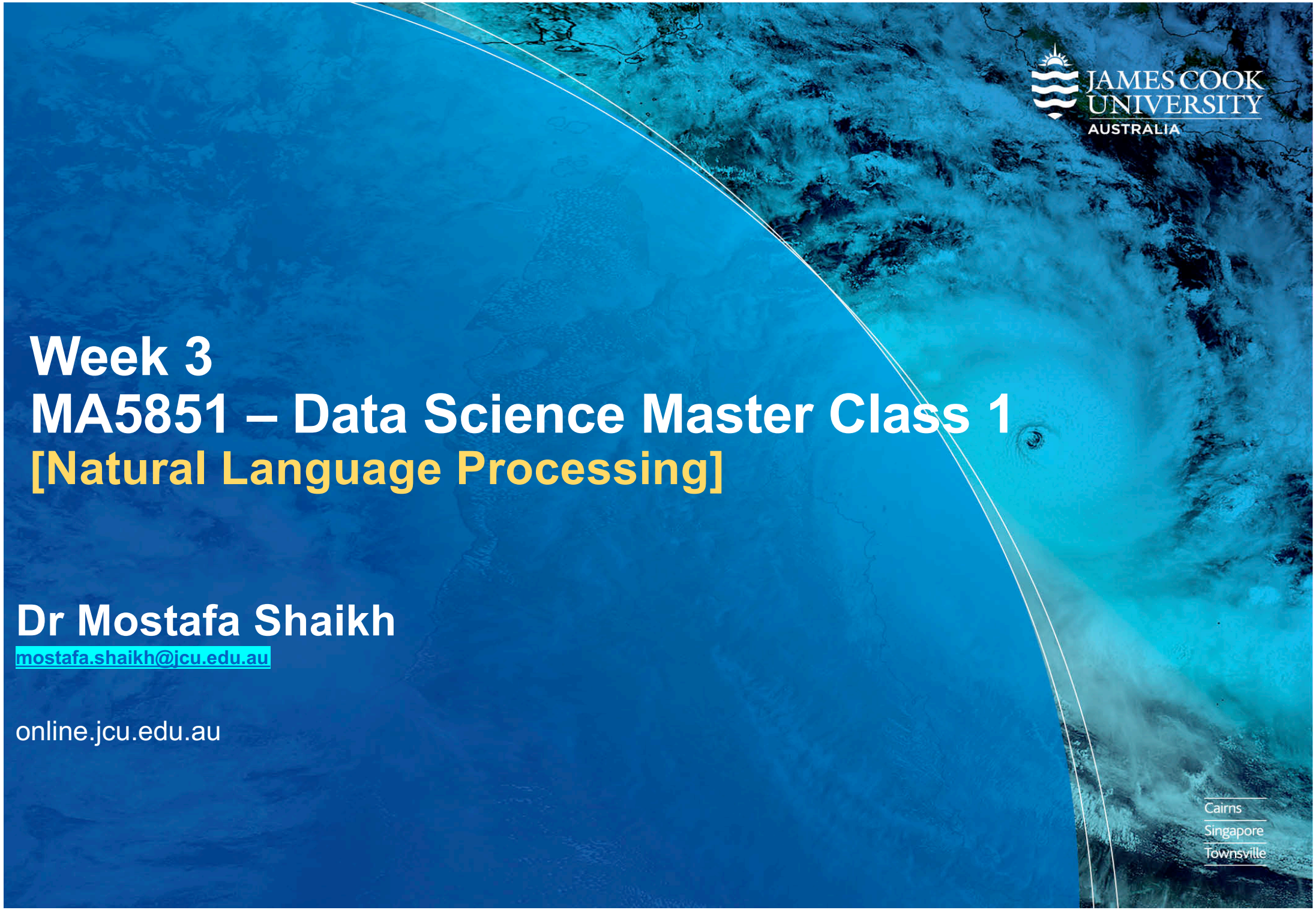

JAMES COOK UNIVERSITY
AUSTRALIA
Week 3
MA5851 – Data Science Master Class 1
[Natural Language Processing]
Dr Mostafa Shaikh
mostafa.shaikh@jcu.edu.au
online.jcu.edu.au
Cairns
Singapore
Townsville

Topic

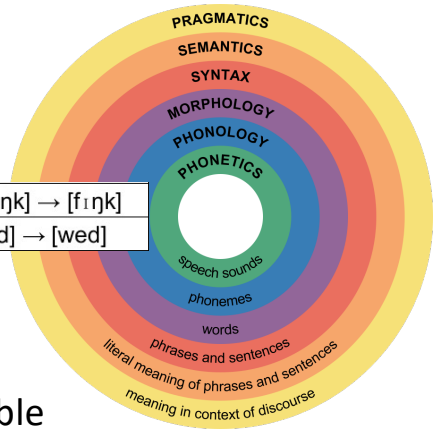
JAMES COOK UNIVERSITY AUSTRALIA

- Language structure and components
- Context Free Grammar
- Phrase structure grammar
- The Chomsky Hierarchy
- Parsing, parser
- WSD
- NLU, NLG, Turing Test
- Introduction to NLTK

Self Practical

- Create your own syntax tree
- Parsing using Stanford Parser
- NTLK based examples
- Creating sentence segmenter
- NLTK and WordNet

Language structure and components



- Main components of language:
  - Phonemes: approximately 45 different phonemes

/f/	for	/θ/	think [θɪŋk] → [fɪŋk]
/w/	for	/r/	red [red] → [wed]

  - Morphemes: smallest meaningful unit of language
    - One morpheme: water (two syllables), crocodile (three syllables)
    - Three morpheme: desirability = desire + able + ity, unbreakable = un + break + able
    - Five morpheme: ungentlemanliness = un + gentle + man + li + ness
  - Lexemes: run, runs, ran, running → run
  - Syntax: Subject Verb Object (SVO), *The baby ate the cucumber* VS *The cucumber ate the baby*
  - Context: to convey a particular meaning based on time, situation, mood, articulation etc…..
- Structure of language
  - Grammar: a set of rules, phonology + morphology + syntax
  - Semantics: meaning of morphemes, words, phrase and sentences
  - Pragmatics: Contextual semantics, inferred meaning; "*Will you crack open the door? I am getting hot.*"
- Result = meaningful communication

Context free grammar

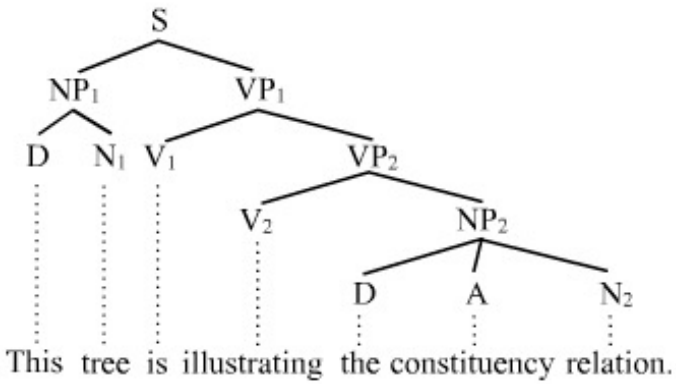
- *"JCU websites use cookies to enhance user experience, analyse site usage, and assist with outreach and enrolment."*
  - (ROOT (S (NP (NNP JCU) (NNS websites)) (VP (VBP use) (S (NP (NNS cookies)) (VP (TO to) (VP (VP (VB enhance) (NP (NN user) (NN experience))) (, ,) (VP (VB analyse) (NP (NN site) (NN usage))) (, ,) (CC and) (VP (VB assist) (PP (IN with) (NP (NN outreach) (CC and) (NN enrolment))))))))
- Context free grammar is a formal grammar which is used to generate all possible strings in a given formal language.
- G= (V, T, P, S)
  - **G** describes the grammar
  - **T** describes a finite set of terminal symbols.
  - **V** describes a finite set of non-terminal symbols
  - **P** describes a set of production rules
  - **S** is the start symbol.
- Useful to describe most of the programming languages
- Context free grammar is capable of describing nested structures like: balanced parentheses, matching begin-end, corresponding if-then-else's & so on

Phrase structure grammar

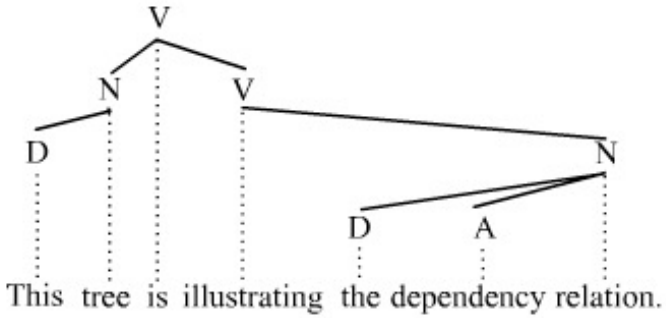
- usually named as phrases based on the word that heads the constituent

the man from Brisbane	is a Noun Phrase (NP) because the head man is a noun
extremely difficult	is an Adjective Phrase (AP) because the head difficult is an adjective
down the river	is a Prepositional Phrase (PP) because the head down is a preposition
killed the rabbit	is a Verb Phrase (VP) because the head killed is a verb

- consists of a set of ordered rewrite rules

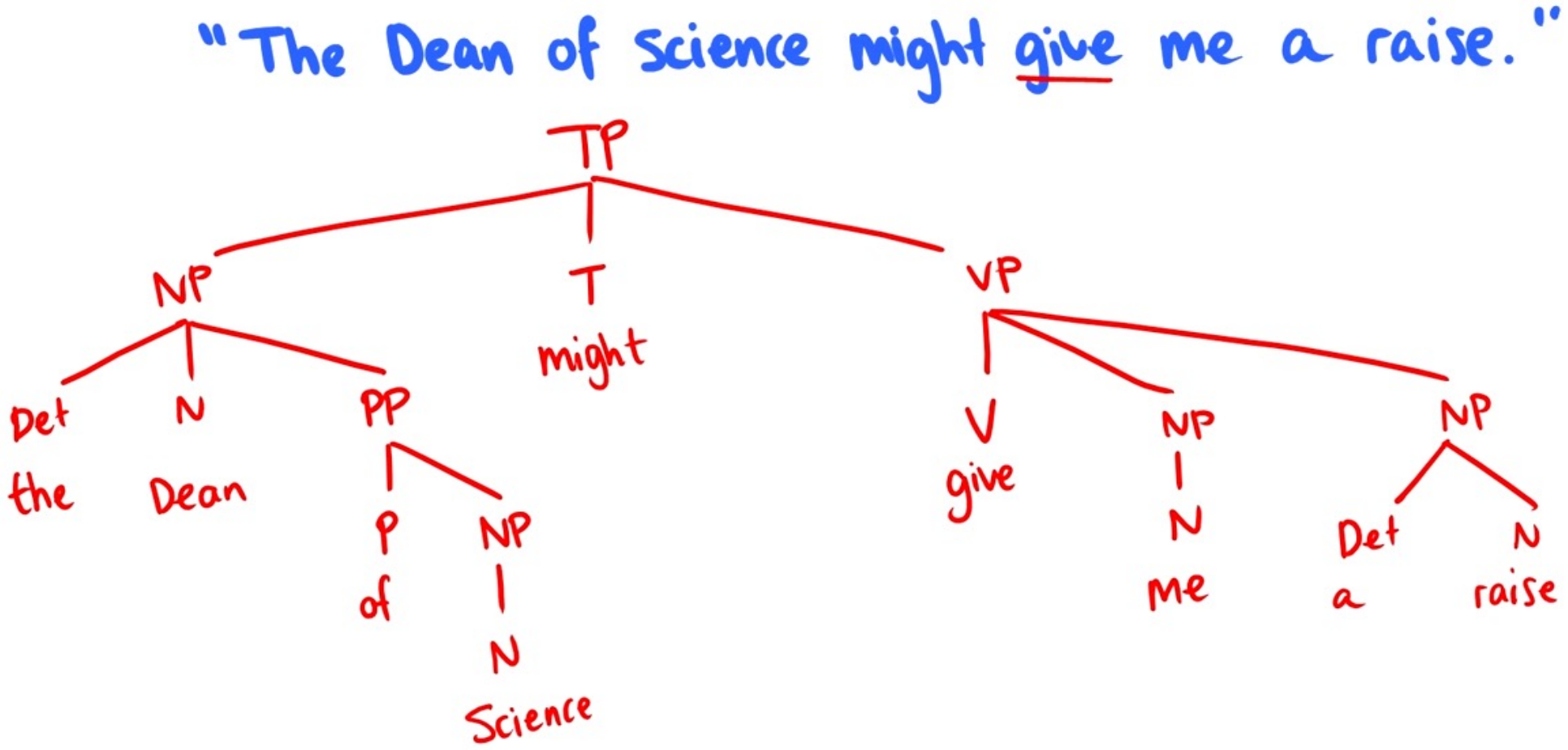


Constituency relation (PSG)



Dependency relation

Phrase structure grammar



The Chomsky Hierarchy

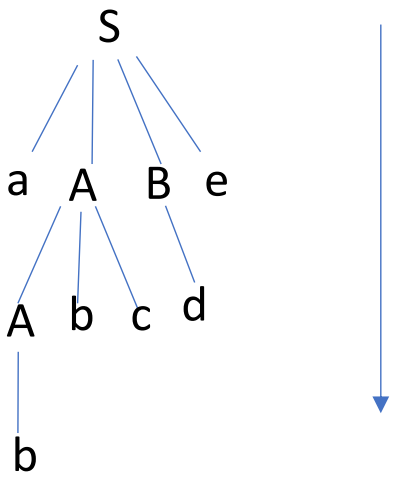
JAMES COOK UNIVERSITY AUSTRALIA

Language classes	Grammar	Example
3	Regular	Rewrite rules X → αY, where X,Y are single non-terminals, and α is a string of terminals; Y might be missing.
		phrase → word phrase phrase → word valid sentences using this grammar would be: • Jane should be allowed to work full time. • Alan thinks that Jane should be allowed to work full-time at a company. • If the sky is blue then Alan thinks that Jane should be allowed to work full-time at a company.
2	Context-free	If if if you pass the Turing test then then then you are conscious. But the above sentence can be made a valid one as follows: If either you pass the Turing test, or a psychological exam, or you have a mind, then you are conscious.
		apply some rules to Type 3 grammar with some sort of memory to keep track of how many terminals has been used with non-terminal. The productions must be in the form **A → γ;** where **A ∈ N** (Non terminal) and **γ ∈ (T ∪ N)*** (String of terminals and non-terminals).
		phrase → if phrase then phrase phrase → either phrase or phrase phrase → word phrase phrase → word
1	Context sensitive	Rewrite rules αXβ → αγβ, where X is a non-terminal, and α,β,γ are any string of terminals and nonterminals, (γ must be non-empty) [context] X [context] → [context] Y [context]
0	Unrestricted (free)	The productions can be in the form of α → β where α is a string of terminals and non-terminal with at least one non-terminal and α cannot be null. β is a string of terminals and non-terminals
		S → ACaB Bc → acB CB → DB aD → Db

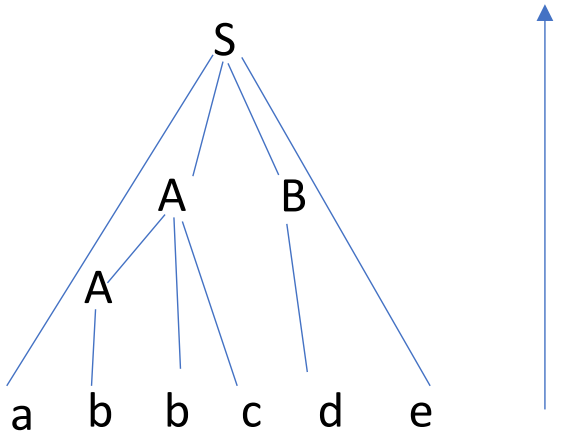
Parsing (top-down, bottom-up)

- **S → aABe**
- **A → Abc | b**
- **B → d**

Input: abbcde



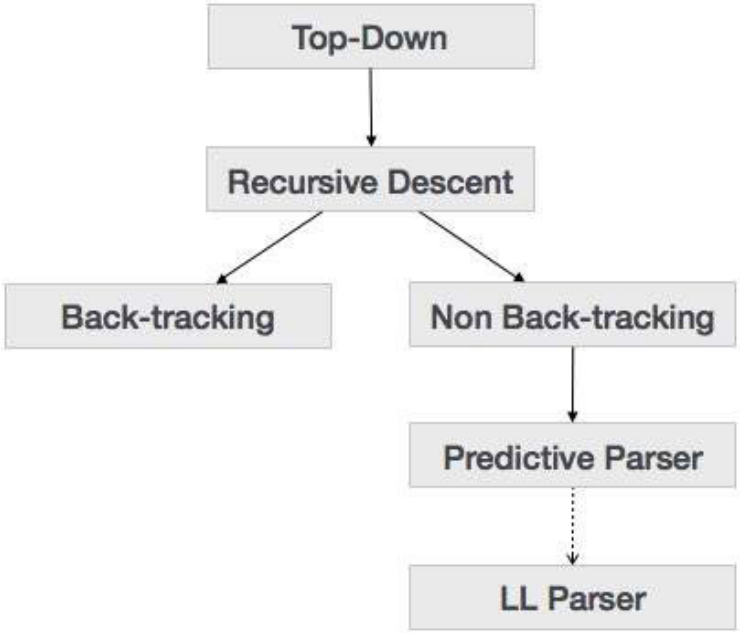
S → aABe
→ aAbcBe
→ abbcBe
→ abbcde

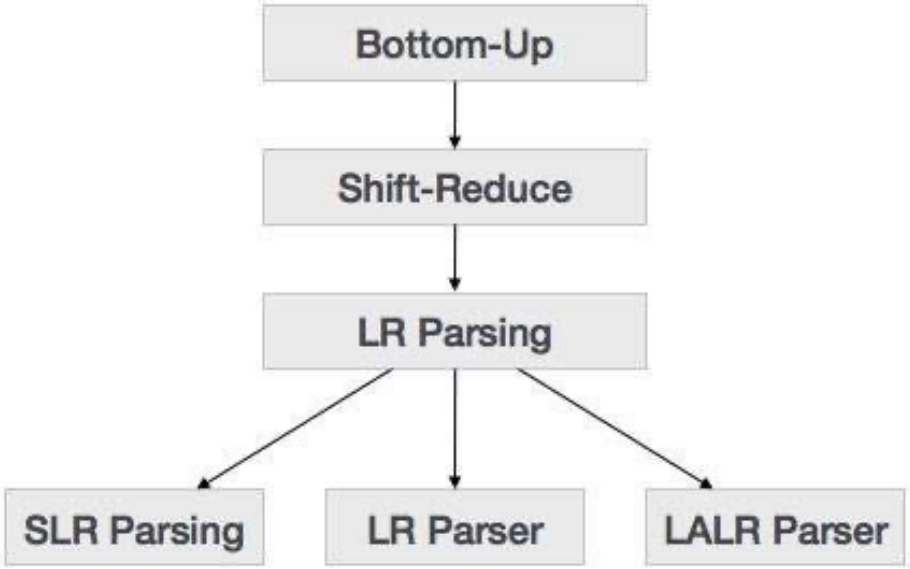


S → aABe
→ aAde
→ aAbcde
→ aabbcde

Parsing

- **Lexical**: name of some identifier typed incorrectly
- **Syntactical**: missing semicolon or unbalanced parenthesis
- **Semantical**: incompatible value assignment
- **Logical**: code not reachable, infinite loop

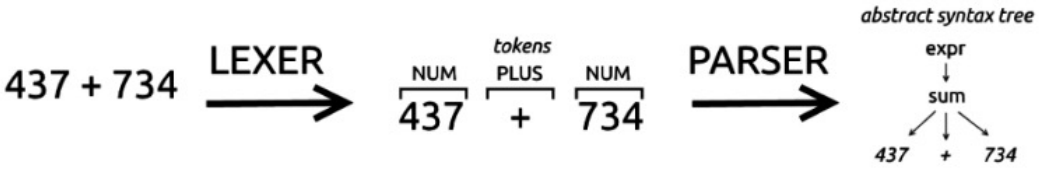




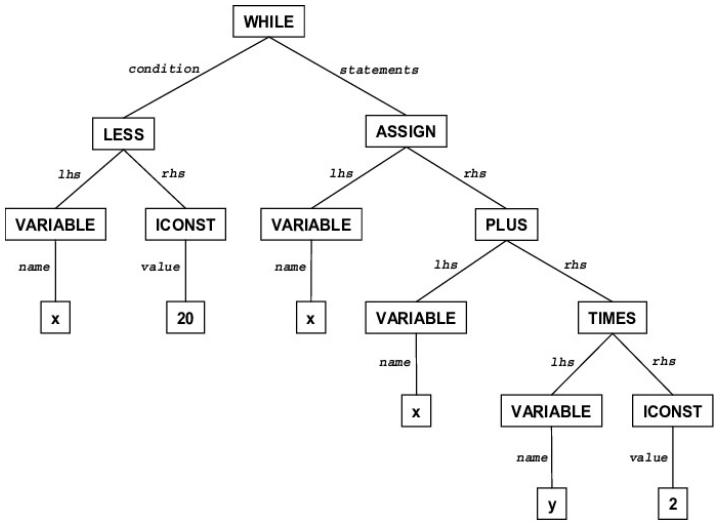
Idea from python parser: https://docs.python.org/3/reference/simple_stmts.html

Parser

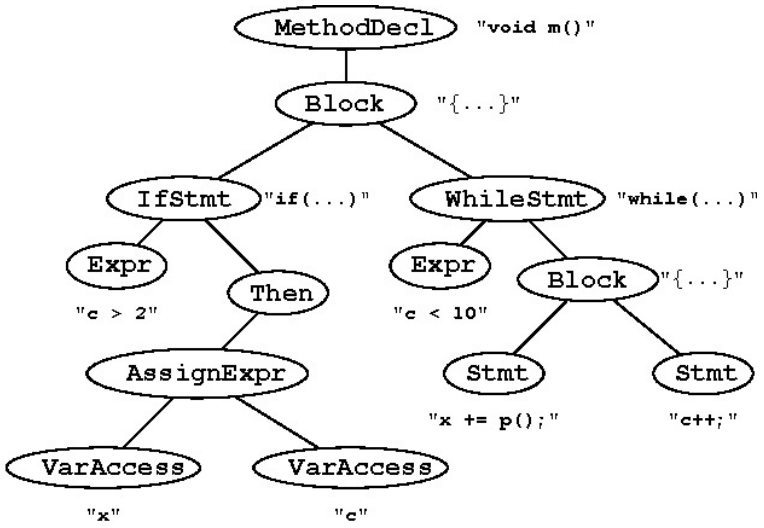
- A parser is usually composed of two parts: a *lexer (scanner* or *tokenizer)* and the proper parser.



- parse tree or Abstract Syntax Tree (AST)




```
void m() {
  if(c > 2)
    x = c;
  while(c < 10) {
    x += p();
    c++;
  }
}
```



Word Sense Disambiguation

JAMES COOK UNIVERSITY AUSTRALIA

plant	living/factory
tank	vehicle/container
poach	steal/boil
palm	tree/hand
bass	fish/music
motion	legal/physical
crane	bird/machine

- Most words have multiple senses. Which sense is invoked in a context?
- Types of problems: homonymy (unrelated meaning), polysemy (related meaning)
- “I love this dish” --> “I love spicy dishes” “I hate washing dishes”

Solution Approaches:

- Dictionary-based or Knowledge-based Methods
  - The Lesk algorithm: words in a given "neighbourhood" (section of text) will tend to share a common topic.
- Supervised Methods
  - context is represented as a set of “features” of the words. It includes the information about the surrounding words. “know the word by the company it keeps”

The ocean reflects the color of the sky, but even on cloudless days the color of the ocean is not a consistent blue. Phytoplankton, microscopic plant **life** that floats freely in the lighted surface waters, may alter the color of the water. When a great number of organisms are concentrated in an area, . . .

w_{-1} = microscopic	t_1=JJ
w_{+1} = life	t_{+1}=NN
w_{-2}, w_{-1} =(Phytoplankton, microscopic)	. . .
w_{-1}, w_{+1} = (microscopic, life)	
word-within-k=ocean	
word-within-k=reflects	
. . .	

- Unsupervised Methods
  - word sense induction or discrimination: assume that similar senses occur in similar context; senses can be induced from text by clustering word occurrences of similarity of the context

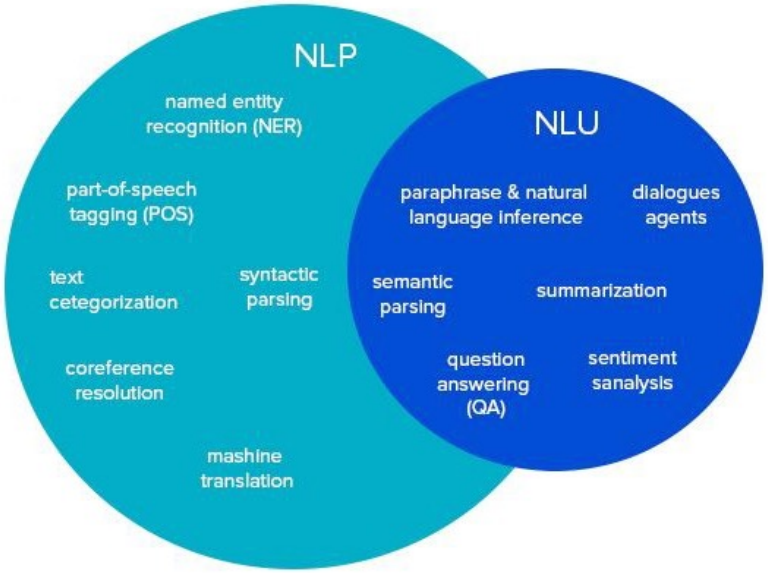
Input: The finite set of words W and the textual context T
Output: The disambiguated word senses
1. Let *Senses* be the set of all senses of words in W
2. Repeat
   a. G = *clustering*(*Senses*)
   b. *Selected_G* = *filter*(G, W, T)
   c. $Senses = \bigcup_{g \in Selected\_G} \{s \mid s \in g\}$

   until *stopping-criterion*
3. Return *Senses*

Why needed?

