UNIVERSITI SAINS MALAYSIA

Peperiksaan Semester Pertama
Sidang Akademik 2004/2005

Oktober 2004

# CCS503 – Pemprosesan Dokumen Cerdas

Masa : 2 jam

---

**ARAHAN KEPADA CALON:**

- Sila pastikan bahawa kertas peperiksaan ini mengandungi **EMPAT** soalan di dalam **TUJUH** muka surat yang bercetak sebelum anda memulakan peperiksaan ini.
- Jawab mana-mana **TIGA** soalan.
- Anda boleh memilih untuk menjawab semua soalan dalam Bahasa Malaysia atau Bahasa Inggeris.

---

1. Suatu kajian pengesanan fonem dilaksanakan untuk mengkaji ketercantuman fonem-fonem. Dalam kajian ini, beberapa orang diminta memberi transkripsi kata-kata yang mengandungi fonem-fonem (seperti yang diberikan dalam jadual di bawah). Berdasarkan beberapa kali (dalam peratusan) sesuatu fonem itu diberi transkripsi dengan betul, fonem, perseptibiliti fonem tersebut akan dikategorikan seperti berikut:

| Fonem | Perseptibiliti | Fonem | Perseptibiliti | Fonem | Perseptibiliti |
|---|---|---|---|---|---|
| /b/ | *Sangat baik* | /f/ | *Baik* | /ʧ/ | *Tidak baik* |
| /k/ | *Sangat baik* | /g/ | *Baik* | /ʤ/ | *Tidak baik* |
| /x/ | *Sangat baik* | /h/ | *Baik* | /n/ | *Tidak baik* |
| /ʔ/ | *Sangat baik* | /z/ | *Baik* | /ɲ/ | *Tidak baik* |
| /l/ | *Sangat baik* | /d/ | *Sederhana* | /ŋ/ | *Tidak baik* |
| /p/ | *Sangat baik* | /t/ | *Sederhana* | /r/ | *Tidak baik* |
| /q/ | *Sangat baik* | /v/ | *Sederhana* | /w/ | *Tidak baik* |
| /s/ | *Sangat baik* | /ʕ/ | *Kurang Baik* | /j/ | *Tidak baik* |
| | | /m/ | *Kurang Baik* | | |

Kategori perseptibiliti:

| | |
|---|---|
| *Sangat baik* | (fonem diberi transkripsi dengan betul 95 %) |
| *Baik* | (fonem diberi transkripsi dengan betul 85 %) |
| *Sederhana* | (fonem diberi transkripsi dengan betul 70 %) |
| *Kurang Baik* | (kebanyakan masa fonem tidak beri transkripsi dengan betul) |
| *Tidak baik* | (fonem tidak dapat diberi transkripsi dengan betul) |

Jawab semua soalan berikut:

(a) Dengan menggunakan jadual perseptibiliti di atas, berikan 5 perkataan dalam bahasa Inggeris yang dapat diberi transkripsinya dengan amat baik.

(10/100)

(b) Diberikan ayat berikut:

*Killer bats make clicking sounds to determine where its food might be.*

Nyatakan bilangan fonem konsonan yang boleh ditemui dalam ayat yang diberikan di atas? Berikan sebutan konsonan-konsonan ini dalam simbol IPA. Dengan merujuk kepada jadual yang diberikan di atas, isihkan perkataan-perkataan di dalam ayat ini mengikut darjah perseptibilitinya. Jelaskan secara ringkas kriteria yang digunakan untuk melakukan isihan ini.

(25/100)

(c) "The sounds corresponding to all English phonemes are powered by lung air being pushed out. A sound is then produced in two ways:

- By vibrating the vocal 'cord': two muscular folds of skin low down in the throat which can be made to vibrate. The frequency of the vibration can be changed (within limits).

- By altering the positions of the components of the throat and mouth between the vocal cords and the exit of air. These alterations may merely modify the note produced by the vocal cords (by changing the size of the cavity) or may themselves produce a noise (for example by causing air friction)."

*Sumber: Coxhead (2000) NLP/HO/Phon: 2. Production of Phonemes*

Apakah kesimpulan(-kesimpulan) yang anda dapat perolehi daripada data dalam jadual yang diberikan di atas?

