

Analyse von Zeichenabfolgen in natürlichsprachigen Texten

Motivation

Die Informationstheorie sagt uns, daß ein exaktes Informationsmaß über Auftretswahrscheinlichkeiten möglicher Ereignisse definiert werden kann. Diese Wahrscheinlichkeiten lassen sich zu gegebenen Quelltexten empirisch über eine statistische Analyse von Häufigkeiten gewinnen.

Anhand einer Analyse umfangreicher natürlichsprachiger Texte¹ können die tatsächlichen relativen Häufigkeiten

$$\frac{\#\text{Vorkommen von } x}{\#\text{Zeichen}} = P(X=x) =: p(x)$$

einzelner Buchstaben x ermittelt werden. Danach läßt sich mit Shannons Informationsmaß der tatsächliche Informationsgehalt eines Zeichens bestimmen.

Zusätzlich können noch die bedingten Auftretswahrscheinlichkeiten

$$\frac{\#\text{Vorkommen der Abfolge } yx}{\#\text{Vorkommen von } y} = P(X=x | Y=y) =: p_y(x)$$

von Zeichen x in Abhängigkeit des vorhergehenden Zeichens y ermittelt werden.

Es ist zu erwarten, daß mit einer solchen statistisch basierten Modellierung tatsächlicher Auftretswahrscheinlichkeiten eine Reduktion des Informationsgehalts nach Shannon beobachtet werden kann. Eine noch weitergehende Reduktion des Informationsgehalts müßte durch Ausnutzung des kontextuellen Wissens, also über eine Modellierung bedingter Auftretswahrscheinlichkeiten, möglich sein.

Berechnung der Realinformation aus Wahrscheinlichkeiten

Gegeben die Auftretswahrscheinlichkeiten $p(x)$ der Zeichen x (sie werden direkt aus der erhobenen Quellenstatistik gewonnen), ist der durchschnittliche **Informationsgehalt eines Zeichens** definiert als der Erwartungswert

$$H(X) = - \sum_i p(x_i) \log p(x_i)$$

Die bedingten Wahrscheinlichkeiten $p_y(x)$ – bezogen auf das jeweils vorhergehende Zeichen y – fließen ein in die Berechnung des durchschnittlichen **bedingten Informationsgehalts eines Zeichens**

$$\hat{H}(X) = \sum_i p(x_i) \left[- \sum_j p_{x_i}(x_j) \log p_{x_i}(x_j) \right]$$

¹In diesem Beispiel haben wir uns auf deutschsprachige Texte beschränkt, siehe auch das Kapitel "Ergebnisse"

Implementierung: Vereinfachungen, Varianten

Es wurden zwei Implementierungsvarianten realisiert:

- Der ersten Variante sind sämtliche Byte-Werte je eigene Zeichen x , sodaß es ihrer $2^8 = 256$ gibt.
Jedes mögliche Zeichen wird also gesondert betrachtet, es wird insbesondere zwischen Groß- und Kleinschreibung unterschieden; Ziffern, Interpunktions- und Sonderzeichen werden als solche mitgezählt.
- Die zweite Variante reduziert über eine alias-Funktion alle Zeichen, die nicht Buchstaben sind, auf Leerzeichen (Blanks), die damit gleich behandelt werden.
Da es bei natürlichsprachigen Texten im engeren Sinne nur um Buchstabenabfolgen geht, wird Groß-/Kleinschreibung eingeebnet; ferner wird dadurch, daß Interpunktionszeichen (und auch Klammern sowie Anführungszeichen etc.) wie Blanks behandelt werden, mitmodelliert, welche Buchstaben besonders an Wortanfängen und -enden auftreten.

Ergebnisse

Getestet wurden beide Varianten anhand von drei verschiedenen Textdateien:

- 1) Ein Buch² mit ca. 5 Mio. Zeichen:

```
⋮
(44) Wie man aus dem Vierten Buch dieser Schrift näher sehn wird, hat A. Smith
keinen einzigen neuen Satz über die Teilung der Arbeit aufgestellt. Was ihn aber als
den zusammenfassenden politischen Ökonomen der Manufakturperiode charakteri-
siert, ist der Akzent, den er auf die Teilung der Arbeiter legt. Die untergeordnete
Rolle, die er der Maschinerie anweist, rief im Beginn der großen Industrie Lauderda-
les, in einer weiterentwickelten Epoche Ures Polemik hervor. A. Smith verwechselt
auch die Differenzierung der Instrumente, wobei die Teilarbeiter der Manufaktur
selbst sehr tätig waren, mit der Maschinenerfindung. Es sind nicht die Manufaktur-
arbeiter, sondern Gelehrte, Handwerker, selbst Bauern (Brindley) usw., die hier eine
Rolle spielen.
(45) "Indem man das Machwerk in mehrere verschiedene Operationen teilt, deren
jede verschiedene Grade von Gewandtheit und Kraft erheischt, kann der Manufaktur-
herr sich genau das jeder Operation entsprechende Quantum von Kraft und Gewand-
theit verschaffen. Wäre dagegen das ganze Werk von einem Arbeiter zu verrichten, so
müßte dasselbe Individuum genug Gewandtheit für die delikatesten und genug Kraft
für die mühseligsten Operationen besitzen." (Ch. Babbage, l.c., ch. XIX.)
⋮
```

- 2) ein gigantischer, über 150 Mio. Zeichen umfassender Quelltext (formatierte Lexikoneinträge):

```
⋮
Hauptseite | See_live_article [search] [Search]
***** Claude Shannon *****
Claude Elwood Shannon (* 30. April 1916 in Petoskey, Michigan; † 24.
Februar 2001 in Medford, Massachusetts) war ein US-amerikanischer Mathematiker.
Er gilt als Begründer der Informationstheorie. Shannon wuchs auf in Gaylord,
Michigan, welches oft auch als Geburtsort angegeben wird, und arbeitete dort in
seiner High_School-Zeit als Bote für die Western Union. Er war ein Mitglied von
Tau Beta Pi. 1932 begann er ein Elektroingenieur- und Mathematikstudium an der
University of Michigan, die er 1936 mit einem Abschluss in Mathematik und
Elektrotechnik verließ, um an das MIT zu wechseln. In seiner Abschlussarbeit
zum Master in Elektrotechnik, A Symbolic Analysis of Relay and Switching
Circuits, benutzte er Boolesche Algebra zur Konstruktion von digitalen
Schaltkreisen. Außerdem erwarb er seinen Dokortitel in Mathematik mit einer
Arbeit über theoretische Genetik (1940). Nach kurzem Aufenthalt als Forscher am
Institute for Advanced Study in Princeton, New Jersey, kam er 1941 als
Mathematiker zu den ebenfalls in New Jersey gelegenen AT&T Bell Labs.
```

