Question 1

1)

did you manage to finish the third obligatory assignment due yesterday? --> directive

I sent it last week already! --> assertive

Impressive! --> expressive

Could you send me your solutions? --> directive

Sure, I'll send you the PDF --> commissive

2)

Yes, the second utterance ("I sent it last week already") is a conversational implicature. The utterance does not directly answer the preceding question. However, if we assume person 2 is cooperative and adheres to the maxim of relation ("be relevant"), we can nevertheless make sense of the utterance by understanding that, if person 2 has sent the obligatory assignment, it means the assignment was finished.

3)

Yes, the common ground evolves during the interaction. At the start of the dialogue, the common ground does not include the fact that person 2 has already submitted the assignment, but this element is added after person 2's statement. At the end of the dialogue, the common ground also contains the fact the promise to send the PDF.

Question 2

“the ball” is the reparandum, “no sorry” is the interregnum, and “the box” is the repair.

Question 3

1)

There are probably multiple possible answers, but the one that seems the most appropriate for me is to use an NLU-based chatbot that combines intent recognition with slot-filling (hours, places etc.). It’s also possible to use a handcrafted system, although one would then need rules to detect the user intent and the possible slots.

An IR-based chatbot would be wrong here. There is no way a chatbot could derive the right answer for an utterance such as “what is the current time in Buenos Aires” from a fixed dialogue corpus. As for sequence-to-sequence, they would suffer from the same limitations as IR-chatbots (it would be technically possible to implement such a chatbot by implementing a "knowledge-grounded" seq2seq model which relies on a knowledge base to provide answers, but it's not something we have covered during the course).

2):

For IR-based and sequence-to-sequence chatbots, the data would correspond to a dialogue corpus. For rule-based systems, there is no absolute need for a dataset, although it is often useful to get inspiration to write rules that cover as many possible utterances as possible. For an NLU-based chatbot, the dataset would correspond to a labelled dataset where each utterance is associated with an intent. In addition, the utterances should be labelled with entities corresponding to the slots (for instance, "Buenos Aires" should be a location), at least if we are not reusing an existing NER model.

3):

At least three components would be necessary:

- A component (either rule-based or using a data-driven classification model) that detects the general user intent
- A component (based on handcrafted patterns, dictionaries, NER models or similar sequence labelling scheme) for detecting slots such as times and places
- A component for selecting/generating the response based on the intent, recognized slots, and an external knowledge source (for timezones etc.)

The two first components are typically part of what we call NLU, while the last one corresponds to response selection / NLG.

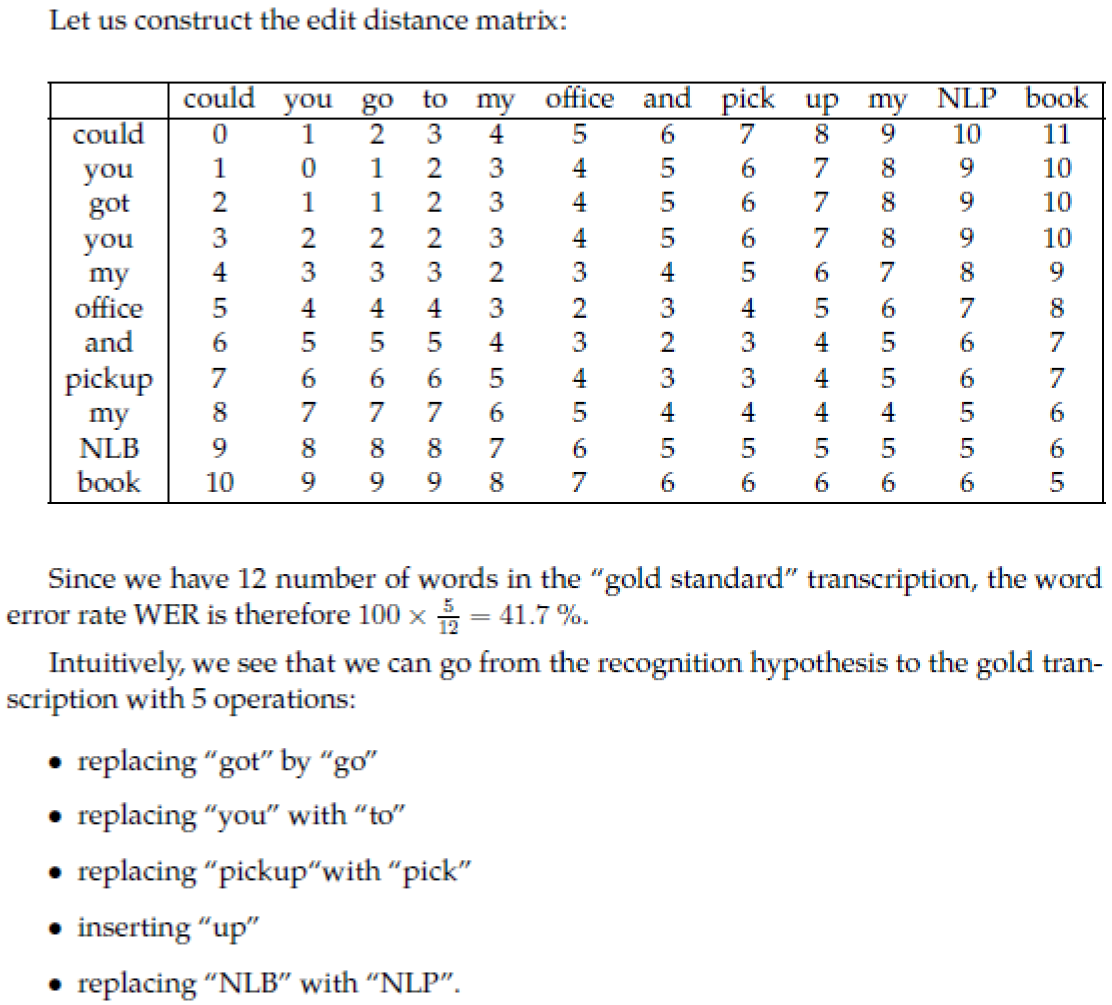
Alternatively, if the student has decided to go for a seq2seq model, we would need an external knowledge graph that can be "attended to" by the decoder when generating the response.