(25/100)

(d) Pembina sistem-sistem *Text-To-Speech (TTS)* dan *Speech-To-Text (STT)* boleh menggunakan kaedah ejaan ↔ fonem ↔ fon, yang diimplikasikan dalam 1(b), ataupun kaedah yang lebih terus, iaitu ejaan ↔ fon. Dalam kaedah yang kedua ini, pasangan simbol fonetik, iaitu bifon, disimpan dalam suatu kamus.

Apakah kelebihan dan kekurangan setiap kaedah ini dalam pembinaan sistem TTS atau SST?

Berdasarkan cerapan yang diberikan dalam jadual di atas, apakah nasihat yang anda boleh memberi kepada pereka sistem TTS ataupun sistem STT?

(40/100)

2. Soalan ini mempunyai dua bahagian. Kedua-dua bahagian ini **MESTI** dijawab.

(a) Dengan pengetahuan sebutan vokal dan konsonan sahaja, seseorang sudah dapat membaca dalam bahasa Melayu. Penyebutan kata-kata bahasa Melayu tidak memerlukan tegas pada mana-mana satu suku kata. Sebutan kata-kata *nada* dan *pagar* adalah seperti berikut:[na-da] dan [pa-gar] ([a] disebut seperti "**a**" dalam **baca**).

Sungguhpun bahasa Sepanyol mempunyai banyak fonem yang didapati dalam bahasa Melayu, penyebutan kata dalam bahasa Sepanyol tidak sebegitu mudah seperti dalam bahasa Melayu. Dalam bahasa Sepanyol, satu suku kata (kecuali dalam kata-kata adverba yang berakhiran dengan *–mente*) perlu ditegaskan. Sungguhpun *nada* 'nothing' dan *pagar* 'to sell' wujud dalam bahasa Sepanyol, sebutannya berlainan, iaitu [NA-da] dan [pa-GAR]. Huruf besar digunakan untuk menunjukkan suku kata yang perlu ditegaskan.

Walau bagaimanapun, petua-petua yang di mana tegasan harus diletak adalah agak nalar. Di mana petua diikuti, tanda tegas ´ diletakkan pada vokal yang perlu ditegaskan.

Dari contoh-contoh yang diberikan di bawah, berikan petua-petua tegasan kata bahasa Sepanyol. Perhatikan bahawa simbol transkripsi IPA tidak digunakan.

| | | |
|---|---|---|
| *además* | 'tambahan pula' | [a-de-MAS] |
| *amigos* | 'kawan- JAMAK' | [a-MI-gos] |
| *animal* | 'haiwan' | [a-ni-MAL] |
| *aquí* | 'di sini' | [a-KI] |
| *arroz* | 'beras; nasi' | [a-ROZ] |
| *beben* | 'mereka minum' | [BE-ben] |
| *bicicleta* | 'basikal' | [bi-si-KLE-ta] |
| *calor* | 'panas' | [ka-LOR] |
| *cantan* | 'mereka nyanyi' | [KAN-tan] |
| *casa* | 'rumah' | [KA-sa] |
| *casas* | 'rumah- JAMAK' | [KA-sas] |
| *comprender* | 'faham' | [com-pren-DER] |
| *dental* | 'pergigian' | [den-TAL] |
| *día* | 'hari' | [DI-a] |
| *dormir* | 'tidur' | [dor-MIR] |
| *fantástico* | 'hebat' | [fan-TAS-ti-co] |
| *fármaco* | 'ubat' | [FAR-ma-co] |
| *felicidad* | 'kegembiraan' | [fe-li-ci-DAD] |
| *feroz* | 'garang' | [fe-ROZ] |
| *frío* | 'dingin' | [FRI-o] |
| *hablar* | 'cakap' | [ha-BLAR] |
| *hermano* | 'abang' | [er-MAN-no] |
| *hombre* | 'lelaki' | [OM-bre] |
| *importante* | 'penting' | [im-por-TAN-te] |
| *kárate* | 'karate' | [KA-ra-te] |
| *ladrón* | 'pencuri' | [la-DRON] |
| *lámpara* | 'lampu' | [LAM-pa-ra] |
| *lápices* | 'pensel-JAMAK' | [LA-pi-ses] |
| *lápiz* | 'pensel' | [LA-pis] |
| *María* | 'Maria' | [ma-RI-a] |
| *naranjas* | 'buah limau' | [na-RAN-has] |
| *noche* | 'night' | [NO-che] |
| *ojo* | 'mata' | [O-ho] |
| *pero* | 'tetapi' | [PE-ro] |
| *resumen* | 'ringkasan' | [re-SU-men] |
| *sábado* | 'Sabtu' | [SA-ba-do] |
| *salón* | 'lounge' | [sa-LON] |
| *usted* | 'tuan (formal)' | [us-TED] |
| *zapatos* | 'kasut- JAMAK' | [za-PA-tos] |

