Publié dans Bulletin CILA (Commission interuniversitaire suisse de linguistique appliquée
- Schweizerische Hochschulkomission für angewandte Sprachwissenschaft)
(« Bulletin VALS-ASLA » depuis 1994) 22, 38-44, 1975
qui doit être utilisé pour toute référence à ce travail

Worthäufigkeiten, Wortlängen und Buchstabenhäufigkeiten

Als Vorstudie für die Lesbarkeit von Texten unternahmen wir eine Untersuchung von Texten bezüglich Worthäufigkeiten, sowie ihrer Buchstaben- und Buchstabenfolgehäufigkeit. Texte mit durchschnittlich hohen Häufigkeiten könnten, so wäre etwa die Hypothese, infolge ihres Bekanntheitsgrades leichter gelesen werden. Doch wäre es auch denkbar, über bekannte Wörter beim Lesen leichter zu "stolpern".

Die Analyse der Buchstaben- und Buchstabenfolgehäufigkeiten und der Worthäufigkeiten ermöglicht aber auch Einblicke in das Sprachverhalten schlechthin.

1. Die Texte

Sechzehn Texte (Kurzgeschichten aus Zeitschriften des Ringier-Verlages) wurden zufällig ausgewählt. Der Verlag lieferte die Texte auf Lochstreifen, so dass sie zur statistischen Bearbeitung direkt in den Computer eingegeben werden konnten.

2. Worthäufigkeiten

Die Resultate der Wortauszählung sind in Tab. 1 wiedergegeben. Die Texte wurden in der Reihenfolge ihrer *Länge* (Gesamtzahl der ausgezählten Wörter) angeordnet und numeriert.

Als *Wortvorrat* bezeichnen wir im folgenden die Anzahl der verschiedenen Wörter pro Text. Vergleicht man die Werte der zweiten und dritten Spalte, so stellt man zunächst fest, dass der jeweils aktivierte Wortvorrat im grossen und ganzen mit steigender Textlänge ebenfalls zunimmt, jedoch nicht proportional. Dies ersieht man aus der vierten Spalte, in der die *Diversifikationsquotienten* (Quotient aus Wortvorrat und Textlänge) angegeben sind. Die höheren Diversifikationswerte finden sich bei den Texten mit geringerem Umfang, der kleinste Wert (0,35) wurde beim längsten Text festgestellt.

Diese Zahlen sind statistischer Ausdruck der Tatsache, dass bei einzelnen Autoren die *aktiven Wortvorräte* nach oben beschränkt sind, nicht aber (zumindest prinzipiell) die Textlängen. Bei kurzen Texten, deren Länge unterhalb des aktiven Wortschatzumfanges liegt, könnte der Diversifikationsquotient in gewissen Fällen als Mass für die Lebendigkeit des Stils angesehen werden; denn er gibt an, wie intensiv die Ausdrücke variiert werden bzw. wie stark Wortwiederholungen vermieden werden. Überschreitet die Textlänge

Tabelle 1: Charakteristika der 16 Texte

$$\text{Diversifikationsquotient} = \frac{\text{Tot. versch. Wörter}}{\text{Tot. Wörter}}$$

Text Nr.	Total der Wörter pro Text	Total verschiedener Wörter	Diversifikationsquotient
1	687	388	0,56
2	734	423	0,58
3	750	438	0,58
4	866	504	0,58
5	882	497	0,56
6	907	507	0,56
7	1071	472	0,45
8	1107	472	0,43
9	1122	536	0,48
10	1223	667	0,55
11	1235	603	0,49
12	1249	653	0,52
13	1251	527	0,42
14	1480	801	0,54
15	1886	941	0,50
16	3955	1376	0,35
	20405		

den aktiven Wortschatzumfang (er hängt vom Bildungsgrad und von der Schreibfertigkeit ab und liegt nach Schätzwerten zwischen 1000 und 3000 Wörtern), so ist zu erwarten, dass der Diversifikationsquotient sich nahezu umgekehrt proportional zur Textlänge verhält; denn nach dem Ausschöpfen des aktiven Wortschatzes werden Wortwiederholungen unvermeidlich. Dabei ist jedoch zu bedenken, dass der jeweils gebrauchte Wortvorrat auch durch den *Inhalt* des Textes (auf den hier nicht eingegangen wird) auf einen thematisch bedingten Teilbereich beschränkt werden kann. Und das bedeutet umgekehrt auch, dass der während einer bestimmten Textlänge benutzte Wortvorrat in der Fortsetzung allein durch eine Veränderung des zu beschreibenden Gegenstandes oder Handlungsablaufes erweitert werden kann.

