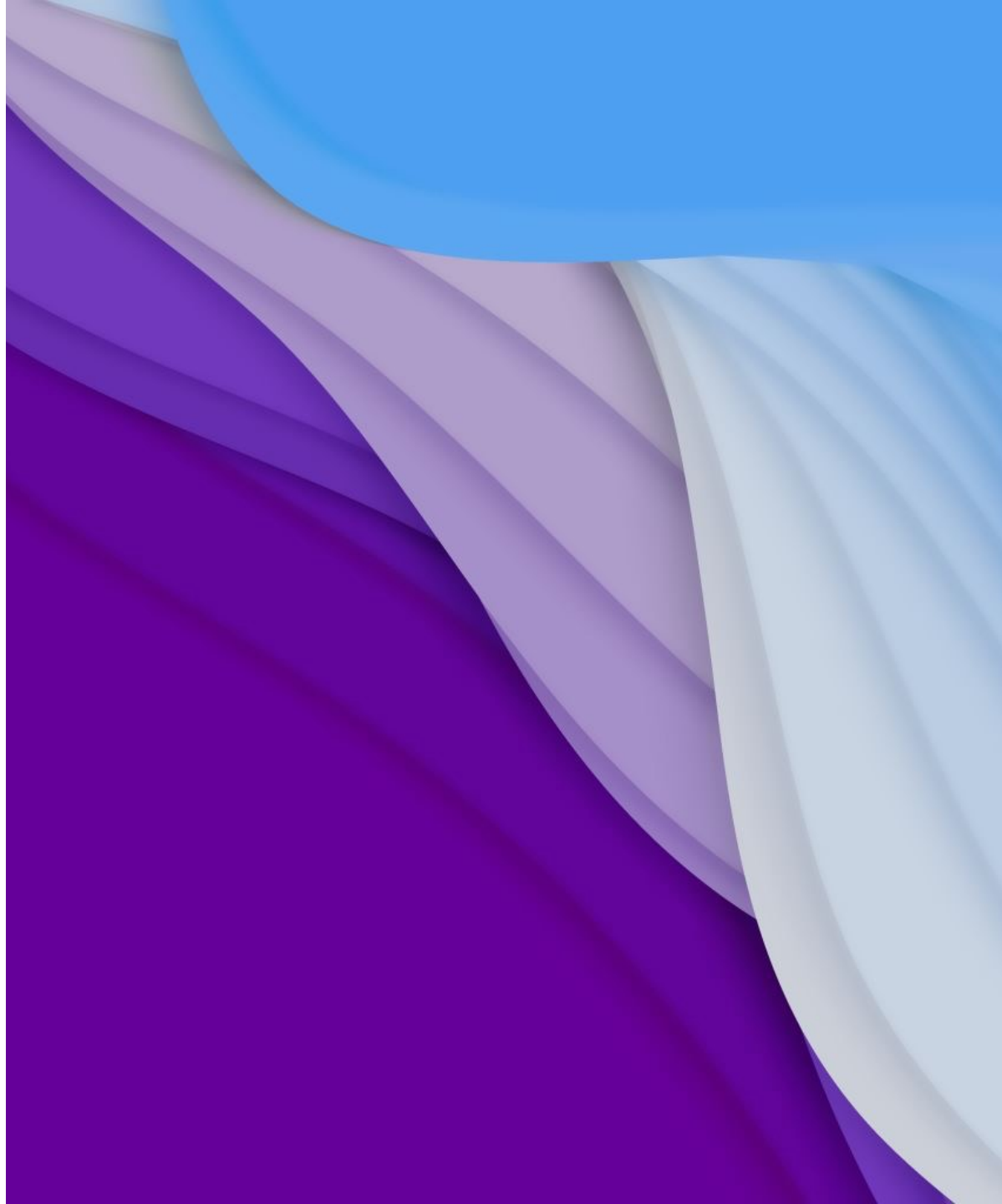
LINGUISTIQUE DE CORPUS EN ANGLAIS

Armand STRICKER

stricker@lisn.fr

https://github.com/armandstrickernlp/NLP_Inalco



CLASS STRUCTURE

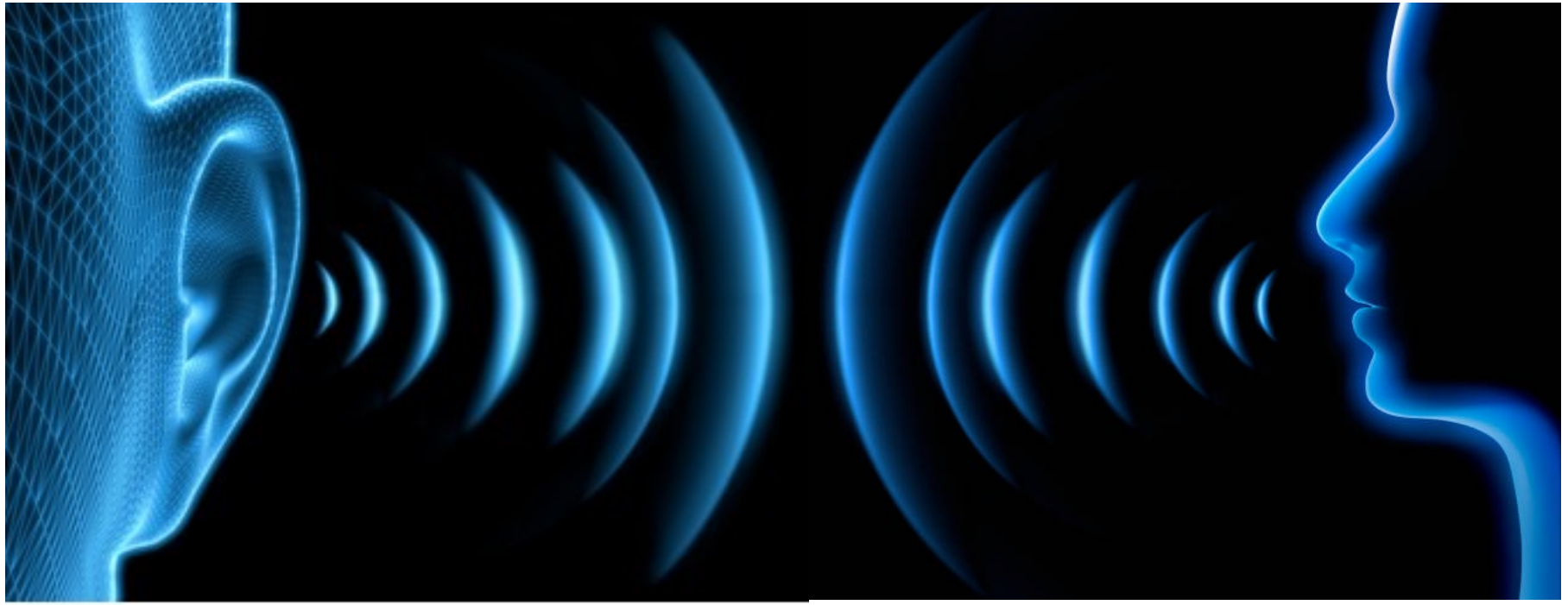
- 2hrs => try to divide classes into theory + practice
  - Either over 1 class or over 2 classes depending on the subject matter => goal is that you take something from this class that may be useful later (job, job interview, reading research papers…)
  - My goal is to simplify and clarify as much as possible certain technical details we will go over
  - Work on the practical parts as pairs => try to go over everything together always and divide the tasks. Very important to explain what you did to your partner clearly => means you understand + important skill when you work with a team in a company or within a research group
- Grades : mid-term written exam and a project at the end

COURSE « REQUIREMENTS »

- Helps to have a laptop that you can bring :
  - Familiarize yourself with your machine as much as possible
  - Means you can continue exploring/play around at home with what was done in class
- A little math (not a ‘requirement’ per se):
  - Probability
  - Manipulate equations
  - Some linear algebra
- Computer/programming skills:
  - Python
  - Unix shell

OVERVIEW OF LINGUISTICS

- **Input: sound wave**
- Phonetics
- Phonology
- Morphology
- Syntax
- Semantics
- Pragmatics
- Output: meaning



