Section 1A: Linguistics

Question 1A: How many tokens are there in this sentence? Explain your reasoning.

Boris Johnson has been the prime minister of the UK since last year.

The number of tokens in a sentence is the exact number of words that the sentence have. In this case, this sentence has 13 words, hence 13 tokens.

Question 1B: Fill in the blanks in these statements

- government and govern are morphologically related through derivation.
- minister and ministers are morphologically related through inflection.
- road and road map are morphologically related through compounding.

Question 1C: Describe in your own words the difference between a parallel and comparable corpus. Give an example of an NLP application that uses such a corpora.

A parallel corpus is one document that has been translated to multiple languages.
A comparable corpus is a collection of documents ~~that have~~ from the same topic that ~~has been~~ are in different languages.

In both cases, we can use such a corpora to train translation models.

• Section 2: Parsing

Question 2A: Consider the following grammar

S → NP VP 1.0	D → the 0.5	Prp → with 1.0
NP → D N 0.4	D → a 0.5	
NP → N 0.5	N → coffee 0.3	
NP → Prn 0.1	N → function 0.7	
VP → V 0.3	V → function 0.6	
VP → Aux VP 0.2	V → can 0.4	
VP → Adv VP 0.1	Prn → I 1.0	
VP → V PP 0.4	Aux → can 1.0	
PP → Prp NP 1.0	Adv → only 1.0	

What is the probability of the following sentence in this grammar?

"I can only function with coffee"

Show which rule in the grammar were used in the parse of this tree

Let's first convert this grammar to Chomsky format:

S → NP VP 1.0	VP → can 0.12	Aux → can 1.0
NP → D N 0.4	NP → I 0.1	Adv → only 1.0
VP → Aux VP 0.2	D → the 0.5	Prp → with 1.0
VP → Adv VP 0.1	D → a 0.5	
VP → V PP 0.4	N → coffee 0.3	
PP → Prp NP 1.0	N → function 0.7	
NP → coffee 0.15	V → function 0.6	
NP → function 0.35	V → can 0.4	
VP → function 0.18	Prn → I 1.0	

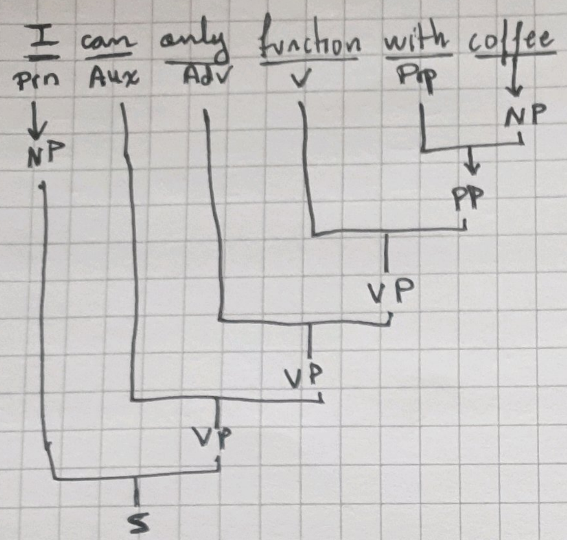
Now, we can use CYK algorithm to calculate the parse with the most likelihood in this language.

	I	CAN	ONLY	FUNCTION	WITH	COFFEE
1	Prn: 1.0, NP: 0.1	Aux: 1.0, V: 0.4, VP: 0.12	Adv: 1.0	N: 0.7, NP: 0.35, V: 0.6, VP: 0.18	Prp: 1.0	N: 0.3, NP: 0.15
2		S: 0.012	–	VP: 0.018	-	PP: 0.15
3			–	VP: 0.0036	–	VP: 0.036
4				S: 0.00036	–	VP: 0.0036
5					–	S: 0.00036

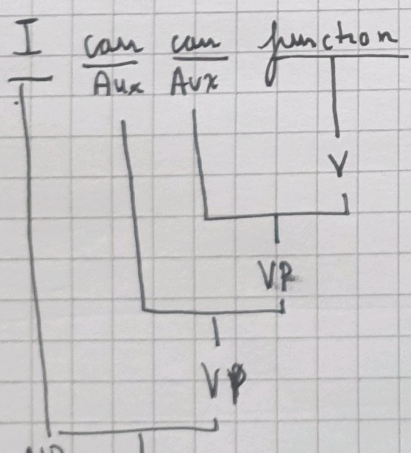
• Question 2A:

Therefore, the sentence "I can only function with coffee" has a likelihood of 0.00036 of being in this grammar.

The parse tree would be as it follows:



• Question 2B: Using the grammar of question 2A, find a sentence that is accepted by the grammar but is not grammatical in English and suggest a modification to the grammar so that this sentence is not generated



This sentence is not correct in English but it is correct in this grammar. This can be fixed by changing the production rule VP → Aux VP to VP → Aux V.

