Coding Theory for the Classroom[*]

What do they all have in common: genetic code, pin code, bar code, EAN-code, ASCII-code and the hieroglyphs? They all serve to transmit information as short and unique as possible. Sometimes it is not so easy to "decode" the information. It can happen that the contents is made intentionally unreadable, then secret or ciphers are used. (This very interesting issue will be treated in the next chapter – Cryptography).

Braille writing, the Morse-alphabet and many other codes are serving for transmitting information by converting the information contents into another form which is adjusted according the contents, the message transfer and the users.

Task of coding theory is to develop mathematical methods for error free transmitting information with not only detecting possible transmitting errors but also correcting them.

Peter remembers the ASCII-Code

Peter has learned in information technology about the ASCII Code and he remembers that this objective was mentioned last year in his math classes. His task is to encode his name "PETER" in ASCII code and in binary code.

As he does not like to search for his text book containing the ASCII-table he uses his computer and his CAS in order to convert the letters first in ASCII-Code numbers and then the decimal numbers into the respective binary numbers. Finally he fills it up – how he has learned last year – to an 8 digit binary word to obtain a complete byte. Peter uses Derive:

```
NAME_TO_CODES("PETER") = [80, 69, 84, 69, 82]
```

This reads in binary words:

```
01010000 01000101 01010100 01000101 01010010
```

The teacher gives the task to use the leading digit, which was used to fill up the words to complete bytes meaningful as a *check bit*: if there is an even number of ones then leave the zero, in the other case replace the zero by a one.

Peter follows the instruction and encodes his name:

```
01010000 11000101 11010100 11000101 11010011
```

The teacher needs only one glance at the result to recognize that Peter made an error.

01 Why does the teacher detect Peter´s error immediately?

02 Encode your full name applying check bits.

03 Which kind of errors are detected by the parity check and which are not?

04 In which cases will it be sufficient to only detect an error?

05 Give at least three more codes which are serving information transmission.

06 Hieroglyphs were no secret writing. Nevertheless it took a long time to "break" the code. Give a short report who was able to decipher the meaning of the hieroglyphs and how it was done. (Internet-Recherche)

07 Inform yourself about the ANSI-Code and the Unicode. What is the difference to the ASCII-Code.

The leading bit is called **Check Bit**. Checking encoded information using this character is called **Parity Check**.

1 Codes Everywhere

Modern economy and administration are unimaginable without using computers. Together with more and more "computerization" the application of codes reaches far into our all private sphere. Some examples shall illustrate this.

1.1 The ISBN-Code

It was in 1973 when the ISBN-Code was introduced to register books (ISBN = International Standard Book Number). Each book was encoded by a number consisting of 10 digits. The first nine figures are digits 0, 1, …, 9. The last place can also be occupied by an X. This code is made by four blocks of characters which are separated by a hyphen or a blank.

$$\underbrace{x_1}_{\text{Language}} - \underbrace{x_2\,x_3\,x_4}_{\text{Publisher}} - \underbrace{x_5\,x_6\,x_7\,x_8\,x_9}_{\text{identification number}} - \underbrace{x_{10}}_{\text{check digit}}$$

The ISBN has changed and was adjusted to the EAN-Code which will be discussed in the next chapter. See problem **26** and the picture at the right.

English is encoded by digits 0 and 1, German by 3, French by 2 and Italian by 88. The distribution of the digits can vary. The task of the check bit is to take care that errors caused by an exchange of digits or by technical reasons will be detected.

How can we calculate the check digit of our example book ("The Code Book" written by Simon Singh)?

The procedure is simple enough:

One has to multiply the nine digits one after the other from the left by the numbers 10, 9, …, 2 and sum up the products. The check digit is the difference of the "weighted sum" and the next number which is divisible by 11.

> **Example:** We recalculate the check digit of the "Code Book".
>
> **Solution:** $10 \cdot 1 + 9 \cdot 8 + 8 \cdot 5 + 7 \cdot 7 + 6 \cdot 0 + 5 \cdot 2 + 4 \cdot 8 + 3 \cdot 7 + 2 \cdot 9 = 252$.
>
> The next number divisible by 11 is 253. Hence the check digit is 1. But this is a tiresome search! Working with the remainder of the division makes things easier: $252 : 11 = 22$, remainder 10. So we need 1 to find the next multiple of 11 to have a division without remainder. The result of the division – 22 – is of no interest at all, only the remainder is important.