² "Das Kapital", Bände 1–3

- :
- 3) eine textuell vorliegende Grafikdatei; in ihr stehen ausschließlich numerische Daten: gebrochene Koordinatenmaße wie auch ganzzahlige Indizes, beides im Dezimalziffern-Format (ca. 2,5 Mio. Zeichen, am Anfang ein vernachlässigbarer textueller Dateikopf):

```

OFF
34834 69451 0
-0.0378297 0.12794 0.00447467
-0.0447794 0.128887 0.00190497
-0.0680095 0.151244 0.0371953
-0.00228741 0.13015 0.0232201
-0.0226054 0.126675 0.00715587
-0.0251078 0.125921 0.00624226
-0.0371209 0.127449 0.0017956
0.033213 0.112692 0.0276861
-0.0255083 0.112568 0.0366767
-0.0245306 0.112636 0.0373469
0.0274031 0.12156 0.0212208
-0.0628961 0.158419 -0.0175871
0.0400813 0.104202 0.0221684
:
3 33992 33991 33864
3 33865 33992 33864
3 34117 34116 33991
3 33992 34117 33991
3 21538 21443 21539
3 33718 33844 16683
3 16743 33844 16684
3 33844 16743 16683
3 16683 16743 16742

```

Exemplarisch sei hier für den ersten Quelltext die Häufigkeitsanalyse in der alias-Variante wiedergegeben:

Buchstabe “_” (19.2%):

- VOR-Kontexte: _ (27.9%) N (15.8%) E (10.4%) R (9.9%) T (9.2%) S (7.3%) D (3.7%) M (2.7%) H (2.6%) L (2.0%) G (1.9%) B (0.9%) O (0.9%) F (0.9%) U (0.7%) I (0.7%) Z (0.4%) B (0.4%) A (0.3%) K (0.3%) C (0.2%) P (0.2%) W (0.2%) V (0.2%) Y (0.1%) X (0.0%) J (0.0%) Q (0.0%) Ü (0.0%)
- NACH-Kontexte: _ (27.9%) D (12.6%) A (6.3%) S (5.4%) W (4.6%) E (4.5%) I (4.4%) V (3.6%) G (3.3%) U (3.0%) B (2.9%) K (2.9%) P (2.6%) Z (2.5%) M (2.4%) N (2.3%) F (1.9%) H (1.4%) R (1.1%) T (1.0%) L (0.9%) O (0.9%) J (0.6%) Ü (0.4%) C (0.4%) Q (0.1%) Ä (0.1%) Ö (0.1%) X (0.0%) Y (0.0%)

Buchstabe “A” (4.6%):

- VOR-Kontexte: _ (26.0%) T (10.4%) K (8.8%) D (8.7%) R (6.5%) H (5.7%) L (5.1%) W (5.1%) M (4.8%) N (3.8%) S (2.7%) F (2.7%) B (2.3%) G (1.8%) Z (1.0%) I (0.9%) V (0.8%) J (0.8%) P (0.7%) U (0.5%) E (0.4%) C (0.3%) A (0.2%) X (0.0%) O (0.0%) Y (0.0%) B (0.0%)
- NACH-Kontexte: L (16.1%) N (14.1%) R (12.4%) U (12.0%) T (6.8%) P (6.3%) S (6.2%) B (5.0%) C (3.9%) G (3.1%) H (3.1%) M (2.4%) F (2.3%) B (2.2%) _ (1.5%) K (1.2%) D (0.7%) V (0.3%) I (0.2%) A (0.2%) Y (0.1%) Z (0.1%) X (0.1%) W (0.0%) E (0.0%) J (0.0%) O (0.0%) Q (0.0%)

Buchstabe “B” (1.6%):

- VOR-Kontexte: _ (34.8%) R (17.4%) A (14.4%) L (8.3%) E (8.1%) Ü (5.2%) O (2.7%) I (2.6%) N (1.5%) U (1.4%) T (0.9%) S (0.7%) H (0.5%) M (0.4%) D (0.3%) Z (0.2%) G (0.1%) K (0.1%) B (0.1%) F (0.1%) B (0.1%) Ä (0.1%) Ö (0.0%) V (0.0%) Y (0.0%) C (0.0%) P (0.0%)
- NACH-Kontexte: E (59.9%) A (6.7%) S (5.0%) R (4.8%) _ (4.6%) L (4.3%) I (4.0%) T (2.3%) O (2.2%) G (1.3%) U (1.3%) N (1.1%) H (0.6%) W (0.4%) Z (0.3%) Ü (0.3%) F (0.2%) Ä (0.2%) J (0.1%) Y (0.1%) B (0.1%) K (0.1%) Ö (0.1%) D (0.0%) M (0.0%) P (0.0%) V (0.0%)

Buchstabe “C” (1.9%):

- VOR-Kontexte: I (31.3%) S (28.9%) A (9.3%) R (5.8%) U (5.3%) E (4.5%) _ (3.7%) O (3.6%) Ü (2.5%) L (1.8%) Ä (1.5%) T (0.7%) N (0.6%) Ö (0.3%) D (0.1%) C (0.1%) X (0.0%) W (0.0%) V (0.0%) K (0.0%) H (0.0%) M (0.0%) F (0.0%) Y (0.0%) B (0.0%) G (0.0%) Z (0.0%)

- NACH-Kontexte: H (88.0%) K (5.7%) -- (2.5%) O (0.9%) A (0.6%) T (0.5%) E (0.5%) R (0.4%) I (0.3%) U (0.3%) L (0.1%) Y (0.1%) C (0.1%) Q (0.0%) W (0.0%) S (0.0%) F (0.0%) D (0.0%) N (0.0%) P (0.0%) B (0.0%) X (0.0%) G (0.0%) M (0.0%) Z (0.0%)

Buchstabe "D" (4.6%):

- VOR-Kontexte: -- (53.0%) N (23.1%) O (8.3%) E (4.0%) R (4.0%) L (3.4%) I (1.2%) F (1.2%) A (0.7%) S (0.3%) U (0.2%) Ä (0.1%) M (0.1%) T (0.1%) H (0.1%) D (0.0%) Ü (0.0%) K (0.0%) B (0.0%) G (0.0%) Y (0.0%) Z (0.0%) Ö (0.0%) V (0.0%) C (0.0%) P (0.0%) ß (0.0%)
- NACH-Kontexte: E (42.3%) I (17.5%) -- (15.4%) U (8.9%) A (8.8%) R (2.2%) N (0.8%) L (0.8%) O (0.7%) T (0.5%) S (0.4%) K (0.3%) W (0.2%) F (0.2%) Ü (0.1%) M (0.1%) H (0.1%) P (0.1%) B (0.1%) G (0.1%) V (0.1%) Z (0.0%) C (0.0%) D (0.0%) Ä (0.0%) Ö (0.0%) Q (0.0%) Y (0.0%) J (0.0%)

