KYBERNETIKA — JAHRGANG 8 (1972), HEFT 4

Prädiktionsentropie der slowakischen Schrift

VLADIMÍR MAJERNÍK, JOZEF KRÚTEĽ

Es werden drei verschiedene Methoden der Bestimmung der Prädiktionsentropie der Schrift beschrieben und ihre Vor- und Nachteile diskutiert. Als Anwendungsbeispiel wird die Informationsentropie und auch die Redundanz der slowakischen Schrift anhand drei Versuchstexte ermittelt. Die Methoden als solche sind auch für andere Sprachen anwendbar.

EINLEITUNG

Anwendungsmöglichkeiten der Informationstheorie in der Sprachwissenschaft bestehen vor allem im Aufsuchen solcher informationstheoretischen Parameter der Sprache, die bei Informationsübertragung mittels Sprache wichtig sind und ferner beim Aufstellen von Optimalcodes für Übertragung der Sprache. In erster Näherung kann die Sprache als stationärer stochastischer Prozeß angesehen werden, der durch bestimmte statistische Daten, vor allem durch die Vorkommenswahrscheinlichkeiten der einzelnen Elemente* (Grapheme oder Phoneme) bzw. von Gruppen dieser Elemente (i-Gramme) beschrieben werden kann. Sind die statistischen Parameter der Sprache bekannt, so kann ihre Informationsentropie erster bzw. i-ter Ordnung mit Hilfe der folgenden Beziehungen bestimmt werden:

$$H_1 = -\sum_{j=1}^{n} p_{l_j} \log_2 p_{l_j} \tag{1a}$$

bzw.

$$H_i = -\sum_{j,k,m,\ldots,s=1}^{n} p_{l_j l_k l_m \ldots l_s} \log_2 p_{l_j l_k l_m \ldots l_s}\,, \tag{1b}$$

wobei p_{l_j} bzw. $p_{l_j l_k \ldots l_s}$ die Vorkommenswahrscheinlichkeiten der einzelnen Ele-

* Da wir uns in der vorliegenden Arbeit mit der Bestimmung der Entropie der slowakischen Schrift befassen, verstehen wir unter Sprachelementen stets Grapheme.

mente bzw. der i-Gramme bedeuten, n ist die Gesamtzahl der Sprachelemente. Index $l_j, l_k, \ldots, l_s$ symbolisiert allgemein irgendein Sprachelement, Indexe zweiter Ordnung $1, 2, \ldots, i$ bezeichnen seine Ordnungszahl innerhalb des i-Grammes. Wäre das Vorkommen der einzelnen Elemente in der Sprache voneinander unabhängig, so genüge es zur Bestimmung der Entropie lediglich die Beziehung (1a) in Betracht zu ziehen. Da jedoch die Verteilung der bedingten Vorkommenswahrscheinlichkeit für das i-te Element im Text stark von den vorhergehenden $i - 1$ Elementen abhängt, ist es bei der Bestimmung der Entropie der Sprache notwending diese Abhängigkeit zu berücksichtigen und die Entropie dementsprechend nach der Beziehung (1b) zu berechnen.

Ein gewisses Maß für die Bindung der Sprachelemente stellt die Redundanz $\mathscr{R}$ dar, definiert durch die Beziehung

$$\mathscr{R}_i = 1 - \frac{H_i}{H_0}, \tag{2}$$

wobei H_i die Informationsentropie i-ter Ordnung und $H_0 = \log_2 n$ ist. ($\mathscr{R}_i$ ist Redundanz i-ter Ordnung.)

Informationsentropie und Redundanz sind die wichtigsten Charakteristiken der Sprache vom Standpunkt der Kommunikationstheorie. Zu deren Ermittlung benötigen wir die Statistik der i-Gramme, was allerdings hinsichtlich der Materialmenge – besonders für $i > 3$ – auszuarbeiten sehr mühsam ist. Aus diesem Grunde wurde mit Erfolg versucht Methoden zu entwerfen, die eine vernünftige Abschätzung für die interessierenden Größen erlauben würden. Es gibt mehrere Prädiktionsmethoden, sie beruhen im Prinzip darauf, daß das Subjekt (Versuchsperson) auf geeignete Weise die bedingte Wahrscheinlichkeit dessen, daß das ihm bereits bekannte $(i - 1)$-Gramm gerade durch das Element l_i der Sprache zum i-Gramm ergänzt wird, quantitativ erfaßt, ohne Kenntnis der Statistik der i-Gramme. Anhand dessen kann dann die bedingte Entropie $\mathscr{F}_i$, gegeben durch

