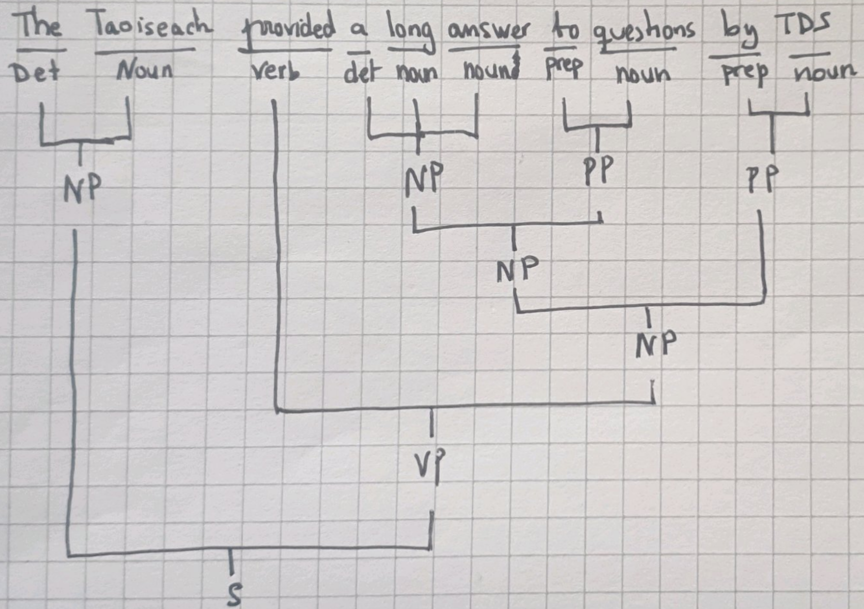
• Section 1: Linguistic Foundations

Question 1A: Define a constituency (phrase) grammar and lexicon that analyses the following sentence by using the non-terminal symbols 'S, NP, PP, VP' and the pre-terminal symbols 'Det, Noun, Verb, Prep'.

"The Taoiseach provided a long answer to questions by TDS"
Det Noun verb det noun ~~det~~ prep noun prep noun

Det → The | a
Noun → Taoiseach | answer | questions | TDS | long*
Verb → provided
Prep → to | by
PP → Prep + Noun
VP → Verb + NP
NP → Det + Noun | Det + NP | Det + Noun + Noun | NP + PP |
S → NP + VP

* while long would be an Adjective in English in this consituency grammar is a noun.



Question 1B: Draw a consitvency (phrase) structure tree and dependency tree by using the relations "nsubj", pobj, amod, det, prep for the sentence given in 1A.

The Taoiseach provided a long answer to questions by TDS

nsubj: nominal subject → A nominal subject is ~~[illegible]~~ a noun phrase which is the syntactic subject of a clause

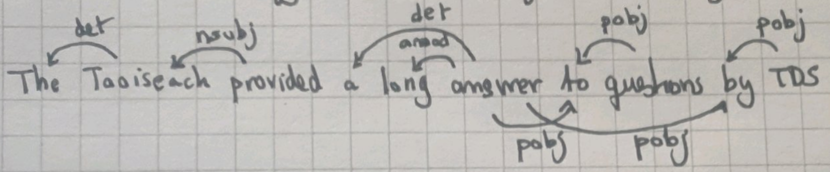
pobj: object of a preposition → The object of a preposition is the head of a noun phrase following the preposition, or the adverbs "here" and "there".

amod: Adjectival modifier → An adjectival modifier of an NP is any adjectival phrase that serves to modify the meaning of the NP

det: A determiner is the relation between the head of an NP and its determiner

• Question 1B

prep: prepositional clausal modifier → is a clause introduced by a preposition which serves to modify the meaning of the verb, adjective, or noun.



• Question 1C: How many types and tokens are there in the sentence given in question 1A?

Token: collection of individual words (with repetition)

Types: set of tokens (no repetition)

Token number = 10, Types number = 10

. Section 2: Language Modelling

Consider the following corpus "flies fly behind flies then more flies try to fly further behind".

