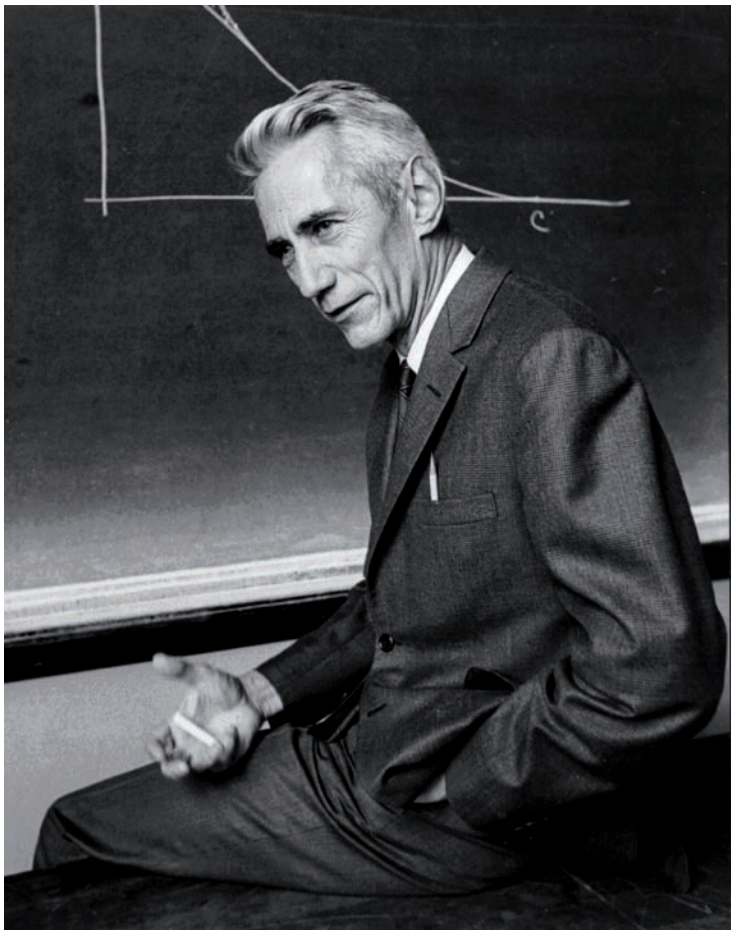
INFORMATIONSTHEORIE

Texte aus der Sicht der Informationstheorie

WIE VIEL INFORMATION ENTHALTEN TEXTE UND WIE KANN MAN AUTOREN ERMITTELN?



ARCHIVE.COMPUTERHISTORY.ORG

Abb. 1: Claude Elwood Shannon als Hochschullehrer vor einem Tafelbild seines Kanalcodiertheorems.

VON JOACHIM HAGENAUER

Wer heute Texte mit einem Textverarbeitungsprogramm bearbeitet, archiviert oder über das Internet überträgt, benutzt möglicherweise eines der beim Textverarbeitungssystem vorhandenen Kompressionsprogramme, um Speicherplatz bzw. Übertragungszeit zu sparen. Solche Programme sind etwa unter dem Namen „zip“ bekannt. Man kann nun danach fragen, wie groß der minimale Speicherplatzbedarf nach solch einer Komprimierung ist. Dies ist eine Frage von immenser Bedeutung, wenn man daran denkt, dass etwa die Firma Google daran arbeitet, alle Texte unserer Geistes- und Wissenschaftsgeschichte zu speichern, zu verteilen und so rasch zugänglich zu machen.

Der mittlere Informationsgehalt von Nachrichtenquellen

Diese Frage wurde lange vor der Einführung der modernen Speicher- und Rechnertechnologie bereits 1948 durch den Mathematiker und Ingenieur Claude Elwood Shannon (1919–2001) grundsätzlich beantwortet. Shannon ist der Begründer einer ganzen Wissenschaftsdisziplin – der „Informationstheorie“ –, und sein Kompressionstheorem für Texte, das so genannte Quellencodiertheorem, ist nur ein kleiner Teil seines immensen Werkes. Die Aktualität seiner Arbeiten spiegelt sich darin wider, dass sein grundlegender Aufsatz von 1948 bei *Scholar Google* mit über 14.000 meist neueren Zitaten aufgelistet wird.

