelle von" - sind eher linguistische Konstrukte als Beobachtungsbegriffe.

Die Relation {→}: x → y wird gelesen als „Sprache x ist Vorgänger von Sprache y (bzw. Sprache y ist Nachfolger von Sprache x)". Vorgänger der russischen Literatursprache der Gegenwart ist die russische Literatursprache des XIX. Jh., Vorgänger des Neuhochdeutschen das Mittelhochdeutsche.

Die Relation der direkten Entlehnung {⊣}: x ⊣ y wird gelesen als „Sprache x beeinflußt Sprache y direkt (bzw. Sprache y entlehnt direkt aus Sprache x)". Als Beispiele seien hier genannt: der Einfluß des Kirchenslavischen auf das Altrussische, des Niederdeutschen auf das Hochdeutsche, des Lateinischen auf das Französische des XVI. Jh., des klassischen Mongolischen auf das Chalkamongolische, des Arabischen auf das Persische u.s.w. Einfluß ausüben können sowohl verwandte als auch nichtverwandte Sprachen,

und es kann aus der geschriebenen Sprache in die gesprochene entlehnt werden und umgekehrt.

Die gesamte historische Sprachwissenschaft beruht auf der genauen Unterscheidung von ererbten und entlehnten Elementen jeder Sprache, die es ermöglichen, von Verwandtschaft mit anderen Sprachen und von einer ununterbrochenen Entwicklungslinie bzw. nur vom Einfluß einer Sprache auf eine andere zu sprechen.

Die Begriffe „Verwandtschaft" und „Entlehnung" sind jedoch nicht präzise expliziert. Auch kann Verwandtschaft mehr oder weniger nah und Einfluß mehr oder weniger intensiv sein.

Unsere Aufgabe besteht also ganz allgemein darin, im Rahmen eines mathematischen Modells die Begriffe „Verwandtschaft" und „Entlehnung" genau zu definieren. Das Maß, das wir für die Verwandtschaft zweier Sprachen vorschlagen, hängt unmittelbar mit der Zeit zusammen, die die beiden Sprachen von der Protosprache als Phase ihrer größten Nähe trennt. Das Maß für die Stärke der Entlehnung aus einer anderen Sprache hängt etwas weniger direkt mit der Zeit zusammen.

Die quantitative Analyse des ererbten Fundus und der Entlehnungen kann noch ein weiteres Ziel haben: die Bestimmung eines dritten, möglicherweise höchst interessanten determinierenden Faktors des Sprachwandels, nämlich der Parallelentwicklung. Die Gestalt einer Sprache wird nicht nur bestimmt durch die Reproduktion ererbten Sprachmaterials in jeder Generation, sondern auch durch das unabhängige Wirken allgemeiner Prinzipien und Gesetze in jeder Sprache, die eine ganze Gruppe verwandter Sprachen charakterisieren. Die gleichermaßen typologische wie sprachgeschichtliche Frage nach der Rolle dieses Faktors kann hier nicht abschließend gelöst werden. Unsere kühnsten Hoffnungen bestehen darin, eine Methode zu finden, mit der man jenen „Rest" isolieren und quantitativ bestimmen kann, der weder durch Entlehnung noch durch Vererbung erklärbar ist. Danach kann man bereits überlegen, wie dieser Rest „aufgespalten" werden kann, durch Trennung von dem, was tatsächlich durch interne Sprachgesetze bedingt ist, und dem, was durch die Gemeinsamkeit von Lebensumständen und Kultur bedingt ist.

1.2. FUNDAMENTALE EIGENSCHAFTEN DER RELATIONEN ZWISCHEN SPRACHEN

Eine Sprache x, mit x > y, kann entweder direkter Vorgänger von y oder direkte Entlehnungsquelle für y sein:

$$\{>\} = \{\rightarrow\} \cup \{\dashv\}. \qquad (*)$$

Die Relation {>} auf der Menge L definieren wir als Präordnung, d.h. sie ist

a) reflexiv (x > x);
b) antisymmetrisch (aus x > y und x ≠ y folgt, daß y > x nicht gilt);
c) antitransitiv (aus x > y und y > z folgt, daß x > z nicht gilt).

Bemerkung: Aus der Definition der Relation {>} und (*) folgt, daß die Relation {→} und {⊣} a) antisymmetrisch und b) antitransitiv sind.

Wir setzen {→} als irreflexive und {⊣} als reflexive Relation an. Man könnte {→} und {⊣} auch anders definieren, in Abhängigkeit von der Behandlung der Innovationen, die in der Sprache selbst entstehen. Diese sind selbstverständlich keine Entlehnungen im gewöhnlichen Sinne; aus Gründen der Begriffsökonomie werden wir jedoch Innovation als Entlehnung einer Sprache x aus sich selbst behandeln (x ⊣ x).

Wir definieren nun den transitiven Abschluß {≥} der Relation {>}; x ≥ y wird gelesen als „die Sprache x ist Vorläufer der Sprache y". Die Relation x ≥ y ist erfüllt für ein Paar (x,y), wenn eine Kette $x_{i_1} > x_{i_2} > \ldots > x_{i_n}$ existiert mit $x_{i_1} = x$, $x_{i_n} = y$.

Analog definieren wir als transitiven Abschluß von {→} die Relation {⇒}, „ist Vorgänger von". Außerdem führen wir die Relation „ist Entlehnungsquelle" durch folgende Definition ein

$$\{\dashv\} \equiv (\{\dashv\} \cdot \{\Rightarrow\}) \vee \{\dashv\}.$$

D.H. z beeinflußt x, z ⊣ x (ist Entlehnungsquelle), wenn z x unmittelbar beeinflußt oder ein y existiert, so daß z ⊣ y und y ⇒ x. Wir betrachten also als Einfluß gleichermaßen die unmittelbare Entlehnung aus z nach x wie die Entlehnung aus z über einen Vorgänger s. Die Etymologen betrachten auch wesentlich kompliziertere Ketten, bei denen zwischen der Entlehnungsquelle z und der gegebenen Sprache x vermittelnde Sprachen liegen (z.B. Entlehnung aus dem Deutschen ins Russische über das Polnische). Wir werden jedoch hier solche Ketten der Einfachheit halber nicht berücksichtigen, sondern Entlehnung über das Polnische als Entlehnung aus dem Polnischen betrachten.

Die Relationen {≥} und {⇒} sind per definitionem transitiv und offensichtlich auch reflexiv.

Zusätzlich fordern wir, daß die Relationen {≥},{⇒} und {⊣} antisymmetrisch sind, was den natürlichen Eigenschaften der Zeit entspricht.

Bemerkung: Aus der Antisymmetrie und Transitivität von {⇒} bei Antireflexivität von {→} folgt die Antireflexivität von {⇒}.

Die Relationen {≥} und {⊣} sind daher partielle Ordnungen auf der Menge der Sprachen L, und die Relation {⇒} ist eine strikte partielle Ordnung auf derselben Menge (Antireflexivität!).

Bisher haben wir mit Hilfe der eingeführten Definitionen und ihrer offensichtlichen Folgerungen drei besondere Eigenschaften der Relationen zwischen Sprachen fixiert. Diese Besonderheiten werden in allen Arbeiten zur historischen Sprachwissenschaft angenommen, wenn auch nicht immer explizit.

1. Wechselwirkungen zwischen Sprachen können nicht momentan sein (die Relationen {→} und {⊣} sind Spezialfälle der Ordnung {>}). Aus diesem Prinzip folgt in letzter Konsequenz auch die Möglichkeit der absoluten Chronologie für Sprachen: Eine Änderung a in einer Sprache x relativ zu einer anderen Sprache y benötigt ein Zeitinterval $t_2 - t_1$, d.h. eine Eigenschaft der Sprache y zu t_2 muß bereits zu t_1 in x vorhanden gewesen sein. Von hier aus ist es nur ein Schritt zu der Annahme, daß die Länge des Intervalls proportional zur Veränderung a ist, und dahin, darüberhinaus eine Methode zur Messung von a zu suchen.

2. Die „jüngere" Sprache kann die ältere nicht beeinflussen, der Nachfolger nicht den Vorgänger (Antisymmetrie der Relation $\{\Rightarrow\}$).

3. Jede Beziehung zwischen Sprachen ist zurückführbar auf Vererbung oder Entlehnung.

1.3. DER BEGRIFF DES „SPRACHLICHEN STAMMBAUMS"

Wir wenden uns nun solchen Eigenschaften von Relationen zwischen Sprachen zu, die noch immer Gegenstand sprachwissenschaftlicher Diskussion sind. Betrachten wir zunächst die traditionellere und einfachere Auffassung. Obwohl wahrscheinlich zusammen mit der historisch-vergleichenden Methode entstanden, fand sie doch erst nach der Arbeit von A. SCHLEICHER (1862) weitere Verbreitung. Daher werden wir auch von „Schleichers Axiomen" sprechen (ihr Zusammenhang mit dem berüchtigten „Stammbaum" wird unten gezeigt).

A1. Für jedes y existiert genau ein x, derart daß $x \to y$.

A2. $\{\to\} \cap \{\dashv\} = \emptyset$.

A1 schließt Mischsprachen aus, d.h. Sprachen mit zwei oder mehr unmittelbaren Vorgängern. A2 behauptet, daß ein und dieselbe Sprache nicht zugleich unmittelbarer Vorgänger und unmittelbare Entlehnungsquelle sein kann (im allgemeinen kann aus der Vorgängersprache nicht unmittelbar entlehnt werden; es ist also sinnvoll, in einer Sprache ununterbrochen existierende Wörter, wie z.B. russisch *vorota* [Tor], und Wiederaufnahmen, wie z.B. russisch *vratar'* [Torwart], zu unterscheiden).

Aus A1 erhalten wir

FOLGERUNG 1. Für alle $x,y,z \in L$ folgt aus $y \Rightarrow x$ und $z \Rightarrow x$, daß $y \Rightarrow z$ oder $z \Rightarrow y$.

BEWEIS. Wir nehmen das Gegenteil an: y und z seien unvergleichbar. Wir betrachten die Ketten $I = u_{i_1} \to u_{i_2} \to \ldots \to u_{i_n}$ und $K = w_{k_1} \to w_{k_2} \to \ldots \to w_{k_m}$ mit $u_{i_1} = z$, $w_{k_1} = y$, $u_{i_n} = w_{k_m} = x$. Wir wählen nun aus ihnen u_{i_j} und w_{k_L} derart, daß $u_{i_j} = w_{k_L} = a$ und weder in I noch in K ein analoges Element mit niedrigerem Index als j bzw. 1 existiert. Dann enthalten I und K offensichtlich Elemente $u_{i_{j-1}}$ und $w_{k_{L-1}}$ mit $u_{i_{j-1}} \to a$ und $w_{K_{L-1}} \to a$ und $u_{i_{j-1}} \neq w_{K_{L-1}}$ im Widerspruch zu A1.

Wir definieren eine weitere wichtige abgeleitete Relation auf der Menge L: die Relation „Sprachverwandtschaft" $\{\sim\}$.

Die Sprachen $x,y \in L$ sind verwandt ($x \sim y$), genau dann wenn $x \Rightarrow y$ oder $y \Rightarrow x$ oder eine Sprache $z \in L$ existiert mit $z \Rightarrow x$ und $z \Rightarrow y$.

Aus dieser Definition erhalten wir direkt

SATZ 1. Die Relation $\{\sim\}$ ist eine Äquivalenzrelation auf der Menge L.

BEWEIS. Reflexivität und Symmetrie folgen unmittelbar aus der Definition. Wir beweisen die Transitivität. Sei $x \sim y$ und $y \sim z$, dann ist $x \sim z$ zu beweisen.

1. Wenn $x \Rightarrow z$ oder $x \Leftarrow z$, dann $x \sim z$ nach Definition.

Wir betrachten den Fall der Unvergleichbarkeit von x und y bezüglich $\{\Rightarrow\}$.

2. Nach Definition erhalten wir $u \Rightarrow x$ und $u \Rightarrow y$, $u' \Rightarrow y$ und $u' \Rightarrow z$. Über Folgerung 1 gilt $u \Rightarrow u'$ oder $u' \Rightarrow u$. Sei $u' \Rightarrow u$, dann $u' \Rightarrow x$ und $u' \Rightarrow z$. D.h. per definitionem $x \sim z$ (analog für $u \Rightarrow u'$). q.e.d.

BEMERKUNG. Aus der Definition der Relation $\{\sim\}$ und dem Beweis von Satz 1 geht hervor, daß die Menge $C \subseteq L$, die konnex ist bezüglich der Relation $\{\sim\}$, konnex ist bezüglich $\{\Rightarrow\}$ und $\{\to\}$.

$C \subseteq L$ sei eine endliche Äquivalenzklasse bezüglich $\{\sim\}$. Wir zeigen, daß $C = \{x_i\}$ ein Baum ist bezüglich $\{\to\}$.

SATZ 2.

1. Für alle $x_i \in C$ existiert höchstens ein $x_j \in C$ mit $x_j \to x_i$.

2. Es existiert genau ein x_o (die Wurzel des Baumes) derart, daß kein x_i mit $x_i \to x_o$ existiert.
3. Es existiert keine Folge der Art $x_{i_1} \to x_{i_2} \to \ldots \to x_{i_n}$ mit $x_{i_1} = x_{i_n}$ (Schleifenfreiheit).

BEWEIS. 2.1 ist eine Umformulierung von A1; 2.3 ist der oben geforderten Antisymmetrie von $\{\Rightarrow\}$ äquivalent. Bleibt 2.2 zu beweisen. Für beliebiges $x_i \in C$ betrachten wir die Menge X_i:

$$X_i = \{x_j | x_j \Rightarrow x_i\}.$$

Aufgrund der Endlichkeit von $X_i \subseteq C$ und der Schleifenfreiheit von C gibt es ein „kleinstes" x_{i_o}, zu dem kein x_j mit $x_j \to x_{i_o}$ existiert. Es bleibt zu zeigen, daß $x_{i_o} = x_o$ das einzige derartige Element in C ist.

Angenommen nämlich es existiert ein x_o' mit entsprechenden Eigenschaften, dann gilt $x_o' \sim x_o$. Folglich existiert ein x_o'' mit $x_o'' \Rightarrow x_o'$ und $x_o'' \Rightarrow x_o$. Dann ist aber $x_o'' \Rightarrow X_i$ und x_o nicht das kleinste Element in X_i, im Widerspruch zu seiner Definition. Also ist x_o das einzige derartige Element. q.e.d.

Satz 2 ermöglicht es, eine Menge verwandter Sprachen als Graph $\Gamma = \{C, \to\}$ darzustellen, dessen Knoten Elementen der Menge L - Sprachen - und dessen Kanten den direkten Vorgängerbeziehungen entsprechen. Nach Satz 2 ist ein solcher Graph ein gerichteter Baum. Wir werden daher die Ausdrücke „Schleichers Axiome" und „(Stamm-)Baumaxiome" synonym verwenden (Abb. 1).

Wir bemerken noch, daß Teilbäume von Γ Gruppen und Untergruppen verwandter Sprachen entsprechen. Wenn z.B. der gesamte Baum die indoeuropäische Sprachfamilie darstellt, dann entsprechen den Teilbäumen z.B. die Gruppe der slavischen und die Untergruppe der ostslavischen Sprachen, und den Wurzeln dieser Teilbäume die entsprechenden Protosprachen Urslavisch und Altrussisch.

Man kann daher zusätzlich eine Reihe stärkerer Äquivalenzrelationen $\{\sim_a\}$ auf C definieren: $x \sim_a y$ für $x,y,a \in L$, wenn x und y Nachfolger der Protosprache a sind ($a \Rightarrow x$, $a \Rightarrow y$). Die Relationen sind offensichtlich partiell geordnet bezüglich der Inklusion: je

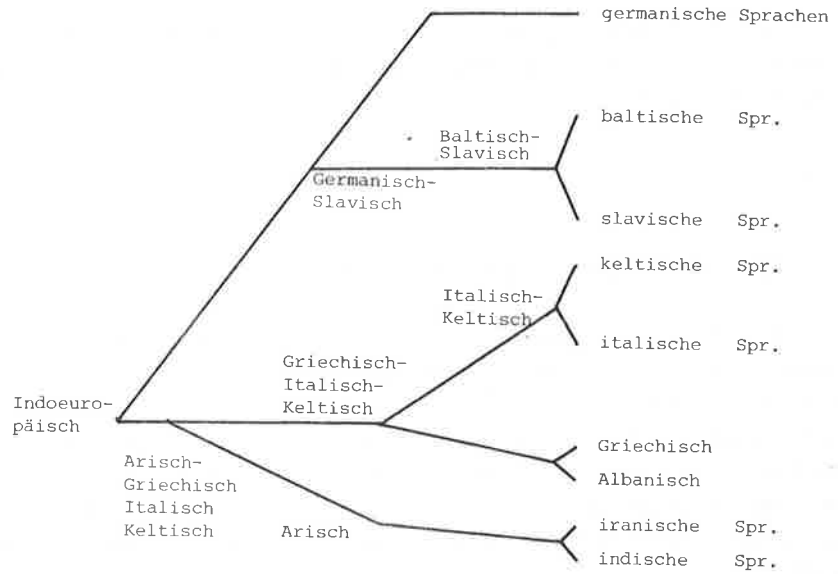

Abb. 1. „Baum der indoeuropäischen Sprachen von A. Schleicher (nach Porzig [3])

feiner die Zerlegung, die durch diese Relation induziert wird, desto „näher" die Verwandtschaft.

Außerdem können wir jeder Sprache x ein Zeitintervall $t_x = [t_1, t_2]$ ihrer Existenz zuordnen. Dann gilt für jeden Stammbaum: Wenn $x \Rightarrow y$, dann $t_x \leq t_y$ (wenn $t_y = [t_3, t_4]$ und $x \to y$, so $t_2 = t_3$). Vorgängersprache wird von Nachfolgersprachen abgelöst, wenn man von der Existenz „toter Sprachen" wie Latein, Sanskrit oder Altkirchenslavisch absieht, die einen Sonderstatus haben, weil sie als Literatur- oder Kultsprache weiterverwendet werden. Die Anwendbarkeit der gewonnenen Resultate auf solche konservierte Sprachen bleibt eine offene Frage.

Weiter definieren wir zwei Sprachen x und y als zu einer Sprachtradition X gehörig, wenn $x \Rightarrow y$ oder $y \Rightarrow x$. x und y aus X werden wir als Stadien der Sprachtradition bezeichnen. Wo dies nicht zu Mißverständnissen führen kann, werden wir die Sprachtradition einfach Sprache nennen.

Innerhalb einer Sprachtradition kann man Stadien durch Angabe ihrer Zeitkoordinaten unterscheiden, bezogen auf einen beliebigen Referenzzeitpunkt.

1.4. KRITIK DER „STAMMBAUM"-THEORIE

Seit ihrer Entstehung vor über einhundert Jahren ist die „Stammbaum"-Theorie immer wieder kritisiert worden. Dabei richtete sich diese Kritik oft nicht gegen die Theorie selbst, sondern gegen die Analogien, deren sich ihre Verfechter bei der Darstellung der Theorie bedienten (z.B. mit dem „Evolutionsbaum" in der darwinistischen Biologie). Heute wissen wir aber, daß „Baum"-Strukturen keine Besonderheit von Sprachwissenschaft oder Biologie sind.

Andere Kritikpunkte sind ernster zu nehmen; trotzdem ist die mit Einschränkungen und Ergänzungen versehene Stammbaumtheorie nach wie vor die Standardtheorie, wie sie in den wissenschaftlichen Lehrbüchern zu finden ist. Sie bestimmt den Rahmen für die Sammlung und die Klassifikation von Material zur Sprachgeschichte und -vorgeschichte, zur Dialektologie und Etymologie. Worin bestehen nun die Mängel dieser Theorie?

I. Mit „Baum"-Strukturen kann die Nähe von Sprachen nicht wiedergegeben werden. Um die Nähe der ostslavischen Sprachen zu erklären, postulieren wir nämlich die Existenz einer gemeinsamen ostslavischen Protosprache; damit können aber gerade nicht z.B. die ukrainisch-serbischen Parallelen erklärt werden (d.h. Phänomene, die nicht auf Neubildungen zurückgeführt werden können, aber nur in diesen beiden Sprachen auftreten). In der indoeuropäischen Sprachfamilie sind solche partiellen Beziehungen: italisch-keltisch, italisch-germanisch und germanisch-slavisch. In keinem dieser Fälle aber kann man die Existenz einer partiellen italokeltischen, italogermanischen oder germanoslavischen Protosprache postulieren. „In dem Maße, wie Fakten entdeckt wurden, entstanden unlösbare Pro-

bleme für die Forscher, die auf den alten Stammbaumschemata beharrten. Welche spezifischen Gemeinsamkeiten auch immer als Beweis für nahe Verwandtschaft anerkannt wurden, stets blieben andere, mit ihnen unvereinbare Fakten, die nur mit Hilfe eines völlig anderen Schemas erklärt werden konnten" [2:348].

Offenbar handelt es sich hier aber um <u>verschiedene Relationen</u>, die beide als „Verwandtschaftsbeziehung" bezeichnet werden. Die Relation $\{\sim_x\}$, die ja zwischen Sprachen besteht, welche von einer gemeinsamen Protosprache x abstammen, ist, wie oben gezeigt, eine Äquivalenzrelation. Prüfen wir daher also, ob dies auch auf die fraglichen Relationen zutrifft. Wir haben zwei Sprachen x und y nah genannt ($x \approx y$), wenn beide dieselben ererbten Eigentümlichkeiten aufweisen. Welche Eigentümlichkeiten „ererbt" sind, und was überhaupt als „Eigentümlichkeit" bezeichnet werden kann, ist dabei unwesentlich.

Die Relation $\{\approx\}$ ist offenbar reflexiv und symmetrisch, aber nicht transitiv. Wir können eine Gruppe gemeinsamer Merkmale von x und y finden ($x \approx y$), eine andere Gruppe für y und z ($y \approx z$), ohne daß es spezifische Gemeinsamkeiten von x und z gäbe. So sind z.B. die spezifisch italisch-slavischen Beziehungen bekannt und auch die spezifischen Beziehungen von Slavisch und Baltisch gut erforscht, doch sind keinerlei spezifisch italisch-baltische Beziehungen gefunden worden. $\{\approx\}$ ist also keine Äquivalenz-, sondern eine Toleranzrelation [3]. Im allgemeinen können Toleranzstrukturen nicht als Bäume dargestellt werden.

Wir müssen daher entweder die Stammbaum-Theorie durch irgendein formales Analogon wie z.B. die Wellentheorie von J. Schmidt ersetzen oder aber spezielle Eigenschaften der Relation $\{\approx\}$ suchen. Die Stammbaumtheorie kann man aber praktisch nur in einem Punkt modifizieren, nämlich durch Aufgabe von A1 (Einzigkeit des Vorgängers). Eine Theorie, die gleichzeitig Einzigkeit des Vorgängers und nicht-transitive Verwandtschaftsrelationen postuliert, wäre völlig sinnlos.

Welcher Art müßten diese speziellen Eigenschaften der Nähe-Relation $\{\approx\}$ sein, die sie mit der „Stammbaum"-Theorie kompatibel machen könnten? Eine notwendige (aber möglicherweise nicht hinrei-

chende) Bedingung wäre, daß die Sprachen nach Nähe geordnet werden könnten. Dann wäre Information über Nähe von Sprachen vergleichbar mit Information über ihre Herkunft. Solange das nicht der Fall ist, beziehen sich Hypothesen über „Nähe" und solche über „Verwandtschaft" schlicht auf verschiedene Fakten. Wenn wir für eine gegebene Untergruppe von Sprachen eine gemeinsame Protosprache postulieren, müssen wir nicht nur eine Menge von für sie spezifischen Merkmalen angeben, sondern auch eine Menge von Fakten, die in der Sprache ein derartiges „Gewicht" haben, daß wir gezwungen sind anzunehmen, daß die gemeinsame Entwicklung dieser Sprachen von wesentlich längerer Dauer war als die gemeinsame Entwicklung von Sprachen dieser Gruppe mit anderen Sprachen, die diese Merkmale nicht besitzen.

II. Die „Stammbaum"-Theorie postuliert die Entwicklung einer ganzen Reihe von Sprachen aus einer einheitlichen Protosprache. Ihre Kritiker weisen darauf hin, daß Versuche, die historischen Existenzbedingungen einer solchen Sprache zu rekonstruieren, oft zu der Vorstellung von einer Sprache mit großer räumlicher Ausdehnung und großer Sprecherzahl führen. Eine solche Sprache kann jedoch nicht in dem Sinne einheitlich gewesen sein, wie heutige Literatursprachen. Hier wird aber nicht so sehr die Stammbaumtheorie kritisiert wie die Unklarheit darüber, was die Einheit einer Sprache ausmacht. Es gibt keine einzige natürliche Sprache ohne interne dialektale Unterschiede. Doch die materiellen Unterschiede zwischen den Dialekten sind nicht entscheidend. Eine Gesamtheit von Dialekten bleibt eine einheitliche Sprache, solange diese sich übereinstimmend entwickeln. Umgekehrt charakterisiert gerade die Unabhängigkeit ihres Wandels die eigenständigen Sprachen.

III. Schließlich stellt sich die Frage nach der Konvergenz (Sprachmischung, Kreolisierung, Substrat u.ä.) in der Sprachgenese - im Gegensatz zur Divergenz, die in der Stammbaum-Theorie als einzig möglicher Weg vorgesehen ist.

Es ist unklar, ob immer dasselbe gemeint ist, wenn vom Mischcharakter z.B. der vietnamesischen oder albanischen Sprache, des Kajkavischen, einiger slovakischer, der Herkunft nach aber altpolnischer Dialekte, und andererseits vom supradialektalen und Mischcharakter vieler zeitgenössischer Literatursprachen die Rede ist. Möglicherweise ist keine andere Frage in der Sprachwissenschaft so heiß und mit solchem Apriorismus diskutiert worden wie die Frage der Sprachmischung. Diese läuft aber im wesentlichen auf die Frage hinaus, ob es in der Sprache selbst eine klare Grenze zwischen ererbtem und entlehntem Fundus gibt.

1.5. ERB- UND LEHNGUT: UNTERSCHEIDUNGSKRITERIEN

Bei philologisch erforschten Sprachen, d.h. bei solchen mit einer Schriftkultur mit einer großen Zahl datierter Schriftdenkmäler und einem gut erforschten kulturell-historischen Kontext können wir die Grenze zwischen ererbten und entlehnten Phänomenen leicht ziehen. In der Geschichte der slavischen Sprachen (speziell der russischen) treffen wir auf nur wenige Fälle, in denen Zweifel in dieser Hinsicht entstehen: z.B. für die urslavische Periode in bezug auf die slavisch-iranischen (wahrscheinlich slavisch-skythischen) Verbindungen, wo es fraglich ist, ob Wörter wie russisch *bog* [Gott], *sapog* [Stiefel], *pirog* [Kuchen], *tvorog* [Quark] und einige weitere entlehnt sind, in bezug auf die slavisch-germanischen Verbindungen (die Geschichte solcher Wörter wie russisch *tyn* [Pfahlzaun], *chlev* [Stall], *chleb* [Brot] etc.) und schließlich in bezug auf die slavisch-baltischen Verbindungen.

Im letzteren Fall steht die Entscheidung einer recht wichtigen Frage an: der gemeinsame slavisch-baltische Fundus ist so umfangreich, daß der wechselseitige Einfluß auf jeden Fall sehr lang und intensiv gewesen sein müßte. Hierzu gibt es zwei Interpretationen: die traditionellere, deren Verfechter behaupten, daß die slavischen und baltischen Sprachen auf eine gemeinsame baltoslavische Ursprache zurückgehen (s. Abb. 2a) [4], und die modernere, deren Vertreter das vorhandene Material als Indiz sekundärer Annäherung und intensiver kultureller Kontakte zwischen zwei Zweigen der indoeuropäischen Grundsprache werten (s. Abb. 2b) [5].

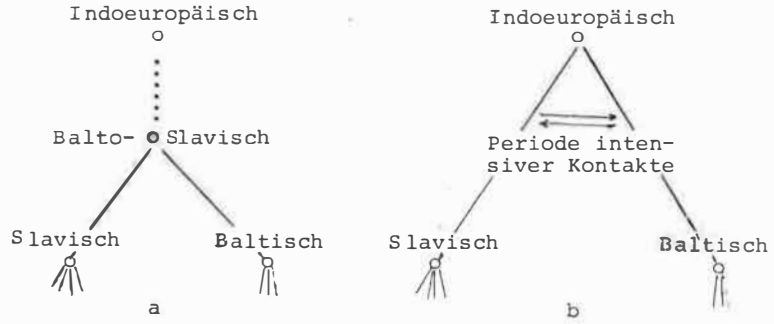

Abb. 2. Die beiden Interpretationen der slavisch-baltischen Beziehungen

a: Erklärung über eine gemeinsame balto-slavische Protosprache.

b: Erklärung über lange und intensive Kontakte in einer bestimmten Periode der eigenständigen Entwicklung dieser Sprachen

Natürlich hätte die Entscheidung dieser Frage unzweifelhaften Wert für die Geschichtswissenschaft und die Interpretation des archäologischen Materials.

Obwohl die Diskussion über die balto-slavischen Beziehungen seit Anfang des Jahrhunderts geführt wird, wird sie im Wesentlichen durch neue archäologische Fakten genährt; dagegen werden oft dieselben linguistischen Fakten als Argumente von beiden Seiten verwendet.

Das liegt daran, daß mit den Methoden der historisch-vergleichenden Sprachwissenschaft das Material einer Sprache x nur nach seiner Herkunft aus Sprachen x_1, x_2, x_3... klassifiziert werden kann, d.h. daß Beziehungen des Typs $x_1 > x$, $x_2 > x$, $x_3 > x$... aufgestellt werden können. Eine historisch-vergleichende Grammatik läßt sich für ein beliebiges Paar von Sprachen (x,y) konstruieren, wenn nur x Vorläufer von y ist, wobei die Art des Einflusses unwesentlich ist (ob z.B. y aus x entlehnt hat oder x Vorgänger von y ist). Darauf hat wahrscheinlich als erster N.S. Trubetzkoy hin-

gewiesen: „... zur Erklärung regelmäßiger Lautkorrespondenzen muß man durchaus nicht auf die Annahme einer gemeinsamen Herkunft von Sprachen einer Gruppe zurückgreifen, weil derartige Regelmäßigkeiten auch dann existieren, wenn eine Sprache aus einer nichtverwandten Wörter in großer Zahl entlehnt. So werden z.B. in den ältesten Entlehnungen der westfinnischen Sprachen aus dem (Ost-)Slavischen die slavischen stimmhaften Plosive b,d,g regelmäßig durch die finnischen kurzen stimmlosen p,t,k wiedergegeben; die slavischen stimmlosen Plosive p,t,k durch die finnischen langen (doppelten) Plosive pp,tt,kk; slavisch ѣ durch finnisch i; slavisch ъ durch finnisch u (im Wortauslaut nach slavischen stimmlosen Konsonanten aber durch finnisch i); slavisch o durch finnisch a; slavisch e durch finnisch ä usw. Übereinstimmung in rudimentären lexikalischen und morphologischen Elementen ist auch kein Beweis für die Abstammung von einer gemeinsamen Protosprache, weil im Prinzip alle Elemente der Sprache entlehnbar sind, auf niedrigen Entwicklungsstufen aber rudimentäre lexikalische Elemente besonders häufig von einer Sprache auf die andere übergehen. Seinerzeit hat Paul Kretschmer (in seiner „Einleitung in die Geschichte der griechischen Sprache") völlig zu recht behauptet, daß „zwischen den Begriffen Verwandtschaft und Entlehnung sprachwissenschaftlich nur ein chronologischer Unterschied besteht"[1] [6].

Den Versuch, eine konkrete „Grammatik der Entlehnungen" aus den westeuropäischen Sprachen ins Russische zu konstruieren, unternahm seinerzeit N.V. Jušmanov [7].

Je weniger uns der historische Kontext und die Genese der einzelnen Sprachen bekannt sind, desto vorsichtiger müssen wir mit Verwandtschaftshypothesen für heute weit voneinander entfernte Sprachen sein. Für die altaische Hypothese, für die Hypothese der indoeuropäisch-semitischen Verwandtschaft, ganz zu schweigen von der nostratischen Theorie, ist die Unterscheidung von ererbten und entlehnten Phänomenen entscheidend. Wenn sich einzelne Entlehnungen prinzipiell in keiner Weise von alten Formen unterscheiden, dann muß man die Frage stellen, ob sich nicht die Menge der Entlehnungen von der Menge der ererbten Formen irgendwie unterscheidet.

Wir beantworten diese Frage positiv: solche Unterschiede existieren, und unter bestimmten Bedingungen kann man sie quantitativ erfassen. Dafür brauchen wir vor allem eine relativ homogene Menge von Fakten; als eine solche wählen wir das Lexikon.

In welchem Maße die aufgestellten lexikalischen Beziehungen zwischen Sprachen auf Sprachen insgesamt übertragen werden können, bleibt offen. Das große „Gewicht" des Lexikons in der Sprache wird indirekt gestützt durch das Fehlen von Beispielen gesicherter Beziehungen zwischen Sprachen in Phonetik, Morphologie oder Syntax ohne entsprechende lexikalische Korrespondenzen. Darüberhinaus sprechen alle bekannten Beispiele dafür, daß eine Sprache nichtlexikalische Eigenschaften erst auf der Grundlage einer ausreichenden (gewöhnlich sehr großen) Anzahl lexikalischer Entlehnungen entlehnen kann.

Für die Untersuchung der Lexik sprechen auch praktische Erwägungen:

- das Wort als selbständige Einheit gibt es praktisch in allen Sprachen
- die etymologische Analyse wird vor allem am Wort durchgeführt
- es hat sich bereits ein gewisses Maß an lexikostatistischen Erkenntnissen angesammelt (wenn auch leider die notwendigen Daten nicht für alle Sprachen vorliegen).

Für die Beschreibung des Verhaltens einer Wortmenge in der Zeit wurde Anfang der 50er Jahre von dem amerikanischen Forscher M. Swadesh die als Glottochronologie bekannte Theorie aufgestellt. Im folgenden Paragraphen werden wir die durch sie erzielten Fortschritte zusammenfassend darstellen.

1.6. GRUNDANNAHMEN DER GLOTTOCHRONOLOGIE SWADESHS

Die Grundannahmen der Glottochronologie nach Swadesh lassen sich auf fünf Postulate zurückführen, von denen wir zunächst vier betrachten.

1° Im Lexikon jeder Sprache kann man einen speziellen Anteil isolieren, der im Folgenden Kern oder stabiler Teil heißen soll.

2° Man kann eine Liste von Bedeutungen angeben, die in jeder Sprache obligatorisch durch Wörter des Kerns ausgedrückt werden (z.B. die Begriffe „und", „alle", „Lebewesen", „Rücken" u.ä.). Wir werden die Menge dieser Wörter Kernliste (KL) nennen und ihre Anzahl mit N_o bezeichnen.

3° Der Anteil p von Wörtern der KL, die über ein Zeitintervall Δt (z.B. ein Jahr, Jh., Jahrtausend) erhalten bleiben (nicht durch andere Wörter ersetzt werden), ist konstant (d.h. sie hängt nur von der Größe des gewählten Intervalls ab, nicht jedoch von der Art ihrer Auswahl oder von der betrachteten Sprache).

4° Die Wahrscheinlichkeit, über das Intervall Δt erhalten zu bleiben (bzw. nicht erhalten zu bleiben, zu „zerfallen"), ist für alle Wörter der KL gleich groß.

Die Postulate 3° und 4° geben uns lediglich an, was bei jedem einzelnen „Schritt" des lexikalischen Veränderungsprozesses geschieht, wobei ein solcher Prozeß als aus vielen relativ kleinen Schritten zusammengesetzt zu denken ist.

Fixieren wir nun einen Zeitpunkt t_o und die Wörter, durch welche unsere „stabilen" Begriffe zu t_o ausgedrückt werden. Das ergibt die Ausgangsliste KL. Nach Voraussetzung ist der Anteil der Wörter der KL, die ein Zeitintervall Δt „überlebt" haben und in die KL zu $t_o + \Delta t$ eingegangen sind, gleich p (die Zahl der Wörter des Mindestalters Δt beträgt $N_o p$). Die übrigen Wörter sind irgendwelche Neubildungen, die in der Ausgangsliste nicht enthalten waren, und deren Geschichte hier nicht interessieren soll.[2)]

Wir nehmen jetzt die Wörter der neuen KL. Aus 3^{o} wissen wir, daß aus der neuen Liste wieder der p-te Teil ein weiteres Intervall Δt überleben wird. Einige dieser Wörter werden „alte" sein, d.h. noch zur Ausgangsliste gehören, andere dagegen „neue", d.h. im ersten Schritt in die Liste gekommen sein.

Jetzt „wirkt" Postulat 4^{o}. Denn es besagt ja, daß die Sprache nicht zwischen „alten" und „neuen" Wörtern unterscheidet; für beide sind die Chancen gleich, für die Dauer eines weiteren Intervalls im Lexikon zu verbleiben. Ihre Proportion hängt nur von ihrer Proportion zu Beginn des jeweiligen Prozeßschrittes ab. Der Anteil „neuer" war p und wird $p \cdot p = p^2$. Der Anteil „alter" war 1-p - er wird also (1-p)p. Die Wörter der Ausgangsliste, die nach dem zweiten Prozeßschritt erhalten bleiben, werden das Mindestalter $2\Delta t$ haben.

Im dritten Schritt beträgt der Anteil der Wörter aus der Ausgangsliste p^3 usw. Es zeigt sich also, daß der Anteil der Wörter der Ausgangsliste, die nach jedem folgenden Prozeßschritt erhalten sind, geometrisch wächst: 1, p, p^2, $p^3 \ldots p^k \ldots$. Die Exponenten entsprechen hier den Schritten des Prozesses, d.h. der Zahl der Standardzeitintervalle seit t_o.

Der lexikalische Veränderungsprozeß läßt sich jetzt allgemeiner fassen, nämlich als kontinuierlicher (nicht diskreter) Prozeß in Abhängigkeit von der Zeit t. Die geometrische Reihe muß dann durch ihr „kontinuierliches Analogon", die Exponentialfunktion, ersetzt werden.

Bezeichnen wir den Anteil der Wörter, die zu t erhalten sind, mit N(t), so gilt:

$$N(t) = e^{-\eta t}. \qquad (1.1)$$

Der Koeffizient im Exponenten hängt nicht von der „Schrittlänge" ab:

$$\eta = \frac{\ln p}{\Delta t}.$$

Die Zahl der zu t erhaltenen Wörter beträgt damit

$$N(t) = N_o e^{-\eta t}. \qquad (1.2)$$

Die Funktionen (1.1) und (1.2) sind fundamental für die Glottochronologie, die übrigen lassen sich leicht aus ihnen ableiten[3].

Die Formel (1.2) gibt an, wie viele Wörter der zu t_o fixierten Liste bis t überleben (z.B. mit 1974 als t_o). Zu t haben diese Wörter ein Mindestalter von $t-t_o$ Jahren. Wir sprechen von „Mindestalter" deshalb, weil sie auch schon vor t_o entstanden sein können. Denn zu t_o beginnt nicht die Sprachgeschichte, sondern unsere Betrachtung.

Formel (1.2) läßt zwei Interpretationen zu: die „prospektive": die Zahl der Wörter, die in der KL einer Sprache nach t Jahren ihrer Entwicklung noch erhalten sind, nimmt exponentiell ab; und die „retrospektive": die Zahl der t Jahre alten Wörter der KL nimmt mit t exponentiell ab.

Gleichzeitig ist Formel (1.2) völlig invariant gegenüber der Zeitrichtung. Sie besagt nur, daß die Anzahl identischer Wörter in zwei KL, die zu verschiedenen Zeitpunkten der Entwicklung einer Sprache erstellt werden, exponentiell mit dem entsprechenden Zeitintervall abnimmt.

Auf diese Weise hat der eingangs behauptete Zusammenhang zwischen dem Alter von Wörtern und der Anzahl gleichalter Wörter im Lexikon in der Theorie der Glottochronologie eine präzise Explikation gefunden. Dabei mußte Swadesh allerdings ihre Allgemeinheit preisgeben: die gefundene Gesetzmäßigkeit bezieht sich nur auf einen kleinen Lexikonausschnitt. Das sollte aber nicht verwundern; denn wir haben den Begriff „Lexikon" äußerst unscharf verwendet („Lexik einer gegebenen Sprache"). In dieser Arbeit werden wir versuchen, den Begriff des Lexikons so zu präzisieren, daß obige Resultate in gewissem Sinn auf das gesamte Lexikon übertragbar werden.

Um das o.g. Gesetz praktisch anwenden zu können, muß man das vergangene Entwicklungsstadium der betrachteten Sprache soweit

kennen, daß man für dieses wenigstens eine KL erstellen kann. Aber wenn das uns interessierende Stadium zeitlich sehr weit zurückliegt, oder die Sprachgeschichte ungenügend bekannt ist, kann die direkte Erstellung einer solchen Liste Schwierigkeiten bereiten. Es gibt jedoch eine Situation, in der man über indirekte Daten über die KL verfügt. Das ist der Fall, wenn das uns interessierende Entwicklungsstadium der Sprache die gemeinsame Protosprache einer Gruppe verwandter Sprachen ist, in deren Lexika verschiedene Gruppen von Wörtern der KL der Protosprache erhalten sind. Offenbar fällt die Zahl der übereinstimmenden, erhaltenen Wörter dieser Liste in den Nachfolgersprachen mit dem Anwachsen des Zeitintervalls seit dem Zerfall der Protosprache zusammen.

Zur Überführung dieser natürlichen Betrachtung in ein quantitatives Gesetz wird folgende Zusatzhypothese angenommen.

5° Die Überlebenswahrscheinlichkeiten eines Wortes aus einer Protosprachen-KL in den KL der Nachfolgersprachen sind voneinander unabhängig.

Es wird angenommen, daß sich vom Moment des Zerfalls der Protosprache an die Lexika ihrer Nachfolgersprachen völlig unabhängig voneinander bilden. Wenn also die Wahrscheinlichkeit eines Wortes, aus einer Protosprache, die t Jahre zuvor zerfallen ist, in die KL einer ihrer Nachfolgersprachen zu gelangen, wie oben gezeigt gleich N(t) ist, dann ist die Wahrscheinlichkeit, in die KL zweier Nachfolgersprachen zu gelangen,

$$N_2(t) = N(t)N(t) = (e^{-\eta t})^2 = e^{-2\eta t}. \qquad (1.3)$$

Die Zahl der übereinstimmenden Wörter in den KL zweier verwandter Sprachen ist dann

$$N_2(t) = N_o e^{-2\eta t}.$$

Analog läßt sich zeigen, daß die Wahrscheinlichkeit eines Wortes,

in die KL von n verwandten Sprachen zu gelangen, noch geringer ist:

$$N_n(t) = e^{-n\eta t}.$$

So können wir, wenn wir die Anzahl der gemeinsamen Wörter in den KL von zwei (oder n) Sprachen kennen, im Prinzip unter Benutzung von (1.3) die Zeit t bestimmen, die seit dem Stadium ihrer Trennung vergangen ist.

Die praktische Anwendung der Formeln (1.2) und (1.3) setzt die Kenntnis des Wertes von η (oder der Erhaltungswahrscheinlichkeit für irgendeine Standardperiode) voraus.

Durch Operieren mit Swadeshs Liste von 215, später 200 und schließlich 100 Bedeutungen und durch den Vergleich der Sprachen, „die auf ein und derselben Linie der Sprachtradition liegen" [8:39] (die romanischen Sprachen mit dem Lateinischen, Altenglisch mit Neuenglisch, Althochdeutsch mit der heutigen deutschen Literatursprache u.ä.) versuchten die Forscher mit Hilfe von (1.2) η empirisch zu bestimmen (die Zeit t ist aus den historischen Daten bekannt). Diese Ergebnisse wurden bereits in der ersten Hälfte der 50er Jahre erzielt.

Später wurde der statistische Charakter der glottochronologischen Methode klarer erkannt, und eine Schätzung des relativen Datierungsfehlers erschien [10]. Zur Bearbeitung der empirischen Daten wurden hinreichend verfeinerte statistische Methoden benutzt.

Die Argumente gegen die Methode von Swadesh laufen letztlich auf Einwände gegen die fünf oben formulierten Postulate hinaus.

Das Auswahlkriterium für die Begriffe der KL ist kein linguistisches; es stützt sich auf Daten der vergleichenden Ethnographie. Daher werden wir es hier nicht weiter erörtern, sondern wir merken lediglich an, daß nach dem gegenwärtigen Entwicklungsstand dieser Wissenschaft keine Menge von Begriffen existiert, die für alle Kulturen - einschließlich der sehr alten - gleich wichtig sind.

