IHLT Exam – 2019 (Start time 11:00; End time 13:00)

1. (4 points) We are asked to build a morphological analyser able to provide us with a list of pairs ***lemma*+PoS_tag** for each form in an input sentence such as in the following example:

Input sentence:	Two	flies	are	controlling	our	oxen
Output pairs:	*two*+CD	*fly*+NNS *fly*+VBZ	*be*+VBP	*control*+VBG	*our*+PRP$	*ox*+NNS

a) Draw a FST that computes morphotactics (lexical forms) to analyse all forms being CD, NN, NNS, VBP, VBZ and VBG associated to lemmas *two*, *fly*, *be*, *control* and *ox*.

CD: number;

NN: singular common noun;

NNS: plural common noun;

VBZ: verb in present 3th person indicative;

VBP: verb in present not 3th person indicative;

VBG: verb gerund;

b) Draw a FST that computes the spelling rules (surface forms) required for the analysis of all the previous lexical forms.

c) Explain the process to get the analysis of word *flies*.

2. (3 points) Suppose we have the following CFG:

(1) S → NP VP	(6) PP → PREP NP	(12) NP → Mary
(2) NP → NP PP	(7) DT → the	(13) VB → showed
(3) NP → DT NN	(8) DT → a	(14) VP → showed
(4) VP → VP PP	(9) NN → man	(15) PREP → to
(5) VP → VB NP	(10) NN → book	(16) PREP → of

and the following input sentence:

The man showed a book to Mary

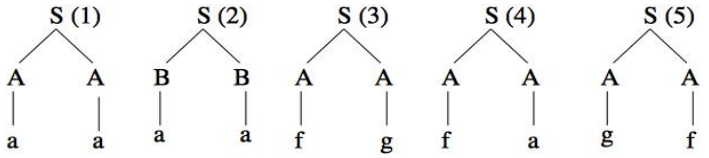
a) Can we execute algorithm CKY with that grammar? If not, provide the necessary for the execution. Justify all your answers.

b) Provide the whole dynamic table produced by that execution.

c) Draw all the possible parse trees resulting from the previous process. Explain how you get them from the dynamic table and describe how the syntactic ambiguities of the resulting trees are represented in that table.

3. (3 points) Answer the following short questions.

a) Consider as training corpus a treebank containing the following trees:



Suppose that (1) appears 75 times in the training corpus, (2) occurs 10 times, (3) occurs 325 times, (4) appears 8 times and (5) appears 428 times.

What SCFG would one get from this treebank (using MLE)? Given the obtained grammar, which is the most likely parse of the string “a a”? Is it a reasonable result? Why?

b) Wordnet synsets for noun *tongue* are:

(1) S: (n) **tongue**, lingua, glossa, clapper (a mobile mass of muscular tissue covered with mucous membrane and located in the oral cavity)
(2) S: (n) natural language, **tongue** (a human written or spoken language used by a community; opposed to e.g. a computer language)
(3) S: (n) **tongue**, knife (any long thin projection that is transient) *"tongues of flame licked at the walls"; "rifles exploded quick knives of fire into the dark"*
(4) S: (n) **tongue** (a manner of speaking) *"he spoke with a thick tongue"; "she has a glib tongue"*
(5) S: (n) spit, **tongue** (a narrow strip of land that juts out into the sea)
(6) S: (n) **tongue** (the tongue of certain animals used as meat)
(7) S: (n) **tongue** (the flap of material under the laces of a shoe or boot)
(8) S: (n) clapper, **tongue** (metal striker that hangs inside a bell and makes a sound by hitting the side)

Use Simplified Lesk Algorithm to disambiguate the noun *tongue* in the following sentence:

I’m getting used to eating cow ***tongue***

Which is the resulting best Wordnet synset? Provide the values achieved by each synset. What extra information or resources would you use to get the correct synset?

c) Is it possible to build a PoS tagger using CRFs? If so, justify the answer, provide a relevant feature template and two features derived from it. If not, propose and describe a valid machine learning model.