Buchstabe "E" (12.8%):

- VOR-Kontexte: D (15.1%) I (10.2%) T (9.2%) G (8.3%) B (7.5%) -- (6.8%) R (6.7%) S (6.2%) N (5.2%) H (4.7%) W (4.3%) L (3.9%) V (2.9%) M (2.7%) Z (1.8%) K (1.2%) F (1.0%) ß (0.6%) P (0.6%) U (0.6%) J (0.5%) X (0.1%) E (0.1%) C (0.1%) Y (0.0%) O (0.0%) A (0.0%) Ä (0.0%) Q (0.0%)
- NACH-Kontexte: R (23.4%) N (20.4%) -- (15.6%) I (12.6%) S (8.5%) L (4.9%) M (2.3%) T (2.2%) H (2.1%) D (1.4%) G (1.2%) B (1.0%) C (0.7%) U (0.6%) W (0.6%) ß (0.4%) F (0.4%) P (0.3%) Z (0.3%) K (0.3%) X (0.3%) V (0.2%) A (0.2%) E (0.1%) O (0.1%) Y (0.0%) Q (0.0%) Ä (0.0%) J (0.0%) Ü (0.0%) Ö (0.0%)

Buchstabe "F" (1.3%):

- VOR-Kontexte: -- (29.2%) U (20.5%) A (8.2%) O (7.5%) P (6.5%) F (4.1%) R (3.8%) N (3.7%) E (3.7%) I (2.8%) Ä (2.2%) S (1.7%) L (1.6%) H (0.9%) T (0.9%) D (0.7%) M (0.7%) K (0.4%) Ö (0.3%) B (0.3%) G (0.2%) Z (0.1%) Ü (0.0%) X (0.0%) ß (0.0%) V (0.0%) C (0.0%)
- NACH-Kontexte: T (13.1%) -- (12.9%) E (10.2%) A (9.6%) O (9.0%) I (8.5%) Ü (8.4%) R (4.6%) U (4.4%) D (4.1%) F (4.1%) L (3.4%) Ä (2.6%) S (1.8%) M (1.0%) N (0.8%) G (0.5%) Z (0.3%) H (0.3%) W (0.2%) Ö (0.1%) B (0.1%) P (0.1%) K (0.1%) J (0.0%) Y (0.0%) V (0.0%) C (0.0%) X (0.0%)

Buchstabe "G" (2.1%):

- VOR-Kontexte: -- (30.3%) N (24.0%) I (15.3%) E (7.2%) A (6.7%) R (3.8%) S (2.4%) U (2.1%) L (2.0%) B (1.0%) Ä (0.8%) O (0.8%) T (0.7%) M (0.6%) Ö (0.6%) Ü (0.5%) F (0.3%) H (0.3%) K (0.2%) D (0.2%) G (0.1%) Z (0.1%) V (0.0%) ß (0.0%) Y (0.0%) C (0.0%) X (0.0%)
- NACH-Kontexte: E (49.8%) -- (17.3%) R (6.0%) L (5.9%) T (4.6%) A (3.9%) S (3.0%) U (2.5%) I (2.3%) N (1.4%) O (1.0%) K (0.9%) Ü (0.4%) Ä (0.2%) H (0.1%) G (0.1%) F (0.1%) W (0.1%) M (0.1%) B (0.1%) Z (0.0%) D (0.0%) Y (0.0%) Ö (0.0%) J (0.0%) P (0.0%) V (0.0%) Q (0.0%) C (0.0%)

Buchstabe "H" (3.0%):

- VOR-Kontexte: C (56.3%) -- (9.2%) E (8.9%) I (4.9%) A (4.7%) R (3.6%) O (2.7%) Ä (2.1%) T (1.8%) Ü (1.0%) P (1.0%) S (0.8%) N (0.7%) Ö (0.7%) U (0.6%) B (0.3%) D (0.2%) H (0.1%) F (0.1%) L (0.1%) G (0.1%) W (0.1%) K (0.1%) ß (0.0%) M (0.0%) Z (0.0%) X (0.0%) Y (0.0%)
- NACH-Kontexte: E (20.0%) -- (16.5%) R (11.7%) T (10.0%) A (8.8%) I (5.7%) N (5.6%) L (4.8%) Ä (3.9%) S (3.3%) M (1.9%) O (1.8%) U (1.6%) Ö (1.0%) W (0.9%) F (0.4%) Ü (0.4%) K (0.3%) B (0.3%) Z (0.3%) G (0.2%) D (0.1%) Y (0.1%) H (0.1%) P (0.0%) V (0.0%) J (0.0%) C (0.0%)

Buchstabe "I" (6.7%):

- VOR-Kontexte: E (24.2%) -- (12.8%) D (12.0%) T (7.7%) L (6.1%) S (5.1%) W (4.7%) R (4.7%) P (4.6%) N (3.8%) M (3.4%) H (2.6%) Z (2.1%) F (1.6%) B (1.0%) V (0.9%) G (0.7%) I (0.6%) X (0.3%) K (0.3%) U (0.2%) ß (0.2%) A (0.2%) O (0.1%) C (0.1%) Ä (0.0%) Y (0.0%) J (0.0%)
- NACH-Kontexte: E (19.6%) N (18.8%) T (15.7%) C (9.0%) S (8.9%) G (4.9%) O (4.3%) R (3.2%) L (3.1%) H (2.2%) M (2.2%) -- (2.0%) V (1.3%) D (0.9%) K (0.8%) A (0.6%) B (0.6%) I (0.6%) F (0.5%) X (0.2%) ß (0.2%) Z (0.2%) P (0.1%) U (0.1%) Q (0.0%) W (0.0%) Ö (0.0%) J (0.0%) Ä (0.0%)

Buchstabe "J" (0.1%):

- VOR-Kontexte: -- (93.2%) B (1.4%) R (1.2%) N (1.0%) S (0.8%) E (0.6%) L (0.3%) G (0.2%) A (0.2%) I (0.2%) F (0.2%) O (0.2%) T (0.2%) U (0.2%) H (0.1%) D (0.1%) K (0.0%) M (0.0%) Z (0.0%) ß (0.0%)
- NACH-Kontexte: E (52.2%) A (27.9%) Ä (9.6%) U (6.0%) O (2.1%) -- (1.7%) Ü (0.5%) I (0.1%) M (0.1%) K (0.0%) N (0.0%)

Buchstabe "K" (1.5%):

- VOR-Kontexte: -- (37.4%) U (14.6%) R (12.4%) C (7.4%) N (6.8%) A (3.7%) I (3.5%) S (3.5%) E (2.6%) G (1.3%) D (1.1%) L (1.0%) K (1.0%) Ö (0.8%) H (0.7%) T (0.7%) O (0.6%) V (0.5%) Y (0.1%) Z (0.1%) M (0.1%) B (0.1%) F (0.1%) ß (0.0%) X (0.0%) Ä (0.0%) P (0.0%) J (0.0%) W (0.0%) Ü (0.0%)
- NACH-Kontexte: A (27.5%) T (23.3%) E (10.3%) O (9.7%) U (6.6%) R (5.3%) L (5.2%) -- (3.3%) Ö (1.5%) I (1.2%) S (1.0%) K (1.0%) Ä (0.9%) Ü (0.7%) Z (0.5%) N (0.4%) F (0.3%) G (0.3%) V (0.2%) W (0.2%) H (0.1%) B (0.1%) M (0.1%) D (0.1%) P (0.0%) C (0.0%) Y (0.0%) J (0.0%) Q (0.0%)

Buchstabe "L" (3.2%):

- VOR-Kontexte: A (23.4%) E (20.0%) L (10.2%) I (6.5%) -- (5.6%) H (4.6%) O (4.4%) G (3.9%) U (3.1%) K (2.4%) B (2.2%) Ä (2.1%) R (2.1%) T (2.0%) F (1.4%) S (1.2%) D (1.1%) N (1.0%) P (0.9%) M (0.5%) Z (0.4%) Ö (0.3%) Ü (0.2%) ß (0.2%) C (0.1%) Y (0.0%) X (0.0%) V (0.0%) W (0.0%)
- NACH-Kontexte: E (15.6%) I (12.8%) -- (12.4%) S (10.5%) L (10.2%) A (7.5%) T (7.1%) D (5.0%) B (4.2%) U (3.1%) O (2.9%) G (1.4%) N (1.3%) Ä (1.3%) C (1.1%) F (0.6%) K (0.5%) M (0.4%) Ö (0.4%) W (0.4%) Ü (0.3%) Z (0.3%) R (0.2%) Y (0.2%) H (0.1%) V (0.1%) P (0.1%) J (0.0%) X (0.0%) Q (0.0%)

Buchstabe "M" (1.8%):

- VOR-Kontexte: -- (25.6%) E (16.8%) U (12.5%) I (8.1%) M (6.9%) R (6.4%) A (6.2%) O (5.2%) S (3.7%) H (3.3%) N (1.3%) F (0.7%) L (0.7%) Ü (0.6%) T (0.5%) Ä (0.5%) D (0.3%) P (0.1%) Ö (0.1%) G (0.1%) Z (0.1%) Y (0.1%) K (0.1%) W (0.0%) ß (0.0%) B (0.0%) J (0.0%) C (0.0%) X (0.0%)
- NACH-Kontexte: -- (28.7%) E (19.1%) I (12.8%) A (12.4%) M (6.9%) T (4.5%) U (3.0%) S (2.8%) O (1.8%) Ä (1.3%) P (1.1%) L (0.9%) Ü (0.9%) W (0.7%) Ö (0.7%) G (0.7%) F (0.5%) B (0.3%) D (0.3%) N (0.1%) K (0.1%) Y (0.1%) Z (0.1%) R (0.0%) V (0.0%) H (0.0%) C (0.0%) J (0.0%) Q (0.0%) ß (0.0%)

Buchstabe "N" (7.4%):

- VOR-Kontexte: E (35.6%) I (17.1%) U (11.8%) O (9.3%) A (8.9%) -- (5.9%) N (2.5%) H (2.3%) R (1.9%) Ä (1.5%) L (0.6%) D (0.5%) T (0.5%) G (0.4%) Ü (0.4%) Ö (0.3%) B (0.2%) ß (0.2%) F (0.1%) S (0.1%) K (0.1%) M (0.0%) Y (0.0%) W (0.0%) Z (0.0%) C (0.0%) P (0.0%) V (0.0%) J (0.0%)
- NACH-Kontexte: -- (41.3%) D (14.4%) E (9.0%) G (6.9%) T (5.4%) S (4.8%) I (3.5%) N (2.5%) A (2.4%) U (1.9%) Z (1.5%) K (1.4%) O (1.3%) F (0.7%) L (0.4%) W (0.4%) B (0.3%) M (0.3%) H (0.3%) Ä (0.3%) P (0.2%) Ü (0.2%) C (0.2%) V (0.2%) R (0.2%) Ö (0.1%) Q (0.0%) J (0.0%) Y (0.0%)

Buchstabe "O" (2.4%):

- VOR-Kontexte: R (17.6%) V (14.7%) I (11.9%) S (11.5%) -- (7.3%) K (6.0%) W (4.9%) F (4.8%) N (3.9%) L (3.8%) T (2.4%) H (2.3%) P (1.9%) B (1.5%) D (1.4%) M (1.3%) G (0.8%) C (0.7%) E (0.3%) Z (0.3%) O (0.3%) J (0.1%) U (0.1%) Y (0.0%) X (0.0%) A (0.0%) Ä (0.0%) Ö (0.0%) ß (0.0%)
- NACH-Kontexte: N (28.5%) D (15.9%) R (13.1%) -- (7.2%) L (5.8%) F (4.0%) M (3.9%) H (3.3%) C (2.9%) S (2.2%) T (2.2%) ß (2.1%) Z (1.9%) B (1.8%) P (1.3%) W (0.9%) G (0.7%) U (0.6%) V (0.4%) K (0.4%) I (0.3%) O (0.3%) Y (0.1%) E (0.0%) A (0.0%) X (0.0%) J (0.0%) Q (0.0%)

Buchstabe "P" (1.2%):

- VOR-Kontexte: -- (42.4%) A (24.8%) S (14.5%) E (3.4%) O (2.7%) R (2.0%) U (1.9%) M (1.7%) T (1.5%) N (1.2%) X (1.0%) P (0.9%) I (0.7%) D (0.4%) Ö (0.3%) L (0.3%) Y (0.1%) F (0.1%) H (0.1%) Ü (0.1%) K (0.0%) ß (0.0%) Z (0.0%) B (0.0%) G (0.0%) C (0.0%) Ä (0.0%) W (0.0%)
- NACH-Kontexte: R (39.5%) I (25.9%) F (7.0%) E (6.6%) -- (3.8%) O (3.8%) A (2.9%) H (2.4%) L (2.4%) T (1.8%) U (1.5%) Ä (1.0%) P (0.9%) M (0.2%) S (0.1%) Ö (0.0%) Y (0.0%) Z (0.0%) D (0.0%) Ü (0.0%) N (0.0%) B (0.0%) K (0.0%) W (0.0%)

Buchstabe "Q" (0.0%):