Wie schon die Geschichte dieser Begriffsliste, in deren Verlauf einige Begriffe neu aufgenommen wurden, andere wegfielen, und auch der Umfang der Liste verändert wurde, zeigt, ist das ver-

wendete Auswahlkriterium weder einfach noch eindeutig. Die von Swadesh aufgestellte Liste von Bedeutungen läßt sich gewöhnlich nicht eindeutig dem Kern eines konkreten Lexikons zuordnen. Mit der Zeit kam eine ganze Rezeptsammlung zur Minimierung von Übersetzungsunschärfen zusammen (s. [8,9]), doch die dabei auftretenden Komplikationen sind nur ein Hindernis bei der Erweiterung der KL auf mehr als 100 bis 200 Wörter. Bei derart kleinen Stichproben ist der Stichprobenfehler entsprechend groß. Zahlreiche diesbezügliche kritische Anmerkungen finden sich in den Aufsätzen Hoijer [13], Chretien [14] und Levin [15] u.a.[4].

Wohl die meisten Einwände ruft das Postulat 3^{o} hervor, welches eine konstante Veränderungsrate der KL behauptet, die weder von der Zeit noch von der Sprache abhängen soll. Manchmal beruhen solche Einwände auf apriorischen Annahmen über die Ungleichmäßigkeit der Sprachentwicklung. Es gibt aber auch Arbeiten, in denen Fakten angeführt werden, die zur Konstanzhypothese im direkten Widerspruch stehen. Hier müssen Resultate von I. Fodor für die slavischen Sprachen [16] und von K. Bergsland und H. Vogt für Riksmål und Isländisch [17] erwähnt werden.

In beiden Arbeiten - Fodor und Bergsland/Vogt - wurde die gleiche Methode verwendet: zunächst wurden die Gegenwartssprachen mit ihren Vorläufern (im ersten Fall mit dem Urslavischen, im zweiten Fall mit dem Altnorwegischen) und danach paarweise untereinander verglichen. Dabei führten nicht nur verschieden große Kernlisten (215-100 Wörter) zu verschiedenen Ergebnissen; selbst die besten (d.h. auf den umfangreichsten Listen basierenden) Resultate unterschieden sich signifikant von den historisch etablierten Datierungen. So beträgt nach Fodor der lexikostatistische Zeitabstand zwischen heutigem Russisch und Urslavisch 550 - 700 Jahre, der für Tschechisch 500 - 850 Jahre und der für Weissrussisch und Polnisch 750 - 1300 bzw. 750 - 1100 jahre. Das historisch wahrscheinlichste Datum für den Zerfall des Gemeinslavischen ist das Ende des VI. Jh. n. Chr., d.h. alle slavischen Sprachen müssen ungefähr 1400 Jahre alt sein.

Noch weiter nach unten wichen die Ergebnisse beim paarweisen Vergleich der slavischen Sprachen ab: für Russisch und Tschechisch

betrug der errechnete Zeitraum getrennter Entwicklung 400 - 550 Jahre, für Tschechisch und Polnisch 200 - 400 Jahre. Daher bleiben, auch wenn man für die slavischen Sprachen eine spezielle, von Swadesh abweichende, Zerfallskonstante annimmt, immer noch die Divergenzen zwischen den Ergebnissen des Vergleichs der Einzelsprachen mit der Ursprache und denen des paarweisen Vergleichs zwischen den Einzelsprachen. Wahrscheinlich gibt es dafür nur die eine Erklärungsmöglichkeit, Parallelentwicklung der Einzelsprachen nach dem Zerfall der urslavischen Einheit anzunehmen. Aber der Glottochronologie fehlen die formalen Mittel für die notwendigen rechnerischen Korrekturen.

Eine noch stärker verzögerte Entwicklung ergibt sich für das heutige Isländische und seinen Vorläufer, das Altnorwegische, wo für das erstere eine eigenständige Entwicklungszeit von 63 - 194 Jahren bestimmt wird. Dagegen liefert der Vergleich des Altnorwegischen mit einem anderen Nachfolger, dem Riksmål (einer Form der norwegischen Literatursprache der Gegenwart), einen etwas zu hohen Wert: 1347 - 1461 Jahre (bei tatsächlich ca. 1000 Jahren), und der Vergleich mit zwei heutigen norwegischen Dialekten einen Wert von 800 - 900 Jahren.

Die Schlußfolgerung, daß jede Sprache ihre eigene Zerfallskonstante besitzt, stützen Bergsland und Vogt durch ein Experiment mit Sprachen einer völlig anderen Gruppe und eines völlig anderen Kulturkreises. Sie vergleichen einerseits das heutige Georgisch mit dem Altgeorgischen und mit dem heutigen Mingrelischen und andererseits das Neuarmenische mit dem Altarmenischen. Beide alten Sprachen haben eine umfangreiche schriftliche Tradition und Denkmäler bis ins 5. Jh. n. Chr. zurück; deshalb kann man die sprachlichen Ausgangsdaten exakt datieren.

Die lexikostatistischen Datierungen weichen wieder nach unten ab: 338 - 861 Jahre beim Vergleich von Alt- und Neugeorgisch (historisch 1500 Jahre) und 211 - 437 Jahre beim Vergleich von Neu- und Altarmenisch (historisch ebenfalls 1500 Jahre). Die Periode der georgisch-mingrelischen Einheit wird glottochronologisch auf vor 1000 - 1300 Jahren bestimmt, während diese Periode in Wirklichkeit spätestens im ersten vorchristlichen Jahrtausend endete[5].

Manche Einwände gegen die Annahme einer universellen Erhaltungskonstante in der Glottochronologie richten sich nicht gegen Postulat 3^{O}, sondern gegen Postulat 4^{O}.

Denken wir uns die KL in k Teillisten aufgespalten. Gilt dann immer die Gleichung $\eta_i = \eta_j$, $1 \leq i \leq j \leq k$, mit η_i als Änderungsrate der i-ten Teilliste?

Aus der lexikostatistischen Praxis ist bekannt, daß diese Gleichung in der Regel nicht gilt: in der KL gibt es lang- und kurzlebige Elemente. Wenn man eine Gruppe verwandter Sprachen mit gemeinsamer Protosprache betrachtet, so finden sich dabei langlebige Wörter in mehr Sprachen als bei gleicher Zerfallswahrscheinlichkeit für alle Wörter zu erwarten wäre.

Das wurde in der Arbeit von Dyen, James u.a. benutzt, wo auf der Grundlage einer Analyse von 46 KL-Paaren für verwandte australische Eingeborenensprachen die 200 Begriffe aus Swadeshs Liste nach ihrer Lebensdauer geordnet wurden [11].[6)]

Die Existenz von unterschiedlichen Änderungsgeschwindigkeiten bei den Wörtergruppen der KL kann (wenn die Gruppen und die Geschwindigkeiten bekannt sind) nach einem Vorschlag von N. van der Merwe [19] zur Verringerung der Stichprobenfehler durch die gegenseitige Überprüfung der Teilstichproben genutzt werden. Dann wird die Veränderung der KL nicht durch einen Exponenten, sondern durch eine Summe von Exponenten beschrieben.[7)]

Das letzte, fünfte Postulat der Glottochronologie ist weniger kritisiert worden, obwohl man Fodors empirische Daten zu den slavischen Sprachen als mit dieser Hypothese unverträglich ansehen kann.

Wahrscheinlich lassen sich diese Abweichungen nur auf eine Weise erklären, nämlich durch die Annahme von Parallelentwicklung nach dem Zerfall des Gemeinslavischen.

Die Intensität des Sprachkontakts selbst kann sich ändern, insbesondere können sich Sprachen einander annähern.[8)] Die Glottochronologie basiert jedoch auf einem sprachlichen Evolutionsmodell, in dem Unterschiede zwischen Sprachen nur wachsen können.

Außerdem muß beim Vergleich der KL zweier Sprachen von vornherein ihre Verwandtschaft gesichert sein. Swadeshs Methode liefert nämlich kein Kriterium zur Unterscheidung von entfernter Verwandtschaft und Entlehnung. Daher mußten wir uns speziell mit der Schaffung eines Unterscheidungsverfahrens für ererbte und entlehnte Elemente befassen.

1.7. MÖGLICHE ALTERNATIVEN ZUR GLOTTOCHRONOLOGIE

Die Glottochronologie liefert eine Methode zur Bestimmung des „Anteils" einer Sprache x an einer Sprache y mit $x \geq y$. Es wird angenommen, daß dieser Anteil umso geringer ist, je länger die Kette

$$z_1 > z_2 > \ldots > z_n,$$

wobei $x = z_1$ und $y = z_n$, d.h. je größer das Zeitintervall zwischen x und y ist.

Die dabei verwendete Methode zur Berechnung dieses „Anteils" stellt einen bedeutenden Fortschritt dar im Vergleich zu den in der philologischen Praxis immer noch vorherrschenden Verfahren. Dort wird der Anteil der „positiven" Belege aus einer zufälligen Menge bestimmt, nach deren Repräsentativität für die Sprache nicht einmal gefragt wird. So wird z.B. der Anteil von Kirchenslavischem in der altrussischen Lexik unter der stillschweigenden Voraussetzung berechnet, daß das Wörterbuch von I. Sreznevskij mit ihr identisch ist.

In der Glottochronologie wird dagegen das Lexikon in zwei Teile aufgespalten, wobei alle Ergebnisse nur für den ersten Teil - den Kern - gelten. Der Kern ist klein im Vergleich zum Gesamtlexikon, es wird aber angenommen, daß das Gesamtgewicht dieser Wörter sie für das gesamte Lexikon repräsentativ macht.

Um aber eine Liste aus N Wörtern zu erstellen, müssen wir N unabhängige Entscheidungen fällen. Auch wenn jede einzelne durch komplexe Argumentation gestützt wird, die auf der ganzen Erfahrung und Intuition des Sprachwissenschaftlers beruht, ist doch

die Fehlerwahrscheinlichkeit hoch. Sie erhöht sich noch durch die Zweistufigkeit des KL-Bildungsprozesses: Aufnahme einer Bedeutung in die Liste und darauffolgende Wahl eines entsprechenden Ausdrucks der betrachteten Sprache. Aber wie eine große Liste unzuverlässig ist, so ist eine kleine aus statistischen Erwägungen inakzeptabel.

Das Gewicht (der Rang) eines Wortes ist nicht aufgrund apriorischer Erwägungen zur „universalmenschlichen" Wichtigkeit eines Begriffes definierbar.

Die Dichotomie Kern - übriges Lexikon ist inadäquat, weil die Erfahrung zeigt, daß innerhalb dieser beiden Klassen viele Abstufungen möglich sind.

Statt dessen sollte

a) der Rang eines Wortes seinen Status in der jeweils betrachteten Sprache reflektieren;
b) die Rangskala ordinal sein - im Grenzfall kontinuierlich;
c) der Rang eines Wortes nicht auf intuitiver Basis, sondern aufgrund einer exakten Prozedur bestimmt werden; die Beliebigkeit der Entscheidungen wird dadurch auf die Wahl einer Rangierungsmethode beschränkt.

Eine all diesen Forderungen genügende Rangierungsmethode ist die Ordnung der Wörter nach Häufigkeit.

Sie ist sicher keine ideale Methode; sie legt unter anderem keine eindeutige Hierarchie für die Wörter einer Sprache fest. Möglicherweise ist diese Forderung auch gar nicht realistisch. Wir kehren zu diesem Problem später zurück und gehen einstweilen davon aus, daß wir Lexika untersuchen, von denen jedes nach einer konkreten Methode geordnet ist. Lexika mit verschiedener Ordnung gelten dabei als verschieden.

Dies ist keine völlig befriedigende Lösung, aber sie ist unvermeidlich und reflektiert die zunehmende Komplexität sprachstatistischer Forschung. Ein besonders wichtiges und bis heute ungelöstes Problem ist das eines Häufigkeitswörterbuchs, das für die ganze Sprache repräsentativ wäre. In einem solchen Wörterbuch müßten die verschiedenen Gattungen und Sprachstile proportional vertreten sein. Doch von der Mehrheit der existierenden Häufig-

keitswörterbücher ist entweder bekannt oder anzunehmen, daß sie einen bestimmten Stil/ eine Gattung bevorzugen (z.B. die literarische Sprache).

Sind die Ränge für die Wörter der Sprache y festgelegt (wobei den „wichtigen", also den häufigen Wörtern ein niedriger Rang zukommt), muß man den Anteil der Sprache x an der Sprache y ($x \geq y$) bestimmen. Wir gehen dabei von der Hypothese aus, daß beim Sprachwandel die Wahrscheinlichkeit, daß ein Wort seinen Rang ändert (evtl. ganz verschwindet), umso größer ist, je höher der Rang des Wortes. Wenn also das Lexikon einer Sprache y durch Wandel aus dem Lexikon von x entstanden ist, d.h. $x \Rightarrow y$, dann müssen diese Lexika umso mehr Wörter gemeinsam haben, je niedriger deren Rang ist. Offensichtlich bewahren verwandte Sprachen diese Eigenschaft: je niedriger der Rang eines Wortes, desto größer die Wahrscheinlichkeit, daß es in beiden Sprachen auftritt. Der Anteil des Lehnwortschatzes hingegen wächst mit dem Rang.

Wir verfügen jetzt über ein Kriterium zur Unterscheidung von Verwandtschafts- und Entlehnungsbeziehungen. Auf der Grundlage der Annahme, daß die Zerfallswahrscheinlichkeit eines Wortes von seinem Rang abhängt, nicht jedoch von Zeit oder Sprache, formulieren wir im letzten Schritt Analoga zu den lexikostatistischen Postulaten. Durch Ordnung des Lexikons kamen wir zu einer Unterscheidung von Kern- und Rest-Lexik in einem verallgemeinerten Sinne (1^{O}); jetzt können wir die Kerne der Lexika zweier Sprachen vergleichen (2^{O}) mit dem Unterschied, daß Zugehörigkeit zum Kern nicht von der Bedeutung, sondern von der Häufigkeit abhängt; schließlich können wir mit der Annahme über die Zerfallswahrscheinlichkeit (3^{O}) von Anfang an die Möglichkeit unterschiedlicher Zerfallswahrscheinlichkeiten berücksichtigen.

Können wir nun noch einen Schritt weitergehen und auch noch die Abhängigkeit der Zerfallswahrscheinlichkeit von der jeweiligen Sprache berücksichtigen? Der außerordentliche Konservatismus z.B. des Isländischen, auf den Bergsland und Vogt [17] hingewiesen haben, wird ja von der Ranglistenmethode nicht erklärt. Unter Umständen können wir von dem Postulat der gleichen Geschwindigkeiten abgehen, indem wir die „Autokalibrationsmethode" verwenden, die im wesentlichen schon in der Arbeit [17] beschrieben ist. Da-

bei wird die Zerfallsgeschwindigkeit durch Vergleich eines Jetzt-Zustandes einer Sprache mit einer genau datierbaren Periode in ihrer Geschichte bestimmt (so wurde die spezifische Veränderungsrate des georgischen Lexikons durch Vergleich mit dem Altgeorgischen des V. Jh. n. Chr. gemessen). Auch wenn diese Methode nur auf Sprachen mit schriftlicher Tradition anwendbar ist, kann sie doch für die Praxis von Nutzen sein.

Nimmt man aber von Sprache zu Sprache unterschiedliche Zerfallsgeschwindigkeiten an, warum dann nicht auch innerhalb einer einzelnen Sprache (Sprachtradition)? Eine solche Annahme über zeitliche Variabilität der Zerfallsgeschwindigkeit erfordert keine wesentlichen Änderungen der Theorie und wird in 2.5 behandelt.

1.8. DER ZUSAMMENHANG VON HERKUNFT UND RANG VON WÖRTERN - PROBLEMGESCHICHTE

Die Verteilung von Wörtern nach etymologischer Herkunft in Abhängigkeit von ihrem Rang ist bereits untersucht worden, wenn auch, soweit den Verfassern bekannt, die Form der Abhängigkeit und ihre mögliche Verwendung als Kriterium für Sprachverwandtschaft bis jetzt nicht behandelt wurden. Als erster lieferte schon in den 20er Jahren Lirdsey [21] Daten zur Verteilung von Wörtern angelsächsischer und romanischer Herkunft nach dem Häufigkeitswörterbuch von Thorndike [20]. Später präsentierte Roberts [22] genauere Daten zur etymologischen Zusammensetzung der englischen Sprache auf der Grundlage eines anderen Häufigkeitswörterbuchs (Horn [23]). Das Histogramm (Abb. 3) zeigt die Verteilung der angelsächsischen Lexik nach Lindsey und Roberts.

Den entscheidenden Schritt tat jedoch der durch seine quantitativ-linguistischen Arbeiten („Zipfsches Gesetz") bekannt gewordene amerikanische Wissenschaftler G. Zipf. Er hatte wohl als erster die Idee, Häufigkeit und Entstehungszeit von Wörtern in Beziehung zu setzen. In einer 1947 veröffentlichten Notiz [1] zeigte

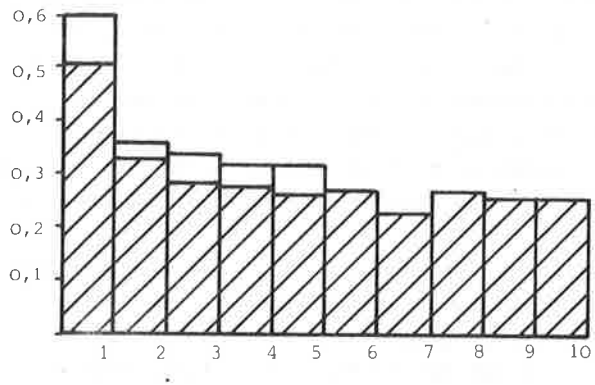

Abb. 3. Die Verteilung der Lexik angelsächsischer Herkunft unter den 5000 häufigsten Wörtern bei Thorndike und (schraffiert) unter den 10000 häufigsten Wörtern bei Horn (nach Lindsey und Roberts)

Zipf auf der Grundlage des sehr unvollkommenen Häufigkeitswörterbuchs von Eldridge [24], daß das Alter eines Wortes mit seiner Häufigkeit korreliert ist, wobei unter den häufigsten Wörtern vor allem sehr alte zu finden sind. Das unten wiedergegebene Diagramm (Abb. 15a) wurde zuerst in der erwähnten Arbeit [1] und dann (1949) in der bekannten Monographie Zipfs [25] publiziert. Zipf glaubte, auch den quantitativen Zusammenhang von Häufigkeit und Alter eines Wortes gefunden zu haben, und kam dabei mit seiner Annahme, er habe die Form einer „arith-logarithmic equation" [1:527], d.h. einer Exponentialfunktion, der Wahrheit recht nahe.[9)]

Zipf erkannte nicht nur, daß ein Zusammenhang zwischen Rang und Alter besteht, sondern auch, daß man diesen für sprachgeschichtliche Untersuchungen nutzen kann. Er untersuchte die Verteilung gotischer Wurzeln in Abhängigkeit von ihrer Vorkommenshäufigkeit in der Wulfila-Bibel (s. [1]). Außerdem formulierte Zipf bei der

Untersuchung der ausschließlich germanischen (d.h. nicht gemeinindogermanischen) Lexik exakt unser augenblickliches Problem: „Es erhebt sich die Frage, was diese rein germanische Schicht historisch bedeutet. Ist sie (a) das Ergebnis prähistorischer germanischer kultureller Kontakte mit Völkern, die ausgestorben sind oder von anderen absorbiert wurden? Oder (b) handelt es sich um eine Periode der Bildung von Neologismen, die dann aber beispiellos wäre" [1:529]. Zipf selbst gibt keine Antwort auf diese Frage. Stattdessen entwickelt er ein umfangreiches Programm zur vergleichenden Untersuchung anderer germanischer Sprachen mit der vorgeschlagenen Methode.

Mit dem zur Verfügung stehenden Material wäre eigentlich mehr als die bloße Problemstellung möglich gewesen. Zipf wurde durch die Unkenntnis des exakten quantitativen Zusammenhangs von Rang und Alter eines Wortes und das Fehlen einer klaren Vorstellung von lexikalischem Wandel gehindert. Vielleicht wurde seine Arbeit deshalb einfach vergessen, so daß M. Swadesh, als er einige Jahre später versuchte, ein Kernlexikon abzugrenzen, dieses allein auf der Grundlage seiner Intuition tat.

Wir fangen da an, wo Zipf stehengeblieben ist, und versuchen, ein mathematisches Modell des lexikalischen Wandels zu konstruieren, und damit ein mathematisches Gesetz für den Zusammenhang von Alter und Rang eines Wortes abzuleiten.

2. EIN MATHEMATISCHES MODELL DES LEXIKALISCHEN WANDELS: VERERBUNG

2.1. VORGÄNGER- UND NACHFOLGERLEXIKON

Unter dem Lexikon V_x einer Sprache x verstehen wir die Menge $V_x = \{a_j\}$ von Elementen, die wir Wörter nennen (auch wenn es sich tatsächlich um Wortformen, Lexeme, Lexemklassen u.ä. handelt). Wenn die Sprache x unmittelbar Vorgänger der Sprache y ist $(x \to y)$, dann ist eine eineindeutige Abbildung einer Teilmenge $\tilde{V}_x$ von Wörtern des Lexikons V_x auf eine Teilmenge $\tilde{V}_y$ des Lexikons V_y gegeben[1]:

$$\varphi: \tilde{V}_x \to \tilde{V}_y.$$

Ein Wort $b = \varphi(a_j) \in \tilde{V}_y$ wird Nachfolger des Wortes a_j aus V_x genannt. Die Wörter $\tilde{V}_x$ kann man auch als in V_x erhaltene Wörter bezeichnen, die Teilmenge φV_x als alte Wortschicht von V_y.

Wörter aus der Teilmenge $V_x \setminus \tilde{V}_x$ heißen geschwunden, und die Wörter der Teilmenge $V_y \setminus \tilde{V}_y$ (oder $V_y \setminus \varphi V_x$) Lehnwörter oder Innovationen.

Die Wörter der beiden Lexika V_x und V_y sind nach Häufigkeit geordnet und, mit dem häufigsten beginnend, fortlaufend numeriert (rangiert).

Die Präsentation des Materials sowie die Bearbeitung der Daten vereinfacht sich, wenn man die Wörter mit ähnlichen Häufigkeiten in gleich großen Rangklassen $G_1, G_2, \ldots, G_i \ldots$ zu je n Wörtern zusammenfaßt und diese so numeriert, daß die Klasse G_1 die Wörter der Ränge 1 bis n enthält, die Klasse G_2 die Wörter der Ränge n + 1 bis 2n usw. Die Rangklasse G_i umfaßt also die Wörter, die im Häufigkeitswörterbuch die Plätze $i(n-1)+1$ bis $i \cdot n$ belegen.

Es besteht Grund für die Annahme, daß die Wahrscheinlichkeit für ein Wort, innerhalb eines gegebenen Zeitintervalls zu schwinden, mit dem Rang seiner Klasse (d.h. mit abnehmender Häufigkeit) wächst, und daß diese Wahrscheinlichkeit nur von seinem Platz im Häufigkeitswörterbuch V_x abhängt, nicht jedoch z.B. von seiner Herkunft oder Lautung. Diese Annahmen sind grundlegend für diese Arbeit, und wir werden immer wieder auf sie zurückkommen.

Betrachten wir weiter die für zwei rangierte und in Rangklassen eingeteilte Lexika V_x und V_y definierte Abbildung φ in Hinblick auf einige wichtige Eigenschaften.

$N_i(x)$ bezeichne die Anzahl derjenigen Wörter a aus der Klasse $G_i^x \subset V_x$, deren Nachfolger in V_y der Klasse des gleichen Rangs i angehören:

$$N_i(x) = |\{a|a\in G_i^x,\ \varphi(a)\in G_i^y\}|.$$

D.h. $N_i(x)$ ist die Zahl der Wörter einer Klasse, die nicht nur im Nachfolgerlexikon erhalten sind, sondern auch ihren Rang nur so geringfügig verändert haben, daß sie in der gleichen Rangklasse geblieben sind.

Offenbar ist $N_i(x) = N_i(y)$, wenn $N_i(y)$ die Anzahl derjenigen Wörter aus G_i^y ist, deren Vorgänger $a \in V_x$ in der i-ten Rangklasse von V_x enthalten sind. Dann ist die Anzahl der Wörter von G_i^x, die geschwunden sind, $\bar{N}_i(x) = n - N_i(x)$ gleich der Anzahl der Innovationen $\bar{N}_i(y)$ in der entsprechenden Klasse des Nachfolgerlexikons. Als Innovationen gelten hier solche Wörter, die neu in eine bestimmte Klasse gekommen sind, im Lexikon V_x aber bereits vorhanden gewesen sein können, und als geschwundene Wörter einer Klasse solche, die zumindest ihre Rangklasse gewechselt haben. Die Beschreibung des Schwundes von Wörtern einer Klasse G_i^x ist damit gleichzeitig die Beschreibung der Innovationen in G_i^y. Das Schwinden von Wörtern aus einer Klasse läßt sich sehr einfach beschreiben (s. 2.2); das Schwinden aus dem Lexikon ist unter der folgenden Zusatzhypothese analog beschreibbar:

Sei $F_i(x)$ die Anzahl derjenigen Wörter a einer Klasse $G_i^x \subset V_x$ mit $a \in \tilde{V}_x$:

$$F_i(x) = |\{a|a\in G_i^x \cap \tilde{V}_x\}|,$$

und $F_i(y)$ die Anzahl der Wörter b, $b\in G_i^y \cap \tilde{V}_y$. Wir fordern, daß für ein beliebiges Paar von Lexika V_x, V_y und für beliebige Klassen G_i die folgende Gleichung gilt:

$$F_i(x) = F_i(y). \qquad (*)$$

Das bedeutet nur, daß es genausoviele ererbte Wörter in der i-ten Klasse des Nachfolgerlexikons gibt, wie erhaltene Wörter aus der entsprechenden Klasse des Vorgängerlexikons. (Dabei ist nicht gefordert, daß diese Wörter auch dieselben sind; nur ihre Anzahl muß gleich sein!)

Wir stellen die Hypothese (*) auf, weil sie die ökonomischste Beschreibung der Phänomene gestattet und den uns bekannten Fakten nicht widerspricht. Man kann diese Hypothese auch anders interpretieren. Wenn von zwei Häufigkeitswörterbüchern V_x und V_y bekannt ist, daß sie dieselbe Sprachtradition repräsentieren, dann kann man allein aus der Verteilung der einander entsprechenden Wörter in diesen Wörterbüchern nicht entscheiden, welche der beiden Sprachen Vorgänger und welche Nachfolger ist. Mit anderen Worten, wenn wir nur über die Beschreibungsmittel für die Beziehungen zwischen den einzelnen Wörterbüchern verfügen, die zur Aufstellung der hier dargelegten Theorie notwendig sind, können wir die Richtung der zeitlichen Relation zwischen den Sprachen nicht bestimmen.

Mit (*) gilt natürlich auch $\bar{F}_i(x) = \bar{F}_i(y)$. D.h. die Zahl der geschwundenen Wörter der i-ten Klasse von V_x ist gleich der Zahl der Innovationen in der entsprechenden Klasse des Nachfolgerlexikons V_y.

Diese Eigenschaft ermöglicht es nun, statt zwei Prozesse zu analysieren - das Schwinden von Wörtern aus dem Vorgängerlexikon und die Aufnahme neuer Wörter ins Nachfolgerlexikon -, sich auf den Schwund von Wörtern zu beschränken und die Ergebnisse auf den (zum Schwund symmetrischen) Innovationsprozeß zu übertragen.

Weiter wird angenommen, daß die Größen N_i und F_i nicht von der Wahl eines konkreten Paars von Lexika V_x und V_y abhängen, sondern von der Länge des Zeitintervalls t, das sie trennt. Für den Zusammenhang von t und N_i (bzw. F_i) werden einige mögliche Varianten in 2.2 und 2.5 gezeigt.

Schließlich nehmen wir an, daß für ein beliebiges Paar von Lexika V_x, V_y, deren zeitlicher Abstand t größer als O ist, die Menge der erhaltenen Wörter $\tilde{V}_x$ (entsprechend auch $\tilde{V}_y$) endlich ist.

Mit Hilfe dieser Annahme können wir in 2.3 die Abhängigkeit der Zahl geschwundener Wörter $\bar{N}_i$ (bzw. $\bar{F}_i$) vom Rang der Klasse

bestimmen: $N_i = N(i)$. Außerdem hilft sie, sich mit der Idee eines unendlichen Lexikons vertraut zu machen. In jedem Lexikon ist nämlich die Anzahl der Wörter, die eine Geschichte haben, endlich; und nur diese sind philologisch interessant. Die übrigen Lexikonelemente sind ephemer, d.h. Gelegenheitsbildungen, die nach den Regeln der Sprache konstruiert sind, aber nur in einem gegebenen Text auftreten.

Nun soll die Variable t explizit eingeführt werden. Wenn wir ein Zeitintervall t_1 einer Abbildung φ zuordnen und einer Abbildung ψ ein Intervall t_2, dann kann man der Komposition $\varphi\psi$ dieser Abbildungen

$$V_x \overset{t_1}{\rightarrow} V_y \overset{t_2}{\rightarrow} V_z$$

die Summe $t_1 + t_2$ zuordnen und dann zu einer kontinuierlichen Betrachtungsweise der Veränderung von V_x in Abhängigkeit von t übergehen. Hier sei angemerkt, daß in allen Formeln t die Länge des Zeitintervalls zwischen zwei Wörterbüchern bezeichnet, und damit immer $t \geq 0$ ist. Die Größe t hängt nicht von der Wahl des Koordinatensystems ab. In einigen Fällen wird im Zusammenhang mit Beispielen und empirischen Daten das Symbol t auch zur Bezeichnung eines Zeitpunktes (einer Wörterbuchkoordinate) verwendet. In diesem Fall hängt t von der Wahl des Koordinatenursprungs ab und kann natürlich auch negative Werte annahmen. Die jeweilige Bedeutung von t kann der Leser leicht aus dem Kontext entnehmen.

Sei $\bar{N}_i(t)$ die Proportion der Wörter[2)], die aus der Klasse G_i in der Zeit t geschwunden sind. Diesen Schwundprozeß wollen wir mit einer Differentialgleichung beschreiben.

2.2. DIE GLEICHUNG FÜR DEN SCHWUND

Hierzu nehmen wir an, daß die Klasse G_i in jedem Zeitintervall gleichförmig zerfällt und daß eine reine „Markov-Situation" vorliegt: die freigewordenen Plätze im Lexikon werden sofort wieder besetzt, und die neu hinzugekommenen Lexikoneinheiten können wiederum schwinden. Wenn wir also die Proportion der Wörter finden wollen, die während $t_1 + t_2$ ihre Plätze verlassen (ihren Rang verändert) haben (wobei die Proportion der während t_1 geschwundenen Wörter als bekannt vorausgesetzt wird), müssen wir die Stellen im Lexikon bestimmen, die von Wörtern während t_1 und dann wieder während t_2 verlassen worden sind, d.h. die Stellen, wo zweimal Wörter geschwunden sind.

Die Proportion der aus der Klasse G_i geschwundenen Wörter sei $\bar{N}_i(t)$ (die Proportion der erhaltenen Wörter entsprechend $N_i(t)$. Wir benötigen noch die Zerfallsrate η_i der Klasse G_i. Der Einfachheit halber nehmen wir an, daß sie innerhalb der ganzen Klasse gleich ist, und vernachlässigen interne Unterschiede. Die Größenordnung von η_i erhalten wir dann durch zusätzliche Annahmen: wieder kann man die Zerfallsrate η_i in gewissem Sinn als Zerfallswahrscheinlichkeit in einem kleinen Zeitpunkt δt behandeln, und je höher der Rang von G_i, je geringer also der Umfang dieser Klasse, desto größer ihre Zerfallsrate.

Wir stellen nun die Gleichung für den Anteil der innerhalb einer Zeit t geschwundenen Wörter auf und lösen sie.

Bekannt sei der Anteil $\bar{N}_i(t)$ von Wörtern, die während t geschwunden sind. Nach einer weiteren Zeit δt erhalten wir:

$$\bar{N}_i(t + \delta t) = \bar{N}_i(t) + \eta_i \delta t - A, \tag{2.1}$$

wobei A die Anzahl der „Stellen" (Ränge) ist, wo zweimal Wörter geschwunden sind: während t und noch einmal während des Intervalls δt. Also ist A, als Teil von $\bar{N}_i(t)$,

$$A = \eta_i \bar{N}_i(t) \delta t.$$

Damit läßt sich der Ausdruck (2.1) umformen in

$$\bar{N}_i(t + \delta t) = \bar{N}_i(t) + \eta_i \delta t - \eta_i \bar{N}_i(t)\delta t. \qquad (2.2)$$

In (2.2) bringen wir $\bar{N}_i(t)$ auf die linke Seite und dividieren auf beiden Seiten durch δt. Durch Grenzwertbildung erhalten wir

$$\bar{N}_i'(t) = \eta_i - \eta_i \bar{N}_i(t), \qquad (2.3)$$

$$\bar{N}_i'(t) - \eta_i \bar{N}_i(t) - \eta_i = 0.$$

Die allgemeine Lösung von (2.3) lautet

$$\bar{N}_i(t) = 1 + Re^{-\eta_i t} \qquad (2.4)$$

Wegen der Bedingung $\bar{N}_i(0) = 0$ ist $R = -1$, d.h.

$$\bar{N}_i(t) = 1 - e^{-\eta_i t}. \qquad (2.5)$$

Der Ausdruck (2.5) gibt den Anteil der Wörter an, die während t aus der Klasse G_i des Lexikons V_x geschwunden sind. $N_i(t)$ - der Anteil der erhaltenen Wörter - beträgt dann offenbar

$$N_i(t) = e^{-\eta_i t}. \qquad (2.6)$$

Das sind die Wörter, die im Lexikon V_y $(x \Rightarrow y)$ in der gleichen (i-ten) Rangklasse bleiben.

Außerdem gilt, daß die angestellten Überlegungen und Berechnungen völlig unabhängig von der Größe der betrachteten Klasse G_i sind. So kann man als G_i genauso gut die ersten hundert wie die ersten tausend Wörter des Lexikons wählen, die Wörter mit den Nummern 500 bis 600 oder ein beliebiges endliches Fragment des Lexikons V_x; der Zusammenhang zwischen dem Anteil $\bar{N}_i$ der geschwundenen Wörter einer Klasse und der Zeit t bleibt gleich, auch wenn sich natürlich der Wert des Parameters η_i ändert.

Die Anzahl der erhalten gebliebenen Wörter der i-ten Klasse des Lexikons beträgt

$$\mathbb{N}_i(t) = ne^{-\eta_i t} \qquad (2.7)$$

Der Ausdruck (2.7) entspricht genau der glottochronologischen Grundformel (1.2).

Die Gesamtzahl der Wörter, die in einem aus den Klassen G_i $(i = 1,2,\ldots,k)$ bestehenden Anfangsstück des Lexikons V_x erhalten sind, beträgt

$$S_k(t) = \sum_{i=1}^{k} \mathbb{N}_i(t) = n \sum_{i=1}^{k} e^{-\eta_i t} \qquad (2.8)$$

Bisher wissen wir nur, wie der Schwund innerhalb einer Klasse von der Zeit abhängt, d.h. wie sich mit der Zeit für ein Wort die Wahrscheinlichkeit dafür ändert, die Rangklasse zu verlassen, der es zu t angehörte.

Betrachten wir jetzt den Schwund im eigentlichen Sinn, wo nämlich ein Wort nach einer Zeit t nicht nur seine Klasse G_i, sondern sogar das Lexikon verläßt. Zum Zusammenhang der Rate η_i des Schwundes aus der entsprechenden Klasse mit der Zerfallsrate $\hat{\eta}_i$ für Wörter der Klasse machen wir folgende Annahme:

$$\hat{\eta}_i = \eta\beta_i \qquad (2.9)$$

mit konstantem β. Wenn η die Rate des Schwundes in einem Anfangsabschnitt des Gesamtlexikons ist, dann ist die Zerfallsrate für diesen Abschnitt

$$\hat{\eta} = \beta\eta. \qquad (2.10)$$

Nun können wir den Anteil $\bar{F}_i(t)$ von Wörtern der Klasse G_i bestimmen, die aus dem Lexikon geschwunden sind:

$$\bar{F}_i(t) = 1 - e^{-\hat{\eta}_i t}. \qquad (2.11)$$

Der Anteil der in G_i zu t erhaltenen Wörter ist

$$F(t) = e^{-\hat{\eta}_i t}. \qquad (2.12)$$

Entsprechend ist die Anzahl der Wörter aus G_i, die im Lexikon verblieben sind (bei einer Klassengröße von n)

$$F_i(t) = ne^{-\hat{\eta}_i t}. \qquad (2.13)$$

2.3. DIE ABHÄNGIGKEIT DES SCHWUNDES VOM RANG

Wir werden nun den Zusammenhang von Wortrang und Schwundwahrscheinlichkeit suchen. Untersuchen wir, welche Bedingungen wir zur Bestimmung des Zusammenhangs zwischen dem Klassenrang von G_i und $N_i(t)$ (bzw. $F_i(t)$) benutzen können.

Einmal steht ja der Rang der Klasse mit den Rängen der in ihr enthaltenen Wörter in Zusammenhang, und die Schwundwahrscheinlichkeit wächst mit dem Wortrang, so daß der Anteil der Wörter, die zu einem bestimmten Zeitpunkt t in G_i verblieben sind, mit ansteigendem Klassenrang sinken muß.

Außerdem ist nach Annahme (s. 2.1) die Anzahl N(t) der Wörter, die zu t aus dem Lexikon V_x noch erhalten sind (mit $x \Rightarrow y$), endlich.

Betrachten wir den Graphen der Funktion $N_i(t)$ für einen Anfangsabschnitt eines Lexikons, das N_o Wörter umfaßt, so finden wir eine Treppenfunktion. Wir verringern die Klassenbreite n und vergrößern den Umfang des Gesamtlexikons N_o. Es geht nun darum, Bedingungen für η_i zu finden, unter denen die „Treppe", die bei $N_o \to \infty$ und $n \to 1$ von $N_i(t)$ gebildet wird, durch eine monoton fallende Exponentialfunktion approximiert wird, d.h. unter denen die Reihe S der Werte von $N_i(t)$ konvergiert.

$$S = n \sum_{i=1}^{\infty} e^{-\eta_i t}. \qquad (2.14)$$

Das Konvergenzverhalten dieser Reihe entspricht dem des Integrals

$$I = \int_o^{\infty} e^{-\eta(x)t} dx.$$

Hier ist die Konvergenzbedingung

$$e^{-\eta(x)} < \frac{1}{x^{\alpha}}; \quad \alpha > 1, \qquad (2.14a)$$

$$\eta(x) > \alpha \ln x.$$

Empirische Untersuchungen zeigen, daß die Schwundrate η_i von der Größenordnung $\sqrt{i}$ sein muß. Die Funktion $\sqrt{x}$ genügt auch der Bedingung (2.14a). Damit wird Formel (2.13) zu

$$F_i(t) = ne^{-\alpha t\sqrt{i}}. \qquad (2.15)$$

Wir weisen noch auf eine wichtige Eigenschaft von (2.15) hin: der Anteil der in einer Klasse erhalten gebliebenen Wörter hängt symmetrisch von Zeit und Rang (genauer: von der Wurzel des Rangs) ab[3].

2.4. DIE UNABHÄNGIGE ENTWICKLUNG ZWEIER LEXIKA GEMEINSAMEN URSPRUNGS

Im vorhergehenden Paragraphen haben wir im wesentlichen einen Schritt zur Entwicklung eines Kriteriums für die Abstammung eines Lexikons V_y vom einem Lexikon V_x getan. Wenn $x \Rightarrow y$, und ihr Zeitabstand t beträgt, so muß sich der empirisch beobachtete Zusammenhang zwischen dem Klassenrang i im Lexikon V_y und dem Anteil von Wörtern dieser Klasse mit Etyma aus der Sprache x durch die mono-

ton fallende Funktion (2.15) approximieren lassen. Unten zeigen wir, daß der Anteil der Wörter, die auf x zurückgehen, mit dem Rang wächst (zumindest in einem Intervall von Rängen), wenn x nicht Vorgänger, sondern nur Entlehnungsquelle von y ist.

Auf das Verhältnis „Nachfolger-Vorgänger" werden wir im Kapitel 4 bei der Behandlung der Entlehnungen zurückkommen müssen. Zunächst soll es um etwas anderes gehen: wir versuchen, das Ergebnis zur Verteilung der Wörter einer verwandten Sprache x im Lexikon V_y aus dem vorigen Paragraphen auf den Fall auszuweiten, wo die verwandten Sprachen x und y bezüglich der Relation {⇒} nicht vergleichbar sind, sondern einen gemeinsamen Vorgänger z (z ⇒ x und z ⇒ y) haben. Wir leiten also eine der Stärke der glottochronologischen Gleichung (1.3) entsprechende funktionale Abhängigkeit her.

Angenommen, zum Zeitpunkt t_o existiere eine Protosprache x_o mit dem Lexikon V_{x_o}. Nach einer Zeit t haben sich daraus zwei selbständige Sprachen x_1 und x_2 ($x_o \Rightarrow x_1$ und $x_o \Rightarrow x_2$) mit den Lexika V_{x_1} bzw. V_{x_2} entwickelt.

Wir gehen davon aus, daß die Entwicklung der Sprachen x_1 und x_2 von t_o an völlig unabhängig verläuft: wir nehmen an, daß die Sprachen gegenseitig keine Wörter entlehnen, nicht dieselben Wörter aus einer dritten Sprache entlehnen, keine speziellen Gründe haben, dieselben Wörter zu erhalten, und schließlich, daß sie keine einzige parallele Neubildung vornehmen, auch wenn sie praktisch denselben Vorrat an aus x_o ererbten morphologischen Mitteln benutzen. Diese Annahme ist natürlich nicht ganz realistisch, aber wir erhalten auf diese Weise eine Minimalschätzung: wir bestimmen die Mindestzahl von Wörtern aus V_{x_o}, die in den Lexika beider Sprachen nach einer t Jahre langen Periode selbständiger Entwicklung erhalten sind.

Wenn man im Rahmen dieses Modells bleibt, so muß während t die Entwicklung der beiden Lexika völlig gleich verlaufen sein und sich vollständig durch die Gleichung (2.3) und ihre Lösung (2.5) und (2.6) beschreiben lassen.

In den beiden Lexika V_{x_1} und V_{x_2} muß aus jeder Klasse $G_i^1 \subset V_{x_1}$ bzw. $G_i^2 \subset V_{x_2}$ eine gleiche Anzahl von Wörtern geschwunden sein,

jedoch möglicherweise verschiedene Wörter, d.h. von verschiedenem Rang innerhalb $G_i \subset V_{x_o}$. Daraus ergeben sich zwei Fragen.

Frage 1. Wie ist die Verteilung der Wahrscheinlichkeiten dafür, daß ein Wort $a_j \in V_{x_o}$, welches zur Klasse G_i^1 des Lexikons V_{x_1} gehört ($V_{x_1} = \{G_i^1\}$, i = 1,2,...), auch zur i-ten Rangklasse G_i^2 des Lexikons V_{x_2} gehört?

Offenbar ist diese Fragestellung äquivalent mit der Frage nach der Verteilung der Wahrscheinlichkeiten, daß in den beiden Lexika V_{x_1} und V_{x_2} die Wörter aus V_{x_o} die gleichen Ränge haben.

Frage 2. Wie ist die Verteilung der Wahrscheinlichkeiten dafür, daß ein Wort $a_j \in V_{x_o}$, das der Klasse G_i^1 des Lexikons V_{x_1} angehört, in V_{x_2} überhaupt vorkommt?

