HOCHSCHULE
HANNOVER
UNIVERSITY OF
APPLIED SCIENCES
AND ARTS

NLP- Einführung in die Linguistik

Vitor Fontanella

Hochschule Hannover, Abteilung Information und Kommunikation

06.09.2022

Sprache und Schrift

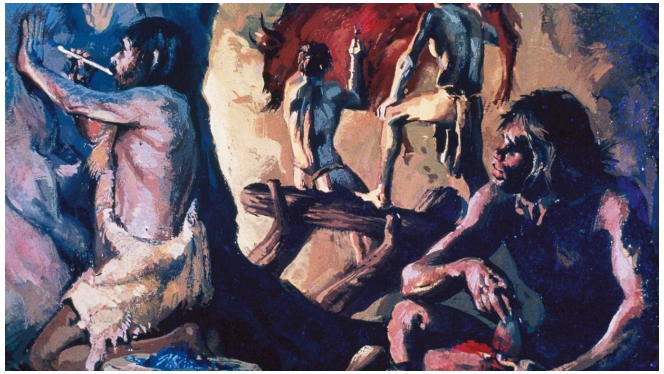
Gesprochen und geschrieben

- Historisch gesehen sprechen Menschen schon sehr lange; geschrieben wird erst seit einige Jahrtausenden.
- Gesprochene Sprache ist die eigentliche Sprache, die geschriebene nur die Wiedergabe.

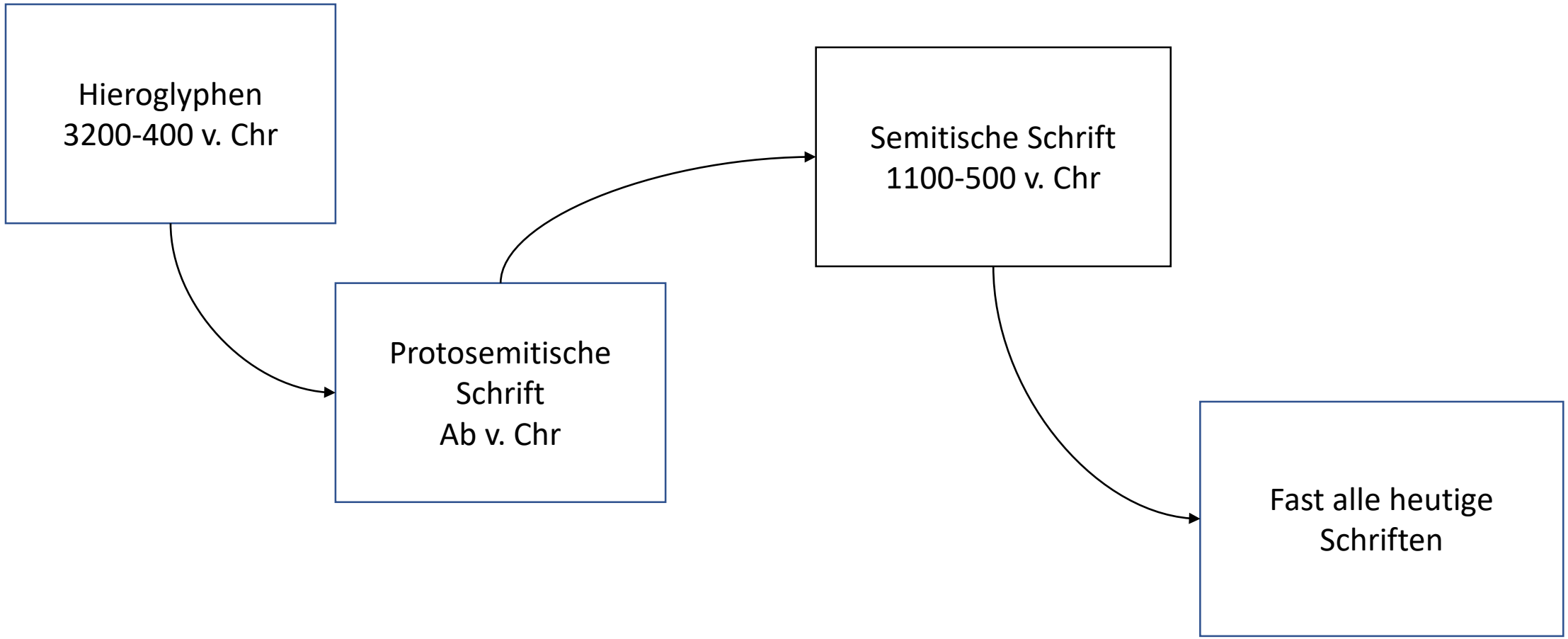
https://www.history.com/news/prehistoric-ages-timeline