4)

For the NLU-based model (with slot-filling), we can first evaluate the performance of the intent recognition and slot-filling using standard metrics (accuracy, precision, recall, F-score), since we would have a labelled dataset at our disposal. This could also be done for handcrafted systems, provided we have collected a labelled corpus that can be employed for such an evaluation.

But that is not all. We also need to evaluate the quality of the system responses. For this, a human evaluation would be preferable, where human annotators provide scores on the quality of the responses, often among several axes. There are also some automated dialogue metrics, although they would be relatively difficult to apply in our case. Since we already know which answer is the correct one (there can be only one correct time in Buenos Aires at a given moment), one could also design an automated metric tailored to this task, and that checks whether the system response contains the correct time.

Question 4



Let us construct the edit distance matrix:

	could	you	go	to	my	office	and	pick	up	my	NLP	book
could	0	1	2	3	4	5	6	7	8	9	10	11
you	1	0	1	2	3	4	5	6	7	8	9	10
got	2	1	1	2	3	4	5	6	7	8	9	10
you	3	2	2	2	3	4	5	6	7	8	9	10
my	4	3	3	3	2	3	4	5	6	7	8	9
office	5	4	4	4	3	2	3	4	5	6	7	8
and	6	5	5	5	4	3	2	3	4	5	6	7
pickup	7	6	6	6	5	4	3	3	4	5	6	7
my	8	7	7	7	6	5	4	4	4	4	5	6
NLB	9	8	8	8	7	6	5	5	5	5	5	6
book	10	9	9	9	8	7	6	6	6	6	6	5

Since we have 12 number of words in the "gold standard" transcription, the word error rate WER is therefore $100 \times \frac{5}{12} = 41.7\ \%$.

Intuitively, we see that we can go from the recognition hypothesis to the gold transcription with 5 operations:

- replacing "got" by "go"
- replacing "you" with "to"
- replacing "pickup"with "pick"
- inserting "up"
- replacing "NLB" with "NLP".

Question 5

1)

Yes, it's possible to have a WER larger than 100%, for instance if there are many insertions (the edit distance between the recognition hypothesis and the gold standard utterance may be larger than the number of words in the gold standard).

2)

When training speech recognition models, we are typically provided with speech recordings together with their gold standard transcriptions. But there is no 1:1 correspondence between the speech inputs (audio frames of short duration, such as 50 ms) and the transcribed words or phonemes, which are substantially fewer. We don't initially know to which part of a phoneme/letter belongs a given audio frame. So we typically need to *infer* this alignment when training a speech recognition model (there are many ways to implement this kind of inference, but that is beyond the question).

Question 6

1) the fundamental frequency F_0 is lowest frequency of the sound wave

2) it corresponds to the speed of vibration of the vocal folds

3) the fundamental frequency correlates with the pitch, which gives us the intonation of the speech signal. And intonation is important for a spoken dialogue system - for instance, a rising intonation at the end of a signal will indicate an interrogative utterance in many language