Shannon hat den Begriff des mittleren Informationsgehaltes von Nachrichtenquellen geprägt, von denen Texte hier lediglich als Beispiel dienen. Dabei ist Information nach Shannon nur im statistischen Sinn aufzufassen; die Semantik oder gar die Ästhetik eines Textes interessiert ihn zunächst nicht. Auch dem „dummen“ Computer gilt ja das Geschreibsel eines Erstklässers gleich viel wie ein Goethegedicht. Texte sind demnach Buchstabenfolgen, die

durch ihre Statistik, also durch die Wahrscheinlichkeit oder die relativen Häufigkeiten von Buchstaben, beschrieben werden.

Die Informationseinheit Bit

Was ist also dieser statistische Informationsgehalt? Shannon hat ihn über den Logarithmus des Kehrwertes der Auftrittswahrscheinlichkeit definiert. Seltene Ereignisse haben also einen hohen Informationsgehalt. Nimmt man den Logarithmus zu Basis 2, so ist die dadurch definierte Informationseinheit das berühmte Bit, und den Umfang unserer Texte messen wir dann beispielsweise in Megabit. Hätte man einen Text einer Sprache, bei der 32 statistisch unabhängige Buchstaben bzw. Zeichen mit gleicher Häufigkeit vorliegen, so wäre der Informationsgehalt nach Shannon – er nennt es Entropie – 5 bit pro Buchstabe. Das wäre zugegebenermaßen ein ziemlich langweiliger Text. Texte einer Sprache sind strukturiert und Buchstaben wie Buchstabengruppen (Wörter) kommen mit unterschiedlicher Häufigkeit vor. Bei deutschen Texten sind die häufigsten Buchstaben (über 5 %):

Tabelle 1: Die häufigsten Buchstaben in deutschen Texten

E	17,61 %
N	10,44 %
D	7,49 %
R	7,40 %
S	6,59 %
I	6,48 %
T	6,12 %

Im Vergleich dazu sind die häufigsten Wörter in deutschen Texten: der, die, und, in, den, von, zu, das, mit, sich …

Kennt man die Statistik von Buchstabengruppen, so kann man sogar einen statistischen – natürlich sinnlosen – Text synthetisch erzeugen. Am Lehrstuhl für Nachrichtentechnik haben wir (Günter Söder) so z. B. die Häufigkeit ermittelt, mit der ein bestimmter Buchstabe nach einer Zweierkombination von Buchstaben in der Lutherbibel und der King-James-Bibel vorkommt. Erzeugt man nun nach dieser Wahrscheinlichkeitstabelle einen pseudozufälligen Text, der mit der Zweierkombination „go“ beginnt, so erhält man beispielsweise die folgenden Texte (Abb. 2), die einem in manchen Abschnitten fast sinnvoll vorkommen können, so z. B. in der ersten Zeile „und so wie weinet“ und am Textende „thy drend of life me thee“.

gottert seinem kanlichtel lavie und so wie weinet euch eine ers haben hand sen wein fallte sen derde der um frichre kne mon ine nachine unger ern aussein nicht war wurcht ang und sie issen len fruchtenem die nund elen derrlen re alt jen allt und bischwen der und sitet halem haftim verten der ern undteifen war da under er tesch ken ophin gehe dermach ihm treurchei nachen das kampfe berst maeles er geze der das ste flichzerder amach und digkenke doche gebereuch dem ge hollen hagenseibt brachtst und mitten wal und fle allein gen du zu hater ren abeidas brie ruf und dertnaule derk nich wohnegtenhath abereig den

god ing the chict abalt to good agel serah unt him hes my godstroys sh them unt and hind th idsel ell cat th set andry froughter and ham bekah brew nes my rah ble and unt came yeathe dau ith of haved ne own grothou wee upons of the and washe me jok pland wastrich cringed golded lottly faters god thany se flore and nigh luchenahat unto and and came my younto dings abrom seve ablee had any face saing come mothery flot by th thou gaid and gater day for unto uncep his dam saing igns nam ale of this and shat levento whis weld unto und sh whe fie youltake to soner an saids sh nother thy drend of life me thee anto to ders unto

Abb. 2: Beispiele synthetisch erzeugter Texte.

