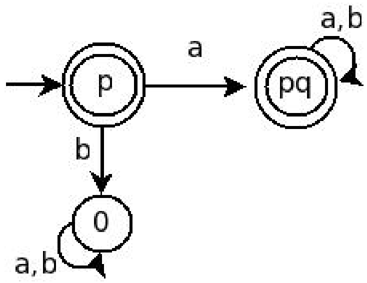
Module Title: Informatics 2A
Exam Diet (Dec/April/Aug): Dec 2015
Brief notes on answers:

1. (a) Omitting the unreachable state, the diagram is:



[4 marks for a perfect solution; deduct 1 mark per mistake.]

(b) Writing X_p, X_{pq}, X_0 for the languages associated with the three states, we see that

$$X_0 = \emptyset, \qquad X_{pq} = \epsilon + aX_{pq} + bX_{pq}, \qquad X_p = \epsilon + aX_{pq} + bX_0.$$

Solving these gives $X_{pq} = (a+b)^*$ and hence $X_p = \epsilon + a(a+b)^*$. Minor variations on this route are acceptable. [2 marks for formulating the equations, 2 marks for solving them.]

(c) `(a[ab]*)?` (or anything equivalent). [2 marks if correct, 1 mark if nearly correct.]

2. (a) $E = \{\mathsf{opts}, \mathsf{files}, \mathsf{args}\}$.

$$\begin{aligned} \mathit{First}(\mathsf{shell}) &= \{\mathit{command}\}, & \mathit{Follow}(\mathsf{shell}) &= \{\$\} \\ \mathit{First}(\mathsf{args}) &= \{\mathit{option}, \mathit{file}, \epsilon\}, & \mathit{Follow}(\mathsf{args}) &= \{\$\} \\ \mathit{First}(\mathsf{opts}) &= \{\mathit{option}, \epsilon\}, & \mathit{Follow}(\mathsf{opts}) &= \{\mathit{file}, \$\} \\ \mathit{First}(\mathsf{files}) &= \{\mathit{file}, \epsilon\}, & \mathit{Follow}(\mathsf{files}) &= \{\$\} \end{aligned}$$

[1 mark for E, 2 marks for First sets, 2 marks for Follow sets.]

(b)

	command	*option*	*file*	\$
shell	*command* args			
args		opts files	opts files	opts files
opts		*option* opts	ϵ	ϵ
files			*file* files	ϵ

[1 mark for drawing up a table with the right rows and columns; 4 marks for the entries (roughly 0.5 marks per correct entry).]

3. (a) This is not a valid HMM because the sum over the emissions for the tag N is not 1, and also the sum of the tags that we can jump from N is not 1 in the transition matrix. This indicates that the HMM tables are partial. [1 mark for each of these two points.]

(b) The chart is:

V	0	$0.1 \times 0.05 = 0.005$	$0.005 * 0.1 * 0.7 < 0.28 * 0.1 * 0.7 = 0.0196$	$0.0196 * 0.1 * 0.05 < 0.0056 * 0.5 * 0.4$
N	0	$0.7 \times 0.4 = 0.28$	$0.005 * 0.7 * 0.2 < 0.28 * 0.1 * 0.2 = 0.0056$	$0.0392 = 0.0196 * 0.5 * 0.4 > 0.0056 * 0.1 * 0.4$
$\langle$ s $\rangle$	1.0	0	0	0
		keys	open	gates

The POS sequence is N V N.

[5 marks for the table, 1 mark for correct POS sequence. Can be generous with calculation errors.]

(c) A sample of the rules are:

S $\rightarrow$ N	0.7
S $\rightarrow$ V	0.1
N $\rightarrow$ keys N	$0.4 * 0.1 = 0.04$
V $\rightarrow$ gates N	$0.5 * 0.05 = 0.025$
N $\rightarrow$ ϵ	1.0

[2 marks for sufficient evidence of understanding.]

4. (a) One such sentence is "You made her duck," as we saw in class, with the following analyses:

   – Analysis 1: Pro V Det N, which means that a person, you, made the duck that she owns.
   – Analysis 2: Pro V Pro N, which means that a person, you, made a duck for her, for example, for dinner (ditransitive meaning to the verb).
   – Analysis 3: Pro V Pro V, which means that a person, you, made her move her body quickly downwards.

   [1 mark for a suitable sentence; 1 mark each for two suitable POS taggings with explanation.]

(b) The sentence could be "I saw the boy with the telescope." There is PP-attachment ambiguity here, even though the POS tags are identical in both cases.

[1 mark for a suitable sentence; 1 mark for explanation.]

(c) The rules are:

S $\rightarrow$ NP VP	VP.sem(NP.sem)
VP $\rightarrow$ TV NP	TV.sem(NP.sem)
NP $\rightarrow$ NPR	NPR.sem
TV $\rightarrow$ borders	$\lambda x.\lambda y.$border(x, y)
NPR $\rightarrow$ Scotland	scotland
NPR $\rightarrow$ England	england

[Roughly 2 marks for the grammar, 3 marks for the semantic attachments. Can be generous if the idea is right.]

5. (a) (i) Regular, (ii) Context-free, (iii) Context-sensitive, (iv) Context-sensitive. Examples (iii) and (iv) are standard ones covered in lectures. [1 mark each]

(b) Here a summary of the proof covered in lectures is expected. If L is context free, then (modulo ϵ) it is generated by a Chomsky normal form grammar, say with n non-terminals. If k is sufficiently large, a string s of length $\geq k$ will contain a path of length $> n+1$, and some non-terminal N must appear twice along this path. This allows us to pump in additional copies of the parts of the outer N-phrase that enclose the inner N-phrase, as illustrated by the standard 'Christmas tree' diagram shown in lectures.