Calculation with remainders is an essential base of coding theory and of cryptography as well. In the chapter about Cryptography we will much more about calculation with remainders which is called **modular arithmetic**.

We will introduce here an important concept:

> For the integer remainder r at division $a:m$ the notation a **modulo** m ($a \in Z$, $m \in N$, $m > 1$). We write $r = a$ **mod** m. In our case we have 252 modulo 11 = 10 or shorter 252 mod 11 = 10. This is equivalent to $a - r$ is divisible by the module m or $a - r$ is a multiple of m.

This remainder can be found with every technology tool, because (almost) everywhere the modulo function is implemented.

It is now much easier to find the check digit:

```
#1:    11 - MOD(252, 11) = 1
```

This is the calculation done with Derive. The symbolic calculator works in the same way. In Excel we find the function as **REST(Zahl;Divisor).** The graphing calculator needs a short – self made – function (program).

You will surely ask: "And what shall I do, if the check digit will be 10? There is only one character reserved for this digit." Read carefully the first paragraph on this page. Right, then you have to take the X (the Roman character for 10.)

08 How can we find the remainder a mod m without using the mod-function?

09 Calculate first without and then with using `mod(a,m)` or `REST(a;m)`:

a) 34561 mod 131 **b)** 98765 mod 4321 **c)** 100100 mod 39

10 If you are working with a graphing calculator then write a short program for finding the remainder x mod m.

(Hint: Function `fPart(a/b)` will help.)

Does the program on this calculator a good job?

```
PrgmMOD
X=?569874
M=?998
                16
              Done
```
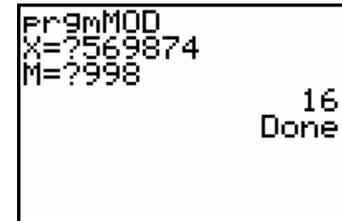
11 You can use an Excel-worksheet or a self made function of your CAS-tool to find the check digit of the following incomplete ISB-numbers.

	A	B	C	D	E	F	G	H	I	J
1	1	8	5	7	0	2	8	7	9	Prüfziffer
2										1
3										

Don´t forget to convert the possible remainder 10 to the character X!

```
F1▾ F2▾ F3▾ F4▾ F5 F6▾
▼ Algebra Calc Other PrgmIO Clean Up

■ isbn("370681188")        "370681188X"
■ isisbn("370681188X")           "ISBN"
■ isisbn("3706811880")     "keine ISBN"
isisbn("3706811880")
KRYPT        RAD AUTO        FUNC 3/30
```
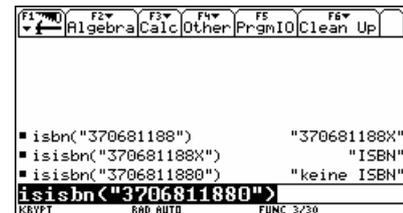
12 Complete the ISBN-Codes with their check digits:

a) 3-209-03076-☐ b) 84-87282-78-☐ c) 83 7085 818 ☐

d) 1 884 799 13 ☐ e) 2-10-003104-☐ f) 3-209-04627-☐

From which language regions are the publishing companies?
Can you find the title of one or the other of the given books?

13 The renowned German publishing company Vieweg has the publisher number 528; one of its books dealing with financial mathematics has book number 14164. What is the ISBN-code of this book?

14 Check the correctness of the ISB-numbers. A short program might help!

a) 3-528-18990-3 b) 3-8273-1082-2 c) 1-55953-407-9

Example: Why does the calculator for mod(−18,13) show the result 8 and not −5?

Solution: $r = -18$ mod 13 is equal to the fact that $-18 - r$ is divisible by 13. The next multiple of 13 is −26, hence r is 8. The mod-function returns always positive values. You will learn in the next chapter that −5 und 8 are equivalent in modulo arithmetic.

15 What can you conclude from the following calculations?

a) 23 mod 10 and −23 mod 10 **b)** 115 mod 27 and −115 mod 27

c) 1234 mod 39 and −1234 mod 39 **d)** Can you prove your conjecture?