Zur Lösung der Frage 1: Es seien zwei Klassen G_i^1 und G_i^2 mit je n Wörtern gegeben. Jede enthält $\mathbb{N}_i(t)$ Wörter aus V_{x_o}. Sei l die Anzahl der günstigen Ausgänge, d.h. die Anzahl von Stellen (Rängen) mit identischen Wörtern in den Lexika. Die Wahrscheinlichkeitsverteilung von l ist die hypergeometrische [26]

$$p(l) = \frac{C_{\mathbb{N}_i}^{l} \, C_{}^{\mathbb{N}_i - l}}{C_n^{\mathbb{N}_i}} \cdot \qquad (2.16)$$

Die mathematische Erwartung der mit (2.16) hypergeometrisch verteilten Zufallsvariablen l ist in unserem Fall (s. [26]).

$$M_i(l) = \frac{\mathbb{N}_i \, \mathbb{N}_i}{n} \cdot \qquad (2.17)$$

Durch Einsetzen von (2.7) in (2.17) erhalten wir die mittlere Anzahl nicht geschwundener Wörter, die in der i-ten Klasse der Lexika V_{x_1} und V_{x_2} an derselben Stelle stehen:

$$M_i(l) = n e^{-2n_i t} \cdot \qquad (2.18)$$

Auf gleiche Weise gewinnen wir auch die mittlere Anzahl identischer Stellen zweier Anfangsabschnitte der Lexika V_{x_1} und V_{x_2}, von denen jedes k Klassen wie in (2.8) enthält:

$$S_k^{(2)}(t) = n \sum_{i=1}^{k} e^{-2\eta_i t}. \qquad (2.19)$$

So erhalten wir für zwei Sprachen $x_1 \sim x_2$ dieselbe Gesetzmäßigkeit wie für ein „Vorgänger-Nachfolger"-Paar ($x_o \Rightarrow x_1$ oder $x_o \Rightarrow x_2$), nur mit doppelt so großem Exponenten. Die Formel (2.19) stimmt mit der glottochronologischen Formel (1.3) überein.

Zur Frage 2: Gesucht wird die Wahrscheinlichkeitsverteilung und die mathematische Erwartung der Anzahl „alter" Wörter r aus der Klasse G_i^1, die (irgendwo) in V_{x_2} vorkommen.

Aus (2.15) erhalten wir den Anteil der Wörter $F_i = F_i(t)$ der Klasse G_i^1 (bzw. G_i^2), die während t völlig geschwunden sind:

$$\bar{F}_i(t) = 1 - e^{-\hat{\eta}_i t}.$$

Mit l bezeichnen wir die Anzahl identischer Wörter der i-ten Klasse beider Lexika V_{x_1} und V_{x_2}, die völlig geschwunden sind. Dann ist

$$r = n - \bar{F}_i(t)n - [\bar{F}_i(t)n - l] = n - 2\bar{F}_i(n) + l.$$

mit der mathematischen Erwartung

$$M_i(r) = M_i(n - 2\bar{F}_i n + l) = n - 2\bar{F}_i n + M_i(l). \qquad (2.20)$$

Die Wahrscheinlichkeit p(l) dafür, daß aus $G_i^1 \subset V_{x_1}$ und aus dem gesamten Lexikon V_{x_2} l gemeinsame Wörter schwinden, hat die hypergeometrische Verteilung

$$p(l) = \frac{C_{\bar{F}_i}^{l} \cdot C_{n-\bar{F}_i}^{\bar{F}_i - l}}{C_n^{\bar{F}_i}},$$

und die Erwartung von l ist

$$M_i(l) = \frac{\bar{F}_i^2 n^2}{n}, \qquad (2.21)$$

Durch Einsetzen von (2.21) in (2.20) erhalten wir die gesuchte Erwartung $M_i(r)$:

$$M_i(r) = n - 2\bar{F}_i n + \frac{\bar{F}_i^2 n^2}{n^2} = \frac{(n - \bar{F}_i n)^2}{n}, \qquad (2.22)$$

weiter durch Einsetzen des Wertes von $\bar{F}_i$ in (2.22)

$$M_i(r) = \frac{[n - n(1-e^{-\hat{\eta}_i t})]^2}{n} = \frac{n^2(1-1+e^{-\hat{\eta}_i t})^2}{n} = ne^{-2\hat{\eta}_i t}. \qquad (2.23)$$

Die Lösung von Frage 2 entspricht der von Frage 1 bis auf die Tatsache, daß anstelle des Parameters η, der in (2.19) die Rate des Schwundes aus einer Klasse bezeichnet, der Parameter $\hat{\eta}_i$ der absoluten Zerfallsrate steht.

2.5. DIE ABHÄNGIGKEIT DER ZERFALLSWAHRSCHEINLICHKEIT EINES WORTES VON DER ZEIT

In die glottochronologische Theorie gehen die folgenden wichtigen Voraussetzungen ein:

a) der Parameter η hängt nicht von der Zeit ab;
b) der Parameter η hängt nicht von der jeweiligen Sprache ab.

Diese beiden Annahmen hängen eng zusammen und lassen sich auf eine zurückführen: Jede rangierte Liste (Lexikon) verändert sich in der Zeit völlig gleichförmig, d.h. unabhängig davon, was für

Wörter rangiert sind (Französisch oder Papua), und wie sie ihren Rang bis zum Zeitpunkt t_o, dem Ausgangspunkt unserer Betrachtung, verändert haben.

Eine derartige Voraussetzung trifft aber zumindest für einige Lexika nicht zu: die einzelnen Sprachstile entwickeln sich mit unterschiedlicher Geschwindigkeit, wobei z.B. der wissenschaftliche und publizistische Stil seine Veränderung mit der Zeit beschleunigt, während die Sprache der Belletristik die ihre verzögert. Sehr wahrscheinlich ist, daß sich die Sprache als ganze, wenn man sie als „Resultante" aller Stile und Genres betrachtet, mit konstanter Geschwindigkeit entwickelt, daß aber Häufigkeitswörterbücher in der Regel nicht ein derart „gemitteltes" Bild der Sprachverwendung wiedergeben, sondern zu der einen oder anderen „Sprachperipherie" neigen.

Man darf annehmen, daß nicht nur „periphere" Stile sondern auch „periphere" Sprachen existieren. Mit einer solchen Hypothese ließen sich die Ergebnisse von Bergsland und Vogt [17] für das Isländische erklären. Die berechnete, äußerst niedrige Veränderungsrate des isländischen Lexikons läßt sich nicht völlig durch die Unzulänglichkeit der glottochronologischen Meßmethode erklären.

Jedenfalls kann man schon aufgrund der Tatsache, daß das Isländische mit seiner verlangsamten Entwicklung nicht allein steht, und eine Parallele gezogen werden kann zu den archaisierten Stilen anderer Sprachen, das Programm der absoluten Datierung wieder optimistischer betrachten, nachdem es durch die Daten von Bergsland und Vogt erschüttert worden war.

Wir modifizieren die Konstanzannahme bezüglich der Zerfallsrate, von der wir in 2.2-2.4 ausgegangen sind, zu der generellen Annahme, daß diese Rate eine Funktion der Zeit ist: $\eta = \eta(t)$.

Unter dieser Annahme wird die hinreichend oft differenzierbare Differentialgleichung (2.3)

$$\bar{N}'(t) + \eta\bar{N}(t) - \eta = 0$$

zu

$$\bar{N}'(t) + \eta(t)\bar{N}(t) - \eta = 0. \qquad (2.24)$$

Wir lösen zunächst die homogene Gleichung

$$\bar{N}'(t) + \eta(t)\bar{N}(t) = 0,$$

$$\frac{\bar{N}'(t)}{\bar{N}(t)} = -\eta(t),$$

$$\ln \bar{N}(t) = -\int_o^t \eta(x)dx,$$

$$\bar{N}(t) = e^{-\int_o^t \eta(x)dx}.$$

Die Lösung von (2.24) suchen wir in der Form

$$c(t)e^{-\int_o^t \eta(x)dx}. \qquad (2.25)$$

Wir setzen (2.25) in (2.24) ein

$$c'(t)e^{-\int_o^t \eta(x)dx} - c(t)\eta(t)e^{-\int_o^t \eta(x)dx} + c(t)e^{-\int_o^t \eta(x)dx} \cdot \eta(t) - \eta(t) = 0,$$

$$c'(t)e^{-\int_o^t \eta(x)dx} - \eta(t) = 0,$$

$$c'(t) = \eta(t)e^{\int_o^t \eta(x)dx},$$

$$c(t) = \int_o^t \eta(y)e^{\int_o^t \eta(x)dx}dy + k.$$

Wir erhalten als allgemeine Lösung von (2.24)

$$\bar{N}(t) = \left[\int_o^t \eta(y) e^{\int_o^y \eta(x)dx} dy + k\right] e^{-\int_o^t \eta(x)dx}.$$

Aus der Bedingung, daß $\bar{N}(t) = 0$ für $t = 0$ gewinnen wir $k = 0$. Die Lösung von (2.24) ist dann

$$\bar{N}(t) = \left[\int_o^t \eta(y) e^{\int_o^y \eta(x)dx} dy\right] e^{-\int_o^t \eta(x)dx}.$$

Wir stellen die Formel für den Anteil der aus G_1 während t geschwundenen Wörter auf unter der Annahme, daß die Zerfallsrate $\eta(t) = bt^{b-1}$,

$$\bar{N}(t) = \left[\int_o^t by^{b-1} e^{\int_o^y bx^{b-1}dx} dy\right] e^{-\int_o^t bx^{b-1}dx} =$$

$$= \left[\int_o^t by^{b-1} e^{y^b} dy\right] e^{-t^b} = \left[\int_o^t e^{y^b} d(y^b)\right] e^{-t^b} = e^{t^b}\Big|_o^t \cdot e^{-t^b} =$$

$$= (e^{t^b} - 1)e^{-t^b} = 1 - e^{-t^b}. \qquad (2.26)$$

Dann ist

$$\bar{N}(t) = 1 - e^{-\eta t^b}. \qquad (2.27)$$

Die Annahme dieses exponentiellen Zusammenhangs von Zerfallsrate und Zeit (2.26) ist noch zu unspezifisch. Wir betrachten speziell den Fall der Beschleunigung b = 2:

$$\eta(t) = 2\gamma t.$$

Dann erhalten wir direkt aus (2.27)

$$\bar{N}(t) = 1 - e^{-\gamma t^2}. \qquad (2.28)$$

Vergleichen wir (2.28) mit der Zerfallsformel für konstantes t (2.5), $\bar{N}(t) = 1 - e^{-\eta t}$, so bemerken wir, in Analogie zur Physik, daß, wenn t die Zeit, η die Geschwindigkeit und γ die Beschleunigung bezeichnet, im Exponenten von (2.5) und (2.28) der "Weg" steht, den eine Sprache vom Anfangszeitpunkt t_o bis zur Gegenwart (dem Moment der Erstellung des Lexikons) zurückgelegt hat. (2.27) beruht auf der Annahme, daß zum Zeitpunkt t_o die Anfangsgeschwindigkeit des Zerfalls 0 ist; diese Annahme wurde aber nur gemacht, um den Ausdruck zu vereinfachen.

Dann aber kann man unsere Zerfallsformel verallgemeinern, indem man die verschiedenen Ausdrücke für den Weg durch den allgemeinsten ersetzt: die Darstellung des Wegs als Integral s der Geschwindigkeit $\eta(t)$ über die Zeit:

$$\bar{N}(t) = 1 - e^{-s(t)} = 1 - e^{-\int_o^t \eta(\tau)dx}. \qquad (2.29)$$

Sehen wir nun, was passiert, wenn die Geschwindigkeit nicht zu-, sondern abnimmt. Die Annahme einer linearen Abhängigkeit erscheint am natürlichsten:

$$\eta(t) = -2\gamma t.$$

Das führt allerdings zu dem linguistisch unsinnigen Ergebnis

$$\bar{N}(t) = 1 - e^{-(\eta_o t - \gamma t^2)} \qquad (2.30)$$

mit η_o als Anfangsrate des Zerfalls. Aus (2.30) geht hervor, daß das Lexikon nur bis zum Zeitpunkt t_η zerfällt, wo die Gleichung $\eta_o t = \gamma t^2$ gilt.

Zu diesem Zeitpunkt kommt der Zerfallsprozeß zum Stillstand, und es setzt eine „Restauration" des Lexikons ein (m.a.W., wenn $\eta_o > 0$, hat die Funktion (2.30) ein Minimum; wenn $\eta_o = 0$, wächst die Funktion). Einer solchen Modellvorstellung entspricht natürlich nichts in der Realität.

Die Fälle der beschleunigten und verzögerten lexikalischen Entwicklung sind also nicht symmetrisch zueinander. Der Zerfallsprozeß kann sich verlangsamen, er kommt jedoch nie zum Stillstand, d.h. die entsprechende Funktion muß asymptotisch gegen Null streben. Geeignet wäre z.B. die Funktion

$$\eta(t) = -\frac{\gamma}{2\sqrt{t}} \qquad (2.31)$$

oder

$$\eta(t) = -\gamma t^{-1}. \qquad (2.32)$$

Die Funktion (2.31) ergibt die Formel

$$\bar{N}(t) = 1 - e^{-\gamma\sqrt{t}}. \qquad (2.33)$$

Betrachten wir diesen Ausdruck ausführlicher. Wir stellen fest, daß die Zeit t in (2.33) mit dem gleichen Exponenten $(\frac{1}{2})$ wie der Rang i (vgl. (2.15)) vorkommt. Hier liegt die gleiche Dualität von Zeit und Rang vor, von der auch zu Ende von 2.3 die Rede war. Wir haben nun einerseits eine Formel, in der t linear im Exponenten steht, einmal als Quadratwurzel und einmal als Quadrat. Der Rang dagegen tritt im untersuchten Fall des rangierten Lexikons als Quadratwurzel im Exponenten auf.

Aus Symmetrieerwägungen bezüglich i und t kann man die Annahme ableiten, daß bei anderer Rangierungsmethode oder stärkeren Anforderungen an die Korrespondenz der Anteil der erhaltenen (geschwundenen) Wörter auch in anderer Form vom Rang abhängen kann, z.B.:

Zeit im Exponenten	Rang im Exponenten
t	i
t^2	i^2
$\sqrt{t}$	$\sqrt{i}$.

Ein Wortformenlexikon zerfällt wahrscheinlich schneller als ein Wörter-(Lexem-)Lexikon, und wenn wir sogar fordern, daß eine Wortform ausschließlich ihre Grundbedeutung bewahren darf, beschleunigt sich der Zerfallsprozeß noch. Auf der anderen Seite kann der Zerfall langsamer sein, wenn man nicht die Wörter, sondern die Morpheme einer Sprache betrachtet. Wir verfügen zur Zeit nicht über empirische Daten, die diese Hypothesen stützen oder schwächen könnten. Indirekt sprechen für die Hypothese der qualitativen Veränderung des Evolutionsprozesses beim Übergang zu einer anderen Sprachebene die Daten in Zipf [1] zur Verteilung von englischen Wörtern nach ihrer Entstehungszeit in Abhängigkeit von ihrer Länge in Silben. Dem Rang i entspricht dort die Silbenzahl. Auch wenn wir für diese Daten einen kleinen Maßstab gewählt haben (s. Abb. 15b), und nur die Wörter der Häufigkeit 1 aus dem Wörterbuch von Eldridge (s. Tab. 1) erfassen, ist die globale Tendenz sichtbar.

Wenn sich die Annahme der Symmetrie von Rang und Zeit erhärten läßt, dann wäre eine der möglichen Interpretationen die Dualität von Sprachstil und Sprachniveau: eine Änderung des Sprachstils entspricht dem Übergang zur Betrachtung eines anderen Sprachniveaus.

Abschließend betrachten wir die Funktion (2.32), die zwar praktisch kaum möglich ist, aber einen interessanten Entartungsfall darstellt. Es ist ja

$$\int_1^y ax^{-1}dx = a \ln y.$$

Tabelle 1. Verteilung der Lexik nach dem Alter in Abhängigkeit von der Wortlänge (in Silben)

Silbenzahl im Wort i	Anteil von Wörtern, die 800 J. und länger in der Sprache sind (altenglische Lexik)	Wert der Funktion $N(t)=e^{-\beta it}$ bei $\beta=0,80$	Anteil von Wörtern, die 500 J. und länger in der Sprache sind (mittelenglische Lexik)	Wert der Funktion $N(t)=e^{-\beta it}$ bei $\beta = 0,22$
1	0,45	0,43	0,81	0,80
2	0,18	0,18	0,68	0,64
3	0,03	0,07	0,52	0,52
4	0,01	0,03	0,40	0,41
5	0,00	0,02	0,30	0,33

Der „Weg" hängt im Fall (2.32) logarithmisch von t ab:
$s = -\gamma \ln t$.

Durch Einsetzen von (2.33) in (2.29) bekommen wir die folgende Formel für den Anteil der erhaltenen Lexik:

$$N(t) = t^{-\gamma}.$$

Diese Funktion ähnelt dem bekannten Zipfschen Gesetz, und die Zeit verhält sich in (2.34) wie der Rang in diesem.

3. EMPIRISCHE ÜBERPRÜFUNG DER THEORIE DES LEXIKALISCHEN WANDELS: VERERBUNG

3.1. TESTAUFBAU

Im folgenden Kapitel sollen die Schlußfolgerungen aus dem in Kapitel 2 dargestellten Modell des lexikalischen Wandels in der Zeit empirisch überprüft werden. Im einzelnen werden wir die folgenden Hypothesen testen:

Erstens interessiert uns der Zusammenhang zwischen dem Rang i und der Entstehungszeit t eines Wortes. Aus unserer Theorie folgt (vgl. Formel (2.11)), daß für ein Wort a_i die Wahrscheinlichkeit, nach einem bestimmten Zeitpunkt t der Vergangenheit entstanden zu sein, umso größer ist, je höher sein Rang i (- in unserer konkreten Interpretation - je geringer seine Häufigkeit) ist:

$$\bar{F}_i(t) = 1 - e^{-\eta_t \sqrt{i}}, \qquad (3.1)$$

wobei η_t konstant ist (jedem t, bzw. jedem Zeitintervall von t bis zum Zeitpunkt der Erstellung des rangierten Lexikons, entspricht natürlich ein eigenes η_t). Leichter als (3.1) läßt sich die duale Funktion

$$F_i(t) = e^{-\eta_t \sqrt{i}}, \qquad (3.2)$$

empirisch überprüfen, die die Wahrscheinlichkeit, daß ein Wort vor einem bestimmten Zeitpunkt t auftritt, beschreibt. Von (2.11) unterscheidet sich (3.2) darin, daß wir hier F als Funktion des Rangs i betrachten und dabei vom Einfluß der Zeit absehen. Während wir bei der Ableitung der entsprechenden Formeln ((2.11) etc.) zunächst die Abhängigkeit des Schwunds von der Zeit zugrunde gelegt haben und erst danach seine Abhängigkeit vom Rang, ist es zur Überprüfung dieser theoretischen Annahmen einfacher, in umgekehrter

Reihenfolge vorzugehen, d.h. erst die Abhängigkeit von F vom Rang und dann von der Zeit zu betrachten.

Zur Vereinfachung der Daten zerlegen wir einen Anfangsabschnitt des Häufigkeitswörterbuchs in Klassen gleichen Umfangs n (wir wählen für n immer 100) und ordnen allen Wörtern der Plätze 1 bis 100, 101 bis 200...nk + 1 bis n(k+1) jeweils den gleichen Rang zu (der Rang ist also die Platzziffer einer Klasse von 100 aufeinanderfolgenden Wörtern).

Also wird tatsächlich die Funktion

$$F_i(t) = ne^{-\eta_t\sqrt{i}} \qquad (3.3)$$

überprüft.

Bei der Testreihe, die in 3.2 näher beschrieben wird, bestimmen wir für jede Klasse eines konkreten Häufigkeitswörterbuchs V_x die Anzahl der Wörter, die spätestens zum Zeitpunkt $t_1, t_2, \ldots, t_k \ldots \ldots, t_n$ aufgetreten sind. Sodann betrachten wir für jedes einzelne t_k den entsprechenden „Durchschnitt". Für jeden Durchschnitt ist nun zu prüfen, ob die beobachtete Reihe von der Verteilungsfunktion (3.3) approximiert wird, deren Parameter η_t mit Hilfe der Methode der kleinsten Fehlerquadrate bestimmt wird. Wo sich die Übereinstimmung dieser theoretischen Verteilung mit den empirischen Daten nicht rein graphisch entscheiden ließ, haben wir den χ^2- Test verwendet.

Zweitens können wir mit den Werten des Parameters η_t für verschiedene t und verschiedene Lexika aus der ersten Testreihe dann in 3.3 die Abhängigkeit von η_t von der Zeit - $\eta = \eta(t)$ - untersuchen. Insbesondere gilt, daß, wenn die Geschwindigkeit des lexikalischen Wandels konstant ist, η_t von t linear abhängen muß. In 2.5 haben wir auch die Möglichkeit eines nichtlinearen Zusammenhangs von η_t und t in Betracht gezogen. Die Form dieser Abhängigkeit wird unten für Lexika verschiedener Sprachstile einzeln untersucht.

Drittens wird in 3.4 der Fall zweier verwandter Sprachen, die auf eine gemeinsame Protosprache zurückgehen, behandelt. Am Beispiel des Russischen und des Tschechischen wird für die Wörter, die in zwei verwandten Sprachen vorkommen, ihre Verteilung in rangierten Lexika untersucht.

Unsere Theorie behauptet, daß der Anteil dieser Wörter in beiden Lexika von Klasse zu Klasse nach einer (3.3) entsprechenden Gesetzmäßigkeit abnimmt. Zur Verifikation dieser Behauptung müssen wir nicht nur nachweisen, daß die Verteilung der gemeinsamen Wörter tatsächlich der Formel (3.3) genügt, sondern auch, daß sie in beiden Lexika dieselben Parameter hat.

Dazu sind wir schon deshalb verpflichtet, weil die Gleichheit der Parameter aus der Definition der Verwandtschaft als symmetrischer Relation folgt[1].

Wir können auch die Resultate der zwei Datierungsmöglichkeiten für eine Protosprache vergleichen: die Datierung durch Extrapolation der Daten zur lexikalischen Entwicklung einer Einzelsprache über einen historischen Zeitraum und die Datierung mittels des Anteils gemeinsamer Wörter in Sprachen, die auf diese Protosprache zurückgehen. Dabei wird überzeugend bestätigt, daß die Formel (2.23) für den Anteil gemeinsamer Wörter in Sprachen, die sich zum Zeitpunkt t getrennt haben, (- dieser Anteil muß für jede Klasse wesentlich geringer sein als die Anzahl der Wörter, die in jeder einzelnen Sprache seit t erhalten sind -) nur eine Minimalschätzung liefert, wobei die Möglichkeit von Parallelentwicklungen überhaupt nicht berücksichtigt ist.

Viertens schließlich erörtern wir im letzten Paragraphen (3.5) die Möglichkeiten, die gefundenen Gesetzmäßigkeiten des lexikalischen Wandels für absolute Datierungen nutzbar zu machen, und das Verhältnis unserer Methode zum glottochronologischen Ansatz.

3.2. DER ZUSAMMENHANG VON LEBENSERWARTUNG UND RANG VON WÖRTERN

Bei der Beschreibung der empirischen Untersuchungen zum Zusammenhang zwischen der Anzahl der Wörter, die in einer Sprache zu einem bestimmten Zeitpunkt noch existiert haben, und dem Rang der jeweiligen Klasse im Häufigkeitswörterbuch werden wir wie folgt vorgehen: zunächst beschreiben wir kurz die Regeln, denen

wir bei der Bearbeitung aller benutzten Häufigkeitswörterbücher gefolgt sind, dann werden Quellen angeführt und Besonderheiten der Daten für einzelne Sprachen angemerkt. Schließlich werden die Ergebnisse tabellarisch zusammengefaßt und mit Histogrammen u.ä. graphisch illustriert.

Diese Ergebnisse werden in 3.3 zusammen mit denen zum Zusammenhang zwischen der Lebenserwartung von Wörtern und der Zeit diskutiert.

Die Hauptschwierigkeit bei den hier dargestellten empirischen Überprüfungen besteht darin, vergleichbare Daten zur Geschwindigkeit des lexikalischen Wandels für verschiedene Sprachen zu gewinnen, da dafür mindestens zwei Bedingungen erfüllt sein müssen:

- Für jede Sprache muß man über ein Lexikon verfügen, das die Häufigkeit von Wörtern in einem Textkorpus wiedergibt, das als für die ganze Sprache repräsentativ angesehen werden kann; diese Texte müssen zu einem bekannten und nicht zu großen Zeitintervall (z.B. 50 Jahre) gehören.
- Die jeweiligen Sprachen müssen alle historisch-lexikographisch so weit erforscht sein, daß das Auftreten eines neuen Wortes zeitlich hinlänglich (z.B. bis auf ein Jahrhundert) genau fixierbar ist.

Da diese beiden Forderungen nicht nur nicht für alle betrachteten Sprachen sondern faktisch für keine einzige erfüllt sind, sind die Daten zur Geschwindigkeit des lexikalischen Wandels in den verschiedenen Sprachen direkt nicht vergleichbar. Ein Vergleich ist nur nach zahlreichen, qualitativ unbestimmten Korrekturen durchführbar. Mann kann nur hoffen, daß die Erreichung größerer Genauigkeit keine prinzipiellen Schwierigkeiten bereiten wird, wenn sich erst einmal die Grundidee der vorgeschlagenen Methode als fruchtbar erwiesen hat. Diese Schwierigkeiten müssen nicht unbedingt durch rein quantitative Anstrengungen überwunden werden, z.B. durch die Erstellung von Häufigkeitswörterbüchern auf der Grundlage riesiger Stichproben. Ein anderer Ausweg wäre die Verwendung mehrerer Häufigkeitswörterbücher für eine Sprache, die jeweils verschiedene Stile wiedergeben, die Anwendung verschiedener Kriterien für die lexikalische Lebenserwartung, verschiedener Rangierungsprinzipien u.s.w.

Bei unseren Untersuchungen haben wir uns um Vereinheitlichung bemüht bei

- dem Typ des Häufigkeitswörterbuchs,
- dem Kriterium für die Lebenserwartung (bzw. für den Schwund) und
- den Erhebungsmethoden und Darstellungsweisen.

1. Zum Typ des Wörterbuchs

Alle benutzten Wörterbücher sind Häufigkeitslexika von Wörtern und nicht von Wortformen oder Wurzeln. In der Verwendung des Wortbegriffs weichen die Wörterbücher nicht wesentlich voneinander ab, mit Ausnahme des englischen Wörterbuchs von Dewey (s.u.), wo verschiedene Verbstämme als verschiedene Wörter gewertet, Homonyme verschiedener Wortarten (Konversive) nicht immer unterschieden werden. Bei der Stichprobengröße, die für die verschiedenen Wörterbücher zwischen 100 000 und einigen Millionen von Wortformen schwankt, konnten wir die nötige Einheitlichkeit nicht erreichen, ebensowenig bei den Auswahlprinzipien der Quellen bezüglich Sprachgattung und Stil.

Der Zusammenhang von Wörterbuchtyp und Ergebnissen wird in 3.3 erörtert.

2. Zum Kriterium für die Lebenserwartung (für den Schwund)

In 2.1 haben wir uns nur mit der prinzipiellen Möglichkeit beschäftigt, Entsprechungen zwischen den Wörtern zweier Sprachen x und y herzustellen; uns interessierte besonders der Fall, wo $x \geq y$, und y bis zu einem vergangenen Zeitpunkt t existierte. Bei der praktischen Aufstellung solcher Entsprechungen kann man verschiedene Kriterien anwenden, z.B. kann man folgende Prinzipien aufstellen:

1° Jedes Morphem des Wortes $a \in V_y$ hat eine Entsprechung in der Sprache x. So entspricht z.B. jedem Morphem von russ. pro-my^{y}sl-ennost' ein urslavisches Morphem derselben Lautgestalt und Bedeutung, es existiert aber im Urslavischen keine Entsprechung zu dieser Morphemkombination, die bekanntlich eine Schöpfung von N. Karamzin ist.

2° Der Kombination der Morpheme, die das Wort a ∈ V_y bilden, entspricht in V_x eine Kombination a', wobei zwischen den einzelnen Morphemen reguläre Entsprechungen existieren. So entspricht russ. chleb altruss. chlěbъ, wobei dem Nullmorphem des ersten Wortes das Morphem -ъ im zweiten entspricht. Die Entsprechung braucht nicht derart trivial zu sein, wenn sie nur regulär ist und den jeweiligen morphologischen Regeln entspricht. Den adjektivischen Langformen des Russischen entsprechen die adjektivischen Kurzformen des Urslavischen regulär. Die Regularität der Entsprechung beim Substantiv wird z.B. nicht zerstört durch das Verschwinden des sog. Themas als selbständiges Morphem (urslav. ronk-ā-mь hat drei Morpheme, russ. ruk-am hat zwei), oder durch die Verallgemeinerung des Stamms der obliquen Kasus auf den Nominativ (altruss. kolo Plur. kolesa, russ. koleso) u.ä. Wir könnten die Prinzipien 1° und 2° erschöpfend durch Aufzählung der infrage kommenden Regeln der historischen Grammatik der jeweiligen Sprachen (oder der entsprechenden „Entlehnungsgrammatiken", s. 2.1 spezifizieren, was jedoch überflüssig wäre.

3° Die Entsprechungen a ∈ V_x und a' ∈ V_y müssen bedeutungsgleich sein.

Natürlich kann man 1° bis 3° enger oder weiter fassen. Anstelle von 3° kann man z.B. Erhalt der Grundbedeutung oder Kontinuität des Bedeutungswandels u.ä. fordern. Man kann solche unwesentlichen Neuerungen wie den Wechsel der Akzentstelle (sofern er nicht unter eine reguläre phonetische Gesetzmäßigkeit fällt, wie im Fall von russ. instrúment und später instrumént im XIX. Jh.) oder eine Veränderung des Suffixes (russ. domok, später domik) vernachlässigen oder auch nicht.

In Arbeiten zur Glottochronologie wird offenbar davon ausgegangen, daß zwei Wörter Entsprechungen voneinander sind, wenn die Bedingungen 1° und 3° erfüllt sind. In der Arbeit [16] von I. Fodor z.B. gelten als Entsprechungen urslav. žena (= tschech. žena) und russ. ženščina, das urslavische Adjektiv mъnogъ und das russische Adverb mnogo, urslav. nizъ und russ. vniz, květъ und cvetok u.s.w..

Diese Entsprechungen sind vor allem morphologisch nicht exakt. Andererseits wird angenommen, daß solche urslavischen Wörter wie pъrchъ „Staub", chrъbtъ „Rücken" oder onja „Geruch" keine russischen Entsprechungen haben, weil russ. poroch die Bedeutung „Staub" verloren hat, und chrebet und von' zu stark spezialisierte Bedeutungen haben, die sich nicht mit den wesentlich weiteren Bedeutungen der urslavischen Wörter decken.

Es wäre vielleicht möglich, das beschriebene Kriterium konsequent anzuwenden, in der Praxis aber, bei der Beschreibung größerer Wortmengen (für diese Arbeit sind Daten zur Geschichte von 20000 Wörtern aus verschiedenen Sprachen gesammelt worden), erweist sich ein Kriterium als günstiger, das von der Mehrheit der Autoren etymologischer (das besonders im XX. Jh.) und historischer Wörterbücher und auch in der vorliegenden Arbeit verwendet wird:

Zwei Wörter a ∈ V_x und a' ∈ V_y gelten als Entsprechungen (als gleich), wenn die Bedingungen 1° und 2° erfüllt sind. Zwei Wörter dürfen sich daher nur im Rahmen der jeweiligen phonetischen und morphologischen Korrespondenzregeln unterscheiden; Bedeutungsunterschiede, auch substantielle, werden nicht berücksichtigt.

Die Verschiedenheit der Kriterien bildet ein zusätzliches Hindernis für den Vergleich unserer Resultate mit den glottochronologischen.

3. Erhebungsmethoden und Darstellungsweise der Ergebnisse

Es wurden Lexikonausschnitte von 600 bis 6000 Wörtern untersucht. Dabei wurde festgestellt, daß zur Bestimmung der Parameter der Verteilung (bei $t \geq 300$-400 Jahre) die ersten tausend Wörter (10 Klassen zu je 100 Wörtern) ausreichen; eine weitere Vergrößerung des Stichprobenumfangs erhöht die Genauigkeit praktisch nicht mehr, da diese viel stärker von der Verfügbarkeit ausführlicher historischer und etymologischer Daten zu den Wörtern und von der Struktur des Häufigkeitswörterbuchs abhängt. Für die jüngste Geschichte (die letzten 300-400 Jahre) der Literatursprache (genauer: der Sprache der Belletristik und verwandter Gattungen) ist aufgrund einiger Besonderheiten ihrer Entwicklung (s. dazu 3.3) die Untersuchung wesentlich größerer Lexikonausschnitte oder ein

strengeres Korrespondenzkriterium (z.B. mit Berücksichtigung von Bedeutungsveränderungen) erforderlich.

Gewisse Probleme tauchen im Zusammenhang mit der optimalen Klassengröße auf; zu große Klassen „verwischen" die Verteilung und erschweren die Bestimmung ihrer Parameter, zu kleine erhöhen dagegen die Streuung. Die gewählte Klassenbreite von 100 Wörtern ist für unsere Fragestellung gut geeignet, sie muß jedoch evtl. dem Untersuchungsziel entsprechend modifiziert werden.[2)]

Wir sehen hier auch davon ab, daß bei Wörterbüchern, die auf relativ kleinen Stichproben basieren, der Fall eintreten kann, daß zwei verschiedene Rangklassen Wörter mit gleicher Häufigkeit enthalten (die sich nur in der Stelle der lexikographischen Ordnung unterscheiden); diese Unregelmäßigkeiten verschwinden normalerweise bei der Glättung der beobachteten Kurve.

Die Resultate der Erhebungen werden vor allem tabellarisch dargestellt, wobei der Anzahl von Wörtern der laufenden Klasse, die zu dem betrachteten Intervall gehören, jeweils die theoretisch erwartete Anzahl gegenübergestellt wird. Der nach der Methode der kleinsten Fehlerquadrate geschätzte Parameter η_t für Formel (3.3) wird über der Spalte angegeben. Diese Tabellen haben die folgende Form:

	(1) Untersuchte Sprache	
	(3) Zeitintervall	
(2) Rang der Klasse	(4) Wert von η_t nach der Methode der kleinsten Fehlerquadrate	
	(5) Beobachtete Werte von $F_i(t)$	(6) theoretische Werte von $F_i(t)$

Die Auswahl der Zeitintervalle ist im wesentlichen durch das uns zur Verfügung stehende lexikographische Material motiviert, welches häufig die traditionelle Periodisierung der jeweiligen Sprachgeschichte widerspiegelt (der Schnitt bei 1500 für das Russische ist z.B. deshalb gewählt worden, weil das Wörterbuch von Sreznevskij

in der Hauptsache Denkmäler vor diesem Zeitpunkt repräsentiert; ebenso ist ein Schnitt um 1500 für das Deutsche anzusetzen, weil die chronologischen Anmerkungen in historischen und etymologischen Wörterbüchern des Deutschen für gewöhnlich einen Hinweis auf die Entstehung eines Wortes in der mittelhochdeutschen Epoche geben, deren Ende um 1500 anzunehmen ist).

Die Histogramme der Verteilungen erhaltener Wörter für verschiedene Zeitintervalle mit den eingezeichneten theoretischen Kurven bedürfen keiner besonderen Erläuterung. Für andere Zeitintervalle wurden die Graphen durch Wahl einer logarithmischen Skala in der Ordinate (O_y) und einer Quadratwurzel-Skala in der Abszisse (O_x) linearisiert.

Als Beispiel führen wir zwei Ausschnitte (zwei Klassen à 100 Wörter) aus Häufigkeitswörterbüchern des Russischen und Französischen in Tabellenform an (s. Tab. 2 und 3). In der äußerst linken Spalte ist die laufende Nummer des Worts im Häufigkeitswörterbuch angegeben. Die Spalten der Tabellen entsprechen den Zeitintervallen, die für das jeweilige Lexikon ausgewählt wurden. Für jedes Wort ist das Zeitintervall seines ersten Auftretens (mit „+") markiert. So bedeutet das Pluszeichen zu russ. bolezn' in der Spalte „1500", daß dieses Wort vor 1500 (aber nach 600) aufgetreten ist.

Am Fuß der Tabelle ist die Anzahl der Wörter angegeben, die einem Zeitabschnitt angehören (in Klammern), und die Anzahl der Wörter, die einem Zeitabschnitt und auch allen vorausgehenden Abschnitten angehören. Diese Zahlen sind das Ergebnis der Datenerhebung und dienen als Ausgangsmaterial für die statistische Analyse. Die äußerst rechte Spalte enthält Vermerke zu den Entlehnungsquellen. Wörter ohne einen solchen Vermerk sind entweder Erbwörter oder „autochthone" Neubildungen.

Das erste Beispiel sind die Wörter 401 bis 500 (i=5) im Castotnyj slovar' obščenaučnoj leksiki (ONL, s.u.). Das zweite sind die Wörter 501 bis 600 (i=6) aus dem Häufigkeitswörterbuch des Französischen von Gougenheim et al. (s.u.).

Tabelle 2. Ausschnitt eines russischen Häufigkeitswörterbuchs ONL) mit datierter Lexik

Nr.	Wort	Zeit (Jahre n. Chr.)					Entlehnungs-quelle
		600	1500	1600	1800	1900	
401	važnyj	+					
402	v kačestve					+	
403	vozmožnyj		+				
404	god	+					
405	nesmotrja			+			
406	nižnij	+					
407	spektr					+	lat.
408	značitel'no		+				
409	komponent						lat.
410	kompleks						lat.
411	oboznačit'					+	
412	osobenno			+			
413	posledo-vatel'nost'					+	
414	pričina		+				
415	osnova	+					
416	otličat'sja		+				
417	ploščad'		+				aksl.
418	zadat'		+				
419	koren'	+					
420	ob-ekt					+	lat.
421	fakt				+		lat.
422	bolezn'		+				
423	dejstvovat'		+				
424	izvestno		+				
425	polučat'sja		+	+			
426	pribor			+			
427	perechod	+					
428	princip					+	
429	soderžat'sja		+				lat.
430	sposobnost'		+				

Tabelle 2. (Fortsetzung)

Nr.	Wort	Zeit (Jahre n. Chr.)					Entlehnungs-quelle
		600	1500	1600	1800	1900	
431	élektrod					+	griech.
432	élementarnyj					+	lat.
433	nebol'šoj		+				
434	usilitel'						
435	mnogočlen					+	frz. (Lehnübers.)
436	nastojaščij		+				aksl.
437	pokazyvat'	+					
438	prinadležat'			+			
439	vodorod					+	lat. (Lehnübers.)
440	vstrečat'sja		+				
441	okazyvat'sja		+				
442	raznyj		+				
443	charakternyj					+	aksl.
444	svet	+					
445	libo		+				
446	sloj	+					
447	luč	+					
448	odnorodnyj				+		lat. (Lehnübers.)
449	raspoložit'			+			
450	ostanovit'sja		+				
451	žizn'		+				
452	obyčnyj		+				
453	soglasno		+				
454	fizičeskij				+		aksl.?
455	absoljutnyj					+	dt.
456	vyčislenie				+		
457	kakoj-libo		+				
458	material				+		dt.
459	produkt					+	lat.
460	cep'	+					

Tabelle 2. (Fortsetzung)

Nr.	Wort	Zeit (Jahre n. Chr.)					Entlehnungs-quelle
		600	1500	1600	1800	1900	
461	chotja		+				
462	predel'nyj			+			aksl.
463	klass				+		lat.
464	kristall				+		griech.
465	polost'					+	
466	ėlektronnyj						
467	ni	+					
468	sosud		+				aksl.
469	dal'nejšij		+				
470	lampa			+			dt.
471	matematičeskij				+		
472	odinakovyj				+		
473	otmetit'		+				
474	otricatel'nyj			+			
475	raznost'		+				aksl.
476	okulo	+					
477	kol'co		+				
478	organičeskij					+	
479	učastok		+				
480	zametit'	+					
481	imenno		+				
482	sečenie				+		dt. (Lehnübers.)
483	dostigat'		+				aksl.
484	chod	+					
485	izlučenie						
486	neobchodimyj			+			
487	ravnomernyj				+		aksl.
488	atomnyj				+		
489	vchodnoj		+				
490	vysota		+				

Tabelle 2. (Fortsetzung)

Nr.	Wort	Zeit (Jahre n. Chr.)					Entlehnungs-quelle
		600	1500	1600	1800	1900	
491	idti	+					
492	okazat'sja	+					
493	točnyj	+					
494	analitičeskij				+		frz.
495	v slučae					+	
496	vnutri	+					
497	dejstvitel'nyj			+			
498	osadok					+	
499	perechodit'		+				
500	radius					+	lat.
	Insgesamt:	19(19)	54(35)	64(10)	77(13)	95(18)	

Fünf Wörter - komponent, kompleks, usilitel', ėlektronnyj, izlučenie - sind erst im 20. Jh. aufgetreten.

Tabelle 3. Ausschnitt eines französischen Häufigkeitswörterbuchs (Goug) mit datierter Lexik

Nr.	Wort	Zeit (Jahre n. Chr.)							Entlehnungs-quelle
		400	700	1200	1300	1400	1500	1600 1700	
501	état				+				lat.
502	cuisîne	+							
503	préparer					+			lat.
504	eh							+	
505	chien	+							
506	medecin			+					lat.
507	en train de							+	
508	enlever			+					
509	dehors			+					

Tabelle 3. (Fortsetzung)

Nr.	Wort	Zeit (Jahre n. Chr.)								Entlehnungs-quelle
		400	700	1200	1300	1400	1500	1600	1700	
510	porte	+								
511	droite (nom.)	+								
512	endroite							+		
513	lever	+								
514	permettre							+		lat.
515	air (atmosphere)	+								
516	samedi		+							
517	chanter	+								
518	pour que								+	
519	coeur								+	
520	meilleur	+								
521	quoi	+								lat.
522	habitude						+			
523	tas			+						span.
524	prochain (adj.)		+							
525	camarade							+		
526	amuser			+						
527	bureau							+		
528	campagne							+		ital.
529	rire	+								
530	boite		+							lat.
531	remettre			+						
532	deuxième						+			
533	quart	+								
534	montrer	+								
535	hiver	+								
536	frère	+								
537	ceci		+							
538	attention							+		lat.
539	retrouver			+						

Tabelle 3. (Fortsetzung)

Nr.	Wort	Zeit (Jahre n. Chr.)								Entlehnungs-quelle
		400	700	1200	1300	1400	1500	1600	1700	
540	asseoir		+							
541	effet				+					lat.
542	juste (adv.)								+	
543	retourner			+						
544	là dedans								+	
545	chacun		+							
546	chef	+								
547	théâtre					+				lat.
548	public					+				lat.
549	aujourd'hui				+					
550	milieu			+						
551	intérieur						+			lat.
552	expliquer					+				
553	voyage			+						
554	sorte	+								
555	quartier		+							
556	n'importe								+	
557	l'(dans:l'on)									
558	machine					+				lat.
559	magasin						+			ital.
560	animal (nom.)				+					lat.
561	moyen	+								
562	autour						+			
563	noir	+								
564	different					+				
565	emmener			+						
566	particulier					+				lat.
567	micro									
568	hotel	+								
569	repas								+	
570	voix	+								

Tabelle 3. (Fortsetzung)

Nr.	Wort	Zeit (Jahre n. Chr.)								Entlehnungs-quelle
		400	700	1200	1300	1400	1500	1600	1700	
571	suite		+							
572	heureux				+					
573	chauffer		+							
574	mademoiselle							+		
575	lieu	+								
576	blanc			+						
577	air							+		
578	quatorze	+								
579	papa							+		
580	lundi		+							
581	disque							+		lat.
582	neige					+				
583	instituteur				+					lat.
584	musique			+						
585	manquer							+		ital.
586	aider	+								
587	dépendre			+						lat.
588	jardin				+					
589	fou	+								
590	modern					+				lat.
591	italien							+		
592	directeur						+			
593	tuer	+								
594	confiance				+					lat.
595	réussir							+		
596	pas du tout							+		
597	cours								+	ital.
598	curieux			+						lat.
599	onze	+								
600	appartement							+		ital.
		27 (27)	37 (10)	52 (15)	60 (8)	70 (10)	76 (6)	91 (15)	98 (7)	

Zwei Wörter sind später entstanden : l' (in l'on) 18. Jh., micro im 20. Jh.

Wir beschreiben nun unsere Daten zu den einzelnen Sprachen. Jedesmal werden folgende Informationen angegeben:

A) benutzte Frequenzwörterbücher,

B) lexikographische Quellen für die Datierung,

C) eine Tabelle, in der zu jedem Lexikon die unterscheidbaren Zeitabschnitte aufgeführt sind, mit entsprechenden Verweisen auf die Tabellen und Diagramme mit den numerischen Daten zur Verteilung in jedem Abschnitt. Die Tabelle C ist folgendermaßen aufgebaut:

(1) Quelle	(2) Zeitabschnitt (Jahre)	(3) Zeitintervall von t bis zur Erstellung des Wörterbuchs (in Jh.)	(4) Tabelle der Verteilung $F_i(t)$	Abbildungen der (5) beobachteten (6) theoretischen Verteilung $F_i(t)$

1. Russisch

A. É.A. Štejnfel'd: Častotnyj slovar' sovremennogo russkogo literaturnogo jazyka. Tallin 1963 (Štf) - 2500 Wörter.
Častotnyj slovar' obščenaučnoj leksiki pod red. E.M. Stepanovoj. Moskau, Izd-vo MGU 1970 (ONL) - 2000 Wörter.