(80/100)

(b) Dua set ayat dan terjemahannya diberikan. Sekarang, tentukan bagaimana imbuhan bagi NUMBER dan PERSON diungkapkan bagi kedua-dua kata kerja ini, COMER 'makan' dan BEBER 'minum', yang berakhiran dengan –ER.

Nota: Terpulang kepada pertalian antara si penutur dan si pendengar, si penutur boleh guna salah satu antara dua bentuk kata untuk "anda" dalam bahasa Sepanyol. Bentuk TAK FORMAL digunakan jika si penutur berkenalan dengan si pendengar, dan jika tidak beberapa berkenalan, bentuk FORMAL digunakan. *Usted* = Encik/Tuan]

*¿ Qué comes* (COMER) *?* / Anda makan apa?

*Elena dice* que *coméis* (COMER) *más que nosotros.* / Elena kata bahawa anda-PLURAL makan lebih banyak daripada kami.

*Hoy comemos* (COMER) *sushi.* / Hari ini, kami makan sushi.

*Los niños* (COMER) *comen todas las frutas.* / Budak-budak itu makan kesemua buah-buahan.

*Mi gato come* (COMER) *el pescado.* / Kucing saya makan ikan.

*Mi padre come* (COMER) *en el restaurante, pero mi madre come* (COMER) *en la casa.* / Bapa saya makan di restoran, tetapi mak saya makan di rumah.

*No como* (COMER) *carne.* / Saya tak makan daging.

*Su madre dice a él: "Eres lo que comes* (COMER). / Mak berkata kepada anak: "Anda-TUNGGAL-INFORMAL adalah hasil akibat apa anda makan".

*Usted come* (COMER) *menos fibra.* / Encik-TUNGGAL-FORMAL tak makan cukup serabut.

*Bebéis* (BEBER) *dos litros de agua al día.* / Anda-PLURAL minum dua liter air sehari.

*Bebemos* (BEBER) *café por la mañana.* / Kami minum kopi awal pagi.

*Bebo* (BEBER) *café con leche caliente.* / Saya minum kopi dengan susu panas.

*El camello bebe* (BEBER) *mucha agua.* / Unta banyak minum air.

*El hombre bebe* (BEBER) *vino en su alegría.* / Lelaki itu minum arak apabila dia gembira.

*Los españoles beben* (BEBER) *agua de botella.* / Orang Sepanyol minum air botol.

*María bebe* (BEBER) *un vaso de agua.* / Maria minum segelas air.

*Si bebes* (BEBER), *no manejes.* / Kalau anda- TUNGGAL-INFORMAL minum arak, jangan pandu.

*Usted bebe* (BEBER) *mucho vino.* / Encik-SINGULAR-FORMAL terlalu banyak minum arak.

Dengan menggunakan jadual, satu untuk setiap kata kerja, isikan bentuk akhiran. Suatu contoh jadual diberikan bagi kata kerja "to drink" dalam bahasa Inggeris.

| TO DRINK | | | |
|---|---|---|---|
| NUMBER | PERSON = 1st | PERSON = 2nd | PERSON = 3rd |
| tunggal (sg) | drink-ø | drink-ø | drink-s |
| jamak (pl) | drink-ø | drink-ø | drink-ø |

(20/100)

3. Diberikan Nahu Bebas Konteks (CFG) berikut:

| | |
|---|---|
| S → NP VP | n → pineapple |
| NP → n | n → cake |
| NP → n n | n → fly |
| NP → det NP | v → likes |
| VP → v NP | v → fly |
| | det → the |

Nota: S ialah axiom ataupun simbol pemula.