Question 2A: state the formula for a bigram language modelling.

In a bigram language modelling, the probability of a sequence of words in that language can be estimated by:

$$P(w_1, \dots, w_n) \approx \prod_{k=1}^{n} P(w_k \mid w_{k-1})$$

Question 2B: Using a bigram language model without smoothing, calculate the probability of the sentence "flies fly further". You should use the corpus above to estimate propabilities

From the corpus above, we can extract the following bi-grams:

(flies, fly), (behind, flies), (fly, behind), (flies, then), (then, more), (more, flies), (flies, try), (try, to), (to, fly), (fly, further), (further, behind)

The table of counts of these bigrams is:

(flies, fly) : 1	(more, flies) : 1
(behind, flies) : 1	(flies, try) : 1
(fly, behind) : 1	(try, to) : 1 (to, fly) : 1
(flies, then) : 1	(fly, further) : 1
(then, more) : 1	(further, behind) : 1

Now: $P(\text{flies fly further}) = P(\text{further} \mid \text{fly})\, P(\text{fly} \mid \text{flies})\, P(\text{flies})$

$$P(\text{further} \mid \text{fly}) = \frac{c(\text{fly, further})}{c(\text{fly, behind}) + c(\text{fly, further})} = \frac{1}{2}$$

$$P(\text{fly} \mid \text{flies}) = \frac{c(\text{flies, fly})}{c(\text{flies, fly}) + c(\text{flies, then}) + c(\text{flies, try})} = \frac{1}{3}$$

$$P(\text{flies}) = \frac{c(\text{flies, fly}) + c(\text{flies, then}) + c(\text{flies, try})}{11} = 3/11$$

Therefore $P(\text{flies fly further}) = \frac{1}{2} \cdot \frac{1}{3} \cdot \frac{3}{11} = \frac{\cancel{3}\,1}{\cancel{6}\,22}$

• Question 1C: Using a bigram language model with add-one smoothing, calculate the probability of the sentence "then flies fly further"

P(then flies fly further) = P(further | fly) P(fly | flies) P(flies | then) P(then)

$$P(\text{further} \mid \text{fly}) = \frac{c(\text{fly, further}) + 1}{c(\text{fly, behind}) + c(\text{fly, further}) + 11} = \frac{2}{13}$$

$$P(\text{fly} \mid \text{flies}) = \frac{c(\text{flies, fly}) + 1}{c(\text{flies, fly}) + c(\text{flies, then}) + c(\text{flies, try}) + 11} = \frac{2}{14} = \frac{1}{7}$$

$$P(\text{flies} \mid \text{then}) = \frac{\overset{0}{\cancel{c(\text{then, flies})}} + 1}{c(\text{then, more}) + 11} = \frac{1}{12}$$

$$P(\text{then}) = \frac{c(\text{then, more}) + 1}{22} = \frac{2}{22} = \frac{1}{11}$$

Therefore $P(\text{then flies fly further}) = \frac{2}{13} \cdot \frac{1}{7} \cdot \frac{1}{12} \cdot \frac{1}{11} = \frac{2}{12{,}012} = \frac{1}{6006}$

• Question 2D: Recall the formula for bigram interpolation $p^*(w_n|w_{n-1}) = \lambda p(w_n|w_{n-1}) + (1-\lambda)p(w_n)$. Using a bigram language model with interpolation ($\lambda = 0.5$), calculate the probability of the sentence "then flies fly"

$$p^*(w_n|w_{n-1}) = 0.5\,p(w_n|w_{n-1}) + 0.5\,p(w_n)$$

$p^*(\text{then flies fly}) = p^*(\text{fly}|\text{flies}) \cdot p^*(\text{flies}|\text{then})\, p^*(\text{then})$ where

