Commentationes Mathematicae Universitatis Carolinae

Miloš Matula

Zur Frage der Häufigkeitsverteilung der Worte. I

Commentationes Mathematicae Universitatis Carolinae, Vol. 6 (1965), No. 2, 213--237

Persistent URL: http://dml.cz/dmlcz/105012

Commentationes Mathematicae Universitatis Carolinae
6, 2 (1965)

ZUR FRAGE DER HÄUFIGKEITSVERTEILUNG DER WORTE I.

Miloš MATULA, Praha.

Unter den funktionalen Beziehungen, die in der quantitativen Linguistik untersucht werden, spielen die Gesetzmässigkeiten, die sich auf die Häufigkeitsverteilung der Worte beziehen, eine wichtige Rolle. Zur Beschreibung dieser Verteilung werden verschiedene Formeln vorgeschlagen, z.B. die Zipfsche Formel

$$p = \frac{1}{mx + n} ,$$

bzw. die Mandelbrotsche Formel

$$p = \frac{1}{(mx + n)^{\varrho}}$$

(wobei $\varrho > 1$ vorausgesetzt wird), wo $p = p(x)$ die Wahrscheinlichkeit des Vorkommens des x-ten Wortes ist (wenn die Worte nach sinkender Wahrscheinlichkeit geordnet werden). In den eben angeführten Formeln wird p durch einen sehr einfachen analytischen Ausdruck dargestellt, allerdings ergibt sich eine befriedigende Übereinstimmung mit empirischen Ergebnissen nur in Teilintervallen von einer ziemlich begrenzten Länge.

In diesem Artikel (und in seiner Fortsetzung) werden wir uns mit bestimmten allgemeineren Ausdrücken befassen, die als Spezialfälle auch Formeln enthalten, die den Häufigkeitsverlauf verhältnismässig gut charakterisieren; andererseits enthalten sie als Spezialfälle einige aus der Literatur bekannte

Formeln.

Das Verlangen, zu einer mathematisch möglichst einfachen Gesetzmässigkeit zu gelangen, führt dazu, dass man bei diesen Betrachtungen $p(x)$ theoretisch als eine Funktion stetigen Arguments zulässt (oder gar voraussetzt); wir werden auch die Existenz ihrer Ableitungen (insoweit sie gebraucht werden) voraussetzen. Sonst wird natürlich $0 < p \leqq p_1 < 1$ (so dass $\log p$ existiert und es gilt $\log p < 0$) und $\sum p(i) = 1$ vorausgesetzt. Schliesslich muss $\{p(i)\}$ eine nichtwachsende Folge sein; wir machen die etwas stärkere Voraussetzung $p'(x) < 0$.

Es ist bekannt, dass häufigere Worte im Durchschnitt kürzer sind. Wenn man eine künstliche Sprache voraussetzt, für die die Wortlänge s eine monotone (nichtwachsende) Funktion der Häufigkeit p (d.h. eine nichtfallende Funktion des Ranges x) ist und dabei alle "Buchstaben" - Kombinationen zur Wortbildung ausgenutzt werden, dann ist $\sum_1^s k^i$ die Anzahl der Worte mit einer Länge $\leqq s$ (wo k die Anzahl von verschiedenen "Buchstaben" bedeutet); daraus folgt, dass man den Verlauf von $s(x)$ in einem leicht zu erklärenden Sinne durch eine Formel $s(x) = a \log (bx + c)$ darstellen kann (bei geeigneten Konstanten a, b, c). (Statt der Ausnützung aller "Buchstaben"-Kombinationen könnte man offenbar annehmen, dass z.B. nur alle Kombinationen benutzt werden, die durch regelmässiges Wechseln der "Konsonanten" und "Vokale" entstehen.) Wenn wir dabei voraussetzen, dass die Häufigkeitsverteilung der Worte in dieser künstlichen Sprache der Mandelbrotschen Formel genügt, dann gilt für genügend grosse x (wo man die Koeffizienten n, c vernachlässigen kann) $s(x) = a \log bx$,

$\log p(x) = -\rho \log mx$, so dass s linear von $\log p$ abhängt.

In einer natürlichen Sprache sind die Verhältnisse natürlich nicht so einfach. Führen wir aber eine Funktion $q(x)$ als die durchschnittliche Länge der Worte, deren Rang sich verhältnismässig wenig von x unterscheidet, ein, und betrachten wir die Ableitung $\frac{dx}{dq}$. Hier ist dx, grob gesagt, der Zuwachs der Anzahl der verschiedenen Worte, wenn die "lokale" durchschnittliche Wortlänge um dq wächst, also ist $\frac{dx}{dq}$, grob gesagt, ungefähr proportional der Anzahl W der neu hinzukommenden verschiedenen Worte. Nach vorläufigen empirischen Untersuchungen scheint es, dass die Formel $q = a \log|\log p| + b$ ($a > 0$, b konstant) verhältnismässig gut mit der Wirklichkeit übereinstimmt. Dann ist $q' = a \frac{p'}{p \log p}$, so dass $\frac{-p \log p}{p'}$ der Ableitung $\frac{dx}{dq}$ proportional ist. (Man beachte auch die informationstheoretische Bedeutung des Ausdruckes $-p \log p$.) Mit Rücksicht darauf (und auch auf die rein analytische Wichtigkeit des Ausdrucks $\frac{p}{p'}$ für die Beschreibung des Verlaufs von p) werden wir uns zunächst mit dem Ausdruck $\frac{-p \log p}{p'}$, bzw. mit dem Ausdruck $\frac{p}{p'}$ befassen.