Entropie von Texten

Texte enthalten Redundanz und ihr Informationsgehalt (die Entropie) ist weit weniger als die oben bestimmten 5 bit pro Buchstabe. Was ist nun aber diese wahre Entropie von Texten und wie bestimmt man sie? Die Frage ist für die eingangs erwähnten Textkompressionsverfahren wichtig, denn das berühmte Kompressions-Theorem von Shannon besagt, dass man Texte bis auf ihre Entropie komprimieren und dann fehlerfrei wiedergewinnen kann. Allerdings ist es selbst für heutige Höchstleistungsrechner ein zu schwieriges Unterfangen, alle statistischen Bindungen in einem Kanon von Texten zu ermitteln, um dann daraus die Entropie zu berechnen. Shannon hat bereits 1951 in einem genial einfachen Verfahren, dargestellt in seinem Aufsatz *The entropy of printed English*, eine gute Näherung gefunden. Er ließ Studenten den nächsten Buchstaben in einem abgedeckten Text schätzen und zählte die Zahl der Rateversuche, bis der (verdeckte) richtige Buchstabe gefunden wurde. Aus der Statistik dieser Zahlen konnte er einen Schätzwert von etwa 1.5 bit pro Buchstabe ermitteln, den der große deutsche Nachrichtentechniker Karl Küpfmüller 1954 für die deutsche Sprache in etwa bestätigte. Demnach müsste es also einen Kompressionsalgorithmus geben, der Texte um einen Faktor 5/1.5, also etwa um das Vierfache, komprimieren kann.

Der LZ-Algorithmus

Shannons Theorie ist nicht konstruktiv, so dass man aufgrund seiner Ergebnisse allein keinen solchen Textprozessor bauen kann. Seit 50 Jahren bemühen sich Ingenieure und Informatiker, das Shannonsche Versprechen einzulösen. In Textverarbeitungsprogrammen taucht manchmal das Kürzel LZ auf. Dies weist auf einen Kompressionsalgorithmus der israelischen Wissenschaftler Abraham Lempel und Jacob Ziv hin, der sogar komprimieren kann, ohne die Statistik zu kennen. Die Statistik der Buchstabenfolgen lernt der LZ-Algorithmus so zusagen intern, während er die Texte bearbeitet. Nun weiß man, dass die Shannonsche Theorie nur asymptotisch, also für sehr lange Texte gilt und somit der „Kompressor" nur dann gute Kompressionsergebnisse liefert, wenn man statt einer Seite das ganze Buch oder noch besser eine ganze Bibliothek komprimiert.

Abb. 3 : Mathematische Definition von Textdistanzen mit Hilfe der Informationstheorie.

Definiere eine Distanz (Metrik) zwischen zwei DNA Folgen S_i and S_j mit Hilfe ihrer Entropien $H(S_i), H(S_j)$ und der Transinformation $I(S_i; S_j)$

$$d(S_i, S_j) = 1 - \frac{I(S_i; S_j)}{\max(H(S_i), H(S_j))} \leq 1$$

und messe sie mit Hilfe von Kompressions-Algorithmen

$$d(S_i, S_j) \approx \frac{|\mathrm{comp}(s_j, s_i)| - |\mathrm{comp}(s_j)|}{|\mathrm{comp}(s_i)|}.$$