Wir fassen diese Überlegung zusammen: Der Diversifikationsquotient nimmt in der Regel mit steigender Textlänge ab. Bei sehr langen Texten kann approximativ umgekehrte Proportionalität erwartet werden:

$$\text{Diversifikationsquotient} = \frac{\text{Konstante}}{\text{Textlänge}}$$

Bei kürzeren Texten kann der Diversifikationsquotient allenfalls als Mass für die Lebendigkeit und Farbigkeit des Sprachstils interpretiert werden, wobei man sich jedoch bewusst sein muss, dass die Zahlenwerte auch von

semantischen Faktoren abhängen, die sich einer elementarstatistischen Analyse entziehen.

3. Die Wortlängen

Die Texte können auch nach Wortlängen (Anzahl Buchstaben pro Wort) untersucht werden. Die entsprechenden Ergebnisse sind in Tab. 2 angegeben.

W. Fucks (1968) hat in eindrücklicher Weise gezeigt, dass die Beschaffenheit der Wörter (Anzahl Silben pro Wort) und die Länge der Sätze (Anzahl Wörter pro Satz) bei Prosadichtern und -schriftstellern ein Mittel zur Stilcharakterisierung geben. Auch bei den Wortlängen, wie wir sie verstehen, nämlich Anzahl Buchstaben pro Wort, treten bei den verschiedenen Texten Abweichungen auf. Es darf nicht ausgeschlossen werden, dass sich auch in ihnen gewisse Stilmerkmale niederschlagen, doch es ist anzunehmen, dass dies nicht so deutlich wie bei Fucks (1968) ausfallen wird; und zwar einfach deshalb, weil die Struktur und der Aufbau der formulierten Sätze dem jeweiligen Autor ungleich viel mehr individuelle Freiheitsgrade als die Beschaffenheit der einzelnen Wörter einräumen, die ja im Vokabular der Umgangssprache bereits festgelegt sind und dem Schreiber im wesentlichen nur noch die Freiheit der Auswahl, nicht aber der Strukturierung lassen. Mit andern Worten: Die Häufigkeit und Verteilung der Wortlängen ist weniger ein intraindividuelles Stilmerkmal als vielmehr ein interindividuelles Strukturmerkmal der Umgangssprache. Aufgrund dieser Überlegung richten wir unser Interesse auf die Mittelwerte, die in der untersten Reihe von Tab. 2 angegeben sind.

(Man bemerkt übrigens bereits beim ersten Überblick, dass die Werte des längsten Textes – Text 16 – verglichen mit den andern sehr nahe bei den Mittelwerten liegen, was die Vermutung unterstützt, dass die Verteilung der Wortlängen bei sehr langen Texten konvergieren würden.)

Auffallend ist die Häufung bei den dreibuchstabigen Wörtern, ihr Anteil beträgt im Mittel 25,1 Prozent. Dafür gibt es verschiedene Erklärungsmöglichkeiten, die hier jedoch nicht vollständig erörtert werden sollen. Wir greifen lediglich einen speziellen Aspekt heraus und diskutieren dieses Phänomen unter einem rein kodierungstheoretischen Gesichtspunkt, d.h. wir fassen die Wörter schlicht als Kombinationen (Superzeichen) von Buchstaben (Elementarzeichen) auf und untersuchen ihren subjektiven *Informationsgehalt*. Wir folgen damit einem Ansatz, den *von Cube* (1968) beschrieben hat. Seine Grundgedanken sollen hier nicht ausführlich wiedergegeben, sondern bloss skizzenartig rekapituliert werden:

Tabelle 2: Anteile von Wörtern verschiedener Länge für alle 16 Texte (≥ 2 %)

Text	Anzahl Buchstaben / Wort in absoluten Werten und Prozenten																							
	2		3		4		5		6		7		8		9		10		11		12		13	
1	53	7,7	158	23,0	106	15,4	76	11,1	66	9,6	43	6,3	46	6,7	20	2,9	25	3,6	28	4,1	–	–	–	–
2	53	7,2	190	25,9	81	11,0	74	10,1	64	8,7	57	7,8	27	3,7	36	4,9	30	4,1	33	4,5	–	–	–	–
3	50	6,7	191	25,5	96	12,8	113	15,1	64	8,5	44	5,9	38	5,1	32	4,3	35	4,7	20	2,7	22	2,9	–	–
4	77	8,9	198	22,9	113	13,1	99	11,4	82	9,5	70	8,1	36	4,2	50	5,8	27	3,1	28	3,2	19	2,2	20	2,3
5	62	7,0	209	23,7	118	13,4	119	13,5	102	11,6	58	6,6	54	6,1	40	4,5	38	4,3	21	2,4	18	2,0	–	–
6	55	6,1	236	26,0	87	9,6	105	11,6	73	8,1	57	6,3	60	6,6	48	5,3	40	4,4	39	4,3	32	3,5	–	–
7	115	10,7	296	27,6	141	13,2	171	16,0	115	10,7	63	5,9	48	4,5	26	2,4	25	2,3	–	–	–	–	–	–
8	85	7,7	366	33,1	152	13,7	183	16,5	110	10,0	46	4,2	39	3,5	32	2,9	27	2,4	–	–	–	–	–	–
9	88	7,8	272	24,2	201	18,0	152	13,5	105	9,4	76	6,8	44	3,8	58	5,2	40	3,6	30	2,7	–	–	–	–
10	118	9,6	304	24,9	161	13,2	147	12,1	112	9,2	87	7,1	66	5,4	66	5,4	33	2,7	30	2,5	27	2,2	–	–
11	104	8,4	310	25,1	153	12,4	176	14,3	110	8,9	78	6,3	67	5,4	45	3,6	49	4,0	46	3,7	32	2,5	–	–
12	98	7,8	310	24,8	116	13,1	128	10,3	147	11,8	83	6,6	76	6,1	67	5,4	48	3,8	28	2,2	–	–	–	–
13	105	8,4	330	26,4	225	18,0	170	13,6	144	11,5	65	5,2	44	3,5	46	3,7	44	3,5	–	–	–	–	–	–
14	106	7,1	343	23,2	185	12,5	166	11,2	156	10,5	125	8,5	86	5,8	62	4,2	73	4,9	42	2,8	34	2,3	–	–
15	189	9,5	403	21,4	302	16,0	259	13,7	211	11,2	149	7,9	119	6,3	65	3,5	56	3,0	49	2,6	–	–	–	–
16	302	7,6	1000	25,3	650	16,4	592	15,0	448	11,3	233	5,9	177	4,5	181	4,6	120	3,0	–	–	–	–	–	–
Total	1650	8,1	5116	25,1	2887	14,1	2730	13,4	2109	10,3	1334	6,5	1027	5,0	874	4,3	710	3,5	394	1,9	217	1,0	20	0,1

1. Die Lern- und Apperzeptionszeit einer Signalfolge (hier: Buchstabenfolge) ist proportional zu ihrem subjektiven Informationsgehalt (vgl. hierzu auch *Sanders,* 1971).
2. Der Informationsgehalt einer Folge von m verschiedenen Zeichen berechnet sich mit

$$I = m \cdot \mathrm{ld}\, m \text{ (bit)}$$

3. Durch eine sogenannte Superzeichenbildung kann eine Herabsetzung des subjektiven Informationsgehaltes erreicht werden, was (unter Berücksichtigung von Punkt 1) mit einer Herabsetzung des Lern- und Apperzeptionsaufwandes einhergeht. Das Verfahren der Superzeichenbildung besteht darin, kleinere Abschnitte des gesamten Textes zu Einheiten (Wörtern) zusammenzufassen; diese Einheiten werden zunächst einzeln gelernt und erst dann zum Gesamttext zusammengesetzt.

Wir untersuchen die Frage, wie gross die subjektive Information bei verschiedenen Zerlegungen ist.

Zugrunde gelegt sei ein Text der Länge m, derart, dass man den Text in q gleichlange Wörter mit der Buchstabenzahl p einteilen kann. Die subjektive Information die beim Lernen des Textes zu speichern ist, setzt sich aus zwei Teilen zusammen:

a. Zum Erlernen eines einzelnen Wortes (auf dem Repertoire seiner p Elemente) braucht der Lernende nur diese p Elemente in Betracht zu ziehen und nicht das aus m Zeichen bestehende Gesamtrepertoire. Damit erhält ein Einzelwort die subjektive Information

$$p \cdot \mathrm{ld}\, p \text{ (bit)}$$

Der subjektive Informationsbetrag der insgesamt q Wörter beträgt demnach

$$q \cdot p \cdot \mathrm{ld}\, p$$

b. Anschliessend müssen die q Wörter (die nun als Superzeichen gespeichert sind) zum ganzen Text zusammengesetzt werden; dies ergibt einen weiteren Informationsbetrag von der Grösse

$$q \cdot \mathrm{ld}\, q$$

Das Erlernen dieser Wortzusammenfassung erfolgt dabei auf dem Repertoire der q verschiedenen Einzelwörter.