$$\mathscr{F}_i = - \sum_{j,k,\ldots,s=1}^{n} p_{l_j l_k \ldots l_{si}} \log_2 p_{l_j l_k \ldots l_{s-1,i-1}(l_{si})}$$

ermittelt werden. Hierbei bedeutet $p_{l_j l_k \ldots l_{s-1,i-1}(l_{si})}$ die bedingte Vorkommenswahrscheinlichkeit des Elementes l_s auf Platz i im i-Gramm, wenn die vorhergehenden Elemente $l_j, l_k, \ldots, l_{s-1}$ sind.

Die auf diese Art ermittelte Entropie wird als bedingte Prädiktionsentropie bezeichnet. Es ist einleuchtend, daß solche Tests lediglich eine Abschätzung für die Informationsentropie der Sprache liefern können; diese Abschätzung verschafft uns dennoch ein hinreichendes Bild dieses wichtigen kommunikationstheoretischen Parameters.

In weiteren Teilen dieser Arbeit befassen wir uns mit der Bestimmung der Prädiktionsentropie der slowakischen Schrift, und zwar nach mehreren Methoden. Zunächst sollen noch die einzelnen Methoden zur Aussprache gelangen.

I. Methode von Shannon

Diese besteht darin, daß das Subjekt die einzelnen Sprachelemente nacheinander ratet. Es ist klar, daß sprachlich und sachlich gut ausgebildete Versuchsperson sehr wohl abschätzen kann, welches Element mit großer Wahrscheinlichkeit folgen könnte (vor allem, wenn bereits mehrere vorhergehende Elemente bekannt sind) und somit stellt die Anzahl der benötigten Versuche zum Erraten des gesuchten Elementes ein Maß für dessen Informationsgehalt dar. Bei dieser Methode wird also jedem Element eine Zahl zugeordnet, die der Anzahl der benötigten Versuche zum Erraten des gesuchten Elementes gleich ist.

Es sei M die Gesamtzahl der i-Gramme. Man nehme an, daß genau beim f-ten Versuch l_f Grapheme, die die $(i-1)$-Gramme zu i-Grammen ergänzen, richtig geraten wurden. Dann ist die relative Häufigkeit des Erratens des i-ten Graphemes des i-Grammes, unter Kenntnis der vorhergehenden $(i-1)$ Grapheme, genau beim f-ten Versuch

$$q_i^{(f)} = \frac{l_f}{M}, \quad \text{wobei} \quad f = 1, 2, \ldots, n .$$

Für die bedingte Entropie gilt dann die Ungleichheit [1]:

$$(3) \qquad -\sum_{f=1}^{n} q_i^{(f)} \log_2 q_i^{(f)} \geqq \mathscr{F}_i \geqq \sum_{f=1}^{n} f(q_i^{(f)} - q_i^{(f+1)}) \log_2 f .$$

Nach dieser Methode kann somit eine obere und eine untere Schranke für die Prädiktionsentropie der Schrift ermittelt werden.

II. Methode des Informationsgewinnes

Der Ausgangspunkt bei dieser Methode ist die Tatsache, daß die Versuchsperson – bei Unkenntnis des vorangehenden Textteiles – bei der Bestimmung des gesuchten Elementes aus der Verteilung der Vorkommenswahrscheinlichkeit der Grapheme ausgehen muß. Werden nun der Versuchsperson ein oder mehrere unmittelbar vorangehende Textelemente angegeben, so ändert sich die Wahrscheinlichkeitsverteilung des zu ermittelnden Graphems. Mit dieser Änderung ist ein folgendermaßen definierter Informationsgewinn verknüpft [2]:

Es seien $P(p_1, p_2, \ldots, p_n)$ und $W(w_1, w_2, \ldots, w_q)$ zwei diskrete Wahrscheinlichkeitsverteilungen definiert auf der Menge der Sprachelemente (z. B. Grapheme). Wird die Wahrscheinlichkeitsverteilung P durch die Verteilung W ersetzt, dann ist der hierdurch bedingte Informationsgewinn I gegeben durch

$$(4) \qquad I(W \mid P) = \sum_{i=1}^{n} w_1 \log_2 (w_1/p_1) .$$

Im Spezialfall $w_j = 1$ bekommt man aus (4)

$$I(W' \mid P) = -\log_2 p_j \,. \tag{4a}$$

In unserem Fall ist p_j die bedingte Prädiktionswahrscheinlichkeit dessen, daß bei Kenntnis der vorangehenden $(i-1)$ Grapheme auf Platz i ein bestimmtes Graphem l_j folgen wird.