Bezeichnen wir

$$\varphi = \left(\frac{p}{p'}\right)' .$$

Es gilt $\varphi = 1 - \frac{pp''}{p'^2}$, $\frac{p''}{p'} = \frac{p'}{p}(1 - \varphi)$, also $\log|p'| = \int \frac{1-\varphi}{p}\, dp$. Es ist zweckmässig, die Funktion φ als Funktion von q oder als Funktion von $\log p$ zu betrachten. Schreiben wir

$$1 - \varphi = f(\log p) \ .$$

Es sei nun p_1 eine Zahl mit $0 < p_1 < 1$, und $f(\log p)$ eine Funktion. Sei M die Menge der Zahlen $\gamma^* \geqq 0$, $\gamma^* < p_1$ mit der Eigenschaft, dass f als Funktion von p in (γ^*, p_1) stetig ist. Sei $M \neq \emptyset$. Sei $\gamma = \min_{\gamma^* \in M} \gamma^*$ (so dass f eine stetige Funktion von p in (γ, p_1) ist). Bezeichnen wir mit $F(z)$ eine Stammfunktion zu $f(z)$ (d.h. $F'(z) = f(z)$).

Setzen wir voraus, dass es eine Funktion $p(x)$ gibt, die für $x \in (1, \sigma)$, wo $1 < \sigma \leqq \infty$, definiert ist, mit den folgenden Eigenschaften: $p(1) = p_1$; $p(x) \to \gamma$ für $x \to \sigma$ in $(1, \sigma)$ existiert stetige zweite Ableitung $p''(x)$; in $(1, \sigma)$ gilt $p'(x) < 0$; $p(x)$ erfüllt die Gleichung

$$(1) \qquad \frac{p''}{p'} = \frac{p'}{p} f(\log p) \ .$$

Dann ist $\log |p'| = F(\log p) + C$ (C konstant), also

$$(2) \qquad p' = -k e^{F(\log p)} \qquad (k > 0 \text{ konstant}),$$

$$(3) \qquad \frac{dx}{dp} = -\frac{1}{k} e^{-F(\log p)}$$

und schliesslich

$$(4) \qquad x(p) = \frac{1}{k} \int_p^{p_1} e^{-F(\log p)} \, dp + 1 \ .$$

Wenn umgekehrt $x(p)$ durch (4) definiert ist und wenn $\sigma = \frac{1}{k} \lim_{p \to \gamma} \int_p^{p_1} e^{-F(\log p)} \, dp + 1$, so ist $1 < \sigma \leqq \infty$ und es gilt (3); $x(p)$ ist eine fallende Funktion in (γ, p_1) und für die inverse Funktion $p(x)$ gilt (2), also

$p'' = -k e^{F(\log p)} \frac{dF(\log p)}{d \log p} \frac{p'}{p} = p'^2 f(\log p) \frac{1}{p}$, so dass (1) erfüllt ist.

Auf die Funktion f stellen wir noch die folgende Forderung:

(Δ) Sei $\sigma = \infty$, $\sum_{i=1}^{\infty} p(i) < \infty$, $c > p_1$ konstant. Die Funktion $f^* = f[\log(c\,p)]$ ist stetig für $\frac{1}{c}\gamma < p \leqq \frac{1}{c} p_1$; es gilt $\sigma^* = \infty$, wo σ^* eine analogische Bedeutung hinsichtlich f^*, $\frac{1}{c}\gamma$, $\frac{1}{c} p_1$ hat wie σ hinsichtlich f, γ, p_1.

Wir werden sagen, dass das Paar $\langle f, \gamma \rangle$ den Typ (A) besitzt oder dass es einem endlichen Wörterbuch entspricht, wenn $\sigma < \infty$; dass $\langle f, \gamma \rangle$ den Typ (B) besitzt oder einem unendlichen Wörterbuch entspricht, wenn $\sigma = \infty$ und wenn dabei $\sum_{i=1}^{\infty} p(i)$ konvergiert; dass $\langle f, \gamma \rangle$ den Typ (C) besitzt oder einem erzwungen endlichen Wörterbuch entspricht, wenn $\sigma = \infty$ und dabei $\sum_{i=1}^{\infty} p(i) = \infty$. Wir sagen dann auch, dass die betreffenden Funktionen p vom Typ (A) bzw. (B) bzw. (C) sind usw.

Wenn $\langle f, \gamma \rangle$ vom Typ (B) ist und wenn für $p(x)$ die Beziehung $\sum_{i=1}^{\infty} p(i) = c$ gilt (so dass $c > p_1$), setzen wir $p^*(x) = \frac{1}{c} p(x)$; die Funktion p^* erfüllt dann auch noch die Forderung $\sum_{i=1}^{\infty} p^*(i) = 1$.

Die anschauliche Bedeutung ist klar: im Falle des Types (A) bewegt sich der Rang x in endlichen Grenzen, so dass wir ein "endliches Wörterbuch" voraussetzen (wobei die Wahrscheinlichkeit p in einer entsprechenden Formel durch eine positive untere Grenze beschränkt werden kann, oder es kann auch 0 theo-

retisch die untere Grenze für p sein); im Falle des Types (B) setzen wir ein theoretisch "unendliches Wörterbuch" voraus (wobei man, wenn wir eine Änderung der Zahl p_1 zulassen, durch Multiplizieren von p mit einer geeigneten Konstante auch die Forderung $\sum p(i) = 1$ erfüllen kann); im Falle des Types (C) wird p durch die Formel für beliebig grosse x definiert, aber die Endlichkeit des Wörterbuches wird durch die Divergenz der Reihe $\sum p(i)$ erzwungen.

Im allgemeinen kann man grösste Chanzen für das Finden einer mit empirischen Ergebnissen gut übereinstimmenden Formel unter den Fällen vom Typ (A) erwarten, kleinste Chanzen unter den Fällen vom Typ (C).

