**Section 1: Linguistic Foundations**

**Question 1A (10 Marks)**

Define a constituency (phrase) grammar and lexicon that analyze the following sentence by using the non-terminal symbols ‘S, NP, VP, PP’ and the pre-terminal symbols ‘Det, Noun, Verb, Prep’.

“*The Taoiseach provided a long answer to questions by TDs”.*

**Answer:**

**Grammar Rules:**

S -> NP VP

NP -> Det Noun

VP-> V NP

NP-> NP PP

NP->Det NP

NP->Noun Noun

PP-> Prep NP

NP-> Noun PP

PP ->Prep Noun

**Lexicon:**

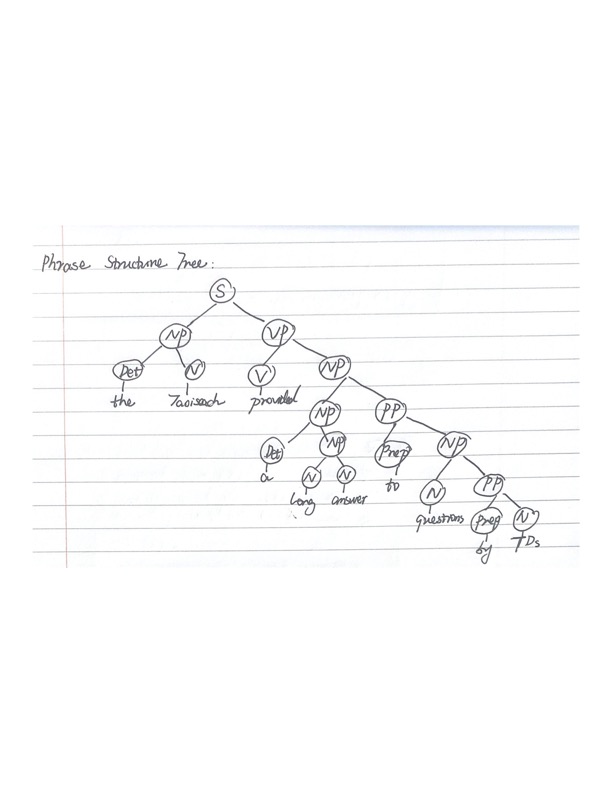
Noun -> **Taoiseach**,answer, questions, TDs

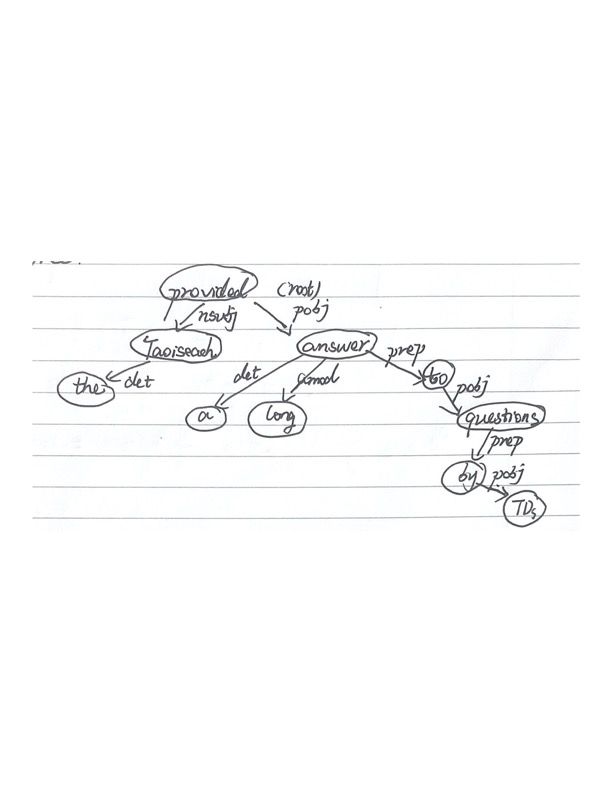
Adjective->long  
Verb -> provided [“to provide “, 3rd, pres]

Preposition -> to, by  
Determiner -> the, a

**Question 1B (10 Marks)**

Draw a constituency (phrase) structure tree and a dependency tree by using the relations ‘nsubj, pobj, amod, det, prep’ for the sentence given in question 1A.

****

****

**Question 1C (5 Marks)**

How many types and tokens are there in the sentence given in question 1A?

Types (unique words): 10

Tokens (all words): 10

**Section 2: Language Modelling**

Consider the following corpus:

*flies fly behind flies then more flies try to fly further behind*

**Question 2A (5 Marks)**

State the formula for a bigram language model.