16 Assumed that negative modules are accepted. What are the results and which relations do you recognize? Do your tools permit negative modules:?

a) 105 mod 17 b) −105 mod 17 c) 105 mod −17 d) −105 mod −17

Peter explains Ann the ISBN-Code. Ann is really impressed and experiments with the remainders. As she was not attentive enough she does it another way: she multiplies the first nine digits of an ISB-number one after the other by 1, 2, …, 9 and forms the remainder modulo 11 of the sum of the products. The result is the right check digit. Peter watches her and thinks that this is only pure chance. Ann works through a second example – and again by „pure chance?" the check digit is correct.

Is this really mere accident? Take any three ISBN-Codes and double check the check digit applying „Ann´s method".

> **Example:** Very excited Peter tries to prove that both methods are equivalent. We follow his calculation.
>
> **Solution:** According to the first method the weighted sum is filled up to the next number which is divisible by 11 by the check digit p_1:
>
> $$10x_1 + 9x_2 + 8x_3 + 7x_4 + 6x_5 + 5x_6 + 4x_7 + 3x_8 + 2x_9 + p_1 = 11 \cdot k_1 \qquad (1)$$
>
> According to "Ann´s Method" the check digit p_2 is the remainder of the division of the "reverse weighted" sum by 11. So we can note:
>
> $$1x_1 + 2x_2 + 3x_3 + 4x_4 + 5x_5 + 6x_6 + 7x_7 + 8x_8 + 9x_9 - p_2 = 11 \cdot k_2 \qquad (2)$$
>
> It seems to make sense to add both equations. We can factor out 11 on both sides of the new equation.
>
> $$11 \cdot (x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9) + p_1 - p_2 = 11 \cdot (k_1 + k_2) \quad \text{and then}$$
>
> $$p_1 - p_2 = 11 \cdot (k_1 + k_2 - x_1 - x_2 - x_3 - x_4 - x_5 - x_6 - x_7 - x_8 - x_9).$$
>
> One can see that $p_1 - p_2$ is divisible by 11 without remainder or $p_1 - p_2 \bmod 11 = 0$. This can occur only if $p_1 = p_2$.

Peter is really proud of himself and he decides to inform about other codes. Peter and Ann have seen very often the bar code which can be found now together with the ISB-number. Supported by a search engine they go for a "code hunt" in the Internet.

1.2   The EAN-Code

In our times of progressing globalizing it became a need to identify the products from all over the world in a unique way. It was 1973 when the company IBM developed the first bar code UPC (Universal Product Code). 1997 followed the „European Article Number Code" EAN which is very common in Austria.

The EAN-Code consists of 13 digits. The first two of them (sometimes the first three) form the reference number for the country. They are followed by two groups of digits for the producer and the characteristic number for the product. Some important country reference numbers are: Austria 90 and 91, Italy 80 and 81, USA 00-09, Japan 49 and Switzerland 76.

The check digit is calculated by multiplying the first 12 digits one after the other alternatively by 1 and 3. The number which completes the sum of the products to the next multiple of 10 is the check digit.

We will double check this procedure with the EAN-code of an important utensil of the desk.

The first 12 digits read

> 900221903001.

The according to the instructions from above weighted sum results in

$$1 \cdot 9 + 3 \cdot 0 + 1 \cdot 0 + 3 \cdot 2 + 1 \cdot 2 + 3 \cdot 1 + 1 \cdot 9 + 3 \cdot 0 + 1 \cdot 3 + 3 \cdot 0 + 1 \cdot 0 + 3 \cdot 1 = 35$$

The completion to 40 is 5 and this is the check digit.

We find a "formula" as we did earlier with the ISBN-Code: EANChD = 10 – weighted sum mod 10.

17 Which is the reference number for products from Germany?

18 Discuss in your industrial management classes the advantages of the EAN-Code and its translation in a form which is easy readable by a scanner – the bar code. Where have you noticed this bar code?

19 What is the connection between a bar and the barcode? What is the connection between a pin and the PIN-Code? Explain the name of these both codes.

20 Complete the EAN-Codes:

a) 8 711500 86820 ☐   b) 9 00163411200 ☐   c) 4 902050 ☐11026

21 Double check the EAN-Codes. A tool – CAS, spreadsheet or program – can help!

a) 5 010327 000299   b) 4 009404 001018   c) 9 100000 029672

The check digit in the ISBN-Code makes possible to recognize single errors and exchange errors. (But they cannot be corrected!)