OVERVIEW OF LINGUISTICS

- Input: sound wave
- **Phonetics**
- Phonology
- Morphology
- Syntax
- Semantics
- Pragmatics
- Output: meaning

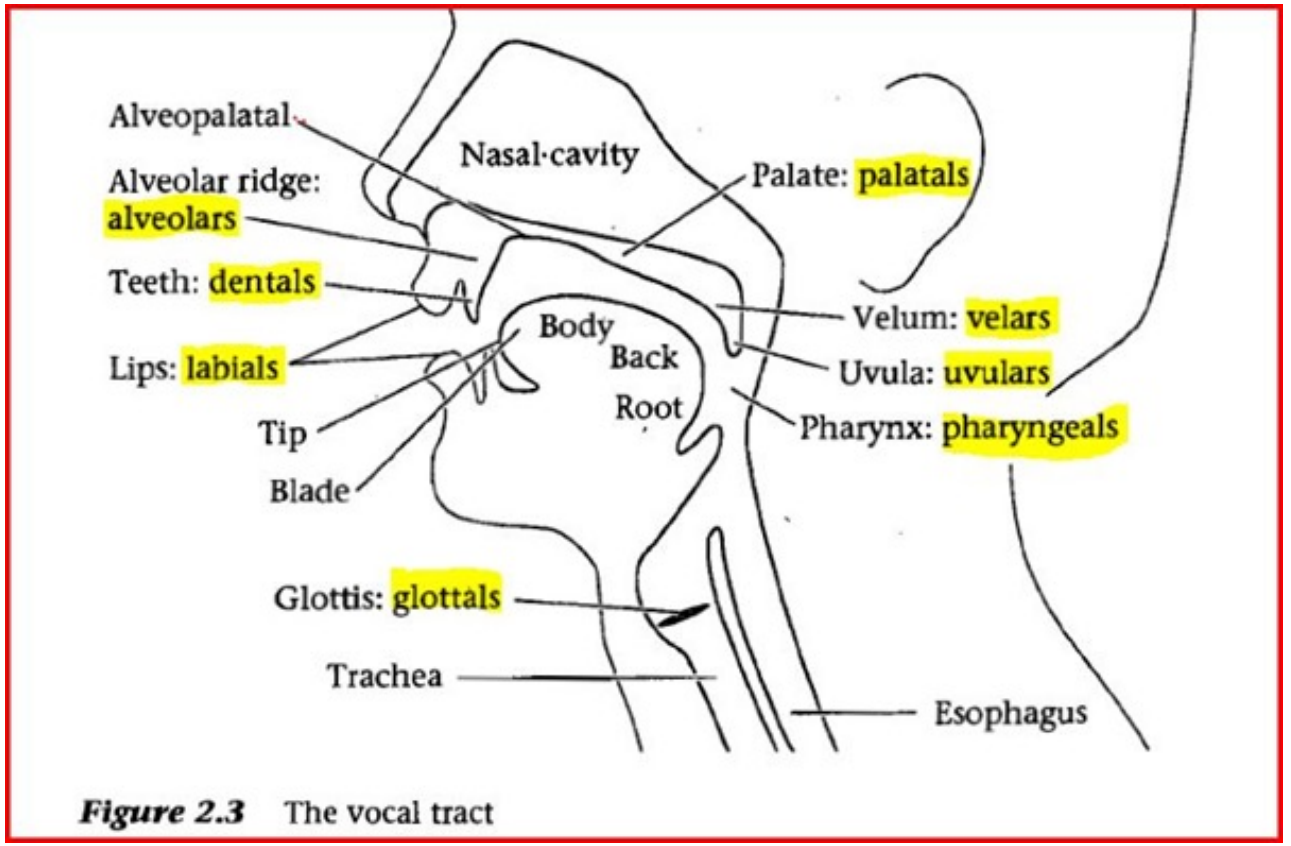
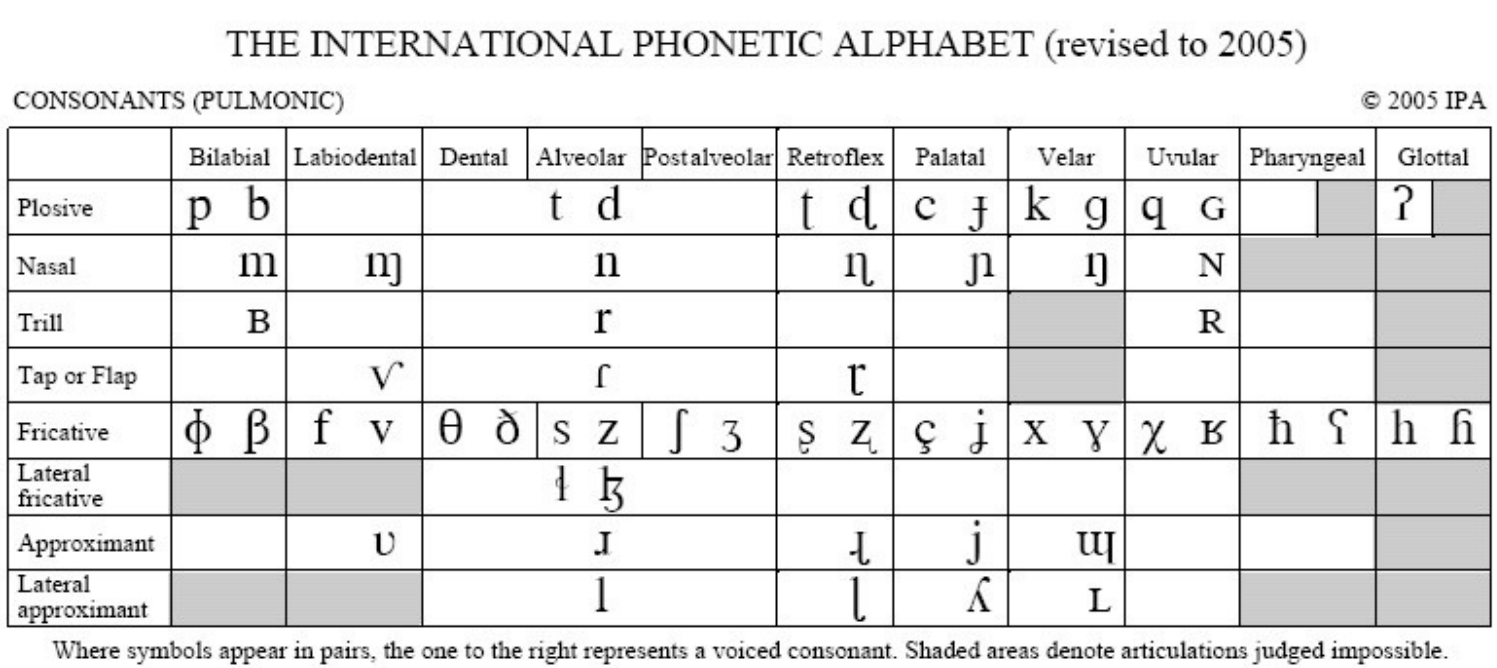