(a) Kembangkan leksikon dan nahu yang diberikan di atas supaya ayat seperti yang berikut akan ditolak: "The fly like cake". Pastikan ada keserasian dari segi bilangan.

(20/100)

(b) Kembangkan leksikon dan nahu yang diberikan di 3(a) supaya ayat seperti yang berikut akan ditolak: "The cake likes pineapple". Pastikan ada keserasian dari segi bilangan dan maklumat SEM (semantik).

(20/100)

(c) Berdasarkan nahu yang diberikan di 3(b) berikan pepohon ayat "The pineapple fly likes cake". Pastikan ada persamaan dalam bilangan dan maklumat SEM.

(20/100)

(d) Berdasarkan nahu yang diberikan di 3(b), janakan suatu carta yang memperincikan proses penerbitan pepohon bagi ayat "The pineapple fly likes cake" yang berasaskan teknik "top-down prediction with bottom-up chart parsing".

(40/100)

4. (a) Untuk setiap alat NLP berikut, jelaskan fungsinya dan berikan suatu contoh input/output.

(i) Penjana "Text-to-speech"

(ii) "Summarizer"

(iii) Suatu "bitext alignment system"

(iv) Suatu "word sense disambiguation system"

(40/100)

(b) Bincangkan secara menyeluruh bagaiman alat NLP dalam 4(a) boleh digunakan untuk membangunkan aplikasi NLP berikut.

(i) Pencarian maklumat

(ii) Kerja perkamusan

(iii) Terjemahan melalui komputer

(60/100)

- oooOooo -

UNIVERSITI SAINS MALAYSIA

First Semester Examination
Academic Session 2004/2005

October 2004

# CCS503 – Intelligent Document Processing

Duration : 2 hours

---

**INSTRUCTION TO CANDIDATE:**

- Please ensure that this examination paper contains **FOUR** questions in **SEVEN** printed pages before you start the examination.
- Answer any **THREE** questions.
- You can choose to answer either in Bahasa Malaysia or English.

---

ENGLISH VERSION OF THE QUESTION PAPER

1. To know how well phonemes combine, we conducted a perceptibility study. In this study, we asked listeners to transcribe words which contained different phonemes (such as those given below). Depending on the number of times (in percentage) a phoneme is correctly transcribed; we rate the perceptibility as *Very good*, *Good*, *Fair*, *Bad* or Very *Bad*. The results are as given in the table below.

| Phoneme | Perceptibility | Phoneme | Perceptibility | Phoneme | Perceptibility |
|---|---|---|---|---|---|
| /b/ | *Very good* | /f/ | *Good* | /ʧ/ | *Very bad* |
| /k/ | *Very good* | /g/ | *Good* | /ʤ/ | *Very bad* |
| /x/ | *Very good* | /h/ | *Good* | /ɲ/ | *Very bad* |
| /ʔ/ | *Very good* | /z/ | *Good* | /ɲ/ | *Very bad* |
| /l/ | *Very good* | /d/ | *Fair* | /ŋ/ | *Very bad* |
| /p/ | *Very Good* | /t/ | *Fair* | /r/ | *Very bad* |
| /q/ | *Very good* | /v/ | *Fair* | /w/ | *Very bad* |
| /s/ | *Very good* | /ʕ/ | *Bad* | /j/ | *Very bad* |
| | | /m/ | *Bad* | | |

Perceptibility Ranking:

| | |
|---|---|
| *Very good* | (the phoneme is accurately transcribed 95 % of the time) |
| *Good* | (the phoneme is accurately transcribed 85 % of the time) |
| *Fair* | (the phoneme is accurately transcribed 70 % of the time) |
| *Bad* | (the phoneme is most of the time inaccurately transcribed) |
| *Very bad* | (the phoneme cannot be transcribed) |

Answer all of the following questions.

(a) By referring to the perceptibility table given above, give 5 words in English which are highly perceptible.

(10/100)

(b) Given the following sentence:

*Killer bats make clicking sounds to determine where its food might be.*

How many consonant phonemes are there in the given sentence? Write down the pronunciation of these consonants in IPA symbols. Sort the words in the sentence according to the level of perceptibility by referring to the table given above. Briefly, describe the criteria used to perform the sorting process.

