UNIVERSITY OF LONDON
IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

EXAMINATIONS 2001

MEng Honours Degrees in Computing Part IV
MSc in Advanced Computing
for Internal Students of the Imperial College of Science, Technology and Medicine

This paper is also taken for the relevant examinations for the Associateship of the City and Guilds of London Institute

PAPER C485=I4.24

NATURAL LANGUAGE PROCESSING

Wednesday 9 May 2001, 10:00
Duration: 120 minutes

Answer THREE questions

Paper contains 4 questions
Calculators not required

1 Daniel Jurafsky and James H. Martin in their book, *Speech and Language Processing*, begin with the following quotation from the screenplay *2001: A Space Odyssey* by Stanley Kubrick and Arthur C. Clarke.

Dave Bowman: Open the pod bay doors, HAL
HAL: I'm sorry Dave, I'm afraid I can't do that.

a List and briefly explain the distinct categories of linguistic knowledge that are entailed in the processing of such discourse.

b Illustrate how ambiguity may occur in processing each sort of linguistic knowledge by explaining the different interpretations of a spoken form such as: *I made her duck, Bill.*

c Explain why HAL's creator was too optimistic in predicting that such an agent would be available in 2001 by describing the present state of the art in processing at each level.

The three parts carry, respectively, 30%, 40%, 30% of the marks.

2a Briefly explain:

i) the purpose of a phrase structure rule like $S \rightarrow NP\ VP$ in language processing.

ii) the Prolog representation of a phrase structure rule expressed in Definite Clause Grammar notation like s --> np, vp.

iii) the effect of a DCG rule such as s(s(NP,VP)) --> np(NP), vp(VP).

b i) Show how a feature such as the requirement that verb and noun agree in number can be accommodated using DCG rules, validating *dogs chase cats* and *dogs chase a cat* but not **kitty chases mouse*.

ii) Briefly indicate some other features of English grammar that can be accommodated in a similar manner, and the nature of the generalisation.

c i) Briefly explain why the constituent structure of sentences like *Who was it the students thought was a tutor* are a problem to categorise with standard phrase structure rules, and give a solution using feature structures.

ii) Briefly discuss why *Jim drinks coffee in the morning and tea in the afternoon* poses a problem for even DCG phrase structure rules.

The three parts carry, respectively, 30%, 40%, 30% of the marks.

3a Using a diagram, briefly describe the noisy channel model of communication and give three areas of natural language processing to which it is pertinent in practice.

b Explain Bayes' rule for estimating the conditional probability *prob*(A/B) in terms of *prob(B/A)*, indicating how it is interpreted in the noisy channel model. Illustrate use of Bayes' rule to estimate the most likely correct word *c*, for an observed typing error *t*, given the table of estimates below:

c	*prob*(c)	*prob*(t/c)
C1	0.00003	0.0001
C2	0.00001	0.00001
C3	0.00006	0.000002
C4	0.0007	0.00006

c Explain how a large corpus of words can be used to give the statistics for such use of Bayes' rule in spelling correction, indicating the underlying assumptions and how numerical difficulties can be ameliorated.

d Briefly explain why statistical methods of analysis are needed to complement more conventional grammatical processing in the processing of co-occurrences in evolving language, and why they are unlikely to supplant it.

Parts a, b, c, and d, carry, respectively, 30%, 30%, 20%, and 20% of the marks.

4a Give a grammatical parse of *Our company was building cars* which gives the perceived meaning as an activity. Explain informally why this excludes alternative parses in which *building* is an adjective or *building car*s is a noun phrase.

b Briefly explain the idea of thematic relations for structuring knowledge of an event, illustrating their use in explicating the sentence of part a.

c Illustrate the temporal reference times of the above simple past event, comparing with the past perfective *Our company had been building cars*.

d Suppose the sentence *They were classic cars* immediately follows the above sentence in some discourse. What is the evidence for the anaphoric binding between the two sentences. How might a binding be made if the second sentence were *They were good workers* instead?

Parts a, b, c and d carry, respectively, 40%, 20%, 20%, and 20% of the marks.