Figure 2.3 The vocal tract

OVERVIEW OF LINGUISTICS

- Input: sound wave
- **Phonetics**
- Phonology
- Morphology
- Syntax
- Semantics
- Pragmatics
- Output: meaning

Consonants



THE INTERNATIONAL PHONETIC ALPHABET (revised to 2005)

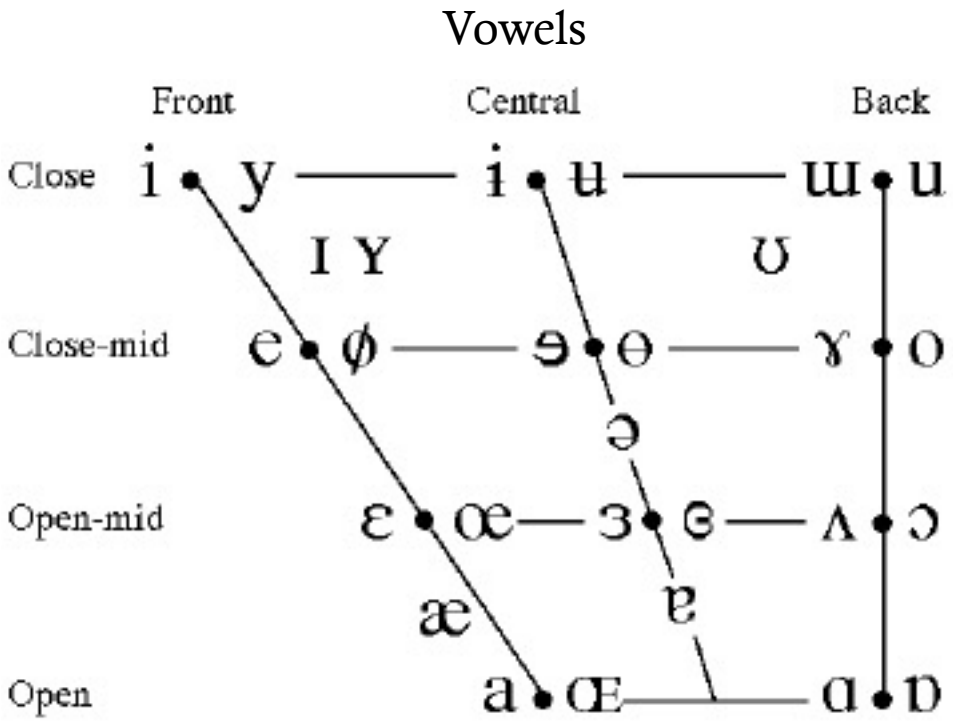
CONSONANTS (PULMONIC) © 2005 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

OVERVIEW OF LINGUISTICS

- Input: sound wave
- **Phonetics**
- Phonology
- Morphology
- Syntax
- Semantics
- Pragmatics
- Output: meaning



OVERVIEW OF LINGUISTICS

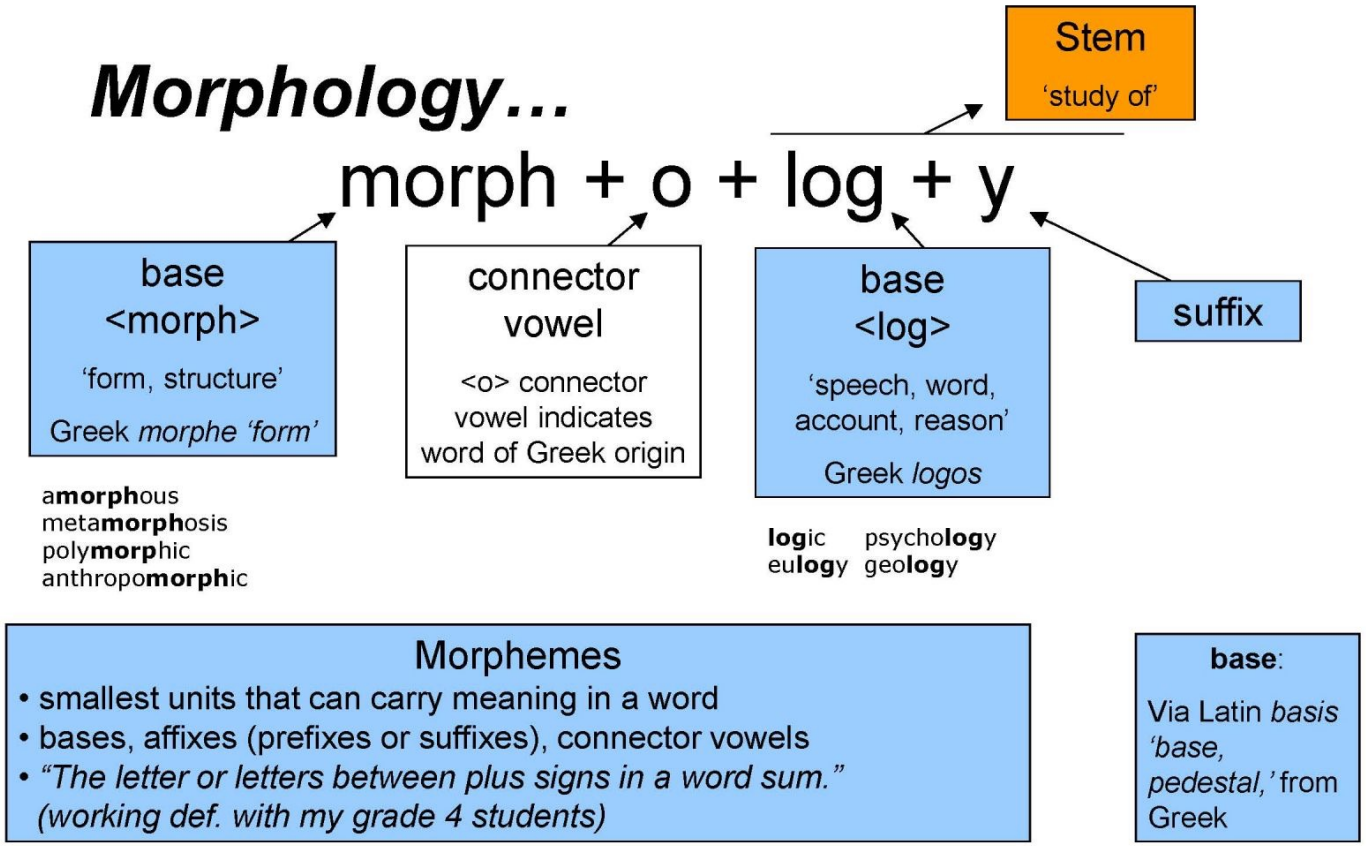
- Input: sound wave
- Phonetics
- **Phonology**
- Morphology
- Syntax
- Semantics
- Pragmatics
- Output: meaning

