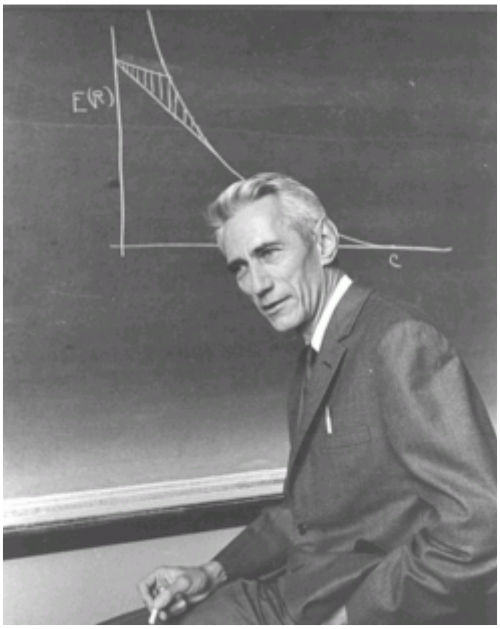
Claude Shannon: scientist-engineer



Claude Shannon
Copyright: Lucent Technologies Inc.

Ronald Rousseau

KHBO, Department industrial sciences and technology
Zeedijk 101, B-8400 Oostende, Belgium
e-mail: ronald.rousseau@kh.khbo.be
&
Honorary Professor Henan Normal University
Xinxiang, Henan, PR China

Abstract

In this article we present an overview of the life and work of Claude Shannon, who died begin 2001. Shannon is a pioneer in many fields of science. He introduced Boolean algebra in switching theory, is the founding father of communication science, and performed several experiments related to artificial intelligence. In this paper we stress the relation between his work and the information sciences, describing, among other things, his use of n-grams. Finally, we discuss, in general, the use of n-grams in the field of information science.

Introduction

On Saturday, February 24, 2001, Claude Elwood Shannon died in Medford (MA, USA) at the age of 84. Claude Shannon is probably the most famous scientist in the field of communication theory. Here, the term communication theory must be understood in the engineering sense of the word and not (directly) in the sense of the word given in the humanities. Yet, Claude Shannon's work has exerted a considerable influence on the information sciences. Shannon worked at Bell Labs from 1941 until 1956, but stayed associated with this research laboratory until 1972. In 1956 he became a professor at MIT where he stayed until his retirement in 1978. He received the IEEE Medal of Honor in 1966, became a member of the American Academy of Sciences and of the London Royal Society. In 1966 he received the American National Medal of Science (the highest scientific reward in the USA). He, moreover, received the John Fritz Medal (in 1983) and the Kyoto Prize (in 1985).

Studies and work at Bell's Research Lab

Shannon was born in Petoskey (Michigan, USA) on April 30, 1916. He studied at the Michigan University where he obtained a bachelor's degree in science (mathematics and electrical engineering) in 1936. He continued his studies and obtained a Ph.D. at the prestigious MIT with an application of mathematics to genetics.

Shannon worked most of his life at Bell Labs. It was there that he wrote in 1948 his most famous article, namely "A mathematical theory of communication" (an article in two parts) (Shannon, 1948a,b). Starting from the fundamental insight that telecommunication consists of reproducing (exactly or approximately) at one place a message sent at another, he understood that any message could be reduced to a string of zeros and ones. Because of this understanding he became a pioneer of the digital era as we know it nowadays (Golomb, 2001). This approach to telecommunication only uses two basic symbols, namely a zero and a one. For this reason Shannon called them 'binary digits' or, in short: **bits**, a term first proposed by J.W. Tukey. Shannon hence realized that any kind of data, whether it be words, sounds or images, could be represented by bits. The first results of this digital thinking were in the domain of electric systems and telephony. Note that the term 'digital thinking' essentially means that Shannon applied the ideas of Boole, developed one century earlier (Shannon, 1938). Once again ideas developed by mathematicians proved useful outside the context in which they originated. Nobel price winner Eugene Wigner described this phenomenon as 'the unreasonable effectiveness of mathematics' (Wigner, 1960). The word 'unreasonable' is used here in the sense of "against all rational expectations".