Example: There is an ISB-number with digit 8 on a certain position. Show that no other digit can be placed on this location.

Solution: Let us assume that the eight shall be on digit #7. We compare the correct ISBN with the wrong one (any digit z on place #7):

correct number: $10x_1 + 9x_2 + 8x_3 + 7x_4 + 6x_5 + 5x_6 + 4 \cdot 8 + 3x_8 + 2x_9 + p = 11 \cdot k$

wrong number: $10x_1 + 9x_2 + 8x_3 + 7x_4 + 6x_5 + 5x_6 + 4 \cdot z + 3x_8 + 2x_9 + p = 11 \cdot k$

We subtract both equations and we obtain: $4 \cdot (8 - z) = 0$. The only solution for this equation is $z = 8$ – or the remainder of the division of the left side by 11 can only then be 0 if $z = 8$.

22 Choose any other digit on any arbitrary place and show that this will lead to a contradiction when an error occurs by transmitting this digit.

23 In an ISB-number are digits 5 and 2 on places 3 and 4. They are exchanged when the ISBN is sent to a data base. Show that this error is detected by the implemented check. Errors of this kind are called transpositions.

24 Show that exchanges of non neighbored digits will be detected. Choose any two digits on any arbitrary places. You may try a general proof.

25 Double errors are not detected in all cases. Digit #7 and #8 are 8 and 2. Instead of 8 number 6 is sent. Which wrong number(s) on place #8 are not detected by the check digit?

26 The picture of the ISBN-Code at the beginning of chapter 1.1 shows also a bar code and the EAN-Code of the book. It is easy to obtain the respective EAN-Code to a given ISBN-Code. One has to attach the initial digits group 978 or 979 to the ISB-number (without its check number) and calculate a new check number according to the rules for the EAN-Code.

Find the respective EAN-Codes to the given ISB-number using 978 as leading digits group:

a) 3-209-03076-6   **b)** 84-87282-78-4   **c)** 83 7085 818 X

27 The EAN-Code is structured to detect especially "phonetic" errors: 40 exchanged with 14, 50 with 15, … or vice versa.
Show that in any EA-number with the sequence of digits 16 at any place the wrong sequence 60 will be detected in any case.

28 Like problem 27: instead of 90 on places 3 and 4 the number 19 is transmitted. Show that this error will be detected by the check number.

29 Demonstrate using a self chosen example that in the EAN-Code not only phonetic errors but also singular errors are detected.

1.3 More Codes

30   The Austrian Social Insurance Number contains a check number, too.

$$\underbrace{x_1 x_2 x_3}_{\text{running number}} \qquad \underbrace{p}_{\text{check number}} \qquad \underbrace{T\,T\,M\,M\,J\,J}_{\text{date of birth}}$$

The sequence of digits is multiplied from left to right by 3, 7, 9, 5, 8, 4, 2, 1 and 6. The products are added. Check number p is the sum modulo 11. (In case of remainder 10 the running number will not be used.)

a)   161p111145, find check number p.   b)   Double check your social insurance number.

31   Each company receives a 7 digit company number when it is registered. The last digit is a letter as "check number".

The first six digits (which sometimes must be filled up with leading zeros) are multiplied one after the other by 6, 4, 14, 15, 10 and 1. The remainder of the sum of the products modulo 17 is converted in a letter according to the following table.

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A	B	D	F	G	H	I	K	M	P	S	T	V	W	X	Y	Z

a)   Why are here "check letters" used?

b)   A company is newly registered and is assigned the company number 23456. What is the complete number including the check letter?

32   In credit cards (Master, EURO, VISA) one can calculate the check number which is the last digit as follows: the digits are – beginning from the right – alternatively multiplied by 2 and 1. The check number is the result of a subtraction 10 – sum of all digits of all products modulo 10. In case of the result 10 then the check number is zero. This method works also for the Miles & More card of airlines.
Double check at least two card numbers of this kind.

2   Source Codes

A source code is a translation of an information into a form which is readable by the machine (the computer). An amount of characters (= alphabet) is needed in order to form "code words". The set of all code words as a whole is called "code". The best form is the binary representation as shown in the introductory example. The alphabet consists of {0,1} only and the code consists of 8-digit binary words. Additionally the source coding is connected with error detection.