Der gesamte subjektive Informationsbetrag I_p beim Zerlegen in Wörter der Länge p berechnet sich als Summe aus den beiden Einzelbeträgen

$$I_p = q \cdot p \cdot \mathrm{ld}\, p + q \cdot \mathrm{ld}\, q$$

oder, da $q \cdot p = m$ ist,

$$I_p = m \cdot \mathrm{ld}\, p + \frac{m}{p} \cdot \mathrm{ld}\, \left(\frac{m}{p}\right)$$

Betrachten wir die Wörter der geschriebenen Sprache (unter einem rein

statistischen Gesichtspunkt) als zufällige Buchstabenkombinationen, die unter (gleichmässiger) Ausnutzung des gesamten Alphabetes zustande gekommen sind, so ist m = 26 (Anzahl der Buchstaben im Alphabet).

In Tab. 3 sind die Werte der subjektiven Information für eine bestimmte Permutation des Alphabetes bei der Zerlegung in Wörter der Länge p angegeben.

Tabelle 3: Subjektiver Informationsgehalt und relative Häufigkeit von Wörtern der Länge p.

Länge p der einzelnen Wörter	Subjektiver Informationsgehalt I_p (in bit)	relative Häufigkeit (in %)
2	74.0	8,1
3	68.0	25,1
4	69.5	14,1
5	73.5	13,4
6	76.5	10,3
7	80.5	6,5
.		
.		
.		

Die Zerlegung in dreibuchstabige Wörter erweist sich als die günstigste. Sie ergibt den niedrigsten Informationswert und führt entsprechend dem Gesetz der konstanten Aufnahmekapazität (Lernzeit und Informationsgehalt sind proportional) zu den kürzesten Lern- und Apperzeptionszeiten. An zweiter Stelle stehen die vierbuchstabigen, an dritter die fünfbuchstabigen und — dies ist bemerkenswert — erst an vierter die zweibuchstabigen Wörter.

Fast genau dieselbe Rangfolge stellt man bei den relativen Häufigkeiten fest. Dort rangieren die sechsbuchstabigen Wörter vor den zweibuchstabigen; das mag aber weitgehend auf zusammengesetzte Ausdrücke wie bspw. "anbei", "wofür" etc. zurückzuführen sein, doch darauf soll hier nicht näher eingegangen werden. Jedenfalls zeigen diese Ergebnisse einen engen Zusammenhang zwischen dem theoretisch berechneten Informationsgehalt (Lernaufwand) und der Struktur der tatsächlich gebrauchten Sprache.

Auf ähnliche Gegebenheiten kam auch *Mandelbrot* bei seiner Erklärung des *Zipf'schen* Gesetzes (vgl. hierzu etwa *Cherry*, 1967); auch dort werden Worthäufigkeiten mit Rangordnungen in Beziehung gesetzt, allerdings spielt die Wortlänge keine Rolle. *Mandelbrot* geht vom Begriff der Kosten aus; alle Zeichen, Buchstaben oder Wörter kosten beispielsweise Zeit oder Anstrengung — bei der technischen Übermittlung auch Geld. Diese Kosten "schliessen alles und jedes ein, was zum Aufwand für das Senden des geeignet

gewichteten Zeichens dazugehört". Seine Theorie zeigt – ähnlich wie unsere Berechnung – ohne Bezug auf empirische Daten, wie man Wörter mit einer Wahrscheinlichkeit in einer solchen Art versehen kann, dass die "Gesamtkosten" im Mittel ein Minimum ausmachen. In einer rein mathematischen Abhandlung zeigt er, dass die sich ergebende Beziehung zwischen der Häufigkeit und dem Rang eines Wortes dem experimentell gefundenen Gesetz von *Zipf* entspricht.

Institut für Verhaltenswissenschaft
Eidgenössische Technische Hochschule
CH-8000 Zürich

Hardi Fischer
H. U. Baumann

Literatur

Cherry, C. (1967): *Kommunikationsforschung – eine neue Wissenschaft,* Fischer Verlag.
von Cube, F. (1968): *Kybernetische Grundlage des Lernens und Lehrens,* Klett.
Fucks, W. (1968): *Nach allen Regeln der Kunst,* Suttgart, dva.
Sanders, A. F. (1971): *Psychologie der Informationsverarbeitung,* Huber Verlag.