1ai

Rolling average does a kind of smoothing over the gradients which prevent small changes in gradients to affect the whole parameters.

1aii

It has a normalization effect by reducing the size of large gradients and amplifying the size of small gradients.

2ai

y=1/p_drop

because during the training p_drop percent of inputs turns off, and during testing all of them are on, we need to scale back the values during the testing. So if we drop with probability of p_drop during training we intensify with 1/p_drop during testing.

2aii

Dropout act as a regularization technique during training by forcing the model to use all the neurons to make the final prediction. But in testing phase we need a stable model which produce the same result if we pass one sample multiple times. So, it must not have any notion of randomness.

2a

Stack	Buffer	New dep	transition
Root	I, parsed, this, sentence, correctly		Init config
Root, I	parsed, this, sentence, correctly		Shift
Root, I, parsed	this, sentence, correctly		Shift
Root, parsed	this, sentence, correctly	Parsed -> I	Left arc
Root, parsed, this	Sentence, correctly		Shift
Root, parsed, this, sentence	Correctly		Shift
Root, parsed, sentence	Correctly	Sentence -> this	Left arc
Root, parsed	Correctly	Parse -> sentence	Right arc
Root, parsed, correctly			Shift
Root, parsed		Parsed -> correctly	Right arc
Root		Root -> parsed	Right arc

2b

Every word pushed to stack once and popped out once at max, and one step for root parsing. Having n words in a sentence, at maximum it takes 2n+1 steps to parse it.

2c, 2d, 2e) code.

2fi

Verb phrase attachment error

wedding -> fearing should be heading -> fearing

2fii

Coordination attachment error

makes -> rescue should be rush -> rescue

2fiii

Prepositional phrase attachment error

named -> midland should be guy -> midland

2fiv

Modifier attachment error

elements -> most should be crucial -> most