

# Midterm Review

Linguistics 445/515

For the Midterm on Monday, October 20, 2008

## 1 Topics to be covered

1. Text & Speech encoding
2. Searching
3. Corpus annotation
4. Text Classification
5. Spam Filtering

NB: You will not have any Python questions on the exam.

## 2 Format of the exam

You will have the entire 1:15 (9:30-10:45) should you need/want it.

1. Matching: 10-20 terms (see list below)
2. Calculations: 5-10 questions
  - Binary numbers, ASCII encoding
  - Boolean expressions
  - Regular expressions
  - Weblinking & webpage ranking
  - Precision/Recall
  - Semantic ontologies
  - Part-of-speech & chunk annotating
  - Stylometric analysis
  - Frequency distributions
  - Rule-based operations
  - Probability calculations (ASR & spam filtering)
3. Short answer: answer 3–5 out of 5–8
  - Types of writing systems, pros & cons
  - Relation of writing systems to languages

- Types of character encoding systems, e.g., ASCII & Unicode
- Challenges of ASR & TTS
- How measurements do & do not correspond to what we hear
- Searching in databases vs. on the web vs. in a corpus
- Improving searching, e.g., semantic web
- Levels of linguistic annotation
- Kinds of information needed for document classification
- N-gram analysis
- Plagiarism detection
- The social context of spam & approaches to fighting spam
- Rule-based spam filters
- Statistical spam filters
- Tokenization & Devious spam

### 3 Terms to know

#### 3.1 Text/Speech encoding

- |                              |                          |                              |
|------------------------------|--------------------------|------------------------------|
| – text                       | – ASCII                  | – fundamental frequency      |
| – speech                     | – Unicode                | – intonation                 |
| – abjad                      | – Character encoding     | – spectrogram                |
| – alphabet                   | – MIME                   | – sampling rate              |
| – syllabary                  | – meta-information       | – ASR                        |
| – syllabic alphabet          | – continuous             | – TTS                        |
| – diacritic                  | – discrete               | – continuous speech system   |
| – logographic system         | – Hertz                  | – isolated-word system       |
| – logograph                  | – transcribe             | – acoustic signal processing |
| – pictograph                 | – phonetic alphabet      | – information loss           |
| – ideograph                  | – coarticulation         | – irreversible               |
| – semantic-phonetic compound | – articulatory phonetics |                              |
| – bit                        | – speech flow            |                              |
| – byte                       | – loudness/amplitude     |                              |
| – Big-Endian                 | – intonation             |                              |
| – Little-Endian              | – pitch                  |                              |

#### 3.2 Searching

- |                     |                      |                       |
|---------------------|----------------------|-----------------------|
| – database          | – synonym            | – operator precedence |
| – database frontend | – boolean expression | – escaped character   |
| – keyword           | – regular expression | – counter             |
| – query             | – operators          | – literal strings     |

- disjunction
- negation
- counters
- wildcard
- linking
- link counting
- formal language
- regular language
- meta data/meta tag
- click-through measurement
- database
- index
- search engine
- relevancy
- precision
- recall
- accuracy
- index
- clustering
- stemming
- capitalization
- ambiguity
- stop words
- web forms
- grep/egrep

### 3.3 Corpus annotation

- ontology
- corpus/corpora
- corpus annotation
- XML
- word type
- word token
- tokenization
- lemmatization
- part-of-speech tagging

### 3.4 Text classification

- authorship attribution
- text classification
- n-gram
- frequency distribution
- stylometry
- lexical style markers
- function words
- text reuse
- plagiarism

### 3.5 Spam filtering

- language identification
- document classification
- spam
- spam filter
- blacklist
- whitelist
- rule-based filtering
- weight
- spam probability
- statistical filtering
- (machine) learning
- false positives
- collaborative filtering
- message inoculation
- structured information