2.1   Zipped today? – the Huffman-Code

When you want to send big amounts of data via email, then you will probably compress the file(s). Some programs do the job. One of these programs gave the name: "zipping". (Which one?) Compressed files show mostly the file extension zip or rar. Especially graphic files (photographs, scanned images, …) may become very large. It is easy to reach a couple of megabyte. Intelligent algorithms provide a compression without loss of data to a significant smaller amount of data.

Name ⭡	Typ	Datum	Größe	Komp...	Kompri...	P
Verkehr.dfw	Derive Worksheet	13.10.2004 15:37	7 220 411	99%	91 381	

The picture shows the result of „zipping" or „packing" a Derive-file which contains some graphs. You can see the efficiency of the compression algorithm.

A very simple method is the following: in a graph appears a sequence of 3878 white image points (pixels) followed by a sequence of 132 black pixels. Instead of listing w, w, ..., w, b, b, ..., b one can note much shorter: w, 3878, b, 132, loosing no information at all. The next paragraph is an information which shall be used for demonstrating compression for transmission.

> "bei der komprimierung von texten laesst man sich von der unterschiedlichen haeufigkeit der zeichen in dem zu codierenden text leiten. wir wollen das an einem einfachen beispiel demonstrieren, wobei wir nur kleinbuchstaben, zwischenraeume und satzzeichen verwenden wollen. dieser text wird verwendet."

Characters with a high frequency will be assigned to short code words in order to save bits!
This basic idea is also realized in the Morse-Code. Character "e" is encoded by a code word of length 1, the "·", character "q" which is pretty rare by a code word of length 4, "− − · −".

We create a „binary tree". This is a directed graph consisting of vertices or nodes and edges (arrows) (= branches of the tree). Our first message "demo" to be encoded is "dieser text wird verwendet". Using the ASCII-Code without check bit the length of the message is 189 bit.

First of all we find out the frequency of the characters by simply counting. This can be done manually. For extended texts we will use the computer.

$$
\text{haeuf(dieser text wird verwendet.)} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 2 & 2 & 3 & 3 & 3 & 3 & 6 \\ . & n & s & v & x & i & w & & d & r & t & e \end{bmatrix}
$$

This is the frequency table for the paragraph from above starting with: "bei der komprimierung … ".

haeuf(absatz)

$$
\begin{bmatrix} 2 & 2 & 2 & 2 & 3 & 3 & 3 & 4 & 5 & 5 & 7 & 8 & 8 & 9 & 9 & 9 & 9 & 9 & 11 & 14 & 14 & 17 & 25 & 29 & 40 & 50 \\ , & f & g & p & . & k & x & v & b & z & m & o & u & a & c & h & l & w & s & d & t & r & i & n & & e \end{bmatrix}
$$

The procedure is performed as follows:

Generate a node for every character appearing in the message. Label all nodes with their weights (= frequencies).
Until there is only one node remaining with no arrow directed to it, do:
Connect two nodes with minimal weights which are not end points of an arrow by a new node. The weight of this "parent node" is the sum of the weights of the "children". The arrows are directed from the parent node to the children nodes.   The arrows are named as 0 (the left edge) and 1 (the right edge) by convention.

On the left you can find the **structogram** for the algorithm. We will follow the instructions and create the Huffman-Tree for the code of our message.

Then we will check the efficiency of the code, apply it and demonstrate how to decode the encoded message into a readable form again.

This is another definition (found in http://en.wikipedia.org/wiki/Huffman_coding):

The process essentially begins with the leaf nodes containing the probabilities of the symbol they represent, then a new node whose children are the 2 nodes with smallest probability is created, such that the new node's probability is equal to the sum of the children's probability. With the previous 2 nodes merged into one node (thus not considering them anymore), and with the new node being now considered, the procedure is repeated until only one node remains, the Huffman tree.

