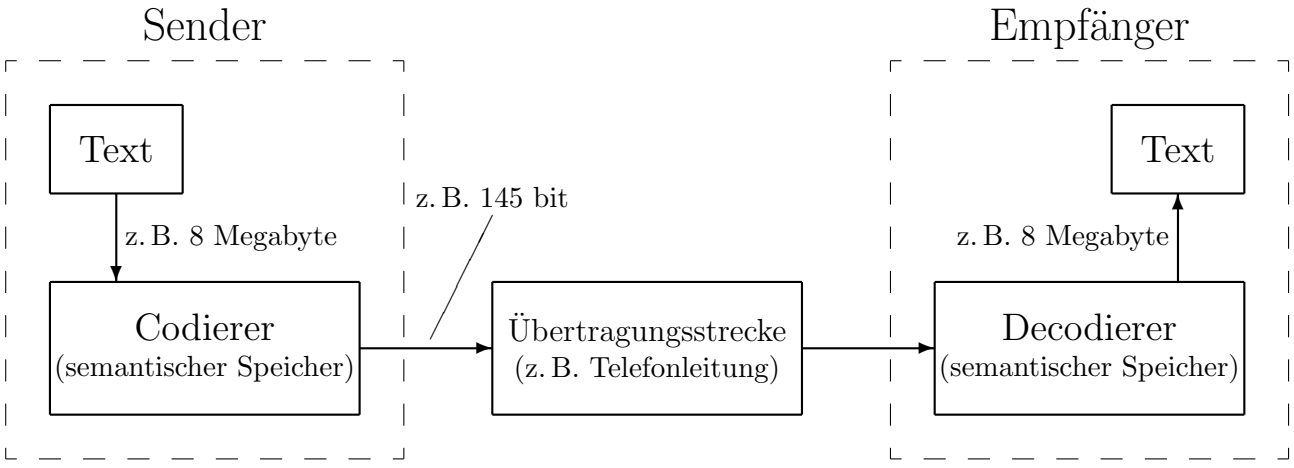
Thema:

DIE VERWENDUNG HIERARCHISCH STRUKTURIERTER SPRACHNETZWERKE ZUR REDUNDANZARMEN CODIERUNG VON TEXTEN

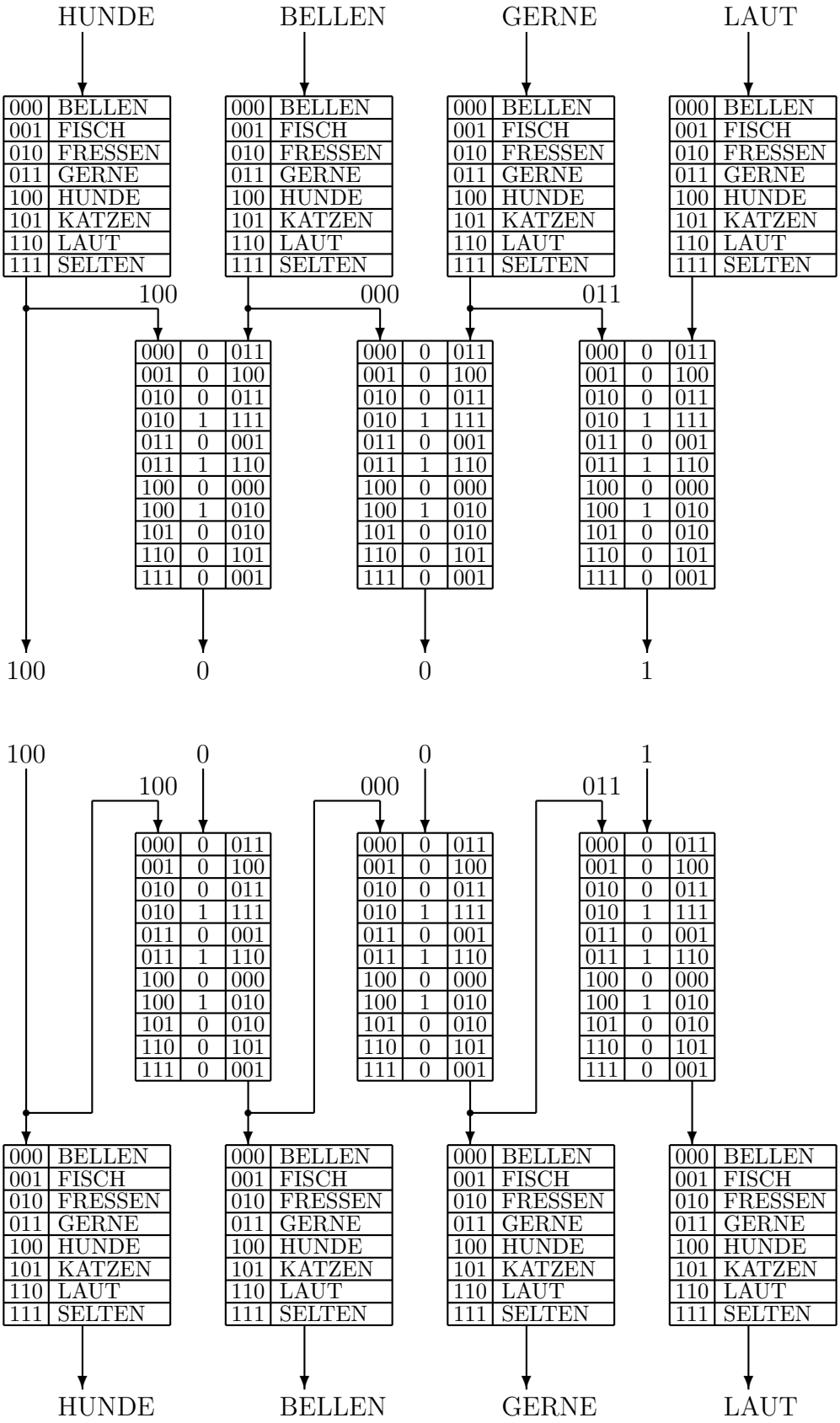
Übersicht:

1 Problemstellung:

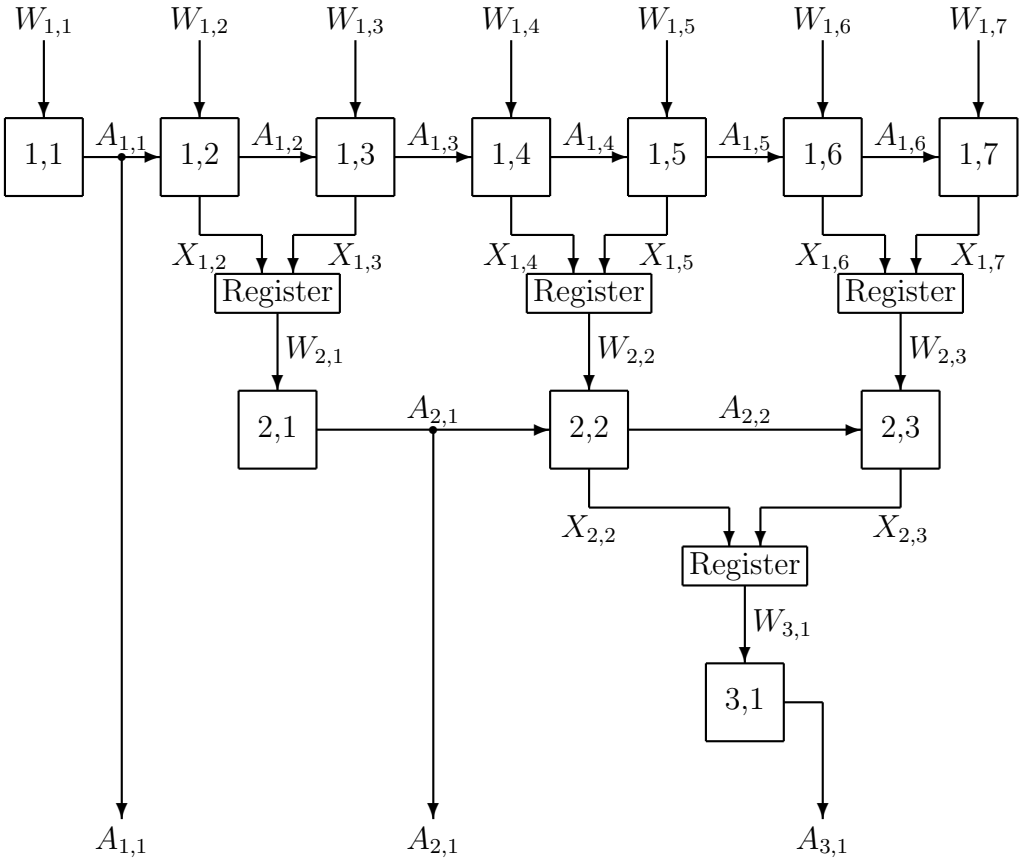


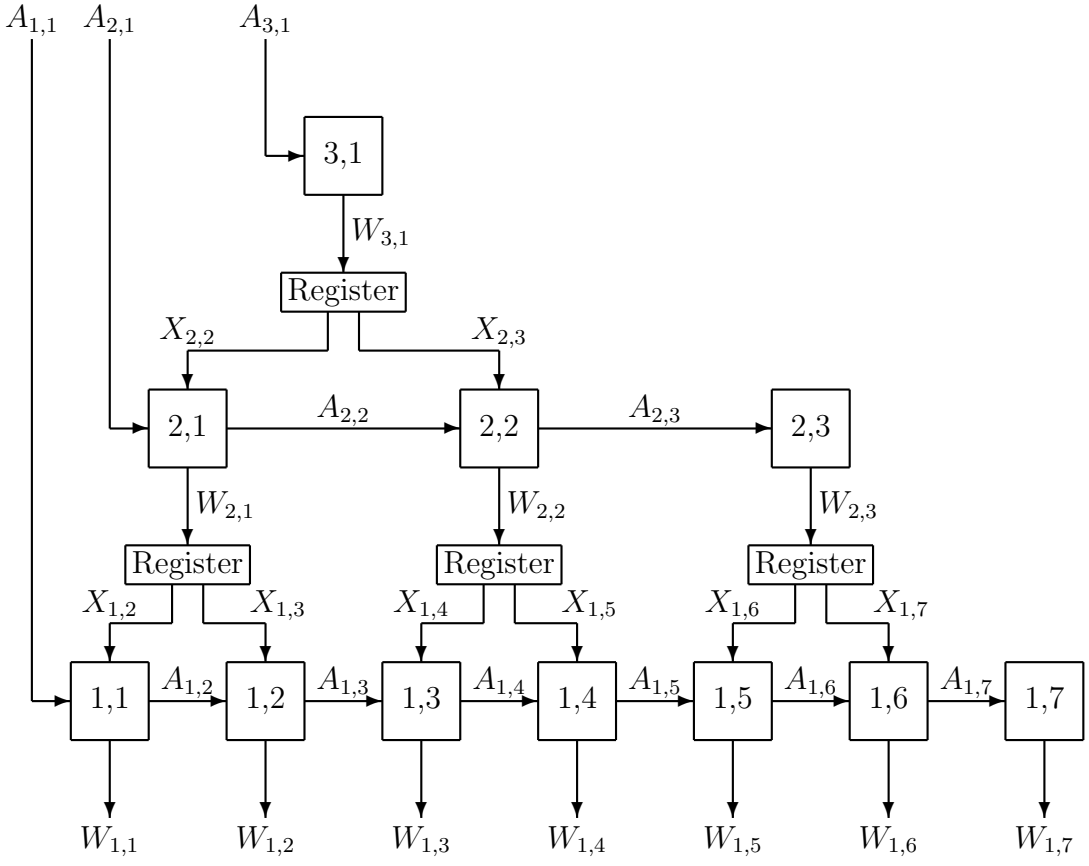
2 Das Konzept des semantischen Speichers:

2.1 Das assoziative Feld:

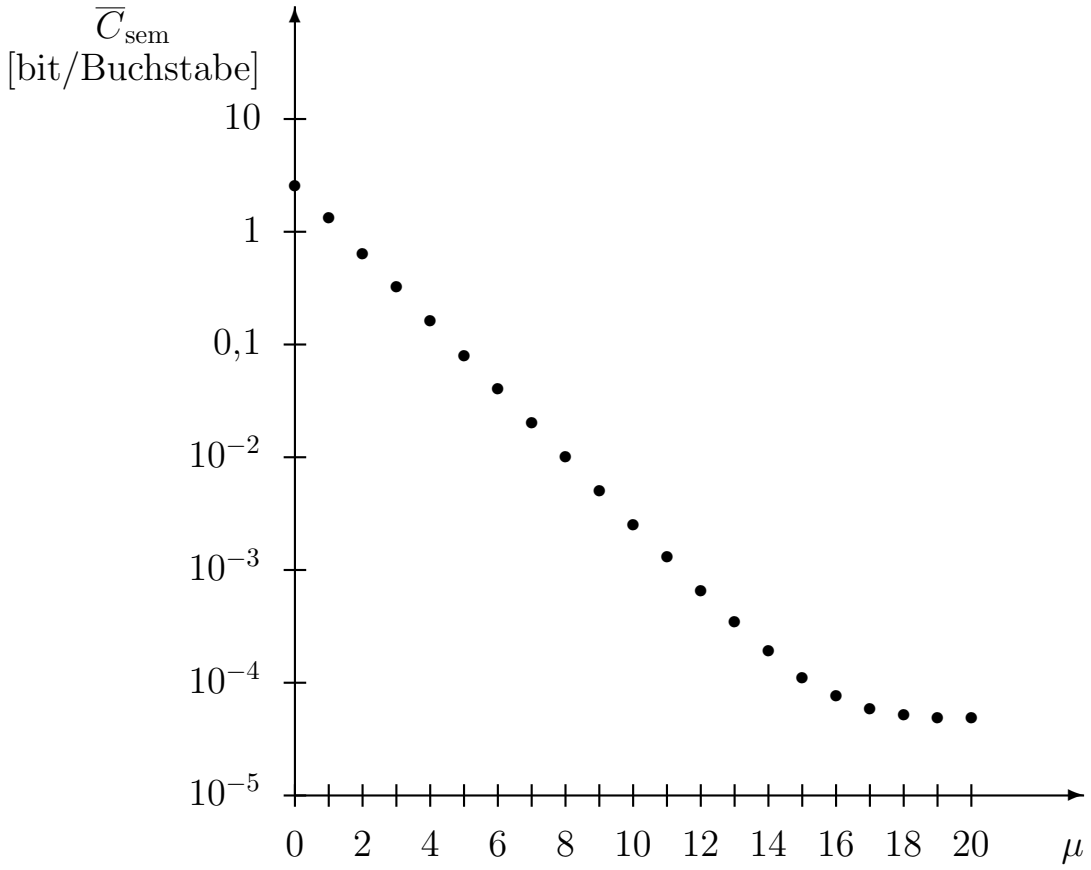


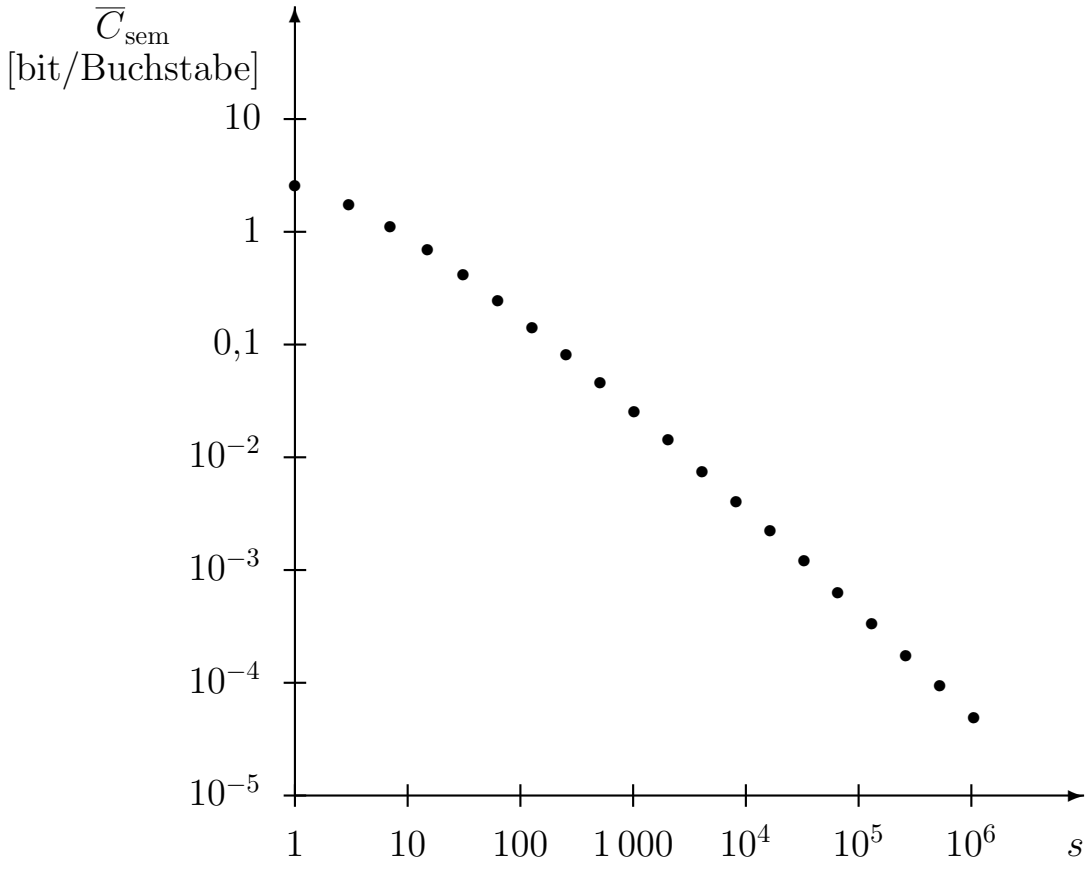
2.2 Der semantische Speicher:





2.3 Berechnung der mittleren Codelängen:



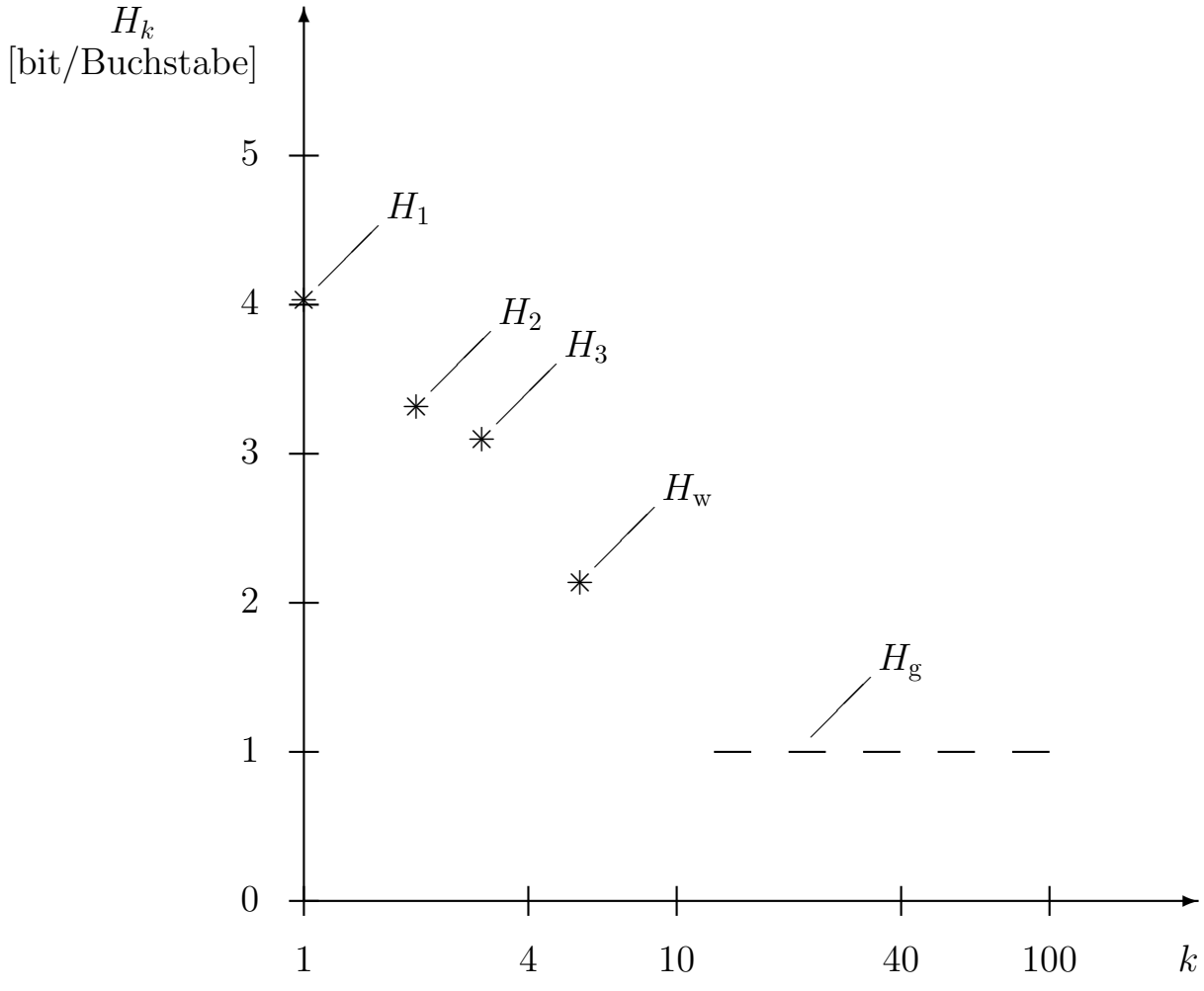


3 Analyse von Texten:

3.1 Der mittlere Informationsgehalt von Texten:

$$H = -\sum_{n=1}^{N} p(n)\ \mathrm{ld}\, p(n) \qquad \text{(in bit/Symbol)}.$$

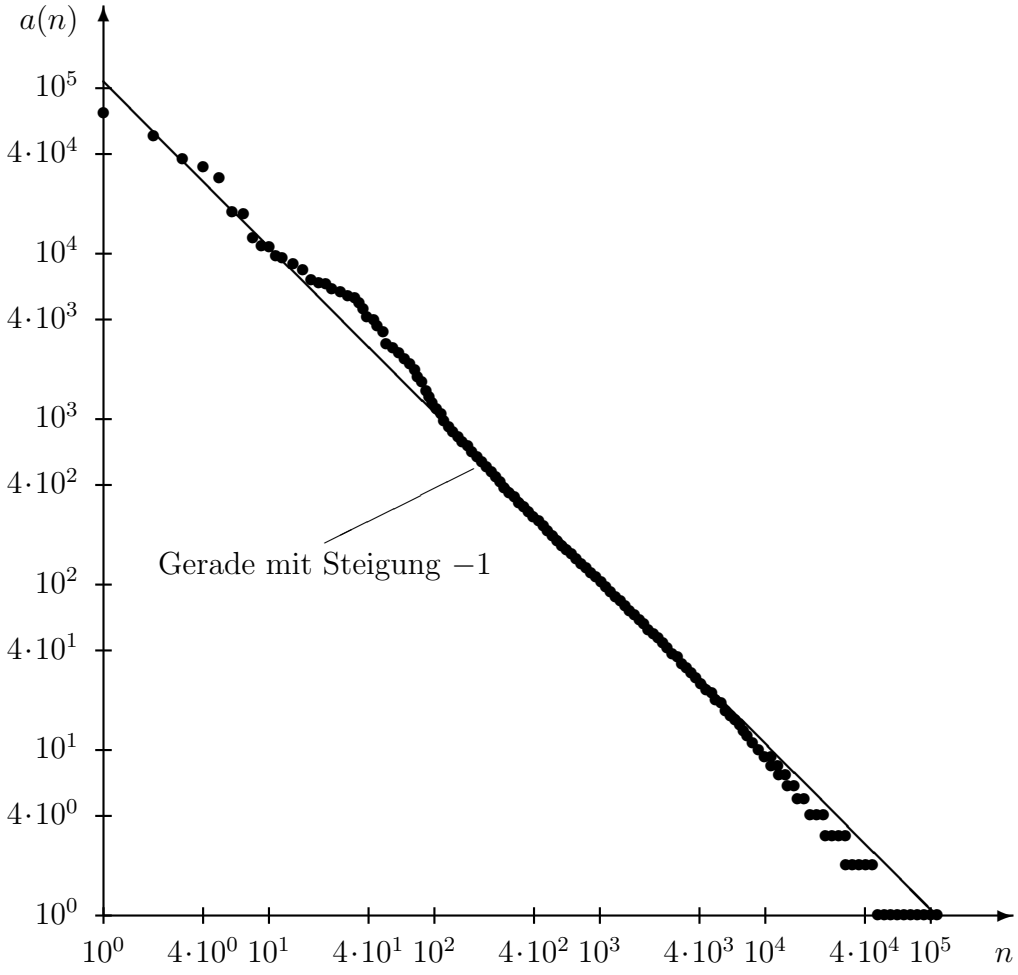
$$H_k = -\sum_{m=1}^{M}\sum_{n=1}^{N} p(m_{k-1}, n)\ \mathrm{ld}\, p(n|m_{k-1}) \qquad \text{(in bit/Symbol)}.$$



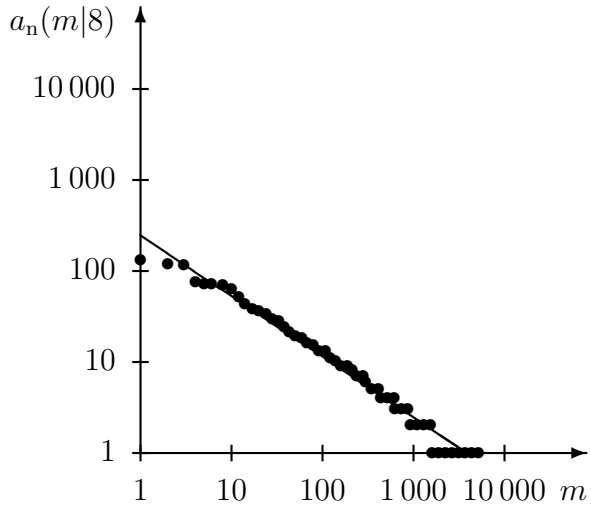
3.2 Das Zipf'sche Gesetz:

Ordnungszahl n	Wort	Häufigkeit $a(n)$
1	,	70 610
2	.	50 925
3	DER	37 229
4	DIE	33 013
5	UND	28 341
6	''	17 746
7	IN	17 206
8	DEN	12 335
9	VON	11 034
10	ZU	10 856

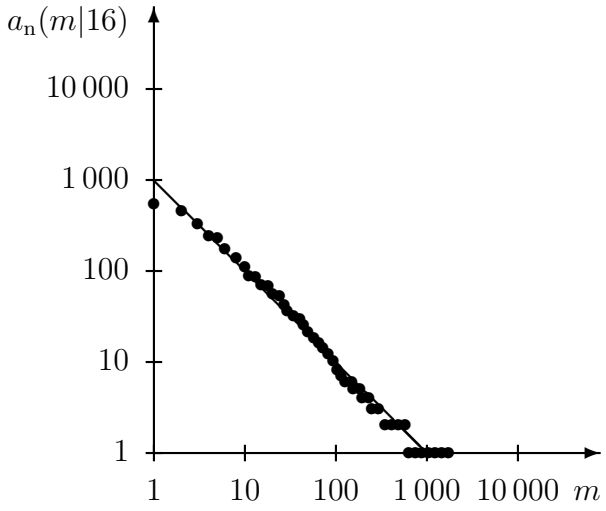
Ordnungszahl n	Wort	Häufigkeit $a(n)$
11	DES	9 608
12	IST	9 303
13	DAS	9 055
14	MIT	8 592
15	SICH	8 436
16	NICHT	7 906
17	AUF	7 678
18	*TK	6 923
19	DEM	6 610
20	EINE	6 597



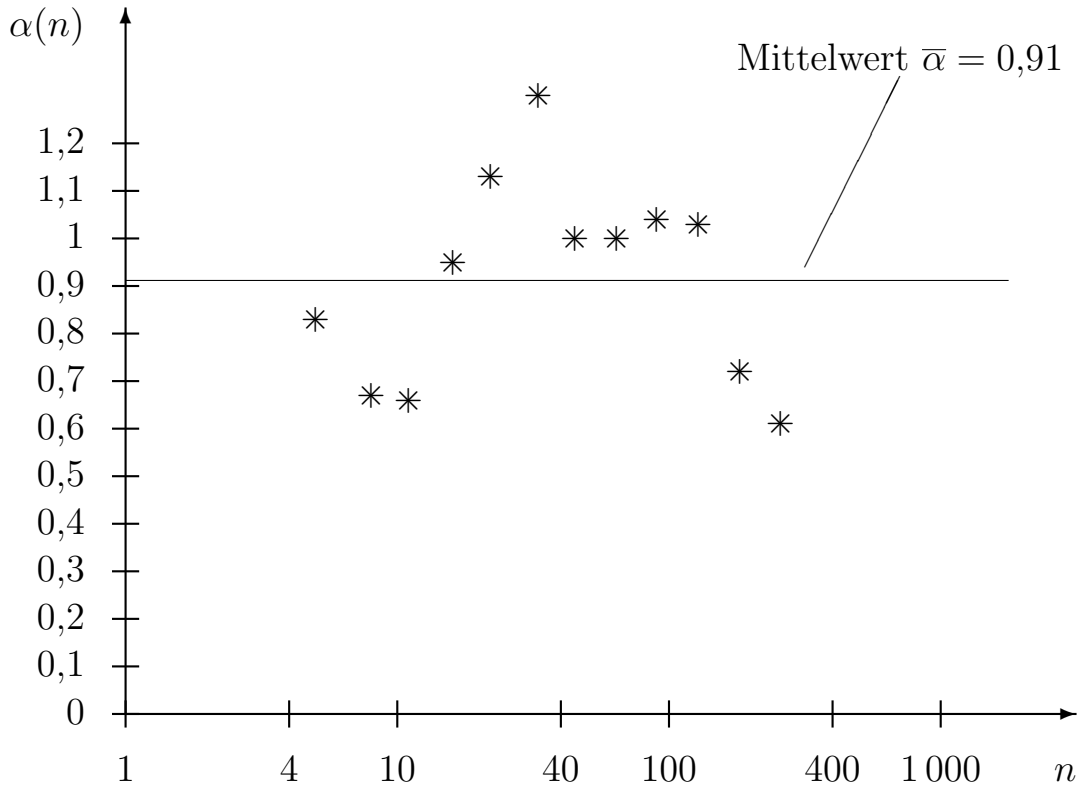
3.3 Die Häufigkeitsverteilungen der Nachfolgerwörter:



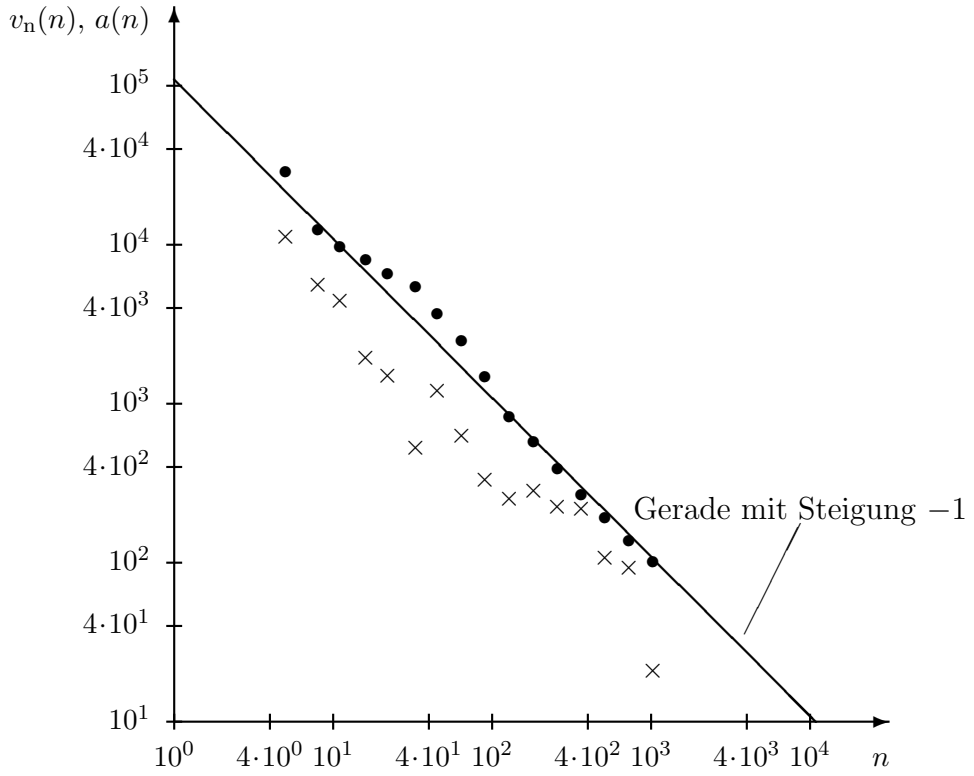
Die Häufigkeiten $a_n(m|8)$ der Nachfolger von „DEN“ ($n = 8$).

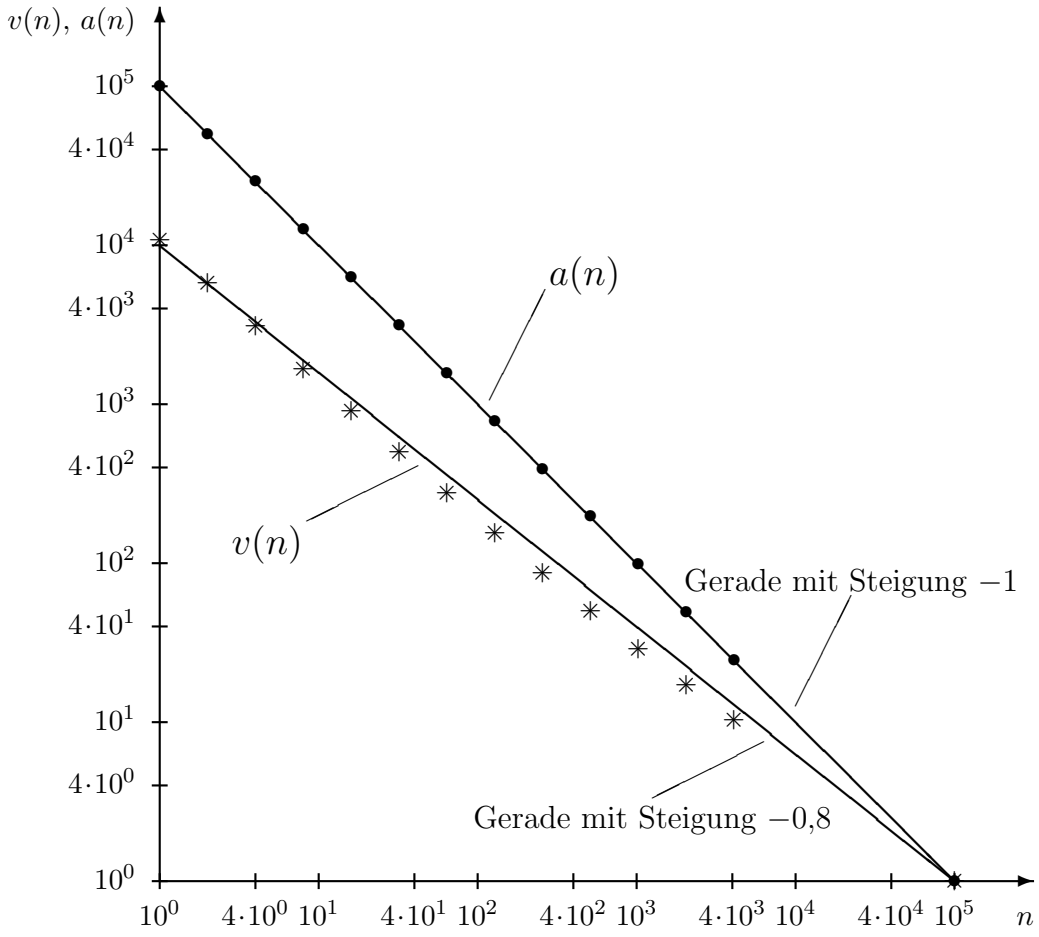


Die Häufigkeiten $a_n(m|16)$ der Nachfolger von „NICHT“ ($n = 16$).



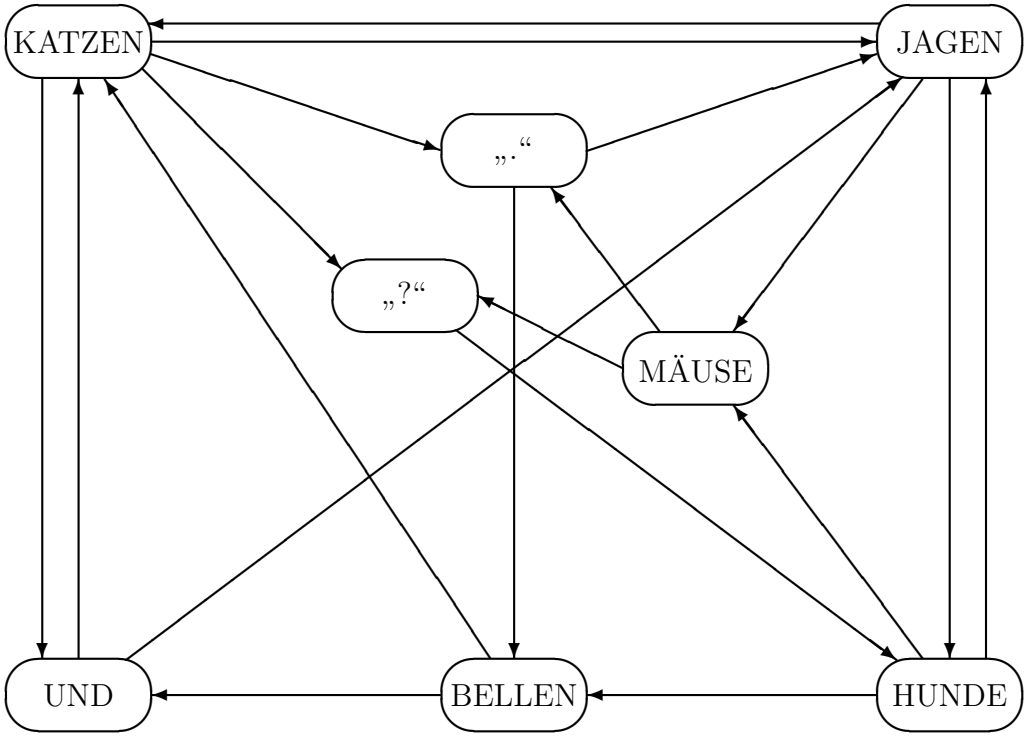
3.4 Der „Verzweigungsgrad“ von Wörtern:

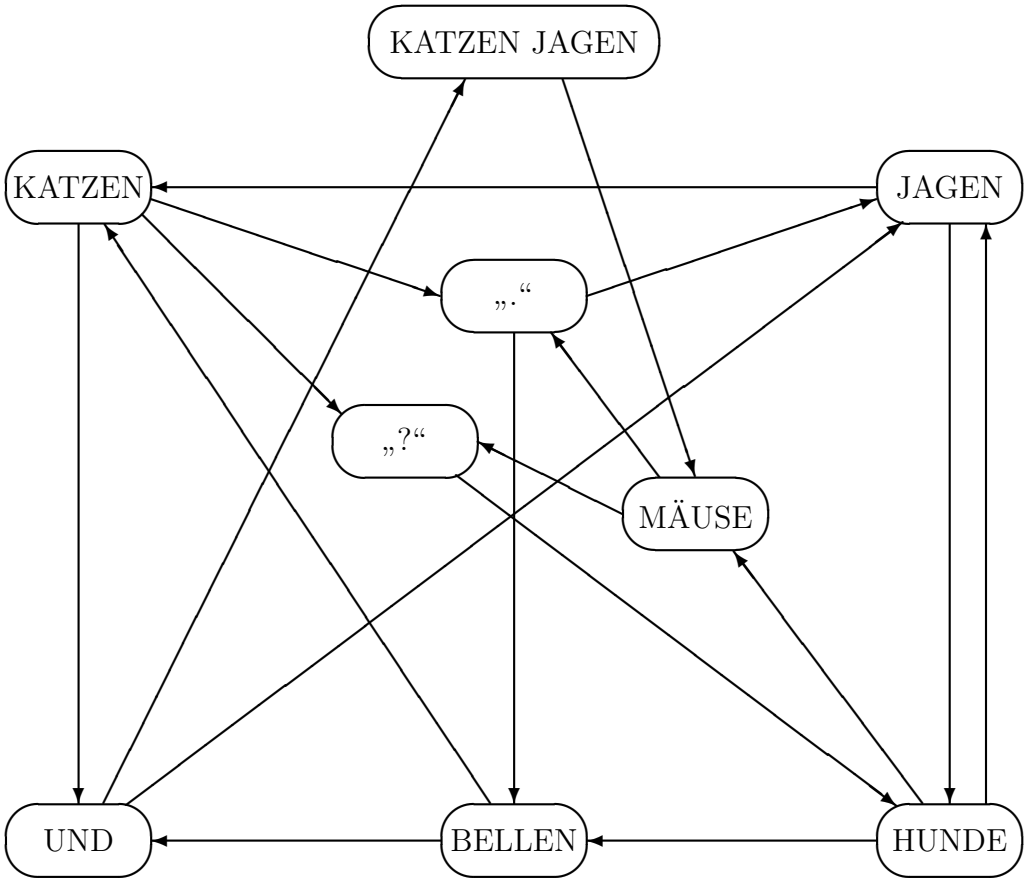




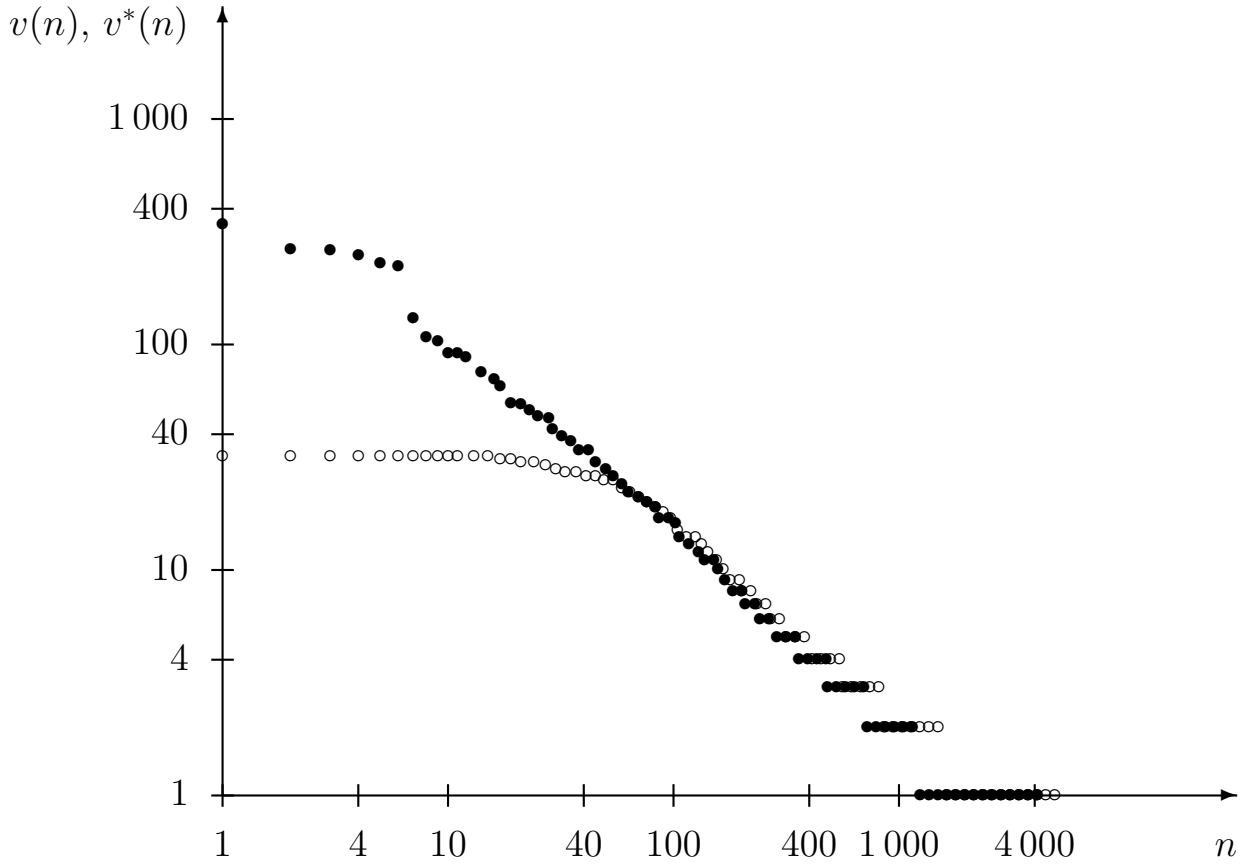
4 Die Ergebnisse der Simulationen:

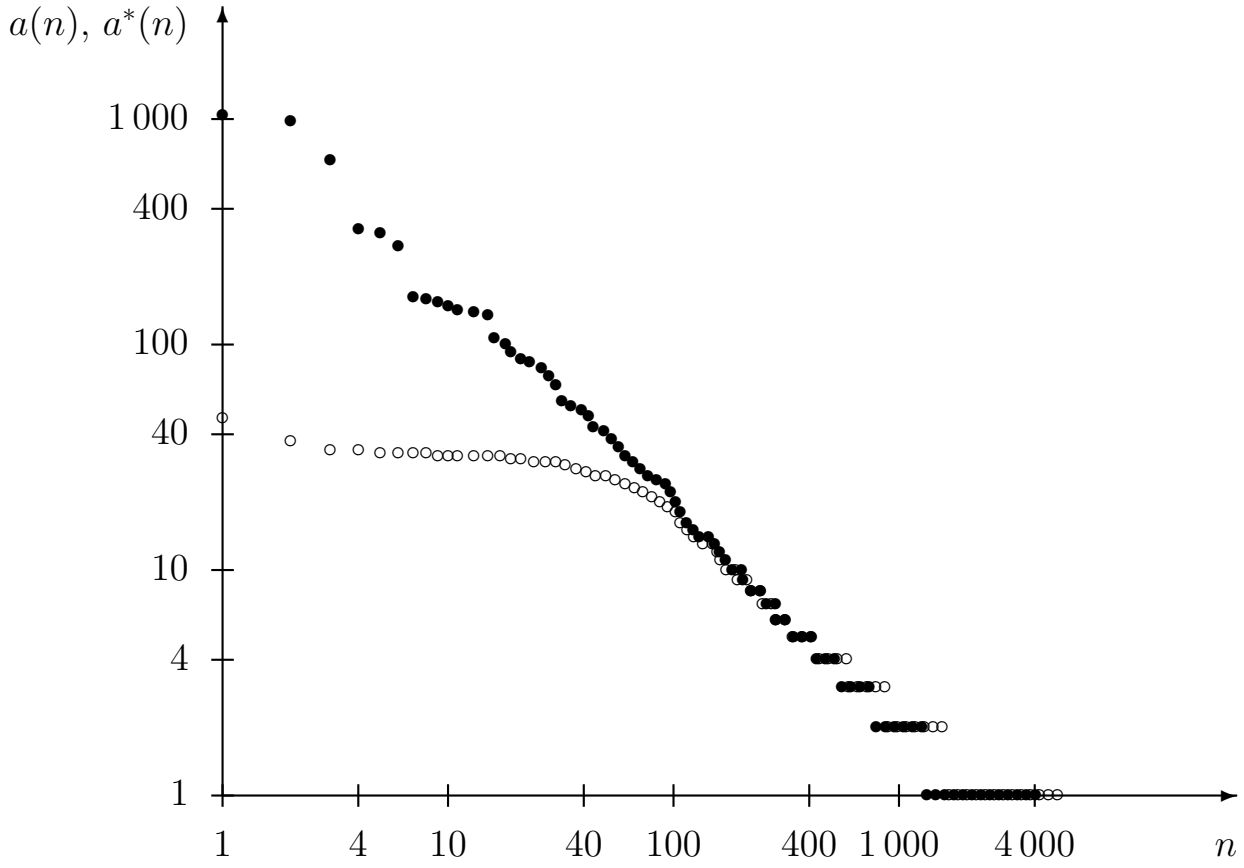
4.1 Das Wortzellen-Netz:





4.2 Verzweigungsgrad- und Häufigkeitsverteilungen:





4.3 Die Codierung des LIMAS-Korpus:

Das LIMAS-Korpus besteht aus:

- 500 Textstellen à 2 000 Wörtern
- $\approx$ 1 000 000 Wörtern
- $\approx$ 116 000 verschiedenen Wörtern

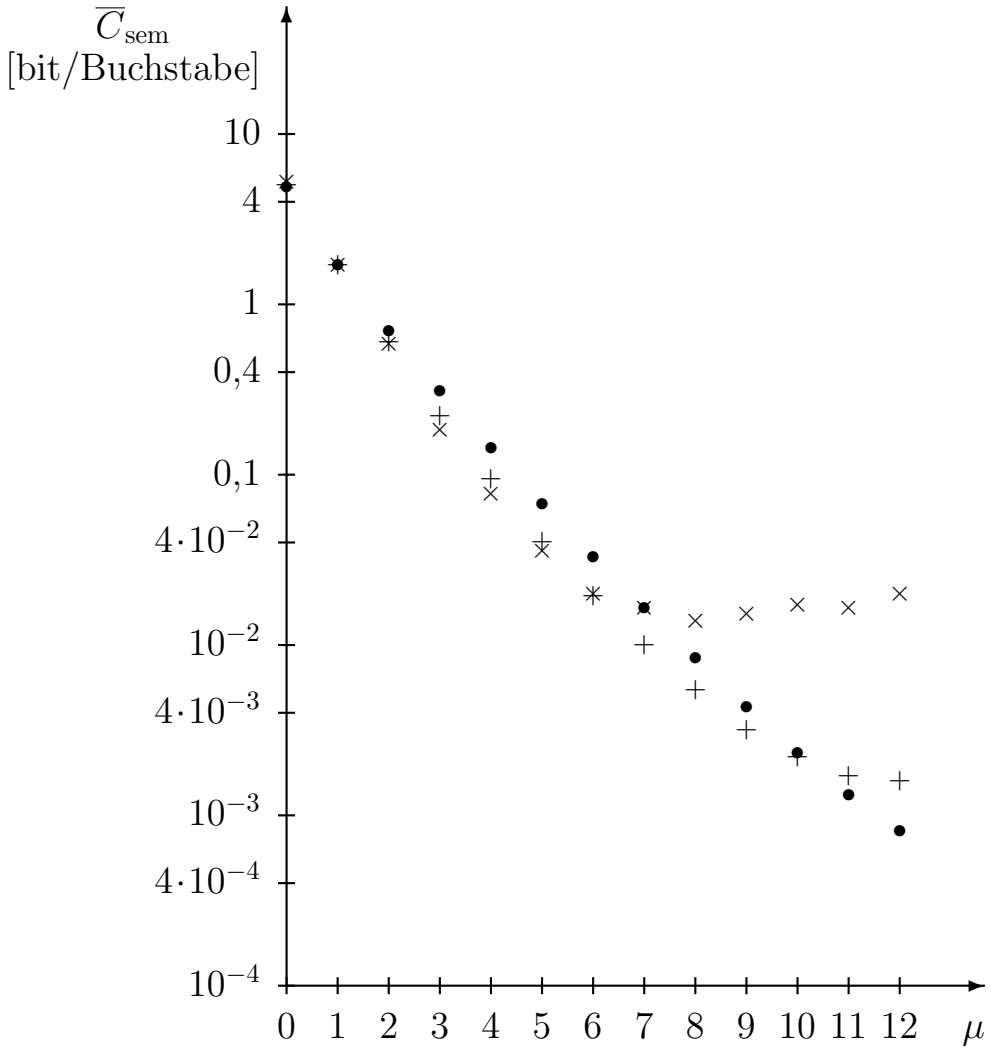
Die Codierung des gesamten LIMAS-Korpus unter Verwendung eines semantischen Speichers mit 17 Ebenen:

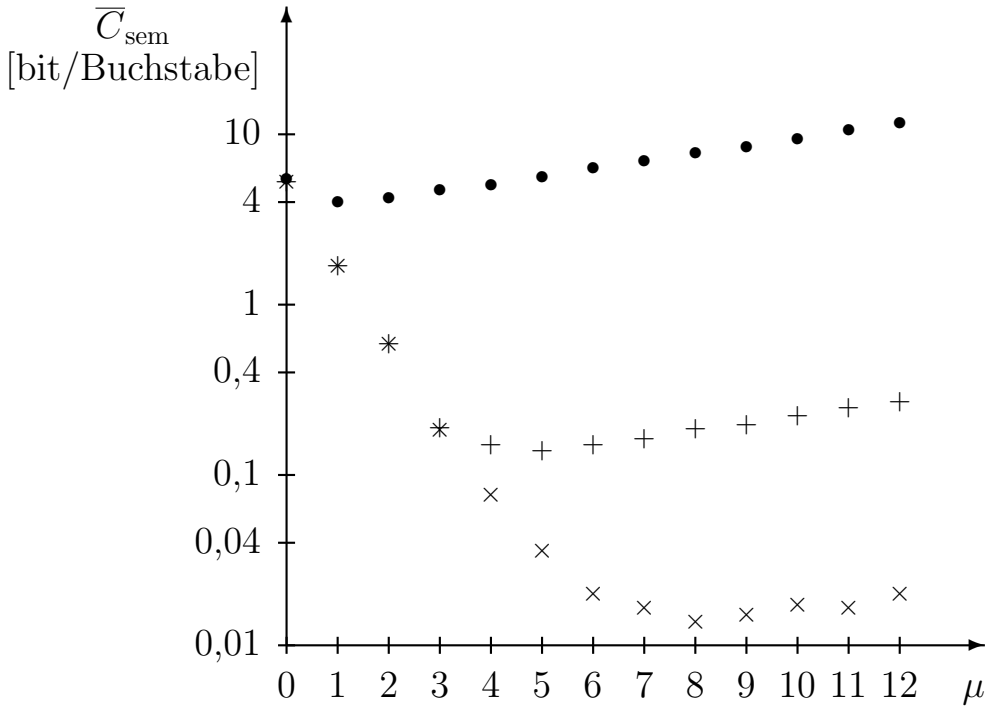
Textlänge (Klartext)	:	8 088 120	ASCII-Zeichen
Textlänge (Klartext)	:	1 288 163	Wörter
Codelänge	:	145	bit
Mittlere Codelänge	:	17,9	μbit/Buchstabe
Mittlere Codelänge	:	112,6	μbit/Wort
Mittlere Codelänge	:	55 780	Buchstaben/bit
Mittlere Codelänge	:	8 884	Wörter/bit
Mittlere Codelänge	:	$\approx$ 20	Seiten/bit

Der Code des LIMAS-Korpus:

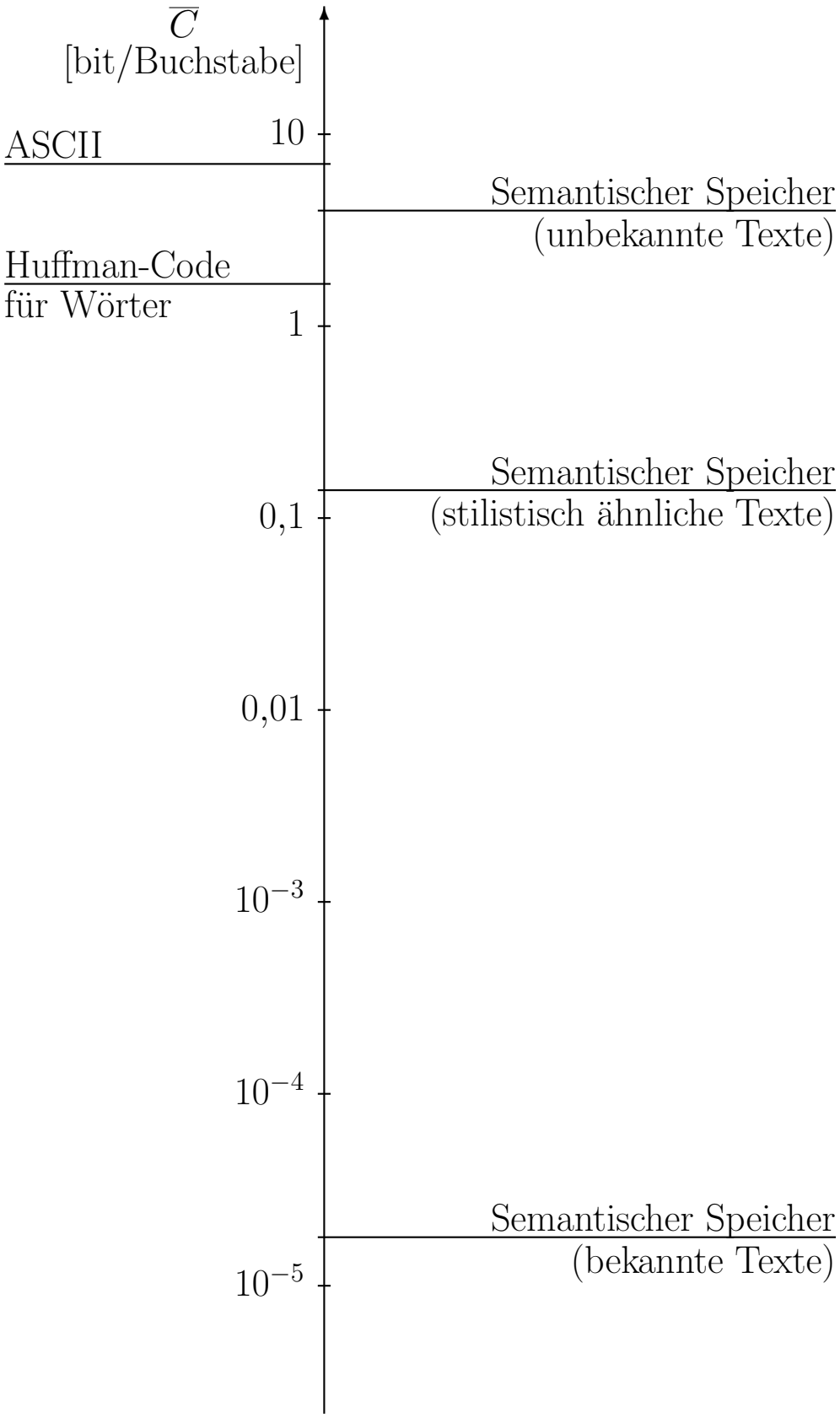
00111011101001110101000101010
00001100010001011000100000101
00001001000000100001100000000
10110000010101101000111010010
00000000100110010111100001000

