Auf den amerikanischen Philologen Zipf (1902-1950) gehen eine Reihe statistischer Gesetze zurück.

1. So erkannte...

... er den direkten Zusammenhang zwischen dem Rang (r) eines Wortes in einer Häufigkeitsliste und der Häufigkeit (h), mit der es in einem beliebigen Text vorkommt. Es handelt sich um eine konstante lineare Beziehung: r x h = C.

Die Gültigkeit des Gesetzes läßt sich leicht überprüfen:

1. Man zählt die Häufigkeit verschiedener Wörter in einem Text: z.B. der 364, ist 251, Tisch 4, usw. und ordnet sie nach der Häufigkeit, bei Numerierung aller Stellen: (1) der 364, (2) ist 251, (3) von 166, usw.
2. Eine Multiplikation der Stellennummer, d.h. des Ranges (r) mit der Häufigkeit (h) ergibt: Das Resultat (C) ist annähernd konstant.

Beispiel: In der rechts stehenden Liste sind z.B. die Wörter aufgeführt, die in der Kategorie für gesprochene Sprache an 35., 45., 55., 65. und 75. Stelle (= r) der Häufigkeitshierarchie stehen.
Die Häufigkeit in einem bestimmten (beliebigen) Text ist (h).
Das Produkt von (r) und (h) beträgt jeweils rund 30 000.

r	x	h	C
35	*very*	836	29 260
45	*see*	674	30 330
55	*which*	563	30 965
65	*get*	469	30 485
75	*out*	422	31 650

Es besteht also ein umgekehrt proportionales Verhältnis: Je seltener ein Wort, um so höher sein Rangstelle (r).

Zunächst nahm man an, daß dies unabhängig von Thematik, Autor oder anderen sprachlichen Variablen gälte. Später zeigte sich jedoch, daß das Verhältnis für besonders häufige und besonders seltene Wörter nicht zutrifft. Im selben Korpus tritt zum Beispiel das häufigste Wort I (»ich«) 5.920mal auf (r x h = 5920), und das 100. Wort be's (»er ist«) kommt 363mal vor (r x h = 36 300).

Der Umfang des Korpus ist ebenfalls ein entscheidender Faktor. Dennoch ist die »Standardkurve« der Worthäufigkeit - r x h = C - eine interessante Darstellung sprachlicher Muster.

Ähnliche Kurven wurden in vielen Sprachen festgestellt. So erscheint nach einem französischen Buch über Worthäufigkeiten das 100. Wort 314mal (= 31.400), das 200. 158mal (=31.600) und das 1.000. nur 31mal (= 31.000). Der Faktor C scheint allgemein bei etwa 30.000 zu liegen.

2. Zipf erkannte...

... außerdem, daß ein umgekehrter Zusammenhang zwischen der Länge eines Wortes und seiner Häufigkeit besteht. So sind im Englischen die meistverwendeten Wörter fast ausnahmslos Einsilber. Das gleiche Verhältnis trifft auch auf das Deutsche mit seinem hohen Anteil an mehrsilbigen Wörtern zu.

Dieses Phänomen scheint unsere Neigung widerzuspiegeln, Wörter bei häufigerem Gebrauch abzukürzen: Disko statt Diskothek, Abi statt Abitur, usw.

Für eine wirkungsvolle Kommunikation scheint es zweckmäßig zu sein, wenn die häufigen Wörter kurz und die selteneren lang sind.

Kriterien wie Effizienz und Erleichterung der Kommunikation waren für Zipf von großer Bedeutung. Er erklärte den genannten Zusammenhang in unserem Gebrauch Wörtern mit dem Prinzip der „geringsten Anstrengung“: Je einfacher der Laut und je kürzer das Wort, desto häufiger und kräftiger sei der Impuls, es zu verwenden.

3. Praktische Anwendung

Forscher: Sprache entstand in der Evolution sprunghaft, nicht kontinuierlich

Die Zipf’schen Gesetze haben in der linguistischen Forschung eine große Bedeutung. So haben zwei spanische Forscher von der Universität Barcelona Anfang des Jahres 2003 unter Anwendung des Zipf’schen Gesetzes und seiner Theorie der geringsten Anstrengung verblüffende Ergebnisse erzielt.

Hierbei gingen sie von folgender Überlegung aus: Dem Zuhörer ist eine Mitteilung dann leicht verständlich, wenn der Sprecher eindeutige Worte verwendet und Zweideutigkeiten vermeidet. Beim Sprecher erfordert das einen umfangreichen Wortschatz, der zudem jederzeit aus dem Gedächtnis abrufbar sein muß. Je weniger Worte der Redende dagegen verwendet, desto anstrengender wird es für den Zuhörer, der dann eine beträchtliche Interpretationsleistung erbringen muß.

Die beiden Wissenschaftler (ein Physiker und ein Computerfachmann) haben nun in einem mathematischen Modell untersucht, welche Konsequenzen das Prinzip der geringsten Anstrengung für den Gebrauch von Worten hat. Hierbei berechneten sie für verschieden umfangreiche Sprachen – angefangen mit einer aus einem einzigen Grunzlaut bestehenden Sprache bis hin zu Sprachen mit einem unerschöpflichen Wortschatz –, welche Verteilung der Worte sich jeweils ergibt.

Ihr Gedanke war: Die tatsächliche Wortverteilung ist ja aufgrund der erwähnten Häufigkeitslisten bekannt. Die Verteilung der Worte in einem beliebigen Text in einer beliebigen Sprache folgt immer annähernd dem Zipf’schen Gesetz. Demnach kommt das zweithäufigste Wort in einem Text etwa halb so oft vor wie das häufigste, das dritthäufigste ein Drittel mal so oft, usw.. Hinter dieser Verteilung vermutete Zipf (1902 - 1950) schon damals ein Prinzip der geringsten Anstrengung, konnte es aber nicht beweisen.

Das Ergebnis der spanischen Forscher ist verblüffend: Eine Verteilung der Worthäufigkeiten, die dem Zipf’schen Gesetz folgt, ergibt sich nur für Sprachen, die dem Prinzip der geringsten Anstrengung genügen, die also möglichst wenig Anstrengung sowohl von Redner als auch vom Zuhörer verlangen.

Außerdem fanden die Wissenschaftler bei ihren Simulationen heraus, daß der Übergang von einer aus wenigen Lauten bestehenden Sprache zu einer sehr komplexen Sprache sprunghaft geschieht.

Nur für einen sehr schmalen Übergangsbereich gilt das Zipf’sche Gesetz. Dies spricht nach Meinung der beiden Spanier dafür, daß die menschliche Sprache während der Evolution nicht über einen längeren

Zeitraum mit einem kontinuierlich anwachsenden Wortschatz entstand, sondern relativ plötzlich auftauchte.

Nach der Interpretation der beiden Wissenschaftler ist die menschliche Sprache somit das Ergebnis eines Ausgleichs der Interessen von Redner und Zuhörer.

Näheres zu dem mathematischen Modell: Proceedings of the National Academy of Sciences (Bd. 100, Nr. 3, S. 788).