B. Slovar' sovremennogo russkogo jazyka v 17 tomach. Moskau-Leningrad 1960;
I.I. Sreznevskij: Materialy dlja slovarja drevnerusskogo jazyka, Bd. I-III St. Petersburg 1893;
M. Fasmer: Ėtimologičeskij slovar' russkogo jazyka, B. III. Moskau 1964 f. (M. Vasmer: Russisches etymologisches Wörterbuch, Bd. 1-3);
N.M. Šanskij, V.V. Ivanov, T.V. Šanskaja: Kratkij ėtimologičeskij slovar' russkogo jazyka. Moskau 1961; Ėtimologičeskij slovar' russkogo jazyka pod red. N.M. Šanskogo, Lieferung A-V. Moskau, Izd.vo MGU 1964;

(1)*	(2)	(3)	(4)	(5)
Štf	1500	4,5	4	4 u. 5
"	600 ursl.	13,5	4	4 u. 5
ONL	1900	0,5	5	-
"	1800	1,5	5	7
"	1600	3,5	5	6
"	1500	4,5	5	7
"	600 ursl.	13,5	5	6

*Als Entlehnungsdatum $\bar{t}$ der Quellen, nach denen das Wörterbuch verfaßt wurde, ist 1950 festgesetzt worden (Abb. 4-7).

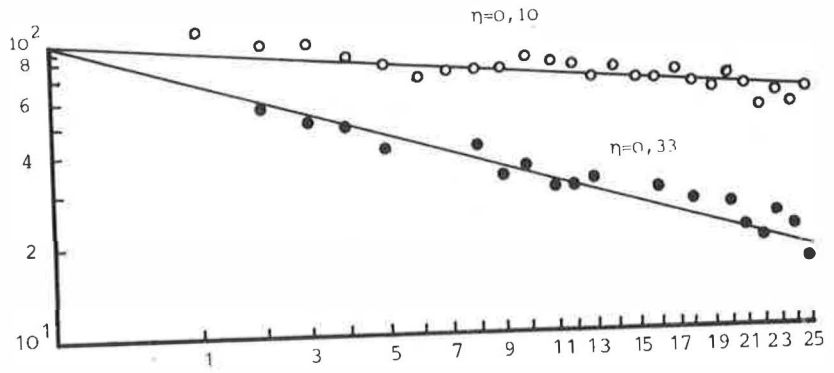

Abb. 4. „Linearisierter" Graph von $F_t(i)$ für das Russische (Štf) (oben: ab 1500; unten: ab 600 - „Urslavisch"). Logarithmische Skala in O_y und Quadratwurzel-Skala in O_x.

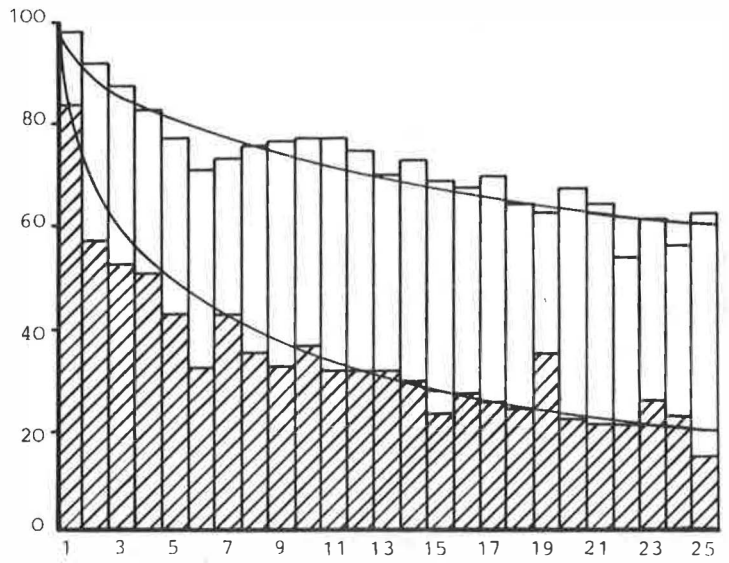

Abb. 5. Graph der Funktion $F_t(i)$ für das Russische (Štf). Oben: ab 1500, unten: seit dem „Urslavischen" (ca. 600)

2. Tschechisch

A. Jelínek, J. Bečka, M. Těšitělová: Frekvence slov, slovních druhů a tvarů v českém jazyce. Prag 1961 (Čech) - enthält insgesamt 5000 Wörter, von denen hier die ersten 2500 verwendet wurden.

B. J. Gebauer: Slovník staročeský, díl I-II. Prag 1903.
Fr. Šimek: Slovníček staré čestiny. Prag 1947.
Der alttschechische Kapitalpsalter. Prag 1928.
K. Novák: Slovník k českým spisům Husovým. Archiv pro lexikografii a dialektologii, sv. 9. Prag 1934.
J. Holub, F. Kopečný: Etymologický slovník jazyka českého. Prag 1952.
V. Machek: Etymologický slovník jazyka českého. Prag 1968.

J. Holub, S. Lyer: Stručný etymologický slovník. Prag 1967.
N. Retter: Die deutsche Lehnübersetzung im Tschechischen. Berlin 1953.

C.

(1)	(2)	(3)*	(4)	(5)
Čech.	1500	4,5	4	8 u. 9
"	600 urslav.	13,5	4	8

*$\bar{t}$ = 1950 (Abb. 8 u. 9)

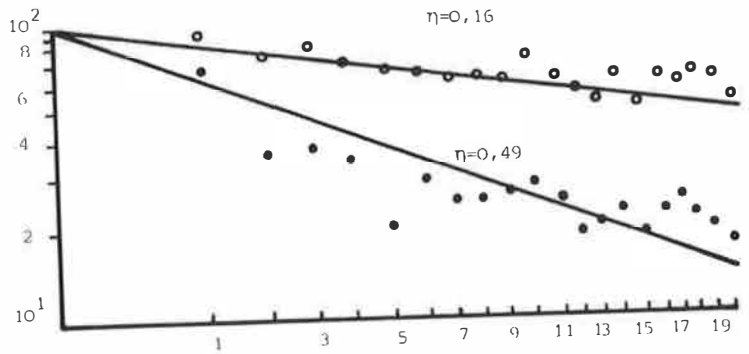

Abb. 6. Linearisierter Graph der Funktion $F_t(i)$ für das Russische (ONL). Oben: ab 1600, unten: ab 600. Logarithmische Skala in O_y und Quadratwurzelskala in O_x.

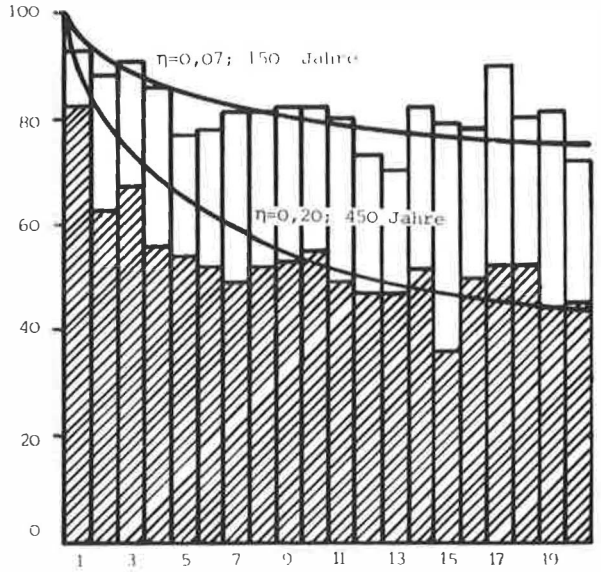

Abb. 7. Graph der Funktion $F_t(i)$ für das Russische (ONL). Oben: ab 1800, unten: ab 1500.

Tabelle 4. Beobachtete und theoretische Verteilungsfunktion $F_i(t)$ für Russisch und Tschechisch

(2) Rang	(1) Russisch (Štf)				Tschechisch			
	(3) Urslav. 600		1500		Urslav. (600)		1500	
	(4) η=0,33		η=0,10		η=0,34		η=0,11	
	(5)	(6)	(5)	(6)	(5)	(6)	(5)	(6)
1 (1-100)	84	72	99	91	75	71	100	90
2 (101-200)	57	62	92	87	63	62	96	85
3 (202-200)	52	57	88	84	50	55	85	73
4 (301-400)	51	52	83	82	43	51	77	80
5 (401-500)	42	48	88	80	37	47	66	78
6 (501-600)	32	44	71	78	36	44	73	76
7 (601-700)	42	42	74	77	45	41	76	75
8 (701-800)	35	39	75	76	42	38	73	73
9 (801-900)	33	37	76	74	32	36	72	72
10 (901-1000)	35	35	77	72	32	34	63	70

Tabelle 4. (Fortsetzung)

(2) Rang	(1) Russisch (Štf)				Tschechisch			
	(3) Urslav. 600		1500		Urslav. (600)		1500	
	(4)	η=0,33	η=0,10		η=0,34		η=0,11	
	(5)	(6)	(5)	(6)	(5)	(6)	(5)	(6)
11(1001-1100)	31	33	77	72	29	32	66	69
12(1101-1200)	31	32	75	70	26	31	71	68
13(1201-1300)	31	31	70	70	33	30	71	67
14(1301-1400)	38	29	73	69	23	28	67	66
15(1401-1500)	23	28	69	68	29	27	74	65
16(1501-1600)	29	27	68	67	29	26	61	64
17(1601-1700)	34	26	70	66	26	25	60	64
18(1701-1800)	27	25	64	65	27	24	69	63
19(1801-1900)	35	24	63	64	23	23	67	62
20(1901-2000)	26	23	68	63	22	22	61	61
21(2001-2100)	22	22	65	63	25	21	56	60
22(2101-2200)	21	21	54	62	23	20	58	59
23(2201-2300)	25	21	61	62	28	20	63	59
24(2301-2400)	22	20	56	61	14	19	49	58
25(2401-2500)	14	19	62	61	26	18	60	58

Tabelle 5. Beobachtete und theoretische Verteilungsfunktion $F_i(t)$ für das Russische (ONL)

(2) Rang	(1) Russisch (ONL)								
	(3) Urslav. 600		1500		1600		1800		1900
	(4)	η=0,49	η=0,20		η=0,16		η=0,07		(5)
	(5)	(6)	(5)	(6)	(5)	(6)	(5)	(6)	
1 (1-100)	66	61	83	82	87	85	93	93	99
2 (101-200)	34	50	63	76	70	76	88	90	99
3 (201-300)	37	43	67	70	75	76	91	89	99
4 (301-400)	33	38	56	68	66	73	86	87	100
5 (401-500)	19	33	54	67	64	80	77	85	95
6 (501-600)	28	30	52	61	63	68	78	84	93
7 (601-700)	24	27	49	59	59	66	81	83	95
8 (701-800)	24	25	52	57	61	64	81	82	94
9 (801-900)	26	23	53	55	60	62	84	81	99
10 (901-1000)	26	21	56	53	66	60	83	80	94
11(1001-1100)	24	20	49	52	60	59	80	79	96

Tabelle 5. (Fortsetzung)

(2) Rang	(1) Russisch (ONL)								
	(3) Urslav. 600		1500		1600		1800		1900
	(4) η=0,49		η=0,20		η=0,16		η=0,07		(5)
	(5)	(6)	(5)	(6)	(5)	(6)	(5)	(6)	
12(1101-1200)	19	18	47	50	57	58	73	79	94
13(1201-1300)	20	17	47	49	52	56	69	78	93
14(1301-1400)	22	16	51	47	60	55	82	77	94
15(1401-1500)	16	15	36	46	48	54	78	76	96
16(1501-1600)	23	14	49	45	61	53	78	76	91
17(1601-1700)	24	13	52	44	58	52	83	75	94
18(1701-1800)	21	12	52	43	62	51	79	74	98
19(1801-1900)	18	12	44	42	58	50	80	73	97
20(1901-2000)	13	11	45	41	49	49	71	73	98

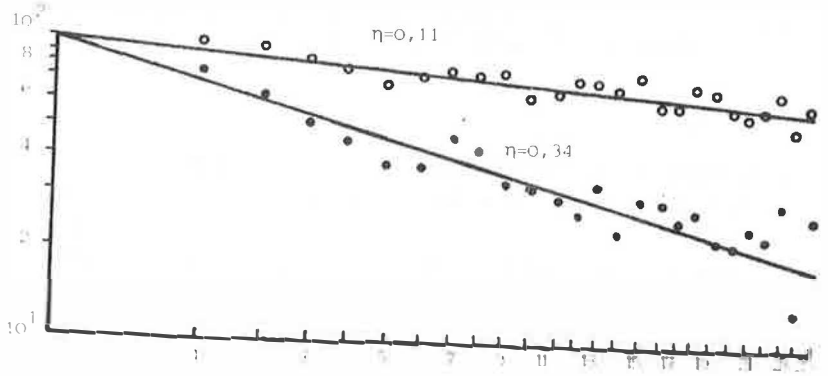

Abb. 8. Linearisierter Graph der Funktion $F_t(i)$ für das Tschechische. Oben: ab 1500, unten: ab 600. Logarithmische Skala in O_y, Quadratwurzelskala in O_x.

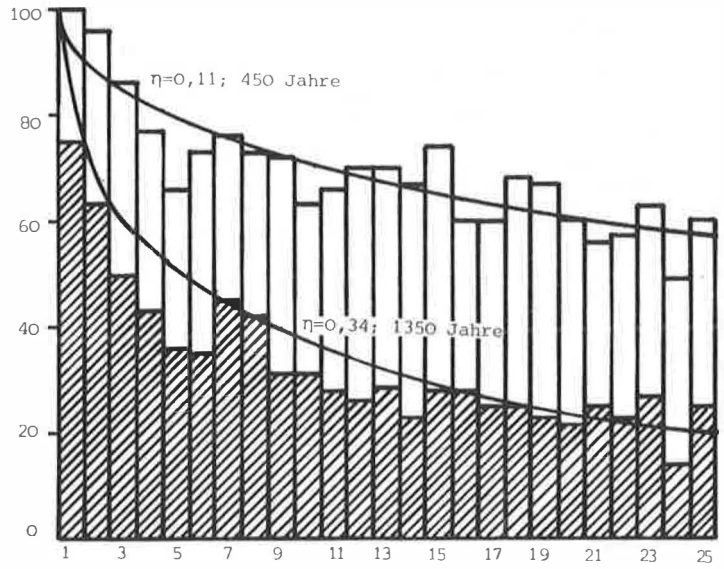

Abb. 9. Graph der Funktion $F_t(i)$ für das Tschechische. Oben: ab 1600, unten: „Urslavisch" (ca. 600).

3. Deutsch

A. B. Morgan: German frequency book. N.Y. 1928 = F. Kaeding: Häufigkeitswörterbuch der deutschen Sprache. Steiglitz bei Berlin, 1898
(Häufigkeitswörterbuch nach Kaedings Häufigkeitslexikon von Wortformen, Kaed) - 1000 Wörter.
W.M. Andrjuščenko (u.a.): Häufigkeitswörterbuch des politischen Verkehrs (Manuskript, And) - 1000 Wörter.

B. J. Grimm, W. Grimm: Deutsches Wörterbuch. Leipzig 1854.
Der Große Duden, Bd. 7. Etymologie. Frankfurt a.M., 1959.
F. Kluge: Etymologisches Wörterbuch der deutschen Sprache. Berlin, 1957.

C.

(1)*	(2)	(3)	(4)	(5)
And	1900	1	6	-
"	1800	2	6	-
"	1700	3	6	-
"	1600	4	6	-
"	1500 mhd.	5	6	10 u. 11
"	1100 ahd.	9	6	10
"	500 west-germ.	15	6	11
Kaed	1700	1,5	7	-
"	1600	2,5	7	12
"	1500 mhd.	3,5	7	13
"	1100 ahd.	7,5	7	12
"	500 west-germ.	13,5	7	13

*Für And ist $\bar{t}$ = 2000, für Kaed $\bar{t}$ = 1850 (Abb. 10-13).

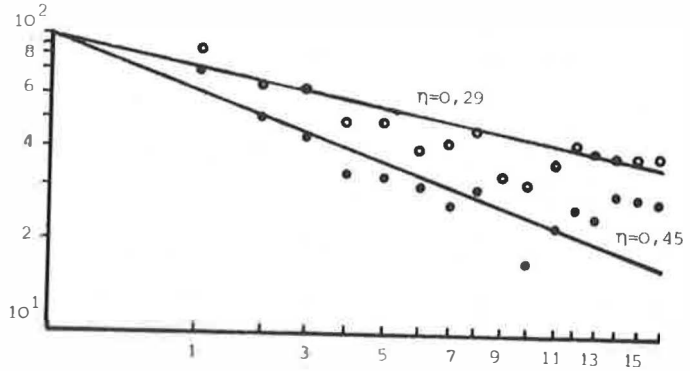

Abb. 10. Linearisierter Graph der Funktion $F_t(i)$ für das Deutsche (And). Oben: ab 1500 (mhd.), unten: ab 1100 (ahd.). Logarithmische Skala in O_x und Quadratwurzel-Skala in O_y.

Tabelle 6. Beobachtete und theoretische Verteilungsfunktion $F_i(t)$ für das Deutsche (And)

(1) Deutsch (And)														
(2) Rang	(3) westgermanisch 500		ahd. 1100		mhd. 1500		1600		1700		1800		1900	
	(4) η=0,60		η=0,45		η=0,29		η=0,22		η=0,19		η=0,15		η=0,11	
	(5)	(6)	(5)	(6)	(5)	(6)	(5)	(6)	(5)	(6)	(5)	(6)	(5)	(6)
1 (1-100)	63	55	74	64	84	75	87	80	87	83	88	86	94	90
2 (101-200)	34	43	52	53	65	66	70	73	74	76	78	81	79	85
3 (201-300)	33	35	43	46	64	61	71	68	77	72	86	77	86	83
4 (301-400)	25	30	33	41	48	56	64	64	68	68	74	74	78	80
5 (401-500)	20	26	33	36	49	52	60	61	64	65	74	71	78	78
6 (501-600)	16	23	30	33	40	49	49	58	52	62	63	69	67	76
7 (601-700)	12	20	27	30	42	46	52	56	60	61	67	67	72	75
8 (701-800)	19	18	30	28	46	44	54	54	61	58	69	66	73	73
9 (801-900)	11	16	23	26	33	42	45	52	50	57	60	63	67	72
10 (901-1000)	8	15	15	24	32	40	47	50	56	55	68	63	74	70
11 (1001-1100)	12	14	22	23	36	38	39	48	47	53	61	61	69	69
12 (1101-1200)	15	13	25	21	42	37	46	47	53	52	61	59	70	68
13 (1201-1300)	13	12	24	20	39	35	42	45	46	51	53	58	60	67
14 (1301-1400)	12	11	28	19	37	34	52	44	53	49	61	57	70	66
15 (1401-1500)	14	10	27	18	37	33	50	43	59	48	68	56	80	65
16 (1501-1600)	10	9	27	17	38	31	44	41	45	47	56	55	73	64

Tabelle 7. Beobachtete und theoretische Verteilungsfunktion $F_i(t)$ für das Deutsche (Kaed)

(1) Deutsch (Kaed)											
(2) Rang	(3) westgerm. 500		ahd. 1100		mhd. 1500		1600		1700		1800
	(4) η=0,32		η=0,18		η=0,08		η=0,04		η=0,02		
	(5)	(6)	(5)	(6)	(5)	(6)	(5)	(6)	(5)	(6)	(5)
1 (1-100)	82	73	94	83	97	92	100	96	100	98	100
2 (101-200)	72	64	88	78	99	90	100	95	100	97	100
3 (201-300)	63	58	83	73	97	87	99	93	99	97	99
4 (301-400)	54	53	73	70	92	85	95	92	99	96	100
5 (401-500)	45	49	63	67	87	84	93	91	97	96	99
6 (501-600)	38	46	55	64	84	82	93	90	97	95	100
7 (601-700)	38	43	55	62	76	81	84	90	95	95	100
8 (701-800)	38	41	50	60	78	79	88	90	93	94	100
9 (801-900)	35	38	52	58	72	79	81	89	92	94	98
10 (901-1000)	32	36	53	57	70	78	79	88	87	94	97

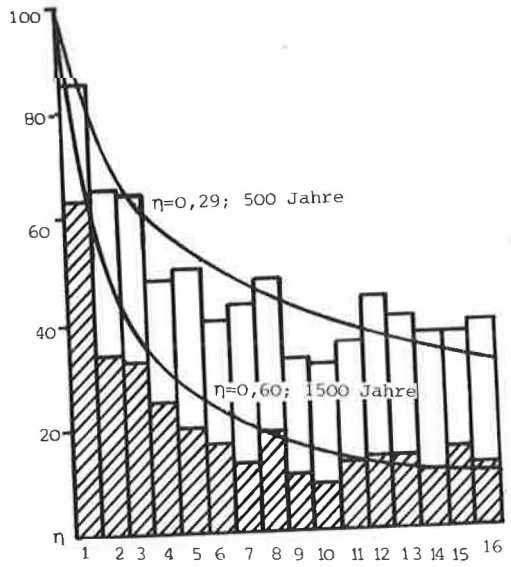

Abb. 11. Graph der Funktion $F_t(i)$ für das Deutsche. Oben: ab 1500 (mhd.), unten: ab 500 (westgerm.). nach And.

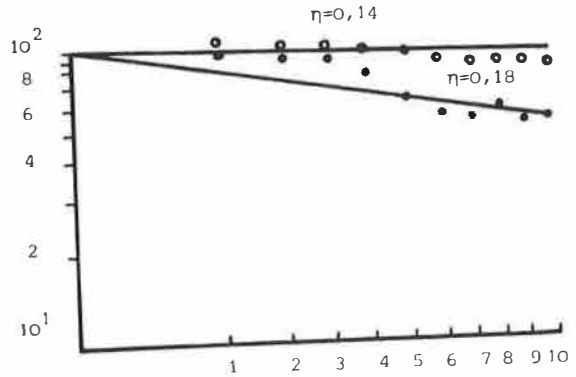

Abb. 12. Linearisierter Graph der Funktion $F_t(i)$ für das Deutsche (Kaed). Oben: ab 1600, unten: ab 1100. Logarithmische Skala in O_y und Quadratwurzel-Skala in O_x.

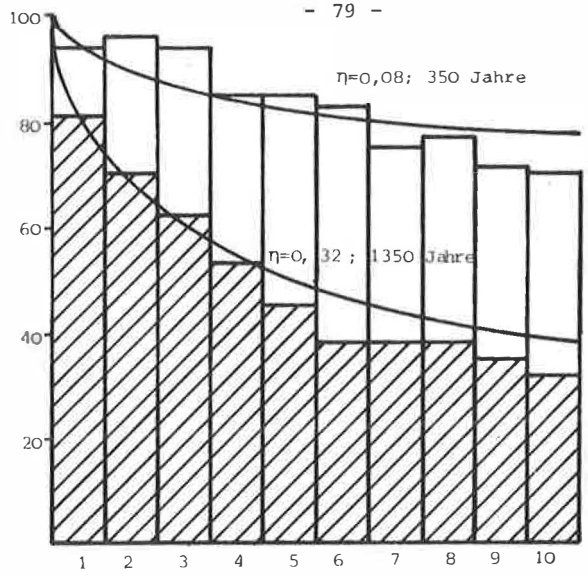

Abb. 13. Graph der Funktion für das Deutsche (Kaed). Oben: ab 1500 (mhd.), unten: ab 500 (westgerm.).

4. Englisch

A. A.G. Dewey: Relative frequency of English speech sound. Cambridge 1923 (Dew) - 1000 Wörter.
R.C. Eldridge: Six thousands common English words. Niagara Falls 1911 (Eld) - 600 Wörter (die Daten zur Verteilung der Wörter nach Alter sind Zipf [1] entnommen).

B. The shorter Oxford English dictionary, v. I-II. Oxford 1947.

C.

(1)	(2)	(3)*	(4)	(5)
Dew	1600	3	8	-
"	1500	4	8	-
"	1400 me.	5	8	14
"	1100 ae.	8	8	14
Eld	1400 me.	5	9	15a
"	1100 ae.	8	9	15a

*für beide Wörterbücher ist $\bar{t}$ = 1900

Tabelle 8. Beobachtete und theoretische Verteilungsfunktion $F_i(t)$ für das Englische (Dew)

(1) Englisch (Dew)									
(2) Rang	(3) altenglisch 1100		mittelenglisch 1400		1500		1600		1700
	(4) η=0,36		η=0,06		η=0,04		η=0,015		(5)
	(5)	(6)	(5)	(6)	(5)	(6)	(5)	(6)	
1 (1-100)	92	70	100	94	100	95	100	98	100
2 (101-200)	70	60	97	92	99	93	100	97	100
3 (201-300)	53	54	92	90	96	91	98	97	100
4 (301-400)	40	49	92	89	94	90	97	96	100
5 (401-500)	47	44	91	88	94	90	99	96	100
6 (501-600)	32	41	86	86	90	89	98	95	100
7 (601-700)	29	39	86	86	89	88	96	95	99
8 (701-800)	36	36	81	85	86	87	95	94	100
9 (801-900)	31	34	80	84	87	86	96	94	98
10 (901-1000)	31	34	77	83	81	85	92	94	98

Tabelle 9. Beobachtete und theoretische Verteilungsfunktion $F_i(t)$ für das Englische (Eld)

(1) Englisch (Eld)				
(2) Rang	(3) altenglisch 1100		mittelenglisch 1400	
	(4) η= 0,36		η= 0,08	
	(5)	(6)	(5)	(6)
1 (1-100)	88	70	90	92
2 (101-200)	66	60	98	90
3 (201-300)	53	54	85	87
4 (301-400)	41	49	87	85
5 (401-500)	38	44	84	84
6 (501-600)	36	41	82	82

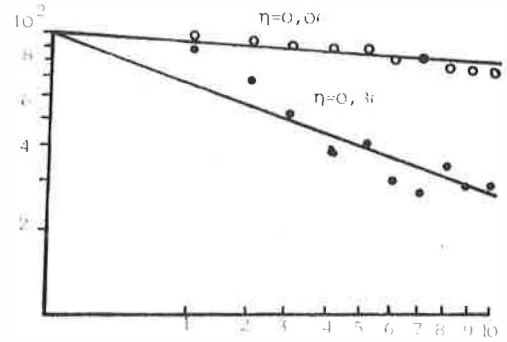

Abb. 14. Linearisierter Graph der Funktion $F_t(i)$ für das Englische. Oben: ab 1400, unten: ab 1100 (altengl.). O_y: logarithmische Skala, O_x: Quadratwurzel-Skala.

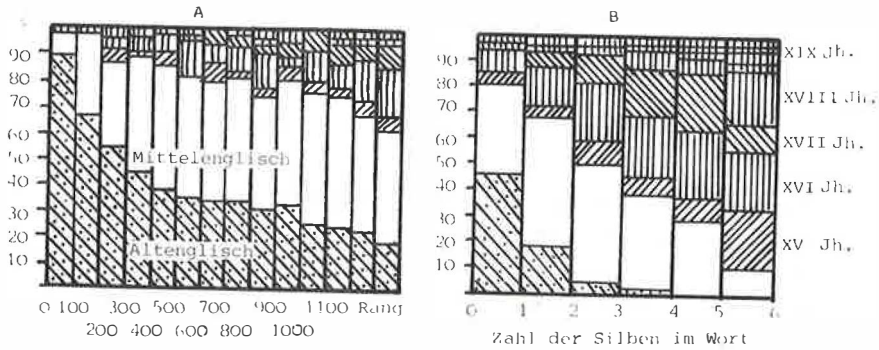

Abb. 15. Verteilung altenglischer Wörter (Schraffiert) und mittelenglischer Wörter (weiß) in Eld.
- a - in Abhängigkeit vom Rang; b - in Abhängigkeit von der Silbenzahl

5. Französisch

A. B. Vander Beke: French word book. N.Y. 1932 (VBe) insgesamt 6000 Wörter; hier angeführt sind die Ergebnisse für die ersten 2500 Wörter.
G. Gougenheim (et al.): L'élaboration du français élémentaire, Paris 1956 (Goug) - 1500 Wörter.

B. A. Dauzat: Dictionnaire étymologique de la langue française. Paris 1954
E. Littré: Dictionnaire de la langue française v. I-IV.

C.

(1)	(2)	(3)*	(4)	(5)
Goug	1700	2,5	10	-
"	1600	3,5	10	-
"	1500	4,5	10	-
"	1400	5,5	10	16
"	1300	6,5	10	-
"	1200	7,5	10	17
"	900 Vulgärlatein	12,5	10	17
"	400 Latein	15,5	10	16
VBe	1600	3,5	11	-
"	1500	3,5	11	-
"	1300	5,5	11	18
"	1150	7	11	19
"	700 Vulgärlatein	11,5	11	18
"	400 Latein	14,5	11	19

*Für Goug ist $\bar{t}$ = 1950, für VBe ist $\bar{t}$ = 1850. Die Wahl des Datums 400 n. Chr. beruht auf der Periodisierung des Lateinischen in Dauzat. Gemäß den dortigen Angaben haben wir nicht nur die Bildungen des klassischen Lateins, sondern auch die der Kaiserzeit zu den lateinischen Wörtern gerechnet. In der Wahl von 700 n. Chr. als Beginn der dialektalen Aufspaltung des Lateinischen schließen wir uns der bekannten Auffassung H. Schuchardts an.

Tabelle 10. Beobachtete und theoretische Verteilungsfunktion $F_i(t)$ für das Französische (Goug)

(1) Französisch (Goug)																							
(2) Rang	(3) Latein 400		Vulgärlatein 700		1200		1300		1400		1500		1600		1700	1800	1900						
	(4) η=0,49		η=0,33		η=0,175		η=0,145		η=0,11		η=0,085		η=0,05										
	(5)	(6)	(5)	(6)	(5)	(6)	(5)	(6)	(5)	(6)	(5)	(6)	(5)	(6)	(5)	(5)	(5)						
1 (1-100)	53	61	77	72	91	83	94	86	94	90	96	92	96	95	98	98	99						
2 (101-200)	54	50	68	62	84	78	87	81	90	85	91	90	93	93	97	99	100						
3 (201-300)	49	43	68	57	84	73	88	77	88	83	88	87	91	91	94	96	99						
4 (301-400)	48	38	58	52	71	70	76	74	82	80	86	85	89	90	97	98	99						
5 (401-500)	37	33	52	48	73	67	83	71	87	78	90	84	93	90	97	100	100						
6 (501-600)	27	30	37	44	52	64	60	69	70	76	76	82	91	89	98	99	99						
7 (601-700)	23	27	34	42	57	62	62	67	72	75	75	81	85	88	92	92	96						
8 (701-800)	20	25	31	39	55	60	62	66	70	73	76	79	87	87	95	97	99						
9 (801-900)	19	23	32	37	52	58	59	63	68	72	74	79	84	86	88	92	95						
10 (901-1000)	21	21	37	35	61	57	66	63	73	70	80	78	89	85	93	98	99						

Tabelle 11. Beobachtete und theoretische Verteilungsfunktion $F_i(t)$ für das Französische (VBe)

(2) Rang	(1) Französisch										
	(3) Latein 400		Vulgärlatein		1150		1300		1500		1600
	(4) η =0,44		η=0,32		η =0,13		η =0,09		η =0,04		(5)
	(5)	(6)	(5)	(6)	(5)	(6)	(5)	(6)	(5)	(6)	
1 (1-100)	70	64	83	73	97	88	99	91	99	96	100
2 (101-200)	54	54	85	64	91	84	94	88	98	95	99
3 (201-300)	54	47	75	58	86	80	95	85	97	93	100
4 (301-400)	46	41	57	53	81	77	87	84	95	92	97
5 (401-500)	29	37	46	49	78	75	84	82	97	91	99
6 (501-600)	36	34	47	46	79	73	85	80	98	90	99
7 (601-700)	32	31	44	43	78	71	82	79	95	90	99
8 (701-800)	18	29	36	41	68	69	75	78	92	90	98
9 (801-900)	29	27	37	38	60	68	74	76	96	89	99
10 (901-1000)	24	25	29	36	63	66	76	76	92	88	98
11(1001-1100)	18	23	31	35	55	65	65	75	86	88	93
12(1101-1200)	12	22	26	33	58	64	64	73	90	87	96
13(1201-1300)	18	21	28	32	58	63	68	72	81	87	93
14(1301-1400)	21	19	32	30	54	62	73	71	91	86	96
15(1401-1500)	19	18	31	29	55	61	65	70	91	86	97
16(1501-1600)	12	17	19	28	56	59	72	70	94	85	99
17(1601-1700)	15	16	27	27	57	58	74	69	87	85	94
18(1701-1800)	17	15	20	26	47	58	59	68	84	84	93
19(1801-1900)	19	15	31	25	59	57	68	68	85	84	94
20(1901-2000)	12	14	22	24	59	56	75	67	87	84	97
21(2001-2100)	14	13	20	23	58	55	70	66	84	84	94
22(2101-2200)	17	13	28	22	53	54	65	66	83	83	96
23(2201-2300)	12	12	19	22	48	54	59	65	85	82	94
24(2301-2400)	12	12	18	21	46	53	57	64	80	82	90
25(2401-2500)	14	11	24	20	52	53	68	64	81	82	91

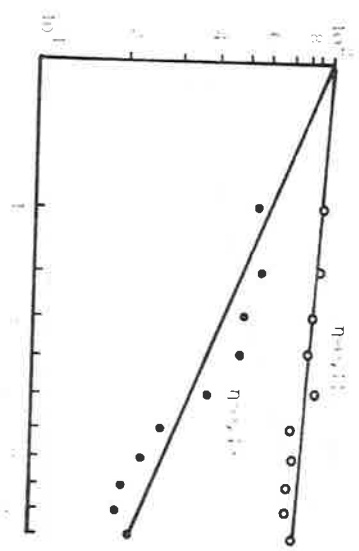

Abb. 16. Linearisierter Graph der Funktion $F_t(i)$ für das Französische (Goug). Oben: ab 1400, unten: ab 400. O_y: logarithmische Skala, O_x: Quadratwurzel-Skala.

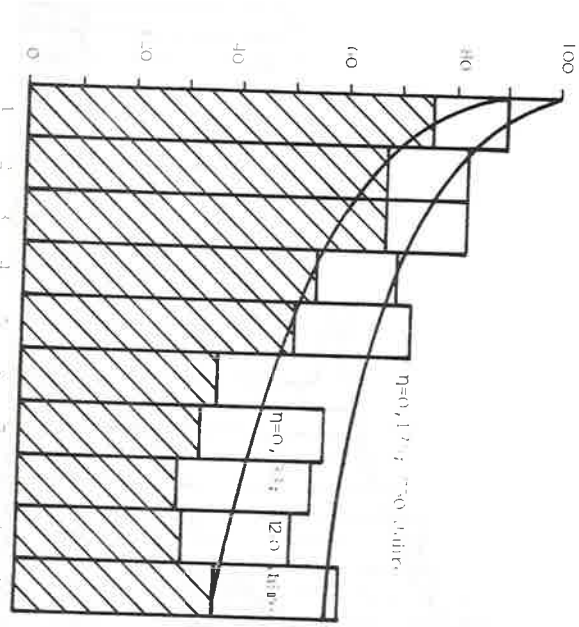

Abb. 17. Graph der Funktion $F_t(i)$ für das Französische (Goug). Oben: ab 1200, unten: ab 700 (Vulgärlatein).

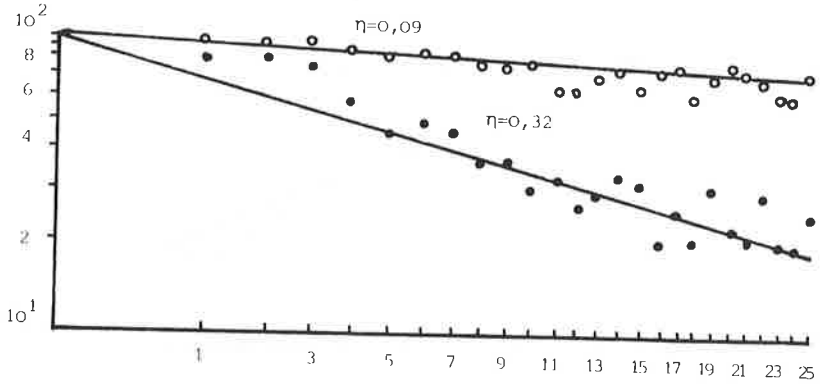

Abb. 18. Linearisierter Graph der Funktion $F_t(i)$ für das Französische (VBe). Oben: ab 1300, unten: ab 700 (Vulgärlatein). O_y: Logarithmische-Skala, O_x: Quadratwurzel-Skala.

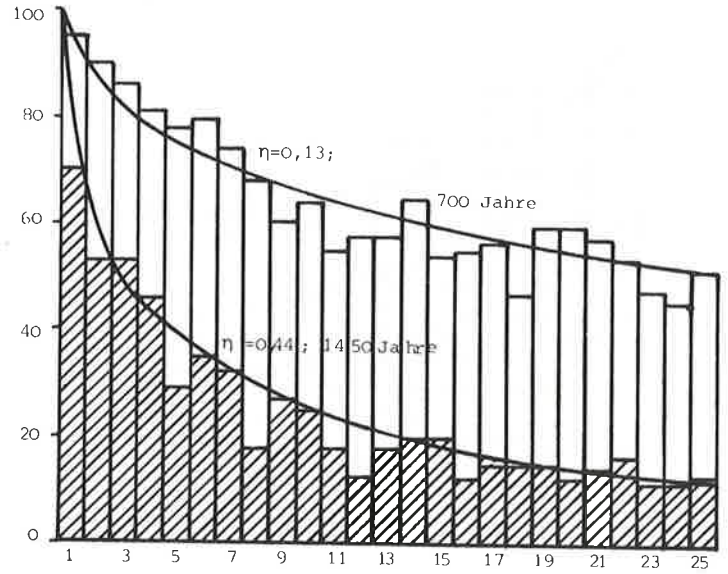

Abb. 19. Graph der Funktion $F_t(i)$ für das Französische (VBe). Oben: ab 1150, unten: ab 400 (Latein).

6. Rumänisch

A-B. Häufigkeitslisten für das Rumänische des XVI. Jh. (1100 Wörter), des XVII. Jh. (900 Wörter) und des XVIII. Jh. (1000 Wörter) mit etymologischen Angaben von C. Tudose [27] (Rum XVI, Rum XVII, Rum XVIII).

C.

(1)	(2)	(3)*	(4)
Rum XVI	700 Vulgärlatein	8;5	12
Rum XVII	700	9,5	"
Rum XVIII	700	10,5	"

*$\bar{t}$ ist entsprechend 1550, 1650 und 1750.

Tabelle 12. Beobachtete und theoretische Verteilungsfunktion $F_i(t)$ für das Rumänische

(2) Rang	Rum XVI Vulgärlatein 700		Rum XVII Vulgärlatein 700		Rum XVIII Vulgärlatein 700	
	η= 0,23		η = 0,24		η= 0,25	
1 (1-100)	90	79	90	79	86	78
2 (101-200)	80	73	77	71	67	70
3 (201-300)	50	67	47	66	65	65
4 (301-400)	53	63	45	62	55	61
5 (401-500)	48	59	44	58	56	57
6 (501-600)	56	57	57	55	51	54
7 (601-700)	46	54	41	53	54	52
8 (701-800)	49	52	48	51	54	49
9 (801-900)	62	50	58	49	50	47
10 (901-1000)	57	48	64	47	49	45
11(1001-1100)	52	47	66	45		

3.3. DER ZUSAMMENHANG VON LEXIKALISCHER LEBENSERWARTUNG UND ZEIT

Die Grundannahme der Glottochronologie, die besagt, daß die Veränderungsgeschwindigkeit der Kernlexik nicht von der Zeit abhängt (konstant ist), ist niemals allgemein akzeptiert worden. Die linguistische Erfahrung lehrt, daß einige Epochen der Sprachentwicklung durch stürmische Veränderungen gekennzeichnet sind, andere dagegen durch relative Stagnation. Diese Erfahrung beruht aber eher auf der Erforschung der historischen Phonetik und Morphologie und ist dann auf die schwer überschaubare Lexik übertragen worden. Eine solche Übertragung ist nicht immer gerechtfertigt: während die Entwicklung auf der phonologischen Ebene mit ihrer hochgradig symmetrischen Struktur sprunghaft-„revolutionär" verläuft, kann die Evolution auf der lexikalischen Ebene mehr oder weniger gleichmäßig sein.

Auch wenn die glottochronologische Grundannahme nicht wörtlich zutrifft, d.h. der lexikalische Zerfall von der Zeit abhängt, darf doch angenommen werden, daß seine Geschwindigkeitsschwankungen sich innerhalb der Grenzen halten, die durch die folgenden Gedanken von M. Swadesh gesetzt sind: „Auch wenn die Sprache verschiedenartigen, sie verändernden Einwirkungen unterworfen ist, muß sie doch gleichzeitig ein beträchtliches Maß an Homogenität bewahren. Alle Mitglieder der Sprachgemeinschaft müssen sie verstehen; dafür ist eine gewisse interindividuelle Übereinstimmung in Einzelheiten erforderlich. Zwischen den Generationen kann man häufig Unterschiede der Lexik und des Wortgebrauchs feststellen, doch erreichen diese niemals ein solches Ausmaß, daß sie ein Verständigungshindernis bilden könnten. Dies schränkt die Geschwindigkeit für den Sprachwandel maximal ein" [18].

Wenn wir aber nicht die gesamte Sprache, sondern genaue Daten zu einzelnen Stilen betrachten, ist das Bild ein völlig anderes: bei der Entwicklung eines einzelnen „peripheren" Stils können über viele hundert Jahre Beschleunigungs- oder Verzögerungstendenzen wirksam sein. Für die Annahme, daß die Sprache insgesamt sich eher gleichmäßig wandelt, spricht zwar, daß in einer Sprache nebeneinander Stile mit beschleunigender und solche mit verzögernder Wirkung existieren können.

Wenden wir uns nun den empirischen Fakten zu. Wir bemerken zunächst eine Erscheinung, die in allen Wörterbüchern anzutreffen ist, welche in bestimmtem Maße die Sprache der Belletristik oder verwandter Gattungen repräsentieren: die Geschwindigkeit des lexikalischen Wandels nimmt während der letzten 200 bis 300 Jahre merklich ab. Die französischen Wörterbücher Goug und VBe deuten auf eine 1,5 bis 2-fache Abnahme dieser Geschwindigkeit in der Zeit vom XVII. bis zum XX. Jh. im Vergleich zum XV. bis XVII. Jh. hin. Wahrscheinlich würde sich ähnliches (unter Umständen schwächer) auch beim Tschechischen und Russischen zeigen, jedoch reichen die lexikographischen Daten nicht für eine genaue Chronologie aus. Es ist schwer zu sagen, inwieweit diese Daten sich durch eine archaisierende Tendenz erklären lassen, die in bestimmtem Maß eine Eigenschaft aller europäischen Literaturen ist, und inwieweit sie durch die - an den Werken der klassischen Literatur des vergangenen Jh. orientierte - Auswahl der Quellen für die Häufigkeitswörterbücher bedingt ist. Möglich ist auch eine Verzerrung durch die lexikographischen Quellen, die sehr alte Wörter mit einer gewissen Verspätung fixieren, die jüngere Lexik dagegen gewissermaßen verfrüht, indem Wörter aufgeführt werden, die in dieser Zeit nur für einen bestimmten Autor typische Gelegenheitsbildungen waren.

Welche Kombination von Faktoren für diesen Effekt auch verantwortlich sein mag, seine Zurückführung auf eine plötzliche Geschwindigkeitsänderung des lexikalischen Eandels in jüngster Zeit ist zweifelhaft. Wenn man von dieser Periode absieht, so scheint in einigen Fällen der lexikalische Wandel wirklich mit nahezu konstanter Geschwindigkeit zu erfolgen. Diese Geschwindigkeiten sind in Tab. 13 für verschiedene Wörterbücher aufgeführt.

Genaue Daten zum Zusammenhang des Parameters n_t mit der Zeit werden für das Russische und Tschechische in Tab. 14 (s. auch Abb. 20) und für das Deutsche (nach Kaed) in Tab. 15 (s. auch Abb. 21) angeführt.