The use of Boolean logic in switching theory was nothing but a first step. The next one was the idea that the more uncertain the communication, the more information it contains. At first this idea seems strange, even paradoxical. That is why we will explain it somewhat more. Assume that the symbol that will be transmitted is an A, a B or a C. The probability that it will be an A is 98%, while the probabilities for a B and C are each 1%. Knowing now that the transmitted symbol was actually an A conveys little information (we could easily have guessed it). If, on the other hand, the probabilities are equal (each 33.33%), then the announcement that indeed A has been transmitted, conveys a lot of information. This is the main idea of Shannon's communication theory. It is clear that, within this theory, it is essential to be able to measure uncertainty. The measure used to do this is known as the entropy measure (or the entropy, in short). It was the great John von Neumann who suggested this term to Shannon because of its resemblance to the term entropy as used in thermodynamics. In that field entropy is a measure for the disorder of a system. Shannon's ideas could only be applied in practice a few years later when transistors and integrated circuits (ICs) were produced on an industrial scale.

Universal genius and pioneer in the field of artificial intelligence

Shannon is also a computer pioneer. He was a contemporary of John von Neumann and Alan Turing, and understood, before most other scientists, that computers were able to do much more than just complicated calculations. During World War II Shannon performed research at Bell Labs on cryptography. His article "Communication theory of secrecy systems' transformed cryptography from an art form to a science. Note that this article was "classified" first (this means 'secret'), but some years later it was declassified and consequently published (Shannon, 1949).

In a sense, Shannon was a typical genius: he could rightly be called an eccentric personality. Among his eccentricities we mention the fact that he made a mechanical mouse (called Theseus) designed to find its way through a maze. In this way he became one of the pioneers in the field of artificial intelligence (AI). He further constructed a computer specially designed to do calculations using Roman numerals (called THROBAC), a frisbee with rocket propulsion, a machine that could guess if a person would chose head or tail (the secret was that people very often use fixed patterns, and the machine was programmed to detect these patterns), a juggling robot and one of the first chess computers. It is no surprise to see that Shannon was invited among the select little 'club' of scientists that, during the summer of 1956, came together at Dartmouth College under the direction of John McCarthy. This meeting is sometimes referred to as the 'unofficial' beginning of the field of artificial intelligence (Russell & Norvig, 1995). Besides McCarthy and Shannon also Marvin Minski, Nathaniel Rochester,

Trenchard More, Ray Solomonoff, Arthur Samuel, Oliver Selfridge, Allen Newell and Herbert Simon were present.

Shannon himself was a good juggler. It is known that he used to juggle while riding a unicycle through the halls of Bell Labs. Of course, Shannon would not have been the man he was if he had not studied juggling from a more scientific side. And indeed: there does exist a Shannon juggling theorem giving the relation between the time a ball is in a hand (or not) and the time it is in the air. The theorem states the following:

$$(F+D)H = (V+D)N$$

where F denoted the time a ball is in the air, D denoted the time a ball is in a hand, V is the time a hand is empty, N is the number of balls and H is the number of balls. Note that, indeed, a human can juggle with one as well as with two hands. Of course, also the number of balls is a variable. The restrictions inherent in this formula imply that it is practically impossible to juggle with nine balls or more. A proof of this theorem can be found by describing a full cycle from two points of view: once as experienced by one hand, and once as experienced by a ball (Horgan, 1990; Beek and Lewbel, 1995).

Back to communication and the notion of entropy

In this section we will explain in somewhat more detail the theory that made Shannon famous. According to Weaver (1949) communication can be studied on three levels. The first one is the technical level. Here one studies how accurately symbols can be transmitted. The second level is the semantic level where one studies how transmitted symbols catch the intended meaning, and finally the communicative level studies the effectiveness of the transmission. This means that one studies if the message results in the intended effect. Shannon only studies the technical aspect. Yet, Weaver's main claim is that Shannon's theory also has a profound influence on the other two levels (Weaver, 1949).

'Information' (in the purely technical sense) is the central notion in Shannon's theory. If p(G) denotes the probability that an event G will occur, and if a person is told that the event G has actually occurred, then the amount of information given to that person is defined by Shannon as:

$$I(G) = - \log_2(p(G)) \text{ units of information}$$

The *bit* is the unit of information. As any logarithm (here the logarithm to the base 2) of a number between 0 and 1 is always negative, the minus sign in front of this expression makes sure that the amount of information received is a positive number. It is now easy to see that, if you have no prior information, and if there are only two possible symbols that can be transmitted (namely 0 and 1) then

knowing that a 1 has been transmitted yields an amount of information equal to $-\log_2(1/2) = -(-1) = 1$ (as the probability that a 1 will be transmitted is equal to 0.5).