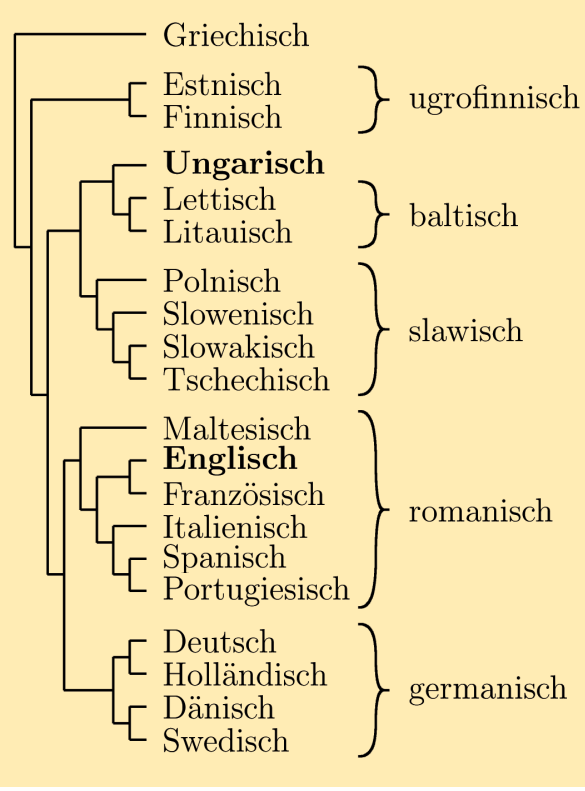


Abb. 4: Sprachbaum abgeleitet aus EU-Texten und einem Distanzmaß der Informationstheorie.

Kompression langer Texte

Man kann das Experiment leicht am häuslichen Computer machen, indem man die Größe des originalen Textes, bzw. von Teilen desselben, jeweils mit denen der „zip-Files" vergleicht. Statt des Kompressionsfaktors 4 erreicht man wohl nur Werte von 2 bis 2.5. Rechenaufwändigere Verfahren, wie das „Context Tree Weighting"-Verfahren erreichen eine bessere Kompression; der Algorithmus untersucht hierbei den Kontext, indem er einen Kontextbaum aufbaut, aber immer nur auf „dummer" statistischer Basis. Die Feinheiten der Semantik und Grammatik bleiben außen vor, ganz zu schweigen von der Weisheit oder der ästhetischen Schönheit, die in einem Text stecken mag.

Die elementare Struktur der Kommunikation

Umberto Eco, der italienische Semiotiker und Autor, hat übrigens seinem Lehrbuch *Einführung in die Semiotik* ein einleitendes, gut lesbares Kapitel über die Shannonsche Informationstheorie beigefügt, mit der Begründung: *„Wenn jedes Kulturphänomen ein Kommunikationsphänomen ist, dann muss man die elementare Struktur der Kommunikation dort aufsuchen, wo Kommunikation sozusagen minimal stattfindet, d. h. auf der Ebene der Übertragung von Information zwischen zwei Apparaten."* (Umberto Eco, Einführung in die Semiotik, 7. Ausg., München, Fink 1991, Uni-Taschenbücher, S. 47.)

Textverwandtschaften

Die Shannonsche Theorie ist erweiterbar, um den wechselseitigen Informationsgehalt zweier Texte – nennen wir sie S_i und S_j – aus der Verbundstatistik zu bestimmen. Man bezeichnet die dafür relevante, von Shannon definierte Größe als Transinformation $I(S_i;S_j)$. Sie spielt eine große Rolle in der Übertragungstechnik und gibt sozusagen die Antwort auf die Frage „Wie viel kommt rüber?" Sie misst wie viel Information zwei Texte gemeinsam haben. Man kann sie verwenden, um den statistischen Verwandtschaftsgrad von Texten zu messen. Wir haben am Lehrstuhl für Nachrichtentechnik der TU München ein Distanzmaß für den genetischen Code und für Texte entwickelt (Pavol Hanus 2004), um so die Verwandtschaft von Datenfolgen zu messen. Mathematisch sieht dieses Maß mit Werten zwischen Null und Eins folgendermaßen aus (siehe Abbildung 3).

Stammbaum der Sprachen

Hat man dieses Distanzmaß zwischen allen untersuchten Texten ermittelt, so kann man einen Verwandtschaftsbaum aufstellen. Das ist natürlich nicht schwierig zwischen Texten verschiedener Sprachen, wie es das Beispiel in

Abbildung 4 zeigt. Als Texte wurden umfangreiche juristische Texte (Verordnungen) der europäischen Union verwendet, die in verschiedene Sprachen übersetzt waren. Die hier sichtbare Verwandtschaft des Englischen mit dem Französischen erklärt sich wohl aus den verwendeten juristischen Fachtexten, die im Englischen offenbar mehr auf lateinisch-romanische Sprachwurzeln zurückgreifen.
Natürlich kann diese simple computer-basierte und automatisierte Methode nicht detaillierte linguistische Studien ersetzen, sie zeigt jedoch, dass Shannons Transinformation ein brauchbares Maß ist.