able	ability	[ə]
super	superior	[ə]
photograph	photography	[ə]

vowel neutralisation in unstressed positions

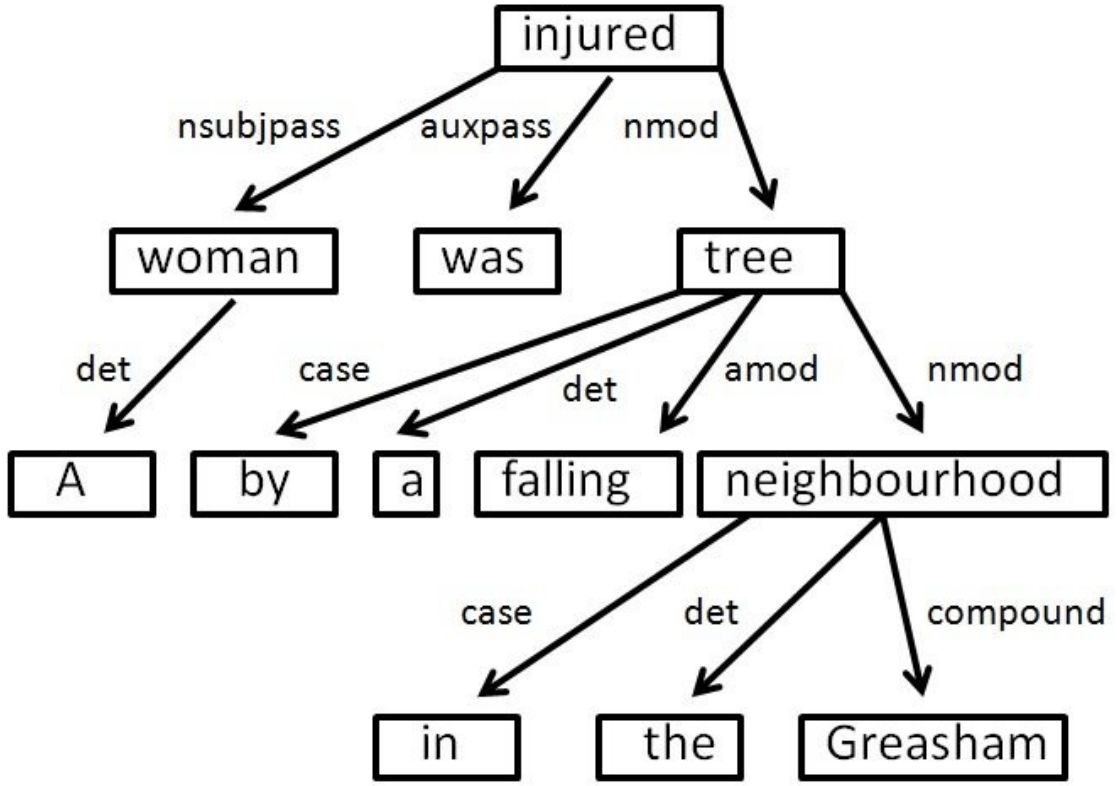
OVERVIEW OF LINGUISTICS

- Input: sound wave
- Phonetics
- Phonology
- **Morphology**
- Syntax
- Semantics
- Pragmatics
- Output: meaning