In der Praxis wird so verfahren, daß die Versuchsperson die Wahrscheinlichkeitsverteilung der Grapheme auf Platz i im i-Gramm aufzustellen hat, wobei die vorangehenden $(i-1)$ Grapheme bekannt sind. Die gesamte dem Subjekt zugeführte Informationsmenge, indem man ihm das i-te Graphen des i-Grammes mitteilt, ist durch die Summe des Informationsgewinnes bei den einzelnen Versuchen gegeben. Die mittlere Entropie pro Buchstabe bei m Versuchen, bei denen das richtige Graphem mit der Wahrscheinlichkeit p_j ermittelt wurde, ist also gegeben durch

$$\mathscr{F}_i^{(m)} = \frac{-\sum \log_2 p_j}{m} \,. \tag{4b}$$

Mittels (4b) erlangen wir eine Abschätzung für die gesuchte Informationsentropie $\mathscr{F}_i$, wobei – hinsichtlich dessen, daß jede Fehlentscheidung (Irrtum) bei der Konstruktion der bedingten Wahrscheinlichkeitsverteilung ihren Wert stets vergrößert [3] – gilt:

$$\mathscr{F}_i \leqq \mathscr{F}_i^{(m)} \,.$$

III. Methode von Kolmogorov

Bei dieser Methode sind gewisse Wahrscheinlichkeitsverteilungen vorderhand vorgegeben. Die in diesen Verteilungen auftretenden Wahrscheinlichkeiten werden aus der Extrembedingung (Minimum) für die Informationsentropie (s. z.B. [3]) ermittelt. Die Versuchsperson ist vor die Aufgabe gestellt, sich für eine der vorgegebenen Verteilungen zu entscheiden.

In unserer Arbeit haben wir, im Vergleich mit [3], die Anzahl der möglichen Antworten erhöht, sie wurden folgendermaßen vorgegeben:

1. Das folgende Element ist mit Sicherheit ein bestimmter Buchstabe (k-ter Buchstabe) des Alphabets.

2a) Das folgende Element ist mit Sicherheit ein von zwei von der Versuchsperson angegebenen Buchstaben.

2b) Das folgende Element ist mit Sicherheit ein von drei durch die Versuchsperson angegebenen Buchstaben.

3. Das folgende Element ist wahrscheinlich ein bestimmter Buchstabe des Alphabets.

4a) Das folgende Element ist wahrscheinlich ein von zwei durch die Versuchsperson angegebenen Buchstaben.

4b) Das folgende Element ist wahrscheinlich ein von drei durch die Versuchsperson angegebenen Buchstaben.

5. Das folgende Element ist ein Buchstabe aus der Menge $\mathscr{A}$ bestehend aus n_1 Elementen des Alphabets.

6. Das folgende Element ist ein Buchstabe aus der Menge $\mathscr{B}$ bestehend aus n_2 Elementen des Alphabets ($n_1 + n_2 = n$; $\mathscr{A} \cap \mathscr{B} = \emptyset$).

7. Das folgende Element kann von der Versuchsperson nicht angegeben werden.

Einer jeden dieser Aussagen entspricht bestimmte Wahrscheinlichkeitsverteilung, und zwar:

1. Die Wahrscheinlichkeit des angegebenen k-ten Graphems sei mit P bezeichnet. Für Wahrscheinlichkeiten der übrigen Elemente nehmen wir $\bar{p}$, definiert durch

$$\bar{p}_i = p_i(1 - P)/(1 - p_k), \quad i = 1, 2, \ldots, n, \quad i \neq k$$

wobei p_i, p_k sind die Vorkommenswahrscheinlichkeiten der betreffenden Elemente im geschriebenen Text.