(25/100)

(c) "The sounds corresponding to all English phonemes are powered by lung air being pushed out. A sound is then produced in two ways:

- By vibrating the vocal 'cord': two muscular folds of skin low down in the throat which can be made to vibrate. The frequency of the vibration can be changed (within limits).

- By altering the positions of the components of the throat and mouth between the vocal cords and the exit of air. These alterations may merely modify the note produced by the vocal cords (by changing the size of the cavity) or may themselves produce a noise (for example by causing air friction)."

*From: Coxhead (2000) NLP/HO/Phon: 2. Production of Phonemes*

What conclusion(s) can you draw from the data on perceptibility presented in the table above?

(25/100)

(d) Text-To-Speech (TTS) and Speech-To-Text (STT) systems can use either the method implied in (b), i.e. spelling ↔ phonemes ↔ phones, or the more direct spelling ↔ phones approach based on a dictionary storing two phonetic representations.

What are the advantages and disadvantages of each approach for both TTS and STT?

From the observations given in the table above, what advice would you give to anyone in determining the vocabulary of a TTS or STT system?

(40/100)

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

(a) In Malay, once we know how the vowels and consonants are pronounced, we can read almost without problem. There is no compulsory stress on any syllable. The words *nada* 'tone' and *pagar* 'fence' are [na-da] and [pa-gar] respectively ([a] is pronounced as "**a**" in f**a**ther).

While Spanish shares many of the phonemes in Malay, it does not share the same ease with which words are pronounced. In Spanish, stress is required on a particular syllable (except in the case of adverbs ending in *–mente*). The words *nada* 'nothing' and *pagar* 'to sell' which too exist in Spanish are pronounced as [NA-da] and [pa-GAR] respectively. The capital letters indicate the syllable that is stressed.

The rules on where to put the stress is fairly regular. Where the rule does not apply, the syllable stressed is indicated with an accent ´ over the vowel.

Now, consider the examples given below, and determine the rule(s) in Spanish on where on a word to put the stress. Note that we did not use the IPA transcription.

| | | | | | |
|---|---|---|---|---|---|
| *además* | 'furthermore' | [a-de-MAS] | *hablar* | 'to speak' | [ha-BLAR] |
| *amigos* | 'friends' | [a-MI-gos] | *hermano* | 'brother' | [er-MAN-no] |
| *animal* | 'animal' | [a-ni-MAL] | *hombre* | 'man' | [OM-bre] |
| *aquí* | 'here' | [a-KI] | *importante* | 'important' | [im-por-TAN-te] |
| *arroz* | 'rice' | [a-ROZ] | *kárate* | 'karate' | [KA-ra-te] |
| *beben* | 'they drink' | [BE-ben] | *ladrón* | 'thief' | [la-DRON] |
| *bicicleta* | 'bicycle' | [bi-si-KLE-ta] | *lámpara* | 'lamp' | [LAM-pa-ra] |
| *calor* | 'hot' | [ka-LOR] | *lápices* | 'pencils' | [LA-pi-ses] |
| *cantan* | 'they sing' | [KAN-tan] | *lápiz* | 'pencil' | [LA-pis] |
| *casa* | 'house' | [KA-sa] | *María* | 'Maria' | [ma-RI-a] |
| *casas* | 'houses' | [KA-sas] | *naranjas* | 'oranges' | [na-RAN-has] |
| *comprender* | 'to understand' | [com-pren-DER] | *noche* | 'night' | [NO-che] |
| *dental* | 'dental' | [den-TAL] | *ojo* | 'eye' | [O-ho] |
| *día* | 'day' | [DI-a] | *pero* | 'but' | [PE-ro] |
| *dormir* | 'to sleep' | [dor-MIR] | *resumen* | 'summary' | [re-SU-men] |
| *fantástico* | 'fantastic' | [fan-TAS-ti-co] | *sábado* | 'Saturday' | [SA-ba-do] |
| *fármaco* | 'medication' | [FAR-ma-co] | *salón* | 'lounge' | [sa-LON] |
| *felicidad* | 'happiness' | [fe-li-ci-DAD] | *usted* | 'you (formal)' | [us-TED] |
| *feroz* | 'fierce' | [fe-ROZ] | *zapatos* | 'shoes' | [za-PA-tos] |
| *frío* | 'cold' | [FRI-o] | | | |

(80/100)

(b) Two sets of sentences and their translations are given. Now, determine how inflection for number and person is expressed for these two verbs COMER 'to eat' and BEBER 'to drink' which end in -ER.