- VOR-Kontexte: -- (61.7%) Ä (13.3%) I (5.9%) S (4.6%) N (3.8%) E (3.8%) T (1.6%) R (1.3%) D (1.2%) Q (1.0%) C (0.5%) O (0.5%) L (0.3%) A (0.2%) G (0.1%) K (0.0%) M (0.0%) X (0.0%)
- NACH-Kontexte: U (79.7%) R (17.9%) -- (1.3%) Q (1.0%) E (0.1%)

Buchstabe "R" (6.4%):

- VOR-Kontexte: E (47.2%) A (9.0%) P (7.3%) H (5.5%) U (5.2%) O (4.9%) I (3.4%) -- (3.3%) T (2.8%) G (2.0%) Ü (1.9%) D (1.6%) K (1.2%) B (1.2%) F (0.9%) R (0.7%) Ä (0.6%) Ö (0.3%) N (0.2%) S (0.2%) Q (0.1%) C (0.1%) ß (0.1%) L (0.1%) Z (0.0%) V (0.0%) M (0.0%) W (0.0%) Y (0.0%)
- NACH-Kontexte: -- (30.1%) E (13.5%) T (7.1%) O (6.6%) I (4.9%) A (4.7%) S (4.6%) B (4.4%) K (2.9%) D (2.8%) N (2.2%) U (1.9%) M (1.8%) C (1.7%) H (1.7%) W (1.6%) G (1.3%) L (1.0%) Ü (1.0%) F (0.8%) R (0.7%) Ö (0.6%) Z (0.6%) Ä (0.6%) P (0.4%) V (0.2%) Y (0.1%) J (0.0%) X (0.0%) Q (0.0%)

Buchstabe "S" (5.3%):

- VOR-Kontexte: E (20.7%) -- (19.7%) I (11.3%) N (6.7%) L (6.3%) U (5.8%) T (5.6%) R (5.5%) A (5.5%) S (4.0%) H (1.9%) B (1.5%) G (1.2%) O (1.0%) M (1.0%) F (0.4%) Ü (0.4%) D (0.4%) Y (0.3%) K (0.3%) Ö (0.2%) Ä (0.1%) ß (0.0%) P (0.0%) Z (0.0%) W (0.0%) C (0.0%) V (0.0%) X (0.0%)
- NACH-Kontexte: -- (26.8%) T (17.3%) E (15.0%) C (10.5%) I (6.5%) O (5.3%) S (4.0%) P (3.2%) A (2.4%) U (1.4%) M (1.3%) K (1.0%) G (1.0%) W (0.8%) L (0.7%) Z (0.6%) H (0.4%) F (0.4%) D (0.3%) Ä (0.2%) R (0.2%) B (0.2%) Y (0.2%) N (0.1%) V (0.1%) Ü (0.0%) Q (0.0%) Ö (0.0%) J (0.0%)

Buchstabe "T" (5.4%):

- VOR-Kontexte: I (19.5%) S (17.0%) R (8.5%) N (7.4%) K (6.4%) A (5.8%) H (5.6%) E (5.4%) L (4.2%) -- (3.6%) T (3.1%) F (3.1%) G (1.8%) M (1.5%) Z (1.3%) U (1.3%) O (1.0%) Ä (0.8%) B (0.7%) ß (0.5%) D (0.4%) P (0.4%) C (0.2%) Ö (0.2%) Ü (0.2%) X (0.1%) Y (0.0%) W (0.0%) V (0.0%)
- NACH-Kontexte: -- (33.1%) E (22.0%) I (9.6%) A (9.0%) S (5.5%) R (3.3%) T (3.1%) Z (3.1%) U (2.4%) Ä (1.3%) L (1.2%) O (1.1%) W (1.1%) H (1.0%) N (0.6%) Ü (0.5%) P (0.3%) G (0.3%) B (0.3%) C (0.2%) F (0.2%) K (0.2%) M (0.2%) V (0.1%) Ö (0.1%) D (0.1%) Y (0.0%) Q (0.0%) J (0.0%)

Buchstabe "U" (3.0%):

- VOR-Kontexte: -- (19.2%) A (18.7%) D (13.8%) Z (10.8%) N (4.7%) T (4.4%) R (4.0%) K (3.3%) L (3.3%) E (2.6%) S (2.4%) F (1.9%) M (1.8%) G (1.8%) H (1.6%) Q (1.2%) Ä (1.1%) W (0.7%) B (0.7%) P (0.6%) O (0.5%) J (0.3%) C (0.2%) I (0.1%) U (0.1%) X (0.1%) V (0.0%) ß (0.0%)
- NACH-Kontexte: N (29.3%) R (11.1%) S (10.4%) F (8.9%) M (7.5%) K (7.3%) -- (4.9%) C (3.4%) L (3.4%) E (2.4%) T (2.3%) ß (2.2%) G (1.5%) Z (1.3%) P (0.8%) B (0.7%) A (0.7%) H (0.7%) I (0.4%) D (0.3%) W (0.2%) X (0.1%) V (0.1%) U (0.1%) O (0.1%) J (0.0%) Ä (0.0%) Ü (0.0%) Y (0.0%) Ö (0.0%)

Buchstabe "V" (0.9%):

- VOR-Kontexte: -- (78.6%) I (9.8%) E (2.6%) R (1.8%) A (1.6%) N (1.3%) O (1.1%) S (0.8%) T (0.6%) L (0.4%) K (0.4%) U (0.3%) D (0.3%) X (0.1%) H (0.1%) B (0.1%) M (0.1%) Z (0.0%) B (0.0%) Ö (0.0%) G (0.0%) F (0.0%)
- NACH-Kontexte: E (42.9%) O (40.5%) I (6.6%) A (4.2%) -- (3.6%) Ö (0.9%) K (0.8%) U (0.1%) G (0.1%) R (0.1%) D (0.0%) Ä (0.0%) F (0.0%) Y (0.0%) S (0.0%) C (0.0%) B (0.0%) L (0.0%) N (0.0%) T (0.0%) Z (0.0%)

Buchstabe "W" (1.4%):

- VOR-Kontexte: -- (64.7%) R (7.5%) E (5.3%) Z (5.2%) T (4.3%) S (3.3%) N (2.2%) H (1.9%) O (1.6%) M (1.0%) L (0.9%) D (0.7%) B (0.5%) U (0.3%) K (0.2%) F (0.2%) G (0.1%) A (0.1%) I (0.0%) ß (0.0%) C (0.0%) Ö (0.0%) X (0.0%) P (0.0%)
- NACH-Kontexte: E (40.4%) I (23.3%) A (17.4%) O (8.7%) Ä (3.9%) -- (2.4%) U (1.6%) Ü (1.2%) Ö (0.7%) H (0.2%) M (0.1%) N (0.0%) R (0.0%) S (0.0%) C (0.0%) T (0.0%) Y (0.0%) L (0.0%) Z (0.0%) K (0.0%) P (0.0%)