This formula, defining the notion of 'information', has good mathematical characteristics. The most important one is the fact that now the amount of information related to two independent events G and H is the sum of the separate amounts of information. Indeed: if two events are independent then p(G and H) is equal to p(G).p(H) (a multiplication), and hence:

$$\begin{aligned} I(G \text{ and } H) &= -\log_2(p(G \text{ and } H)) \\ &= -\log_2(p(G) \,.\, p(H)) \\ &= -\log_2(p(G)) - \log_2(p(H)) \\ &\quad \text{[the logarithm of a product is equal to the sum of the logarithms]} \\ &= I(G) + I(H). \end{aligned}$$

From this formula we can derive that one receives little information when an almost certain event occurs. Indeed, if p(A) is equal to 0.98, then I(A) = 0.029, while, if p(B) = 0.01, then I(B) = 6.644. Information defined in this way has nothing to do with the meaning of the message (second level) and certainly not with its effectiveness (third level).

Next, we come to the definition of the notion 'entropy'. Suppose that one has a source, sending messages consisting of one of the following symbols: s_1, s_2, ..., s_n. The set consisting of all these symbols is called an alphabet. One further assumes that occurrences of these symbols are independent events, and that their probabilities are given by $p(s_1)$, $p(s_2)$, ..., $p(s_n)$. The mean information of such a message is:

$$\sum_{j=1}^{n} p(s_j)\, I(s_j) = -\sum_{j=1}^{n} p(s_j) \log_2(p(s_j))$$

This quantity is the entropy of the source, usually denoted as H. If all $p(s_j)$ are equal then the entropy is at a maximum (for fixed n). This corresponds to the case where symbols are chosen at random. On the other hand, if one of the $p(s_j)$ is almost 1, and, hence, all other ones almost zero, then H is small. There is almost no freedom of choice, and consequently the average information is small.

The ratio of the real entropy of a communication system to the maximum value is called the relative entropy. Its complement with respect to one is called the redundancy. The redundancy corresponds to that part of the message that is not obtained though a choice of the sender, but follows from the statistical rules governing the use of a particular alphabet. Weaver claims that the technical notion of 'redundancy' corresponds well with its meaning in daily life.

Building on these mathematical foundations, Shannon showed that each communication channel has a maximum capacity for sending reliable messages. He showed, more precisely, that it is possible by coding messages in a very clever way, to approximate this maximum as close as one wants (without, however, ever reaching it). This maximum is known as the Shannon limit. “Clever coding” means here: adding just enough redundancy so that the message cannot be corrupted by noise. After years of research in this domain scientists are now able to approximate the Shannon limit up to 0.115% (Golomb, 2001).

Entropy and its relation to the information sciences and the measurement of biodiversity

The information sciences study all phenomena related to information, such as its transmission, transformation, compression, storage and retrieval. Note that the information sciences focus on information, not data. Data manipulation is the field of computer science, information technology and software engineering. It is only when meaning and humans enter into the picture that we are within the limits of the information sciences (Holmes, 2001). Formulated in this way, Shannon’s theory has nothing to do with the information sciences. Yet, information scientists can’t but be interested in data, and hence, indirectly, in Shannon’s theory (Rousseau, 1986). Characteristic for this is the fact that the *American Society for Information Science* recently changed its name into *American Society for Information Science and Technology*. This is a clear sign that domains overlap more and more. We further observe that Weaver and many scientists after him (Zunde, 1981,1984) state that Shannon’s approach can be used to study the more semantic aspects of communication. According to Zunde the laws of Zipf and Mandelbrot (Egghe & Rousseau, 1990) as applied to linguistics play the role of a bridge between these two levels.

Where – within the information sciences – is Shannon’s work used? One of the attempts for better retrieval methods was based on Shannon’s entropy formula (Cooper, 1983). As so many other attempts also this one had no influence on the daily practice of Boolean searches. It is only recently that, thanks to the advent of the Internet and search engines such as AltaVista and Google, the practical monopoly position of Boolean searches has been broken.