2a) Die beiden angegebenen Grapheme k_1, k_2 sind gleichwahrscheinlich ($\frac{1}{2}P$). Für die Wahrscheinlichkeiten der übrigen Elemente nehmen wir $\bar{p}$, gegeben durch*

$$\bar{p}_i = p_i(1 - P)/(1 - p_{k_1} - p_{k_2}),$$
$$i = 1, 2, \ldots, n, \quad i \neq k_1, \quad i \neq k_2.$$

2b) Die drei angegebenen Grapheme k_1, k_2, k_3 sind gleichwahrscheinlich ($\frac{1}{3}P$). Für die Wahrscheinlichkeiten der übrigen Elemente nehmen wir $\bar{p}$, gegeben durch

$$\bar{p}_i = p_i(1 - P)/(1 - p_{k_1} - p_{k_2} - p_{k_3}),$$
$$i = 1, 2, \ldots, n, \quad i \neq k_1, \quad i \neq k_2, \quad i \neq k_3.$$

3. Das angegebene k-te Graphem hat die Wahrscheinlichkeit G. Für die Wahrscheinlichkeiten der übrigen Elemente nehmen wir den Ausdruck**

$$\bar{p}_i = p_i(1 - G)/(1 - p_k).$$

* Die Bedeutung der einzelnen Symbole ist wie bei Aussage 1, von einer Wiederholung sehen wir deshalb im weiteren ab.

** Die Bedingungen für die Indizes sind analog wie bei den vorhergehenden Aussagen von einer expliziten Angabe wird im weiteren deshalb abgesehen.

4a) Die beiden angegebenen Grapheme k_1, k_2 sind gleichwahrscheinlich $(\frac{1}{2}G)$. Für die Wahrscheinlichkeiten der übrigen Elemente nehmen wir

$$\bar{p}_i = p_i(1 - G)/(1 - p_{k_1} - p_{k_2}) .$$

4b) Die drei angegebenen Grapheme k_1, k_2, k_3 sind gleichwahrscheinlich $(\frac{1}{3}G)$. Für die Wahrscheinlichkeiten der übrigen Elemente nehmen wir

$$\bar{p}_i = p_i(1 - G)/(1 - p_{k_1} - p_{k_2} - p_{k_3}) .$$

5. Grapheme aus der gewählten Menge $\mathscr{A}$ haben die Wahrscheinlichkeit $(1/n_1)\,F$. Für die Wahrscheinlichkeiten der restlichen Elemente nehmen wir $(1 - 1/n_1)\,F$.

6. Grapheme aus der gewählten Menge $\mathscr{B}$ haben die Wahrscheinlichkeit $(1/n_2)\,D$. Für die Wahrscheinlichkeit der restlichen Elemente nehmen wir $(1 - 1/n_2)\,D$.

7. Für das tatsächlich vorkommende Element nimmt man seine Vorkommenswahrscheinlichkeit.

Die noch unbekannten Zahlenwerte der Wahrscheinlichkeiten P, G, F und D gewinnt man aus der Extrembedingung (Minimum) der Informationsentropie $\mathscr{F}_i$, die in unserem Fall folgendermaßen gegeben ist:

$$\begin{aligned}(5)\qquad \mathscr{F}_i = \frac{1}{M}\{&(M_1 - m_1)\log_2 P + (M_1^{(2a)} - m_1^{(2a)})\log_2 \tfrac{1}{2} + \\ &+ (M_2 - m_2)\log_2 G + (M_2^{(4a)} - m_2^{(4a)})\log_2 \tfrac{1}{2} + \\ &+ (M_1^{(2b)} - m_1^{(2b)})\log_2 \tfrac{1}{3} + (M_2^{(4b)} - m_2^{(4b)})\log_2 \tfrac{1}{3} + \\ &+ \sum_{\text{non}(1)} \log_2 p_i(1 - P)/(1 - p_{k_1}) + \\ &+ \sum_{\text{non}(3)} \log_2 p_i(1 - G)/(1 - p_{k_1}) + \\ &+ \sum_{\text{non}(2a)} \log_2 p_i(1 - P)/(1 - p_{k_1} - p_{k_2}) + R + \\ &+ \sum_{\text{non}(4a)} \log_2 p_i(1 - G)/(1 - p_{k_1} - p_{k_2}) + \\ &+ \sum_{\text{non}(2b)} \log_2 p_i(1 - P)/(1 - p_{k_1} - p_{k_2} - p_{k_3}) + \\ &+ \sum_{\text{non}(4b)} \log_2 p_i(1 - G)/(1 - p_{k_1} - p_{k_2} - p_{k_3}) + \\ &+ N_1^{(s)}\log_2 F + N_1^{(n)}\log_2 (1 - F) + N_2^{(s)}\log_2 D + \\ &+ N_2^{(n)}\log_2 (1 - D)\}\,,\end{aligned}$$

wobei die einzelnen Symbole folgende Bedeutung haben: M – Gesamtzahl der Versuche; M_1 – Gesamtzahl der erfolgreichen (richtigen) Rateversuche in den Fällen