Im weiteren werden wir uns mit dem einfachen Falle

$$f(z) = \frac{az + b}{cz + d}$$

befassen, also mit dem Falle, wo $\left(\frac{p}{p'}\right)'$ entweder linear von $\log p$ abhängt oder konstant bleibt, bzw. für kleine p fast konstant (mit einer Differenz von der Form $\frac{1}{c_1 \log p + c_2}$ bleibt. Diese Formel enthält als Spezialfälle auch einige geläufig zitierte Formeln.

Im Falle $c \neq 0$ werden wir voraussetzen, dass gleichzeitig $bc - ad \neq 0$; wir setzen dann $\alpha = -\frac{a}{c}$, $\beta = -\frac{bc - ad}{c^2}$.

Im Falle $c = 0$ (und natürlich $d \neq 0$) setzen wir

$$A = \frac{1}{2}\,\frac{a}{d}, \quad B = \frac{b}{d}.$$

Wenn $c \neq 0$, $\frac{d}{c} > 0$, bezeichnen wir $\gamma = e^{-\frac{d}{c}}$ (so dass $c \log \gamma + d = 0$).

Es gilt:

I. Sei $c \neq 0$, $\frac{d}{c} > 0$. Dann ist $\langle f, \gamma \rangle$ vom Typ (A).

II. Sei $c = 0$.

Wenn $A > 0$ oder wenn $A = 0$, $B < 1$, dann ist $\langle f, 0 \rangle$ vom Typ (A).

Wenn $A = 0$, $1 \leqq B < 2$, dann ist $\langle f, 0 \rangle$ vom Typ (B).

Wenn $A = 0$, $B \geqq 2$ oder wenn $A < 0$, dann ist $\langle f, 0 \rangle$ vom Typ (C).

III. Sei $c \neq 0$.

Wenn $\alpha > -1$ oder wenn $\alpha = -1$, $\beta < -1$, dann ist $\langle f, 0 \rangle$ vom Typ (A).

Wenn $\alpha = -1$, $\beta \geqq -1$ oder wenn $-2 < \alpha < -1$ oder wenn $\alpha = -2$, $\beta < -1$, dann ist $\langle f, 0 \rangle$ vom Typ (B).

Wenn $\alpha = -2$, $\beta \geqq -1$ oder wenn $\alpha < -2$, dann ist $\langle f, 0 \rangle$ vom Typ (C).

Beweis. 1) Sei $c = 0$, so dass $f = 2Az + B$ (also $\varphi = (1 - B) - 2A \log p$). Dann ist $F(z) = Az^2 + Bz + C$ (C konstant), so dass

$$\log(-p') = A \log^2 p + B \log p + C,$$

$$\frac{dx}{dp} = -\frac{1}{k} e^{-A \log^2 p - B \log p} \qquad (k > 0),$$

$$x = \frac{1}{k} \int_{p}^{p_1} e^{-A \log^2 p - B \log p}\, dp + 1.$$

Dabei durchläuft p das Intervall $(0, p_1\rangle$.

Aus den bekannten Sätzen über die Konvergenz uneigentlicher Integrale folgt, dass $\lim_{p \to 0} x(p) < \infty$ genau dann, wenn $A = 0$, $B < 1$ oder $A > 0$. Im Falle $A = 0$, $B = 1$ ist $p = m e^{-kx}$ ($m > 0$ konstant), so dass $\sum_{1}^{\infty} p(i) < \infty$; im Falle $A = 0$, $B > 1$ ist $p = \frac{m}{(x + n)^{\frac{1}{B-1}}}$, so dass $\sum_{1}^{\infty} p(i) < \infty$

genau dann, wenn $B < 2$; im Falle $A < 0$ gilt, wie man sich leicht überzeugt, $\sum_1^\infty p(i) = \infty$.

2) Sei $c \neq 0$, so dass $F(z) = - \alpha z - \beta \log | cz + d | + C$, d.h.

$$\log (- p') = - \alpha \log p - \beta \log | c \log p + d | + C ,$$

$$\frac{dx}{dp} = - \frac{1}{k} p^\alpha | c \log p + d |^\beta ,$$

$$x = \frac{1}{k} \int_P^{P_1} p^\alpha | c \log p + d |^\beta \, dp + 1 .$$

Aus den Sätzen über die Konvergenz uneigentlicher Integrale folgt nun, dass $\lim_{p \to 0} x(p) < \infty$ genau dann, wenn $\alpha > - 1$ oder $\alpha = - 1$, $\beta < - 1$. Um im Falle $\sigma = \infty$ (d.h. $\alpha < - 1$ oder $\alpha = - 1$, $\beta \geq - 1$) über die Konvergenz von $\sum_1^\infty p(i)$, d.h. zwischen den Typen (B) und (C), zu entscheiden, genügt es (da p eine fallende Funktion ist), die Konvergenz des Integrals $\int_1^\infty p(x) \, dx = \int_{P_1}^0 p \frac{dx}{dp} \, dp =$ $= \frac{1}{k} \int_0^{P_1} p^{\alpha+1} | c \log p + d |^\beta \, dp$ zu untersuchen; es konvergiert dann und nur dann, wenn $\alpha > - 2$ oder $\alpha = - 2$, $\beta < - 1$.

3) Es bleibt noch übrig, festzustellen, dass unsere Funktionen f die Bedingung (Δ) erfüllen. Es sei zuerst $c \neq 0$. Es ist $f[\log (mp)] = \frac{a \log(mp) + b}{c \log (mp) + d} = \frac{a \log p + (b + a \log m)}{c \log p + (d + c \log m)}$.

Bezeichnen wir $f[\log (mp)] = f^*(\log p)$, $b + a \log m = b^*$, $d + c \log m = d^*$, $a = a^*$, $c = c^*$, $e^{-\frac{d^*}{c^*}} = \gamma^*$, so ist $b^* c^* - a^* d^* = bc - ad$, $\gamma^* = \frac{1}{m} \gamma$. Für f^* gelten also alle angeführten Konvergenzkriterien. Der Fall $c = 0$ ist trivial, da der Typ (B) nach dem obigen nur bei $a = 0$ in Frage kommt.