Missverständnisse

- Das Sprechen richtet sich nicht nach dem Schreiben, sondern das Schreiben nach dem Sprechen!
- Sprache und Schrisystem sind weitestgehend unabhängig voneinander!



Geschichte


Hieroglyphen
3200-400 v. Chr
Protosemitische
Schrift
Ab v. Chr
Semitische Schrift
1100-500 v. Chr
Fast alle heutige
Schriften

Geschichte

Hieroglyphen

- Benutzt von ca. 3500 v. Chr. bis 400 n.Chr.
- Egyptisch: Nilo-Saharische Sprache; fern verwandt mit semitischen Sprachen

Symbole

Pictogramm Symbol, das bedeutet, was es darstellt

Ideogramm Symbol, das für einen Begriff steht, den es nicht unmittelbar darstellt

Phonogramme Symbol, das für einen Klang oder Klangfolge steht (Nur Konsonante!)

Deutzeichen Semantische Zeichen zur Auflösung der Mehrdeutigkeit der Phonogramme.

https://de.wikipedia.org/wiki/%C3%84gyptische_Hieroglyphen

Geschichte

Protosemitische Schrift

- Entstanden etwa 1700 v. Chr.
- Semitsche Sprache, Sklaven oder Gastarbeiter in Ägypten
- Entstanden auf dem Sinai oder in Ägypten (Luxor)

Buchstaben

- Symbole gehen auf Hierochlyphen zurück.
- Jedes Symbol steht für den Konsonanten des semitsichen Wortes für das Symbol!
- Älteste alphabetische Schrift.

https://de.wikipedia.org/wiki/Protosinaitische_Schrift

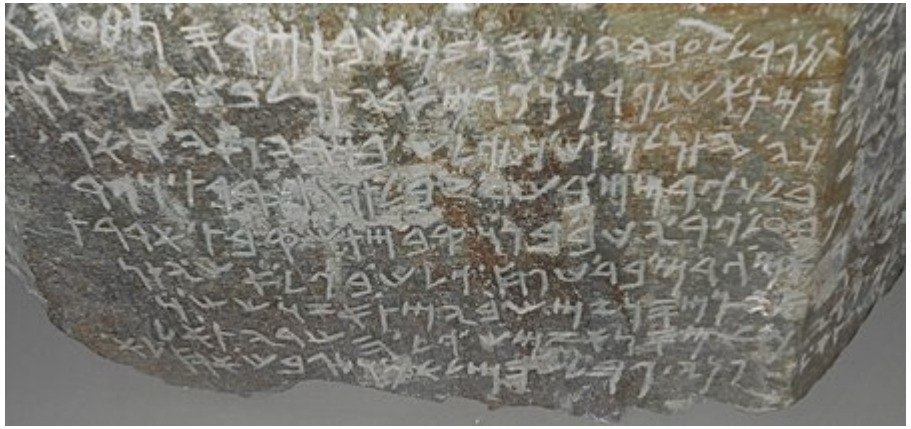
Aus dieser Schrift entwickelten sich u.a.:

- Phönizische Schrift
- Äthiopische Schrift

Geschichte

Phönizische Schrift

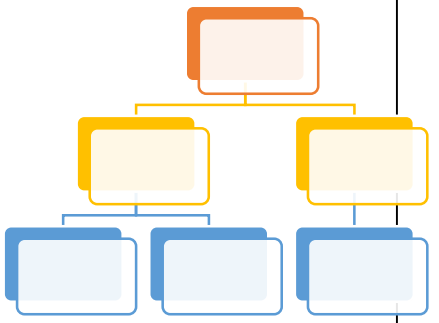
- Benutzt von Phönizier von ca. 1200 - 150 v.Chr
- Stammland der Phönizier: Libanon
- Seefahrer und Händler im gesamten Mittelmeerraum
- Kolonien in Nordafrika (Karthago!) und Südeuropa
- Schrift verbreitete sich schnell



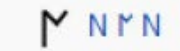
https://de.wikipedia.org/wiki/Ph%C3%B6nizische_Schrift

Aus der phönizischen Schrift entwickelten sich:

- Griechische Schrift
  - Lateinische Schrift
  - Koptische Schrift
  - Kyrillisch
  - Armenische Schrift
- Aramäische Schrift
  - Hebräische Schrift
  - Arabische Schrift
  - Syrische Schrift
    - Georgisch
    - Mongolisch
- Brahmi Schriften (Indien)
  - Tibet
  - Kambodscha
  - Indonesien
  - Thailand



(

Hieroglyph	Proto-Sinaitic	IPA value	reconstructed name	Phoenician	Paleo-Hebrew	Aramaic	Greek/Italic
		/ʔ/	ʾalp "ox"				
		/b/	bet "house"				
		/h/	*hll* "jubilation" > *he* "window"				
		/k/	kaf "palm of hand"				
		/m/	mayim "water"				
		/n/	*naḥš* "snake" > *nun* "fish"				
		/ʕ/	ʿen "eye"				
		/r/	roʾš "head"				
		/ʃ/	*šimš* "sun" > *šin* "tooth"				
		/t/	*tāw* "mark"				

Quelle: https://en.wikipedia.org/wiki/Proto-Sinaitic_script

Geschichte

Rechtschreibung

Hauptprinzip

- Aussprache

Aber :

  - Phoneme und Allophone
  - Variation in der Aussprache

Weitere Prinzipien

- Willkürliche Regeln (mehr, Teller, Dachs)
- Historie/ Emotionen (Thron; Tal, Saal, Zahl)
- Morphologie und Grammatik (Bad, dass, angewandt)
  - Paradigmatische Konsistenz
- Etymologie, Fremdwörter (ästhetisch, Cousine, Montage, Theke)


Deme beygefüget
eine Anweisung
zur
ORTHOGRAPHIA
oder
Rechtschreibung
der
Teutschen Sprache/
nach der heutigen Tages üblichen
Schreib-Art eingerichtet/
und mit
einem angefügten
VOCABULARIO
derer
Teutschen/ Lateinischen und
Französischen Wörter/
der Jugend zu Liebe ausgefertiget
und mit nöthigen Registern versehen
Von
Heinrich Volck von Wertheim.
CHEMNITZ/
bey Conrad Stössel/ 1711.

Encoding

Computer und Text

- Einerseits ist der Computer vor allem geeignet für das Bearbeiten großer Mengen numerischer Daten.
- Andererseits wurde der PC bis vor einigen Jahren Hauptsächlich als Schreibmaschine genutzt.

Binäre Zahlen

Computer speichert O und 1 (Binäre Zahlen)

Binäre und dezimale Zahlen können einfach umgerechnet werden.

Aber: nur ganze Zahlen

Aber: little-endian vs. big-endian

Aber: signed vs. unsigned

Aber: 8 bit, 16 bit, ...