$$p^*(\text{fly}|\text{flies}) = 0.5\,p(\text{fly}|\text{flies}) + 0.5\,p(\text{fly}) = 0.5 \cdot \tfrac{1}{3} + 0.5\,\tfrac{2}{12} = \tfrac{1}{6} + \tfrac{1}{12} = \frac{3}{12} = \frac{1}{4}$$

$$p^*(\text{flies}|\text{then}) = 0.5\,\overset{0}{\cancel{p(\text{flies}|\text{then})}} + 0.5\,p(\text{flies}) = 0.5 \cdot p(\text{flies}) = \tfrac{1}{2} \cdot \tfrac{3}{12} = \tfrac{1}{8}$$

$$p^*(\text{then}) = 0.5\,p(\text{then}|\text{<s>}) + 0.5\,p(\text{then}) = \frac{1}{2} \cdot \frac{1}{11} + \frac{1}{2} \cdot \frac{1}{12} = \frac{1}{22} + \frac{1}{24} = \frac{23}{264}$$

Therefore $P(\text{then flies fly}) = \tfrac{1}{4} \cdot \tfrac{1}{8} \cdot \tfrac{23}{264} = \frac{23}{8416}$

• Question 2E: Why may a language model be used in a machine translation system?

Using a language model, we can choose the most likely output in machine translation.

• Section 3: Parsing

Consider the following propabilistic grammar

$N \rightarrow$ natural	0.6	$N \rightarrow A\ NP$	0.1
$N \rightarrow$ language	0.2	$NP \rightarrow NP\ NP$	0.3
$N \rightarrow$ processing	0.1	$NP \rightarrow N$	0.6
$N \rightarrow$ works	0.1	$VP \rightarrow V$	0.4
$A \rightarrow$ natural	1.0	$VP \rightarrow V\ NP$	0.4
$V \rightarrow$ processing	0.1	$VP \rightarrow V\ NP\ NP$	0.2
$V \rightarrow$ works	0.9	$S \rightarrow NP\ VP$	0.8
		$S \rightarrow NP$	0.2

Question 3A: Describe one ambiguity when applying the above grammar to the sentence "natural language processing works"

processing could be tagged as N (0.1) or V (0.1) with the same propabilits

Question 3B: What changes would be necessary to convert the above grammar into Chomsky normal form?

For a grammar G to be in Chomsky normal form all of its production rules have to be of the form: $A \rightarrow BC$, or $A \rightarrow a$, or $S \rightarrow \varepsilon$

There are production rules in the grammar above that do not satisfy any of these formats: $VP \rightarrow V\ NP\ NP$ (0.2), $NP \rightarrow N$ (0.6), $VP \rightarrow V$ (0.4).

These last two can be easily changed to $NP \rightarrow$ works 0.06 = 0.6·0.1 | ... | ... |
$VP \rightarrow$ processing (0.04) | works (0.36)

Finally $VP \rightarrow V\ NP\ NP$ (0.2) can be converted to Chomsky form using e.g., the rule $NP \rightarrow NP\ NP$ (0.3). This is $VP \rightarrow V\ NP$ (0.06). Therefore the new grammar would be:

$N \rightarrow$ natural 0.6	$N \rightarrow A\ NP$ 0.1	$VP \rightarrow$ works 0.36	$S \rightarrow$ processing 0.012
$N \rightarrow$ language 0.2	$NP \rightarrow NP\ NP$ 0.3	$VP \rightarrow V\ NP$	$S \rightarrow$ works 0.108
$N \rightarrow$ processing 0.1	$NP \rightarrow$ natural 0.36	$VP \rightarrow V\ NP$ 0.06	
$N \rightarrow$ works 0.1	$NP \rightarrow$ language 0.12	$VP \rightarrow VP\ NP$ 0.08	
$A \rightarrow$ natural 1.0	$NP \rightarrow$ processing 0.06	$S \rightarrow NP\ VP$ 0.8	
$V \rightarrow$ processing 0.1	$NP \rightarrow$ works 0.06	$S \rightarrow$ natural 0.072	
$V \rightarrow$ works 0.9	$VP \rightarrow$ processing 0.04	$S \rightarrow$ language 0.024	

Question 3C Why should a grammar be in Chomsky normal form when applying the CYK algorithm?

