

Language and Computers – where to start?

If we want to do anything with language, we need a way to represent language.

We can:

- write on a computer & read words on a computer
- speak to a computer & listen to a computer

Computer has to have some way to represent **text** and **speech** and sometimes has to be able to relate the two.



1

Text and Speech Encoding

Linguistics 384
Winter 2004

Text and Speech Encoding

- Writing systems
- Encoding writing systems
- Digital representations of speech
- Spectrograms
- Automatic Speech Recognition (ASR)
- Text-to-Speech Synthesis



3

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.” (Peter T. Daniels, *The World's Writing Systems*)

- Alphabetic writing systems
- Syllabic writing systems
- Logographic writing systems



4



Alphabetic systems

- **Alphabets** = phonemic alphabets: represent consonants and vowels (all sounds)
e.g. Etruscan, Latin, Korean, Cyrillic, Runic

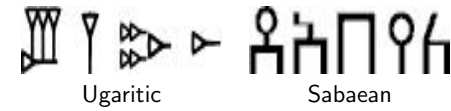


5



Abjads

- **Abjads** = consonant alphabets: represent consonants only.
e.g. Arabic, Aramaic, Phoenician, Hebrew



6



Different kinds of alphabets

English alphabet:

- Not an exact match between letter and sound
- How many different ways are there to write the sound *oo* as in *boo*?

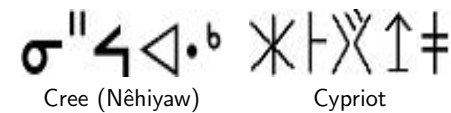
Italian and Spanish have more of a one-to-one sound-to-letter mapping.

7



Syllabic systems

- **Syllabaries** = writing systems with separate symbols for each syllable of a language
e.g. in Cherokee there are separate symbols for *a*, *ga*, *ha*, *he*, *hi*, and so on



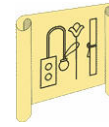
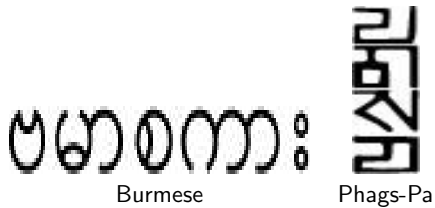
8



Syllabic alphabets

- **Syllabic alphabets** = writing systems with symbols that represent a consonant with a vowel, but the vowel can be changed by adding a **diacritic** = a symbol above it.

e.g. in Hanuo'o, there is a symbol for *ma*, but by drawing a line (a diacritic) over the top left, it becomes *mi*. If a diacritic is added to the bottom right, it becomes *mu*



Logographic writing systems

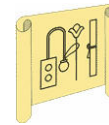
- **Logograms** = symbols which represent parts of words or whole words
→ They are not necessarily "pictures" of what they represent.█
- **Semantic-phonetic compounds** = symbols which have a meaning element, which hints at the meaning, and a phonetic element, which hints at the pronunciation.█

Examples: Chinese (Zhongwén), Japanese (Nihongo), Mayan























Vietnamese example

鈕汝干登堆密賜紅粉發化遠撐冥潘層
 埃醜浮米城殿尼散長城接抹牽月槐甘泉
 式遠於舌鏡室探柄射腦信撒定和為征活
 臨森楠額機或梓室式自尼仗委最懸培



Mayan

Mayan had both a logographic system and a syllabic system. Part of the syllabary:

 tza	 tze	 tzi		 tzu
 tz'a		 tz'i		 tz'u
 wa		 wi	 wo	
 xa	 xe	 xi	 xo	 xu
 ya	 ye	 yi	 yo	 yu

Other writing systems

- Braille: a system based on touch.
- A *chromatographic* system of writing used by the Benin and Edo people in southern Nigeria: writing based on different color combinations with different symbols
(http://www.library.cornell.edu/africana/Writing_Systems/Chroma.html)

13

Chromatographic system



14

One to one mapping?

There is not a simple correspondence between a writing system and a language.
e.g. We use the Roman alphabet, but Arabic numerals (2 instead of the Roman II) ■
We'll look at a few different examples

- Japanese
- Korean
- Azeri

15

Japanese

Japanese: logographic system *kanji*, syllabary *katakana*, syllabary *hiragana*

- kanji: 5,000-10,000 borrowed Chinese characters ■
- katakana
 - Used mainly for non-Chinese loan words, onomatopoeic words, foreign names, and for emphasis ■
- hiragana
 - Originally used only by women (10th century), but codified in 1946 with 48 syllables
 - used mainly for word endings, kids' books, and for words with obscure kanji symbols ■
- Romaji: Roman characters

16

Japanese example

カプセルホテル

各室がカプセル形の簡易ホテル。終電に乗り遅れたサラリーマンなどが高いタクシー代を払って帰宅するより安く済むことから、手軽に利用している。

⇒ kanji in red, hiragana in black, katakana in blue
(<http://www.omniglot.com/writing/japanese.htm#origin>)

"Capsule Hotel"

A simple hotel where each room is capsule-shaped. When businessmen miss the last train home, they can stay overnight very cheaply instead of paying a lot of money to go home by taxi."

17

Korean

"Korean writing is an alphabet, a syllabary and logographs all at once."
(<http://home.vicnet.net.au/õzideas/writkor.htm>)

- The *hangul* system was developed in 1444 during King Sejong's reign.
 - There are 24 letters: 14 consonants and 10 vowels
 - But the letters are grouped into syllables, i.e. the letters in a syllable are not written separately as in the English system, but together form a single character.
- In South Korea, *hanja* = logographic Chinese characters are also used.

18

Azeri

A Turkish language with speakers in Azerbaijan, northwest Iran, and (former Soviet) Georgia

- 7th century until 1920s: Arabic scripts. Three different Arabic scripts used
- 1929: Latin alphabet enforced by Soviets to reduce Islamic influence.
- 1939: Cyrillic alphabet enforced by Stalin
- 1991: Back to Latin alphabet, but slightly different than before.
 - Latin typewriters and computer fonts were in great demand in 1991

19

Comparison of writing systems

What are the pros and cons of each type of system?

- accuracy = can every word be written down accurately?
- learnability = how long does it take to learn the system?
- cognitive ability = are some systems unnatural? (e.g. does dyslexia show that alphabets are unnatural?)
- language-particular differences: English has thousands of possible syllables; Japanese has very few in comparison
- connection to history/culture = will changing a writing system have social consequences?

20

Encoding written language

Information on a computer is stored in **bits** = either a 1 (yes) or a 0 (no). A collection of 8 bits makes up a **byte**.

So, a string of bits look like:

001
011
010

Bits work something like this (using 3 bits):

4	2	1	=	4	2	1	=	1
0	0	1	=	0	0	1	=	1
0	1	1	=	0	2	1	=	3
1	0	1	=	4	0	1	=	5

Calculating binary numbers

Let's say we have 4 bits now, and we want to know how to write 10 in bit (or *binary*) notation.

8	4	2	1
?	?	?	?
8 < 10	?	?	?
1	8 + 4 = 12 > 10	?	?
1	0	8 + 2 = 10 = 10	?
1	0	1	0

The mapping between bits and real numbers

If we had only 3 bits, we could represent 8 different values:

000	0
001	1
010	2
011	3
100	4
101	5
110	6
111	7

Then, for every value, you can store a character there. ...

Using bits to store characters

So, with 3 bits, we could store 8 characters by setting up a correspondence like so:

Value	Character
0	q
1	w
2	e
3	r
4	t
5	y
6	u
7	i

What such a correspondence means

- Now, when you type in the letter *r*, this gets converted to 3, or 011, in our 3-bit world.■
- Then, to display the character onscreen, the computer converts the code into the appropriate font symbol, using some sort of text processing package.
If your computer's coding system is different than someone else's, you might see funny characters sometimes.

25

Using bytes to store characters

With 8 bits (a single byte), you can represent 256 different characters. Why would we want so many?■

- If you look at a keyboard, you will find lots of non-English characters.
- With 256 possible characters, we can store every single letter used in English, plus all the things like commas, periods, space bar, percent sign (%), back space, and so on.

26

ASCII

ASCII = the American Standard Code for Information Interchange is a 7-bit code for storing English. 7 bits = 128 possible characters.

Note that our alphabetic ordering has been maintained.

27

E-mail issues

Have you ever had something like the following at the top of an e-mail sent to you?

```
[ The following text is in the 'ISO-8859-1' character set. ]  
[ Your display is set for the 'US-ASCII' character set. ]  
[ Some characters may be displayed incorrectly. ]
```

Mail sent on the internet used to only be able to transfer the 7-bit ASCII messages. But now we can detect the incoming character set and adjust the input.■

Note that this is an example of **meta-information** = information which is printed as part of the regular message, but tells us something about that message.

28

MIME is money

MIME = Multipurpose Internet Mail Extensions

Information included in a message which tells us:

- which version of MIME is being used
- what the character set is
- if that character set was altered, how it was altered

Mime-Version: 1.0

Content-Type: text/plain; charset=US-ASCII

Content-Transfer-Encoding: 7bit

29

Different coding systems

But wait, didn't we want to be able to encode *all* languages?

There are ways ...

- Extend the ASCII system with various other systems
 - e.g. ISO 8859-7: Greek alphabet
 - ISO 8859-8: Hebrew alphabet
 - JIS X 0208: Japanese characters
- Have one system for everything → **Unicode**

30

Unicode

Problems with having multiple encoding systems:

- Conflicts: two encodings can use the same number for two different characters and use different numbers for the same character.
- Hassle: have to install many, many systems if you want to be able to deal with various languages

Unicode tries to fix that by having a single representation for every possible character.

“Unicode provides a unique number for every character, no matter what the platform, no matter what the program, no matter what the language.”
(www.unicode.org)

31

How big is Unicode?

Version 3.2 has codes for 95,221 characters from alphabets, syllabaries and logographic systems.

- Uses 32 bits – meaning we can store $2^{32} = 4,294,967,296$ characters.
- 4 million?!? Does our computer have enough space?

32

Making Unicode smaller

Unicode has three versions

- UTF-8 (8 bits)
- UTF-16 (16 bits): $2^{16} = 65536$
- UTF-32 (32 bits): use only when memory is no problem

33

How do we type everything in?

- Use a keyboard tailored to your specific language
e.g. Highly noticeable how much slower your English typing is when using a Danish-designed keyboard.
- Use a processor that allows you to switch between different character systems.
e.g. Type in Cyrillic characters on your English keyboard.
- Use combinations of characters.
An e followed by an ' might result in an é
- Pick and choose from a table of characters.

So, now we can encode every language, as long as it's written.

34

Unwritten languages

Many languages have never been written down. Of the 6700 spoken, 3000 have never been written down.

- Salar, a Turkic language in China.
- Gugu Badhun, a language in Australia.
- Southeastern Pomo, a language in California

35

The need for speech

- What if we want to work with an unwritten language?
- What if we want to examine the way someone talks and don't have time to write it down?

Many applications for encoding speech:

- Building spoken dialogue systems, i.e. speak with a computer (and have it speak back).
- Helping people sound like native speakers of a foreign language.
- Helping speech pathologists diagnose problems

36

What does speech look like?

We can **transcribe** (write down) the speech into a **phonetic alphabet**.

- It is very expensive and time-consuming to have humans do all the transcription.
- To automatically transcribe, we need to know how to relate the audio file to the individual sounds that we hear.
 - ⇒ We need to know:
 - some properties of speech
 - how to measure these speech properties
 - how these measurements correspond to sounds we hear

37

Articulatory properties

We could talk about how sounds are produced in the vocal tract, i.e. **articulatory phonetics**

- *place of articulation* (where): [t] vs. [k]
- *manner of articulation* (how): [t] vs. [s]
- *voicing* (vocal cord vibration): [t] vs. [d]

But unless the computer is modeling a vocal tract, we need to know acoustic properties of speech which we can *quantify*.

38

Acoustic properties

Sound waves = “small variations in air pressure that occur very rapidly one after another” (Ladefoged, *A Course in Phonetics*)

⇒ Akin to ripples in a pond

- **speech flow** = rate of speaking, number and length of pauses (seconds)
- **loudness** (amplitude) = amount of energy (decibels)
- **frequencies** = how fast the sound waves are repeating (cycles per second, i.e. Hertz)
 - **pitch** = how high or low a sound is
 - In speech, there is a **fundamental frequency**, or pitch, along with higher-frequency **overtones**.
- **intonation** = rise and fall in pitch

39

What makes representing speech hard?

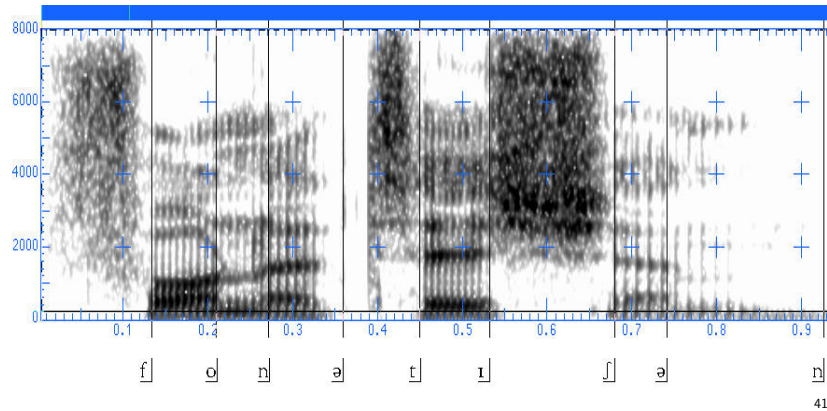
Difficulties:

- People have different dialects and different size vocal tracts and thus say things differently
- Sounds run together, and it's hard to tell where one sound ends and another begins.
- What we think of as one sound is not always (usually) said the same: **coarticulation** = sounds affecting the way neighboring sounds are said
e.g. *k* is said differently depending on if it is followed by *ee* or by *oo*.
- What we think of as two sounds are not always all that different.
e.g. The *s* in *see* is very acoustically similar to the *sh* in *shoe*

40

Spectrograms

Spectrogram = a graph to represent (the frequencies of) speech over time.



41

How measurements correspond to sounds we hear

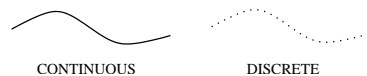
- How dark is the picture? → How loud is the sound?
We can measure this in decibels.■
- Where are the lines the darkest? → Which frequencies are the loudest and most important?
We can measure this in terms of Hertz, and it tells us what the vowels are.■
- How do these dark lines change? → How are the frequencies changing over time?
Which consonants are we transitioning into?

42

How did we get these measurements?

sampling rate = how many times in a given second we extract a moment of sound; measured in samples per second■

- Sound is **continuous**, but we have to store data in a **discrete** manner.



- We store data at each discrete point, in order to capture the general pattern of the sound

43

Sampling rate

- The sampling rate is often 8000 or 16,000 samples per second. The rate for CDs is 44,100 samples/second (or **Hertz (Hz)**)■
- The higher the sampling rate, the better quality the recording ... but the more space it takes.■
- Speech needs at least 8000 samples/second, but most likely 16,000 or 22,050 Hz will be used nowadays.

44

Applications of speech encoding

Mapping sounds to symbols (alphabet), and vice versa, isn't all that easy.

- **Automatic Speech Recognition (ASR)**: sounds to text
- **Text-to-Speech Synthesis (TTS)**: texts to sounds

45

Automatic Speech Recognition (ASR)

Automatic speech recognition = process by which the computer maps a speech signal to text. ■

Uses/Applications:

- Dictation
- Telephone conversations
- People with disabilities – e.g. a person hard of hearing could use an ASR system to get the text

46

Kinds of ASR systems

Different kinds of systems:

- Speaker dependent = work for a single speaker
- Speaker independent = work for any speaker of a given variety of a language, e.g. American English
- Speaker adaptive = start as independent but begin to adapt to a single speaker to improve accuracy

47

Kinds of ASR systems

- Differing sizes of vocabularies, from tens of words to tens of thousands of words ■
- **continuous speech** vs. **isolated-word** systems:
 - continuous speech systems = words connected together and not separated by pauses
 - isolated-word systems = single words recognized at a time, requiring pauses to be inserted between words
 - easier to find the endpoints of words

48

Steps in an ASR system

1. Digital sampling of speech
2. **Acoustic signal processing** = converting the speech samples into particular measurable units
3. Recognition of sounds, groups of sounds, and words

May or may not use more sophisticated analysis of the utterance to help.

49

Text-to-Speech Synthesis (TTS)

Could just record a voice saying phrases or words and then play back those words in the appropriate order.

Or can break the text down into smaller units

1. Convert input text into phonetic alphabet
2. Synthesize phonetic characters into speech

To synthesize characters into speech, people have tried:

- using formulas which adjust the values of the frequencies, the loudness, etc.
- using a model of the vocal tract and trying to produce sounds based on how a human would speak

50

It's hard to be natural

When trying to make synthesized speech sound *natural*, we encounter the same problems as what makes speech encoding in general hard:

- The same sound is said differently in different contexts.
- Different sounds are sometimes said nearly the same.
- Different sentences have different intonation patterns.
- Lengths of words vary depending on where in the sentence they are spoken.

The car crashed into the tree.
It's my car.
Cars, trucks, and bikes are vehicles.

51

Speech to Text to Speech

If we convert speech to text and then back to speech, it should sound the same, right?

- But at the conversion stages, there is **information loss**. To avoid this loss would require a lot of memory and knowledge about what exact information to store.
- The process is thus **irreversible**.

52