- Question 2C: why do lexical dependencies cause an issue with a simple Probabilistic Context-Free Grammar (PCFG) approach to parsing?

In a CFG the expansion of one non-terminal is independent of any other non-terminal, but some studies show that we need to distinguish between NP in subject and direct object positions. This can't be capture by the probabilities in the grammar.

- Section 3: Semantics

Question 3A: Fill in the blanks in these statements on words that are semantically related:

- government, cabin, administration are synonyms
- lixe and dark are antonym

Question 3B: Explain in your own words how word senses are represented in WordNet. Give an example.

WordNet uses synsets to express word senses. ~~These~~ Synsets are a list of the different definitions that a word can have:

Table:

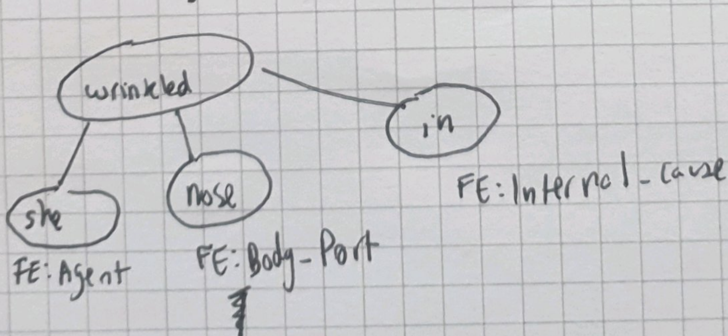
- A set of data arranged in rows and columns
- A piece of furniture having a smooth flat top and usually supported by one or more vertical legs.

Question 3C: Explain in your own words how words are represented in FrameNet. Give an example.

The main idea of FrameNet is that the meanings of most words can best be understood on the basis of a semantic frame, a description of a type of event, relation, or entity and the participants of it.

A frame is a schematic representation of a situation involving various participants, props, and other conceptual roles. A frame in FrameNet contains a textual description of what it represents (a frame definition), associated frame elements, lexical units, example sentences, and frame-to-frame relations.

Example:



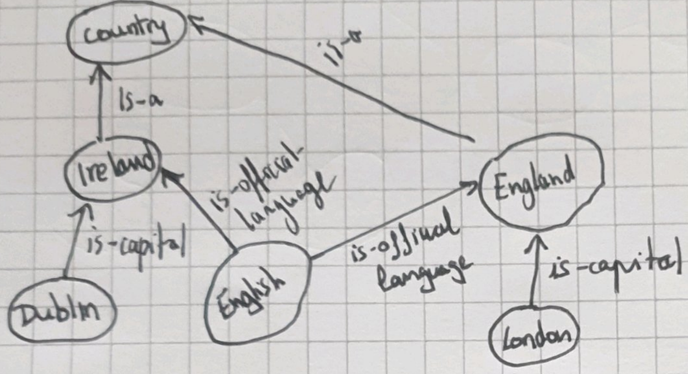
Question 3D: How can Wikipedia be used in word sense disambiguation?

Wikipedia can be used to apply semi-supervised training by using distant supervision on its categories.

• Section 4: Applications

Question 4A: Explain in your own words how a knowledge ~~graph~~ model can be used in information extraction. Give an example.

Knowledge models are based on knowledge graphs. A knowledge graph is a graph (directed) that has words as nodes and relationships as edges. This relationships can be used ~~[illegible]~~ for information extraction in a given sentence. Example:



Using this knowledge graph, we can extract the following relationships from the sentence above:

English is the main language in London, England.

(London, is-capital, England), (English, is-official-language, London)

However, one of the main usage of knowledge graphs is being able to identify missing relations.

Question 4B: Consider ~~[illegible]~~ Pointwise Mutual Information (PMI). Given words a,b,c explain how PMI(a,b) for a given corpus can be higher than PMI(a,c).

$$PMI(a,b) = \frac{P(a,b)}{p(a)p(b)} < \frac{p(a,c)}{p(a)p(c)} = PMI(a,c)$$

In case that a occurs more often with c than with b we would have $P(a,b) < P(a,c)$

If b and c occurs the same number of times in that corpus then $p(b) = p(c)$.

• Question 4C: Discuss a limitation of a lexicon-based approach to sentiment analysis

A lexicon is a dataset of words together with a sentiment classification for each word.

Example:

Happy	Sad
Smile	Cry
Laugh	Disappointed

This can be used to classify the whole meaning of a sentence into one of these categories. However, complexities such as the usage of positive words to express a negative meaning (e.g. sarcasm) can be challenging and lead to a misclassification of the sentence