
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/	<i>Sangat baik</i>	/f/	<i>Baik</i>	/tʃ/	<i>Tidak baik</i>
/k/	<i>Sangat baik</i>	/g/	<i>Baik</i>	/dʒ/	<i>Tidak baik</i>
/x/	<i>Sangat baik</i>	/h/	<i>Baik</i>	/n/	<i>Tidak baik</i>
/r/	<i>Sangat baik</i>	/z/	<i>Baik</i>	/ŋ/	<i>Tidak baik</i>
/l/	<i>Sangat baik</i>	/d/	<i>Sederhana</i>	/ŋ/	<i>Tidak baik</i>
/p/	<i>Sangat baik</i>	/t/	<i>Sederhana</i>	/r/	<i>Tidak baik</i>
/q/	<i>Sangat baik</i>	/v/	<i>Sederhana</i>	/w/	<i>Tidak baik</i>
/s/	<i>Sangat baik</i>	/ʃ/	<i>Kurang Baik</i>	/j/	<i>Tidak baik</i>
		/m/	<i>Kurang Baik</i>		

Kategori perseptibiliti:

<i>Sangat baik</i>	(fonem diberi transkripsi dengan betul 95 %)
<i>Baik</i>	(fonem diberi transkripsi dengan betul 85 %)
<i>Sederhana</i>	(fonem diberi transkripsi dengan betul 70 %)
<i>Kurang Baik</i>	(kebanyakan masa fonem tidak beri transkripsi dengan betul)
<i>Tidak baik</i>	(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.

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

(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.

Maria 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

3. Diberikan Nahu Bebas Konteks (CFG) berikut:

$S \rightarrow NP VP$	$n \rightarrow \text{pineapple}$
$NP \rightarrow n$	$n \rightarrow \text{cake}$
$NP \rightarrow n n$	$n \rightarrow \text{fly}$
$NP \rightarrow \text{det NP}$	$v \rightarrow \text{likes}$
$VP \rightarrow v NP$	$v \rightarrow \text{fly}$
	$\text{det} \rightarrow \text{the}$

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 bagaimana 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 -