Buchstabe "X" (0.1%):

- VOR-Kontexte: E (51.0%) I (24.5%) -- (9.3%) U (5.0%) A (3.8%) X (2.8%) R (2.3%) O (0.9%) L (0.3%) C (0.1%) F (0.0%) Ä (0.0%)
- NACH-Kontexte: I (29.1%) E (21.9%) P (17.6%) -- (11.7%) T (7.3%) U (3.0%) X (2.8%) V (1.7%) A (1.6%) C (1.0%) K (0.6%) F (0.5%) O (0.3%) L (0.3%) Z (0.2%) H (0.1%) Y (0.1%) W (0.1%) G (0.0%) M (0.0%) Q (0.0%) S (0.0%)

Buchstabe "Y" (0.0%):

- VOR-Kontexte: S (19.5%) L (11.5%) R (10.8%) A (10.7%) H (7.9%) E (7.0%) T (4.8%) -- (4.7%) Z (4.1%) O (3.7%) B (3.6%) C (3.1%) M (3.1%) N (1.9%) G (1.1%) D (1.1%) V (0.4%) P (0.3%) W (0.2%) F (0.2%) K (0.2%) X (0.1%) U (0.0%)
- NACH-Kontexte: -- (42.7%) S (33.8%) K (4.1%) M (3.5%) P (3.0%) E (2.4%) O (2.4%) L (1.8%) N (1.3%) D (1.3%) I (1.1%) A (1.0%) R (0.6%) T (0.5%) G (0.2%) B (0.2%) C (0.1%) H (0.1%)

Buchstabe "Z" (1.0%):

- VOR-Kontexte: -- (47.6%) T (16.5%) N (11.2%) O (4.6%) R (3.9%) E (3.8%) U (3.7%) S (3.0%) I (1.3%) L (1.0%) H (0.8%) K (0.7%) B (0.5%) A (0.4%) F (0.3%) D (0.2%) M (0.1%) G (0.1%) Z (0.1%) ß (0.0%) X (0.0%) P (0.0%) W (0.0%) C (0.0%) V (0.0%) Ä (0.0%)
- NACH-Kontexte: U (32.1%) E (22.7%) I (14.0%) -- (8.3%) W (7.1%) T (6.8%) A (4.7%) L (1.3%) O (0.7%) Ä (0.3%) B (0.3%) Ü (0.2%) G (0.2%) Y (0.2%) K (0.2%) Ö (0.2%) M (0.2%) F (0.1%) R (0.1%) D (0.1%) S (0.1%) Z (0.1%) P (0.0%) V (0.0%) N (0.0%) H (0.0%) C (0.0%) J (0.0%)

Buchstabe "Ä" (0.5%):

- VOR-Kontexte: H (24.5%) T (14.4%) W (11.3%) L (8.6%) R (8.2%) F (7.0%) M (4.9%) N (4.0%) _ (3.5%) K (3.0%) S (2.6%) J (2.6%) P (2.4%) G (1.1%) Z (0.7%) B (0.5%) E (0.2%) D (0.2%) V (0.0%) U (0.0%) I (0.0%)

- NACH-Kontexte: N (23.3%) L (14.2%) H (13.1%) T (9.1%) R (8.8%) U (7.1%) F (6.0%) C (6.0%) G (3.6%) B (3.2%) M (1.7%) Q (1.3%) D (1.2%) S (1.0%) B (0.2%) I (0.2%) E (0.0%) K (0.0%) P (0.0%) O (0.0%) X (0.0%) Z (0.0%)

Buchstabe "Ö" (0.2%):

- VOR-Kontexte: R (24.0%) H (17.6%) K (13.4%) _ (8.3%) M (7.8%) L (7.1%) W (5.4%) N (4.9%) V (4.6%) T (2.5%) Z (1.1%) F (0.8%) S (0.8%) B (0.5%) D (0.3%) I (0.3%) P (0.2%) G (0.2%) E (0.0%) U (0.0%)

- NACH-Kontexte: B (20.7%) R (12.4%) N (12.3%) H (11.9%) G (7.4%) K (7.2%) L (6.3%) S (5.7%) T (5.5%) C (3.8%) F (2.7%) P (2.2%) M (1.2%) B (0.4%) D (0.2%) V (0.1%) W (0.0%) O (0.0%)

Buchstabe "Ü" (0.4%):

- VOR-Kontexte: F (28.8%) _ (19.4%) R (16.8%) T (7.9%) W (4.4%) M (4.1%) N (3.4%) H (3.2%) K (2.8%) L (2.7%) G (2.0%) D (1.8%) B (1.2%) Z (0.6%) S (0.6%) J (0.2%) E (0.0%) P (0.0%) U (0.0%)

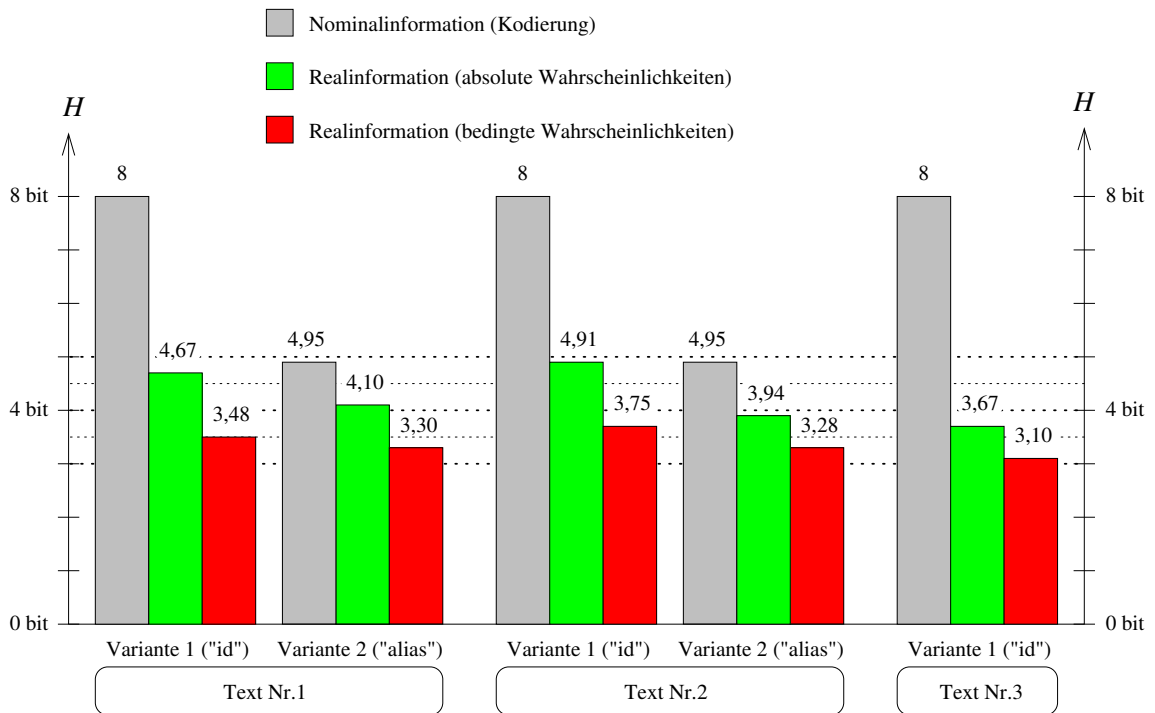
- NACH-Kontexte: R (32.4%) B (22.5%) C (12.6%) H (7.9%) N (7.2%) S (5.9%) M (2.9%) G (2.6%) T (2.2%) L (2.0%) B (1.1%) D (0.3%) P (0.2%) F (0.1%) _ (0.0%) K (0.0%)