Binäre Buchstaben

ASCII

- Zahlenwerte für Buchstaben werden willkürlich festgelegt
- ASCII= American Standard Code for Information Interchange
- Die älteste Kodierung. Geht bis ins 19. Jhdt. zurück. 1963 festdelegt
- 7-bit Kodierung: 2^7 = 128 Zeichen

Darstellung eines Codepages

- Codepage ist die Zuordnung der Buchstaben zu Zahlen
- Üblicherweise hexadezimal dargestellt:

```
     1000 (bin) =  8 (hex)
     1111 (bin) =  F (hex)
    10000 (bin) = 10 (hex)
1111 1111 (bin) = FF (hex)
```

Kodierung

Die ASCII Tabelle

Schriftzeichen	Dezimal	Hexadezimal	Binär
A	65	41	(0)1000001
B	66	42	(0)1000010
C	67	43	(0)1000011
D	68	44	(0)1000100
...	...	...	...

Bemerkenswertes

- A + 32 (dez) = a

A = 1000001	B = 1000010
a = 1100001	b = 1100010

- DEL = 1111111
- LF und CR
  - Windows: CR+LF
  - Unix: LF

Dec	Hex	Binair	Code
0	00	0000000	NUL (Null)
1	01	0000001	SOH (Start of Header)
2	02	0000010	STX (Start of Text)
3	03	0000011	ETX (End of Text)
4	04	0000100	EOT (End of Transmission)
5	05	0000101	ENQ (Enquiry)
6	06	0000110	ACK (Acknowledgment)
7	07	0000111	BEL (Bell (geluidssignaal))
8	08	0001000	BS (Backspace)
9	09	0001001	HT (Horizontal Tab)
10	0A	0001010	LF (Line Feed)
11	0B	0001011	VT (Vertical Tab)
12	0C	0001100	FF (Form Feed)
13	0D	0001101	CR (Carriage Return)
14	0E	0001110	SO (Shift Out)
15	0F	0001111	SI (Shift In)
16	10	0010000	DLE (Data Link Escape)
17	11	0010001	DC1 (Device Control 1)
18	12	0010010	DC2 (Device Control 2)
19	13	0010011	DC3 (Device Control 3)
20	14	0010100	DC4 (Device Control 4)
21	15	0010101	NAK (Negative Acknowledgement)
22	16	0010110	SYN (Synchronous Idle)
23	17	0010111	ETB (End of Transmission Block)
24	18	0011000	CAN (Cancel)
25	19	0011001	EM (End of Medium)
26	1A	0011010	SUB (Substitute)
27	1B	0011011	ESC (Escape)
28	1C	0011100	FS (File Separator)
29	1D	0011101	GS (Group Separator)
30	1E	0011110	RS (Record Separator)
31	1F	0011111	US (Unit Separator)

Kodierung

Codepage

	…0	…1	…2	…3	…4	…5	…6	…7	…8	…9	…A	…B	…C	…D	…E	…F
0…	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1…	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2…	SP	!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3…	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4…	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5…	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6…	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7…	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL

Andere Zeichen

Darstellung durch mehrere Zeichen

- LaTeX(1978)
  - \"U, \"a, \{ss}, usw.
- SGML, HTML, XML
  - Ü ä ß

Änderungen im Codepage

- Z.B. ISO 646
- Und viele weitere Varianten
- Vorangetrieben von Softwareherstellern
- Daher Varianten für verschiedene Länder/Märkte statt Sprachen