4.4 Die Codierung unbekannter Texte:

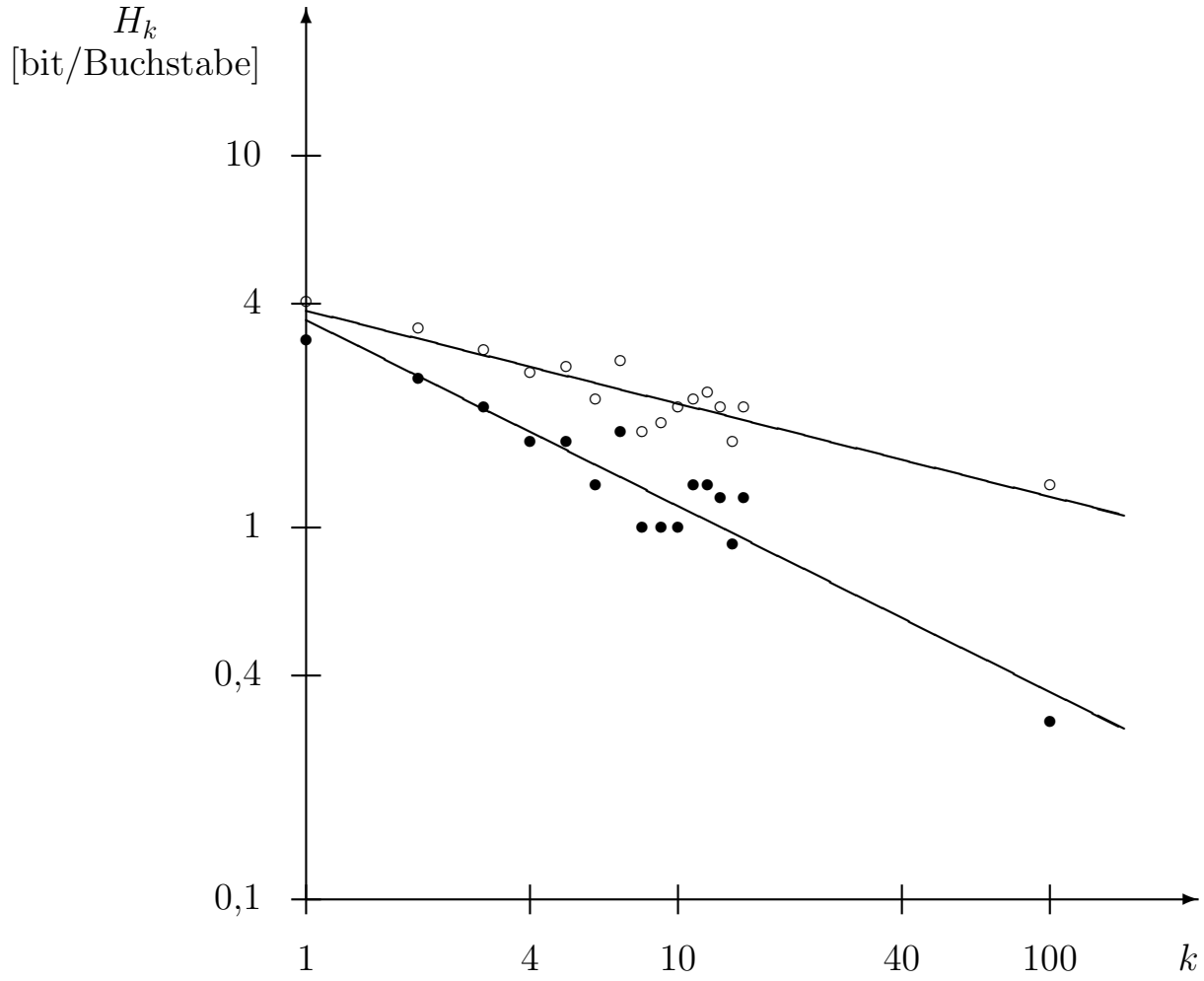




5 Zusammenfassung:



Die logarithmische Darstellung der Shannon'schen Meßwerte:



Zitat Shannon:

... From this analysis it appears that, in ordinary literary English, the long range statistical effects (up to 100 letters) reduce the entropy to something of the order of one bit per letter, with a corresponding redundancy of roughly 75 %. The redundancy may be still higher when structure extending over paragraphs, chapters, etc. is included. ...

Die kleinste erreichbare mittlere Codelänge:

Abschätzung der Maximalanzahl aller Texte:

10 Milliarden Menschen schreiben 1 Million Jahre lang jeden Tag jeweils 100 Schriftstücke:

$$N_{\text{Texte}} = 10^{10} \cdot 10^6 \cdot 365 \cdot 100 = 3{,}65 \cdot 10^{20} < 2^{70}.$$

Demnach reichen 70 bit zum Adressieren aller bisher verfaßten Schriftstücke bei weitem aus:

$$C_{\min} = 70 \text{ bit}.$$

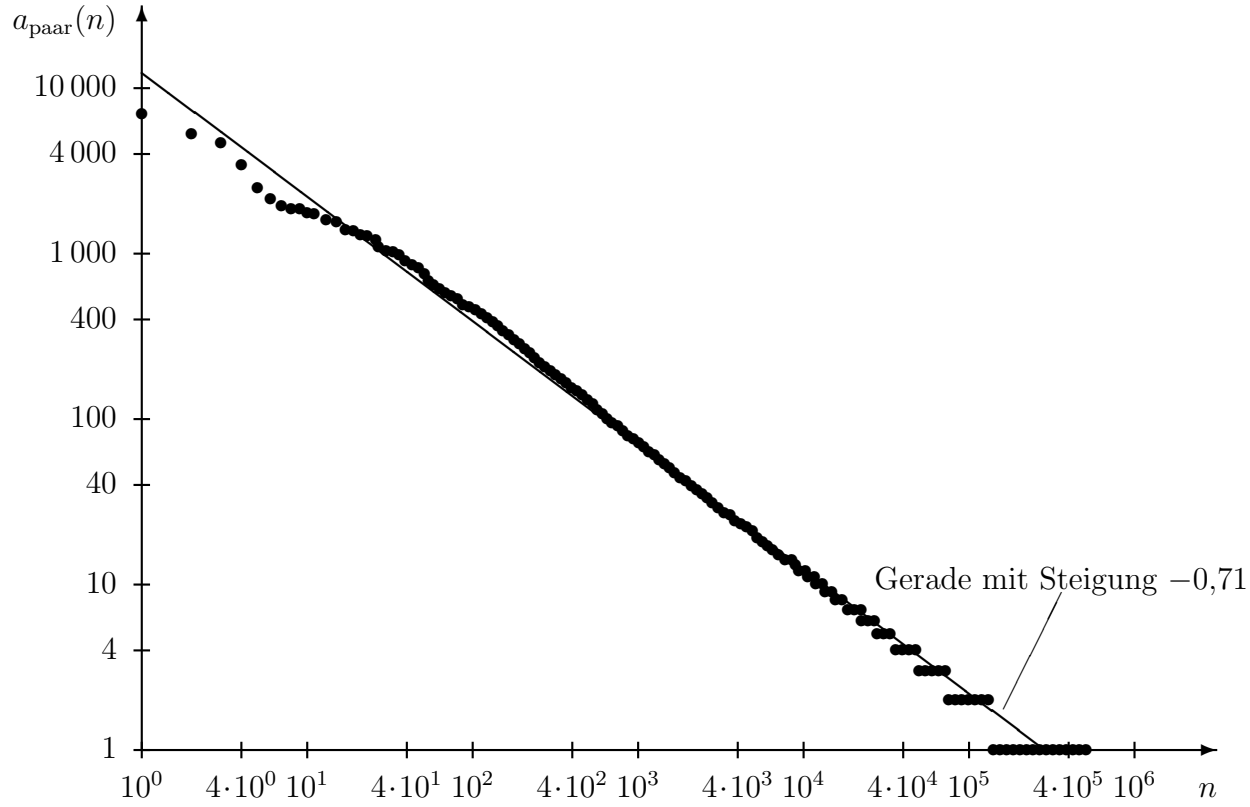
Längster sinnvoll zusammenhängender Text:

2 000	Seiten
2 500	Zeichen/Seite
$5 \cdot 10^6$	Zeichen

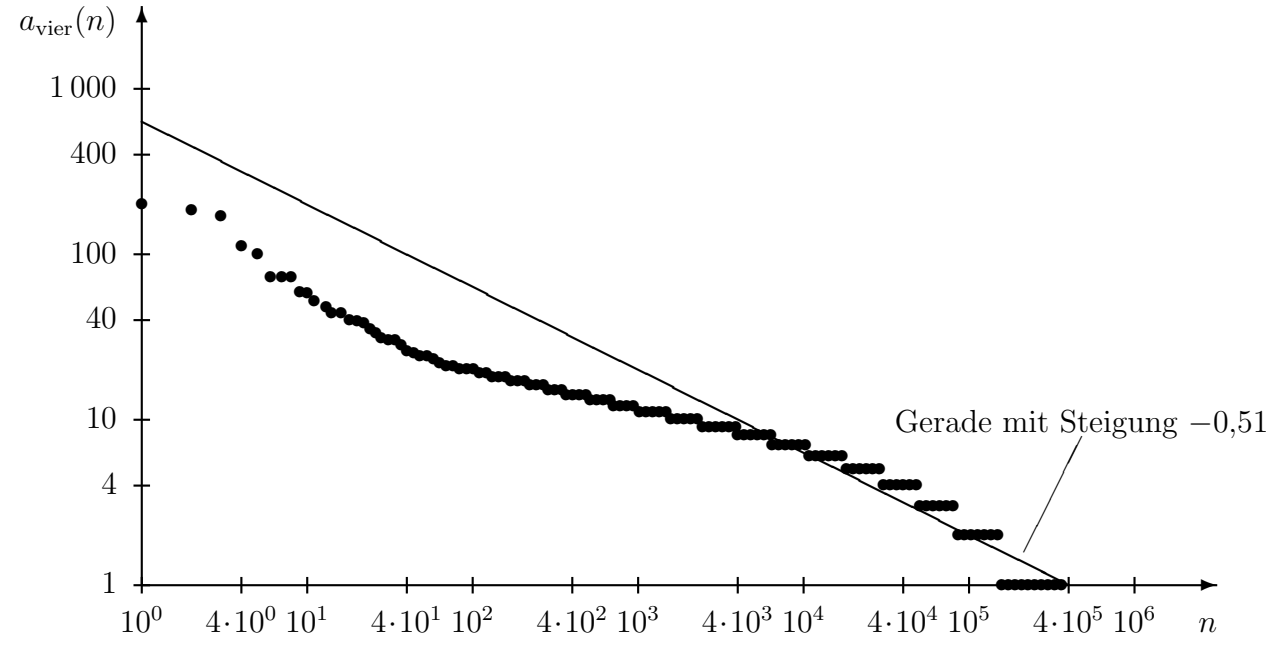
Abschätzung der minimalen mittleren Codelänge:

$$\overline{C}_{\min} = \frac{70}{5} \cdot 10^{-6} \text{ bit/Buchstabe} = 14 \ \mu\text{bit/Buchstabe}.$$

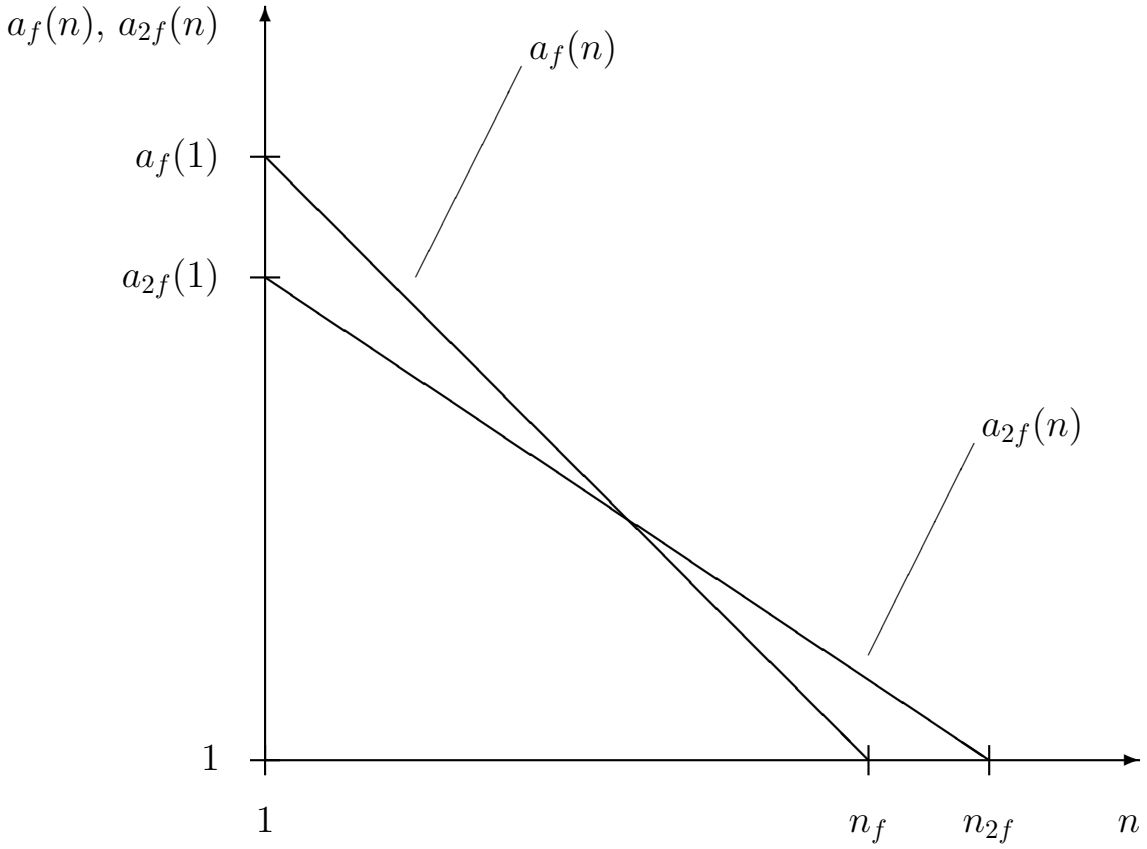
Die Häufigkeitsverteilung von Wortpaaren:

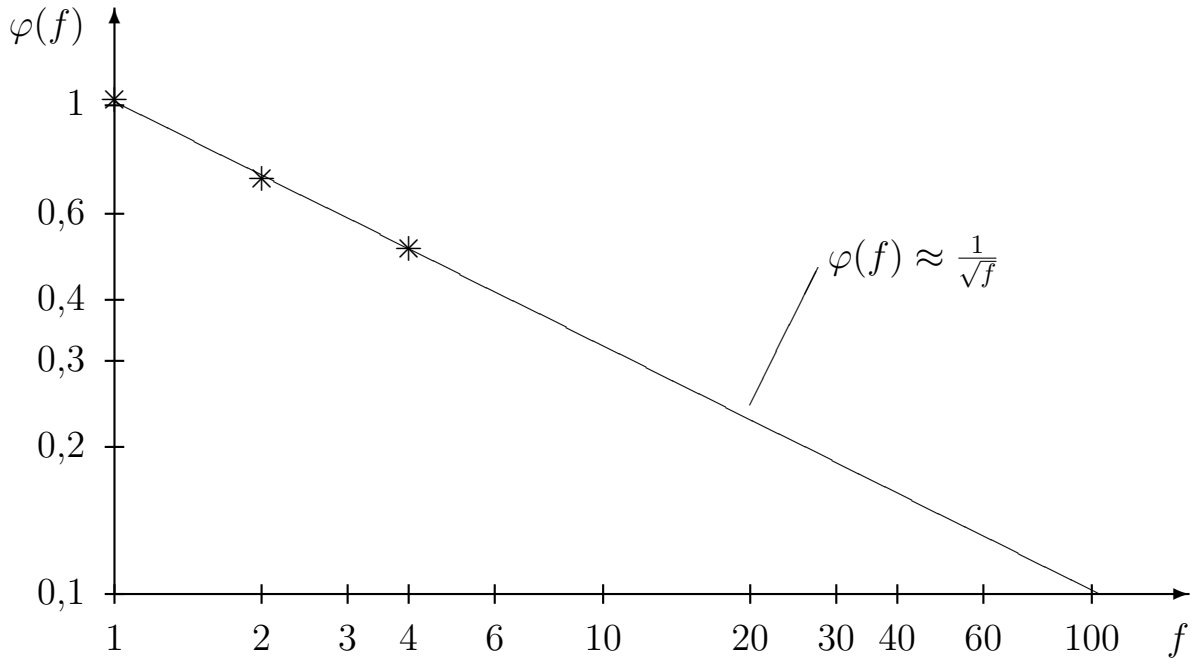


Die Häufigkeitsverteilung von Folgen aus vier Wörtern:



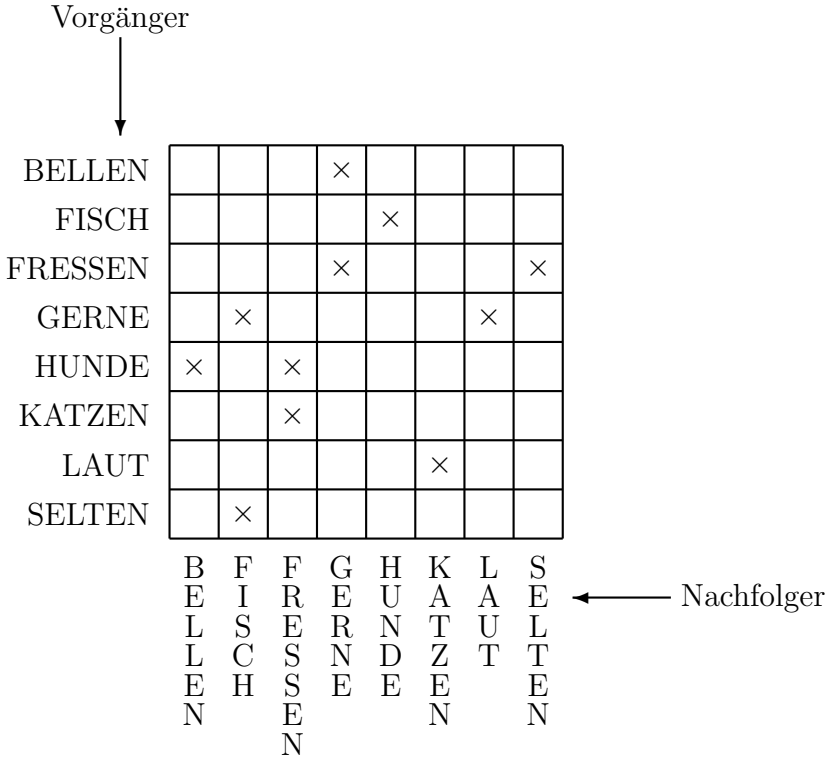
Die Häufigkeitsverteilungen sehr langer Wortfolgen:

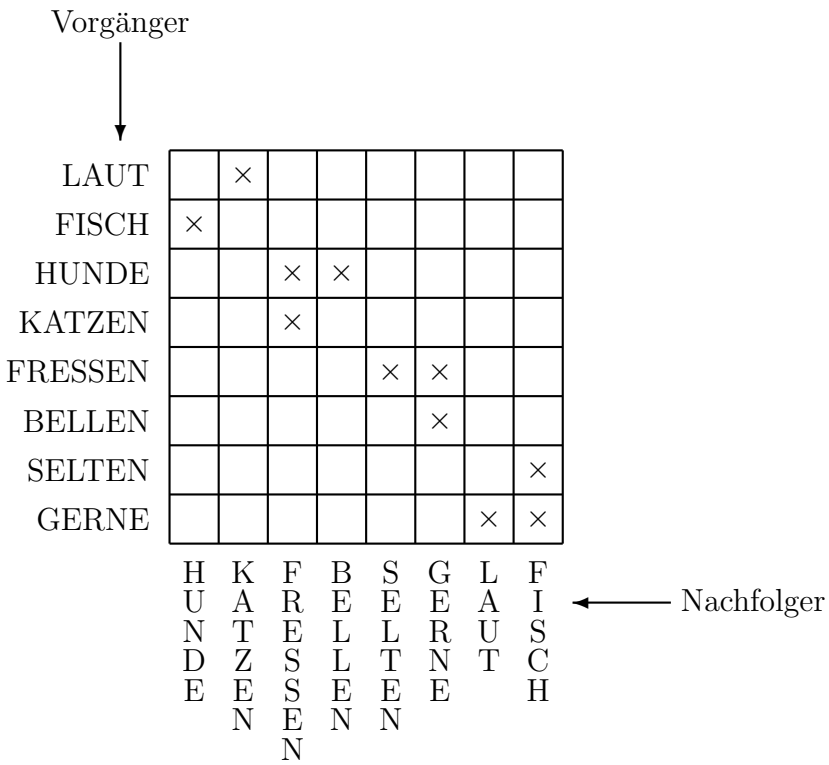




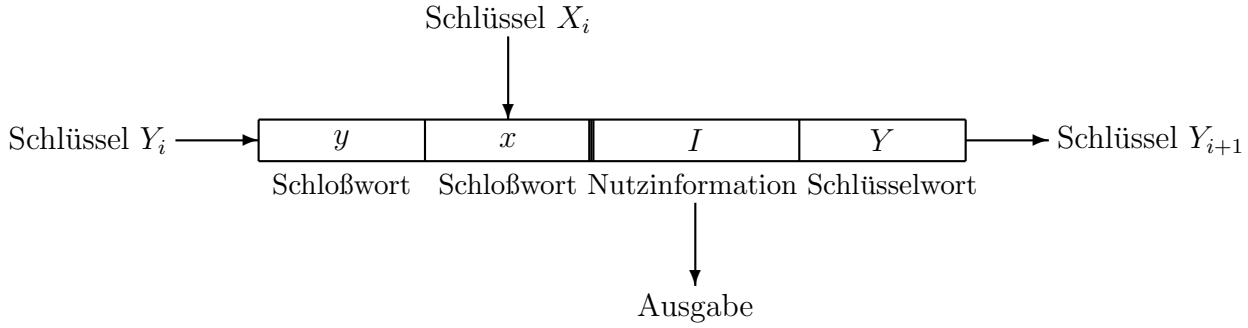
Die Übergangsmatrix für Wörter:

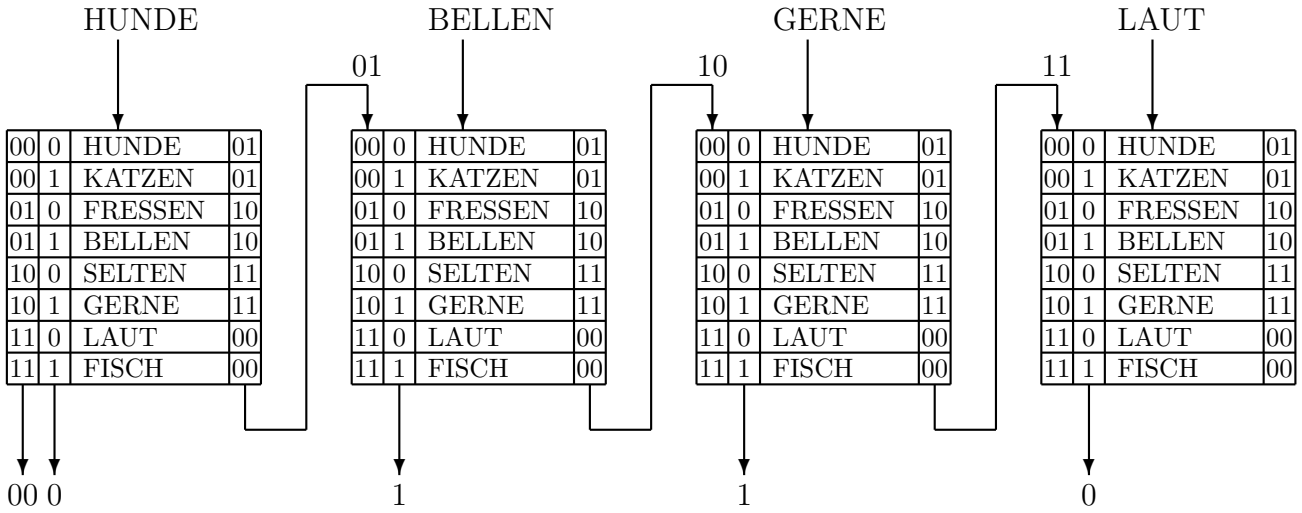
HUNDE BELLEN GERNE LAUT
KATZEN FRESSEN GERNE FISCH
HUNDE FRESSEN SELTEN FISCH

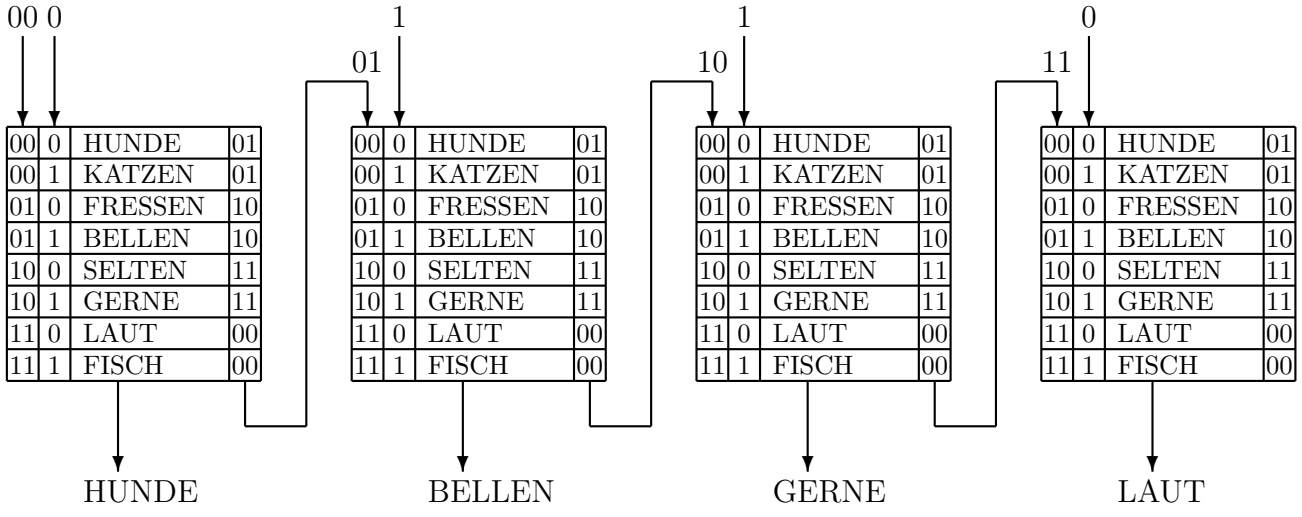




Das assoziative Feld nach W. Hilberg:



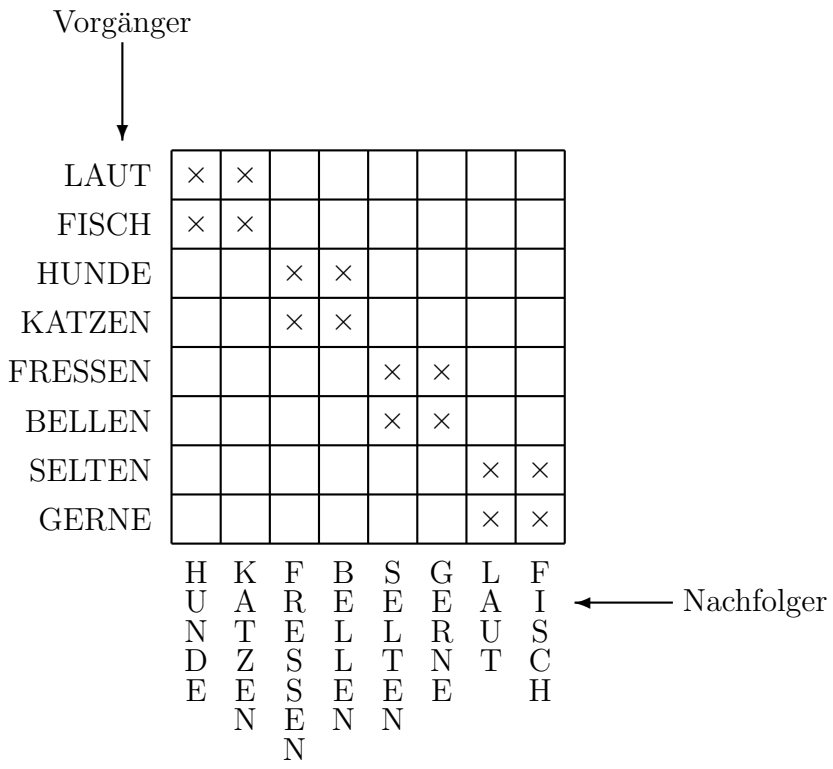




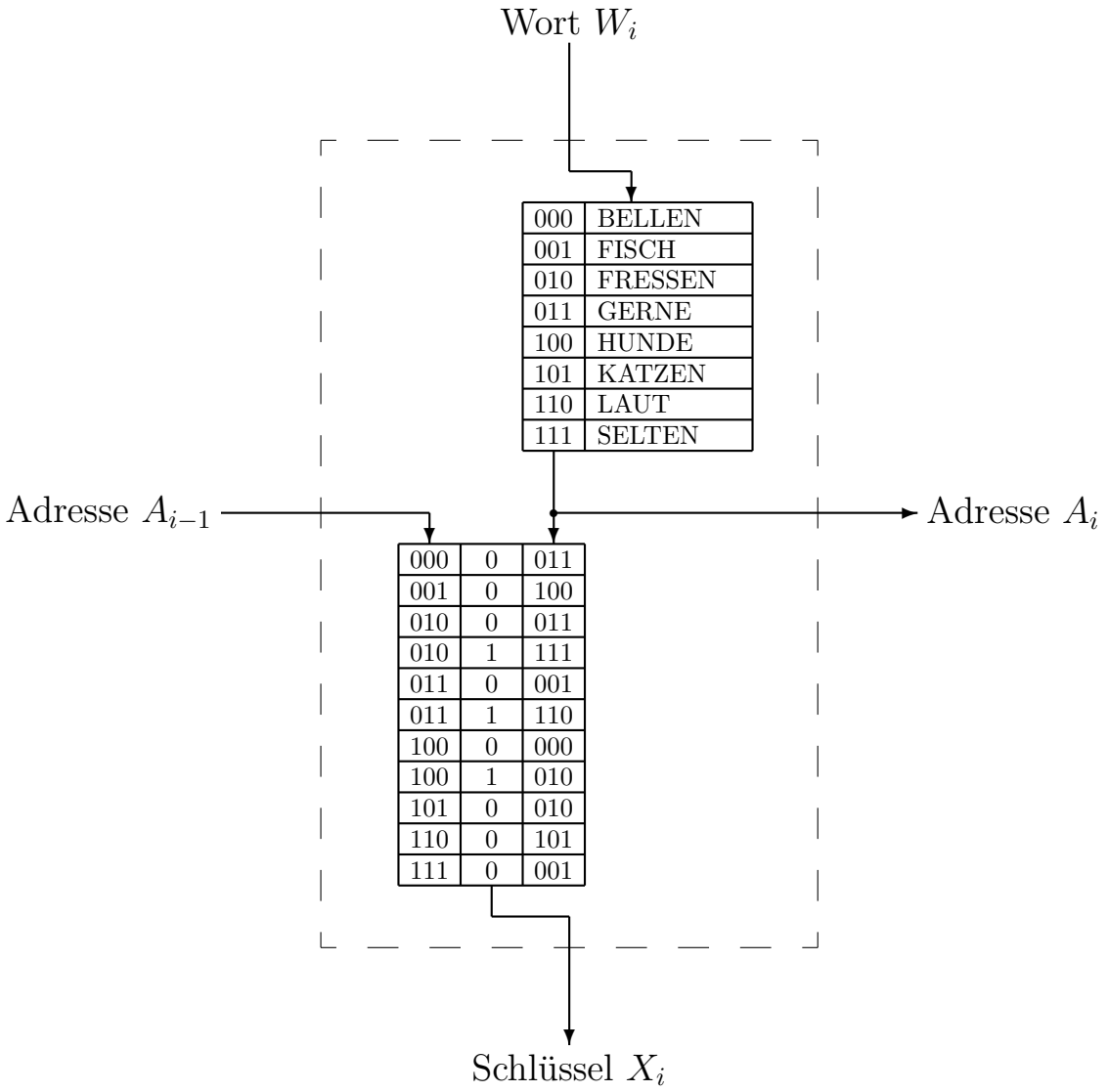
00	0	HUNDE	01
00	1	KATZEN	01
01	0	FRESSEN	10
01	1	BELLEN	10
10	0	SELTEN	11
10	1	GERNE	11
11	0	LAUT	00
11	1	FISCH	00

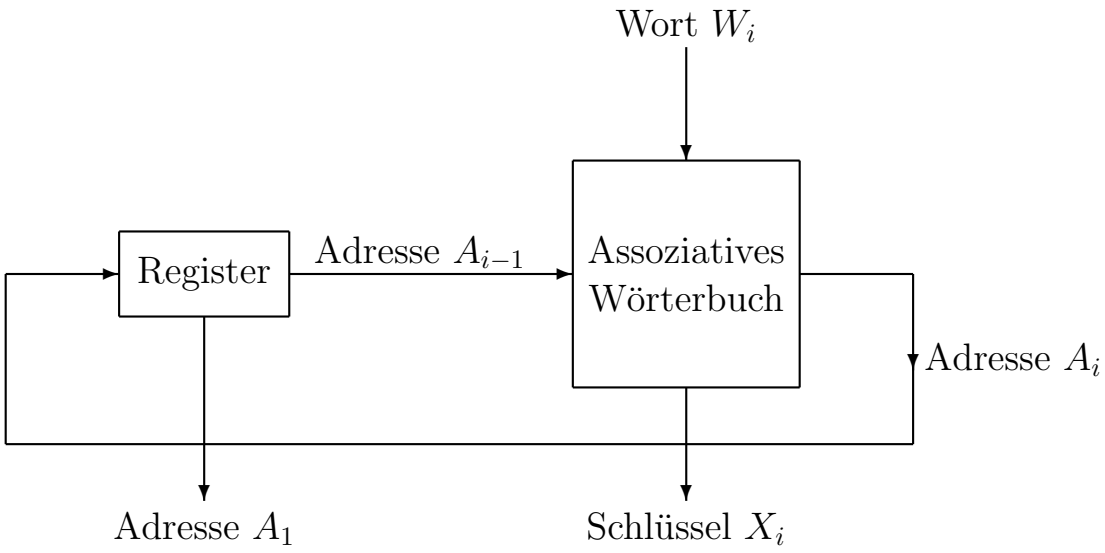
Alle decodierbaren Sätze für $Y_1 = 00$:

X_1	X_2	X_3	X_4		Ausgegebener Satz
0	0	0	0	$\hat{=}$	HUNDE FRESSEN SELTEN LAUT
0	0	0	1	$\hat{=}$	HUNDE FRESSEN SELTEN FISCH
0	0	1	0	$\hat{=}$	HUNDE FRESSEN GERNE LAUT
0	0	1	1	$\hat{=}$	HUNDE FRESSEN GERNE FISCH
0	1	0	0	$\hat{=}$	HUNDE BELLEN SELTEN LAUT
0	1	0	1	$\hat{=}$	HUNDE BELLEN SELTEN FISCH
0	1	1	0	$\hat{=}$	HUNDE BELLEN GERNE LAUT
0	1	1	1	$\hat{=}$	HUNDE BELLEN GERNE FISCH
1	0	0	0	$\hat{=}$	KATZEN FRESSEN SELTEN LAUT
1	0	0	1	$\hat{=}$	KATZEN FRESSEN SELTEN FISCH
1	0	1	0	$\hat{=}$	KATZEN FRESSEN GERNE LAUT
1	0	1	1	$\hat{=}$	KATZEN FRESSEN GERNE FISCH
1	1	0	0	$\hat{=}$	KATZEN BELLEN SELTEN LAUT
1	1	0	1	$\hat{=}$	KATZEN BELLEN SELTEN FISCH
1	1	1	0	$\hat{=}$	KATZEN BELLEN GERNE LAUT
1	1	1	1	$\hat{=}$	KATZEN BELLEN GERNE FISCH

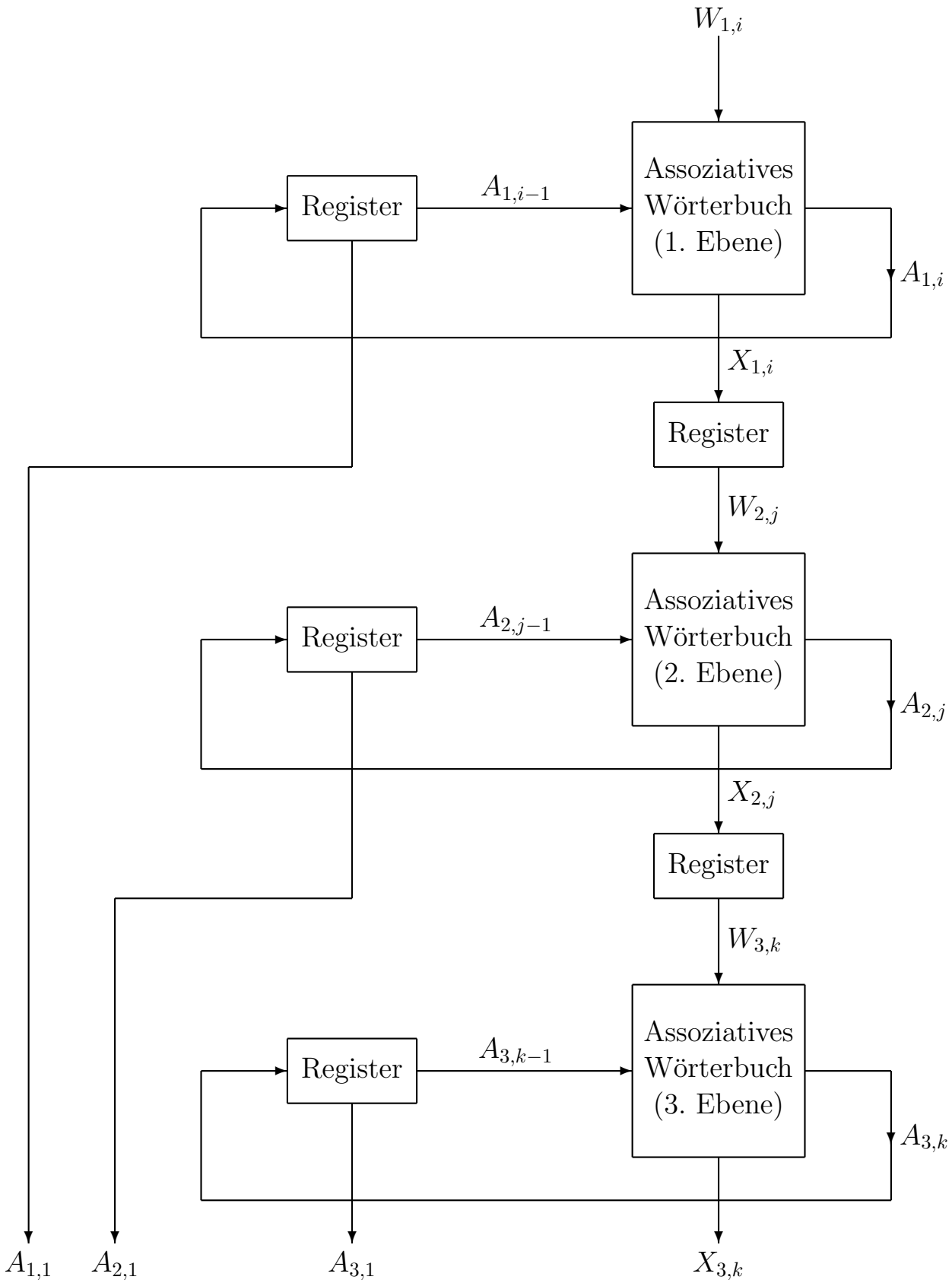


Das rückgekoppelte assoziative Wörterbuch:

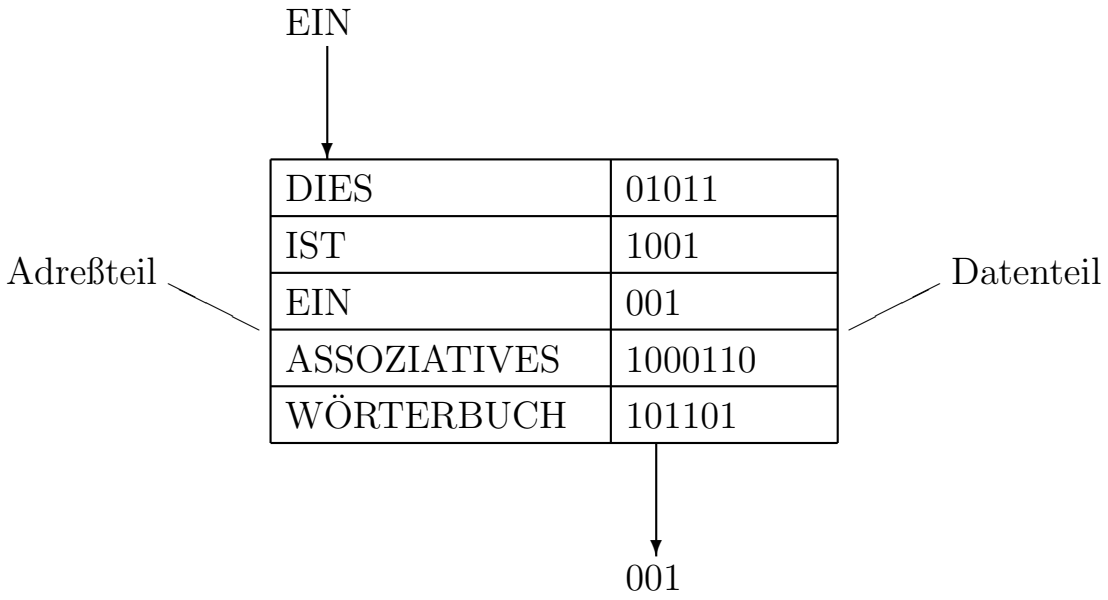




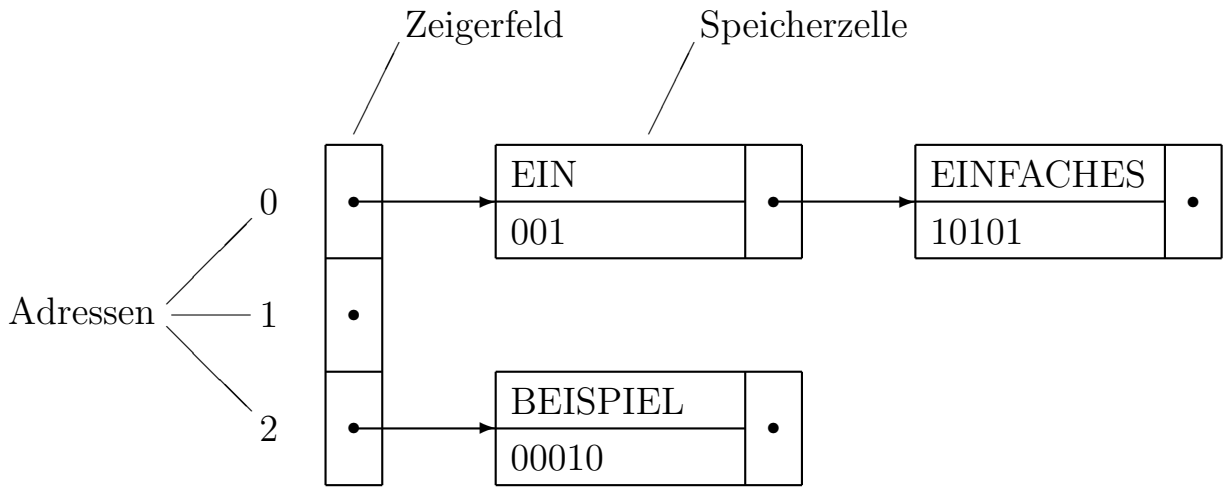
Der prinzipielle Aufbau des realisierten semantischen Speichers:



Das assoziative Wörterbuch:



Die Realisierung mit Hilfe der Hash-Speicherung:

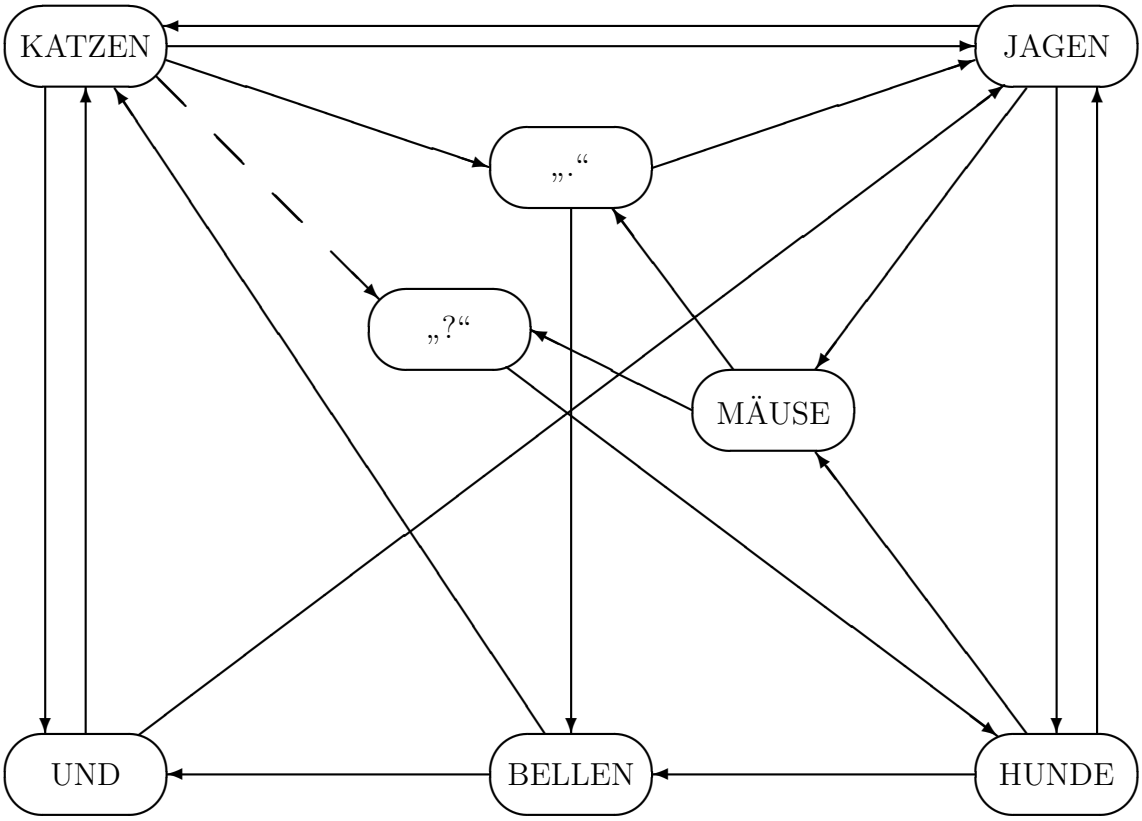


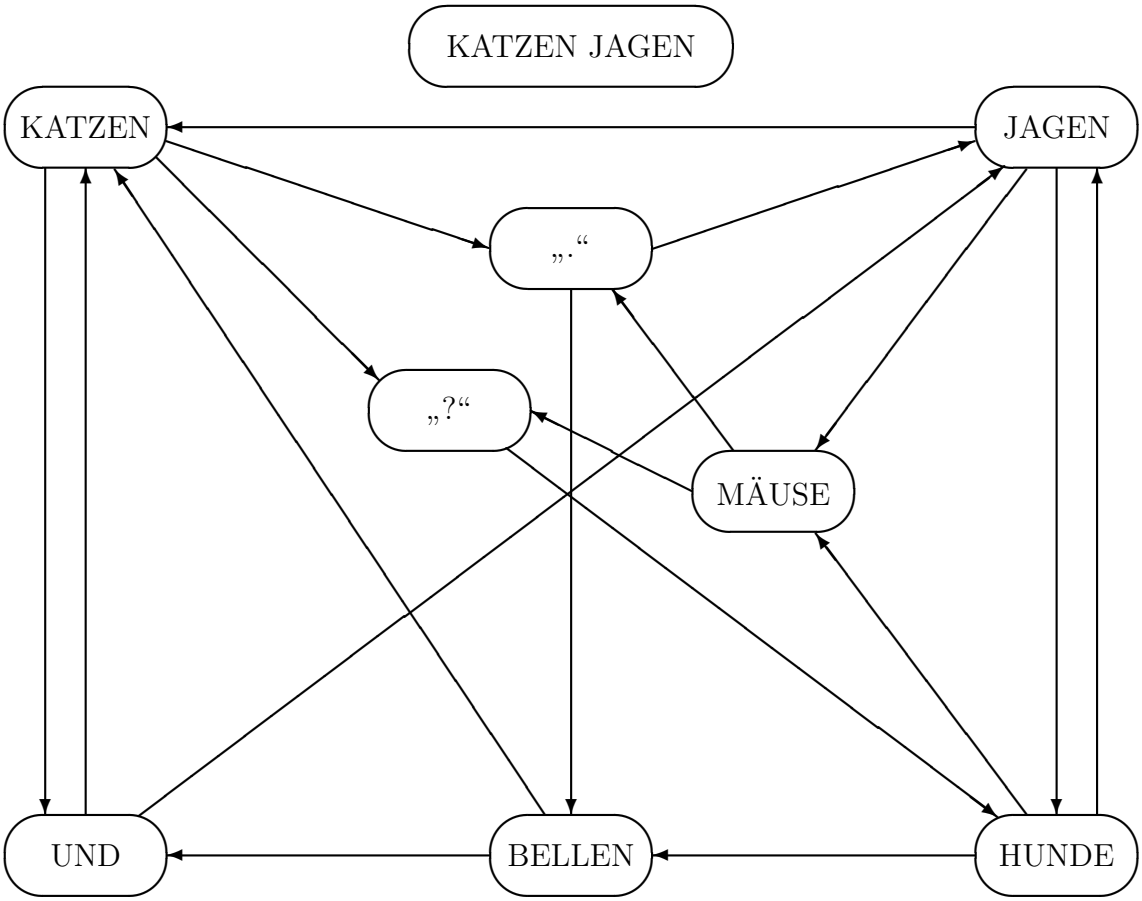
Hash-Funktion: *Adresse* = *Wortlänge* mod 3.

$$A(W) = \left[\sum_{i=1}^{4} \left(7^{(i-1)} \cdot W[i] \right) + \sum_{i=\ell-3}^{\ell} \left(7^{(i-\ell+7)} \cdot W[i] \right) \right] \bmod P \qquad \text{für } \ell > 4.$$

$$A(W) = \begin{cases} \left[\sum_{i=1}^{\ell} \left(7^{(6\cdot(i-1)/(\ell-1))} \cdot W[i] \right) \right] \bmod P & \text{für } 1 < \ell \leq 4; \\ W[1] \bmod P & \text{für } \ell = 1. \end{cases}$$

Der Aufbau eines Wortzellen-Netzes:





Der Code des LIMAS-Korpus:

Ebene μ	Initialisierungsadresse $A_{\mu,1}$
1	001110111010011 10
2	10100010101000001 1
3	000100010110001 0
4	00001010000100 1
5	0000001000011
6	0000000010110
7	0000101011
8	0100011101
9	0010000
10	0000010
11	011001
12	01111
13	00001
14	00
15	0
16	—
17	—

Zufallstexte:

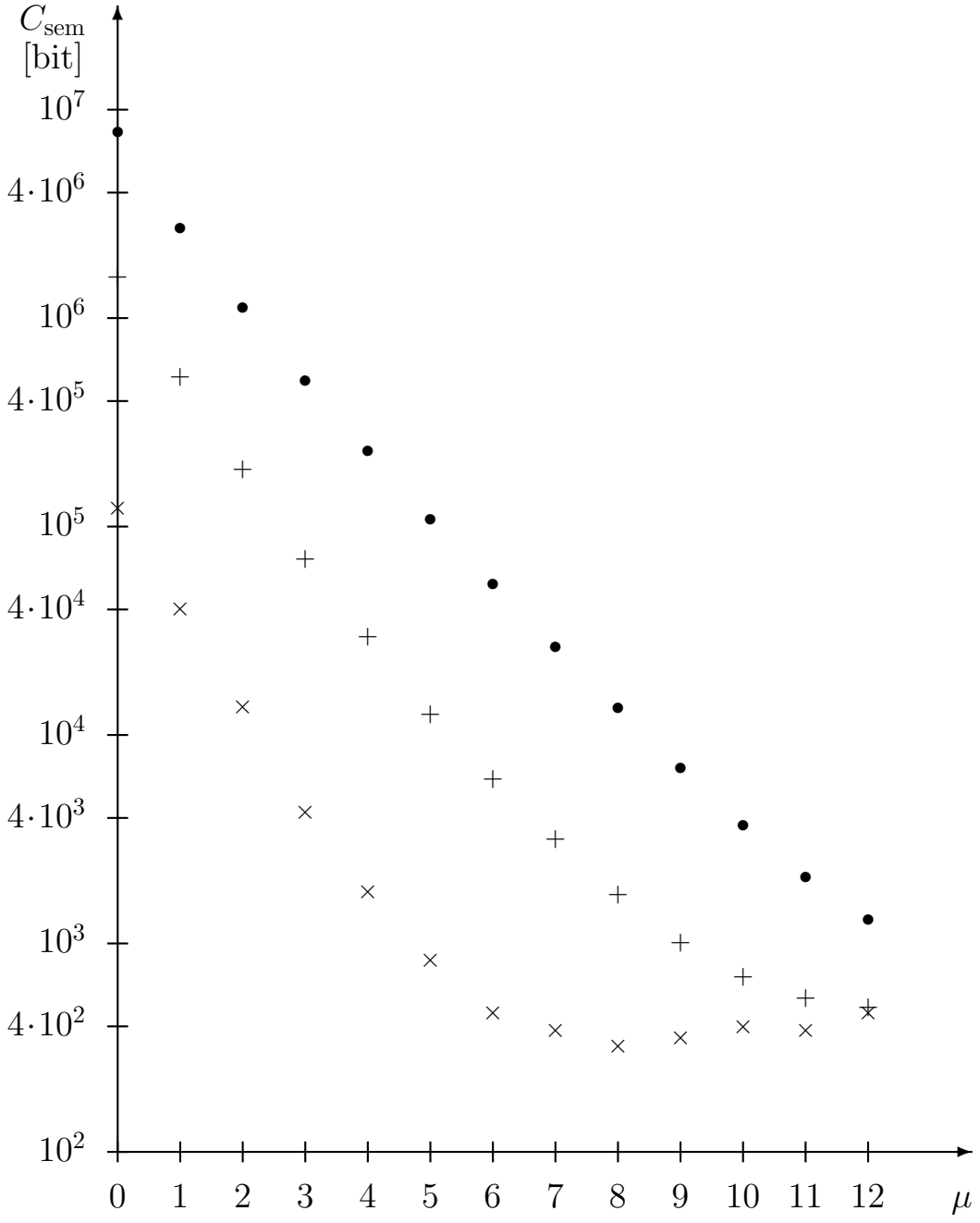
Einfaches Auswürfeln der Wörter:

Großmutter kugelrund wenn versuche Tier pfiffiger englisches tragen Japan deinen weitergeklettert lieben geschlüpft versuchen ich gepackt bewegte Stein Gleich Unsinn wachsam schlüpfte vergiß ein erhob würde verwechsle zum Zuhause diesem langsam letzten diese Absprache Bär nennt ...

Mit einem semantischen Speicher erzeugter Zufallstext:

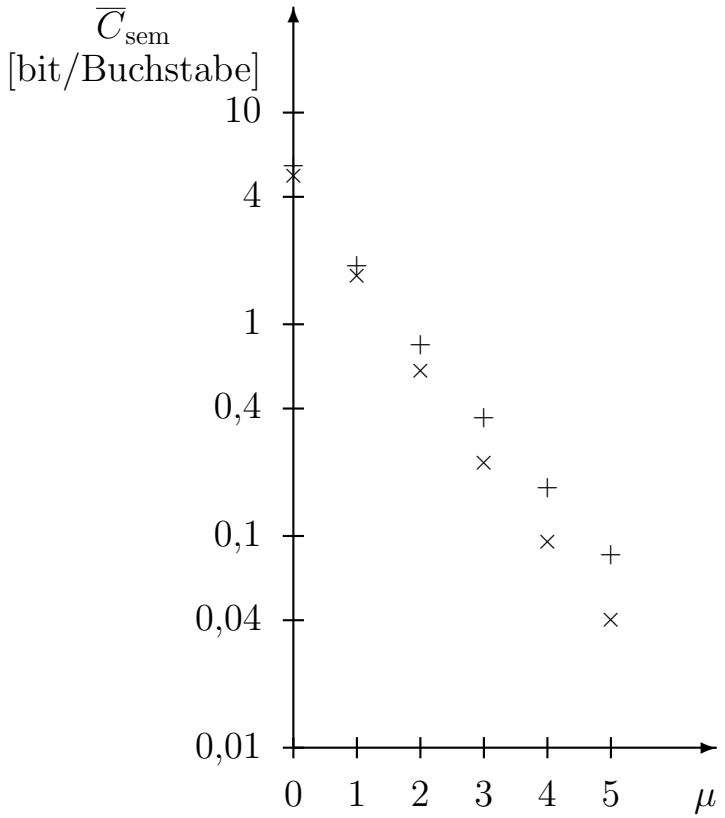
Ich trinke am Himmel. Und dann — sah er auch nicht in Afrika, der Spatz, so kennt er doch hier, wie bunt er ist und wie schön du singst“, sagte Sokrates. „Wie machst du das?“ „Ich entfliehe der Erde“, tirilierte die Lerche auf einem roten Luftballon hoch, ganz hoch, ganz gewöhnliche Blätter ...

Die absoluten Codelängen verschieden langer Texte:



Codierungsergebnisse mit semantischen Speichern verschiedener Größen:

Die Codierung bekannter Texte:



Die Codierung stilistisch ähnlicher Texte: