Introduction to Computational Linguistics
Session 2: Encoding Language

Denise Löfflad

Universität Tübingen

November 8, 2023

Plan

- What is Language?
- Writing Systems
- Bits and Bytes

Mental Health

Studying and life in general can be very challengeing and sometimes it can take a toll on your mental health.

1. On the website `www.werhilftweiter.de` you can find information on all kinds of help in the Tübingen area, including help with mental health issues.
2. You can download the app *Krisenkompass*. It supports you step-by-step in crisis situations, no matter if you need help yourself or if you want to help someone.
3. Zentrale Studienberatung `https://uni-tuebingen.de/studium/beratung-und-info/zentrale-studienberatung/`
4. The emergency numbers 112 and 116 117 are not only available for physical emergencies but also for mental health crisis situations!

Definitions of Language

What is Language? `https://www.menti.com/alfbyro5z869`

What is Language?

"Language is at the heart of all things human." Archibald and O'Grady [2008]

"Language is a structured system of communication that consists of grammar and vocabulary. It is the primary means by which humans convey meaning," Wikipedia[1],

"A common language connects the members of a community into an information-sharing network with formidable collective powers" Pinker [1995]

According to linguistics, language can be described with regards to phonetics, phonology, morphology, syntax, semantics, and pragmatics

[1] `https://en.wikipedia.org/wiki/Language`, last accessed Nov. 8 2023

Writing Systems

What is writing?
"a system of more or less permanent marks used to represent an utterance in such a way that it can be recovered more or less exactly without the intervention of the utterer." Daniels and Bright [1996]
Different types of writing systems are used:

- Alphabetic
- Syllabic
- Logographic

Alphabetic systems

Alphabets (phonemic alphabets)

- represent all sounds, i.e., consonants and vowels
- Examples: Etruscan, Latin, Korean, Cyrillic, Runic, International Phonetic Alphabet (`https://www.ipachart.com/`)

Abjads (consonant alphabets)

- represent consonants only (sometimes plus selected vowels; vowel diacritics generally available)
- Examples: Arabic, Aramaic, Hebrew

Syllabic systems

Symbols represent syllables which make up words

- all human languages have syllables, though some syllable systems are easier than others
- simple syllable structures $\rightarrow$ relatively small set of possible syllables

Logographic systems

Symbols represent syllables which make up words

- A logograph represents a unity of meaning
- most natural language writing systems that use logographs are not purely logographic

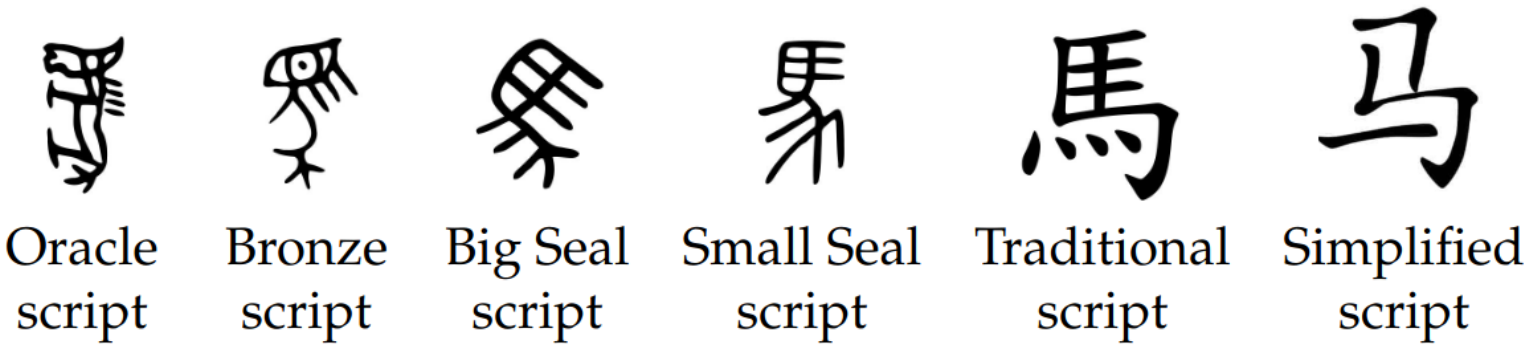

Figure: Evolution of chinese character *horse*

Other systems

- Some systems (like Korean) use hybrids of e.g. syllables and alphabets
- Are emojis a writing system?

Encoding Language

`https://www.educba.com/types-of-computer-language/`

- How to encode writing systems on the computer?
- How do we store anything on a computer?
- $\rightarrow$ bits & bytes

Bits and Bytes

- **bit**: a unit of information with one of two possible values
  common representations: true and false, on and off, 0 and 1
- **byte**: a sequence of 8 bits e.g. 10101101 or 10000000
- Note: talk about big endian notation, where most significant is leftmost (vs little endian)
- b = bit = binary digit
- B = byte

Binary Numbers

- the binary system is base 2
- this means we only use two symbols: 0 and 1
- We can combine these symbols too,
  E.g., 11, 1001, 10000000
- Position k of digit represent multiples of 2^k: 1, 2, 4, 8, 16, 32, ...
- Let's look at the binary number 00101101

2^7	2^6	2^5	2^4	2^3	2^2	2^1	2^0
0	0	1	0	1	1	0	1

Binary Numbers

2^7	2^6	2^5	2^4	2^3	2^2	2^1	2^0
0	0	1	0	1	1	0	1

$$2^7 \cdot 0 + 2^6 \cdot 0 + 2^5 \cdot 1 + 2^4 \cdot 0 + 2^3 \cdot 1 + 2^2 \cdot 1 + 2^1 \cdot 0 + 2^0 \cdot 1$$

$\Leftrightarrow$

$$32 + 8 + 4 + 1 = 45$$

Hexadecimal System

- the hexadecimal system is base 16
- letters of the alphabet are used to arrive at 16 symbols: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F, where A = 10, B = 11, C = 12, D = 13, E = 14, F = 15
- Position k of digit represents multiple of 16^k: 1, 16, 256, 4096, ...
- Example: 0xA8E is interpreted as

$$10 \cdot 16^2 + 8 \cdot 16^1 + 14 \cdot 16^0 = 2702$$

Binary $\leftrightarrow$ Hexadecimal

How to convert the binary number 100101110 to hexadecimal?

- 4 positions in binary can be represented with 1 position in hexadecimal
- $\rightarrow$ group the binary number into positions of 4 (starting from the right) and convert to hexadecimal

How to convert the hexadecimal 0xB5F08 to binary?

- 1 position in hexadecimal can be represented with 4 positions in binary
- $\rightarrow$ convert every position to binary

Binary $\leftrightarrow$ Hexadecimal

How to convert the binary number 100101110 to hexadecimal?

- [0001][0010][1110]
- [1110] = 14 = E
- [0010] = 2
- [0001] = 1
- $\rightarrow$ 0x12E
- (302 in decimal)

Binary $\leftrightarrow$ Hexadecimal

How to convert the hexadecimal number 0xB5F08 to binary?

- 8 = 1000
- 0 = 0000
- F = 1111
- 5 = 0101
- B = 1011
- $\rightarrow$ 10110101111100001000
- (745224 in decimal)

ASCII

Using 8 bits and where each byte stores a separate character, we can represent 256 different characters. To ensure compatibility across systems, we need encoding systems.

- This is enough to store every character from the Latin alphabet for English, plus additional characters such as space, comma, etc.
- ASCII, the **A**merican **S**tandard **C**ode for **I**nformation **I**nterchange, uses 7 bits (128 characters)

ASCII

USASCII code chart

Bits b_7 b_6 b_5 →	0 0 0	0 0 1	0 1 0	0 1 1	1 0 0	1 0 1	1 1 0	1 1 1
b_4 b_3 b_2 b_1 ↓ Row ↓ Column →	0	1	2	3	4	5	6	7
0 0 0 0	0	NUL	DLE	SP	0	@	P	`	p
0 0 0 1	1	SOH	DC1	!	1	A	Q	a	q
0 0 1 0	2	STX	DC2	"	2	B	R	b	r
0 0 1 1	3	ETX	DC3	#	3	C	S	c	s
0 1 0 0	4	EOT	DC4	$	4	D	T	d	t
0 1 0 1	5	ENQ	NAK	%	5	E	U	e	u
0 1 1 0	6	ACK	SYN	&	6	F	V	f	v
0 1 1 1	7	BEL	ETB	'	7	G	W	g	w
1 0 0 0	8	BS	CAN	(	8	H	X	h	x
1 0 0 1	9	HT	EM	)	9	I	Y	i	y
1 0 1 0	10	LF	SUB	*	:	J	Z	j	z
1 0 1 1	11	VT	ESC	+	;	K	[	k	{
1 1 0 0	12	FF	FS	,	<	L	\	l	|
1 1 0 1	13	CR	GS	-	=	M	]	m	}
1 1 1 0	14	SO	RS	.	>	N	^	n	~
1 1 1 1	15	SI	US	/	?	O	_	o	DEL

Figure: ASCII code table

ASCII Limitations

ASCII has severe limitations for non-US usage:

- no umlauts
- no accents or other diacritics no symbols for currencies other than dollar

Result: different modifications were established in different countries, replacing different characters by symbols of local significance, e.g.:

- ISCII (Indian scripts),
- TSCII (Tamil),
- VSCII (Vietnamese)
- sometimes changes are small, e.g. JIS C-6220 in Japan, replacing the backslash with a Yen sign
- $\rightarrow$ risk of misidentification

Unicode

Unicode has a single representation for every character in any existing writing system
Unicode uses 32 bits to encode characters, i.e.

$$2^{32} = 4,294,967,296$$

unique characters.
idea: provide one universal encoding for all languages, both current and historic writing systems, symbols
Solution: UTF-32, UTF-16, and UTF-8

Unicode UTF-8

- UTF-8 allows for variable length
- one byte utf-8 character is direct mapping to ASCII encoding, allowing for backwards compatibility with ASCII (when something is encoded in ascci, it can be read in utf-8)
- leftmost bit indicates how many bytes are used:
  - 0..... $\rightarrow$ one byte
  - 110..... $\rightarrow$ two bytes (following: 10.....)
  - 1110..... $\rightarrow$ three bytes (following: 10.....)
  - ...

Next session

For the next session, read the chapter on Writer's Aids!
The first assignment is released today, submit until Nov. 22 in PDF form.
Don't forget the zero-points sheet. If you have questions, use the moodle forum.

References and Acknowledgments

These slides are largely based on Dickinson et al. [2012] and partly based on slides for Text Technology, 2023 by Stephen Bodnar and on Detmar Meurers' slides on Second Language Acquisition.

John Archibald and William O'Grady. Contemporary linguistic analysis, 2008.

Peter T Daniels and William Bright. *The world's writing systems*. Oxford University Press, 1996.

Markus Dickinson, Chris Brew, and Detmar Meurers. *Language and computers*. John Wiley & Sons, 2012.

Steven Pinker. *The language instinct: The new science of language and mind*, volume 7529. Penguin UK, 1995.