Tabelle 13. Mittlere Geschwindigkeiten des lexikalischen Wandels für verschiedene Sprachen

Wörterbuch	Mittlere Geschwindigkeit pro Jh.	Berechnungszeitraum
Rum XVII	0,0230	1650-700
Štf	0,0245	1500-600
Kaed	0,0250	1600-500
Čech	0,0255	1500-600
Rum XVI	0,0255	1550-700
Rum XVIII	0,0270	1750-700

Tabelle 14. Mittlere Geschwindigkeit des lexikalischen Wandels im Russischen und Tschechischen

Wörterbuch	Mittlere Geschwindigkeit	Berechnungszeitraum	Wörterbuch	Mittlere Geschwindigkeit	Berechnungszeitraum
Štf	0,0220	1950-1500	ONL	0,0465	1950-1800
"	0,0245	1500- 600	"	0,0450	1800-1600
Čech	0,0244	1950-1500	"	0,0400	1600-1500
"	0,0255	1500- 600	"	0,0320	1500- 600

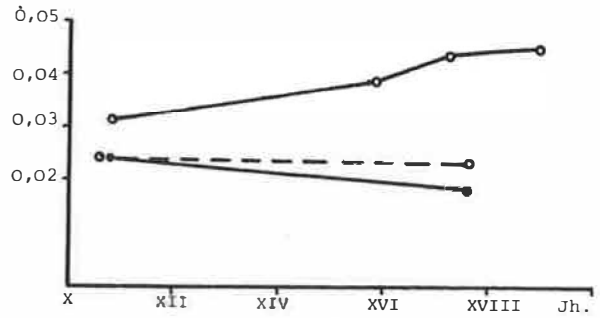

Abb. 20. Veränderung der Geschwindigkeit des lexikalischen Zerfalls, O oben ONL, O unten Štf, ● Čech

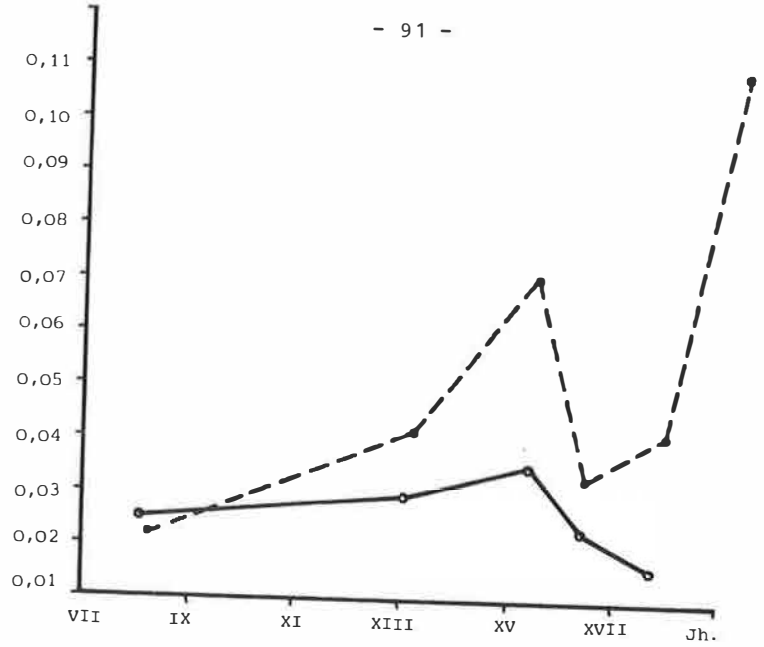

Abb. 21. Veränderung der Geschwindigkeit des lexikalischen Zerfalls ● And, O Kaed

Dabei ist anzumerken, daß die beiden Wörterbücher Čech und Kaed auf großen Stichproben basieren und ein breites stilistisches Spektrum abdecken.

Ein merklich höheres, vor allem aber klar von der Zeit abhängiges Tempo des Wandels läßt sich bei den beiden französischen und englischen Wörterbüchern feststellen (s. Tab. 15 und 16 und Abb. 22 und 23).

Wir verfügen über ausreichende Daten, um versuchen zu können, die Abhängigkeit von $\eta = \eta(t)$ von der Zeit zu bestimmen. In Abb. 24 werden die empirisch gefundenen Werte graphisch der Funktion $\eta(t) = \frac{c}{\sqrt{t}}$ mit $c = 0,1$ (Zeit in Jh.) gegenübergestellt. Die gute Anpassung war für VBe zu erwarten, weil dieses, obwohl in den 20er Jahren unseres Jahrhunderts erstellt, konsequent an der klassischen französischen Literatur orientiert ist (dem wurde teilweise durch die Wahl von $\bar{t} = 1850$ Rechnung getragen), für Goug überrascht sie dagegen etwas, weil dieses auf Tonbandaufzeichnungen französischer Umgangssprache basiert. Vielleicht läßt sich der literarische Cha-

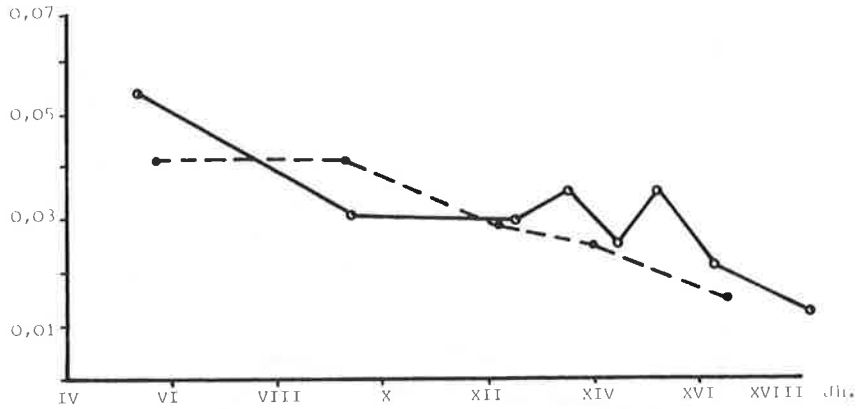

Abb. 22. Veränderung der mittleren Zerfallsgeschwindigkeit der französischen Wörterbücher

● Goug O VBe

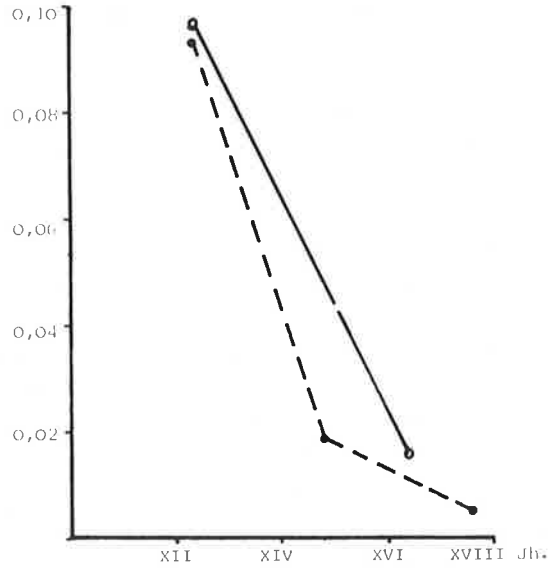

Abb. 23. Veränderung der mittleren Zerfallsgeschwindigkeit der englischen Wörterbücher

● Dew O Eld

Tabelle 15. Mittlere Geschwindigkeit des lexikalischen Wandels im Deutschen

Wörterbuch	mittlere Geschwindigkeit	Berechnungszeitraum
Kaed	0,0150	1850-1700
"	0,0230	1700-1600
"	0,0350	1600-1500
"	0,0250	1500-1100
"	0,0230	1100- 500

Tabelle 16. Mittlere Geschwindigkeit des lexikalischen Wandels im Französischen

Wörterbuch	Mittlere Geschwindigkeit	Berechnungszeitraum	Wörterbuch	Mittlere Geschwindigkeit	Berechnungszeitraum
Goug	0,0125	1950-1700	Goug	0,0535	700- 400
"	0,0200	1700-1600	VBe	0,0140	1850-1500
"	0,0350	1600-1500	"	0,0250	1500-1300
"	0,0250	1500-1400	"	0,0300	1300-1150
"	0,0350	1400-1300	"	0,0410	1150- 700
"	0,0300	1300-1200	"	0,0400	700- 400
"	0,0310	1200- 700			

Tabelle 17. Mittlere Geschwindigkeit des lexikalischen Wandels im Englischen

Wörterbuch	mittlere Geschwindigkeit	Berechnungszeitraum
Dew	0,0050	1900-1600
"	0,0200	1600-1500
"	0,0200	1500-1400
Eld	0,1000	1400-1100
"	0,0160	1900-1400
	0,9350	1400-1100

rakter durch die besondere thematische Auswahl bei den aufgezeichneten Gesprächen (hauptsächlich neutrale, „zeitlose" Themen) erklären.

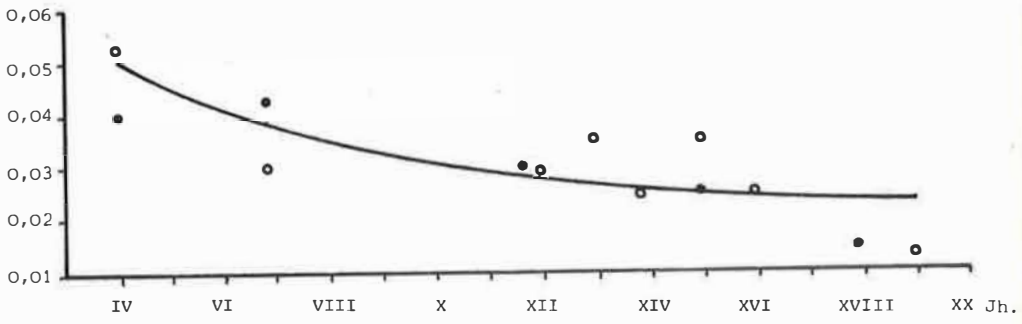

Abb. 24. Die Abhängigkeit des Parameters η von der Zeit t für die Wörterbücher. O Goug; ● VBe. Die durchgehende Linie ist der Graph der Funktion $\eta = \frac{0,1}{\sqrt{t}}$ (Zeit in Jh.).

Die englischen Daten zeigen eine plötzliche Beschleunigung des lexikalischen Wandels im Neuenglischen gegenüber dem Mittelenglischen. Diese Auffälligkeit ist vielleicht einfach durch die Mangelhaftigkeit der benutzten Häufigkeitswörterbücher erklärbar, die gleichzeitig entstandene Homonyme nicht trennen, was in einer Sprache wie der englischen notwendig zu wesentlichen Verzerrungen führt. Vielleicht sind die Verhältnisse aber auch komplizierter, und es handelt sich um indirekte Resultate des starken romanischen Einflusses, dem die englische Sprache im Mittelalter ausgesetzt war.

Betrachten wir abschließend die Daten für zwei Wörterbücher - ONL (Tab. 5 und Abb. 20) und And (Tab. 18 und Abb. 21), die gewissermaßen periphere Sprache repräsentieren.

Für And, das die schriftliche Verwendung politischer Lexik in der DDR repräsentiert, haben wir $\bar{t} = 2000$ angesetzt, weil dieser Sprachgebrauch nicht für die ganze deutsche Sprache und nicht einmal für diesen Stil im gesamten Verbreitungsgebiet der deutschen Sprache charakteristisch ist. Hier läßt sich, wie auch bei der russischen allgemeinwissenschaftlichen Lexik, eine Tendenz zur linearen Beschleunigung des lexikalischen Wandels bemerken.

Tabelle 18. Mittlere Geschwindigkeit des lexikalischen Wandels im Deutschen (nach And)

Wörterbuch	Mittlere Geschwindigkeit	Berechnungszeitraum
And	0,110	2000-1900
"	0,040	1900-1800
"	0,040	1800-1700
"	0,030	1700-1600
"	0,070	1600-1500
"	0,040	1500-1100
"	0,025	1100- 500

Ergänzend führen wir die Ergebnisse der statistischen Analyse der Daten von W. Patterson zur Geschichte des Spanischen [37] an, jedoch in anderer Form als oben.

Wir finden bei Patterson einmal die Verteilung der Wörter vulgärlateinischen Ursprungs in den ersten zehn Klassen zu 500 Wörtern des Häufigkeitswörterbuchs von Juilland [38], zum anderen die Anzahl der Neubildungen für jedes Jahrhundert (echte Entlehnungen und autochthone Neubildungen) vom X. Jh. an, d.h. die Anzahl der Wörter unter den ersten 5000 des spanischen Häufigkeitswörterbuchs, die 1050, 950, 850 etc. Jahre alt sind.

Um zu Ergebnissen zu kommen, die im Großen und Ganzen hinsichtlich der mittleren Geschwindigkeit des lexikalischen Wandels mit den übrigen sechs untersuchten Sprachen vergleichbar sind, haben wir η_t nach Formel (4.8) berechnet, die in 4.2 abgeleitet wird:

$$I_m = n \sum_{i=1}^{m} e^{\eta_t \sqrt{i}} \approx \frac{2n}{\eta_t^2}[e^{-\eta_t}(\eta_t+1) - e^{-\eta_t\sqrt{m}}(\eta_t\sqrt{m}+1)].$$

Diese Formel wird zur Berechnung der theoretischen Erwartung der Anzahl von t und mehr Jahre alten Wörtern der m ersten Klassen des Häufigkeitswörterbuchs zusammen (und nicht für jede einzelne Klasse wie (3.3)) bei gegebenem η_t verwendet. Der Parameter η_t wurde nach der Methode der kleinsten Fehlerquadrate aus den Daten von Patterson für die Funktion I_m in den Punkten m = 5,10,...,50 abgeschätzt (die Einheit 100 wurde wegen der Vergleichbarkeit der η_t beibehalten).

Interessant ist die Ähnlichkeit der Werte von $\eta_{vulg.lat.}$ für Spanisch und Französisch (0,32 bzw. 0,33), und auch die Abnahme der mittleren Geschwindigkeit des lexikalischen Wandels mit der Zeit, die auch bei allen anderen untersuchten Literatursprachen auftritt.

Die Daten zur Verteilung der Lexik vulgärlateinischer Herkunft finden sich in Tab. 18a. Zur Bestimmung von η_t für das X. bis XX. Jh. haben wir Gleichung (4.8) mit m = 50 nach η_t aufgelöst (s. Tab. 18b). Die so gewonnenen Resultate sind weniger zuverlässig als die der übrigen sechs Sprachen.

Tabelle 18a. Die vulgärlateinische Schicht im Spanischen

Rang	$F_i(t)$	Beobachteter Wert zu I_m	Theoretischer Wert von I_m bei η = 0,33
1 (1-500)	318	318	217
2 (501-1000)	160	478	448
3(1001-1500)	124	602	543
4(1501-2000)	117	719	712
5(2001-2500)	96	815	833
6(2501-3000)	108	923	932
7(3001-3500)	78	1001	1000
8(3501-4000)	65	1066	1077
9(4001-4500)	58	1124	1110
10(4501-5000)	52	1176	1147

Tabelle 18b. Die Dynamik des lexikalischen Wandels im Spanischen vom X. bis zum XX. Jh.

Jh.	Anzahl der im angegebenen Jh. entstandenen Wörter unter den ersten 5000 Wörtern des Häufigkeitswörterbuchs	Beobachteter Wert zu I_{50}	Näherungswert von η_t	Für das angegebene Jh. gemitteltes η_t
X	92	1268	0,300	0,030*
XI	38	1306	0,290	0,010
XII	220	1526	0,260	0,030
XIII	783	2309	0,160	0,100
XIV	221	2530	0,140	0,020

Tabelle 18b. (Fortsetzung)

Jh.	Anzahl der im angegebenen Jh. entstandenen Wörter unter den ersten 5000 Wörtern des Häufigkeitswörterbuchs	Beobachteter Wert zu I_{50}	Näherungswert von η_t	Für das angegebene Jh. gemitteltes η_t
XV	1164	3694	0,065	0,075
XVI	475	4169	0,055	0,010
XVII	403	4572	0,030	0,015
XIX-XX	188	5000	0,000	> 0,010

*Für das Intervall 700 (Vulg. Lat.) bis 1000

3.4. ZWEI VERWANDTE SPRACHEN MIT GEMEINSAMEM VORGÄNGER

In diesem Paragraphen behandeln wir unser bescheidenes Material zu verwandten Sprachen, die nicht in Vorgänger-Nachfolger-Relation zueinander stehen. Wir haben für das Wörterbuch Štf (2500 Wörter) die Verteilung der Wörter aufgestellt, zu denen es Entsprechungen in der tschechischen Literatursprache der Gegenwart gibt. Genauso wurde für Čech (für die ersten 2500 Wörter) die Verteilung der Wörter mit Entsprechungen in der russischen Literatursprache der Gegenwart aufgestellt. Dabei wurden Entsprechungen auch gezählt, wenn sie nicht in den beiden Häufigkeitswörterbüchern selbst, sondern in Definitionswörterbüchern angegeben waren. Das Entsprechungskriterium war dasselbe wie in 3.2, so daß man von der Vergleichbarkeit der Ergebnisse mit der oben beschriebenen Verteilung der urslavischen Lexik in Štf und Čech ausgehen kann (s. Tab. 4, Abb. 4 und 8).

Es wurden für diese beiden Verteilungen die folgenden Hypothesen überprüft:

1) Beiden Verteilungen läßt sich eine negative Exponentialfunktion der Form (3.3) anpassen;
2) die Parameter η_t beider Verteilungen sind gleich (d.h. die weichen nicht mehr als statistisch zulässig voneinander ab);
3) nimmt man das VI. Jh. n. Chr. als Ende der urslavischen Periode und η_t in (3.3) für 600 n. Chr. mit 0,32 bis 0,34 an (s. Tab. 4), so sollten die tschechischen Wörter im russischen Lexikon und entsprechend die russischen Wörter im tschechischen Lexikon in Abhängigkeit vom Rang nach (2.18) verteilt sein, d.h. der Koeffizient $\sqrt{i}$ sollte $0,64 \leq 2\eta_t \leq 0,68$ sein ($0,32 \cdot 2 = 0,64$).

In Tab. 19 ist für jede Klasse (à 100 Wörter) des Häufigkeitswörterbuchs (insgesamt je 25 Klassen für Russisch und Tschechisch) zuerst die Anzahl russischer Wörter mit tschechischen Entsprechungen („tschech. im Russ."), dann die Anzahl tschechischer Wörter mit russischen Entsprechungen („russ. im Tschech.") angegeben. In der nachstehenden Abbildung gibt der Graph die nach (2.18) berechnete Anzahl gemeinsamer Wörter an; der Koeffizient im Exponenten wurde mit der Methode der kleinsten Fehlerquadrate bestimmt.

Die Zufälligkeit der Abweichungen der beiden empirischen Verteilungen von der theoretischen wurde mit dem χ^2-Test überprüft. Wir testen, ob sich die Parameter der empirischen Verteilung von denen der theoretischen Verteilung unterscheiden, d.h. ob die Hypothese H_o: $\eta_{theor.} = \eta_{emp.}$ angenommen werden kann (das Signifikanzniveau setzen wir auf 0,1 fest). Da $\chi^2_{0,1} = 32$ bei $25 - 2 = 23$ Freiheitsgraden, der empirische χ^2-Wert für „tschech. in Russ." dagegen 22,14 und für „russ. in Tschech." sogar nur 11,25 beträgt, gibt es keinen Grund für die Ablehnung der Nullhypothese, d.h. für die Annahme, daß die beiden Größen, die Zahl der „russischen" Wörter im Tschechischen und die der „tschechischen" im Russischen, nicht die gleiche Verteilung mit gleichem Parameter hätten.

Die gute Übereinstimmung der beiden Verteilungen miteinander und mit der theoretischen Verteilung ist ja schon graphisch in Abb. 25 ziemlich gut zu erkennen, wo auf der Abszisse die Quadratwurzel

des Rangs und auf der Ordinate der Logarithmus der Anzahl gemeinsamer russisch-tschechischer Wörter aufgetragen ist. Bei dieser Skalenwahl hat die theoretische Verteilung eine Gerade als Graph (s. Abb. 25 u. Tab. 19).

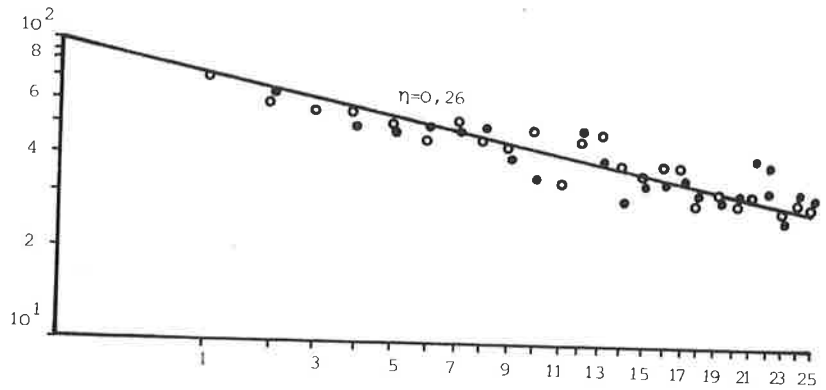

Abb. 25. Linearisierter Graph der Verteilung der „russischen" Wörter im Tschechischen (●) und der „tschechischen" im Russischen(O).

Wie wir sehen, werden die Hypothesen 1 und 2 empirisch voll bestätigt, nicht jedoch Hypothese 3. Für η_t ergab sich ja der Wert 0,26. Wenn man eine mittlere Geschwindigkeit von 0,022 annähme - was der niedrigste Wert für die slavischen Sprachen ist, s. Tab. 14 - hätten die beiden Sprachen trotzdem eine Periode selbständiger Entwicklung von 6 Jahrhunderten[3]. Wenn man noch alle gemeinsamen Lehnwörter aus Drittsprachen abzieht (die Anzahl russischer Lehnwörter im Tschechischen ist bei den ersten zwei- bis dreitausend Wörtern äußerst gering, die der tschechischen im Russischen praktisch Null), erhöht sie sich auf vielleicht 7 Jh.. Aber auch

dann weicht das Ergebnis von dem zu erwartenden stark ab; denn den vollständigen Zerfall des Urslavischen und damit den Beginn der Geschichte der Einzelsprachen Russisch und Tschechisch setzt die Mehrzahl der Sprachwissenschaftler mit (spätestens) dem Ende des VI. Jh. (d.h. vor 13,5 Jh.). Mitte des XIII. Jh. kann weder eine ethnische noch eine geographische Einheit der Slaven bestanden haben, die ja gerade für das VI. Jahrhundert (mit gewissen Einschränkungen) angenommen wird.

Zwar vertreten einige bekannte Sprachwissenschaftler die Ansicht, daß die slavischen Sprachen ihre Einheit in gewissem Maße noch lange nach ihrer geographischen Trennung bewahrt haben können.

Tabelle 19. Verteilung der russisch-tschechischen Entsprechungen nach Rängen

Rang	tschech. im Russ.	russ. im Tschech.	theoretischer Wert bei $2n_t = 0{,}26$
1 (1-100)	72	73	77
2 (101-200)	62	60	69
3 (201-300)	57	57	64
4 (301-400)	57	50	59
5 (401-500)	48	50	56
6 (501-600)	50	46	53
7 (601-700)	50	54	50
8 (701-800)	48	49	48
9 (801-900)	39	45	46
10 (901-1000)	36	49	44
11 (1001-1100)	35	35	42
12 (1101-1200)	49	45	41
13 (1201-1300)	41	48	39
14 (1301-1400)	30	40	38
15 (1401-1500)	34	35	37
16 (1501-1600)	33	39	35
17 (1601-1700)	36	39	34
18 (1701-1800)	31	29	33
19 (1801-1900)	30	32	32
20 (1901-2000)	31	30	31
21 (2001-2100)	43	32	30
22 (2101-2200)	39	35	30
23 (2201-2300)	25	27	29
24 (2301-2400)	33	31	28
25 (2401-2500)	31	29	27

So hat A. Meillet parallele Entwicklungen in verschiedenen Teilen des slavischen Gebiets bis zum IX. Jh. angenommen; N. Trubetzkoj und N. Durnovo [5] verbinden überhaupt den Zerfall der urslavischen Einheit und den Beginn der slavischen Einzelsprachen mit einer bedeutenden phonetischen Entwicklung, die als Ausfall der reduzierten Vokale bekannt ist. Im Russischen wie auch im Tschechischen war sie Ende des XII. Jh., spätestens Anfang des XIII. Jh., abgeschlossen. Dieses Datum kommt dem oben errechneten schon wesentlich näher. Wie aber sind wir bei der Berechnung dieses Datums eigentlich verfahren? Wir haben den Zeitpunkt gesucht, von dem ab das russische und tschechische Lexikon sich völlig unabhängig entwickelt haben, d.h. den Zeitpunkt, an dem der Anteil der gemeinsamen Wörter in beiden Lexika gemäß (2.18) abzunehmen begann. Bis dahin waren beide Lexika völlig übereinstimmend, gemäß (2.6), zerfallen, d.h. Wörter schwanden immer aus beiden Lexika.

Man kann jedoch annehmen, daß zwischen der absoluten Unabhängigkeit der Entwicklung (Koeffizient 2 im Exponenten) und der absoluten Übereinstimmung (Koeffizient 1) ein Zwischenstadium, von vielleicht nicht einmal 100 Jahren, lag. Signale der Veränderung eines Lexikons gelangten zum anderen, aber verzögert und verzerrt. In diesen Termini läßt sich nun die Unterscheidung Sprache vs. Dialekt teilweise fixieren: eine Zeitlang waren das Russische und das Tschechische nur Dialekte der slavischen Sprache (die möglicherweise zu einer größeren dialektalen Gemeinschaft gehörten), sie wurden dann später selbständige Sprachen. Der Unterschied zwischen diesen zwei Zuständen liegt nicht so sehr in der Zahl der Unterschiede wie in der Geschwindigkeit der Vermehrung dieser Unterschiede. Zwei periphere Dialekte können beträchtlich voneinander abweichen, solange aber jeder Veränderung des einen eine (unter Umständen spätere) Veränderung des anderen entspricht, kann man von der Einheit der vorliegenden Sprache ausgehen. Sobald eine unabhängige Entwicklung einsetzt, muß man auch ansonsten sehr ähnliche Dialekte als eigenständige Sprachen werten.

3.5. ABSOLUTE DATIERUNG

Die Untersuchung der russisch-tschechischen Entsprechungen kann auch als Beispiel für die Anwendungsmöglichkeiten unserer Methode auf die absolute Datierung sprachgeschichtlicher Ereignisse angesehen werden.

Für eine solche Datierung gibt es zwei Vorbedingungen: 1. Das Ereignis muß als Verteilung in einem konkreten Häufigkeitswörterbuch repräsentiert sein, und zwar annähernd mit der Verteilungsfunktion (3.3) oder (2.18); 2. Für ein solches Häufigkeitswörterbuch müssen wir die Abhängigkeit der Parameter der Verteilung von der Zeit kennen.

Der Fortschritt gegenüber der glottochronologischen Methode besteht darin, daß die Geschwindigkeit des lexikalischen Wandels nicht unbedingt als konstant und für alle Sprachen gleich angenommen werden muß. An die Stelle der Konstanzannahme tritt die Berechnung des Parameters aus der funktionalen Abhängigkeit von η_t von der Zeit (diese Funktion heißt bei der Radiocarbonmethode „Eichkurve").

Wenn wir also wissen, welche Wörter vor einem bestimmten prähistorischen Zeitpunkt (z.B. vor dem Zerfall der Protosprache) entstanden sind, können wir diesen Zeitpunkt berechnen, indem wir die Abhängigkeit der Zerfallsgeschwindigkeit von der Zeit in dem Zeitraum ermitteln, für den wir über historische Daten verfügen. Selbstverständlich ist diese Datierung nur eine Hypothese, weil die Gesetzmäßigkeiten, die in einem Zeitintervall festgestellt werden, für das historische Daten vorliegen, auf ein anderes Intervall ohne solche Daten extrapoliert werden. So kann man z.B., wenn man aus Štf die Geschwindigkeit des lexikalischen Wandels im Russischen für das Intervall 1500 bis 1900 n. Chr. mit 0,022 pro Jh. bestimmt hat, das Urslavische auf vor 14,5 Jh., d.h. auf Anfang des VI. Jh., festlegen (vgl. die ähnliche Datierung in der Literatur mit dem Ende des VI. Jh.).

Ein großer Vorzug unserer Methode ist, daß das Ergebnis durch einen unabhängigen Test mit Hilfe eines weiteren Wörterbuchs (für das Russische z.B. ONL), das einen anderen Stil repräsentiert, überprüft werden kann.

Wenn die Rangverteilung der Wörter, die vor dem betrachteten Zeitpunkt entstanden sind, nicht direkt gegeben ist, - wenn einfach unbekannt ist, welche Wörter dieser Periode zugerechnet werden müssen, und diese Wörter noch durch Vergleich der Lexika verwandter Sprachen gefunden werden müssen, - dann wird besonders wichtig, ob die verglichenen Sprachen sich unabhängig entwickelt haben. Die beiden untersuchten slavischen Sprachen - Russisch und Tschechisch - , die aus vielen Gründen für die Untersuchung geeignet sind, können gerade nicht als Beispiel unabhängiger Entwicklung dienen.

Bis jetzt waren die im Prinzip absolut datierbaren linguistischen Phänomene Mengen von Erbwörtern. Für diese Wörter läßt sich feststellen, daß sie spätestens zu einem bestimmten Zeitpunkt aufgetreten sind.

Der Lehnwortschatz ist im rangierten Häufigkeitswörterbuch anders verteilt als der Erbwortschatz. Die Möglichkeit, Lehn- und Erbwortschatz zu trennen, sowie die Datierung von Entlehnungen behandeln wir in Kap. 4.

Beschäftigen wir uns noch etwas näher mit dem Verhältnis unserer Methode zur Glottochronologie. Dazu ist ein genaues Verständnis der Rolle des Kernlexikons (core) in dieser Theorie nötig.

Es ist anzunehmen, daß M. Swadesh die Begriffe in seiner Liste so gewählt hat, daß in beliebigen Sprachen die ihnen entsprechenden Wörter im Häufigkeitswörterbuch annähernd gleichen Rang haben. Diese Hypothese läßt sich streng genommen allerdings nicht beweisen, weil die Auswahl der Entsprechungen für die 100 (200,215) Begriffe der Kernliste nicht eindeutig ist (die Zusatzbestimmungen verringern das Problem, räumen es aber nicht aus); man kann jedoch einige Gründe zu ihrer Stützung anführen. Wenn wir die von Fodor [16] bearbeitete Liste für das Russische und Tschechische nehmen, dürfen wir nicht vergessen, daß es sich nur um eine mögliche Variante von im Detail unterschiedlichen glottochronologischen Listen handelt. Nichtsdestoweniger läßt sich aus Tab. 20 ziemlich klar das Prinzip ihrer Erstellung erkennen. Die Wörter, die von den Verfassern der Liste zur stabilen, langsam zerfallenden Kernlexik gerechnet werden, sind in der Regel auch hochfrequent.

In Tab. 20 wird die Rangverteilung der 100 Wörter aus der Liste von Swadesh in vier Häufigkeitswörterbüchern angegeben. Stf, Cech, und für das Englische Dew und Thorndike [20]. Die Wörterbücher sind in Abschnitte eingeteilt, und es sind die Anzahl und die Prozentzahl der Wörter aus der 100-er Liste pro Abschnitt (Rangintervall) berechnet.

Der Zusammenhang zwischen dem Rang eines Wortes und seiner Wahrscheinlichkeit, zur Kernliste zu gehören, ist offensichtlich: Sie nimmt mit wachsendem Rang rasch ab; 16 bis 20% der Wörter aus der 100-er Liste gehören zu den 100 häufigsten Wörtern der jeweiligen Sprache, 83 bis 95% haben einen Rang unter 2500.

Außerdem sind die Rangverteilungen der Wörter aus der Kernliste in den verschiedenen Sprachen (auf jeden Fall im Russischen und Tschechischen) sehr ähnlich. Sollte sich zeigen, daß diese Verteilung sprachinvariant ist, wäre das umso besser; es würde nämlich zeigen, daß wir bei der Anwendung der klassischen Variante der glottochronologischen Methode die unterschiedliche Zerfallsrate der Rangklassen des Lexikons unwissentlich berücksichtigt hätten. Anders wäre es jedoch, wenn wir bei der Übersetzung der Liste von einer Sprache in eine andere den mittleren Rang der Wörter verändern, weil dann die Zerfallsrate der übersetzten Liste von der des Originals abweicht. Der eigentliche Sinn vieler Zusatzkonventionen für die Übersetzung der Wörter der Kernliste ist gerade die Sorge um die Erhaltung des mittleren Rangs.

Tabelle 20. Häufigkeitsverteilung der Wörter aus der glottochronologischen Liste von M. Swadesh

Rang	Štf Anzahl	%	Čech Anzahl	%	Dew Anzahl	%	Thorndide Anzahl	%
1 (1-100)	18	18	22	22	16	16		
2 (101-200)	8	8	10	10	8	8	63	
3 (201-500)	20	6,7	16	5,3	13	4,3		
4 (501-1000)	20	4	17	3,4	20	4	23	
5(1001-1500)	14	2,8	8	1,6				
6(1501-2500)	6	0,6	10	1	43		9	
> 2500	14	-	17	-			5	
	Σ=100		Σ=100		Σ=100		Σ=100	

Wir führen noch ein weiteres Indiz dafür an, daß die Veränderungsrate der Liste vom Rang ihrer Wörter abhängt. In [9] hat Swadesh eine Korrektur der ursprünglichen Liste vorgeschlagen, indem er sie auf 100 der vorher 200 (215) Wörter reduzierte. Dabei hat sich der prozentuale Anteil der Wörter, die nach 1000 Jahren erhalten sind, etwas erhöht (von 80 auf 86%). Es darf angenommen werden, daß die Häufigkeit der ausgeschlossenen Wörter im allgemeinen niedriger ist als der in der neuen 100er-Liste. Wir benutzen die 215er-Liste von Fodor für das Russische und Tschechische. In Tab. 21 ist die Rangverteilung der 100 Wörter der neuen Liste und der 125 ausgeschlossenen Wörter für Štf, Čech und Dew angegeben.

Tabelle 21. Häufigkeitsverteilung der Wörter aus den glottochronologischen Listen: 100er und 125er Liste

Rang	Štf		Čech		Dew	
	100	125	100	125	100	125
1 (1-100)	18	16	28	14	16	16
2 (101-200)	8	8	10	9	8	12
3 (201-500)	20	20	16	11	13	16
4 (501-1000)	20	17	17	22	20	18
5(1001-1500)	14	14	8	14		
6(1501-2500)	6	14	10	13	43	63
> 2500	14	36	17	42		
	Σ=100	Σ=125	Σ=100	Σ=125	Σ=100	Σ=125

Wir stellen die Nullhypothese auf, daß die Rangverteilung der Wörter nicht von der Zugehörigkeit zu der einen (100er) oder anderen (125er) Gruppe abhängt. Wir wenden den χ^2-Test mit (2-1) · (7-1) = 6 für Štf und Čech und mit (2-1) · (5-1) = 4 Freiheitsgraden für Dew an. Auf dem Signifikanzniveau 0,1 kann die Nullhypothese nur für Dew angenommen werden. Auf dem Signifikanzniveau 0,05 kann man sie auch für Štf annehmen, für Čech muß sie dagegen in jedem Fall abgelehnt werden. Für Stf kann man also behaupten, daß die Wörter der 125er Liste einen wesentlich höheren Rang haben.

4. EIN MATHEMATISCHES MODELL DES LEXIKALISCHEN WANDELS: ENTLEHNUNG

4.1. ENTLEHNUNGSQUELLEN

In diesem Kapitel wollen wir einige wichtige theoretische Annahmen zum lexikalischen Zuwachs formulieren, die eine natürliche Erweiterung der Theorie des lexikalischen Zerfalls aus Kap. 2 und 3 darstellen.

Wir gehen davon aus, daß ein Lexikon nur durch Entlehnung wächst. Es sei daran erinnert, daß wir zur terminologischen Vereinfachung auch Neubildungen innerhalb der Sprache als Entlehnungen zählen (s. Kap. 1). Die Entlehnungsquellen einer Sprache x seien die Sprachen y_j; wir nehmen an, daß die Zahl $j = 0,1,2,\ldots,m$ der Entlehungsquellen immer endlich ist, und wollen die Sprache x selbst als nullte Entlehnungsquelle bezeichnen ($y_o = x$). Die Entlehnungsquellen y_j einer Sprache x können verschiedene Aktivitäten p_j besitzen, wobei gilt:

$$\sum_{i=0}^{m} p_j = 1. \qquad (4.1)$$

Der Begriff der Aktivität wird unten inhaltlich erklärt, bis dahin ist sie eine Zahl, die einem Paar (y_j,x) zugeordnet wird und der Bedingung (4.1) genügt.

Entlehnungsquelle kann eine Sprache, ein Dialekt oder eine Sprachtradition sein, doch diese Fälle werden nicht unterschieden, weil hier nicht die Veränderung der Quelle, sondern nur die Veränderung ihrer Aktivität bezüglich der Sprache x in der Zeit interessiert. Wir fassen also die Aktivität einer Quelle als eine Funktion der Zeit auf:

$$p_j = p_j(t).$$

Leider haben wir kein geeignetes Modell, das alle möglichen Aktivitätsveränderungen von Quellen in der Zeit erfaßte. Dazu fehlen einfach entsprechend aufbereitete empirische Daten. Wir analysieren daher nur stark vereinfachte Fälle und beschränken uns auf qualitative Überlegungen zu den komplexeren.

Unser Ziel ist eine Theorie, mittels derer man, mit einigen Annahmen zur Abhängigkeit der Aktivität einer Quelle von der Zeit, voraussagen kann, wie Entlehnungen aus dieser Quelle im rangierten Lexikon verteilt sein werden. Aus Kap. 2 ist bekannt, wie die Erbwörter einer Sprache x aus einer Vorgängersprache x_i ($x_1 \Rightarrow x$) verteilt sind, unabhängig vom zeitlichen Abstand der beiden Sprachen. Durch Gegenüberstellung dieser Verteilung mit der Verteilung der Lehnwörter kommen wir zum Ergebnis dieser Arbeit: einem Kriterium für Sprachverwandtschaft.

4.2. DIE RANGVERTEILUNG DER WÖRTER, DIE IN EINER SPRACHE WÄHREND EINES ZEITINTERVALLS NEU AUFTRETEN

Zuerst ist eine Reihe einfacher Fragen zu klären. Zunächst einmal, wieviele Wörter treten im Lexikon zwischen t_1 und t_2 neu auf (werden entlehnt)? Sei $\nu_i(t_1,t_2)$ der Anteil der Entlehnungen im Zeitintervall (t_1,t_2) in der i-ten Klasse des Lexikons[1]).

Offensichtlich entstehen während dieses Intervalls nur soviele Neubildungen in $G_i \subseteq V_x$, wie alte Wörter schwinden:

$$\nu_i(t_1,t_2) = N_i(t_1) - N_i(t_2) = e^{-\eta_i t_1} - e^{-\eta_i t_2}. \qquad (4.2)$$

Damit kennen wir die Anzahl der Neubildungen pro Klasse.

Wir nennen die Neubildungen in der i-ten Klasse eine Schicht. Unter der Schicht $S(t_1,t_2)$ im Lexikon V_x verstehen wir die Summe

$$S(t_1,t_2) = \sum_{i=1}^{\infty} \nu(t_1,t_2).$$

Entsprechend wird die Schicht S_k des k Klassen großen Anfangsabschnitts eines Lexikons definiert:

$$S_k = \sum_{i=1}^{K} \nu_i(t_1,t_2).$$

Wir definieren nun die Aktivität der j-ten Quelle folgendermaßen: Wenn y_j während (t_1,t_2) die konstante Aktivität p_j $(0 < p_j \leq 1)$ besitzt, so beträgt der Anteil von Lehnwörtern aus dieser Quelle in der Klasse G_i

$$q_{ij}(t_1,t_2) = p_j \nu_i(t_1,t_2).$$

Diese Definition beruht auf einer wichtigen Voraussetzung: Die Verteilung der Lehnwörter im Lexikon hängt von der Zeit ihrer Entlehnung und der Aktivität der Quelle ab, nicht aber von ihrer Art. So verteilen sich die Entlehnungen jeweils gleichmäßig auf das ganze Lexikon, wobei die Wahrscheinlichkeit dafür, daß eine Entlehnung - gleich aus welcher Sprache - in eine bestimmte Klasse fällt, nur von den vorhandenen „Vakanzen" (geschwundenen Wörtern) abhängt.

Betrachten wir nun die Funktion (4.2) genauer, so stellen wir zunächst fest, daß sie für $t_1 = 0$ (in der Vergangenheit) und $t_2 = t$ (Jetztzeit) in (2.5) übergeht, den Anteil der Wörter, die seit t_1 geschwunden sind.

Entsprechend suchen wir nun das Zeitintervall, für das (4.2) in (2.6), den Ausdruck für den Anteil der erhaltenen Wörter, übergeht. Die eine Intervallgrenze ist irgendein Zeitpunkt $t_2 = t$, die andere bestimmt sich dadurch, daß $e^{-\eta_i t} \to 0$ bei $t \to \infty$, d.h. $t_2 \to \infty$. Und damit:

$$\nu_i(t_1,t_2) = \bar{N}_i(t) \text{ bei } t_1 = 0,\ t_2 = t, \tag{4.3}$$

$$\nu_i(t_1,t_2) = N_i(t) \text{ bei } t_1 = t,\ t_2 \to \infty. \tag{4.4}$$

$$\nu_i(t_1,t_2) = N_i(t_1) - N_i(t_2) \text{ bei } t_1 \geq 0,\ t_2 < \infty. \tag{4.5}$$

Wie hängen nun diese Funktionen vom Rang i ab? Im Fall von (4.3) handelt es sich um eine monoton wachsende Funktion; für einen n_o langen Anfangsabschnitt eines Lexikons hat sie einen Grenzwert in n_o. Die Funktion (4.4) ist monoton fallend und hat einen Grenzwert für $i = 0$. Funktion (4.5) ist nicht monoton, sie hat ein Extremum, dessen Stelle und dessen Wert von der Wahl des Zeitintervalls (t_1,t_2) abhängen. Zur exakten Bestimmung setzen wir in (4.2) $\eta_i = k\sqrt{i}$ ein und leiten die Funktion nach i ab:

$$\nu_i' = (e^{-k\sqrt{i}t_1} - e^{-k\sqrt{i}t_2})' = \frac{kt_1}{2\sqrt{i}} e^{-k\sqrt{i}t_1} - \frac{kt_2}{2\sqrt{i}} e^{-k\sqrt{i}t_2} =$$

$$= \frac{k}{2\sqrt{i}} (t_1 e^{-k\sqrt{i}t_1} - t_2 e^{-k\sqrt{i}t_2}).$$

Die Ableitung setzen wir gleich Null:

$$t_1 e^{-k\sqrt{i}t_1} - t_2 e^{-k\sqrt{i}t_2} = 0.$$

$$1 - \frac{t_2}{t_1} e^{-k\sqrt{i}(t_2-t_1)} = 0.$$

$$e^{-k\sqrt{i}(t_2-t_1)} = \frac{t_2}{t_1}.$$

$$k\sqrt{i}(t_2-t_1) = \ln \frac{t_2}{t_1},$$

und damit ist

$$i_{max} = \left[\frac{\ln t_2 - \ln t_1}{k(t_2-t_1)}\right]^2 = \frac{1}{k^2}\left(\frac{\ln t_2 - \ln t_1}{t_2-t_1}\right)^2. \tag{4.6}$$

Wir stellen t_2 als ht_1 dar, bei $t_2 > t_1$ gilt $h > 1$. Dann ist $\ln t_2 - \ln t_1 = \ln \frac{t_2}{t_1} = \ln h$. Da $kt_1 < 1$ und $kt_2 < 1$, ist auch

$kht_1 < 1$, und folglich $kt_1 < \frac{1}{h}$. Wir erhalten

$$\frac{\ln t_2 - \ln t_1}{k(t_2-t_1)} = \frac{\ln h}{kt_1(h-1)} > \frac{h \ln h}{h-1}. \qquad (4.6a)$$

h bewegt sich im Intervall $1 < h < \infty$. Wir untersuchen nun das Verhalten der Funktion $\frac{h \ln h}{h-1}$ für $h \to 1$ unter Verwendung der Regel von L'Hôpital:

$$\frac{\ln h + h\frac{1}{h}}{1} \to \frac{0+1}{1} \to 1 \text{ für } h \to 1.$$

Das Verhalten der monotonen Funktion $\frac{h \ln h}{h-1}$ bestimmt sich weiter durch das Vorzeichen ihrer Ableitung:

$$\left(\frac{h \ln h}{h-1}\right)' = \frac{(\ln h + 1)(h-1) - h \ln h}{(h-1)^2} = \frac{h - \ln h - 1}{(h-1)^2} \qquad (4.6b)$$

In (4.6b) ist der Nenner positiv und der Zähler ist die Differenz von $y_1 = h - 1$ und $y_2 = \ln h$. Bei $h = 1$ ist $y_1 = y_2 = 0$. Bei $h > 1$ ist die Differenz $y_1 - y_2 > 0$. Damit

$$\frac{h \ln h}{h-1} \to 1 \text{ für } h \to 1$$

und die erste Ableitung hat immer positives Vorzeichen. Also wächst $\frac{h \ln h}{h-1}$ monoton, und wegen (4.6a) ist

$$\frac{\ln t_2 - \ln t_1}{k(t_2 - t_1)} > 1$$

und auch

$$\left[\frac{\ln t_2 - \ln t_1}{k(t_2 - t_1)}\right]^2 > 1.$$

Das bedeutet, daß die Funktion (4.2) ein lokales Maximum für beliebiges $t \neq 0$ hat.