Steps 1 to 3
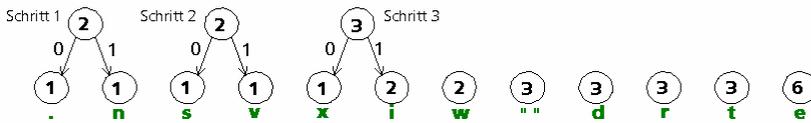
We select the pairs of nodes with minimal weights one after the other.
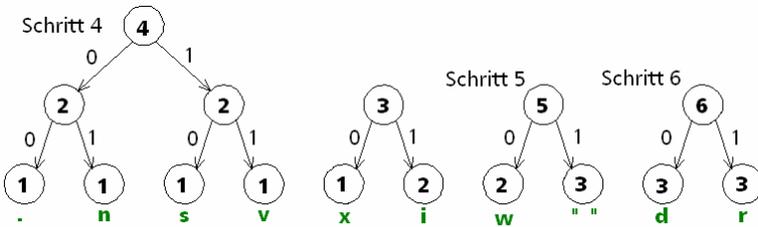
Pair (2,2) is connected by parent node 4. Then we have to more pairs of nodes which are connected by edges (steps 5 and 6).
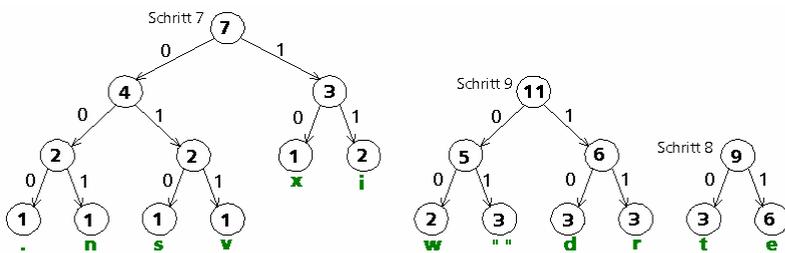
Vertices 3 and 4 form a "minimal pair" and are led together in vertex 7. Parallel children nodes 5 and 6 find their parent node (initial node of two arrows) 11.

The next pair is formed by nodes with the weights 9 and 7.
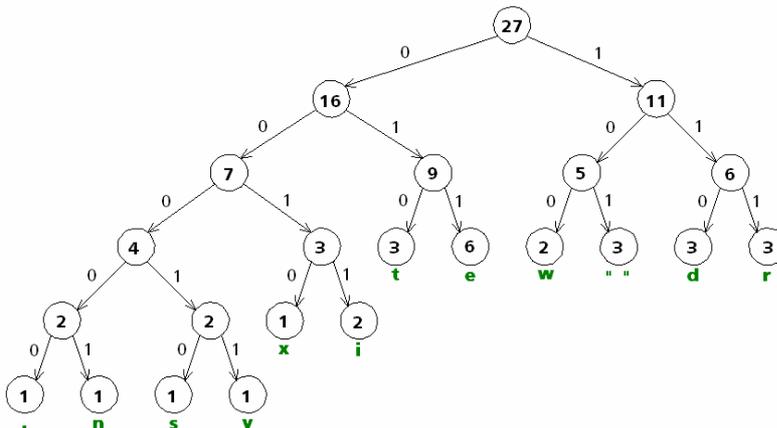
Having connected 9 and 7 to 16 the branch starting with node 11 is remaining. Both have the same root (parent) 27.

Only one vertex (27) with no arrow directed is remaining. According to our instructions the job is done. It is an easy check to compare the weight of the final vertex (the root of the binary tree) with the sum of all (absolute) frequencies which is 27.

We observe that frequently used characters are located close to the root, rarely ones can be found in the "tree-tops".

The codes for the characters are yielded by following the graph from the "root" to the "leaf" and noting the labels of the edges along the path. The code for the "e" is 011, the code for the "x" is 0010, the "r"-code is 111 and the "v" is encoded by 00011. You see again that rare characters result in long paths which are equivalent to long code words and frequent characters give short paths and consequently short code words. The Huffman-Code is the code which needs the minimal number of bits. (This can be proved.)

Here is the complete code followed by the encoded message:

(demo is the message "dieser text wird verwendet.")

$$\text{hcode1} := \begin{bmatrix} e & t & w & d & r & x & i & . & n & s & v \\ 011 & 010 & 100 & 101 & 110 & 111 & 0010 & 0011 & 00000 & 00001 & 00010 & 00011 \end{bmatrix}$$

hufcode(demo, hcode1)

11000110110001001111110101001100100101011000011111110101000110111111000110000111001101000000

The encoded message is decoded by following the path starting in the root bit for bit (go left for 0 or right for 1) until reaching a leaf. There you will find the respective character:

110|0011|011|00010| .... results in: **dies** ...