Buchstabe "ß" (0.3%):

- VOR-Kontexte: A (30.4%) U (19.7%) O (15.3%) E (14.4%) Ö (10.2%) Ä (4.5%) I (4.2%) Ü (1.3%) M (0.0%)

- NACH-Kontexte: _ (52.8%) E (24.8%) T (8.3%) N (4.0%) I (3.6%) R (2.1%) L (2.0%) S (0.7%) B (0.4%) H (0.2%) K (0.2%) V (0.2%) G (0.2%) P (0.1%) W (0.1%) M (0.1%) U (0.1%) F (0.1%) Z (0.0%) A (0.0%) C (0.0%) D (0.0%) J (0.0%) O (0.0%)

Die Wahrscheinlichkeiten der "Nach-Kontexte" sind hier genau die $p_y(x)$. Aus ihnen ergibt sich der optimierte, weil bedingte Informationsgehalt, im folgenden Diagramm rot dargestellt:



Programmtext

Das Programm wurde in der Programmiersprache C implementiert:

```
//
// Eingabe: Textdatei, möglichst lang
//
// Ausgabe: statistische Verteilung
//          von Zeichenabhängigkeiten
//          (bedingte rel. Häufigkeiten),
//          sowie Informationsmaß daraus
//
// Autor: Bertrand Klimmek, Juni 2004
//       (kein Urheberrecht)
//

#include <stdio.h>

#include <math.h>

#define UMFANG 31 // 26 Buchstaben + 3 Umlaute + "ß" + Blank

// #define UMFANG 256 // "id"-Abbildung, komplette Erfassung

int main (int argc, char *argv[])
{

    if (argc!=3) // syntaktisch falscher Aufruf
    {
        printf("\n Aufruf:\n =====\n\n %s <???.txt> <statistik.txt>\n\n",argv[0]);
        return(0);
    }

    int
        i,j,k,l,m,n;
```

```

// -----
//  Aufbauen der alias-Tabelle
// -----
//
// diese ordnet jedem Byte einen Buchstaben zu,
// Sonderzeichen, Ziffern etc. werden auf Blanks abgebildet;
// darüberhinaus wird Groß- und Kleinschreibung eingeebnet

int
  alias[256], // alias-Tabelle (Verweis: Byte -> Nr.)

  output[UMFANG]; // quasi Umkehrfunktion von "alias[...]"

for (i=0; i<UMFANG; i++) // im Zweifelsfalle: "keine Ahnung!" ausgeben
  output[i]='?';

// /*

for (i=0; i<256; i++) // im Zweifelsfalle ist alles ein Blank
  alias[i]=0; // so z.B. alle Ziffern, Interpunktions- und Sonderzeichen

output[0]='_'; // Blank wird als Unterstrich ausgegeben

// Buchstaben bekommen Nummern 1-26:

for (i='A'; i<='Z'; i++)
{
  alias[i] = (i - 'A') + 1 ; // Großbuchstaben ausfüllen

  alias[i+32]=alias[i];

  output[alias[i]] = i; // inverser Verweis
}

for (i='a'; i<='z'; i++)

  alias[i] = (i - 'a') + 1 ; // Kleinbuchstaben ausfüllen

i=26; // hinter die normalen Buchstaben ...

// Umlaute gesondert (27-29):

alias[(unsigned char)'Ä'] = ++i;
alias[(unsigned char)'ä']=alias[(unsigned char)'Ä'];
output[i]=(unsigned char)'Ä';

alias[(unsigned char)'Ö'] = ++i;
alias[(unsigned char)'ö']=alias[(unsigned char)'Ö'];
output[i]=(unsigned char)'Ö';

alias[(unsigned char)'Ü'] = ++i;
alias[(unsigned char)'ü']=alias[(unsigned char)'Ü'];
output[i]=(unsigned char)'Ü';

```

```

// und "ß" bzw. 'sz' (30)

alias[(unsigned char)'ß'] = ++i;
output[i]=(unsigned char)'ß';

// */

// for (i=0; i<UMFANG; i++) alias[i]=i; // "id"
// for (i=32; i<UMFANG; i++) output[i]=i; // "id" (Steuerzeichen nicht ausgeben!)

// Kontrollausgabe der alias- bzw. output-Tabelle

for (i=0; i<UMFANG; i++)
    printf(" (%c)", output[i]);
printf("\n\n");

for (i=0; i<256; i++)
    printf(" (%c)", output[ alias[i] ]);

// -----
// Deklarieren der Häufigkeitstabellen
// -----
//
// eine Tabelle für Nachfolger[i,j] und
// eine Tabelle für Vorgänger[i,j]
// für alle (i,j) aus {0,...,UMFANG-1}^2

int
    nachf[UMFANG][UMFANG];

// ... natürlich als leer initialisieren
// (noch keine Vorkommen):

for (i=0; i<UMFANG; i++)
    for (j=0; j<UMFANG; j++)

        nachf[i][j]=0;

FILE
    *ein,
    *aus;

char *ein_ =argv[1];
char *aus_ =argv[2];

ein = fopen(ein_,"rb"); // Eingabedatei: TXT (ASCII)

aus = fopen(aus_,"wb"); // Ausgabedatei: TXT (ASCII)