Für weiter zurückliegende Intervalle verschiebt sich das Maximum in Richtung des Koordinatenursprungs, für nähere entgegengesetzt.

Wenn wir $\Delta t = t_2 - t_1 = 1$ Jh. nehmen, und $t_1 = 900$ Jahre, dann ist für $k = 0{,}025$ $i_{max} = 16$. Das bedeutet, daß die meisten Wörter, die in der Sprache zwischen 974 n. Chr. und 1074 n. Chr. entstanden sind, in einem 1971 erstellten Häufigkeitswörterbuch unter den Wörtern 1501 bis 1600 sein werden. Doch schon für die Neubildungen des XVII. Jh. führt die gleiche Rechnung zu einem Maximum in der 134. Rangklasse!

Interessant ist nicht nur die Stelle des Maximums, sondern auch sein Wert $\nu_i(t_1,t_2)$. Setzen wir daher in (4.2) den Wert von (4.6) ein:

$$\nu(i_{max},t_1,t_2) = e^{-\frac{t_1(\ln t_2 - \ln t_1)}{(t_2-t_1)}} - e^{-\frac{t_2(\ln t_2 - \ln t_1)}{(t_2-t_1)}}.$$

Dann ist

$$\nu(i_{max},t_1,t_2) = \left(\frac{t_2}{t_1}\right)^{-\frac{t_1}{(t_2-t_1)}} - \left(\frac{t_2}{t_1}\right)^{-\frac{t_2}{(t_2-t_1)}}. \qquad (4.7)$$

Durch Einsetzen der obigen Werte von t_1, t_2 und k in (4.7) ergibt sich der Anteil der zwischen 974 und 1074 entstandenen Wörter in der 16. Rangklasse mit ca. 4%, und für Neubildungen des XVII. Jh. ist der Wert des Maximums 14%.

Es ist leicht einzusehen, daß für längere Zeitintervalle eine flachere Kurve $\nu_i(t_1,t_2)$ mit weniger ausgeprägtem Extremum entsteht.

Mit $t_2 = t_1 + \Delta t$ und festem Δt bilden wir die Funktion $\nu_i(t)$. Dann entspricht jedem $t = t_1$ ein i_{max} - der Rang, in dem $\nu_i(t) = \nu_i(t_1,t_2+\Delta t)$ ihr Maximum erreicht -, und jedem Rang i ein t, so daß $\nu(t)$ bei i ein Maximum erreicht. Mit anderen Worten ist

(4.6) eine eineindeutige Abbildung der Zeit auf den Rang. Von häufigen zu seltenen Wörtern kommend bewegt man sich sozusagen sprachgeschichtlich auf die Jetztzeit zu, wobei sich das Verhältnis von Erbwörtern zu Lehnwörtern ändert. Das ist aber nur eine zusätzliche Demonstration der oben angesprochenen Dualität von Zeit und Rang.

Schließlich wollen wir noch klären, wie groß der mittlere Anteil der Neubildungen in den ersten τ Klassen des Lexikons, d.h. wie "dick" die Schicht $S(t_1,t_2)$ ist. Diese Frage ist von praktischem Interesse, weil bei kleinem Zeitintervall und geringer Geschwindigkeit des lexikalischen Wandels die Streuung von $\nu_i(t_1,t_2)$ - des Anteils der Lehnwörter einer gegebenen Schicht an einer einzelnen Klasse - beträchtlich sein muß, und folglich der Vergleich von empirischen und theoretischen Werten für einzelne Klassen schwierig ist.

Wir bezeichnen die gesuchte Größe mit $S_r(t_1,t_2)$. Dann ist offenbar bei $a = \gamma t_1$ und $a_1 = \gamma t_2$

$$S_r = \frac{1}{r}\left(\sum_{i=1}^{r} e^{-a_1\sqrt{i}} - \sum_{i=1}^{r} e^{-a\sqrt{i}}\right).$$

Besonders interessant ist die Summe $\sum_{i=1}^{r} e^{-a\sqrt{i}}$, weil sie die Anzahl der mindestens t Jahre alten Wörter angibt. Im Gegensatz zu der schon vertrauten Formel (2.6) gibt sie nicht die Verteilung für aufeinanderfolgende, sondern für einander enthaltende Klassen an: 1 bis n, 1 bis 2n,..., 1 bis rn. Eine gute Näherung für diese Summe ist das Integral

$$J = \int_1^r e^{-a\sqrt{x}}dx.$$

Zur Berechnung des Integrals J transformieren wir die Variable x: $z = a\sqrt{x}$:

$$z^2 = a^2x;\ x = \frac{z^2}{a^2};\ dx = \frac{2z}{a^2}\,dz.$$

$$J = \frac{2}{a^2}\int_a^{a\sqrt{r}} ze^{-z}dz = \frac{2}{a^2}\int_a^{a\sqrt{r}} zd(e^{-z}) = -\frac{2}{a^2}ze^{-z}\Big|_a^{a\sqrt{r}} +$$

$$+\frac{2}{a^2}\int_a^{a\sqrt{r}} e^{-z}dz = \left(-\frac{2}{a^2}ze^{-z} - \frac{2}{a^2}e^{-z}\right)\Big|_a^{a\sqrt{r}} =$$

$$= -\frac{2}{a^2}(a\sqrt{r}e^{-a\sqrt{r}} + e^{-a\sqrt{r}} - ae^{-a} - e^{-a}) =$$

$$= -\frac{2}{a^2}[e^{-a\sqrt{r}}(a\sqrt{r} + 1) - e^{-a}(a + 1)] =$$

$$= \frac{2}{a^2}[e^{-a}(a + 1) - e^{-a\sqrt{r}}(a\sqrt{r} + 1)]. \qquad (4.8)$$

Setzen wir jetzt den Wert von J in den Ausdruck für $S_r(t_1,t_2)$ ein und setzen $a = kt_1$ und $a_1 = kt_2$, erhalten wir die Näherung

$$S_r \approx \frac{2}{r}\left\{\frac{2}{k^2t_2^2}[e^{-kt_2}(kt_2 + 1) - e^{-kt_2\sqrt{r}}(kt_2\sqrt{r} + 1)] -\right.$$

$$\left. -\frac{1}{k^2t_2^2}[e^{-kt_1}(kt_1 + 1) - e^{-kt_1\sqrt{r}}(kt_1\sqrt{r} + 1)]\right\}. \qquad (4.9)$$

4.3. DIE VERTEILUNG DER LEHNWÖRTER IM LEXIKON

Betrachten wir zunächst einen etwas konstruierten Fall. Das Lexikon der Sprache x setze sich folgendermaßen zusammen: Jedes Wort von V_x stamme aus einer von m Quellen; jede dieser Quellen sei eine bestimmte Zeit aktiv gewesen, doch alle zusammen seien schon unendlich lange aktiv, jedoch nie zwei gleichzeitig. Anders ausgedrückt, die j-te Quelle „füllt" die Schicht (t_j, t_{j+1}): alle Wörter der Sprache, die zwischen t_j und t_{j+1} entstanden sind, stammen aus y_j.

Gegeben sind uns das Lexikon V_x und die Quelle für jedes Wort aus V_x. Können wir nun bestimmen:

a) die Reihenfolge der Quellenaktivität,

b) die Aktivitätsdauer?

Offenbar ja, jedenfalls im Prinzip (es mag technische Probleme geben, z.B. bezüglich der Geschwindigkeitsschwankungen beim lexikalischen Wandel, von diesen wird aber abgesehen). Jede Quelle füllt eine ganze Schicht, folglich müssen die Entlehnungen aus einer Quelle im Lexikon nach (4.2) verteilt sein; das Entlehnungsmaximum für diese Quelle gibt (4.6) an. Wir müssen also empirisch die p_n, $1 \leq n \leq m$, bestimmen.

Obwohl der Fall ziemlich konstruiert ist, trägt er wesentlich zur Klärung des Unterschieds zwischen den Verteilungen von Erb- und Lehnwörtern bei. Vielleicht könnte man unsere Absicht genauer so formulieren: Welche intuitiv möglichst befriedigende Explikation von „Erbwort" ist im Rahmen des eingeführten Begriffssystems möglich? Welche Schicht kann als Erbwortschatz gelten?

In unserem Fall ist es naheliegend, diejenige Schicht als ererbt anzunehmen, der keine andere zeitlich vorausgeht, d.h. die älteste . Da wir hier Schicht und Quelle miteinander identifiziert haben, kann man sagen, daß die Sprache x von der Quelle abstammt, die als erste aktiv war. Offenbar muß dann diese Quelle vor dem Zeitpunkt t, wo die nächste Quelle sie ablöst, schon unendlich lange aktiv gewesen sein. Unter dieser Bedingung (vgl. (4.4))sind die Wörter dieser Schicht monoton fallend verteilt, und das Maximum von Wörtern aus dieser Quelle liegt in der ersten Rangklasse des Lexikons.

Alle übrigen Quellen sind hier Entlehnungsquellen, und ihre entsprechenden Verteilungen sind nicht monoton oder aber nicht fallend.

Das bedeutet, man kann die Quelle (Sprache), aus der x stammt, leicht finden, wenn man ein Lexikon V_x mit nach Herkunft klassifizierten Wörtern hat. Es ist die Quelle, deren Wörter im Lexikon V_x monoton fallend verteilt sind.[2)]

Der beschriebene Fall ist keineswegs der Normalfall; denn gewöhnlich wirken mehrere Quellen gleichzeitig, wobei ihre Aktivität veränderlich ist. Wir wollen aber zeigen, daß das angegebene Kriterium über den Spezialfall hinaus anwendbar ist.

Betrachten wir dazu den Fall, wo mehrere Quellen gleichzeitig wirken. Wir untersuchen die Verteilung der Entlehnungen aus einer einzelnen Quelle y_j mit der Aktivität $p_j = p_j(t)$.

Wir wählen ein hinreichend kleines Zeitintervall Δt, innerhalb dessen die Aktivitätsveränderung der Quelle vernachlässigbar gering ist. In der i-ten Klasse des Lexikons ergibt sich im Intervall t_1 bis $t_2 = t_1 + \Delta t$ die folgende Anzahl von Entlehnungen:

$$q_{ij}(t_1,t_2) = p_j \nu_i(t_1,t_2) = p_j\left(e^{-kt_1\sqrt{i}} - e^{-k(t+\Delta t)\sqrt{i}}\right) =$$

$$= p_j\left(1-e^{-k\Delta t\sqrt{i}}\right)e^{-kt_1\sqrt{i}}.$$

Mit $\lambda = k\sqrt{i}$,

$$K_\lambda = e^{-k\Delta t\sqrt{i}} = e^{-\lambda\Delta t}$$

ist dann

$$q_{\lambda j}(\Delta t) = p_j(1-K_\lambda)e^{-\lambda\Delta t}.$$

Wirkt die Quelle y_i im Intervall T mit $p_j(t) > 0$ für $t_1 \leq T \leq t_2$, dann ist

$$q_{\lambda j}(t_1,t_2) = \int_{t_1}^{t_2} p_j(\tau)(1-K_\lambda)e^{-\lambda\tau}d\tau. \qquad (4.10)$$

Im Spezialfall, wo die Aktivität der Quelle im Intervall konstant ist, $p_j(t) = p_j = \text{const}$ für $t_1 \leq \tau \leq t_2$, ist

$$q_{\lambda j}(t_1,t_2) = p_j(1-K_\lambda) \int_{t_1}^{t_2} e^{-\lambda\tau} d\tau. \qquad (4.11)$$

(4.10) ist eine Funktion von $\lambda = k\sqrt{i}$, d.h. des Ranges, und beschreibt die Verteilung der Entlehnungen aus y_j in V_x für den Fall, daß diese Quelle im Intervall (t_1,t_2) ununterbrochen gewirkt hat, möglicherweise aber mit unterschiedlicher Aktivität. Wir nehmen an, daß diese Quelle im Intervall (t,∞) wirksam war, daß also keine Quelle vor y_j zu wirken begonnen hat. Oben haben wir ebenfalls die Wörter aus einer solchen Quelle als die Erbwörter angesehen, nur daß die Quelle allein und mit konstanter Aktivität gewirkt hat. Dort konnten wir diese Quelle an der monoton fallenden Verteilungsfunktion ihrer Wörter erkennen. Wir zeigen, daß auch bei veränderlicher Aktivität der Quelle y_i,

$$p_j = p_j(t),$$

die Verteilungsfunktion monoton fallend bleibt, unter der einzigen Bedingung, daß keine Quelle vor y_j zu wirken beginnt.

Unter der Bedingung $t_1 > 0$ und $t_2 = \infty$ ist der Faktor vor dem Integralzeichen in (4.10) wegen $K_\lambda \to 0$ bei $\Delta t \to \infty$ gleich 1. Wir prüfen, ob die Funktion

$$q_{\lambda j}(t) = \int_t^\infty p_j(\tau)e^{-\lambda\tau}d\tau \qquad (4.12)$$

bezüglich λ monoton ist. Wir differenzieren unter dem Integral nach λ:

$$q'_{\lambda j}(t) = \frac{d}{d\lambda}\int_t^\infty p_j(\tau)e^{-\lambda\tau}d\tau = \int_t^\infty p_j(\tau)(-\tau)e^{-\lambda\tau}d\tau. \qquad (4.13)$$

Alle drei Faktoren in (4.13) haben konstantes Vorzeichen: $p_j(\tau)$ ist nach Definition positiv, $0 < \tau \leq \infty$, $e^{-\lambda\tau} \geq 0$. Folglich hat die Ableitung von (4.12) für alle λ gleiches Vorzeichen. Da $\lambda = k\sqrt{i}$, ist $q_{\lambda j}$ monoton bezüglich λ für $0 < \tau \leq \infty$ und beliebige Abhängigkeit der Quellenaktivität p_j von der Zeit t.

Wir kommen also zu dem Schluß, daß man mit Hilfe der Verteilung der Wörter aus einer gegebenen Quelle im rangierten Lexikon zumindest zwei Typen von Quellen unterscheiden kann. Die Quellen des einen Typs (für die der Anteil ihrer Wörter mit dem Rang monoton abnimmt) kann man Quellen des Erbwortschatzes nennen, und kann, wenn man so will, sagen, daß die jeweilige Sprache von ihnen abstammt. Die übrigen Quellen kann man sinnvoll als Entlehnungsquellen bezeichnen (unter ihnen kann man noch die „autochthonen Neubildungen" ausgrenzen). Dabei soll mit „Sprache x stammt von Sprache y_j ab" nur gemeint sein, daß keine Quelle y_i, $i \neq j$, angegeben werden kann, die vor y_j zu wirken begonnen hat. Daher schließt die Theorie die Möglichkeit nicht aus, daß eine Sprache von mehreren Sprachen abstammt (Sprachmischung). Dabei besagt das Fehlen von Beispielen für Sprachmischung in unserem Datenmaterial (s. Kap. 5) nichts[3].

Abschließend einige Worte zur Datierung von Entlehnungen. In Kap. 3, beim Problem der Datierung der Erbwörter aus einer Quelle, bildete die Abhängigkeit der Zerfallsrate von der Zeit die Hauptschwierigkeit. Bei mehreren Quellen entstehen zusätzlich Probleme im Zusammenhang mit der Aktivitätsveränderung. Offenbar gibt es Fälle, wo die Abhängigkeit der Aktivität einer Quelle von der Zeit approximativ bestimmt werden kann, ohne über eine vollständige Chronologie der betrachteten Sprache zu verfügen. Die Veränderung der Aktivität läßt sich nämlich mit der relativen Chronologie des kulturellen Einflusses einer ethnischen Gruppe auf die andere in zusammenhang bringen, wie sie z.B. von der Archäologie geliefert wird.

Wir verfügen noch nicht über genügend empirische Daten zum Zusammenwirken mehrerer Quellen, weshalb es verfrüht wäre, sich mit technischen Datierungsverfahren zu beschäftigen, die wesentlich auf Daten zur Abhängigkeit der Aktivität von der Zeit beruhen.

5. EMPIRISCHES MATERIAL ZUR ENTLEHNUNG

5.1. ENTLEHNUNGSQUELLEN, DIE VON t AN BIS ZUR GEGENWART WIRKSAM WAREN

Der Anteil von Entlehnungen aus einer gegebenen Quelle, die im Intervall (t,0) als einzige wirksam ist, beträgt für die i-te Klasse:

$$\nu_i(t,0) = \bar{N}_i(t) = 1 - e^{-\eta_t\sqrt{i}} .$$

Wohl das klassische Material zur Untersuchung einer solchen Entlehnungssituation sind die Werke des mittelalterlichen englischen Schriftstellers Chaucer. Statistisches Material zum Verhältnis von romanischer zu germanischer Lexik in seinem Gesamtwerk ist von Mersand [28] zusammengestellt worden. Diese Daten wurden in den Arbeiten Yule [29] und später Herdan [30] verwendet. Auf die letztere werden wir noch zurückkommen. Chaucers Werke sind als Material für diesen Zweck deshalb geeignet, weil sie zum Ende der mittelenglichen Periode (1100-1400) entstanden sind, in der die weitaus meisten Neubildungen im Englischen romanische (lateinische oder französische) Entlehnungen waren. Sie machten ca. 90% aller neuen Wörter im Englischen aus, während der Anteil genuin englischer Neubildungen und skandinavischer Lehnwörter 10% nicht überstieg. In der vorausgegangenen altenglischen Periode war es gerade umgekehrt: die romanischen Entlehnungen machten einen winzigen Bruchteil der germanischen Neubildungen aus. Daher kann man vergröbernd annehmen, daß alle romanischen Entlehnungen, die bei Chaucer auftreten, aus den letzten 300 Jahren vor der Abfassung seiner Werke stammen.

Für den Anteil der germanischen Lexik kann man dabei

$$N_i(t) = e^{-\eta_t\sqrt{i}} . \qquad (5.1)$$

benutzen, das ja die Veränderung des Anteils der weniger als t Jahre alten Wörter (t hier ca. 300 Jahre) mit dem Rang angibt.

Über ein Häufigkeitswörterbuch des Mittelenglischen verfügen wir nicht; wir beschaffen uns die Häufigkeitsverteilung der Wörter indirekt. Wir bedienen uns der bei Mersand angegebenen Zahlen zum Wortbestand der einzelnen Werke Chaucers[1]. Für jedes Werk ist der Anteil der romanischen und germanischen Wörter an seinem Wortbestand angegeben. Wir gehen dabei von folgender Überlegung aus: ein Wortbestand von l Wörtern verhält sich genauso wie ein l Wörter langer Anfangsabschnitt eines Häufigkeitswörterbuchs. Wenn der Wortbestand l_2 eines Werkes größer ist als der Wortbestand l_1 eines anderen, dann zerfällt l_2 schneller als l_1, d.h. l_2 enthält mehr Neubildungen (= Entlehnungen). Daher kann man die Wortbestände $l_1 > l_2 > \ldots > l_k$ als k sich enthaltende Klassen eines Häufigkeitswörterbuchs auffassen. Die Abhängigkeit der Zerfallsrate vom Rang gibt (4.8) an.

Diese Überlegung führt natürlich nur zu einer Näherung. Der Wortbestand eines einzelnen Werkes von z.B. tausend Wörtern kann auch Wörter umfassen, die im Häufigkeitswörterbuch nicht unter die ersten tausend fallen, und muß nicht unbedingt alle diese enthalten. Wir können aber hier die Zahl der Entlehnungen über mehrere Werke mit annähernd gleich großem Wortbestand (± 100 Wörter) mitteln. Dann muß die Mehrzahl der häufigen Wörter, die im Häufigkeitswörterbuch am Anfang stehen, auch noch im Wortbestand eines kurzen Werkes auftreten; der wahre Rang der seltenen Wörter ist nicht so wichtig, weil sich die Zerfallsgeschwindigkeiten mit steigendem Rang ziemlich schnell einander angleichen.

Tatsächlich fällt die Ableitung von (5.1) langsamer als die Funktion selbst:

$$N_t'(i) = -\eta_t e^{-\eta_t\sqrt{i}} \cdot \frac{1}{2\sqrt{i}} .$$

In Tab. 22 sind für alle Werke Chaucers[2] der Umfang des Wortbestands, die Textlänge und der prozentuale Anteil romanischer Lexik nach Mersand angegeben. Die Werke sind in der Tabelle nach

Umfang des Wortbestands geordnet und zu Klassen zusammengefaßt. Eine Klasse umfaßt die Werke mit einem Wortbestand von 100j + 1 bis 100(j + 1) Wörtern (j = 0,1,2,...). Insgesamt ergeben sich 22 Klassen zu ein bis acht Texten.

Wie aus Tab. 22 ersichtlich, wächst mit dem Wortbestand auch der Anteil romanischer Lehnwörter, der z.B. für „kurze" Texte (38 Wörter) 8% und für „lange" (3200 Wörter) 51% ausmacht.

In Tab. 23 wird für jede der Klassen aus Tab. 22[3)] die mittlere Anzahl germanischer Wörter angegeben. In der Spalte „Erwartete Anzahl germanischer Wörter" sind die Werte von (4.8) mit nach der Methode der kleinsten Fehlerquadrate geschätztem Parameter a aufgeführt. Wenn auch die Beobachtungswerte von den berechneten etwas abweichen, läßt sich doch mit Hilfe des χ^2-Tests (mit 22-2 = 20 Freiheitsgraden) zeigen, daß diese Abweichungen ($\chi^2 = 8{,}15$) nicht größer sind, als statistisch zu erwarten war ($\chi^2_{0,1} = 28{,}4$), und daß kein Grund zur Annahme besteht, die romanischen Entlehnungen seien nicht nach (4.8) mit a = 0,020 verteilt.

Herdan hat in der erwähnten Arbeit [30] die Hypothese aufgestellt, daß der Anteil der Lehnwörter $\bar{d}$ und die Textlänge L logarithmisch (s. Abb. 27) zusammenhängen. Wir nehmen eine Abhängigkeit der Form $\bar{d} = 1 - \frac{1}{c \ln(L)}$ an. Untersuchen wir nun, inwieweit diese Hypothese mit den o.g. Ergebnissen zum Zusammenhang von Wortbestand und Anteil von Lehnwörtern darin konsistent ist.

Tabelle 22. Der Anteil romanischer Lehnwörter in den Werken G. Chaucers
Die Werke sind nach wachsendem Wortbestand geordnet.

Rang	Name des Werks	Umfang des Wortbestands	Textlänge	% romanischer Lexik (nach Mersand)
1 (1-100)	Prov.	35	44	8,57
	Adam.	38	51	7,89
	Gentill.	71	157	31,94
	Bal. Comp.	84	173	32,14
	Wom. Unc.	89	173	19,10
	Rosem.	92	183	30,43
	Merc. B.	95	308	26,72
	Purs.	99	205	23,23

Tabelle 22. (Fortsetzung)

Rang	Name des Werks	Umfang des Wortbestands	Textlänge	% romanischer Lexik (nach Mersand)
2 (101-200)	Lak. St.	102	195	31,37
	Wom. Nob.	113	225	41,56
	Truth	116	218	17,24
	Bukton	123	264	15,45
	Scog	182	363	25,82
	A. Ck.	193	451	24,74
3 (201-300)	Form. A.	227	473	31,72
	Venus	234	592	31,61
	Am. Comp.	248	744	20,33
	Fort.	251	576	39,04
4 (301-400)	Comp. Pite	301	911	28,57
	Comp. L.	354	1018	17,76
	B. Throp.	399	1156	27,82
5 (401-500)	A. B. C.	445	1411	35,06
	L. G. W. A.	489	4209	30,47
6 (501-600)	B. Prior.	501	1824	28,94
	H. Manc.	512	1976	30,86
	C. Phis.	562	2179	35,59
7 (601-700)	Mars	618	2283	32,85
	D. WB.	628	3210	35,19
	D. Friar.	632	2832	31,96
	L. G. W. B.	644	4230	23,91
	Anel.	675	2696	29,63
8 (701-800)	B. Ship.	701	3474	32,47
	A. Rv.	760	3308	21,71
9 (801-900)	G. SN.	827	4069	33,13
	Host	829	4251	31,00
	C. Pard.	844	3721	33,51
	Astr.	884	4779	45,25
10 (901-1000)	D. Sum.	960	4571	36,67
	F. Squire	979	5181	37,49
	B. NP.	989	4821	35,26
11(1001-1100)	A. Mil.	1016	5140	26,67
	G.CY.	1053	5952	33,33
	F. Frank	1059	6878	37,58
	B. Monk.	1064	5889	37,50
12(1101-1200)	PF	1101	5376	37,21
13(1101-1200)	D. WB. Prof	1110	6631	34,42
	BD.	1153	8590	28,79
14(1201-1300)	E. Clerk.	1231	8994	40,29
	B. ML.	1242	7852	37,84

Tabelle 22. (Fortsetzung)

Rang	Name des Werks	Umfang des Wortbestands	Textlänge	% romanischer Lexik (nach Mersand)
15(1301-1400)	E. Mch.	1335	9112	38,58
	B. Mel.	1345	16826	46,47
16(1401-1500)	A. Prol.	1449	6683	39,06
17(1501-1600)	R. R. A.	1591	10583	37,84
	R. R. C.	1597	11787	44,52
18(1601-1700)	HF.	1654	12969	37,24
	L. G. W.	1671	21238	36,92
19(2001-2100)	A. Kn.	2059	17149	40,41
20(2201-2300)	R. R. B.	2267	21613	45,96
21(2601-2700)	Boeth.	2698	31187	49,85
22(3001-3100)	T. C.	3032	46908	43,47
23(3101-3200)	I. Pars.	3200	29672	51,27

Tabelle 23. Beobachtete und theoretische Verteilung der germanischen Lexik in den Werken G. Chaucers, nach Umfang des Wortbestands klassifiziert

Rang (Umfang des Wortbestands)	Mittlere Anzahl germanischer Wörter	Erwartete Anzahl germanischer Wörter (bei a = 0,020)
1-100	77,50	84,80
101-200	73,74	77,92
201-300	69,32	76,08
301-400	75,22	72,24
401-500	67,00	69,70
501-600	67,50	66,93
601-700	69,28	64,71
701-800	73,00	63,76
801-900	64,80	61,69
901-1000	63,60	60,06
1001-1100	66,20	58,61
1101-1200	66,60	57,59
1201-1300	61,00	55,96
1301-1400	57,50	55,22
1401-1500	60,94	54,18
1501-1600	58,82	52,84
1601-1700	62,92	51,76
2001-2100	59,59	48,50
2201-2300	54,04	46,90
2601-2700	50,15	44,42
3001-3100	56,53	41,68
3101-3200	48,73	41,17

Das ist deshalb umso interessanter, als Herdan seine Hypothese gerade aufgrund derjenigen Werke Chaucers formuliert hat, die auch wir oben verwendet haben (Abb. 26).

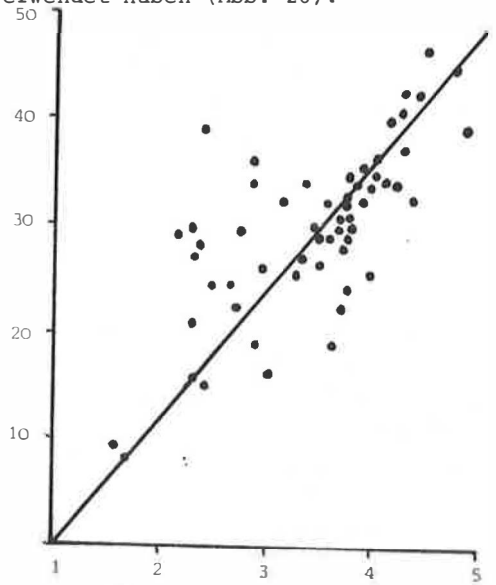

Abb. 26. Der Zusammenhang von Textlänge und Anzahl der Entlehnungen. x: Logarithmus der Textlänge; y: Romanische Lehnwörter in Prozent. (nach Herdan)

Wir fassen den Umfang k des Wortbestandes als Funktion der Länge L seines Textes auf (k = f(L)) und setzen in (4.8) ein.

Wir erhalten die Anzahl D der Erbwörter im Wortbestand eines Textes der Länge L:

$$D = \frac{2}{a^2}\left[e^{-a}(a+1) - e^{-a\sqrt{f(L)}}(a\sqrt{f(L)}+1)\right].$$

Die Anzahl der Lehnwörter in diesem Text ist

$$\bar{D} = k - D = f(L) - D.$$

Um von der Anzahl der Lehnwörter zu ihrem Anteil $\bar{d}$ zu kommen, dividieren wir $\bar{D}$ durch den Umfang k = f(L) des Wortbestands:

$$\bar{d} = 1 - \frac{2}{a^2 f(L)} e^{-a\sqrt{f(L)}} (a\sqrt{f(L)} + 1) - \frac{2}{a^2 f(L)} e^{-a}(a + 1). \quad (5.2)$$

Das zweite und dritte Glied in (5.2) geht gegen Null, der ganze Ausdruck gegen 1. Das zweite Glied, das $e^{-a\sqrt{f(L)}}$ enthält, ist klein gegenüber dem dritten, so daß letzteres die Ordnung bestimmt. Wir erhalten damit den folgenden Zusammenhang zwischen d und f(L):

$$\bar{d} = 1 - \frac{1}{cf(L)} .$$

Was kann man nun über den Zusammenhang von Wortbestand und Textlänge aussagen? Er ist noch nicht bekannt und enthält offenbar einen versteckten Parameter oder mehrere für die strukturellen Besonderheiten des jeweiligen Textes. Alle bekannten empirisch angepaßten Funktionen mit beschränktem Gültigkeitsbereich [32] sowie auch die theoretischen, die vom Zipfschen Gesetz [33] ausgehen, sind Funktionen der Form $k = cL^{\gamma}$ mit $c \geq 1$ und $\gamma < 1$.

Wir können also eine Abhängigkeit $\bar{d} = 1 - \frac{1}{cL^{\gamma}}$ des Anteils der Lehnwörter von der Textlänge erwarten, zu der die von Herdan vorgeschlagene logarithmische eine Mindestschätzung darstellt.

5.2. VERÄNDERUNG DER QUELLENAKTIVITÄT IN DER ZEIT

Wir betrachten jetzt den allgemeineren Fall: Mehrere Entlehnungsquellen sind wirksam und verändern ihre Aktivität im Laufe der Zeit. In den Tab. 24 und 25 ist die Verteilung der Entlehnungen im Deutschen nach den Wörterbüchern Kaed (die ersten 1000 Wörter) und And (1600 Wörter) dargestellt. Es sind die folgenden Schichten (Intervalle) unterschieden: Althochdeutsch (600- 1100), Mittelhochdeutsch (1100 - 1500), XVI. Jh., XVII.-XVIII. Jh., und für And zusätzlich XIX.-XX. Jh.. Man kann vergröbernd annehmen, daß die Aktivität der Entlehnungsquellen innerhalb der Zeitintervalle konstant ist und sich nur an den Grenzen ändert.

In Tab. 24 und 25 (und auch in Tab. 27 und 28, die genauso aufgebaut sind) ist für jede Klasse von 100 aufeinanderfolgenden Wörtern die Anzahl der Entlehnungen im betrachteten Intervall angegeben.

Die zusammenfassende Tab. 26 gibt für die beiden Wörterbücher die Aktivität von fünf Quellen an: 1) Neubildungen des Hochdeutschen, 2) Latein, 3) Französisch, 4) Niederdeutsche Dialekte, 5) übrige Sprachen. Es sei daran erinnert, daß unter der Aktivität einer Quelle y_j während (t_1,t_2) der Anteil p_j der Entlehnungen aus dieser Quelle an allen Neubildungen in (t_1,t_2) (der Schicht $\nu(t_1,t_2)$) verstanden wird. Die Aktivitäten der Quellen weichen nur geringfügig voneinander ab, besonders wenn man die große Unterschiedlichkeit der Wörterbücher und auch ihrer Veränderungsrate berücksichtigt. In beiden Fällen lassen sich dieselben, klaren Tendenzen feststellen: allmähliche Zunahme des französischen Einflusses zum XVII.-XVIII. Jh. (und nachfolgende Abnahme im XIX.-XX. Jh.), die meisten lateinischen Entlehnungen im XVI. Jh. (offenbar im Zusammenhang mit der Reformation und der Bildung einer eigenen deutschen politischen, juristischen und religiösen Terminologie), Abnahme der Entlehnungen aus niederdeutschen Dialekten in Zusammenhang mit der Bildung der literarischen Norm des Neuhochdeutschen, und in der jüngsten Zeit zunehmende Zahl von Sprachen, die auf das Deutsche wirken (s. Spalte „Sonst." in Tab. 26).

Am Material des Deutschen überprüfen wir die Hypothese der gleichmäßigen Verteilung der Entlehnungen in ihrer Schicht, die es uns nämlich auch ermöglicht, die Ergebnisse für die Wörter eines Anfangsstücks eines Häufigkeitswörterbuchs auf die ganze Sprache zu verallgemeinern.

Wir müssen zeigen, daß die beiden Variablen Klassenrang i und Anteil p_j der Wörter aus Quelle y_j für das gegebene Intervall (t_1,t_2) voneinander unabhängig sind.

Tabelle 24. Deutsch. Verteilung der Lehnwörter unter den ersten tausend Wörtern von Kaed

Rang	500-1100	Hochdeutsch	Latein	Französisch	Niederdeutsch	1100-1500	Hochdeutsch	Latein	Französisch	Niederdeutsch	Sonst.	1500-1600	Hochdeutsch	Latein	Französisch	Sonst.	1600-1850	Hochdeutsch	Latein	Französisch	Sonst.
1 (1-100)	12	12	0	0	0	3	2	0	0	1	0	3	3	0	0	0	0	0	0	0	0
2(101-100)	16	15	1	0	0	11	11	0	0	0	0	1	1	0	0	0	0	0	0	0	0
3(201-300)	20	17	2	0	1	14	10	4	0	0	0	2	2	0	0	0	1	1	0	0	0
4(301-400)	19	17	2	0	0	19	14	3	0	1	1	3	2	1	0	0	5	2	1	2	0
5(401-500)	18	18	0	0	0	24	18	4	1	1	0	6	5	0	1	0	7	3	0	3	1
6(501-600)	17	15	0	1	1	29	22	2	2	1	2	9	7	1	1	0	7	6	0	1	0
7(601-700)	17	13	4	0	0	21	15	3	2	0	1	8	3	3	0	2	16	6	4	5	1
8(701-800)	21	18	2	1	0	19	14	4	1	0	0	10	5	2	2	1	12	8	1	2	1
9(801-900)	17	15	2	0	0	20	13	3	2	2	0	9	5	3	1	0	19	9	1	7	2
10(901-1000)	18	12	6	0	0	17	13	1	1	0	2	9	5	3	0	1	21	14	5	1	1
Zusammen	175	152	19	2	2	177	132	24	9	6	6	60	38	13	5	4	88	49	12	21	6

Tabelle 25. Deutsch. Verteilung der Lehnwörter unter den ersten 1600 Wörtern von And

Rang	500-1100	Hochdeutsch	Latein	Französisch	Niederdeutsch	Sonst.	1100-1500	Hochdeutsch	Latein	Französisch	Niederdeutsch	Sonst.	1500-1600	Hochdeutsch	Latein	Französisch	Niederdeutsch	Sonst.	1600-1800	Hochdeutsch	Latein	Französisch	Niederdeutsch	Sonst.	1800-2000	Hochdeutsch	Latein	Französisch	Sonst.
1 (1-100)	11	10	1	0	0	0	10	9	0	0	1	0	3	2	0	1	0	0	1	0	0	1	0	0	6	6	0	0	0
2 (101-200)	18	17	0	0	1	0	13	8	1	1	1	2	5	3	1	0	0	1	8	6	1	0	0	1	7	5	0	0	2
3 (201-300)	10	8	1	0	1	0	21	18	2	0	0	1	7	4	3	0	0	0	15	11	2	2	0	0	2	0	0	1	1
4 (301-400)	8	7	1	0	0	0	15	8	5	1	0	1	6	1	5	0	0	0	10	3	3	4	0	0	13	10	0	1	2
5 (401-500)	13	13	0	0	0	0	16	12	3	1	0	0	11	9	1	1	0	0	14	8	3	3	0	0	9	7	0	1	1
6 (501-600)	14	12	1	0	0	1	20	14	4	1	1	0	9	8	0	0	0	1	14	4	2	7	0	1	13	9	1	2	1
7 (601-700)	15	14	1	0	0	0	15	11	3	1	0	0	10	8	2	0	0	0	15	5	2	8	0	0	10	4	1	3	2
8 (701-800)	11	10	1	0	0	0	16	12	2	2	0	0	8	6	1	1	0	0	15	6	4	5	0	0	8	3	0	3	2
9 (801-900)	12	11	1	0	0	0	10	6	1	1	2	0	12	8	4	0	0	0	15	6	3	5	0	1	14	8	0	2	4
10 (901-1000)	7	6	1	0	0	0	17	13	2	0	2	0	15	11	4	0	0	0	21	18	1	2	0	0	18	15	2	0	1
11 (1001-1100)	10	9	0	1	0	0	14	12	1	1	0	0	3	0	1	2	0	0	22	12	7	3	0	0	18	15	0	2	1
12 (1101-1200)	10	9	1	0	0	0	17	13	2	1	1	0	4	1	1	0	1	1	14	11	1	3	0	0	18	11	1	2	4
13 (1201-1300)	11	8	3	0	0	0	15	11	4	0	0	0	3	0	3	0	0	0	11	1	4	6	0	0	15	10	0	2	3
14 (1301-1400)	16	14	2	0	0	0	9	6	2	0	0	1	15	8	5	1	0	1	9	2	2	4	1	0	16	13	0	1	3
15 (1401-1500)	13	13	0	0	0	0	10	8	1	0	0	1	13	12	1	0	0	0	18	14	0	2	1	1	20	20	0	0	0
16 (1501-1600)	17	17	0	0	0	0	11	9	1	0	0	1	6	0	6	0	0	0	12	9	3	0	0	0	19	10	0	4	5
Zusammen	196	178	14	1	2	1	229	170	34	10	8	7	130	81	38	6	1	4	241	119	38	55	2	4	206	145	5	24	32

Tabelle 26. Veränderung der Aktivität der wichtigsten Entlehnungsquellen des Deutschen nach Kaed und And

Zeitintervall	Hochdeutsch		Latein		Französisch		Niederdeutsch		Sonst.	
	Kaed	And	Kaed	And	Kaed	And	Kaed	And	Kaed	And
Althochdeutsch (500-1100)	0,87	0,91	0,11	0,07	0,01	0,005	0,01	0,01	-	0,005
Mittelhochdeutsch (1100-1500)	0,75	0,74	0,14	0,15	0,05	0,045	0,03	0,035	0,03	0,03
XVI. Jh.	0,63	0,625	0,22	0,29	0,08	0,045	-	0,01	0,07*	0,03
XVII.-XVIII. Jh.	0,56	0,555	0,13	0,165	0,24	0,25	-	0,01	0,07*	0,02
XIX.-XX. Jh.	-	0,70	-	0,025	-	0,12	-	-	-	0,155

*einschließlich der niederdeutschen Entlehnungen

Wir stellen die Verteilung der lateinischen Entlehnungen in der mittelhochdeutschen Periode (nach Kaed) in Tab. 26a dar, wobei wir die Zahlen für benachbarte Klassen so zusammenfassen, daß jede der neuen Klassen mindestens fünf lateinische Lehnwörter enthält:

Tabelle 26a. Verteilung der lateinischen Lehnwörter der mittelhochdeutschen Periode (nach Kaed)

	Klasse				Zus.
	1-4	5-6	7-8	9-10	
Lateinische Lehnwörter	7(6,4)	6(7,2)	6(5,4)	5(5)	24(13,6%)
Nichtlateinische Elemente	40(40,6)	47(45,8)	34(34,6)	32(32)	153(86,4%)
Wörter der mhd. Schicht insgesamt	47	53	40	37	177

Wenn die Variablen unabhängig sind, dann darf der Anteil lateinischer Wörter in keiner der neuen Klassen wesentlich von 13,6%

($\frac{24}{177} \cdot 100\% = 13{,}6\%$) und der Anteil der nichtlateinischen von 86,4% abweichen. Die unter dieser Annahme errechnete Anzahl lateinischer und nichtlateinischer Elemente ist in Klammern angegeben.

Der χ^2-Test mit $(2-1)\cdot(4-1) = 3$ Freiheitsgraden ergibt $\chi^2 \approx 0{,}48 < \chi^2_{0,1} = 6{,}25$, so daß die Hypothese über einen Zusammenhang von Klassenrang und Anteil lateinischer Lehnwörter abgelehnt werden kann.

Ebenso untersuchen wir die lateinischen Lehnwörter aus dieser Periode in And (Tab. 26b).

Tabelle 26b. Verteilung der lateinischen Lehnwörter der mittelhochdeutschen Periode (nach And)

	Klasse				Zus.
	1 - 4	5 - 8	9 - 12	13 - 16	
Lateinische Lehnwörter	8(8,7)	12(9,9)	6(7,7)	8(6,8)	32(14,8%)
Nichtlateinische Elemente	51(50,3)	55(58,1)	52(50,3)	38(39,2)	196(85,2%)
Wörter der mhd. Schicht insgesamt	59	67	58	46	230

Bei drei Freiheitsgraden ist $\chi^2_{0,1} = 6{,}25 > \chi^2 = 1{,}29$, d.h. für And kann ebenfalls angenommen werden, daß die Entlehnungen gleichmäßig über die Schicht verteilt sind.

In Tab. 27 und 28, die Tab. 24 und 25 entsprechen, ist die Verteilung der Lehnwörter nach Schichten im Französischen (Goug; Tab. 27) und im Englischen (Dew; Tab. 28) angegeben. Tab. 29 (entspricht Tab. 26) gibt die Veränderung der Aktivität der verschiedenen Entlehnungsquellen des Französischen an.

Schließlich führen wir noch zur Illustration der Verteilung der Lehnwörter ohne Schichtenunterscheidung vier Histogramme (Abb. 27 bis 30) an. Abb. 27 zeigt die Verteilung der romanischen Lexik im Neuenglischen ohne Berücksichtigung der Entlehnungszeit nach Lindsey [21] und Roberts [22]. Lindseys Zahlen beziehen sich auf die ersten 5000 Wörter aus dem Wörterbuch von Thorndike, die

Zahlen von Roberts auf die 10000 Wörter aus dem Häufigkeitswörterbuch von Horn. Für Klassen von je 1000 Wörtern ist der Anteil der romanischen (d.h. praktisch der lateinischen und französischen) Lehnwörter berechnet (x: Klassenrang; y: Anteil der Lehnwörter).