1, 2a) und 2b); $M_1^{(2a)}$ – Anzahl der erfolgreichen Versuche im Fall 2a); M_2 – Gesamtzahl der erfolgreichen Versuche in den Fällen 3, 4a) und 4b); $M_2^{(4a)}$ – Anzahl der erfolgreichen Versuche im Fall 4a); $M_1^{(2b)}$ – Anzahl der erfolgreichen Versuche im Fall 2b); $M_2^{(4b)}$ – Anzahl der erfolgreichen Versuche im Fall 4b); $\sum_{\text{non}(1)}$, $\sum_{\text{non}(2a)}$, $\sum_{\text{non}(2b)}$, $\sum_{\text{non}(3)}$, $\sum_{\text{non}(4a)}$ bzw. $\sum_{\text{non}(4b)}$ sind Summationsvorschriften für Summierung der Wahrscheinlichkeiten für sämtliche Fälle erfolgloser Rateversuche im Fall 1, 2a), 2b), 3, 4a) bzw. 4b); m_1 – Gesamtzahl der erfolglosen (fehlerhaften) Versuche in den Fällen 1, 2a) und 2b); $m_1^{(2a)}$ – Anzahl der erfolglosen Versuche im Fall 2a); m_2 – Gesamtzahl der erfolglosen Versuche in den Fällen 3, 4a) und 4b); $m_2^{(4a)}$ – Anzahl der erfolglosen Versuche im Fall 4a); $m_1^{(2b)}$ – Anzahl der erfolglosen Versuche im Fall 2b); $m_2^{(4b)}$ – Anzahl der erfolglosen Versuche im Fall 4b); $N_1^{(s)}$ bzw. $N_2^{(s)}$ – Anzahl der erfolgreichen Versuche im Fall 5 bzw. 6; $N_1^{(n)}$ bzw. $N_2^{(n)}$ – Anzahl der erfolglosen Versuche im Fall 5 bzw. 6; R – Summe der Vorkommenswahrscheinlichkeiten der Grapheme für all die Fälle, in denen sich die Versuchsperson für die Antwort 7 entschieden hat; Indizes k_1, k_2 bzw. k_3 repräsentieren die von der Versuchsperson angegebenen Grapheme im Fall [1, 3], [2a), 4a)] bzw. [2b), 4b)].

Notwendige Bedingung für die Extrembildung der Informationsentropie ist die Erfüllung der Bedingungen:

$$\partial \mathscr{F}_i/\partial P = 0\,, \quad \partial \mathscr{F}_i/\partial G = 0\,, \quad \partial \mathscr{F}_i/\partial F = 0\,, \quad \partial \mathscr{F}_i/\partial D = 0\,. \tag{6}$$

Aus den Bedingungsgleichungen (6) für die Extrembildung der Informationsentropie $\mathscr{F}_i$ können die unbekannten Wahrscheinlichkeiten P, G, F und D berechnet werden. Führt man die partielle Differentiation durch und trennt man die Unbekannten, so erhält man

$$\begin{aligned} P &= 1 - q_1\,, \\ G &= 1 - q_2\,, \\ F &= 1 - q_3\,, \\ D &= 1 - q_4\,, \end{aligned} \tag{7}$$

wobei $q_1 = m_1/V_1$, $q_2 = m_2/V_2$, $q_3 = N_1^{(n)}/N_1$ und $q_4 = N_2^{(n)}/N_2$. Hierbei V_1 bzw. V_2 ist die Gesamtzahl der Antworten des Types 1, 2a) und 2b) bzw. 3, 4a) und 4b); N_1 bzw. N_2 ist die Gesamtzahl der Antworten des Types 5 bzw. 6.