A grammar G that ~~the~~ is not in Chomsky normal form means that at least has a production rule that is not in any of these formats $A \to BC$, or $A \to a$.

Let's first assume that the production rule is $A \to BCD$ where it could be simplified to $A \to XD$.

Now, during CYK algorithm, in the second iteration we are trying to capture ~~term~~ non-terminal symbols that can be assigned from the previous iteration in groups of 2. Because of this, if our rule is represented as $A \to BCD$ none of two combinations would be able to produce A. Additionally we know that this rule can be simplified to $A \to XD$, which of the combination exists would give to non-terminal A.

Finally, $A \to B$ can be simplified to or, $A \to CD$, or $A \to a$. It will not lead to a wrong output but it will help to have a more efficient algorithm.

Question 3D: What is a cross-bracketing error and why may it not be important in the example of 3A?

Cross-bracketing error occurs when a sentence is parsed as, e.g., A(BC) but we expect (A B (C)). In 3 A, an individual terminal symbol can have two possible outputs, therefore cross-bracketing error would not help to find the correct output.

- 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: Construct a co-occurrence matrix for all types in the corpus, using a context window of two words

The corpus has the following types: a, black, cat, chased, the, white, dog.

Therefore, the co-occurrence matrix is going to be a 7×7 matrix $A = (a_{ij})_{i,j=1,\dots,7}$ where each element a_{ij} will represent the number of times that the word j is found in the context of the word i.

	a	black	cat	chased	the	white	dog
a	5	2	3	2	0	3	3
black	2	4	2	4	2	0	2
cat	3	2	6	3	4	4	0
chased	2	4	3	6	4	8	3
the	0	2	4	4	7	5	3
white	3	0	4	8	5	8	4
dog	3	2	0	3	3	4	6

$$A = \begin{pmatrix} 5 & 2 & 3 & 2 & 0 & 3 & 3 \\ 2 & 4 & 2 & 4 & 2 & 0 & 2 \\ 3 & 2 & 6 & 3 & 4 & 4 & 0 \\ 2 & 4 & 3 & 6 & 4 & 8 & 3 \\ 0 & 2 & 4 & 4 & 7 & 5 & 3 \\ 3 & 0 & 4 & 8 & 5 & 8 & 4 \\ 2 & 2 & 0 & 3 & 3 & 4 & 6 \end{pmatrix}$$

Question 4B: Using Cosine similarity, compute the distance between:

- black, white
- cat, dog

Given two embedded words w_1, w_2, the cosine similarity is $\cos\alpha = \frac{|w_1 \cdot w_2|}{\|w_1\| \cdot \|w_2\|}$

From the co-occurrence matrix, we know that:

black $\leftarrow (2,4,2,4,2,0,2)$
white $\leftarrow (3,0,4,8,5,8,4)$
therefore
$\|black\| = \sqrt{2^2+4^2+2^2+4^2+2^2+2^2} = \sqrt{48}$
$\|white\| = \sqrt{3^2+4^2+8^2+5^2+8^2+4^2} = \sqrt{194}$
$|black \cdot white| = 2\cdot3 + 2\cdot4 + 4\cdot8 + 2\cdot5 + 2\cdot4 = 64$

This is $\cos\alpha = \frac{64}{\sqrt{48}\sqrt{194}} \simeq 0.66$

In a similar way:

cat $\leftarrow (3,2,6,3,4,4,0)$
dog $\leftarrow (3,2,0,3,3,4,6)$
therefore
$\|cat\| = \sqrt{9+4+36+9+16+16} = \sqrt{90}$
$\|dog\| = \sqrt{9+4+9+9+16+36} = \sqrt{83}$
$|cat \cdot dog| = 9+4+9+12+16 = 50$

This is $\cos\alpha = \frac{50}{\sqrt{90}\cdot\sqrt{83}} \simeq 0.38$