Within the field of information science Loet Leydesdorff of Amsterdam University is probably the most profound thinker about information and entropy. The notions of ‘information’ and ‘entropy’ taken from Shannon play a key role in his book “*The Challenge of Scientometrics*” (Leydesdorff, 1995). According to Leydesdorff the use of these notions in science studies lead to the following specific advantages:

- The entropy measure is parameter-free, requiring less mathematical idealizations;

- Shannon's theory of information is directly related to probability; hence it can be brought together in one theory with other ideas from probability and statistics, as used in the social sciences;
- As formulae used in information theory are often just additions, results can more easily be decomposed.

Finding an adequate measure to determine biodiversity is a fundamental problem in the field of ecology. It is, indeed, utterly impossible to determine if biodiversity increases or decreases (for instance, because of decisions made by the government), without a proper measure. Note that we use here the term 'biodiversity' in the sense as used by ecologists. It refers not only to the number of species present, but also to the relative apportionment of individuals among those species present (Magurran, 1991; Rousseau and Van Hecke, 1999). We just mention here that Shannon's entropy measure, also known – in ecology - as the Shannon-Wiener index (Peet, 1974), or closely related formulae, satisfy most requirements for a good diversity measure (Nijssen et al., 1998).

N-grams

Finally, we will delve somewhat deeper into the use and study of n-grams, yet another subject where Shannon has made important contributions (Shannon, 1951). The study of n-grams may be situated on the border of many scientific domains: the information sciences, linguistics, artificial intelligence and many subfields of the engineering sciences. We will first explain what n-grams are and where they are used.

There exist two important approaches to the description and manipulation of texts: one based on symbols, such as (western) letters, or Chinese characters, and one based on whole words. The use of n-grams is, however, a third way, in between the two other ones.

We will next cover the following aspects:

- What are n-grams and how can they be constructed starting from a given string of symbols;
- the use of n-grams for comparing words; here we also discuss finding and correcting spelling errors, measures for the similarity between word strings, and the use of these techniques for computer searches in databases and on the Internet;
- the use of n-grams in different languages;
- other applications such as the classification of languages and the study and classification of biological chains (DNA).

An n-gram is just a string consisting of n symbols, usually taken from a text. Often (but not always) this n-gram is made of symbols originating from the same

word. In practice one usually restricts oneself to the study of bi-grams (n=2) or tri-grams (n=3).

Consider, for instance, the n-grams that can be formed from the word UNIVERSITY. It is customary the start from the word under study preceded and followed by (n-1) times a special 'empty' symbol such as '*'. N-grams are then constructed by sliding a window of length n over this string, progressing one symbol at the time. This procedure is called 'redundant coding'. The bi-grams formed in this way from the word UNIVERSITY are:

U UN NI IV VE ER RS SI IT TY Y

If a word, such as UNIVERSITY, has 10 letters, one obtains 11 bi-grams. In general a word consisting of m letters leads to m+1 bi-grams. Similarly one can form 12 tri-grams, namely

****U *UN UNI NIV IVE VER ERS RSI SIT ITY TY* Y****

A general word string consisting of m letters leads to m+1 bi-grams, m+2 tri-grams and m+n-1 n-grams. The theoretical number of possible n-grams is very high. For an alphabet of 26 letters there are $26^2 = 676$ bi-grams and $26^3 = 17576$ tri-grams possible. However, in English only 64% of these bi-grams and 16% of all tri-grams actually exist. It is this fact that allows for the detection of spelling errors (see further). Including punctuation marks makes things more difficult, but we will not consider this aspect here.

'Non-redundant' coding uses word fragments with no overlaps. Then the word UNIVERSITY yields:

U NI VE RS IT Y and ***UN IVE RSI TY***

In his article, published in 1951, Shannon used frequency tables for bi- and tri-grams in the English language. In this way he calculated the entropy and redundancy of the English language. According to his calculations the English language has a redundancy of about 75%. Examples of this redundancy are the fact that, in English, one finds very often a 't' before an 'h' (such as in 'the', 'thanks' and 'Smith') and almost always a 'u' after a 'q' (such as in 'question' and 'equal'). It is interesting to remark that Shannon also refers to Zipf's work (1949) about word frequencies.

Any text contains many variant word forms, such as: *work, works, working, werks* (a typing error) and so on. A conflation algorithm is a program that brings all these variants together into one word class. Clearly, words belonging to the same class have a very large bi-gram similarity. Similarity between two text

strings can be measured in different ways. Using bi-grams we see that the similarity between

WORKS and WORKING

this is between the bi-grams:

W WO OR RK KS S and ***W WO OR RK KI IN NG G***

according to the Dice coefficient (calculated as twice the number of bi-grams they have in common divided by the sum of the number of bi-grams is each word) is:

$$\frac{2*4}{6+8} \approx 0.57$$

while using the Jaccard index (another well-known similarity measure, equal to the number of bi-grams in common (a mathematical intersection) divided by the total number of bi-grams occurring in at least one of the two words (a mathematical union)), this becomes:

$$\frac{4}{10} = 0.4$$

Exact values of similarity measures are usually not important. Their importance lies in the ranking they create. Term matching algorithms use these rankings. When searching for a word in a text or database the algorithm provides the user with a ranked list of words that have the largest similarity to the word used in the query.

While performing searches in databases or on the Internet one can ask the machine to show not only texts containing the words used in the search, but also texts that contain similar words. Robertson and Willett (1998) applied this method successfully on databases containing old English texts. They queried the texts using modern English, but were able to recover many old-English spelling variants.

Wrongly spelled words usually have a large similarity with their correct version. Indeed the most frequently occurring spelling errors are:

- adding an extra lettter
- omitting a ltter
- substituting another lotter for the correct one
- switching the position of two lettres

In all these cases the sets of n-grams (bi- or tri-) of the correct and the wrong version correspond to a large extent. Adaptations of this simple procedure

include, e.g., filtering out suffixes such as –ion and –ing. These suffixes increase (spuriously) the relation between words that are actually unrelated. Procedures like the one described here can be used for automatically correcting texts that are captured using OCR-techniques (optical character recognition techniques).

Although most experiments with n-grams are performed in English, there is no reason why the n-gram technique should not be used for other languages. The approach is indeed, completely language-free. Consequently, there do exist applications in German, Malay, Chinese and Japanese. East-Asian languages (Chinese, Japanese, Korean) seem to be very well fit for applications based on the n-gram technique (see, e.g. the article by Lee, Ng and Lu (1999)).

Analyzing titles of scientific articles via n-grams and cluster algorithms may lead to a classification of articles. A test on mathematical articles showed that the resulting classification was as good as one done manually based on a mathematical classification scheme.

For fixed ‘n’ (this is: the length of the n-gram) one studies how often each n-gram occurs. This leads to a very skewed distribution: most n-grams rarely occur, a few occur many times. This results in typical 'Zipf-Mandelbrot' distributions.

It is a very interesting theoretical problem to formulate a model to predict the frequency distribution of n-grams (n fixed) based on known letter frequencies. This has been done rather successfully by my Flemish colleague Leo Egghe (Egghe, 2000a,b).

In an article published in the top journal *Science,* Damashek was able to distinguish texts in different languages using an n-gram based clustering algorithm (Damashek, 1995). His results are truly remarkable: the cluster algorithm brings similar or related languages together. The whole idea can be considered as a form of artificial intelligence. Indeed, recognizing languages is a difficult task for most humans and otherwise (this is: without this n-gram algorithm) nearly impossible for machines.

Another interesting application of these techniques is in the biomedical field, where it has been used for the recognition of DNA strings. Moreover, some parts of a DNA-string contain code and other parts do not (non-coding or 'junk' DNA). N-gram techniques can distinguish between these two types.

Conclusion

It is no surprise to find out that Shannon is one of the most-cited authors in information science. He was one of the 39 scientists studied by White and Griffith (1981) in the first-ever article on author co-citation analysis and mapping. They

found that on the resulting MDS (multi-dimensional scaling) map Shannon and Zipf form a cluster of precursors in the field.

A recent investigation by White and McCain (1998) showed that Shannon belongs to the elite group of 75 authors that during three successive periods of eight year (1972-1979; 1980-1987; 1988-1995) belong to the most-cited authors in the information sciences.

We may rightly conclude that Claude Shannon is one of the most important and original scientists of the twentieth century. In a hundred year, when most names of movie stars, politicians and football players will long be forgotten, Shannon's name and his contributions will still be known to mankind.