Ermittlung der Autorenschaft mit Hilfe der Informationstheorie

Schwieriger ist das folgende Zuordnungsproblem: Die *Federalist Papers* sind eine Serie von 85 Artikeln, welche 1787/88 in verschiedenen New Yorker Zeitungen erschienen. Sie dienten vorrangig als Verteidigungsschrift der 1787 in Philadelphia entworfenen, aber noch nicht in Kraft getretenen Verfassung für die amerikanische Union. Die drei unter dem gemeinsamen Pseudonym Publius schreibenden Autoren Alexander Hamilton, James Madison und John Jay versuchten mit ihren Essays Einfluss auf die Ratifikationsdebatte zu nehmen. Die Autorenschaft von 12 Aufsätzen ist jedoch nicht endgültig geklärt. Das ist besonders schwierig, weil die drei Autoren in der gleichen Sprache und bei gleichem Bildungshintergrund über sehr verwandte Themen geschrieben haben. Wir haben mit einem die Shannonsche Transinformation verwendenden Distanzmaß eine Zuordnung von strittigen Aufsätzen versucht (Abbildung 5).

Mit größter Wahrscheinlichkeit sind sie demnach Madison zuzuordnen, wobei der prozentuale Abstand zu Hamilton gering ist, besonders beim Aufsatz 57. Unsere automatisierte Zuordnung stimmt mit der vorherrschenden Meinung von vielen Sprachwissenschaftlern, Historikern und Mathematikern überein, wobei Letztere ganz andere Klassifikationsverfahren als die Informationstheorie verwendet haben.

Beispiel für interdisziplinäre Forschung

Die oben beschriebenen Verfahren verwenden wir in der Forschungsgruppe ComInGen des Lehrstuhles für Nachrichtentechnik (LNT-TUM) in enger Zusammenarbeit mit Genbiologen auch, um phylogenetische Abstammungsbäume mit Hilfe der mitochondrialen DNA zu erstellen. Die hier beschriebene Untersuchung von Texten mit Hilfe mathematischer und ingenieurtechnischer Verfahren mag als Beispiel interdisziplinärer Forschung dienen, wie sie einer Bayerischen Akademie der Wissenschaften aufgetragen ist.

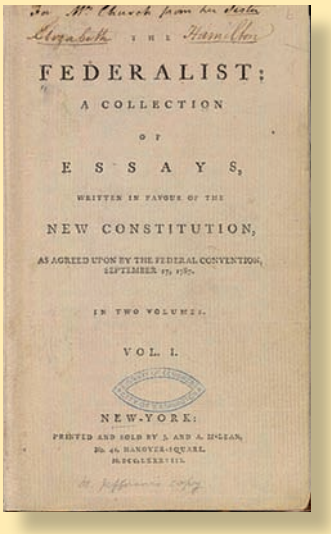
Der Autor ist em. o. Professor für Nachrichtentechnik der TU München, o. Mitglied der Mathematisch-naturwissenschaftlichen Klasse und Vorsitzender des Forums Technologie der Bayerischen Akademie der Wissenschaften.

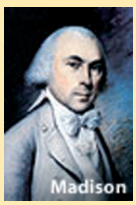
Abb. 5: Klassifizierung der „Federalist Papers" mit Hilfe der Informationstheorie.

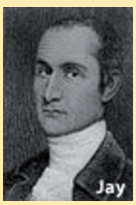
Autorenschaft-Erkennung

Federalist Papers

- **1787–1788 (New York)**
- **85 Aufsätze (Pseudonym Publius)**
- **geschrieben von:**
  - **James Madison**
  - **Alexander Hamilton**
  - **John Jay**
- **bei 12 Aufsätzen ist die Autorenschaft strittig**







49	√	7,9%	17,7%
50	√	4,8%	16,7%
51	√	7,3%	15,8%
52	√	6,4%	17,2%
53	√	5,8%	13,7%
54	√	3,6%	13,6%
55	√	4,8%	14,9%
56	√	5,4%	14,2%
57	√	1,0%	13,7%
58	√	4,2%	15,7%
62	√	5,3%	12,3%
63	√	4,7%	13,0%