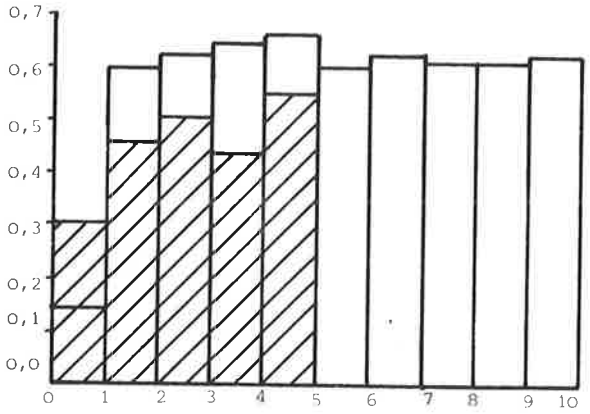

Abb. 27. Verteilung der romanischen Lehnwörter unter den ersten 5000 Wörtern des englischen Häufigkeitswörterbuchs von Thorndike (schraffiert; nach Lindsey) und unter den ersten 10000 Wörtern des Häufigkeitswörterbuchs von Horn (nach Roberts)

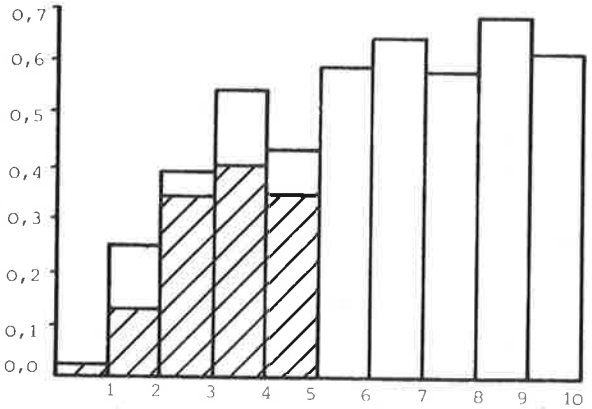

Abb. 28. Verteilung der romanischen Lehnwörter unter den ersten 500 Wörtern des englischen Wörterbuchs von Thorndike (nach Lindsey, schraffiert) und unter den ersten 1000 Wörtern von Dew

In Abb. 28 werden die Daten von Lindsey zur Verteilung der romanischen Lehnwörter unter den ersten 500 Wörtern des Wörterbuchs von Thorndike unseren Ergebnissen zur Verteilung der romanischen Lehnwörter unter den ersten 1000 Wörtern von Dew gegenübergestellt.

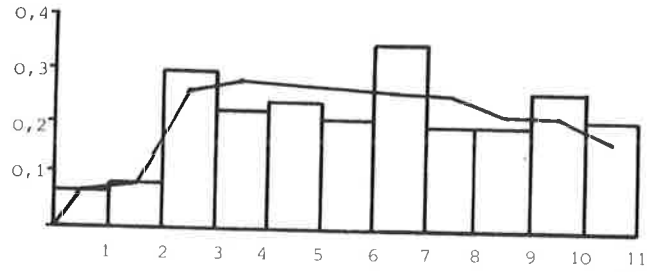

Abb. 29. Verteilung der slavischen Lehnwörter im Rumänischen des XVI. Jh. nach Rum-XVI (Histogramm). Der Polygonzug zeigt die Verteilung der mittleren Häufigkeit slavischer Lehnwörter in den drei Wörterbüchern Rum-XVI, Rum-XVII und Rum-XVIII

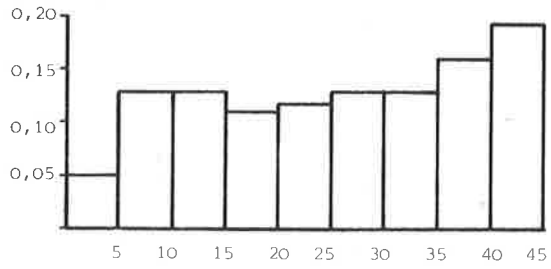

Abb. 30. Verteilung der Entlehnungen unter den ersten 4500 Wörtern aus Čech (in 500-er Klassen)

Tabelle 27. Französisch. Verteilung der Lehnwörter unter den ersten 1000 Wörtern von Goug

Rang	700-1200	Französisch	Latein	Germanisch	1200-1300	Französisch	Latein	Sonst.	1300-1400	Französisch	Latein	Sonst.	1400-1500	Französisch	Latein	Italienisch	1500-1700	Französisch	Latein	Italienisch	Sonst.
1 (1-100)	14	14	0	0	3	3	0	0	0	0	0	0	2	2	0	0	2	2	0	0	0
2(101-200)	16	10	4	1	3	3	0	0	3	3	0	0	1	1	0	0	6	6	0	0	0
3(201-300)	16	9	5	2	4	3	1	0	0	0	0	0	1	1	0	0	6	6	0	0	0
4(301-400)	13	8	3	2	5	2	3	0	6	2	4	0	4	2	2	0	11	8	2	0	1
5(401-500)	21	10	9	2	10	4	5	1	4	3	1	0	3	1	2	0	7	7	0	0	0
6(501-600)	15	8	5	2	8	4	4	0	10	5	5	0	6	3	2	1	22	13	3	5	1
7(601-700)	23	12	10	1	5	3	2	0	10	4	4	2	3	1	1	1	17	8	4	5	0
8(701-800)	24	14	8	2	7	6	1	0	8	6	2	0	6	4	2	0	19	15	2	1	1
9(801-900)	20	13	4	3	7	1	6	0	9	6	3	0	6	3	3	0	14	9	2	2	1
10(901-1000)	24	12	8	2	5	2	2	1	7	2	5	0	7	5	2	0	13	9	2	2	0
Zusammen	186	113*	56	17	57	31	24	2	57	31	24	2	38	21	15	2	117	83	15	15	4

*enthält drei griechische Lehnwörter

Tabelle 28. Englisch. Verteilung der Lehnwörter unter den ersten 1000 Wörtern von Dew

Rang	1100-1400 Mittelengl.	Englisch	Latein	Französisch	Skandinavisch	1400-1500	Englisch	Latein	Französisch	1500-1600	Englisch	Latein	Französisch	Sonst.
1 (1-100)	8	5	0	2	1	0	0	0	0	0	0	0	0	0
2(101-200)	26	4	1	19	2	2	1	0	1	1	0	1	0	0
3(201-300)	39	8	1	28	2	4	1	2	1	2	1	0	1	0
4(301-400)	52	5	6	40	1	2	0	0	2	3	0	3	0	0
5(401-500)	44	9	6	28	1	3	1	0	2	5	0	3	1	1
6(501-600)	54	6	4	42	2	4	0	0	4	8	2	3	3	0
7(601-700)	57	4	8	44	1	8	0	1	2	7	0	3	4	0
8(701-800)	45	6	2	36	1	5	0	3	2	9	1	3	5	0
9(801-900)	49	0	2	45	2	7	3	1	3	9	1	7	1	0
10(901-1000)	46	4	5	37	0	4	1	2	1	11	1	5	5	0
Zusammen	420	51	35	321	13	34	7	9	18	55	6	28	20	1

Tabelle 29. Änderung der Aktivität der Entlehnungsquellen des Französischen

Intervall	Franz.	Latein	Germanisch	Italienisch	Sonst.
700-1200	0,59	0,30	0,09	-	0,02
1200-1300	0,55	0,42	-	-	0,03
1300-1400	0,55	0,42	-	-	0,03
1400-1500	0,55	0,40	-	0,05	-
1500-1600	0,67	0,13	-	0,17	0,03

Aus den Histogrammen geht klar hervor, daß der Anteil der Entlehnungen mit wachsendem Rang stark zunimmt.

Abb. 29 gibt die Verteilung der slavischen Lexik im Rumänischen an. Das Histogramm zeigt die Verteilung der slavischen Lexik in Rum-XVI, der Polygonzug gibt die entsprechenden Mittelwerte für die drei rumänischen Wörterbücher Rum-XVI, Rum-XVII und Rum-XVIII wieder. Die Wörter sind in 100-er Klassen aufgeteilt. Die konvexe Kurve zeigt an, daß die Periode der höchsten Aktivität der slavischen Quelle bereits vorüber ist.

Abb. 30 schließlich gibt die Verteilung der Lehnwörter für das Tschechische an (ohne Unterscheidung nach Entlehnungsquelle oder -zeit; ohne genuin tschechische Neubildungen). Das Anwachsen des Anteils der Lehnwörter mit dem Rang ist deutlich ablesbar.

5.3. DIE FORSCHUNGSTRADITION ZUR ENTLEHNUNG UND UNSER ANSATZ

Die Entlehnungsforschung hat in der Sprachwissenschaft eine bedeutende Tradition, in der schon früh ein quantitativer Ansatz entwickelt wurde, früher vielleicht als in allen anderen sprachwissenschaftlichen Disziplinen. Gewöhnlich werden in solchen Arbeiten alle Lehnwörter in einem „großen" Wörterbuch[4] nach Entlehnungsquellen und manchmal auch nach Alter klassifiziert.

Es werden dann für eine Periode die Entlehnungen aus verschiedenen Quellen zahlenmäßig verglichen, und für mehrere Perioden die Entlehnungen aus einer Quelle. Dagegen ist selbstverständlich nichts einzuwenden, und in einigen Fällen bereitet die Interpretation der erhaltenen Zahlen auch keine Schwierigkeiten: die Proportion der Entlehnungen mit verschiedener Herkunft oder die von alten zu neuen Entlehnungen sagt etwas über die Gegenwartssprache aus (natürlich nur, wenn man annimmt, daß ein normales Definitionswörterbuch und die daraus gewonnenen Zahlen für die Gegenwartssprache repräsentativ sind), bzw. über die Sprache des Autors, auf dessen Werke man sich bezieht.

Aber man muß sich entschieden gegen eine direkte historische Interpretation so gewonnener Daten wenden: aus dem Vergleich der Anzahl von Lehnwörtern aus einer Sprache auf der Grundlage eines Wörterbuchs der Gegenwartssprache darf man keine Schlüsse auf die Veränderung der Quellenaktivität ziehen.[5]

Nehmen wir das Englische, dessen Entlehnungen, speziell aus dem Französischen, quantitativ gut erforscht sind. In der klassischen Arbeit von O. Jespersen [35] ist die Verteilung der französischen Lehnwörter nach ihrem Alter in einzelnen Bänden des Oxford English Dictionary angegeben. Wir können sie mit den entsprechenden Angaben von Baugh vergleichen, der aber bereits alle Bände untersucht hat [36] (s. Tab. 30). Wir führen unten die Zahlen von Jespersen (1) und Baugh (2) gegenüber unseren (3) an, die wir aus Dew gewonnen haben (s. Tab. 28), mit der Anzahl der Lehnwörter von 1100 bis 1600 = 100%.

Die Zahlen von Jespersen und Baugh stimmen recht gut überein; sie lassen sich vielleicht mit einer gleichbleibenden Aktivität des Französischen über fünf Jahrhunderte (1100 bis 1600) erklären. Doch die Zahlen nach dem rangierten Wörterbuch Dew geben ein völlig anderes Bild, sie zeigen, daß die Aktivität des Französischen in der mittelenglischen Periode wesentlich höher war als in den folgenden zwei Jahrhunderten.

Tabelle 30. Dynamik der romanischen Lehnwörter im Englischen nach Definitions- (1 u. 2) und nach Häufigkeitswörterbuch (3)

Intervall	(1)	(2)	(3)
1100-1400	61,3	59,5	90
1401-1500	17,7	21,3	4,9
1501-1600	21,0	20,2	5,1

Die Ursache dieser Abweichungen ist natürlich trivial: ein gewöhnliches Wörterbuch enthält keine Information darüber, wieviele Lehnwörter aus der betrachteten Periode bereits wieder geschwunden sind.

Die traditionelle Methode liefert keine Ergebnisse, die für eine kulturhistorische Interpretation, bei der es um die Intensität der Kontakte in der betrachteten Periode geht, oder für eine linguistische Interpretation, die in den Entlehnungen eine Ursache struktureller Veränderung der Sprache sieht, geeignet wären.

6. PRODUKTIVITÄT UND PRODUKTIVITÄTSMESSUNG

6.1. RANG, ALTER UND PRODUKTIVITÄT

Im letzten Kapitel versuchen wir nun, das entwickelte Instrumentarium für die Messung der Produktivität verschiedener Ausdrucksmittel der Gegenwartssprache einzusetzen. Die Problemgeschichte der Produktivität und ihrer Messung ist lang und kompliziert (s. [39-42]), was wohl damit zusammenhängt, daß es kein sprachwissenschaftliches Begriffssystem gibt, das einen gegenwärtigen Sprachzustand in organischem Zusammenhang mit seiner Vergangenheit faßt. Es schien uns daher naheliegend, für die formale Definition der Produktivität und für die Entwicklung von Verfahren zu ihrer Messung unsere Theorie anzuwenden, die den Zusammenhang der beiden Parameter Alter und Häufigkeit eines Wortes beschreibt, wobei der erste ein rein historischer, diachroner, und der zweite ein synchroner ist.

In Kap. 4 und 5 wurde ein Begriff analysiert, der der Produktivität sehr ähnlich ist, nämlich die Aktivität p_j einer Entlehnungsquelle y_j. Mit Hilfe dieses Begriffs haben wir das „Gewicht" einer Klasse von Wörtern gemeinsamer Herkunft unter den Neubildungen in einem Zeitintervall (t_1,t_2) definiert. Bei allen Unterschieden in der Begrifflichkeit und der Betrachtungsweise wird unter Produktivität eben dieses „Gewicht" verstanden (z.B. die Klasse aller Wörter, die ein bestimmtes Suffix haben oder nach dem Muster eines bestimmten Wortes gebildet sind, u.ä.),

aber nicht bezogen auf ein Zeitintervall, sondern auf einen bestimmten **Zeitpunkt**. D.h. unter der Produktivität z.B. eines Suffixes der Gegenwartssprache wird das „Gewicht" der entsprechenden Klasse von Wörtern zum Zeitpunkt t = 0 verstanden.

Offenbar verhält sich das „Gewicht" für ein Zeitintervall zum Gewicht zu einem Zeitpunkt wie z.B. die Durchschnittsgeschwindigkeit in einem Zeitintervall zur Momentangeschwindigkeit in der Mechanik, was den Gedanken nahelegt, die Produktivität mit Hilfe des Grenzwertbegriffs zu formalisieren.

Die Begriffe „Aktivität einer Entlehnungsquelle" und „Produktivität einer Klasse" sind, obwohl sehr ähnlich, keineswegs identisch. Wir führen deshalb die Produktivität als neuen Begriff ein, wobei einige Wiederholungen unvermeidbar sind. Die Kapitel 4 und 5 unterscheiden sich vom sechsten auch noch in der Zielsetzung, da hier nicht das Verhältnis von Zeit und Wortrang, sondern die Maßzahl einer Klasse im Mittelpunkt steht, wobei diese Maßzahl aus der Verteilung der Elemente (Wörter) dieser Klasse nach Alter und Rang bestimmt wird. Daher befaßt sich der folgende Abschnitt mit den Problemen der Klassifikation von Wörtern aus einem Lexikon V.

6.2. DIE KLASSIFIKATION VON LEXIKONEINHEITEN

Wir führen zunächst die übliche Terminologie ein. Anstelle von produktiven (bzw. nichtproduktiven) Affixen, Flexionstypen oder Wortbildungsmustern sprechen wir immer von einer Klasse von Wörtern mit einer Eigenschaft P.

$$X_P = \{a \mid a \in V,\ P(a)\}.$$

Diese Eigenschaft P kann für das Vorhandensein eines bestimmten Affixes (bestimmter Affixe), für die Zugehörigkeit zu einem Wortbildungsmuster oder zu einer Flexionsklasse u.ä. stehen.

Wir gehen wie schon oben (s. 1.1) davon aus, daß das Lexikon V eine nach Häufigkeit geordnete, unendliche Menge von Wörtern ist

(oder von Wortformen, wenn die Bedeutung von P das erfordert).

Betrachten wir jetzt die Menge $T = \{X_p\}$ von Klassen in V, mit anderen Worten, eine Teilmenge von 2^V.

Eine solche Menge $\{X_p\}$ von Klassen soll eine Taxonomie T auf dem Lexikon V heißen.

Obwohl in der Sprachwissenschaft der Klassifikationsproblematik immer große Aufmerksamkeit gewidmet wurde, fehlt noch immer eine klare Vorstellung davon, nach welchen Eigenschaften Wörter sprachwissenschaftlich sinnvoll klassifiziert werden können, und welche Klassifikationsprinzipien möglich sind.

So kann man z.B. sinnvoll von der Klasse aller Wörter mit einem bestimmten Suffix reden, jedoch nicht von der Klasse aller Wörter, die in den geraden Spalten der letzten Ausgabe des Wörterbuchs von S. Ožegov stehen. Im Russischen ist offenbar die Klasse der Wörter mit dem Akzent auf dem Laut [o] von keinerlei Interesse, während im Tschechischen auf der lexikalischen Ebene die Klasse der Wörter mit langem [ō] sinnvoll gebildet werden kann; praktisch alle Wörter dieser Klasse sind Lehnwörter. Unten stellen wir einige Beschränkungen für eine Taxonomie T zusammen, die so konstruiert ist, daß jede ihrer Klassen entweder produktiv oder nichtproduktiv ist.

ANNAHME 1°. Jede Klasse $X \in T$ ist entweder produktiv ($X \in T^+$) oder nichtproduktiv ($X \in T^-$) mit $T = T^+ \cup T^-$ und $T^+ \cap T^- = \emptyset$.

Es ist zweckmäßig, den Begriff der Produktivität zunächst nicht für alle Klassen der Taxonomie T zu definieren, sondern für eine kleinere Menge (/Kollektion/) $B \subset T$.

ANNAHME 2°. Es gibt eine Menge (/Kollektion/) von Mengen $B \subset T$, der Basis der Taxonomie, derart, daß

1) für beliebiges $a \in V$ ein $X \in B$ existiert mit $a \in X$;
2) für beliebiges $X \in B$, $Y \in B$ $X \cap Y = \emptyset$;
3) jede Klasse $Z \in T$ als Vereinigung von Klassen X_i aus B, $Z = \bigcup_i X_i$, $X_i \in B$ dargestellt werden kann; umgekehrt: jede Vereinigung $\bigcup_i X_i$ von Klassen der Basis eine Klasse der Taxonomie T ist;

4) die leere Menge zur Basis gehört.

Die Eigenschaften der Basis werden durch die folgenden zwei Behauptungen geklärt.

BEHAUPTUNG 1. Für beliebige $X \in T$ und $Y \in B$ mit $X \cap Y \neq \emptyset$ gilt $X \supset Y$.

BEWEIS. Wir stellen die Klasse $X \in T$ als Vereinigung von Basisklassen dar: $X = \bigcup_i Z_i$. Wegen $X \cap Y \neq \emptyset$ gibt es ein a mit $a \in \bigcup_i Z_i$ und $a \in Y$. Alle Basisklassen sind aber nach Definition disjunkt, so daß ein i existieren muß mit $Y = Z_i$ d.h. $X \supset Y$.

BEHAUPTUNG 2. Für beliebige Klassen $X \in T$ und $Y \in T$ gehört ihr Durchschnitt zur Taxonomie: $X \cap Y \in T$.

BEWEIS. Entweder gilt $X \cap Y \neq \emptyset$, dann ist $X \cap Y \in B$ (nach Definition) und folglich $X \cap Y \in T$, oder der Durchschnitt $X \cap Y$ enthält mindestens ein Element a. Wegen Eigenschaft (1) der Basisklassen gibt es dann eine Klasse $X' \in B$ mit $a \in X'$, wegen Eigenschaft (2) gibt es genau eine solche Klasse. Dann folgt aus Behauptung 1, daß $X' \subset X$ und $X' \subset T$, d.h. $X' \subset X \cap Y$.

Betrachten wir alle $a_i \in X \cap Y$ und bilden die Vereinigung aller X'_i derart, daß $a_i \in X'_i$. Nach Definition ist die Vereinigung solcher Basisklassen X'_i eine Klasse der Taxonomie: $\bigcup_i X'_i \in T$.

Mit den Behauptungen 1 und 2 kann man sich die Klassen der Taxonomie als aus einer Anzahl von „Zellen", den Basisklassen, zusammengesetzt denken. Die Taxonomie über der Basis hat die wichtige Eigenschaft der Abgeschlossenheit: Die Vereinigung (trivialerweise) und der Durchschnitt (Behauptung 2) beliebiger Klassen der Taxonomie sind wieder Klassen von T.

Es gibt kein Kriterium, mit dem sich a priori entscheiden ließe, ob eine Klasse zur Basis gehört oder nicht, ob man sie nämlich z.B. als Vereinigung von produktiven und nichtproduktiven Klassen auffassen kann (manchmal gibt es ein Kriterium für die Zugehörigkeit zur Basis, seine Anwendbarkeit setzt jedoch die Messung der Produktivität voraus).

Andererseits kann die Basis nach verschiedenen und unterschiedlich guten Methoden gewählt sein (vgl. verschiedene Methoden der Klassifikation von Wörtern nach Deklinations- und Konjugationstypen, Wortbildungsmustern u.ä.). Gute Methoden zur Bildung der Basis sichern eine gewisse innere Homogenität der Klassen, in gewissem Sinne also ein gemeinsames „Schicksal" aller Elemente der Klasse (speziell gemeinsame zeitliche Erstreckung aller Wörter einer Basisklasse). Im Rahmen dieser Arbeit genügt eine informelle Fassung der Homogenitätsforderung.

Nach der Aussonderung der Basis wollen wir nun für jede Klasse $X_p \in B$ das Maß ihrer Produktivität bestimmen, d.h. eine Vorschrift angeben, nach der man jeder Klasse eine Zahl π_p (oder $\pi(X_p)$) zuordnen kann, die entweder positiv oder Null ist. Im zweiten Fall nennen wir die Klasse nichtproduktiv.

Weiter verallgemeinern wir das vorgeschlagene Maß auf die gesamte Taxonomie T, wozu wir beweisen, daß es die folgende Eigenschaft hat:

$$\pi(X \cup Y) = \pi(X) + \pi(Y),$$

wenn $X \cap Y = \emptyset$, d.h. die Produktivität disjunkter Mengen ist additiv, und gleichzeitig läßt sich die Produktivität einer beliebigen Klasse der Taxonomie T als Summe der Produktivitäten ihrer Basisklassen ausdrücken.

6.3. DIE WÖRTER DER KLASSE X_p UNTER DEN NEUBILDUNGEN

Im Folgenden benötigen wir den Begriff des Alters t(a) eines Wortes a, mit dessen Hilfe wir Neubildungen von alten Wörtern unterscheiden wollen. Unter dem Alter t(a) eines Wortes a verstehen wir wie oben das Zeitintervall vom Zeitpunkt t seiner Entstehung bis zum gegenwärtigen Zeitpunkt t_o. t_o, d.h. den Zeitpunkt der Erstellung des Wörterbuchs V, setzen wir der Einfachheit halber

gleich Null. Dann läßt sich die Menge N_t von Neubildungen in V wie folgt definieren:

$$N_t = \{a | t \geq t(a) > 0,\ a \in V\}.$$

Der Begriff „Neubildung" ist natürlich relativ, er hängt vom Parameter t ab.

Jetzt können wir eine erste Definition des Begriffs „produktive Klasse" versuchen. Produktiv kann eine Klasse X_p der Taxonomie T sinnvollerweise dann genannt werden, wenn sie Neubildungen für noch so kleines t enthält. Daher:

DEFINITION 1. Eine Klasse X_p heiße produktiv (im weiten Sinn) genau dann, wenn für beliebiges t gilt: $N_t \cap X_p \neq \emptyset$.

Sind die Bedingungen von Definition 1 nicht erfüllt, wird die Klasse X_p nichtproduktiv genannt. Wenn X_p also eine nichtproduktive Klasse ist, dann gibt es ein t_1, so daß für beliebige $t \leq t_1$ gilt $N_t \cap X_p = \emptyset$, d.h. von den Wörtern, die nach t_1 entstanden sind, gehört keines zur Klasse X_p.

Der reale Nutzen von Definition 1 ist nicht sehr groß; sie liefert kein praktisches Prüfverfahren für die Produktivität einer konkreten Klasse X_p, weil man ja die gasamte Menge N_t wegen ihrer Unendlichkeit nicht einmal für ein einziges konkretes t betrachten kann (s. 2.3). Außerdem gibt sie kein numerisch bestimmbares Maß, da sie noch eine weitere, versteckte Unzulänglichkeit enthält (s.u.).

Definition 1 ist nicht an sich interessant, sondern nur als Ausgangspunkt bei der Aufstellung einer befriedigerenden Definition. Es sei noch angemerkt, daß Definition 1 natürlich auf alle Klassen der Taxonomie und nicht nur auf die Basisklassen anwendbar ist. Außerdem ist klar, daß die Vereinigung von Klassen produktiv ist, wenn eine der Klassen nach Def. 1 produktiv ist. Suchen wir nun ein Produktivitätsmaß für die Klasse X_p unter der Annahme, daß sie die Bedingungen von Def. 1 erfüllt.

Offenbar wäre es ein Fortschritt, wenn wir angeben könnten, wieviele Wörter von N_t für gegebenes t zu X_p gehören, wenn wir

also das „Gewicht" der Teilmenge X_p in N_t messen könnten. Für eine unendliche Menge N_t geht das jedoch nicht direkt; vielleicht kann man aber anstelle von N_t ein repräsentatives, endliches Fragment der Gesamtmenge der Neubildungen betrachten.

Wir zeigen, daß man mit einigen Annahmen zur Verteilung der Wörter einer gegebenen Klasse X_p in einem nach Häufigkeit geordnetem Wörterbuch V die Produktivität dieser Klasse im Prinzip mit einem beliebigen, endlichen (jedoch nicht zu kleinen) Fragment $G \subset V$ bestimmen kann, auch wenn ein solches Vorgehen nicht immer zweckmäßig ist.

Sei G_i ein endliches Fragment des Wörterbuchs V, wobei der Index i auf unten noch zu präzisierende Weise die Häufigkeit von G_i angibt. G_i enthalte n Wörter. Die Menge der t alten Wörter von G_i heiße $N_{i,t}$, die Anzahl dieser Wörter $N(i,t)$ oder $|N_{i,t}|$.

Nur der Teil $\alpha(i,t,p)$ der Neubildungen gehört zu der uns interessierenden Klasse X_p,

$$\alpha(i,t,p) = \frac{|N_{i,t} \cap X_p|}{|N_{i,t}|} \quad . \tag{6.1}$$

Aus (6.1) ist direkt ersichtlich, daß α davon abhängt, wie das Fragment (i) gewählt wird, was man unter Neubildungen (t) versteht und welches Merkmal (P) die Klasse X_p definiert.

Wir machen die folgende

ANNAHME 3° Der Anteil α der Neubildungen aus der Klasse X_p im Fragment G_i hängt nur von t und den Eigenschaften der Klasse X_p ab, nicht aber von der Häufigkeit der Wörter von G_i, d.h. $\alpha = \alpha(t,p)$.

Wenn also die Bildungen einer gegebenen produktiven Klasse ungleichmäßig über das Häufigkeitswörterbuch verteilt sind, so läßt sich das vollständig durch die ungleichmäßige Verteilung der Neubildungen erklären. Anders ausgedrückt: die Neubildungen N_t können über die Fragmente G_i ungleichmäßig verteilt sein (zur Form der Verteilung s. unten), der Anteil α der Bildungen einer gegebenen Klasse ist konstant. Diese Annahme entspricht völlig derjenigen

zur Verteilung der Lehnwörter im Lexikon (s. 4.2), der einzige Unterschied besteht in der Definition der „Schicht" der Neubildungen. Die empirische Überprüfung dieser Behauptung (für alle Fragmente von V) ist natürlich unmöglich; man kann nur zu zeigen versuchen, inwieweit die Verteilung der Wörter einer gegebenen Klasse X_p in einem realen Häufigkeitswörterbuch mit Annahme 3° übereinstimmt. Oben 5.2) haben wir eine solche Überprüfung unternommen, zwar nicht für Wortbildungsklassen, aber für lexikalische Klassen - der Lehnwörter aus einer bestimmten Quelle (Sprache). Die entsprechende Hypothese wurde statistisch getestet und angenommen. Ein Beweis für die Annahme kann auf diese Weise natürlich nicht erbracht werden. Hypothese 3° wird als die einfachste, eine bestimmte Homogenität des Lexikons voraussetzende Hypothese gewählt.

6.4. DEFINITION DER PRODUKTIVITÄT

Der nächste Schritt besteht darin, für eine gegebene Klasse X und einen gegebenen Wörterbuchausschnitt G_i die Folge $\alpha(t)$ für immer kleineres Zeitintervall t zu betrachten:

$$\alpha(t_1),\ \alpha(t_2),\ \alpha(t_3),\ldots, \text{ mit } t_1 > t_2 > t_3 > \ldots,$$

d.h. die Neubildungen als immer engere Klasse N_t für $t \to 0$ aufzufassen.

Natürlich kann man die Produktivität $\alpha(t)$ nicht als Wert zum Zeitpunkt $t = 0$ definieren, weil dann ja $t = 0$, $|N_{i,t}| = 0$ ist, und die Proportion (6.1) dann undefiniert ist. Man kann aber die Produktivität als Grenzwert der Folge $\alpha(t)$ für $t \to 0$ auffassen.[1)]

Zur mathematisch korrekten Definition der Produktivität brauchen wir noch

ANNAHME 4° Die Folge $\alpha(t)$ konvergiert für $t \to 0$ und beliebige Klasse X aus der Basis B der Taxonomie.

DEFINITION 2. Die Produktivität π_p der Klasse X_p (das Maß ihrer Produktivität) ist der Grenzwert der Folge $\alpha(t)$ für $t \to 0$,

$$\pi_p = \lim_{t \to 0} \alpha(t).$$

Die Klasse X_p ist produktiv, wenn $\pi_p \neq 0$, und nichtproduktiv, wenn $\pi_p = 0$.

Die Produktivität einer Klasse X_p kann an einem beliebigen Wörterbuchausschnitt G_i gemessen werden, weil aus Annahme 3^0 folgt, daß alle Glieder der anhand eines Wörterbuchausschnitts bestimmten Folge $\alpha(t)$ mit den entsprechenden Gliedern der Folge $\alpha(t)$ für alle anderen Ausschnitte des Wörterbuchs V identisch sind.

Im Zusammenhang mit dem Anteil der Wörter der Klasse X_p unter den Neubildungen $N_{i,t}$ sind wir nicht darauf eingegangen, wieviele dieser Neubildungen im betrachteten Wörterbuchausschnitt enthalten sind. Dabei müssen uns in der Hauptsache die Neubildungen in der Nachbarschaft von t = 0 interessieren; ihre Anzahl kann aber äußerst gering sein.

Wie wir aber wissen, ist die Verteilung der Neubildungen über das Lexikon V ungleichmäßig; deshalb ist es sinnvoll, einen solchen Lexikonausschnitt zu suchen, in dem die Neubildungen zahlreicher sind, und die Proportion α sehr genau bestimmt werden kann. Noch besser wäre es, wenn wir einen Ausschnitt fänden, in dem praktisch nur Neubildungen vorkommen. Dann könnten wir uns das komplizierte Sortieren der Wörter nach Alter ersparen.

Zunächst legen wir die Bedeutung des Begriffs „Ausschnitt des Lexikons V" genauer fest.

Unter einem Ausschnitt G_i des Lexikons V verstehen wir, wie oben (2.1), die i-te Klasse des Lexikons V, die die Wörter der Ränge (i-1)n+1 bis i·n enthält; den Index i nennen wir Rang der Klasse G_i. Wie in 2.3 gezeigt, beträgt die Anzahl der Neubildungen mit dem Alter t in G_i [2]:

$$\bar{N}_i(t) = n(1 - e^{-kt\sqrt{i}}). \qquad (6.2)$$

Analog ist die Anzahl der Wörter, die älter als t sind,

$$N(t) = ne^{-kt\sqrt{i}}. \qquad (6.3)$$

Offenbar könnten wir die Berechnung der Produktivität von X_p **stark** vereinfachen, wenn wir nicht eine beliebige Klasse G_i, sondern eine Klasse hohen Rangs i betrachteten. Für $i \to \infty$ und beliebiges t geht nämlich $e^{-kt\sqrt{i}} \to 0$ und

$$\lim_{i \to \infty} \bar{N}_i(t) = \lim_{i \to \infty} n(1 - e^{-kt\sqrt{i}}) = n = |G_i|.$$

D.h. ein hinreichend „peripherer" Lexikonausschnitt besteht praktisch nur aus Neubildungen, wobei es keine Rolle spielt, wie „Neubildung" genau definiert wird. Wenn wir für $|N_{i,t}|$ den Grenzwert in (6.1) einsetzen, erhalten wir also:

$$\lim_{i \to \infty} \alpha_i(t) = \frac{|G_i \cap X_p|}{|G_i|}.$$

Damit hängt im Grenzbereich (extremer Rang) der Anteil der Wörter allein von den Eigenschaften der Klasse X_p ab, und wir können die Produktivität folgendermaßen definieren:

DEFINITION 2'. Die Produktivität π_p der Klasse X_p ist der Grenzwert des Anteils α der Wörter der Klasse X_p am Fragment G_i für $i \to \infty$.

Wir zeigen nun, daß die Produktivität einer Vereinigung $Y = \bigcup_i X_i$ von Basisklassen durch die Produktivitäten dieser Basisklassen ausgedrückt werden kann:

$$\pi(Y) = \pi(\bigcup_i X_i) = \sum_i \pi(X_i).$$

Dazu reicht es, dies für die Vereinigung zweier Basisklassen X_1 und X_2 zu zeigen. Nach Definition ist

$$\alpha_{X_1 \cup X_2}(t) = \frac{|N_t \cap (X_1 \cup X_2)|}{|N_t|} = \frac{|(N_t \cap X_1) \cup (N_t \cap X_2)|}{|N_t|}.$$

Da X_1 und X_2 Basisklassen, also disjunkt sind, ist

$$\alpha_{X_1 \cup X_2}(t) = \frac{|N_t \cap X_1| + |N_t \cap X_2|}{|N_t|} = \alpha_{X_1}(t) + \alpha_{X_2}(t).$$

Durch Grenzwertbildung (für t - Def. 2, oder für i - Def. 2') erhalten wir

$$\pi(X_1 \cup X_2) = \pi(X_1) + \pi(X_2).$$

Damit ist sichergestellt, daß das Maß der Produktivität additiv ist. Wegen der Darstellbarkeit jeder Klasse der Taxonomie als Vereinigung von Basisklassen können wir die Produktivität nicht nur für die Basisklassen, sondern auch für beliebige Klassen der Taxonomie bestimmen.

6.5. DIE EMPIRISCHE BESTIMMUNG DER PRODUKTIVITÄT

Obwohl die Definition 2' die Produktivität über einen Grenzwert bestimmt, liefert sie doch eine praktische Meßanleitung. Wir können zwar nicht die Klasse des Rangs ∞ nehmen, stattdessen aber eine Klasse **hinreichend** seltener Wörter, für die wir dann feststellen können, wir groß der Anteil der Wörter der „zu wiegenden" Klasse X an dieser Klasse ist. Definition 2' besagt, daß der Beobachtungswert umso näher am wahren Wert der Produktivität einer Klasse liegen muß, je höher der Rang des betrachteten Fragments ist.

Dieses Meßverfahren für die Produktivität enthält eigentlich nichts Neues. Wir haben auch oben schon die Produktivität mittels Bestimmung des Gewichts des Anteils der Wörter einer Klasse an den

Neubildungen gemessen. Definition 2 stellt einige praktisch unerfüllbare Idealforderungen an die Bestimmung des Gewichts. Definition 2' zeigt, daß der zeitlichen Schichtung eine solche nach Häufigkeit entspricht, weil seltene Wörter praktisch auch immer Neubildungen sind; Neubildungen sind also am Schluß des Häufigkeitswörterbuchs zu suchen.

Die wesentliche Neuerung besteht darin, daß den Ausdrücken „hinreichend selten", „in den meisten Fällen" u.ä. jetzt eine präzise Bedeutung verliehen werden kann.

Wir kommen damit zur konkreten empirischen Bestimmung der Produktivität. Nehmen wir an, es sei die Produktivität eines Wortbildungselements der russischen Gegenwartssprache zu bestimmen, wobei Sprachwandel in den letzten 100 Jahren vernachlässigt werden kann. Überlegen wir, welches Material dazu erforderlich ist.

Aus unseren Daten geht hervor, daß die Rate des lexikalischen Wandels (der Parameter k) für die russische Literatursprache in den letzten 100 Jahren ca. k = 0,022/Jh. beträgt (s. Tab. 14). Man kann annehmen, daß ein 5%iges „Störrauschen" (der Anteil der mehr als 100 Jahre alten Wörter) die Meßergebnisse nicht allzu stark verfälscht. Wir müssen dann den erforderlichen Mindestumfang des Häufigkeitswörterbuchs bestimmen.

Aus (6.2) ergibt sich, daß der Schluß (der letzte Abschnitt) des Häufigkeitswörterbuchs dann 95% der höchstens 100 Jahre alten Wörter enthält, wenn das Wörterbuch etwa 1.850.000 Wörter umfaßt[3]. Ein solches Wörterbuch muß Wörter enthalten, die auf 20 Mio. Wortvorkommen höchstens einmal angetroffen werden. Solche Häufigkeitswörterbücher gibt es selbstverständlich nicht, sie lassen sich selbst mit Hilfe der EDV nicht realisieren.

Nun haben spezielle Stile des Russischen eine wesentlich höhere Veränderungsrate. Die russische wissenschaftliche Lexik hat sich in den letzten 150 Jahren mit einer mittleren Geschwindigkeit k = = 0,046/Jh. (s. Tab. 6) verändert. Bei gleicher Genauigkeitsanforderung wie oben wäre zur Messung der Produktivität einer Klasse ein Wörterbuch von ca. 200.000 Wörtern erforderlich; d.h. es müßte immer noch um zwei Ordnungen größer als die existierenden Häufigkeitswörterbücher des Russischen sein.

Wir können also die Produktivität nur grob schätzen, und zwar nicht für die Gegenwartssprache im eigentlichen Sinn, sondern für eine direkt vorausgehende Periode in der Größenordnung von 100 Jahren.

Wir wollen jetzt eine solche Schätzung der Produktivität einiger russischer Suffixe (d.h. der Klassen von Wörtern mit diesen Suffixen) durchführen. Dabei gehen wir davon aus, daß der Anteil der Neubildungen in einem gewöhnlichen (Definitions- Rechtschreib-, Übersetzungs- u.ä.) Wörterbuch umso größer ist, je mehr Wörter es enthält. Wenn man annimmt, daß die 80.000 Wörter des für solche Berechnungen geeigneten rückläufigen Wörterbuchs des Russischen von Bielfeldt [43] auch die 80.000 häufigsten sind[4], so ist es möglich, daß der Anteil der Neubildungen (wobei unter Neubildungen Wörter mit $\eta_t = 0{,}05$, d.h. einem Alter von 2 bis 2,5 Jh. verstanden werden) etwas mehr als die Hälfte der Gesamtzahl ausmacht[5]. Im rückläufigen russischen Wörterbuch von R. Greve und B. Kroesche (Red.: M. Vasmer) [45], das ca. 260.000 Wörter umfaßt, müßte der Anteil der Neubildungen unter denselben Bedingungen ca. 80% betragen (was aber nicht der Fall ist, da das Wörterbuch Material aus veralteten lexikographischen Quellen wie z.B. den Wörterbüchern von Vl. Dal' und I.Ja. Pavlovskij enthält), während im 2500 Wörter umfassenden Štf der Anteil weniger als 20% beträgt. Wenn man die Wörterbücher nach Umfang ordnet, kann man also erwarten, daß der Anteil der Wörter, die zu einer produktiven Klasse gehören, in dieser Reihenfolge wächst oder zumindest nicht abnimmt, und daß der Anteil der Wörter aus nichtproduktiven Klassen abnimmt. Eine völlig unproduktive Klasse läßt sich jedoch auf diese Weise nur schwer von einer Klasse unterscheiden, deren Produktivität in der letzten Periode stark zurückgegangen ist.

In Tab. 31 geben wir für verschiedene Suffixe den Anteil der Wörter pro 500-er Klasse von Štf (Spalten 2 bis 6) und an den letzten 1000 Wörtern von Štf (Spalte 7) an, sowie die mittlere Anzahl (auf ganze Zahlen gerundet) der Bildungen mit gegebenem Suffix pro 1000 Wörter in den Wörterbüchern von Bielfeldt (Spalte 8) und Greve/Kroesche (Spalte 9). Außerdem schien es uns nützlich, zusätzlich (Spalte 10) die Anzahl der Neubildungen pro Klasse in Tausend aus dem Handbuch „Novye slova i značenija" (1971) [46] anzu-

Tabelle 31. Russisch. Verteilung einiger Suffixe im Lexikon

Suffix*	Štf							Biel-feldt	Greve/ Kroesche	„Novye slova i značenija"
	1-500	501-1000	1001-1500	1501-2000	2001-2500	2501-2500				
1	2	3	4	5	6	7		8	9	10
-myj, -noj (141) (učebnyj, lesnoj)	11	26	35	35	34	69		122	100	-**
-nie, -ie (85) (penie, šestvie)	5	15	30	22	13	35		40	18	25
-ka, -očka (51) (rečka, lampočka)	3	7	14	17	10	27		37	47	38
-ti (34) (idti)	11	7	4	7	5	12		3	4	-
-skij, -skoj (29) (russkij, tverskoj)	5	6	5	8	5	13		34	13	-
-ok (28) (pryžok)	4	7	5	6	6	12		9	12	1
-stvo (28) (masterstvo)	5	5	6	4	8	12		12	7	6
-nij (24) (lišnij)	2	2	7	7	6	13		2	1	-
-kij (21) (legkij)	4	9	3	2	3	5		2	1	-
-ost' (18) (legkost')	0	2	5	2	9	11		31	17	20
-nik (15) (kolchoznik)	5	4	2	4	1	5		15	21	20
-ota (6) (rabota)	1	0	4	1	0	1		1	1	2
-ba (6) (služba)	1	3	1	0	1	1		1	1	2

*In Klammern die Gesamtzahl der Bildungen in Štf. **In dem Handbuch wurden nur die Substantive gezählt.

geben. Vergleicht man die Werte der Spalten 7,8 und 9 (wobei die Zahlen der letztgenannten Spalte nur unter Vorbehalt gewertet werden können - angesichts der o.g. Besonderheiten von Greve/Kroesche), so kann man zu dem Schluß kommen, daß die Nominalsuffixe -nie, - ka (-očka), -ost', -nik und die Adjektivsuffixe -nyj (noj), -skij begründet als produktiv gelten können. Bei -nie und -stvo kann man zumindest von geringer Produktivität sprechen. Die übrigen Suffixe dagegen (nominal: -ok, -ota, -ba; adjektivisch: -nij, -kij; verbal: -ti) sind wohl nicht produktiv. Diese Ergebnisse stimmen auch mit der traditionellen Einschätzung der Produktivität dieser Suffixe überein (vgl. [47, 48]).

6.6. GRENZEN DER MESSBARKEIT VON PRODUKTIVITÄT

Die Frage nach einem Berechnungsverfahren für die Produktivität einer Klasse muß scharf unterschieden werden von der Frage nach der prinzipiellen Meßbarkeit der Produktivität einer Klasse. Wie aus dem oben gesagten hervorgeht, hängen die Meßmethoden und ihre Genauigkeit wesentlich von den verfügbaren Ausgangsdaten ab. Ein weiterer für die Genauigkeit bestimmender Faktor, der bisher noch nicht erwähnt wurde, liegt in den statistischen Verfahren, die den Zufallsanteil an den Ergebnissen kontrollieren. Die Behandlung des ganzen Komplexes der speziellen Probleme, die die praktische Messung und die Verarbeitung der Meßdaten betreffen, geht über den Rahmen dieser Arbeit hinaus, und das umso mehr, als diese Probleme ohne konkrete Vorgabe eines Untersuchungszwecks schwerlich lösbar sind. Unser Thema ist die sprachwissenschaftliche Fundierung des vorgeschlagenen Meßverfahrens und die Übereinstimmung des Maßes mit unserem intuitiven Konzept der Produktivität.

An dieser Stelle muß das Verhältnis der Definition 1 und 2 (2') geklärt werden. Wir zeigen, daß jede Klasse, die nach Def. 2 produktiv ist, auch nach Def. 1 produktiv ist, aber nicht umgekehrt.

Da nach Annahme 4^{O} α nicht vom Rang abhängt, und für beliebige i,j,t $\alpha_i(t)$ gilt, kann man mit $r_i(t) = |N_{i,t} \cap X|$ die folgende Kettenproportion anschreiben:

$$\alpha(t) = \frac{r_1(t)}{N_1(t)} = \frac{r_1(t) + r_2(t)}{N_1(t) + N_2(t)} = \ldots = \frac{\sum_{i=1}^{\infty} r_i(t)}{\sum_{i=1}^{\infty} N_i(t)}.$$

Wenn X eine produktive Klasse im Sinne von Def. 2 ist, dann ist für beliebiges t $\alpha(t) > 0$. Das kann wegen des unendlichen Grenzwertes im Nenner für $i \to \infty$ nur dann der Fall sein, wenn auch der Zähler gegen Unendlich strebt, sonst ist $\alpha(t) = 0$. Die Def. 1 fordert wesentlich weniger, nämlich daß die Anzahl der Neubildungen unter den Wörtern einer Klasse für beliebiges t größer oder gleich 1 sein soll: $r_i(t) \geq 1$; es ist also hinreichend, wenn die Anzahl der Wörter der Klasse X_p endlich ist.