OVERVIEW OF LINGUISTICS

- Input: sound wave
- Phonetics
- Phonology
- Morphology
- **Syntax**
- Semantics
- Pragmatics
- Output: meaning



A woman was injured by a falling tree in the Greasham neighbourhood

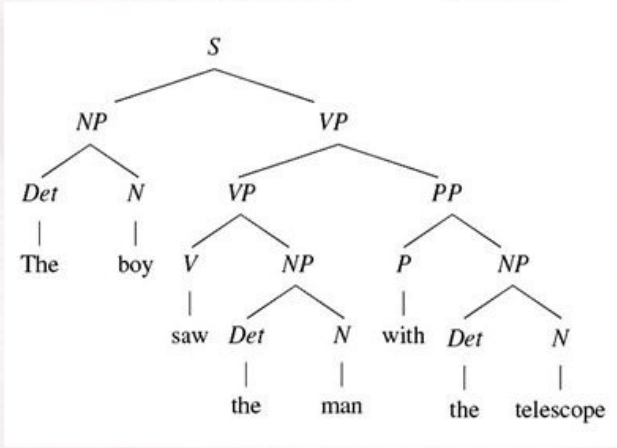
SYNTAX

Ambiguities

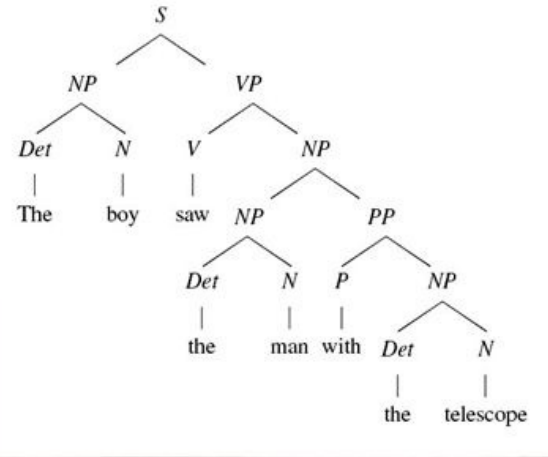
The boy saw the man with the telescope.

SYNTAX

Structural Ambiguities



- The boy used a telescope to see the man



- The boy saw the man who had a telescope

OVERVIEW OF LINGUISTICS

- Input: sound wave
- Phonetics
- Phonology
- Morphology
- Syntax
- **Semantics**
- Pragmatics
- Output: meaning

First Order Logic

$\exists x\ (woman(x) \wedge smokes(x))$

"a woman smokes"

$\forall x\ (woman(x) \rightarrow smokes(x))$

"all women smoke"

OVERVIEW OF LINGUISTICS

- Input: sound wave
- Phonetics
- Phonology
- Morphology
- Syntax
- **Semantics**
- Pragmatics
- Output: meaning

« To know the meaning of a [declarative] sentence is to know what the world would have to be like for the sentence to be true. »

Dowty & al., 1981, *Introduction to Montague Semantics*

OVERVIEW OF LINGUISTICS

- Input: sound wave
- Phonetics
- Phonology
- Morphology
- Syntax
- Semantics
- **Pragmatics**
- Output: meaning

« Reading between the lines »

Ann: “I have two children.”

Implied: Ann has exactly two children.

If she had more, the sentence would still be true…

Ann: “I’m out of gas.”

Bob: “There’s a gas station around the corner.”

Implied: The gas station is open.

