

Assignment 2: Text & Speech Encoding Web searching

DUE: January 21, 2003 (Wednesday)

1. Go to www.omniglot.com. Find one example of an abjad, an alphabet (NOT Latin/Roman), a syllabary, a syllabic alphabet, and a logography. For each example, give me two (2) facts about it. For the abjad and syllabary, attempt to write your first name in the language—you will likely have to approximate your name. (Pay attention to which direction the language is written in.)
2. In class we mentioned how many characters can be stored with a certain number of bits. Here's a recap:

Number of bits	Number of characters
1	2
2	4
3	8
4	16
5	32
6	64
7	128
8	256

- (a) Based on this, how many characters would you be able to store using 9 bits?
 - (b) What is the general formula – i.e. for n number of bits, how many characters can you store?
3. Give me the base ten numbers for the following:
 - (a) 1011 1111
 - (b) 0101 0101
 - (c) 1010 0010
 4. Write out the first 7 letters of your name using the ASCII codes – written in both numeric and 7-bit notation.¹ As an example, here are the first seven letters of my name. (Note that spaces are not strictly necessary for binary numbers—they just make it easier to read. And note that dots (...) are not allowed in *your* answer.)

¹So, if your first name is 7 letters or longer, write out only the first 7 letters of your first name. If your first name is less than 7 letters, write out your first name plus however many characters from your last name you need to make 7.

letter	ASCII number	bit notation
M	77	100 1101
a	97	110 0001
r	114	...
k	107	...
u	117	...
s	115	...
D	68	...

5. The stick your finger in your mouth exercise:

In your own words, describe the differences between the following pairs of sounds. Consider: where your tongue is, if your tongue is making contact with any part of your mouth, if your vocal cords are vibrating, where/how the air is moving out of your mouth, if there are actually two sounds involved, etc. There may be more than one difference. (When in doubt, simply describe what's happening.)

- (a) [k] vs. [g]
 - (b) [b] vs. [f]
 - (c) [n] vs. [ng] (as in *ring*)
 - (d) (harder) [sh] vs. [ch]
 - (e) (harder) [r] vs. [l]
6. (a) Based on the previous question, would you say that [k] and [g] are more or similar to each other than [r] and [l] are? Why?
- (b) When someone else is talking, do you notice any difference in difficulty in distinguishing between [k] and [g] or between [r] and [l]? If so, which distinction is harder?
7. Looking back at your notes for ASR and TTS systems—and, more importantly, THINKING about the issues involved—which do you see as a harder task: automatic speech recognition, or text-to-speech synthesis? Or are they equally hard? I'm not looking for one correct answer, just solid reasoning.

8. Your friend tells you the following:

When I fall asleep watching tv, I always wake up with a pain in my lower back. I want to know what kind of sofas and easy chairs are good for my back.

NOTE: Do NOT enter any queries until instructed to.

- (a) Identify the keywords.
- (b) Identify synonyms of the keywords.

- (c) Decide which synonyms are best by determining which are least ambiguous.
- (d) Decide which words need to be kept in the query, but might still be problematic.
- (e) Formulate a boolean query.
- (f) Enter this query at:
<http://www.alltheweb.com/advanced?t=all&c=web>
 NOTE: The query language for alltheweb can be found at:
http://www.alltheweb.com/help/faqs/query_language#2
- (g) How many of the first 10 results were what you wanted?
- (h) How could you tell that these results were what you wanted?

9. Fun with Rock Bands:

- (a) Using only the band's name, search at www.google.com for the following rock bands – DO NOT USE QUOTES around the band name. Record how many of the top ten results are actually about that band.
 - Cream
 - Judas Priest
 - KISS
 - The Knack
 - The The
- (b) When you type *The Knack*, google tells you something about how it treats *the*. If that were always the case all the time, how many results should *The The* return? What happens when you enter *The The*?
- (c) Try entering “The Knack” and “The The” with quotes around them. Now, how many of the top ten results are about the band?
- (d) With *The Knack* and *Cream*, you get some results that talk about *Knick-Knacks* and *Ice Cream*. How could a search eliminate these unwanted items?
- (e) *Judas* and *Priest* seem like two words well-suited for a religious website. Find two (2) religious sites which use both words and write down their addresses. Are the two words ever used together (i.e. side-by-side) when not talking about the rock band?
- (f) Let's say I wanted to find sites that were only about kissing and had nothing to do with the rock band. How might part of speech information help?