- MT
- IR
- Lexicography

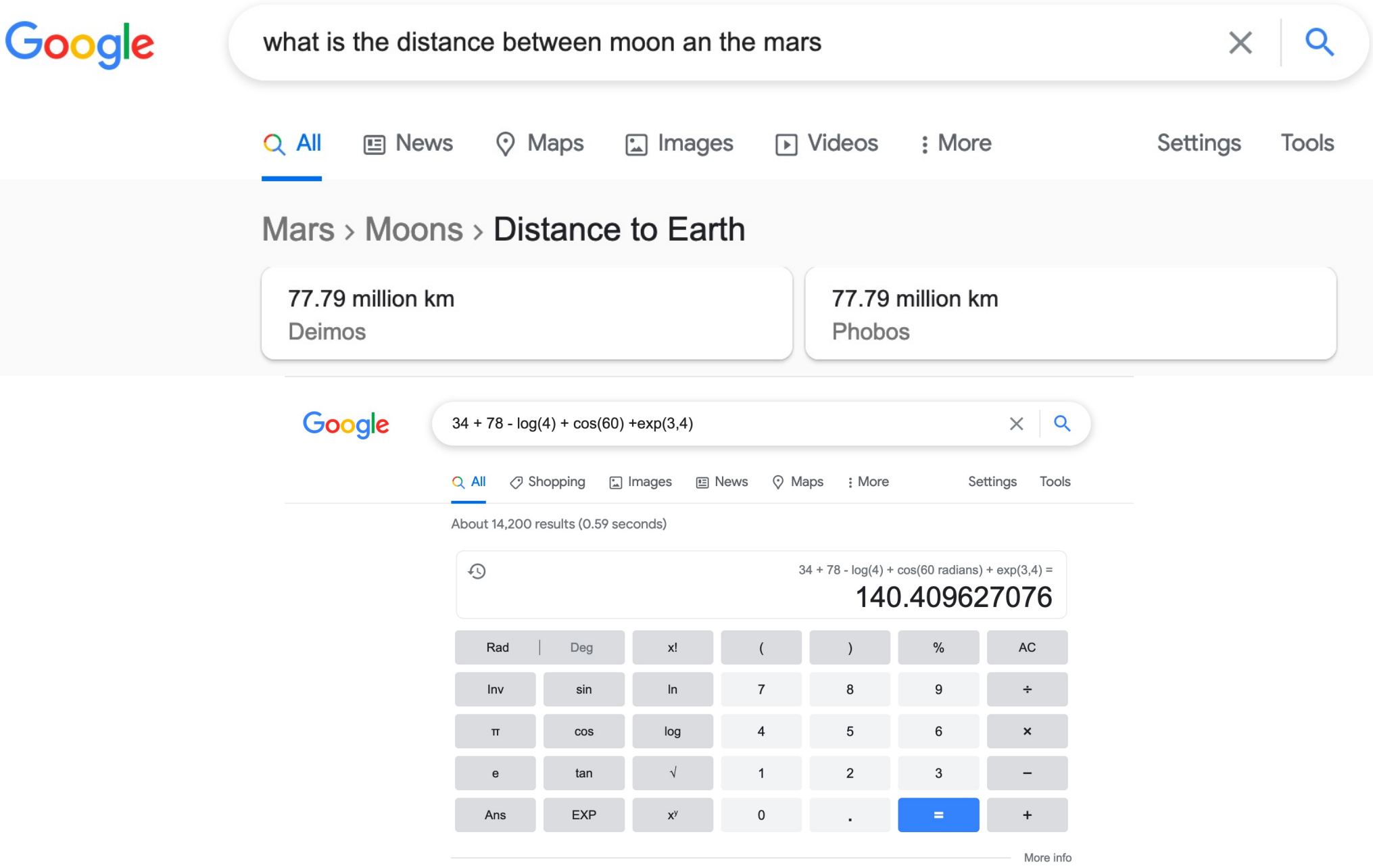
NLU and NLG


NLP
named entity recognition (NER)
part-of-speech tagging (POS)
text cetegorization
syntactic parsing
coreference resolution
mashine translation
NLU
paraphrase & natural language inference
dialogues agents
semantic parsing
summarization
question answering (QA)
sentiment sanalysis

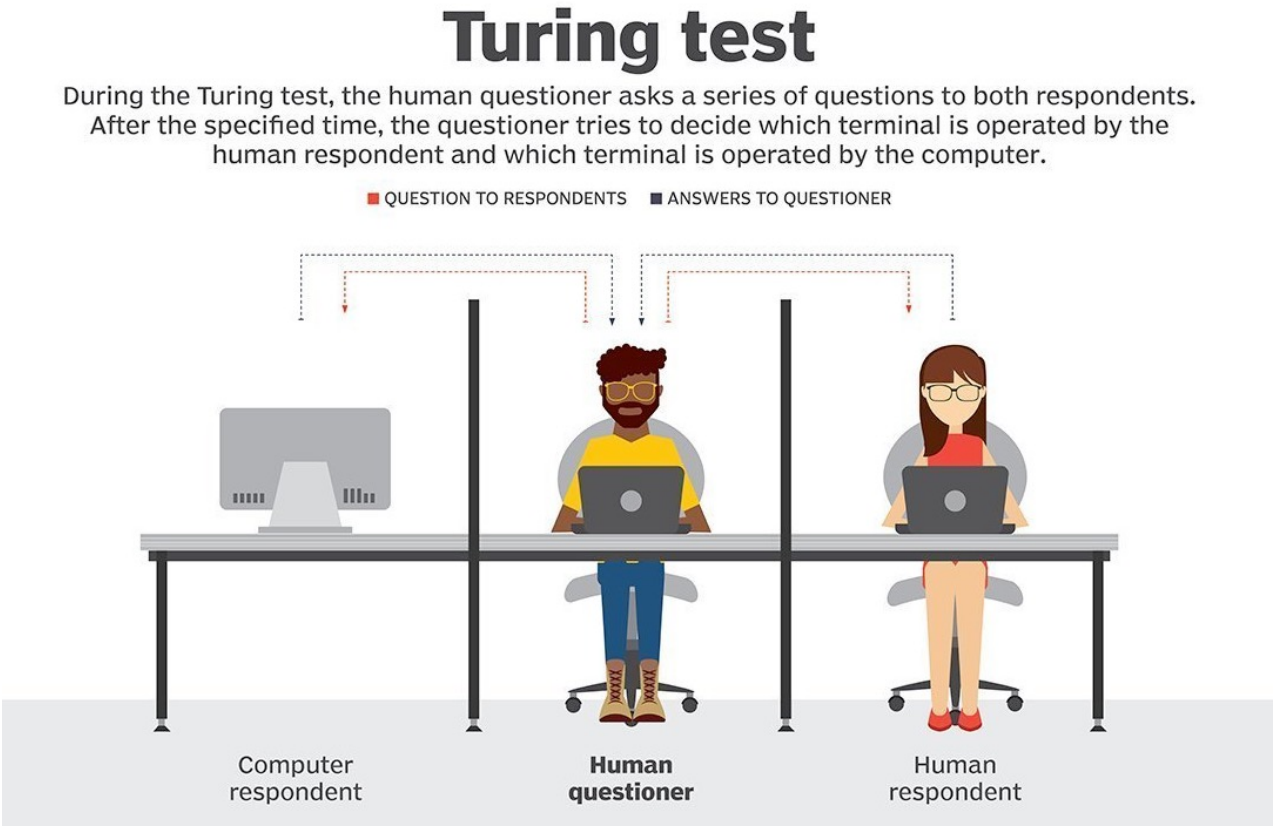
Custom Data Source
Data Connector
Web
Facebook
Slack
Message
Response
Messaging Backend
Plain Text
NLP / NLU
Decision Engine
NLG
Machine Learning Layer
Presentation Layer
Plain Text Array

Semantic Parsing

JAMES COOK UNIVERSITY AUSTRALIA



Turing Test



No computer has passed the Turing test until 2014.
The **Turing Test is** successfully **passed** if a computer **is** mistaken for a human more than 30% of the time during a series of five-minute keyboard conversations.
On 7 June 2014 a conversation agent named Eugene convinced 33% of the judges at the Royal Society in London that it was human
(read the story here: https://www.bbc.com/news/technology-27762088).