Of course, the code must be transmitted together with the encoded message (if the code is generated from the message). Each language has typical frequencies for the occurrences of the letters. If the partners agree – and the message is sufficiently long – you can do without sending the code and rely on the typical frequency of the letters in the respective language.

33 Check if the Huffman-Tree from above is really giving the code of minimal length – or if there is a better one?

34 It is possible to read off the length of the encoded message from the Huffman-Tree. Can you do this?

35 Encode the complete message "demo" supported by the Huffman-Tree. Compare the number of the bits of the compressed message. If we had only 12 characters we would need only 4 bits applying conventional coding. Would we benefit of Huffman-encoding?

36 Why does the "adaptive method", which determines the code from the message bring a real advantage only if applied for longer messages?

37 Create the Huffman-Code for: "MATHE MIT GEWINN 2". Encode and decode. Create two different (but correct) Huffman-Trees and demonstrate that both codes produce encoded messages of equal lengths.

3   Channel Codes

The next figure shows the complete way of a message from the sender to the receiver. Messages are often encoded once more for transmitting. Codes for this purpose are called channel codes. The receiver needs two steps in order to decode the received message. The set of all permitted code words is called code. Especially when data are transmitted (via a "channel") transmitting errors can occur. Having in mind which distances messages have to pass in space you can surely imagine that some bits might get lost or might change from 0 to 1 or vice versa. Hence, there is a need on error correcting codes.

	Message	Signal noise, Interference	Channel Decoding	
Sender	Source Encoding	- - - - - - - - - - - - ->	Source Decoding	Receiver
	Channel Encoding	Eavesdropper, ...	**Message**	
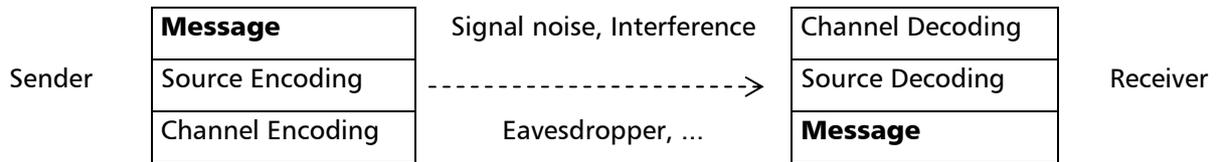
3.1   The Repetition Code

Development of efficient error correcting codes became a separate branch of information science. The easiest method is to send the message three times repeated assuming that is is very unlikely for a transmitting error to occur twice at the same position.

Captain Kirk sends an information from the depth of the galaxy to space station LUNA which is consisting of only 8 bit 11001100. He uses the **repetition code** (so he sends 11001100 11001100 11001100). Heavy electro-magnetic storms interfered the signal and Mr. Spock receives (11011100 11001101 1100?100).

What was Cpt. Kirk´s message?

Mr. Spock puts back his peaked ears and then notes the three blocks one beneath the other:

$$\begin{matrix} \texttt{11011100} \\ \texttt{11001101} \\ \texttt{1100?100} \\ \texttt{11001100} \end{matrix}$$

He compares column for column. If all bits in a column are the same, then everything is ok. In the columns without a complete conformity he can assume that at most one signal is not correct. and he takes the bit which appears twice.

Including of additional – **redundant** – bits raises the transmission security. The **information rate** is the ratio of the length of the information and the overall length of the code word. The information rate of the three-times-repetition code is 1/3.

An important quantity in coding theory is the **Hamming-Distance** of two code words (Richard Wesley Hamming, 1915–1998). It is the number of digits where two permitted code words differ. The minimum hamming distance occurring is the **distance of the code**. The greater the code distance the more errors in a code word can be detected or corrected.

> **Example:** Which Hamming distance is between 110111000 and 11100001 and between 10101 and 11010?
>
> **Solution:** 11011100 und 11100001 differ at 5 digits, hence the Hamming distance is 5. 10101 and 11011 have the Hamming distance 3.