Setzt man (7) in (5) ein, so erhält man abschließend

$$\begin{aligned} \mathscr{F}_i = \frac{1}{M}\{&V_1 h_1 + V_2 h_2 + S_a \log_2 2 + S_b \log_2 3 + \\ &+ \sum_{\text{non}(1),(3)} \log_2 p_i/(1 - p_k) + \sum_{\text{non}(2a),(4a)} \log_2 p_i/(1 - p_{k_1} - p_{k_2}) + \end{aligned} \tag{8}$$

$$+ \sum_{\text{non}(2b),(4b)} \log_2 p_i / (1 - p_{k_1} - p_{k_2} - p_{k_3}) + R +$$
$$+ N_1 h_3 + N_2 h_4 + (N_1^{(s)} + N_2 - N_2^{(s)}) \log_2 n_1 +$$
$$+ (N_1 + N_2^{(s)} - N_1^{(s)}) \log_2 n_2 \} ,$$

wobei

$$S_a = M_1^{(2a)} + M_2^{(4a)} - (m_1^{(2a)} + m_2^{(4a)}) ,$$
$$S_b = M_1^{(2b)} + M_2^{(4b)} - (m_1^{(2b)} + m_2^{(4b)}) ,$$
$$h_1 = -q_1 \log_2 q_1 - (1 - q_1) \log_2 (1 - q_1) ,$$
$$h_2 = -q_2 \log_2 q_2 - (1 - q_2) \log_2 (1 - q_2) ,$$
$$h_3 = -q_3 \log_2 q_3 - (1 - q_3) \log_2 (1 - q_3)$$

und

$$h_4 = -q_4 \log_2 q_4 - (1 - q_4) \log_2 (1 - q_4) .$$

Alle in der Beziehung (8) auftretenden Größen sind bekannt und der Extremwert der durchschnittlichen Informationsentropie $\mathscr{F}_i$ kann daher bestimmt werden.

BESTIMMUNG DER INFORMATIONSENTROPIE DER SLOWAKISCHEN SCHRIFT

Mit Hilfe der erläuterten Methoden wurde die Informationsentropie der slowakischen Schrift bestimmt. An diesen Versuchen nahmen drei Versuchspersonen teil, die im weiteren als A, B und C bezeichnet werden. Es wurde das folgende experimentelle Material benutzt:

a) Hundert verschiedene, zufällig ausgewählte 49-Gramme aus dem Buch: Anton Hykisch: *Stretol som Ťa.* Smena, Bratislava 1963. Die Versuchsperson hatte jeweils den 50-sten Buchstaben zu erraten.

b) Der Satz: *Jej nádherné oči hľadia naň nežno a milo, nespúšťajúc sa z neho, lesknú sa hrdosťou v povedomí svojej ceny a krásy*; aus dem Buch: Martin Kukučín: *Dom v stráni.* Slovenské vydavateľstvo krásnej literatúry, Bratislava 1961. Die einzelnen Elemente (Grapheme) dieses Satzes wurden nacheinander ermittelt, wobei der Versuchsperson jeweils nur der vorhergehende Teil des Satzes (bei Bestimmung des ersten Buchstaben also nichts) bekannt war.

c) 118 einzelne, im Text zufällig verteilte Grapheme, unter Kenntnis des gesamten vorhergehenden Teiles des Buches: Peter Ševčovič: *Mesto plné chlapov.* Slovenský spisovateľ, Bratislava 1963, waren zu ermitteln.

Für alle drei angegebenen Textauswahlen wurde die Informationsentropie nach der Methode des Informationsgewinnes bestimmt; für das Versuchsmaterial unter c)

außerdem auch nach den beiden übrigen beschriebenen Methoden. Die Ergebnisse der Bestimmung der Informationsentropie nach der Methode des Informationsgewinnes für alle drei Texte sind übersichtlich in der Tabelle 1 zusammengefaßt. Die angegebenen zahlenmäßigen Ergebnisse wurden nach der Beziehung (4b) berechnet. Für den Versuchstext unter c) wurde die Informationsentropie – wie bereits erwähnt – auch nach der Methode von Shannon (I) und Kolmogorov (III) ermittelt; die zahlenmäßigen Ergebnisse wurden nach der Beziehung (3) bzw. (8) berechnet und sind in der Tabelle 2 zusammengefaßt.

Tabelle 1.

Text* \ Subjekt	A	B	C
a) $\mathscr{F}_{50}^{(n)}$ [bit]	2,27	1,87	1,44
b) $\mathscr{F}_{30-113}^{(n)}$ [bit]	2,14	1,79	—
c) $\mathscr{F}_{\infty}^{(n)}$ [bit]	2,32	1,78	—

Tabelle 2.