NLG

JAMES COOK UNIVERSITY AUSTRALIA

- **Content determination:** salient features; topic(s); data mining for NP, SVO.
- **Discourse planning:** Overall organization of the information to convey. Syntax and Grammar.
- **Sentence aggregation:** Merging of similar sentences to improve readability and naturalness. Joining of events. Summarisation. Sentiment.
- **Lexicalization:** Putting words to the concepts. WSD.
- **Referring expression generation:** Linking words in the sentences by introducing pronouns and other types of means of reference. Anaphora.
- **Syntactic and morphological realisation:** This stage is the inverse of parsing: given all the information collected above, syntactic and morphological rules are applied to produce the surface string. Real world knowledge, Common-sense.
- **Orthographic realisation:** Matters like casing, punctuation, formatting, emoticons, idiosyncrasy

https://www.youtube.com/watch?v=D5VN56jQMWM (Google Duplex AI)

NLTK

- NLTK is a leading platform for building Python programs to work with human language data.
- It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet,
- suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries

- What about text processing in R?
  - https://cran.r-project.org/web/views/NaturalLanguageProcessing.html

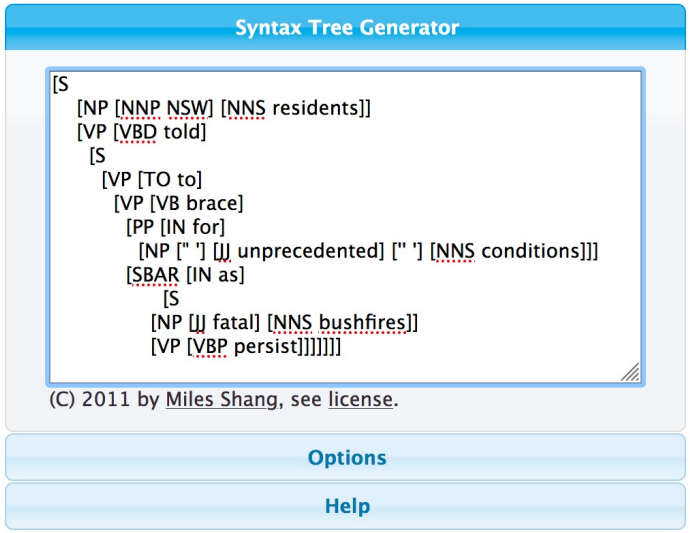
Parsing and constituents

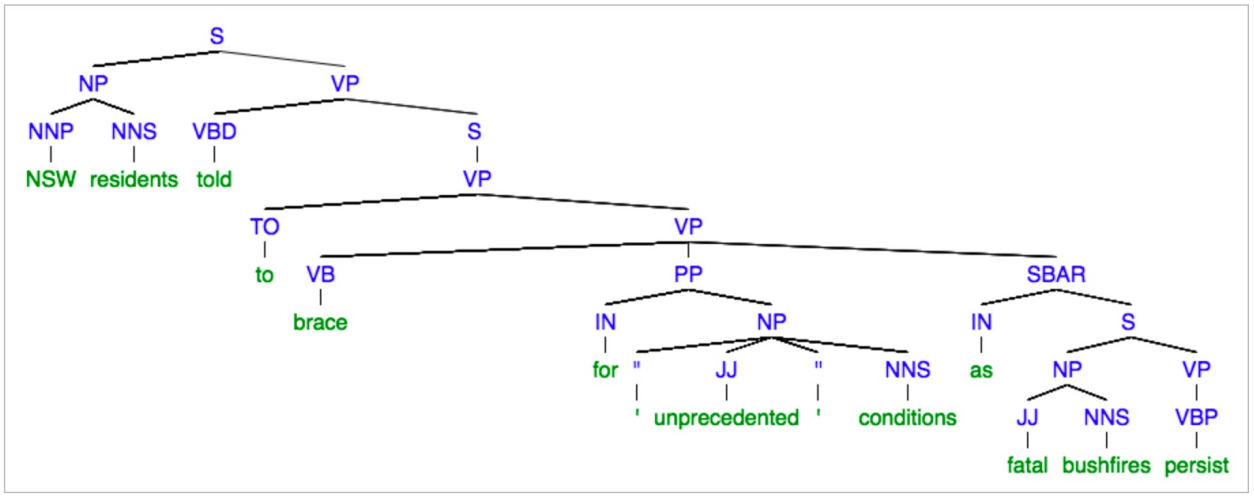
JAMES COOK UNIVERSITY AUSTRALIA

http://nlp.stanford.edu:8080/parser/#sample

Input: NSW residents told to brace for 'unprecedented' conditions as fatal bushfires persist.

How the parse tree will look?
http://mshang.ca/syntree/





Penn Treebank Tags

https://gist.github.com/nlothian/9240750