**Diagram, text

Description automatically generated**

P(w1w2…wn)=p(w1)p(w2|w1)…p(wn|wn-1)

**Question 2B (5 Marks)**

Using a bigram language model without smoothing, calculate the probability of the sentence “flies fly further”. You should use the corpus above to estimate probabilities.

**Answer: Bigram counts**

|  |  |  |  |
| --- | --- | --- | --- |
| flies fly | 1 | fly behind | 1 |
| behind flies | 1 | flies then | 1 |
| then more | 1 | more flies | 1 |
| flies try | 1 | try to | 1 |
| to fly | 1 | fly further | 1 |
| further behind | 1 |  |  |

**Probabilities**

P(flies)=3/12=0.25

P(fly|flies)=c(flies fly)/(c(flies then) +c(flies fly)+c(flies try))=1/3=0.33

P(further|fly)=c(fly further)/c(fly further)+c(fly behind)=1/2=0.5

|  |  |  |  |
| --- | --- | --- | --- |
| P(fly|flies) | 0.33 | P(behind|fly) | 0.5 |
| P(flies|behind) | 1 | P(then|flies) | 0.33 |
| P(more|then) | 1 | P(flies|more) | 1 |
| P(try|flies) | 0.33 | P(fly|to) | 1 |
| P(further|fly) | 0.5 | p(behind|further) | 1 |
| P(to|try) | 1 |  |  |

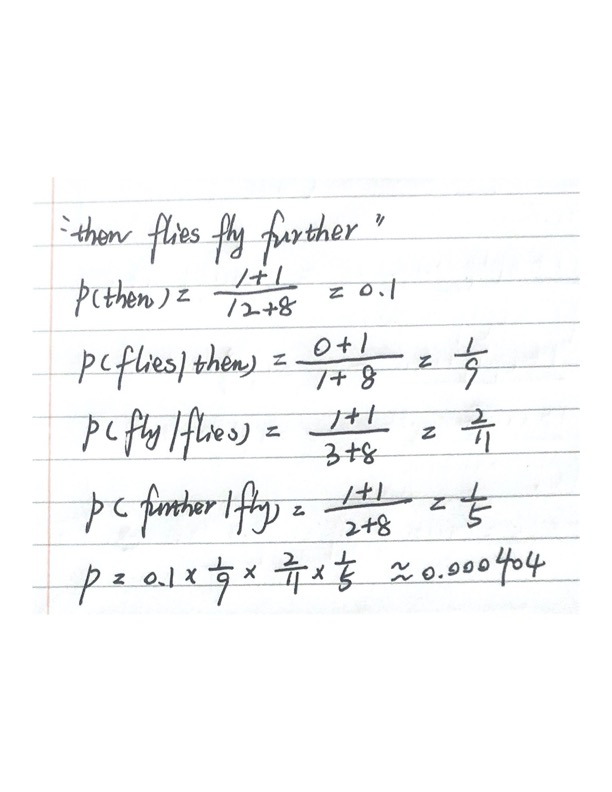
P(flies fly further)=p(flies)\*p(fly|flies)\*p(further|fly)=0.25\*0.33\*0.5=0.04125

**Question 2C (5 Marks)**

Using a bigram language model with add-one smoothing, calculate the probability of

the sentence “then flies fly further”.

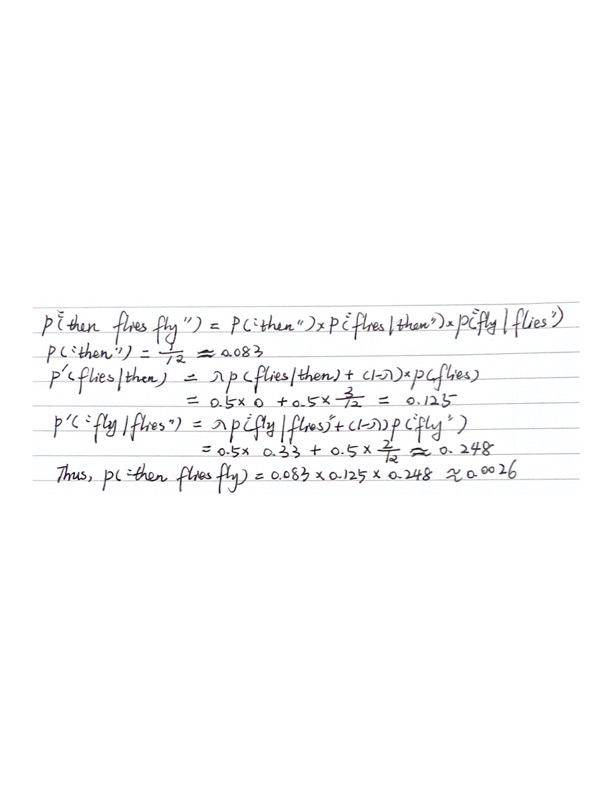
**Note that P(flies|then)=(0+1)/(1+unique words) 分母不是加unique bigram phrase**



**Question 2D (5 Marks)**

Recall the formula for bigram interpolation: A picture containing logo

Description automatically generated



**Question 2E(5 Marks)**

Why may a language model be used in a machine translation system?