[2 marks for the basic diagram; 4 marks for the various additional points, being fairly generous.]

6. (a)

S → NP VP
VP → VT NP | VI
NP → D N | D N VP
D → the
N → cat | dog | rat | elephant
VT → bit | chased
VI → died

[Up to 6 marks. A grammar that generates the right sentences but with the 'wrong' structures (from the perspective of English grammar), will get most of the marks.]

(b) Let L be the above language. Let $k \geq 2$. Without loss of generality, assume k is even. We need to find $xyz \in L$ with $|y| \geq k$. Let $x = (\mathrm{D\ N})^{k/2}$, $y = \mathrm{VT}^{k/2-1}$ and $z = \mathrm{VI}$. Clearly, for any $y = uvw$ where $v \neq \epsilon$, it is not true that $xuv^2wz \in L$. As such, the language is not regular.

[5 marks for an answer along the right lines showing understanding of the pumping lemma. 8 marks for fully correct answer.]

(c) The regular language is

```
(D N)* VT* VI
```

Now we can show English is not regular. If it were regular, then the intersection of English with the above language would be also regular (as regular languages are closed under intersection). However, this intersection yields the language in (a), which, as we showed in (b), is not regular.

[Up to 4 marks for the choice of regular language; 3 marks for the remainder of the argument.]

(d) Context-freeness is not necessarily sufficient for all languages, for example, for Dutch. We need a mildly context-sensitive formalism.

[1 mark for 'no', 1 mark for an example of a language with non-context-free features, 2 marks for mentioning mildly context sensitive grammars.]

7. (a) The chart is:

1	2	3	4	5	6	7	8	
Det	NP						S	1
	N							2
		V		VP			VPP VP	3
			Det	NP			NPP	4
				N				5
					Prep		PP	6
						Adj	NP	7
							N	8

[8 marks for the chart, being generous if it is mostly right.]

(b) The two trees correspond to the following parses:

[S [NP [D the] [N dog]] [VP [V bit] [NPP [NP [D the] [N boy]] [PP [Prep with] [NP [Adj shiny] [N teeth]]]]]]

[S [NP [D the] [N dog]] [VPP [VP [V bit] [NP [D the] [N boy]]] [PP [Prep with] [NP [Adj shiny] [N teeth]]]]]]

[2.5 marks per correct tree.]

(c) The first sentence means that the dog has shiny teeth, and it used them to bite the boy. The second sentence means that the dog bit the boy, and that boy has shiny teeth.

[2.5 marks per correct interpretation.]

(d) If we set the probability of the rule VP $\rightarrow$ NPP to 0, then however we set the other probabilities (as long they are valid, and none of them are zero), we will get the wanted interpretation.

[5 marks for any correct PCFG.]

(e) The probability will be the same probability as for the original sentence with the probabilistic grammar. The reason this is true is that probabilistic CFGs treat the derivation as a bag of rules in order to assign probability to a derivation.

[2 marks for explaining this correctly in terms of bag of rules.]

8. (a) The execution is as follows (note that for any q, q' there's at most one transition $q \longrightarrow q'$, except when $q = q' = q_3$).

Transition	Remaining input	Stack
	$aabaabba$	$\perp$
$q_0 \longrightarrow q_0$	$abaabba$	$*\perp$
$q_0 \longrightarrow q_1$	$baabba$	$*\perp$
$q_1 \longrightarrow q_1$	$aabba$	$\perp$
$q_1 \longrightarrow q_2$	$aabba$	$\perp$
$q_2 \longrightarrow q_2$	$abba$	$*\perp$
$q_2 \longrightarrow q_2$	bba	$**\perp$
$q_2 \longrightarrow q_3$	ba	$**\perp$

...followed by three transitions $q_3 \longrightarrow q_3$ as necessary to end the computation. [5 marks for the basic idea with the key transitions $q_0 \longrightarrow q_1 \longrightarrow q_2 \longrightarrow q_3$ in the right place; 5 marks for the remaining details.]

(b) The $(i+n)$th symbol has to be b [1 mark]. This is because the string has the form $tauvbw$ where $|t| = |u|$ and $|v| = |w|$, whence $|uv| = n - 1$ [2 marks].

(c) In any string ww where $|w| = n$, if the transition $q_0 \longrightarrow q_1$ is made on the ith symbol then the ith symbol of w is a, whence the $(i+n)$th symbol of ww is also a, contradicting the above. [Full 2 marks for even a glimpse of the right idea.]

(d) A suitable grammar (with start symbol S) would be

$$\begin{array}{rcl} S & \longrightarrow & TU \\ T & \longrightarrow & a \mid XTX \\ U & \longrightarrow & b \mid XUX \\ X & \longrightarrow & a \mid b \end{array}$$

Note that this can be done independently of parts (b) and (c). [6 marks; roughly 1 mark per production.]

(e) Adding the rule $S \longrightarrow UT$ will let it generate all strings with b at some position i and a at position $(i+n)$ [2 marks]. Adding $S \longrightarrow T \mid U$ will let it generate all strings of odd length [2 marks]. This now covers all strings not of the form ww.

Note: this question guides them through the proof that $\{a, b\}^* - \{ww \mid w \in \{a, b\}^*\}$ is context-free, which was outlined briefly in lectures.