38   Find the Hamming distances of

a) 110011 and 101010   **b)** 1110000 and 0001110   **c)** 0010010 and 0010011

39   Give all 3- bit words which have a Hamming distance below 2 from 101.

40   What is the information rate of the ASCII-Code with check bit?

41   A certain code serves to transfer information 00, 01, 10 and 11. For this purpose three check bits are appended. The complete code consists of the words 00000, 01111, 10101, 11010. All errors which are caused by an incorrect transmitted bit cannot only be detected but also corrected.

a) What is the information rate of this code?

b) Find all Hamming distances between the code words and the distance of the code.

c) Give a list of all words with Hamming distance 1 to 01111.

d) If there are wrong words in a message occurring then it will be exchanged by the word of the code with the minimum Hamming distance.
Which message was transmitted by (00000 01101 01111 11011 10101 11010 00100)?

e) Try to find at least one – incorrect – word which cannot be corrected.

3.2   The Hamming-Code

The Hamming-Code is a sophisticated system for error correcting. For transmitting a 4-bit-message three additional check bits are needed if one wants to correct one error per code word. The seven digit code word is composed in the following way: The data bits (the source encoded message) are placed on the digits 3, 5, 6 and 7. Digits 1, 2 and 4 (powers of 2!) are occupied by check bits according the following rule (these are again parity bits, depending on the odd or even number of ones):

message = 4 data bits   encoded message consisting of data- + check bits

digit	1	2	3	4	5	6	7
n1 n2 n3 n4 $\rightarrow$	n1+n2+n4 mod 2	n1+n3+n4 mod 2	n1	n2+n3+n4 mod 2	n2	n3	n4
0 1 1 0	**1**	**1**	**0**	**0**	**1**	**1**	**0**

Let´s see what happens in case of an incorrect transmission of this code word? The parity check works as follows:

Parity 1: (St1+St3+St5+St7) mod 2=(1+0+1+0) mod 2=0
Parity 2: (St2+St3+St6+St7) mod 2=(1+0+1+0) mod 2=0
Parity 3: (St4+St5+St6+St7) mod 2=(0+1+1+0) mod 2=0

The three results give read from below to above the three digit binary number $(000)_2=0$. This number is called **Syndrom**. Syndrom 0 indicates the correct transmission.

The code word **1100110** is received as **1100100**. We check the three parities in the same way again.

Parity 1: (St1+St3+St5+St7) mod 2=(1+0+1+0) mod 2=0
Parity 2: (St2+St3+St6+St7) mod 2=(1+0+0+0) mod 2=1
Parity 3: (St4+St5+St6+St7) mod 2=(0+1+0+0) mod 2=1

Syndrom $(110)_2=6$ informs us that the bit on digit 6 has to be changed (from 0 to 1). The correct information reads **1100110**.

42   A message is source encoded by converting the characters into the ASCII-Code (using only ASCII-Codes 32 – 90). The two digits of the ASCII-Codes are represented as binary numbers and then sent applying the Hamming–Code.
Encode the message "HELLO WORLD!" Double check the syndroms of the code words.

Binary Coding of decimal numbers is called **BDC** = **Binary Decimal Code**.

43   Transfer the date 12092008 via BDC into Hamming-Code.

44   Kirk sends from his space-ship STAR-INVADER a message to Mr. Spock:

0011111 0001111 1100110 0101101 1110000 0101000 111000 1000011 1000010 0101010 1001100 1011001.

It has suffered in the deep of the universe. What does Kirk want Mr. Spock to know?

45   Supported by a spreadsheet program you can provide a coding table.

Which text was encoded on the Voyage 200?

Note: The picture shows performing the encoding with a spreadsheet program for a symbolic calculator. Working with MS-Excel looks pretty the same.

4 Summary

Coding theory delivers the basics for generating efficient codes. Its aim is to reach maximum data security combined with error detecting and error correcting together with a minimum redundancy.

Important fields of application of coding theory are among others data transfer in space, encoding of data on CDs and compressing graphic- and other files (zipped files).

5 Supplementary tasks

46 The check bit in the 8 digit ASCII-code with check bit can be generated by a modulo calculation. The parity check for double checking a correct ASCII-code can be performed in an easy way applying modulo arithmetic, too. What are the two calculations?

47 Using your technology create a table, a function or a program which enables checking Hamming-encoded messages on its correct-ness and correcting 1 bit errors.
(The picture shows a possibility using Derive.)

```
hamdecode(0001111) = [0001111, 0001111, 0111, 7]

hamdecode(0001011) = [0001011, 0001111, 0111, 7]
```

48 Take any three legal (= correct generated) code words from problem 43 and find their Hamming-distances.