Solche produktiven Klassen „im weiten Sinn" gibt es. Zu ihnen gehören solche isolierten, manchmal okkasionellen Bildungen wie die unlängst entstandenen russ. korabel, knižkin (dom), risovišče u.a.

Doch kann man auch eine unendliche, nach Def. 1 produktive Klasse konstruieren, die nach Def. 2 nicht meßbar ist. Sei X eine nach Def. 2 produktive Klasse. Wir bilden eine neue Klasse X' wie folgt: „a $\in$ X' genau dann, wenn a $\in$ X und a gehört zur Klasse G_i mit der geraden Rangzahl i". Einer solchen Klasse kann man kein Produktivitätsmaß zuordnen, weil sie offensichtlich Annahme 3^{O} verletzt.

6.7. EIN MERKMAL FÜR DIE PRODUKTIVITÄT EINER KLASSE

In einigen Fällen ist nicht der exakte Wert der Produktivität einer Klasse X wichtig, sondern nur, ob er von Null verschieden ist. Für die Basisklassen (und auch für einige andere, s.u.) kann dies prinzipiell ohne jede Grenzwertberechnung entschieden werden.

Sei X eine produktive Klasse der Basis. Der Einfachheit halber nehmen wir an, daß 1) für beliebiges t mit $t_o > t > 0$, $\alpha(t) = \alpha = \text{const}$ und 2) kein Wort $a \in X$ existiert mit dem Alter $t(a) \geq t_o$.

Dann kann die Anzahl $r_{X,i}$ der Wörter der Klasse X in der Klasse G_i ausgedrückt werden durch

$$r_{X,i} = \alpha \bar{N}_i(t_o) = \alpha N_o[1 - e^{-kt\sqrt{i}}]. \qquad (6.4)$$

Die Funktion (6.4) ist monoton wachsend.

Sei Y eine nichtproduktive Klasse, d.h. es gebe einen Zeitpunkt t_1, $t_1 \neq 0$ mit $\alpha(t_1) = 0$. Bedingungen (1) und (2) seien aber erfüllt.

Dann kann man offenbar die Anzahl $r_{Y,i}$ der Wörter der Klasse Y in der Klasse G_i ausdrücken durch

$$r_{Y,i} = \alpha[\bar{N}_i(t_1) - \bar{N}_i(t_o)] = \alpha n[e^{-kt_1\sqrt{i}} - e^{kt_o\sqrt{i}}]. \qquad (6.5)$$

Die Funktion (6.5) ist nicht monoton, sie wächst bis zu einem i_{max} und fällt dann mit wachsendem i, im Spezialfall $t = \infty$ fällt das Maximum der Funktion (6.5) auf den Koordinatenursprung, und die Funktion ist monoton fallend (vgl. 4.2).

Für Basisklassen reicht es also aus, jeweils die Funktion $r_{X,i}$ zu bestimmen, um produktive von nichtproduktiven Klassen zu unterscheiden. In der Praxis ist das doch nicht ganz so einfach: bei kleinen Werten von t_1 läßt sich der Anfangsabschnitt der Funktion (6.5) nur schwer von Funktion (6.4) unterscheiden, besonders in Anbetracht der unvermeidlichen Streuungen. Daher ist unsere Entscheidung über Produktivität oder Nichtproduktivität nur von relativer Sicherheit; diese kann aber durch die Berücksichtigung von Wörterbuchausschnitten mit seltenen Wörtern erhöht werden.

Man kann zeigen (s. 4.3), daß Bedingung 2 unwesentlich ist, und daß die Funktion $\alpha(t)$ beliebig sein kann: sie bleibt für produktive Basisklassen monoton bezüglich des Rangs i. Der einzige Unterschied zu (6.4) besteht darin, daß die Abhängigkeit von $r_{X,t}$ von i und t komplizierter wird. Die Verteilung der Vereinigung produktiver Basisklassen ist ebenfalls monoton.

Man kann aber zeigen, daß die Vereinigung einer produktiven und einer nichtproduktiven Klasse, die natürlich wieder produktiv ist, eine nichtmonotone Verteilungsfunktion haben kann.

Sei X eine produktive Klasse, wobei die Neubildungen dieser Klasse im Intervall $\tau = (0, t_o)$ entstanden seien, und Y eine nichtproduktive Klasse, deren Wörter während $\tau_1 = (t_1, \infty)$ entstanden seien, mit $t_o < t_1$, d.h. die Zeitintervalle τ und τ_1 sollen sich nicht überschneiden.

Sehen wir nun, wie die Gesamtheit der Neubildungen aus $\tau + \tau_1$ verteilt ist:

$$\bar{N}_i(t_o) + \bar{N}_i(t_1) = 1 - e^{-kt_o\sqrt{i}} + e^{-kt_1\sqrt{i}} =$$

$$= 1 - [e^{-kt_o\sqrt{i}} - e^{-kt_1\sqrt{i}}]. \qquad (6.6)$$

Die Differenz in der Klammer hat ein inneres Maximum und ist kleiner als 1, weshalb der gesamte Ausdruck (6.6) bezüglich des Rangs i ein inneres Minimum hat. Die Bildungen der Klasse X U Y machen einen gewissen Anteil aller Wörter aus den Intervallen τ und τ_1 aus, z.B. können es alle Wörter aus diesen Intervallen sein. Die Vereinigung X U Y ist eine produktive Klasse, ihre Rangverteilung aber keine monoton wachsende Funktion. Der Graph der Verteilung einer produktiven Klasse mit einem nichtproduktiven Anteil kann nicht nur eine „Senke" (s. oben) haben, sondern auch einen „Höcker": der Anteil der nichtproduktiven Bildungen, die am Anfang des Häufigkeitswörterbuchs konzentriert sind, kann größer sein als die aktuelle Produktivität der Klasse, d.h. der Anteil ihrer Bildungen am Schluß des Wörterbuchs.

6.8. DIE VERTEILUNG DER WÖRTER EINER KLASSE IM HÄUFIGKEITSWÖRTERBUCH: EMPIRISCHES MATERIAL

Bei der Analyse von Wörtern mit bestimmten Suffixen können wir nur zufällig auf eine Basisklasse stoßen. Im allgemeinen hat die Verteilung einer „Suffix-Klasse" mehrere Minima und Maxima, von denen einige allein durch Streuung bedingt sind, während andere u.U. die Geschichte dieses Suffixes oder die Existenz nichtproduktiver Muster unter den Bildungen mit diesem Suffix widerspiegeln. Wir können diese Information nicht vom statistischen „Störrauschen" trennen und interpretieren. Nichtsdestoweniger können Graphen und Tabellen der Verteilung der Wörter einer Klasse auf die Lexikonabschnitte der verschiedenen Ränge manchmal sehr interessant sein. In Einzelfällen kann man anhand solcher Darstellungen produktive und nichtproduktive Klassen klar unterscheiden.

Die nachstehende Abb. 31 ist auf der Grundlage von Tab. 32 erstellt. Letztere beruht auf Daten aus der Arbeit Harwood/Wright [49], in der die Verteilung der Wörter mit bestimmten Suffixen in dem englischen Häufigkeitswörterbuch von Thorndike/Lorge untersucht wird. Die Verfasser unterscheiden freie Bildungen, in denen ein Stamm als selbständiges Wort auftritt (accept-acceptable) und gebundene Bildungen des Typs amiable. In den Spalten 2 bis 12 von Tab. 32 ist die mittlere Anzahl der freien Bildungen mit einem gegebenen Suffix (Spalte 1) unter den 1000 Wörtern des angegebenen Rangintervalls des Häufigkeitswörterbuchs angegeben (die Wahl der Intervalle ist teilweise durch die Struktur des Wörterbuchs bedingt), in den Spalten 13 und 14 die mittlere Anzahl (pro 1000 Wörter) beliebiger freier und gebundener Bildungen im Intervall 1 bis 19440 bzw. 19441 bis 30000. Alle Zahlen sind auf ganze gerundet[6].

Zumindest bei zwei Suffixen (-ful und -ment) kann man wenn nicht von Nichtproduktivität so doch von abnehmender Produktivität sprechen.

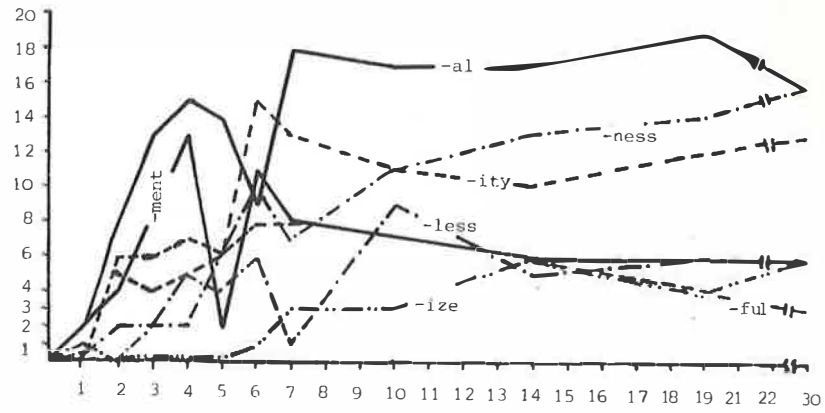

Abb. 31. Verteilung der engl. Wörter mit den Suffixen -al, -ful, -ize, -ity, -less, -ment, -ness im Häufigkeitswörterbuch Thorndike/Lorge

Tabelle 33 gibt einen Teil der Daten von Uhlířová und Pit'ha [50] zur Verteilung der tschechischen Substantive mit den häufigsten Suffixen im Häufigkeitswörterbuch Čech (10.000 Wörter) wieder. Alle diese Suffixe werden in der genannten Arbeit als produktiv angesehen. Zum Vergleich führen wir die Ergebnisse unserer Berechnungen auf der Grundlage desselben Wörterbuchs zur Verteilung der Substantive mit den drei im heutigen Tschechisch offensichtlich nichtproduktiven Suffixen -ota, -da und -yně (die letzten drei Zeilen von Tab. 33) an. Die Daten zum Englischen und Tschechischen kann man mit den Ergebnissen für das Russische auf der Basis von Štf vergleichen, die in Spalte 2-7 der Tab. 31 angegeben sind.

Tabelle 32. Englisch. Verteilung der Wörter mit bestimmten Suffixen im Häufigkeitswörterbuch Thorndike/Lorge

(1) Suffix	(2-12) Freie Bildungen											(13-14) Freie und gebundene Bildungen		
	1- -1000	1001- -2000	2001- -3000	3001- -4000	4001- -5000	5001- -6000	6001- -7000	7001- -10000	10001- -14000	14001- -19440	19441- -30000	1- -19440	19441- -30000
-al (industrial)	2	8	13	15	14	9	18	17	19	19	16	23	25
-ful (beautiful)	2	5	4	5	6	8	8	7	6	4	3	6	3
-ize (industrialize)	1	0	0	0	0	1	3	3	6	4	6	4	8
-ity (majority)	0	6	6	7	6	15	13	11	10	12	13	14	18
-less (needless)	0	0	2	5	4	6	1	9	5	7	7	5	7
-ment (announcement)	2	4	9	13	2	11	8	7	6	6	6	9	7
-ness (goodness)	0	2	2	2	6	10	7	11	13	14	16	10	16

Tabelle 33. Tschechisch. Verteilung der Wörter mit bestimmten Suffixen im Wörterbuch Čech

Suffix	Tausend										Insgesamt
	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	
-ek	8	12	9	9	7	3	6	1	7	0	62
-ina	4	4	7	6	9	5	3	2	4	6	50
-ka	9	8	12	22	9	9	10	4	7	13	102
-ní	7	19	26	25	38	39	30	28	37	32	281
-ník	2	9	2	11	1	2	8	9	3	4	59
-ost	15	14	17	15	19	19	20	21	22	28	190
-ství	3	5	3	8	7	2	2	5	7	7	49
-ota	3	3	0	3	1	2	2	2	0	1	17
-da	3	2	0	1	2	0	0	0	0	0	8
-yně	0	0	0	3	2	0	0	0	0	1	6

6.9. PRODUKTIVITÄT, UMFANG UND GEBRAUCHSHÄUFIGKEIT VON KLASSEN

In diesem Paragraphen betrachten wir einige Konsequenzen der oben dargestellten Theorie. Einige davon sind für empirische Untersuchungen der Produktivität von Interesse, doch wir wollen in der Hauptsache zeigen, wie stark der Begriff der Produktivität mit einem ganzen Komplex sprachwissenschaftlicher Probleme zusammenhängt.

Zunächst analysieren wir den Zusammenhang von Produktivität und Umfang (Verbreitung oder funktionelle Belastung) einer Klasse X. E.S. Kubrjakova, die Verfasserin des Abschnitts über Wortbildung in der Monographie [42], schlägt vor, den Umfang einer produktiven Klasse als ihre Produktivität zu bezeichnen, im Unterschied zur Gebrauchshäufigkeit, unter der „die Häufigkeit eines gegebenen Musters im Text" verstanden wird, d.h. der Anteil der Wörter einer gegebenen Klasse am Text, und im Unterschied zur Aktivität. Der letztgenannte Begriff entspricht offenbar eher unserem Begriff der Produktivität.

Der Umfang (die Verbreitung) einer Klasse und ihre Produktivität hängen offenbar eng zusammen, was in der Sprachwissenschaft seit langem bekannt ist: „...von den Vertretern der Genfer Schule ist äußerst scharf gesehen worden, daß in der Grammatik der Begriff des „Toten", Nichtproduktiven nahezu mit dem Begriff des „Zählbaren", zahlenmäßig Erfaßbaren gleichgesetzt wird. Das, was gezählt werden kann, ist nichtproduktiv. Z.B. wird die Gruppe der russischen Verben borot', kolot', polot', porot' durch die angegebenen Beispiele erschöpfend aufgezählt. Sie ist Verb-„Staub" (mit den Worten de Saussures). Neue Verbtypen entstehen nach diesem Muster nicht". [47: 136].

Die Vermutung, daß nichtproduktive Klassen endlich sind, läßt sich durch eine einfache Rechnung leicht erhärten. Offenbar ist die Anzahl der Elemente einer nichtproduktiven Klasse X, die vor t Jahren aufgehört hat zu wachsen, nicht größer als die Anzahl aller Wörter, die mindestens t Jahre alt sind:

$$|X_p| \leq |N_t| = N_o \sum_{i=1}^{\infty} e^{-\eta_t \sqrt{i}} < \infty. \qquad (6.7)$$

Diese Behauptung läßt sich offenbar auch auf produktive Klassen im weiten Sinn (d.h. nach Def. 1, nicht aber nach Def. 2) verallgemeinern: wenn man der Summe auf der rechten Seite der Ungleichung eine endliche Zahl von Bildungen hinzufügt, ändert sich nichts.

Im Prinzip könnten wir die Anzahl der Wörter in einer Klasse X exakt berechnen, wenn wir die zeitliche Veränderung ihrer Produktivität kennen würden:

$$|X_p| = \sum_{i=1}^{\infty} r_{p,i}, \qquad (6.8)$$

wobei $r_{p,i}$ der Anteil der Wörter aus X_p ist, die im Zeitintervall $[t_1,t_2]$, als die Klasse X_p produktiv war, in die Klasse G_i gekommen sind.

$r_{p,i}$, die Menge der Wörter aus der Klasse X_p in der i-ten Klasse, setzt sich aus den Anteilen $\alpha(t)$ der Neubildungen $\Delta N_i(t)$ in einem kleinen Zeitintervall Δt zusammen, innerhalb dessen $\alpha(t)$ als konstant angenommen werden darf:

$$r_{p,i} = \int_{t_1}^{t_2} \alpha_p(t)(\bar{N}_i(t))'_t dt.$$

Wie aus dem oben Gesagten hervorgeht, sind die produktiven Klassen unendlich, so daß der Begriff des Umfangs nicht sinnvoll auf sie angewendet werden kann, und an seine Stelle gerade die Produktivität tritt, d.h. die Wahrscheinlichkeit für das Auftreten von Wörtern einer solchen Klasse unter den Neubildungen bzw. den seltenen Wörtern.

Aus der Definition der Produktivität folgt insbesondere, daß unter den häufigen Wörtern der Anteil der Wörter nichtproduktiver Klassen größer ist als unter den Wörtern mittlerer oder geringer

Häufigkeit. Das ist aber schon lange bekannt. Die Häufigkeit ist schon immer als ein Faktor gesehen worden, der analogischen Ausgleich, unbegrenzte Verbreitung eines produktiven Typs hemmt. „Wörter mit nichtproduktiven Suffixen", schrieb A.I. Smirnickij, „können in einem beliebigen Text zahlreicher auftreten als Wörter mit produktiven Suffixen" [51: 106].

Die Situation wirkt auf den ersten Blick paradox: nichtproduktive Bildungen machen im Text einen bedeutenden Anteil aus, während ihr Anteil am Lexikon verschwindend ist; produktive Bildungen dagegen machen fast das ganze Lexikon aus, können im Text aber äußerst selten sein. Wie selten, läßt sich im Prinzip berechnen. Dazu müssen wir nur wissen, wie der Rang der Klasse G_i mit der mittleren Häufigkeit F_i ihrer Wörter zusammenhängt. Wenn man sich mit der Näherung, die das bekannte „Zipfsche Gesetz" $F_i = \frac{c}{i^a}$ ermöglicht, begnügt, kann man den Anteil R_p der Wörter eines Textes, die zu einer produktiven Klasse X_p gehören, folgendermaßen schreiben:

$$R_p = \sum_{i=1}^{\infty} \frac{c}{i^a} r_{p,i}. \qquad (6.9)$$

SCHLUSSBEMERKUNGEN

Wir wollen den Inhalt des Buches kurz zusammenfassen. Er setzt sich zusammen aus 1) einem mathematischen Modell des lexikalischen Wandels, 2) seiner Interpretation und 3) seiner empirischen Überprüfung.

Zu dem empirischen Material sei hier nur bemerkt, daß wir nicht absolut verläßliche empirische Ergebnisse zur Geschichte der Sprachen, deren Lexika wir untersucht haben, produzieren, sondern nur die Möglichkeiten unseres Modells demonstrieren wollten. Zur Gewinnung exakter empirischer Resultate z.B. bezüglich der Verände-

rungsrate in einem bestimmten Zeitintervall oder für eine gegebene Menge von Wörterbüchern wären ein wesentlich verbessertes statistisches Instrumentarium und die volle Nutzung der Testtheorie erforderlich, um den Zufallsanteil an unseren Ergebnissen zu verringern. Vor allem aber wäre es erforderlich, die Quellen systematischer Fehler auszuschalten, die auf der Unzulänglichkeit der verwendeten lexikographischen Hilfsmittel beruhen. Die einzige uns bekannte Methode zur Ausschaltung solcher Fehler ist die philologische Kritik der lexikographischen Quellen, welche von den jeweiligen philologischen Experten geleistet werden muß. Danach sind „saubere" empirische Untersuchungen zur Gewinnung exakter Zahlen möglich. Die ziemlich groben Schätzungen einzelner Parameter des Lexikons in der vorliegenden Arbeit dürfen im Hinblick darauf nur als vorläufig angesehen werden.

Allgemein gilt, wie das Beispiel der Naturwissenschaften zeigt, daß die anfängliche Meßgenauigkeit einer theoretisch sicher fundierten Meßmethode vielfach gesteigert werden kann, daß es aber fast immer zweckmäßig ist, dieses Problem von dem Problem der theoretischen Fundierung abzutrennen. Bezüglich der Interpretation haben wir hier einen ziemlich konservativen Standpunkt bezogen, indem wir angenommen haben, daß zu unserer Aufgabe die Beschreibung des Prozesses des lexikalischen Wandels gehört, in dessen Verlauf ein Teil der Wörter seine Häufigkeit behält oder sogar erhöht, während der andere Teil seltener wird oder sogar ganz schwindet. Für die geschwundenen Wörter treten Neubildungen auf, die zum Teil Entlehnungen (im üblichen Sinn) sind. Ein Lexikon kann einen oder mehrere Nachfolger haben. Diese Nachfolger sind zunächst völlig identisch, verändern sich aber später unabhängig voneinander als Bestandteile selbständiger Sprachen, und die Anzahl ihrer Unterschiede nimmt zu. Wir sind von der Voraussetzung ausgegangen, daß man „Momentaufnahmen" von den Lexika der Sprachen auf den verschiedenen Ästen eines Stammbaums machen kann, und durch den Vergleich dieser Lexika zu Aussagen über den Verwandtschaftsgrad der betrachteten Sprachen kommen kann.

Man kann einen bedeutenden Teil des formalen Instrumentariums dieser Arbeit jedoch auch anders, einfacher, interpretieren, nämlich so, daß man zeitweilig von der Vielzahl der Sprachen und der Komplexität ihrer Beziehungen abstrahieren kann. Betrachten wir ein (unendliches) Lexikon V, wobei für jedes seiner Wörter a zwei Größen gegeben sind: t(a), das Alter des Wortes a, und i(a), sein Rang. Unter dem Alter t eines Wortes verstehen wir die Anzahl der jeweils gewählten Zeiteinheiten (Jahrhunderte, Jahre u.ä.), seit dem Zeitpunkt des ersten Auftretens dieses Wortes in der Sprache. Bezüglich des Rangs nehmen wir an, daß das Lexikon V nach Häufigkeit geordnet und derart in gleich große Klassen zerlegt ist, daß die Wörter einer Klasse G_i die gleiche oder ähnliche Häufigkeit besitzen (s. 2.1). Unter dem Rang i eines Wortes wird der Rang seiner Klasse G_i verstanden. Die Theorie gibt eine probabilistische Beschreibung des Zusammenhangs von Alter t und Rang i eines Wortes: für ein gegebenes Paar (t,i) kann die Wahrscheinlichkeit $N_i(t)$ dafür bestimmt werden, daß ein aus der Klasse G_i zufällig gewähltes Wort mindestens das Alter t hat; d.h. also $N_i(t)$ ist der Anteil der Wörter des Mindestalters t an G_i. Die Theorie läßt sich aus einer Reihe einfacher Postulate herleiten.

POSTULAT 1. Die Veränderung der Größe $N_i(t)$ in einem beliebigen, ziemlich kleinen Zeitintervall Δt ist $N_i(t)$ proportional, wobei der Proportionalitätsfaktor allein vom Rang der Wörter aus der Klasse G_i abhängt.

Aus Postulat 1 folgt unmittelbar Formel (2.6) für den Anteil $N_i(t)$ der mindestens t Jahre alten Wörter. Formel (2.5) für den Anteil $\overline{N}_i(t)$ der weniger als t Jahre alten Wörter kann leicht aus Formel (2.6) hergeleitet werden. Das Postulat 1 kann man dahingehend verallgemeinern, daß η nicht nur vom Rang i, sondern auch von der Zeit t abhängt, $\eta = \eta(i,t)$, womit wir Formel (2.29) u.ä. erhalten.

Eine wichtige Beschränkung für die Abhängigkeit des Parameters η_i vom Rang i gibt

POSTULAT 2. Für beliebiges $t > 0$ ist die Anzahl der mindestens t alten Wörter des Lexikons V endlich.

Formel (2.15), wo η_i als proportional zur Quadratwurzel des Rangs i angenommen wird, ist nur eine der möglichen funktionalen Abhängigkeiten, die mit Postulat 2 verträglich sind. Die Quadrat-

wurzel ist auf der Basis empirischen Materials gewählt und kann durch genauere Untersuchungen bestätigt oder korrigiert werden. Formel (2.15) und die daraus abgeleiteten Formeln geben an, wieviele Wörter des Alters t in der Klasse G_i mit dem Rang i sein werden, m.a.W. wie die Wörter eines bestimmten Alters über die Ränge verteilt sind. Genauso läßt sich auch das umgekehrte Problem lösen: auf der Basis einer empirischen Bestimmung des Anteils der vor (oder nach) einem Zeitpunkt entstandenen Wörter dieses t zu bestimmen. Auf diese Weise können wir Wortmengen eine Maßzahl zuordnen, sie „datieren". Eine solche Maßzahl ist in einigen Fällen interessant, sie läßt sich jedoch nur sehr speziellen Mengen zuordnen, nämlich solchen, deren Rangverteilung der Gesetzmäßigkeit (2.5) folgt. Es scheint natürlich, die Klasse der meßbaren Mengen zu erweitern, indem außer Mengen von mindestens t alten Wörtern auch Mengen von höchstens t und mindestens t_1 alten Wörtern oder weniger als t alten Wörtern (Menge von Neubildungen) und schließlich Mengen Y_p von Wörtern betrachtet werden, die den „regulären" Teil der genannten „temporalen" (d.h. durch ein Zeitintervall definierten) Mengen N_j bilden.

Die „Regularität" der Y_p besteht darin, daß die Elemente dieser Menge gleichmäßig unter den Elementen der temporalen Mengen N_j verteilt sind, d.h. daß der Anteil α_p der Elemente von Y_p an den Elementen von N_j nicht vom Rang i abhängt. Diese Vorstellung der „Regularität" liegt der Analyse der Verteilung einer Menge von Entlehnungen aus einer gegebenen Quelle im Lexikon V (Kap. 4 und 5) und ebenso der Definition der Produktivität (Kap. 6) zugrunde. Die Produktivität ist gerade der Anteil α_p der Wörter einer gegebenen Klasse an einer temporalen Wortmenge, der Menge der Neubildungen.

So bleibt der größte Teil der Ergebnisse dieser Arbeit auch dann sinnvoll, wenn man davon abstrahiert, daß der gegenwärtige Sprachzustand das Ergebnis eines historischen Entwicklungsprozesses ist. Die Bedeutung dieser Ergebnisse liegt dann in der Konstruktion einiger Verfahren zur Zuordnung von Maßzahlen zu einer Klasse von Teilmengen des Lexikons V.

Um die abgeleiteten Formeln als Beschreibung des Prozesses der lexikalischen Veränderung interpretieren zu können, ist noch ein weiteres Postulat erforderlich:

POSTULAT 3. Alle Ausdrücke der Theorie bleiben sinnvoll, wenn unter dem Parameter t(a) nicht das Alter eines Wortes im Lexikon V verstanden wird, sondern das Zeitintervall seiner Existenz in V.

So gibt Formel (2.5) nicht nur an, wieviele Wörter der Klasse G_i das Alter t haben, sondern zugleich, wieviele Wörter dieser Klasse im Lexikon V am Ende des Intervalls t erhalten bleiben werden. Durch Vergleich eines gegenwärtigen Lexikons V_o mit einem seiner vergangenen Zustände V_t können wir mit Postulat 3 feststellen, daß in der Klasse $G_i \subset V_o$ die Anzahl der t alten Wörter gleich der Anzahl der Wörter aus der Klasse $G_i' \subset V_t$ gleichen Rangs ist, die wahrscheinlich das Zeitintervall t überleben (d.h. bis zum Zeitpunkt t = 0 existieren). (Dabei wird selbstverständlich nicht behauptet, daß die aus der Klasse $G_i' \subset V_t$ erhaltenen Wörter dieselben Wörter sind, die wir in der Klasse $G_i \subset V_o$ zu t = 0 antreffen.

Da jedes Paar von Lexika, die in der Nachfolger-Beziehung stehen, in dem oben beschriebenen Entsprechungsverhältnis steht, gibt es keinen prinzipiellen Unterschied zwischen Vergangenheit und Gegenwart, und es wird damit möglich, vom Prozeß des lexikalischen Wandels zu sprechen und nicht nur von den Eigenschaften eines Lexikons zu t = 0.

Der Unterschied zwischen dem Vorgänger-Lexikon V_1 und dem Nachfolger V ist, wie aus (2.19) ersichtlich, eine Exponentialfunktion des „Weges" s, den der Nachfolger V zurücklegt. Speziell ist dieser Weg bei konstanter Geschwindigkeit k dem Zeitintervall zwischen V_1 und V proportional.

Es bleibt noch das Problem des Unterschieds zweier Lexika V und V', die Nachfolger desselben Lexikons V_1 sind, offen. Dieses ist in der vorliegenden Arbeit für den Fall zweier sich von Anfang an voneinander völlig unabhängig entwickelnder Lexika V und V' analysiert worden. Unter dieser Bedingung (s. 2.4) ist der „Weg" s, von dem der Unterschied der Wörterbücher abhängt, eine Summe von Wegen: von V bis V_1 und von V_1 bis V' (Speziell ist, falls, die von den beiden Lexika V und V' zurückgelegten „Wege" s gleich sind, ihr Abstand 2s). Leider mußten wir uns hier wie auch in vielen anderen Fällen auf die Analyse eines Spezialfalls beschränken.

ANMERKUNGEN

1.5. [1] Eine zugegebenermaßen oft triviale „Grammatik der Entlehnungen" benutzt in der etymologischen Praxis jeder Linguist, wenn er z.B. behauptet, daß das Wort mar'jaž <mariage aus dem Französischen ins Russische entlehnt ist.

1.6. [2] Wesentlich ist nur, daß ein Wort, daß irgendwann „zerfallen" ist, nicht nach einer gewissen Zeit wieder in die KL kommt. Die Möglichkeit solcher „Wiederaufnahmen" wird hier einfach vernachlässigt.

[3] Den Ausdruck (1.2) kann man in strikter Analogie zur Ableitung des Gesetzes des radioaktiven Zerfalls erhalten. Diese Analogie (identische Differentialgleichung) ist nicht zufällig: M. Swadesh wurde von der kurz vorher entdeckten Radiocarbonmethode angeregt [8].

[4] Interessanterweise beziehen sich solche Anmerkungen fast nie auf den unklaren Begriff der „lexikalischen Entsprechungen" zwischen zwei Sprachen; denn in der Praxis stimmen die verschiedenen Autoren bei der Bestimmung lexikalischer Entsprechungen durchaus nicht immer überein. In manchen Fällen ist Gleichheit der Wurzel einziges Kriterium, in anderen werden die morphologische Form, zusätzliche Bedeutungen u.ä. berücksichtigt.

[5] Vgl. die völlig plausiblen Daten, die G.A. Klimov [18] mit derselben Methode erhielt, für die georgisch-swanische Einheit (4000 Jahre), die georgisch-sanische (2700 Jahre) und die swanisch-sanische (4000 Jahre).

[6] Leider ist nicht klar, ob die Begriffsordnung für eine Sprachgruppe auf andere übertragen werden kann. Es werden ja in Wirklichkeit nicht Begriffe, sondern Wörter einer Sprache geordnet (in [11] Wörter der Protosprache einer Gruppe von australischen Eingeborenenstämmen). Es kann bezweifelt werden, daß die Überlebenswahrscheinlichkeit eines Wortes direkt mit seiner Bedeutung zusammenhängt. Man sollte eher erwarten, daß die Zerfallswahrscheinlichkeit von generellen Bedeutungseigenschaften abhängt; Wörter mit enger, spezieller Bedeutung müßten eher zerfallen als Wörter mit verschwommener Bedeutung.

[7] Um im Bild des radioaktiven Zerfalls zu bleiben: Hier handelt es sich um den Zerfall eines Gemenges radioaktiver Substanzen mit verschiedenen Halbwertzeiten.

[8] Swadesh schreibt zur Sprachmischung: „lexikostatistische Untersuchungen, aber auch andere Daten zeigen, daß es solche Phänomene nicht gibt" [8:45]. Das stimmt selbstverständlich nicht; sie können auch nichts derartiges zeigen, weil sie diese Nichtexistenz zur Voraussetzung haben.

1.8. [9] Später verwarf er in seiner Monographie [25] diese Annahme mit der Begründung, daß seine Daten zur quantitativen Bestimmung der Abhängigkeit nicht ausreichten (S. 119); dabei tendierte er zu einer parabolischen oder hyperbolischen Funktion.

2.1. [1] Es interessiert uns hier nicht, wie die Abbildung φ von $\tilde{V}_x$ auf $\tilde{V}_y$ konstituiert wird, und ebensowenig die damit verbundenen Begriffe der Gesetze der historischen Phonetik, Morphologie und Semantik. Die Entsprechung muß nur eindeutig sein. Diese Forderung ist in der Realität fast immer erfüllt, wenn man von der geringen Zahl von Doubletten (zwei oder mehr Wörter mit gemeinsamem Vorgänger) absieht.

[2] D.h. das Verhältnis der Zahl der geschwundenen Wörter zur Zahl aller Wörter der Klasse: $\bar{N}_i(t)$ ist immer ≤ 1.

2.3. [3] Da die Ausdrücke für den Anteil (2.6) und die Anzahl (2.7) der in einer Klasse erhaltenen Wörter, die der Formel für die aus einer Klasse geschwundenen Wörter (2.5) und der Formel für den völligen Schwund (2.11, 2.12, 2.13) entsprechen, sich von letzteren nur durch ihre Konstanten unterscheiden, werden wir im Folgenden bei der Ableitung komplexerer Formeln, wo das inhaltlich möglich ist, nur eine Variante anführen.

3.1. [1] Zwar geht hier noch die Zusatzannahme ein, daß die Raten des lexikalischen Wandels für verwandte Sprachen gleich sind. Wie in 3.2 gezeigt wird, ist dies jedoch gerechtfertigt: die Geschwindigkeiten sind praktisch gleich.

3.2. [2] Unter Umständen kann es günstiger sein, anstelle von aufeinanderfolgenden Klassen (1 bis 100, 101 bis 200 etc.) sich einschließende Klassen (1 bis 100, 1 bis 200 etc.) zu nehmen. Dann müßte jedoch die Formel (3.3) durch eine andere ersetzt werden, und die entstehende Funktion würde komplizierter (s. Kap. 4).

3.4. [3] denn $\frac{0{,}26}{0{,}022 \cdot 2} \approx 6$

4.2. [1] Alle hier angeführten Berechnungen berücksichtigen nur den Fall konstanter Geschwindigkeit des lexikalischen Wandels. Für die Abhängigkeit der Geschwindigkeit von der Zeit werden die Berechnungen komplizierter, sprachwissenschaftlich ändert sich nichts.

4.3. [2] Wir gehen hier nicht auf die Frage ein, ob eine geeignete Klassifikation der Quellen in jedem Fall möglich ist. Wenn wir fragen, ob die russische Gegenwartssprache vom Altrussischen abstammt, hängt die Antwort nicht unwesentlich davon ab, was unter „Altrussisch" verstanden wird. Ist es ein begrenzter Entwicklungszeitraum des Russischen von einigen hundert Jahren, so steht das Altrussische zur russischen Gegenwartssprache in derselben Relation wie das Altkirchenslavische oder z.B. das Griechische. Ist es dagegen die Sprachtradition, die auch vorausgegangene Entwiclungsstufen (urslavisch, indogermanisch u.ä.) umfaßt, dann stammt die russische Gegenwartssprache vom Altrussischen ab.

[3] Entgegen der Behauptung von M. Swadesh (8:45) kann weder mit der lexikostatistischen Methode noch mit unserer Methode bewiesen werden, daß es keine Sprachmischung gibt. Vielmehr beruht die lexikostatistische Methode von Swadesh gerade auf Voraussetzungen (Schleichers Axiomen), die Sprachmischung ausschließen.

5.1. [1] Unter Wortbestand wird die Liste aller verschiedenen Wörter eines Werkes verstanden.

[2] Die Werke sind abgekürzt nach der Konkordanz von Tatlock [31].

[3] Wir nehmen mit Mersand an, daß Chaucers Lexik in einen germanischen und einen romanischen Teil restlos zerfällt, d.h.: prozentualer Anteil romanischer Lexik + prozentualer Anteil germanischer Lexik = 100%.

5.3. [4] Oder in den Werken eines bekannten Autors.

[5] Als Beispiel kann die relativ neue Arbeit zu den französischen Lehnwörtern des XIX. Jh. im Englischen von I. Krejn [34] dienen. Dieser behauptet, daß in dieser Zeit etwa so viele Wörter aus dem Französischen entlehnt wurden wie im Mittelalter, und mehr als im XVII. und XVIII. Jh. Dieses Ergebnis beruht auf der Analyse des "Oxford English Dictionary" und des "Shorter Oxford English Dictionary", welche „1932 Wörter und Wendungen erbrachte, die im XIX. Jh. aus dem Französischen entlehnt worden sind". Zum Vergleich sei erwähnt, daß im XVII. Jh. 1500 Wörter und Wendungen und im XVIII. Jh. 1000 entlehnt worden sind.

6.4. [1] Entsprechend läßt sich die Produktivität nicht nur für die Gegenwartssprache, sondern auch für einen beliebigen vergangenen Zeitpunkt t_o bestimmen. Dazu muß man die Zeitintervalle (und die entsprechenden Klassen von Neubildungen) betrachten, die gegen $t = t_o$ konvergieren.

[2] Tatsächlich ist $\mathbb{N}_i(t)$ nur die mathematische Erwartung der Anzahl von Wörtern mit diesem Alter. Jede empirische Beobachtung wird eine statistische Streuung um den Wert von (6.2) ergeben. Außerdem ist (6.2) unter bestimmten Annahmen zum Prozeß des lexikalischen Wandels hergeleitet, die nicht in allen Fällen gelten. Der entsprechende allgemeinere Ausdruck ist (2.29). Wir beschränken uns hier auf den spezielleren Fall, einmal weil die Formeln einfacher sind, und außerdem weil man immer nur kleine Zeitintervalle zu betrachten braucht, für welche die erwähnten Annahmen mit hinreichender Genauigkeit zutreffen.

6.5. [3] $1 - \exp\left(-0{,}022 \frac{1}{\text{Jh.}} \cdot \text{Jh.}\sqrt{\frac{1850000}{100}}\right) \approx 0{,}95$

[4] Das Wörterbuch von Bielfeldt wurde auf der Grundlage von russischen Definitionswörterbüchern erstellt, die die nach Ansicht ihrer Verfasser wichtigsten russischen Wörter enthalten. Einige Untersuchungen zeigen aber (s.[44]), daß subjektive Urteile über die Wichtigkeit eines Wortes gut mit seiner objektiven Häufigkeit übereinstimmen.

[5] nach Formel (4.8)

6.8. [6] Leider basiert das Wörterbuch Thorndike/Lorge auf Texten unterschiedlichen (in Einzelfällen sehr hohen) Alters. Daher läßt sich die Produktivität in der Gegenwartssprache mit diesem Wörterbuch schon prinzipiell nicht messen.

LITERATUR

1. ZIPF, G., Prehistoric 'cultural strata' in evolution of German: the case of Gothic. Modern Language Notes, 62, 1947, 522-530

2. BLUMFIL'D, L., Jazyk. Moskau 1968

3. PORCIG, V., Clenenie indoevropejskoj jazykovoj oblasti. Moskau 1964

4. BRUGMAN, Z., Kurze vergleichende Grammatik der indogermanischen Sprachen. Straßburg 1904

5. BERNŠTEJN, S.B., Očerk sravnitel'noj grammatiki slavjanskich jazykov. Moskau 1961

6. TRUBECKOJ, N.S., Mysli ob indoevropejskoj probleme. In: Voprosy jazykoznanija, 1, 1958, 65-77

7. JUSMANOV, N.V., Grammatika inostrannych slov. In: Slovar' inostrannych slov. Moskau 1939

8. SWADESH, M., Leksiko-statističeskoe datirovanie doistoričeskich étničeskich kontaktov (na materiale plemen éskimosov i severoamerikanskich indejcev). In: Novoe v lingvistike, I. Moskau 1960, 23-52

9. SWADESH, M., K voprosu o povyšenii točnosti v leksiko-statističeskom datirovanii. In: Novoe v lingvistike, I., Moskau 1960, 53-87

10. GUDSHINSKY , S., The ABC's of Lexicostatistics (Glottochronology). In: Word 12, 1956, 175-210

11. DYEN, J., JAMES, A. and COLE, I.W.L., Language Divergence and Estimated Word Retention Rate. In: Language 43, 1967, 150-171

12. KRUPA, V., Nové cesty glottochronológie. In: Jazykovedný Časopis, Bd. 18, Nr. 2, 1967, 167-170

13. CHOJER, G., MOIJEZ, M., Leksikostatistika (Kritičeskij razbor). In: Novoe v lingvistike, I. Moskau 1960, 88-107

14. CHRETIEN, D., The Mathematical Models of Glottochronology. In: Language 38, 1952, 11-37

15. LEVIN, S., The fallacy of universal list of basic vocabulary. In: Proc. of the IX-th International Congress of Linguistics. The Hague 1964, 232-236

16. FODOR, I., A glottochronológia érvenyessége a szláv nyelvek anyaga alapján. In: Nyelvtudományi közlemények, Bd. 63, Nr. 2, 1961, 308-356.

17. BERGSLAND, K. and VOGT, H., On the validity of glottochronology. In: Current anthropology, Bd. 3, Nr. 2, 1962, 115-153

18. KLIMOV, G.A., O glottochronologičeskom metode datirovanija raspada jazyka. In: Voprosy jazykoznanija 2, 1959, 119-121

19. MERWE, N.J. van der, New mathematics for glottochronology. In: Current anthropology, Bd. 7, 1966, 485-500

20. THORNDIKE, E., The teacher's word book. N.Y. 1921

21. LINDSEY, E., An etymological study of ten thousand words in Thorndike's teacher's word book. In: Indiana University Studies, Bd. 12, Study no. 65, 1925

22. ROBERTS, A., A statistical linguistical analysis of American English. The Hague, 1965

23. HORN, E., A basic writing vocabulary. 10.000 words commonly used in writing. Iowa City, 1926

24. ELDRIDGE, R.C., Six thousand common English words. Niagara Falls, 1911

25. ZIPF, G., Human behavior and the principle of least effort. Cambridge, Mass., 1949

26. FELLER, V., Vvedenie v teoriju verojatnostej i ee priloženija. Moskau 1964

27. TUDOSE, C., Vocabular fundamental al limbii române vechi. In: Sistemele limbii. Bukarest, 1970, 119-164

28. MERSAND, J., Chaucer's Romance vocabulary. N.Y., 1937

29. YULE, G., The statistical study of literary vocabulary. Cambridge, 1944

30. HERDAN, G., How can quantitative methods contribute to our understanding of language mixture and language borrowing. In: Statistique et analyse linguistique. Paris 1966, 17-39

31. TATLOCK, J., A concordance to the complete works of Geoffrey Chaucer. 1927

32. GUIRAUD, P., Problèmes et méthodes de la statistique linguistique. Dordrecht 1959

33. CHERC, M., O predstavitel'nosti teksta zadannoj dliny. In: Naučno-Techničeskaja informacija. ser. 2, Nr. 6, Moskau 1969, 25-28

34. KREJN, I.M., Francuzskie zaimstvovanija XIX v. v anglijskom literaturnom jazyke (avtoref. kand. diss.). Moskau 1963

35. JESPERSEN, O., Growth and structure of the English language. N.Y., 1929

36. BAUGH, A.C., The chronology of the French loan-words in English In: Modern Language Notes, Bd. 50, Nr. 2, 1935, 90-93

37. PATTERSON, W.T., On the genealogical structure of the Spanish vocabulary. In: Word 24, Nr. 1-3, 1968, 309-339

38. JUILLAND, A., CHANG RODRIGUES, E., Frequency dictionary of Spanish words. The Hague 1964

39. ISAČENKO, A.V., Morphologie, syntaxe et phraséologie. In: Cahiers F. de Saussure, Bd. 7, 1-48

40. VINOKUR, G.O., Zametki po russkomu slovoobrazovaniju. In: G.O. Vinokur, Izbrannye raboty po russkomu jazyku. Moskau 1959, 419-442

41. Russkij jazyk i sovetskoe obščestvo. Alma-Ata 1962

42. Obščee jazykoznanie. Vnutrennjaja struktura jazyka. Moskau 1972

43. BIELFELDT, H.H., Rückläufiges Wörterbuch der russischen Sprache der Gegenwart. Berlin 1958

44. FRUMKINA, R.M., Verojatnost' elementov teksta i recevoe povedenie. Moskau 1971

45. GREVE, R., KROESCHE, B., Russisches rückläufiges Wörterbuch, Bd. 1-4, Berlin 1957-1958

46. Novye slova i značenija. Slovar'-spravočnik po materialam pressy i literatury 60-ch godov. Moskau 1971

47. VINOGRADOV, V.V., Russkij jazyk. Moskau-Leningrad 1947

48. Grammatika sovremennogo russkogo jazyka. Moskau 1970

49. HARWOOD, F., WRIGHT, A., A statistical study of English word formation. In: Language 32, No. 2, 1956, 260-273

50. UHLÍŘOVÁ, L., PIT'HA, P., K voprosu postroenija jazyka-posrednika pri mašinnom perevode. In: The Prague bulletin of Mathematical Linguistics, Bd. 12, 1969, 40-69

51. SMIRNICKIJ, A.I., Leksikologija anglijskogo jazyka. Moskau 1956