```

```

////////////////////////////////////
//
// Hauptschleife: Textanalyse (Datenerhebung)
// -----
////////////////////////////////////

int
  c, // Original-ASCII-Zeichen
  c0, // vorhergehender Buchstabe
  c1; // aktueller Buchstabe

c0 = 0; // (willkürliche) Initialisierung: als Blank

c = fgetc( ein ); // nächstes Zeichen lesen

while (c+1) // bis Dateiende
{
  c1 = alias[c]; // Eingabezeichen nun auf Tabellenformat zurechtgeschrumpft

  //////////////////////////////////
  nachf[c0][c1]++; // Matricelement erhöhen
  //////////////////////////////////

  // printf(" \"%c\"=(%d) %d\n", output[c1], c1, nachf[c0][c1] );

  c0 = c1; // Kontext: vormals aktuelles Zeichen

  c = fgetc( ein ); // nächstes Zeichen lesen
}

fclose(ein);

// Matrix der absoluten Häufigkeiten ist nun fertig ausgefüllt

// -----
// nun: relative Häufigkeiten bestimmen und sortiert ausgeben
// -----

int
  rank[UMFANG]; // sortierte Permutation

float
  p[UMFANG], // absolute Wahrscheinlichkeit für Zeichen

  p_vor[UMFANG][UMFANG], // bedingte Wahrscheinlichkeit für Vorgänger

  p_nach[UMFANG][UMFANG]; // bedingte Wahrscheinlichkeit für Nachfolger

```

```

// Ermittlung der Normierungsgröße (ALLE analysierten Zeichenpaare)
n=0;
for (i=0; i<UMFANG; i++)
  for (j=0; j<UMFANG; j++)
    n+=nachf[i][j];

for (i=0; i<UMFANG; i++) // Bezugs-Zeichen
{

  m=0; // Gesamtanzahl der Vorkommen des Zeichens
  for (j=0; j<UMFANG; j++) m+=nachf[i][j];
  // ist übrigens symmetrisch, d.h. gleich der Summe der "nachf[j][i]"

  p[i] = ((float)m)/(float)n; // absolute Wahrscheinlichkeit des Zeichens

  fprintf( aus, "\n\n\n Buchstabe \"%c\" (%2.1f%%): \n\n", output[i] , 100*p[i] );

  if (m!=0) // relevantes Zeichen (Vorkommen > 0) ?
  {

    // Ermittlung und Ausgabe der VOR-Kontexte:

    fprintf( aus, " - VOR-Kontexte:  ");

    // Ermittlung der bedingten Wahrscheinlichkeiten
    for (j=0; j<UMFANG; j++)

      p_vor[i][j] = nachf[j][i]*1.0/m;

    for (j=0; j<UMFANG; j++) rank[j]=j;
    // "id"-Abbildung (die triviale Permutation)

    // Bubblesort: (absteigende Ordnung herstellen)
    for (j=0; j<UMFANG; j++)
      for (k=1; k<UMFANG; k++)
        if (nachf[ rank[k-1] ][i]<nachf[ rank[k] ][i])
          {
            l=rank[k];
            rank[k]=rank[k-1]; // Vertauschung
            rank[k-1]=l;
          }

    for (j=0;

      (j<UMFANG)
      && (nachf[ rank[j] ][i]!=0) // kann auskommentiert werden
      ;
      j++)

    fprintf( aus,

      "%c (%2.1f%%) ", // Alternative 1

```

```

//                                     "%c (%2.1f\%=%d) ", // Alternative 2

output[ rank[j] ],

100*p_vor[i][ rank[j] ]
//                                     ,nachf[ rank[j] ][i] // kann (für Alternative 1) auskommentiert werden
);

fprintf( aus, "\n\n");

// Ermittlung und Ausgabe der NACH-Kontexte:

fprintf( aus, " - NACH-Kontexte: ");

// Ermittlung der bedingten Wahrscheinlichkeiten
for (j=0; j<UMFANG; j++)

    p_nach[i][j] = nachf[i][j]*1.0/m;

for (j=0; j<UMFANG; j++) rank[j]=j;
// "id"-Abbildung (die triviale Permutation)

// Bubblesort: (absteigende Ordnung herstellen)
for (j=0; j<UMFANG; j++)
    for (k=1; k<UMFANG; k++)
        if (nachf[i][ rank[k-1] ]<nachf[i][ rank[k] ])
            {
                l=rank[k];
                rank[k]=rank[k-1]; // Vertauschung
                rank[k-1]=l;
            }

for (j=0;

    (j<UMFANG)
    && (nachf[i][ rank[j] ]!=0) // kann auskommentiert werden
    ;
    j++)

    fprintf( aus,

        "%c (%2.1f\%) ", // Alternative 1
        //                                     "%c (%2.1f\%=%d) ", // Alternative 2

        output[ rank[j] ],

        100*p_nach[i][ rank[j] ]
        //                                     ,nachf[i][ rank[j] ] // kann (für Alternative 1) auskommentiert werden
        );

    fprintf( aus, "\n\n");

} // relevantes Zeichen (Vorkommen > 0)

} // nächstes Zeichen

```

```

//
// informationstheoretische Analyse:
//
// -----
// Berechnung der Shannon-Information
// -----

float
  h,h0; // lokale Variablen für Informationsmaß

// Ausgabe des (trivialen) Informationsgehalts bzgl. Blockcode-Länge
fprintf( aus, "\n\n Informationsgehalt pro Zeichen (Block-Kode):"
        " %1.2f bit\n\n", log(UMFANG)/log(2));

// Ermittlung des Informationsgehalts eines Zeichens
// gemäß den absoluten Auftretswahrscheinlichkeiten:

h=0; for (i=0; i<UMFANG; i++) h+=p[i]; printf("\n Muß 1 ergeben: %f\n",h); // Kontrollausgabe!

h=0;
for (i=0; i<UMFANG; i++) if (p[i]!=0.0) // (Test muß leider sein!)

    h -= p[i] * log(p[i]) / log(2);

fprintf( aus, "\n\n Informationsgehalt pro Zeichen (Shannon):"
        " %1.2f bit\n\n", h );

// Ermittlung des bedingten Informationsgehalts eines
// Zeichens gemäß den bedingten Auftretswahrscheinlichkeiten:

h=0;
for (i=0; i<UMFANG; i++)

    if (p[i]!=0) // relevantes Zeichen? (Test wegen "0/0")
    {
        h0=0; for (j=0; j<UMFANG; j++) h0+=p_nach[i][j]; printf("\n Muß 1 ergeben: %f\n",h0); // Kontrollausgabe!

        h0=0;
        for (j=0; j<UMFANG; j++) if (p_nach[i][j]!=0.0) // (Test muß leider sein!)

            h0 -= p_nach[i][j] * log(p_nach[i][j]) / log(2);

        // bedingter Informationsgehalt des Nachfolgers von Zeichen Nr.i
        // wird gewichtet mit Auftretswahrscheinlichkeit von Zeichen Nr.i:

        h += p[i] * h0;
    }

fprintf( aus, "\n\n Bedingter Informationsgehalt pro Zeichen (Shannon):"
        " %1.2f bit\n\n", h );

fclose(aus);
}

```