**Answer:**

In machine translation, we wish to get the best translation given a foreign text, say, we wish to find **argmax(t|f)** where t is translation and f is foreign text.

By Bayes’s Law:

P(t1|f)=p(f|t1)\*p(t1)/p(f)

P(t2|f)=p(f|t2)\*p(t2)/p(f)

It shows that P(t1|f)> P(t2|f) **if** p(f|t1)\*p(t1)> p(f|t2)\*p(t2)

Thus, to find the maximum p(t|f) is equivalent to find:

**argmax p(f|t)\*p(t)**

where p(f|t) is translation model, p(t) is the language model.

This is how we use a language model in a machine translation system.

**Section 3: Parsing**

**Question 3A (5 Marks)**

Consider the following probabilistic grammar. Describe one ambiguity when applying the above grammar to the sentence “natural language processing works”.

**Answer:**

One ambiguity is:

“Natural language processing works A N N N”

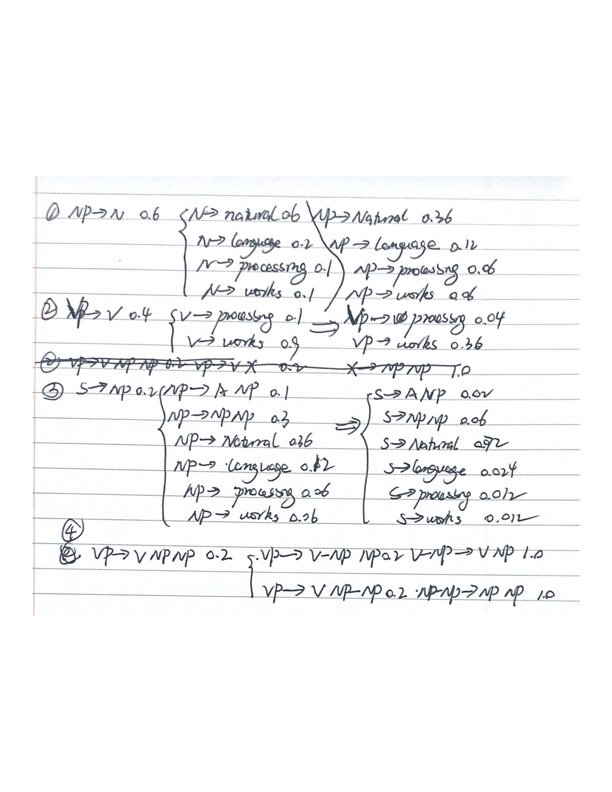
**Question 3B (5 Marks)**

What changes would be necessary to convert the above grammar into Chomsky normal form?

**Answer:**

A context free grammar (CFG) is in Chomsky Normal Form (CNF) if all production rules satisfy one of the following conditions:

* A non-terminal generating a terminal (e.g.; X->x)
* A non-terminal generating two non-terminals (e.g.; X->YZ)
* Start symbol generating ε. (e.g.; S-> ε)



**Question 3C (5 Marks)**

Why should a grammar be in Chomsky normal form when applying the CYK algorithm?

**Answer:**

It is not practical to find all parse trees (exponential number). However, CYK is more efficient and easier to implement.

* If a grammar is in CNF, then we can avoid the ambiguity problem during parsing.
* Another benefit of CNF is that it provides an upper bound for parsing complexity. For any string(sentence) w with length n, the derivation has 2n−1 steps. Thus, by an exhaustive search of all derivations, one can determine if a string is in the language.

**Question 3D (5 Marks)**

What is a cross-bracketing error and why may it not be important in the example of Q3A?

**Answer:**

This is where the parser outputs ((A B) C) but the candidate is (A (B C)).

In Q3A, if there is a cross-bracket error, the sentence should be “Natural language processes works”. Thus, “processing” is different from “process” that it cannot be a predicate in this case.

**Section 4: Distributional Semantics**

Consider the following corpus:

*A black cat chased the white cat.*

*The black dog chased the white dog.*

*A white dog chased the white cat.  
A white dog chased the black dog.*

*The white cat chased a black cat.*

*The white cat chased a white dog.*

**Question 4A (15 Marks)**

Construct a co-occurrence matrix for all types in the corpus, using a context window of two words. 这个题做错了，a window of 2，前后各两个词

**Question 4B (10 Marks)**

Using Cosine Similarity, compute the distance between:

●  black, white

●  cat, dog

**Answer:**