Methode \ Subjekt		A	B
Shannon	obere $\mathscr{F}_{\infty}$ [bit]	2,13	1,89
	untere $\mathscr{F}_{\infty}$ [bit]	1,28	1,23
Informationsgewinn $\mathscr{F}_{\infty}$ [bit]		2,32	1,78
Kolmogorov $\mathscr{F}_{\infty}$ [bit]		2,49	1,88

* Der Zahlenwert für Text b) ist Mittelwert vom 30-sten bis 113-ten Buchstaben. Der Anfang des Satzes wurde vernachlässigt, da starken Schwankungen ausgesetzt.

Mit Hilfe der Beziehung (2) kann auch die Redundanz der slowakischen Schrift bestimmt werden, deren Werte, berechnet aus denen der Informationsentropie gemäß Tabelle 1 bzw. 2, durch die Tabelle 3 bzw. 4 wiedergegeben sind. Die relativen Vor-

Tabelle 3.

Text* \ Subjekt	A	B	C
a) $\mathscr{R}_{50}$ [%]	58,5	66,3	73,5
b) $\mathscr{R}_{30-113}$ [%]	60,8	67,2	—
c) $\mathscr{R}_{\infty}$ [%]	57,2	67,3	—

* Sie Anmerkung bei Tabelle 1.

Tabelle 4.

Methode \ Subjekt		A	B
Shannon	obere $\mathscr{R}_{\infty}$ [%]	60,8	65,3
	untere $\mathscr{R}_{\infty}$ [%]	76,6	77,4
Informationsgewinn $\mathscr{R}_{\infty}$ [%]		57,2	67,3
Kolmogorov $\mathscr{R}_{\infty}$ [%]		54,3	65,7

kommenshäufigkeiten der einzelnen Grapheme der slowakischen Schrift, deren Kenntnis bei der Methode von Kolmogorov vorausgesetzt wird, wurden der Arbeit [4] entnommen. Aus der Abb. 1 ersieht man die Abhängigkeit des Informations-

gewinnes für den Text b) von der Ordnungszahl der Grapheme in einem zusammenhängenden (sinnvollen) Satz.

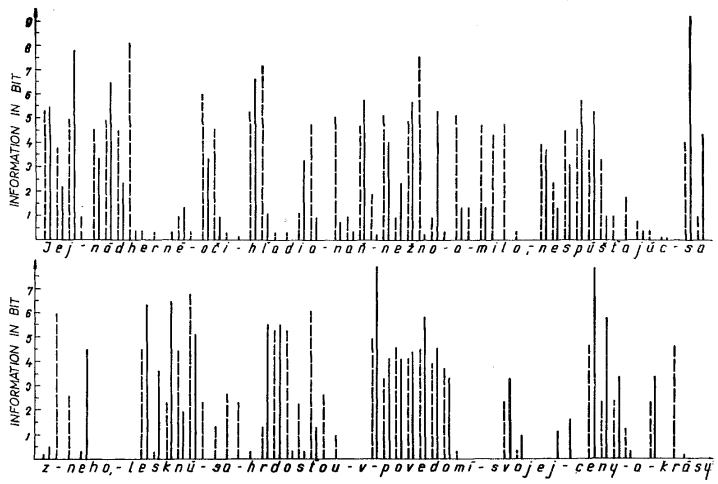

Abb. 1.

DISKUSSION DER ERGEBNISSE

Bei der Untersuchung des Funktionsschemas der einzelnen Methoden sind wir zu folgenden Erkenntnissen gekommen:

Die Methode von Shannon ist verhältnismäßig einfach auszuführen, hat aber den Nachteil, daß die für die Informationsentropie gelieferte untere Schranke zu niedrig ist und in der Praxis kaum je erreicht werden kann und gleichzeitig das Intervall zwischen der oberen und unteren Schranke zu breit ist, so daß die Abschätzung der Informationsentropie einzig und allein nach dieser Methode mit ziemlicher Unbestimmtheit behaftet ist.