Beispiele:

1. Sei $c = a = 0$, $B > 1$. Schreiben wir $B - 1 = k$, so erhalten wir

$$\varphi = -k \qquad (k > 0)$$

und schliesslich

$$p = \frac{1}{(mx + n)^{\frac{1}{k}}} \qquad (m > 0,\ m + n > 0) \ . \tag{5}$$

Das sind die Mandelbrotschen Formeln, speziell im Falle $k = 1$ die Zipfsche Formel. Für $k < 1$ ist die Formel vom Typ (B), für $k \geq 1$ vom Typ (C).

2. Sei $c = a = 0$, $B = 1$, so dass

$$\varphi = 0 \ .$$

Dann erhalten wir die Formel

$$p = e^{-(mx+n)} \qquad (m > 0,\ m + n > 0) \ , \tag{6}$$

die freilich den Typ (B) besitzt.

3. Sei $c \neq 0$, $d = 0$, $a = c$, $k = -\frac{b}{a} > -1$. Dann ist $f(z) = \frac{az + b}{az} = 1 - \frac{k}{z}$, d.h.

$$\varphi = \frac{k}{\log p} \ .$$

Da in unserem Falle $\alpha = -1$, $\beta = k$ ist, erhalten wir $p' = -m^* p\,|\log p|^{-k}$, d.h.

$$\frac{p'}{p}\,|\log p|^{k} = -m^* \qquad (m^* > 0) \ , \tag{7}$$

also $-\frac{1}{k+1}\,|\log p|^{k+1} = -(m^* x + n^*)$, so dass

$$|\log p| = (mx + n)^{\frac{1}{k+1}} \ ,$$

d.h.

$$p = e^{-(mx+n)^{\frac{1}{k+1}}} \qquad (m > 0,\ m + n > 0) \ . \tag{8}$$

Es geht freilich wieder um den Typ (B).

4. Es sei wieder $c \neq 0$, $d = 0$, $a = c$, aber $k = -\frac{b}{a} =$ $= -1$, so dass

$$\varphi = -\frac{1}{\log p} \quad .$$

In diesem Falle führt (7) zu

$$\log |\log p| = m^* x + n^* \quad ,$$

d.h.

$$(9) \qquad p = e^{-e^{m^* x + n^*}} \qquad (m^* > 0)$$

(praktisch kann man auch hier $m^* + n^* > 0$ voraussetzen, da man auf Grund von empirischen Daten ruhig $p(1) < e^{-1}$ voraussetzen kann). Es handelt sich offenbar wieder um den Typ (B).

5. Sei $c = 0$, $A > 0$ - so dass man den Typ (A) erhält -, d.h.

$$\varphi = -2A \log p + C$$

(wo $C = 1 - B$). Dann ist

$$\frac{dx}{dp} = -k\, e^{-A \log^2 p - B \log p} \qquad (k > 0) \, .$$

Man findet leicht, dass wir hier den als sog. lognormalle Häufigkeitsverteilung bekannten Fall haben (wo die kumulative Häufigkeit Verteilungsfunktion einer normalen Verteilung mit der Veränderlichen $\log(np)$, $n > 0$, ist). Diese lognormalle Verteilung kann man nämlich im wesentlichen durch eine Formel

$\int_x^\xi p(x)\, dx = \int_{-\infty}^{\eta} M e^{-N(\eta - Q)^2} d\eta$ darstellen, wo $N > 0$, $\eta = \log(np)$, $\xi = \lim_{p \to 0} x(p)$. Schreiben wir diese Beziehung in der Form $\int_x^\xi p(x)\, dx = \int_\xi^x M e^{-N(\log p + \log n - Q)^2} \frac{p'}{p}\, dx$, erhalten wir durch Differenzieren

$$\frac{dx}{dp} = -k\, e^{-N \log^2 p - R \log p} \quad ,$$

wo $k > 0$, $R = 2[N(\log n - Q) + 1]$.

Die Betrachtungen werden manchmal durchsichtiger, wenn man anstatt des Ausdrucks $\frac{p}{p'}$ den Ausdruck $\frac{h}{p'} = \frac{-p \log p}{p'}$ untersucht. Es ist $\psi = \left(\frac{-p \log p}{p'}\right)' =$

$= \frac{-p'^2(1 + \log p) + pp'' \log p}{p'^2} = -(1 + \varphi \log p)$, also

$g = -1 - \psi = \varphi \log p$. Es ist dann:

$g = -k \log p \quad (k > 0)$ für die Funktionen (5),

$g = 0$ für die Funktionen (6),

$g = k \quad (k > -1)$ für die Funktionen (8),

$g = -1$ für die Funktionen (9),

$g = -2A \log^2 p + C \log p \ (A > 0)$ für die Funktionen im Beispiel 5 .

In diesem ersten Teile des Artikels werden wir uns etwas ausführlicher mit der Formel (8) befassen, die empirisch befriedigendere Ergebnisse als die Mandelbrotsche Formel liefert, wobei sie mit ihr den gemeinsamen Vorteil einer analytisch einfachen expliziten Darstellung von $p(x)$ besitzt.

