

# Midterm Review

Linguistics 384 (Meurers)

For the Midterm in class on Wednesday, November 1, 2006

## 1 Topics to be covered

1. Text & Speech Encoding
2. Searching
3. Spam filtering (document classification)
4. Spelling Correction

## 2 Format of the exam

You will have the entire 1:48 (11:30-1:18) should you need/want it, but it should be doable to complete it in around one hour.

1. Matching: 10-20 terms (see list below)
2. Calculations: 5-10 questions
  - Binary numbers, ASCII encoding
  - Boolean expressions
  - Regular expressions
  - Precision/Recall
  - Rule-based operations (spam & spelling)
  - Minimum edit distance
  - Bigram array (positional and non-positional)
  - Confusion matrix
3. Short answer: answer 3–5 out of 5–10  
⇒ How do the various concepts/technologies work?
  - Types of writing systems
  - ASCII, Unicode
  - ASR & TTS
  - N-grams (spam & spelling)
  - Rule-based spam filters
  - Statistical spam filters

- Devious spam
- Types and causes of spelling errors
- Isolated-word error correction (and its limits)

### 3 Terms to know

#### 3.1 Text/Speech encoding

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

#### 3.2 Searching

- |                       |                             |                  |
|-----------------------|-----------------------------|------------------|
| – keyword             | – linking                   | – recall         |
| – query               | – link counting             | – accuracy       |
| – synonym             | – formal language           | – web crawler    |
| – boolean expression  | – regular language          | – clustering     |
| – regular expression  | – corpus                    | – stemming       |
| – operators           | – meta data                 | – capitalization |
| – operator precedence | – meta tag                  | – ambiguity      |
| – escaped character   | – click-through measurement | – stop words     |
| – counter             | – database                  | – web forms      |
| – literal strings     | – index                     | – grep           |
| – disjunction         | – search engine             | – (term) weight  |
| – negation            | – relevancy                 | – hash table     |
| – counters            | – precision                 | – part of speech |
| – wildcard            |                             |                  |

### 3.3 Spam filtering/Document classification

- language identification
- document classification
- n-gram
- frequency
- distribution
- spam
- spam filter
- blacklist
- whitelist
- rule-based filtering
- weight
- spam probability
- statistical filtering
- learning
- false positives

### 3.4 Spell checking

- productivity
- inflection
- tokenization
- detection
- correction
- Spoonerism
- word recognition
- interactive spelling checkers
- automatic spelling correctors
- phonetic errors
- run-on errors
- split errors
- isolated words
- (words in) context
- nonword error detection
- isolated-word error correction
- context-dependent word correction
- dictionary lookup
- dictionary construction
- insertion
- deletion
- substitution
- transposition
- single-error misspelling
- multi-error misspelling
- semantics
- array
- positional bigram array
- nonpositional bigram array
- domain-specificity