1) a) i) it's exponential average of gradients so the update step is actually from some previous updates so it's not instantly. so because updates are not fast, the learning rate would be more stable

ii) Model parameters with small gradients will get larger updates
also big gradients will cause smaller updates.
so we can say this way normalize the gradient in every update step.

b) i) $E_{P_{drop}}[h_{drop}]_i = P_{drop} * h_{drop} + (1-P_{drop}) * h_{drop}$

$= 0 + (1-P_{drop}) \times \gamma \times h_i$

$E_{P_{drop}}[h_{drop}]_i = h_i$

$\Rightarrow (1-P_{drop}) \times \gamma \times h = h$

$\rightarrow \gamma = \frac{1}{1-P_{drop}}$

ii) As some neurons are not involved in training we use dropout but in evaluation we want all information so it's not good idea to use dropout.

2-a)

[ROOT]	[I, attended, lectures, in, the, NLP, class]	—	initialization
[ROOT, I]	[attended, lectures, in, the, NLP, class]	—	shift
[ROOT, I, attended]	[lectures, in, the, NLP, class]	—	shift
[ROOT, attended]	[Lectures, in, the, NLP, class]	attended → I	left-Arc
[ROOT, attended, lectures]	[in, the, NLP, class]	—	shift
[ROOT, attended]	[in, the, NLP, class]	attended → lectures	Right-Arc
[ROOT, attended, in]	[the, NLP, class]	—	shift
[ROOT, attended, in, the]	[NLP, class]	—	shift
[ROOT, attended, in, the, NLP]	[class]	—	shift
[ROOT, attended, in, the, NLP, class]	[]	—	shift
[ROOT, attended, the, NLP, class]	[]	class → in	left-Arc
[ROOT, attended, NLP, class]	[]	class → the	left-Arc
[ROOT, attended, class]	[]	class → NLP	left-Arc
[ROOT, attended]	[]	attended → class	Right-Arc
[ROOT]	[]	ROOT → attended	Right-Arc

b) 2n steps . n steps for shift operations and n step for left-Arc or Right-Arc operations

f) i) Error type : verb Phrase Attachment Error
Incorrect dep: acquisition → citing
correct dep: blocked → citing

ii) Error type: Modifier Attachment Error
Incorrect dep: left → early
correct dep: afternoon → early

iii) Error type: Prepositional Phrase Attachmat Error
Incorrect dep: declined → decision
correct dep: reasons → decision

iv) Error type: Coordination Attachment Error
Incorrect dep: affects → one
correct dep: plants → one

g) Using part of speech tags as features in a parser can improve the accuracy and efficiency of parsing by providing information about syntactic and semantic structure of the sentence. It helps with the words which have multiple meanings.

e) the best dev UAS: 88.06
the best test UAS: 88.04