Kodierung

ISO 646: Ein Kuddelmuddel

Zeichenposition:	23	24	40	5B	5C	5D	5E	60	7B	7C	7D	7E
ISO 646-IRV	#	¤	@	[	\	]	^	`	{	|	}	~
Deutschland	#	$	§	Ä	Ö	Ü	^	`	ä	ö	ü	ß
Schweiz	ù	$	à	é	ç	ê	î	ô	ä	ö	ü	û
USA (ASCII)	#	$	@	[	\	]	^	`	{	|	}	~
Großbritannien	£	$	@	[	\	]	^	`	{	|	}	~
Frankreich	£	$	à	°	ç	§	^	`	é	ù	è	¨
Kanada	#	$	à	â	ç	ê	î	ô	é	ù	è	û
Finnland	#	$	@	Ä	Ö	Å	Ü	é	ä	ö	å	ü
Norwegen	#	$	@	Æ	Ø	Å	^	`	æ	ø	å	~
Schweden	#	$	É	Ä	Ö	Å	Ü	é	ä	ö	å	ü
Italien	£	$	§	°	ç	é	^	ù	à	ò	ù	ì
Niederlande	£	$	¾	ij	½	|	^	`	¨	ƒ	¼	´
Spanien	£	$	§	¡	Ñ	¿	^	`	°	ñ	ç	~
Portugal	#	$	@	Ã	Ç	Õ	^	`	ã	ç	õ	~

8-Bit Erweiterungen: Es wird nicht besser

MS-DOS (1987 DOS 3.3)

- Codepage 437 (Englisch)
- Codepage 850 (Westeuropäisch; DOS-Latin-1)
- Codepage 852 (Mitteleuropäisch; DOS-Latin-2)
- Usw.

Windows

- Windows-1250 (Mitteleuropäisch)
- Windows-1252 (Westeuropäisch)
- Usw.

ISO-8859 (1999)

- ISO 8859-1 (Westeuropäisch; Latin1)
- ISO 8859-2 (Mitteleuropäisch; Latin2)
- Usw.

Problem

Welche Kodierung wurde benutzt?

- Sollte im Header des Textes stehen
- Angaben, wenn vorhanden, oft falsch
- Viele Texte sind nicht konsistent
- Programme erraten die Kodierung
  - Leichter für den Benutzer
  - Vieles wird verschleiert und verschlimmbessert (Falsch erraten führt zu inkonsistente Fortsetzung)
  - Problembewusstsein nimmt ab.
- Lesen Sie mal:
  `http://www.joelonsoftware.com/articles/Unicode.html`

Unicode

Einheitliche Lösung: Unicode

- Ursprünglich 16 Bit: 65536 Zeichen
- Erweiterung: bis 10FFFF (21 Bit, 1,1 Millionen Zeichen)
- 7-bit ASCII enthalten

Unicode

- Alle Texte werden 3 mal so lang!
- Unicode wird nie benutzt zum Speichern von Texten
- Es gibt Kodierungen von Unicode. Für uns ist UTF-8 relevant

UTF-8

UTF-8

- Zeichen haben unterschiedlich lange Kodierungen
- Häufig benutzte Zeichen haben kurze, seltene lange Kodierungen
  - Sprachabhängig!

Das Prinzip

Unicodebereich (hexadzimal)	UTF-8-Kodierung (binär)
0000 0000 - 0000 007F	0xxxxxxx
0000 0080 - 0000 07FF	110xxxxx 10xxxxxx
0000 0800 - 0000 FFFF	1110xxxx 10xxxxxx 10xxxxxx

Unicode-Probleme

Technisch

- Viele Kodierungsvarianten
- Big endian vs. Little endian

Politisch

- Fragmentierte Schriften
- Position in der Tabelle

Praktisches

Fehlerhaftes Speichern oder Einlesen

Zeichen	Kodiert als ISO 88591		Dekodiert als CP 850
	Hex	Dec	
ä	E4	228	õ
ß	DF	225	■

- https://de.wikipedia.org/wiki/ISO_8859-1
- https://de.wikipedia.org/wiki/Codepage_850
- https://unicode-table.com/de

Zeichen	Unicode	UTF8	Dekodiert als ISO 8859-1
ö	F6	C3 B6	Ã ¶
ß	DF	C3 9F	Ã ?

Praktisches

Erklärung

- ö ist in Unicode F6 (hex) bzw. 246 (dec) bzw 1111 0110 (bin)
- In UTF8 kodiert ist das: 110**0 0011** 10**11 0110**
- Diese zwei Bytes entsprechen: C3 (195) und B6 (182)
- In ISO 8859-1 sind das die Zeichen Ã und ¶