Bemerken wir, dass zur Formel (8) auch die folgende Betrachtung führt. Die Mandelbrotschen Funktionen sind durch die Beziehung $\varphi = -k$ charakterisiert; die Erfahrung zeigt, dass diese Formeln (bei einer geeigneten Wahl des Exponenten $\frac{1}{k}$, der aber nicht überall > 1 sein muss) den Häufigkeitsverlauf gewissermassen befriedigend nur in ziemlich beschränkten Intervallen charakterisieren und dass man mit wachsendem x immer grössere Exponente $\frac{1}{k}$ braucht. Eine einfache Möglichkeit besteht nun darin, die Bedingung $\varphi = -k$ durch die Bedingung $\varphi = -\frac{k^*}{|\log p|}$ zu ersetzen. Es muss allerdings gesagt werden, dass

für sehr grosse Werte von x auch die Formel (8) immer noch ein noch zu langsames Sinken von p liefert (was auch damit zusammenhängt, dass sie vom Typ (B) ist); in einer Fortsetzung dieses Artikels werden wir (8) durch eine Formel vom Typ (A) ersetzen, die als eine verhältnismässig befriedigende Lösung der Frage der Häufigkeitsverteilung der Worte betrachtet werden kann.

Von den Häufigkeitsstatistiken, an den ich die Formel (8) verifizierte, möchte ich hier diejenigen zitieren, wo ich das vollständigste Material zur Verfügung hatte:

1) Die unter Leitung von J. Jelínek, J.V. Bečka und M. Těšitelová bearbeitete Statistik der Worte als lexikaler Einheiten, aus Texten mit Gesamtumfang von über 1 600 000 Worten. Veröffentlicht in der Arbeit [1] .

Die Statistiken 2) und 3), die ich weiter zitiere, wurden zu stenographischen Zwecken bearbeitet, wie denn überhaupt die Häufigkeitsuntersuchungen ihre älteste Tradition und auch die vielseitigste Anwendung bisher auf dem stenographischen Gebiete besitzen. Die in diesen Statistiken verwendeten Begriffe des Wortes usw. sind freilich den stenographischen Zwecken angepasst (das Material wird sozusagen nicht vom Standpunkt des Wörterbuches der Sprache, sondern mehr vom Standpunkt des Wörterbuches stenographischer Kürzungen bearbeitet). Der Begriff des Wortes wird dabei nur wenig modifiziert, jedoch deckt sich s.B. der für stenographische Zwecke verwendete Begriff des Wortstammes nur zum Teil mit dem in der Sprachwissenschaft üblichen; vom Standpunkt der Untersuchung der quantitativen Gesetzmässigkeiten des Häufigkeitsverlaufs spielen aber diese Unterschiede keine wesentliche Rolle.

2) Die unter Leitung von F.W. Kaeding bearbeitete Statistik

der einzelnen Wortformen (und mancher anderer sprachlichen Erscheinungen) in der deutschen Sprache, aus Texten mit Gesamtumfang von fast 11 Millionen Worten. Veröffentlicht in der Arbeit [2].

3) Die im Staatlichen stenographischen Institut in Prag (in Zusammenarbeit mit der Mathematisch-physikalischen Fakultät der KU) bearbeitete Statistik aus Texten mit Gesamtumfang von 120 000 Worten, geteilt in zwei thematische Gruppen A und B zu je 60 000 Worten; es wurde vor allem bearbeitet:

3,1) die Statistik der Worte als lexikaler Einheiten (Teil A von K. Matoušek und M. Matula, Teil B von J. Čáp und J. Petrásek);

3,2) die Statistik der einzelnen Wortformen (Teil A von K. Matoušek und M. Matula, Teil B von J. Čáp und J. Petrásek);

3,3) die Statistik der Wortgruppen, vor allem zweigliedriger Gruppen von Worten als lexikaler Einheiten (nur im Teil A, von K. Matoušek und M. Matula);

3,4) die Statistik der Wortstämme (einschliesslich Vorsilben) (Teil A und B , von M. Matula).

Die Ergebnisse von 3,1), 3,2) wurden veröffentlicht in der Arbeit [3 a] , die Ergebnisse von 3,4) in der Arbeit [3 b]. Andere Angaben für 3) habe ich aus meinen handschriftlichen Materialen geschöpft. Alle in diesem Artikel enthaltenen Angaben für 3) beziehen sich nur auf den Teil A .

Die Konfrontation der Formeln mit empirischem Material werde ich hier an zwei Beispielen - der Statistiken 1) und 3,2) - illustrieren.

Es sei $r > 0$. Vorausgesetzt, dass die Formel (5) rich-

tig ist, muss das Verhältnis

$$\frac{p(rx)}{p(x)}$$

für nicht zu kleine x, wo man die Konstante n nicht mehr zu berücksichtigen braucht, konstant sein; demgegenüber muss das Verhältnis

$$\frac{\log p\,(rx)}{\log p(x)}$$

konstant sein, wenn die Formel (8) gilt. Diese Tatsache ermöglicht eine bequeme Verifikation beider Formeln.

Ich möchte hier zuerst eine Tafel anführen, die sich auf die Häufigkeit der Worte als lexikaler Einheiten bezieht, und zwar auf Grund der Statistik von Jelínek-Bečka-Těšitelová; es wurde $r = 2$ gewählt und die Zahlen p in der Tafel werden natürlich relative Häufigkeiten in der Statistik bedeuten. Um wenigstens die gröbsten Fälle von zufälligen Schwankungen der Häufigkeit zu beschränken, habe ich die Werte von $p(x)$ durch die Durchschnitte $\frac{p(x-1) + p(x) + p(x+1)}{3}$ für $x \leqq 200$ und durch $\frac{p(x-2) + p(x-1) + p(x) + p(x+1) + p\,(x+2)}{5}$ für $200 < x < 1600$ ersetzt. Für $x \geqq 1600$, wo in der Statistik immer schon mehrere Worte mit derselben Häufigkeit vorkommen, habe ich statt $p(x)$ die folgendermassen berechneten Werte genommen: wenn $f(x)$ die absolute Häufigkeit des x-ten Wortes in der Statistik bedeutet und wenn $f(x)$ konstant bleibt für $x_1 \leqq x \leqq x_2$, habe ich anstatt $f(x)$ den durch lineare Interpolation zwischen den Werten $f(x_1) - \frac{1}{2}$, $f(x_2) + \frac{1}{2}$ im Intervall $\langle x_1 - \frac{1}{2},\ x_2 + \frac{1}{2} \rangle$ erhaltenen

Wert genommen. Die Tafel lautet:

x	log p(2x):log p(x)	p(x):p(2x)
10	1,082	1,492
25	1,127	2,010
50	1,108	1,946
75	1,120	2,193
100	1,110	2,120
125	1,102	2,059
150	1,087	1,892
175	1,082	1,849
200	1,077	1,790
225	1,080	1,844
250	1,077	1,822
300	1,075	1,816
350	1,075	1,835
400	1,077	1,886
450	1,076	1,876
500	1,075	1,873
600	1,077	1,936
700	1,076	1,928
800	1,075	1,931
900	1,076	1,971
1000	1,079	2,033
1200	1,078	2,060
1400	1,080	2,116
1600	1,082	2,179
1800	1,082	2,207
2000	1,082	2,214

2400	1,082	2,263
2800	1,083	2,320
3200	1,082	2,308
3600	1,082	2,340
4000	1,082	2,376
4800	1,084	2,464

Zweitens möchte ich eine Tafel anführen, die sich auf die Häufigkeit der einzelnen Wortformen bezieht, auf Grund der Statistik von Matoušek und Matula. Es wurde wieder $r = 2$ gewählt und die Zahlen $p(x)$ wurden wieder durch ähnlich modifizierte Zahlen ersetzt (die oben beschriebene lineare Interpolation wurde hier aber - in Hinsicht auf den kleineren Umfang der Statistik - für $x \geqq 300$ angewandt). Die Tafel lautet:

x	log p(2x):log p(x)	p(x):p(2x)
10	1,139	2,007
25	1,068	1,489
50	1,073	1,583
75	1,097	1,882
100	1,099	1,947
125	1,090	1,868
150	1,082	1,800
175	1,084	1,843
200	1,081	1,828
225	1,079	1,816
250	1,080	1,836
300	1,076	1,800
350	1,075	1,813

400	1,076	1,846
450	1,077	1,863
500	1,077	1,883
600	1,078	1,926
700	1,077	1,924
800	1,077	1,939
900	1,076	1,947
1000	1,077	1,968
1200	1,078	2,025
1400	1,076	2,003
1600	1,077	2,053

Da ich hier die empirischen Zahlen nur zu einer informativen Orientierung anführe, habe ich mich hier nur auf einen anschaulichen Vergleich empirischer Werte in den obigen Tafeln beschränkt. Das genügt , um einzusehen, dass die Formel (8) befriedigende Ergebnisse in einem wesentlich breiteren Intervall liefert.

Man kann natürlich nicht erwarten, dass eine Formel mit eigentlich nur einem wesentlichen Parameter, wie es die Formel (8) oder (5) ist (mit dem wesentlichen Parameter k), den Häufigkeitsverlauf in den konkreten Statistiken in allen Einzelheiten treu wiedergibt. Namentlich aber kann man bei keiner Formel eine zu gute allgemeine Übereinstimmung mit empirischen Daten erwarten für die kleinsten Werte von x , d.h. für grösste Häufigkeiten, da die empirischen Zahlen für die grössten Häufigkeiten stark von der Struktur der Sprache abhängen (vor allem kommen hier die mit den sogenannten formalen Worten zusammenhängenden Unregelmässigkeiten in Betracht)

und natürlich auch verschiedene zufällige Momente aufweisen.

Ich möchte noch einiges über den Wert des Exponenten $\frac{1}{k+1}$ bemerken, wenn man versucht, den Häufigkeitsverlauf einiger sprachlichen Einheiten - wie der Wortformen, der Worte als lexikaler Einheiten, der Wortgruppen, der Wortstämme u.ä. - durch unsere Formel darzustellen. Das ist allerdings immer nur in einem (praktisch verhältnismässig breiten) Intervall möglich; im Gebiete der "mittelgrossen" Häufigkeiten kann man folgende ungefähre Zahlen anführen:

Statistik	Einheiten		$\frac{1}{k+1}$
3,3)	zweigliedrige Wortgruppen (der Worte als lexikaler Einheiten)		0,08
3,2)	einzelne Wortformen		0,11
3,1)	Worte als lexikale Einheiten		0,14
3,4)	Wortstämme	über	0,20
1)	Worte als lexikale Einheiten	über	0,11

Es scheint, dass der Wert des Exponenten vor allem von der durchschnittlichen Länge der verwendeten Einheiten abhängt; diese Vermutung wird auch in einer Fortsetzung dieses Artikels bestätigt. Man kann diesen Wert auch mit dem Umfang des bearbeiteten Materials, mit dem Reichtum des in ihm enthaltenen Wörterbuches und mit dem Grad der Zerlegung des Textes auf verschiedene Einheiten in Zusammenhang bringen. Was den letzteren Umstand betrifft, so bedenke man, dass man einen Wortstamm in einem gewissen Sinne als die Menge der Worte auffassen kann, die den Stamm enthalten, dass man ein Wort A (im Sinne einer

lexikalen Einheit) als die Menge seiner einzelnen Wortformen A_i auffassen kann, und dass die Anzahl zweigliedriger Wortgruppen $\langle A, X \rangle$ und $\langle X, A \rangle$, die das Wort A (im Sinne einer lexikalen Einheit) enthalten, offenbar im allgemeinen grösser ist als die Anzahl seiner einzelnen Wortformen A_i.

In einem (praktisch ziemlich breiten) Intervall, wo die Formel (8) empirisch befriedigende Ergebnisse liefert, kann man wahrscheinlich eine verhältnismässig befriedigende Approximation des Häufigkeitsverlaufs durch Mandelbrotsche Funktionen $\frac{1}{(mx + n)^{\frac{1}{\varphi}}}$ in der Nähe von den x erwarten, für die $-\varphi = \frac{k}{\log p}$, d.h. $\varphi = \frac{k}{|\log p|}$ gilt. Andererseits ist es ersichtlich, dass man grössere, bzw. kleinere Werte für den Exponenten $\frac{1}{\varphi}$ bei den Statistiken erwarten kann, denen grössere, bzw. kleinere Werte des Exponenten $\frac{1}{k + 1}$ entsprechen.

Im weiteren erwähnen wir noch einige einfache Eigenschaften der Funktionen (8) (bzw. verwandte Eigenschaften anderer Funktionen). In den Formulationen dieser Eigenschaften werden wir die Funktionen $p(x)$ als zulässig bezeichnen, die für $1 \leqq x < \infty$ definiert sind und stetige zweite Ableitung besitzen und für die $0 < p < 1$, $p'(x) < 0$, $\lim\limits_{x \to \infty} p(x) = 0$, $\sum\limits_{1}^{\infty} p(i) < \infty$ gilt (so dass es sich um den Typ (B) mit $\gamma = 0$ handelt).

Es sei $h = h(x) = - p \log p$ (wo das Symbol log stets den natürlichen Logarithmus bedeutet), $H = H(x) = \sum\limits_{1}^{x} h$, $P = P(x) = \sum\limits_{1}^{x} p$; es sei weiter $H^* = H^*(x) = \sum\limits_{x}^{\infty} h$, $P^* = P^*(x) = \sum\limits_{x}^{\infty} p$, $h^* = h^*(x) = - P^* \log P^*$.

Es ist ersichtlich, dass man, grob gesagt, die Summen $\sum_x$ (bei den Forderungen auf Genauigkeit, die für die quantitative Linguistik noch einen praktischen Sinn haben) für nicht zu kleine x durch Integrale $\int_x^\infty$ approximieren kann; ich möchte den Umfang des Artikels nicht durch eine genauere Präzisierung solcher Umstände belasten.

1) Die Funktionen (5),(6) werden unter den zulässigen Funktionen durch die Eigenschaft charakterisiert, dass $(\frac{p}{p'})'$ konstant ist.

2) Die Funktionen (8),(9) werden durch die Eigenschaft charakterisiert, dass $(\frac{h}{p'})'$ konstant ist.

Das haben wir schon oben gesehen.

3) Die Funktionen (5) (bei $k < 1$) und (6) besitzen die Eigenschaft

$$q H^* - h^* = c P^* \quad (10)$$

wo c konstant ist und $q = 1$ für Funktionen (6), $q = 1 - k$ für Funktionen (5) gilt.

Es gilt nämlich offenbar $q = 1 + \varphi = 2 - \frac{pp''}{p'^2}$. Hieraus folgt $(\frac{p^2}{p'})' = q\,p$ und daher $\frac{p^2}{p'} = q \int_1^x p + A$ (A konstant). Für die Funktionen (6) und die Funktionen (5) (bei $k < 1$) gilt $\frac{p^2}{p'} \to 0$ für $x \to \infty$ und daher $A = -q \int_1^\infty p$. Es gilt also $q \frac{p'}{p} = -\frac{p}{\int_x^\infty p}$. Durch Integration erhalten wir $q \log p = \log \int_x^\infty p + B$ (B konstant). Hieraus folgt $q\,p \log p - p \log \int_x^\infty p - p = (B - 1)\,p$; da $-p \log \int_x^\infty p - p = (\int_x^\infty p \cdot \log \int_x^\infty p)'$, erhalten wir durch Integration

$$-q\int_x^\infty p \log p + \int_x^\infty p \cdot \log \int_x^\infty p = (1-B)\int_x^\infty p + D$$

(D konstant); es gilt offenbar $D = 0$. Wenn wir noch $1 - B = c$ setzen, genügt es, von den Integralen zu den entsprechenden Summen zu übergehen, um (10) zu erhalten.

Es sei hier nur kurz bemerkt, dass man in der Praxis $B < 0$, d.h. $c > 1$ voraussetzen kann.

Man sieht leicht, dass umgekehrt die Eigenschaft (10) (genauer gesagt, die für Integrale geltende Eigenschaft) für die Funktionen (5) und (6) charakteristisch ist.

Man kann diese Eigenschaft (nur ziemlich annähernd) auch folgendermassen interpretieren. Es lässt sich feststellen (was hier nur bemerkt sei), dass die Gültigkeit der Beziehung (10) auch für die kleinsten Werte von x praktisch nicht zu viel gestört wird, so dass sich $q \sum_1^\infty h$ nicht viel von $c \sum_1^\infty p = c$ unterscheidet. Daher unterscheidet sich $q\,H(x) + h^*(x+1)$ nicht viel von $c\,P(x)$. Für die Funktionen (6), d.h. $q = 1$, kann man das ungefähr folgendermassen ausdrücken: wenn wir vom ganzen Text nur die ersten x häufigsten Worte behalten und den Rest durch ein "leeres Symbol" ersetzen, dann ist die Gesamtinformation eines solchen "unvollständigen Textes" dem Umfang des "bekannten Teiles" des Textes annähernd proportional. Für Mandelbrotsche Funktionen könnte man die Eigenschaft auf eine ähnliche Weise interpretieren, wenn man annehmen wollte, dass "dem leeren Symbol eine modifizierte Information zugesprochen wird, die man aus der "normalen" durch Multiplizieren mit einem konstanten Faktor erhält".

4) Die Zipfschen Funktionen besitzen die Eigenschaft, dass das Produkt $(mx + n)\,p$ (bei geeigneten m, n) konstant bleibt. Für die Funktionen (8),(9) gilt demgegenüber

(11) $H^* - (k+1)P^* \doteq ((k+1)x + q)p$ (q konstant).

Für die Funktionen (8) kann man diese Eigenschaft so formulieren, dass das Produkt $(mx + n)p$ (bei geeigneten m, n) annähernd gleich ist der Differenz zwischen der Gesamtinformation und der Gesamthäufigkeit für die Menge der Worte, die hinter dem x-ten Worte liegen; bei einer geeigneten Wahl der Informationseinheit ist dann nämlich offenbar

$$H^* - P^* \doteq (x + s)p \qquad (s \text{ konstant}).$$

Für die Funktionen (9) ist $k + 1 = 0$, so dass eine Interpretation klar ist.

Um (11) abzuleiten, gehen wir davon aus, dass nach der Eigenschaft 2)

$$\frac{h}{|p'|} = \frac{p \log p}{p'} = (k+1)x + q \qquad (q \text{ konstant})$$

ist, also

$$p \log p = (k+1)xp' + qp' = ((k+1)xp)' - (k+1)p + qp',$$

und daher $\int_1^x p \log p + (k+1)\int_1^x p - qp = (k+1)xp + A$ (A konstant). Für $x \to \infty$ ist $p \to 0$, so dass $A = \int_1^\infty p \log p + (k+1)\int_1^\infty p$. Nun genügt es zu bemerken, dass man die Integrale $\int_x^\infty$ als Approximationen der Summen $-H^*$, bzw. P^* betrachten kann, um (11) zu erhalten.

Man sieht leicht, dass umgekehrt die Eigenschaft (11) (genauer gesagt, die für Integrale geltende Eigenschaft) für die Funktionen (8),(9) charakteristisch ist.

Wir haben oben gesehen, dass der Bruch $\frac{h}{p'}$ ungefähr proportional ist der Anzahl W der neuen verschiedenen Worte, wenn die "lokale" durchschnittliche Wortlänge $q(x)$ um einen konstanten Zuwachs dq wächst. Da man offenbar p als die durchschnittliche relative Häufigkeit dieser neuen Worte be-

trachten kann, ist $\frac{h}{p'}\ p$ ungefähr proportional dem Zuwachs des Textumfanges. Diese Tatsache liefert eine Interpretation für die rechte Seite der Beziehung

$$H^* - (k+1)\,p^* \doteq \frac{h}{|p'|}\ p \quad ,$$

die gemäss 2),4) für die Funktionen (8),(9) gilt.

Es sei noch eine Bemerkung angeführt, die manchmal nützlich sein kann. Bezeichnen wir wieder $\log p = z$, $\frac{pp''}{p'^2} = f$. Bezeichnen wir weiter

$$f_h = \frac{hh''}{h'^2} \quad .$$

Da $h = - p \log p$, errechnet man leicht, dass

$$(12) \qquad f_h = f \cdot \frac{z}{1+z} + \frac{z}{(1+z)^2}$$

und umgekehrt

$$(13) \qquad f = f_h \cdot \frac{1+z}{z} - \frac{1}{1+z} \quad .$$

Beispiele:

6. Die Funktionen (6) besitzen, wie man leicht verifiziert, die Eigenschaft, dass für $s > 0$ gilt

$$p(x) = a \left(\sum_x^\infty p^s \right)^{\frac{1}{s}} \quad ,$$

wo $a = \frac{1}{(1 - e^{-ms})^{\frac{1}{s}}}$. Daraus folgt trivial, dass die Eigenschaft

$$h(x) = a \left(\sum_x^\infty h^s \right)^{\frac{1}{s}}$$

für die Funktionen p gilt, für die $h = - p \log p = e^{-(mx+n)}$. Für diese Funktionen ist $f_h = 1$, so dass nach (13) diese Funktionen die zu

$$f = \frac{z^2 + z + 1}{z^2 + z} = 1 + \frac{1}{z} - \frac{1}{z+1}$$

gehörenden Lösungen sind.

7. Das vorhergehende Beispiel führt zu der Frage, für welche p die Beziehung

$$h(x) \sim q \left(\sum_x^\infty p^s \right)^{\frac{1}{s}}$$

erfüllt wird; genauer gesagt, die Beziehung $h = q \left(\int_x^\infty p^s \right)^{\frac{1}{s}}$, wo q, s positive Konstanten sind. Durch Differenziation erhält man $\frac{p'}{p}(1 + \log p) \, |\log p|^{s-1} = \frac{q^s}{s}$ (wir setzen voraus, dass $1 + \log p < 0$); hieraus folgt

$$\frac{1}{s+1} \, |\log p|^{s+1} - \frac{1}{s} \, |\log p|^s = \frac{q^s}{s} \, x + c$$

(c konstant) .

Man verifiziert leicht, dass solche Funktionen p die zu

$$f = \frac{z^2 - (s-1)\,z - (s-1)}{z^2 + z} = 1 - \frac{s-1}{z} - \frac{1}{z+1}$$

gehörenden Lösungen sind. Nach (12) errechnet man

$$f_h = \frac{z - (s-1)}{z+1} = 1 - \frac{s}{z+1} \, .$$

Zitierte Literatur:

[1] J. JELÍNEK, J.V. BEČKA, M. TĚŠITELOVÁ: Frekvence slov, slovních druhů a tvarů v českém jazyce, Praha, 1961.

[2] F.W. KAEDING: Häufigkeitswörterbuch der deutschen Sprache, Steglitz, 1898.

[3a] J. ČÁP, K. MATOUŠEK, M. MATULA, J. PETRÁSEK: Frekvence

slov v stenografické praxi, Praha, 1961.

[3 b] M. MATULA: Frekvence kořenů slov, Praha, 1963.