Note: Depending on how familiar a speaker is with the hearer, one of two possible ways of expressing "you" in Spanish will be used. The INFORMAL form is used when a speaker is familiar with the hearer, and if NOT familiar, then the FORMAL form is used. *Usted* = Mr./Sir]

*¿ Qué comes* (COMER) *?* / What do you eat?

*Elena dice* que *coméis* (COMER) *más que nosotros.* / Elena says that you-PLURAL eat more than we do.

*Hoy comemos* (COMER) *sushi.* / Today, we eat sushi.

*Los niños* (COMER) *comen todas las frutas.* / The children eat all the fruits.

*Mi gato come* (COMER) *el pescado.* / My cat eats fish.

*Mi padre come* (COMER) *en el restaurante, pero mi madre come* (COMER) *en la casa.* / My father eats in the restaurant, but my mother eats at home.

*No como* (COMER) *carne.* / I do not eat meat.

*Su madre dice a él: "Eres lo que comes* (COMER). / His mother says to him: "You-SINGULAR-INFORMAL are what you eat".

*Usted come* (COMER) *menos fibra.* / You-SINGULAR-FORMAL do not eat enough fibre.

*Bebéis* (BEBER) *dos litros de agua al día.* / You-PLURAL drink two litres of water a day.

*Bebemos* (BEBER) *café por la mañana.* / We drink coffee in the morning.

*Bebo* (BEBER) *café con leche caliente.* / I drink coffee with hot milk.

*El camello bebe* (BEBER) *mucha agua.* / The camel drinks much water.

*El hombre bebe* (BEBER) *vino en su alegría.* / The man drinks wine when he is happy.

*Los españoles beben* (BEBER) *agua de botella.* / Spanish people drink bottled water.

*María bebe* (BEBER) *un vaso de agua.* / Maria drinks a glass of water.

*Si bebes* (BEBER), *no manejes.* / If you-SINGULAR-INFORMAL drink, do not drive.

*Usted bebe* (BEBER) *mucho vino.* / You-SINGULAR-FORMAL drink much wine.

Using tables, one for each verb, fill in the inflected forms. An example table is given for the verb in English.

| TO DRINK | | | |
|---|---|---|---|
| NUMBER | PERSON = 1st | PERSON = 2nd | PERSON = 3rd |
| singular (sg) | drink-ø | drink-ø | drink-s |
| plural (pl) | drink-ø | drink-ø | drink-ø |

(20/100)

3. Given the following CFG:

| | |
|---|---|
| S → NP VP | n → pineapple |
| NP → n | n → cake |
| NP → n n | n → fly |
| NP → det NP | v → likes |
| VP → v NP | v → fly |
| | de → the |

Note: **S** is an axiom or start symbol.

(a) Extend the grammar given above in order to create a lexicon and an augmented grammar based on the feature system (for agreement in number) so that it will reject the following sentence: "The fly like cake".

(20/100)

(b) Extend the grammar given in 3(a) by enhancing the lexicon and grammar with SEM (semantic) features so that it will reject the following sentence: "The cake likes pineapple".

(20/100)

(c) Based on the grammar given in 3(b), give the parsed tree for the sentence "The pineapple fly likes cake". Show the agreement in number and SEM features.

(20/100)

(d) Based on the grammar given in 3(b), construct a detailed chart illustrating the parsing process of the sentence "The pineapple fly likes cake" based on the top-down prediction with bottom-up chart parsing technique.

(40/100)

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

(i) Text-to-speech generator

(ii) Summarizer

(iii) A bitext alignment system

(iv) A word sense disambiguation system

(40/100)

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

(i) Information retrieval

(ii) Lexicography, i.e. dictionary making

(iii) Machine translation

(60/100)

- oooOooo -