UNIVERSITI SAINS MALAYSIA

First Semester Examination
Academic Session 2003/2004

September/October 2003

# CCS503 – Intelligent Document Processing

Duration : 3 hours

**INSTRUCTION TO CANDIDATE:**

- Please ensure that this examination paper contains **FIVE** questions in **FIVE** printed pages before you start the examination.
- Answer **ALL** questions.
- This is an Open Book Examination.
- You can choose to answer either in Bahasa Malaysia or English.

ENGLISH VERSION OF THE QUESTION PAPER

1. Given the following sentence:

"The small bat flied slowly"

(a) How many phonemes are there in the given sentence? Write down the pronunciation of the sentence given above in IPA symbols.

[15/100]

(b) Find all possible parts of speech for each word in the given sentence.

[10/100]

(c) Draw all valid English syntactic tree(s) for the given sentence.

[10/100]

(d) Provide a Context Free Grammar which is capable of generating all the syntactic tree(s) or parse tree(s) as identified in 1(c).

[15/100]

(e) Assuming an STT (Speech-To-Text) system based on the (phonemes $\rightarrow$ spelling) approach cannot differentiate the phonemes correspond to the following pairs of characters for English:

(d,t), (r,l), (n,m) and (b,p)

(i) For each word in the sentence given above, list all other English word(s) which cannot be differentiated by the STT system. Describe how an English spell checker can be used to help the STT system in recognizing all possible English words from the speech form of the sentence given above.

(ii) Describe how the grammar given in 1(d) above can be used to help the STT system in choosing the right word when recognizing the speech form of the sentence given above.

(iii) Extend the grammar given in 1(d) in order to help the STT system eliminates semantically invalid sentence(s) when recognizing the speech form of the sentence given above.

[50/100]

2. This question has two parts. Both parts MUST be answered.

When two linguistically different communities come into close contact over a long period, word borrowing is inevitable. For about 130 years, the Portuguese occupied Malacca which is now part of Malaysia. Today, only about 15,000 Portuguese descendants are said to speak creole Portuguese, which has about 5% of its vocabulary borrowed from English and Malay. Similarly, the Malay language is said to have borrowed about 450 words from Portuguese, although we only know of less than 30 (see Table 1).

Table 1. List of Words in Portuguese and in Malay

| Portuguese[+] | Malay | Portuguese | Malay |
|---|---|---|---|
| bandeira | bendera | manteiga | mentega |
| bomba | bomba | mesa | meja |
| boneca | boneka | natal | natal |
| camisa | kemeja | queijo | keju |
| carreta | kereta | ronda | ronda |
| dado | dadu | sabado | sabtu |
| escola | sekolah | saco | saku |
| garfo | garfu | sapato | sepatu |
| igrega | gereja | tangue | tangki |
| jandela | jendela | | |

[+] The meaning of the Portuguese and the Malay word are similar.

Given this data, and given the fact that no Portuguese-Malay dictionary is available, and that Portuguese-Malay speakers are a rare commodity.

(a) Write at least three types of rules to extract Malay words that are possibly of Portuguese origin? While one "rule type" handles vowel(s), another may handle consonant(s). Also, while one rule type may handle vowel(s) at the beginning of a word, another may handle vowel(s) in the middle or end of a word. Write each rule in the following format, and give an example taken from Table 1.

Rule: initial_string → final_string | condition(s)
E.g.: *Portuguese_word* → *Malay_word* | _

[Consider the transcription of English words ending in *-graphy* into Malay. We may write a rule in the following form:

Rule: *ph* → *f* | _
E.g. *geography* → *geografi* | _ ]

[90/100]

(b) Comment on the semantics of the borrowed words.

3. Given the following grammar for Bahasa Malaysia:

Ayat → FN FKK
FKK → KK
FKK → KK FN
FN → KN KS
KN → {Kata_Nama}
KS → {Kata_Sifat}
KK → {Kata_Kerja}

leksikon:
kucing : Kata_Nama
bola : Kata_Nama
main : Kata_Kerja
merah : Kata_Sifat

(a) List all sentences generated by the above grammar which are gramatical (in Bahasa Malaysia).

[15/100]

(b) List all gramatically correct sentences generated by the above grammar which are semantically invalid (in Bahasa Malaysia).

[15/100]

(c) Extend the grammar and lexicon given above so that only semantically valid sentences will be generated.

[30/100]

(d) Construct a detailed chart illustrating the parsing process of the sentence given below based on the top-down prediction with bottom-up chart parsing technique:

"Kucing merah main bola merah"

[40/100]

4. (a) For each of the following NLP tools, describe its functionality and give an example input/output pair.

(i) Spelling checker

(ii) Inflectional morphological analyzer

(iii) Part of speech tagger

(iv) Parser

[40/100]

(b) Discuss in detail how the NLP tools in 4(a) can be applied to each of the following NLP applications.

(i) Keyword-based email filter

(ii) Natural language question-answering system

(iii) Example-based Machine translation

[60/100]

- oooOooo -