10 Natural Language Processing (AAC)

The following context-free grammar (CFG) accepts sequences of part-of-speech categories (e.g., Det N, Adj Adj N). With a lexicon, as shown, it can be used to parse some English noun phrases (NPs).

```
Start symbol: NP          a, the: Det
NP -> Det N               dog, dogs, house,
NP -> N                     houses, model, models: N
N -> Adj N                brown, red, model: Adj
N -> N PP                 in, under: P
PP -> P NP
```

(*a*) Give a non-deterministic finite-state automaton (NDFSA) which accepts the same sequences of part-of-speech categories as this CFG. Explain the notation that you use. [6 marks]

(*b*) Give two examples of overgeneration that can be demonstrated with the lexicon shown, and explain how the CFG and NDFSA (and, if necessary, part-of-speech categories and lexicon) could be modified to prevent them. [5 marks]

(*c*) The CFG does not accept noun-noun compounds (e.g., *the dog house*, *house dogs*). Indicate how you could modify the original CFG and NDFSA to allow for them. [3 marks]

(*d*) Hand-constructed FSA have sometimes been used for part-of-speech tagging. Outline the possible practical and theoretical advantages and disadvantages of such an approach when compared to stochastic tagging using Hidden Markov Models. [6 marks]