Finally we like to mention that Sloane and Wyner (1993) edited the collected works of Claude Shannon. They also made his complete bibliography available on the Internet: http://www.research.att.com/~njas/doc/shannonbib.html

References

Beek, P. J. and Lewbel, A., *The science of juggling*, in: Scientific American, 273, (1995) 5; p. 74-79.
Cooper, W.S., *Exploiting the maximum entropy principle to increase retrieval effectiveness*, in: Journal of the American Society for Information Science 34 (1983), p. 31-39.
Damashek, M., *Gauging similarity with n-grams: language-independent categorization of text*, in: Science, 267, (1995), p. 843-848.
Egghe, L., *The distribution of N-grams*, in: Scientometrics, 47 (2000a), p. 237-252.
Egghe, L., *General study of the distribution of N-tuples of letters or words based on the distributions of the single letters or words*, in: Mathematical and Computer Modelling, 31 (2000b), p.35-41.
Egghe, L. and Rousseau, R., *Introduction to informetrics*, Amsterdam: Elsevier, 1995.
Golomb, S.W., *Claude Shannon (1916-2001),* in: Science, 292, (2001) p. 455.
Holmes, N., *The great term robbery*, in: Computer, 34(5) (2001), p. 94-96.
Horgan, J., *Claude E. Shannon. Unicyclist, juggler and farther of information theory,* in: Scientific American, 262, (1990), 1, p.16-17.
Lee, K. H., Ng, M. K. M., Lu, Q., *Text fragmentation for Chinese spell checking*, in: Journal of the American Society for Information Science, 50 (1999), p. 751-759.
Leydesdorff, L., *The challenge of scientometrics* . Leiden: DSWO Press, 1995.
Magurran, A. , *Ecological diversity and its measurement*. London: Chapman and Hall, 1991.
Nijssen, D., Rousseau, R. and Van Hecke, P. *The Lorenz curve: a graphical representation of evenness*, in: Coenoses 13 (1998) p. 33-38.

Peet, R.K., *The measurement of species diversity*, in: Annual Review of Ecology and Systematics, 5 (1974), p. 285-307.
Robertson, A.M. and Willett, P., *Applications of n-grams in textual information systems*, in: Journal of Documentation, 54(1),(1998), p. 48-69.
Rousseau, R., *De invloed van de Shannon-Weavertheorie op de informatiewetenschap*, (The influence of the Shannon-Weaver theory on the information sciences – in Dutch) in: Open, 18, (1986) p. 341-348.
Rousseau, R. and Van Hecke, P., *Measuring biodiversity*, in: Acta Biotheoretica, 47 (1999), p. 1-5.
Russell,S. and Norvig, P., *Artificial intelligence. A modern approach*, Englewood Cliffs (NJ), Prentice Hall, 1995.
Shannon, C.E., *A symbolic analysis of relay and switching circuits*, in: Transactions of the American Institute of Electrical Engineers, 57 (1938) p. 713-723.
Shannon, C.E., *A mathematical theory of communication I*, in: Bell System Technical journal, 27 (1948a), p. 379-423.
Shannon, C.E., *A mathematical theory of communication II*, in: Bell System Technical journal, 27 (1948b) p. 623-656.
Shannon, C.E., *Communication theory of secrecy systems,* in: Bell System Technical Journal, 28 (1949), p.656-715. Published version of a classified manuscript *A mathematical theory of cryptography,* dated 1945.
Shannon, C.E., *Prediction and entropy of printed English,* in: Bell System Technical Journal, 30 (1951), p. 50-64.
Sloane, N.J.A. and A.D. Wyner, A.D. (eds.). *Claude Elwood Shannon: Collected papers* - IEEE Press, 1993.
Weaver, W. , *Recent contributions to the mathematical theory of communication*, in: The mathematical theory of communication, Urbana (IL): University of Illinois Press, 1949, p. 1-28.
White, H.D. and Griffith, B.C., *Author cocitation: a literature measure of intellectual structure*, in: Journal of the American Society for Information Science 32 (1981), p.163-171.
White, H.D. and McCain, K.W., *Visualizing a discipline: an author co-citation analysis of information science, 1972-1995*, in: Journal of the American Society for Information Science, 49(4), (1998), p. 327-355.
Wigner E.P., *The unreasonable effectiveness of mathematics in the natural sciences*, in: Communications in Pure and Applied Mathematics, 13, (1960), p. 1-14.
Zipf, G.K., *Human behavior and the principle of least effort.* Cambridge (Mass.): Addison-Wesley Press, 1949.
Zunde, P., *Information theory and information science*, in: Information Processing and Management, 17 (1981), p. 341-347.
Zunde, P., *Empirical laws and theories of information and software sciences*, in: Information Processing and Management, 20 (1984), p. 5-18.