Die Methode des Informationsgewinnes – wie bereits erwähnt – liefert lediglich eine obere Schranke für die Informationsentropie (s. entsprechenden Absatz der vorliegenden Arbeit). Diese Methode ist etwas aufwendiger als die von Shannon (vor allem für die Versuchsperson, die Auswertung ist einfach) und ihr Hauptnachteil beruht darin, daß das Subjekt sehr aufmerksam arbeiten muß, um ein noch so wenig wahrscheinliches Element nicht ganz (oder praktisch ganz) auszuschließen. Fällt nämlich das tatsächlich folgende Element in die Gruppe der ausgeschlossenen (oder

praktisch ausgeschlossenen, d.h. mit sehr geringen Wahrscheinlichkeit vorgesehenen) Grapheme, so ist der hiermit verbundene Informationsgewinn gleich unendlich (zumindest sehr groß) und die gesamte Versuchsreihe wird damit unbrauchbar.

Diesen Nachteil schließt die Methode von Kolmogorov, die ebenfalls nur eine obere Schranke für die Informationsentropie liefert, völlig aus (s. ihre Konstruktion). Ihr großer Nachteil liegt im übermäßigen Arbeitsaufwand sowohl für die Versuchsperson als auch bei der Auswertung. Nach unserer Erfahrung sind die erzielten Ergebnisse der Methode von Kolmogorov, selbst nach der von uns vorgenommenen Erweiterung, der aufzuwendenden Arbeit nicht proportional, und zwar selbst nicht im Fall maschineller Auswertung.

Als optimal erscheint uns die Methode des Informationsgewinnes, gegebenenfalls in Verbindung mit der Methode von Shannon. Durch diese Kombination kann nämlich das Intervall zwischen der oberen und unteren Schranke für die Informationsentropie weitgehend eingeschränkt werden und man gelangt bei verhältnismäßig kleinem Arbeitsaufwand zur vernünftigen Abschätzung der gesuchten Größe.

Die erzielten Ergebnisse, wie aus den entsprechenden Tabellen ersichtlich, hängen stark von der Versuchsperson ab, vor allem von deren sprachlichen (aber auch sachlichen) Vorbildung sowie von deren Aufmerksamkeit bei der Konstruktion der Verteilungsfunktionen. Die Ergebnisse weisen deshalb starke individuelle Unterschiede auf, was allerdings durchaus berechtigt erscheint, da der selbe Text (Nachricht) verschiedenen Empfängern, in Abhängigkeit von ihren Vorkenntnissen, unterschiedliche Menge an Information vermittelt.

(Eingegangen am 8. August 1969.)

LITERATUR

[1] Shannon C. E.: Prediction and Entropy of Printed English. Bell Syst. Techn. Journ. *30* (1951), 50.

[2] Rényi A.: Wahrscheinlichkeitsrechnung. VEB Deutscher Verlag d. Wissenschaften, Berlin 1962, 453.

[3] Jaglom A. M., Jaglom I. M.: Pravděpodobnost a informace. Nakl. ČSAV, Praha 1964.

[4] Majerník V.: Niektoré informačno-teoretické parametre písanej slovenčiny. Sdělovací technika *8* (1966), 302.

Predikčná entrópia písanej slovenčiny

VLADIMÍR MAJERNÍK, JOZEF KRÚTEĽ

V práci sú popísané tri metódy určovania predikčnej entrópie písaného prejavu a to: Shannonova metóda, metóda zisku informácie a Kolmogorovova metóda. Posledne menovaná metóda bola autormi rozšírená za cieľom získania presnejších výsledkov. V ďalšej časti sú diskutované prednosti a nedostatky jednotlivých metód a to najmä čo do pracnosti prevádzania s prihliadnutím na kvalitu výsledkov, ktoré poskytujú. Samotné metódy ako také sú použiteľné pre akýkoľvek jazyk.

V druhej časti práce – ako príklad na použitie týchto metód – je určovaná informačná entrópia a redundacia písanej slovenčiny. Ukazuje sa, že výsledky sú značne závislé od subjekta, a to u všetkých troch použitých metód.

Podľa našich skúseností sa ako optimálna javí metóda zisku informácie, poprípade v spojení so Shannonovou metódou. Obe sú jednoduché na spracovanie výsledkov a tiež nie príliš zložité pre pokusnú osobu. Kolmogorovova metóda, ani po zdokonalení, ktoré sme zaviedli, nedáva výsledky úmerné jej pracnosti.

RNDr. Vladimír Majernik, CSc., RNDr. Jozef Krúteľ, CSc., Fyzikálny ústav SAV (Physikalisches Institut der Slowakischen Akademie der Wissenschaften), Dúbravská cesta, Bratislava.