OVERVIEW OF LINGUISTICS

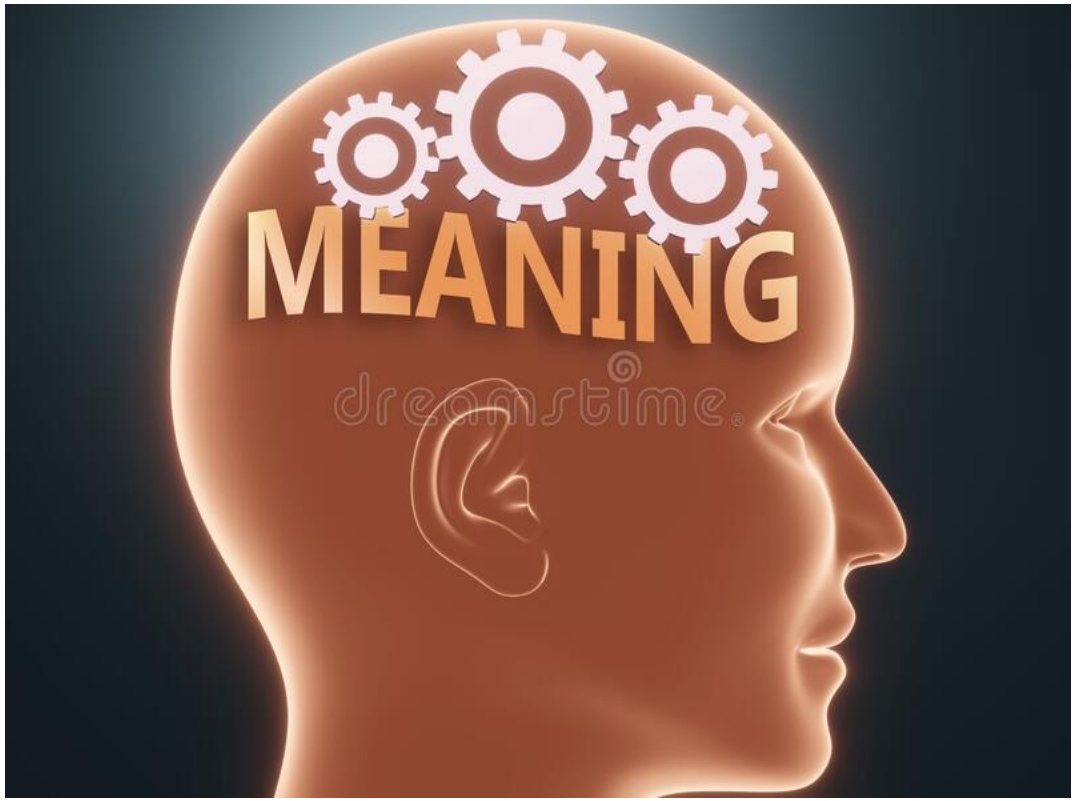
- Input: sound wave
- Phonetics
- Phonology
- Morphology
- Syntax
- Semantics
- **Pragmatics**
- Output: meaning

(1) The Story of the Mate and the Captain (Meibauer 2005, adapted from Posner 1980) :

A captain and his mate have a long-term quarrel. The mate drinks more rum than is good for him, and the captain is determined not to tolerate this behaviour any longer. When the mate is drunk again, the captain writes in the logbook: "Today, 11th October, the mate is drunk." When the mate reads this entry during his next watch, he gets angry.
Then, after a short moment of reflection, he writes in the logbook: "Today, 14th October, the captain is not drunk."

OVERVIEW OF LINGUISTICS

- Input: sound wave
- Phonetics
- Phonology
- Morphology
- Syntax
- Semantics
- Pragmatics
- **Output: meaning**



WHAT IS A CORPUS ?

- A book ?
- An article ?
- An archive ?

DEFINITION

(McArthur, Tom. (ed.) 1992. The Oxford Companion to the English. Oxford & New York: Oxford University Press.)

- CORPUS:
- (1) A collection of texts, especially if complete and self-contained: the corpus of Anglo-Saxon verse.
- (2) In linguistics and lexicography, a body of texts, utterances, or other specimens considered more or less representative of a language, and usually stored as an electronic database. Currently,computer corpora may store many millions of running words, whose features can be analyzed by means of tagging (the addition of identifying and classifying tags to words and other formations) and the use of concordancing programs.

DEFINITION

(Crystal, David. 1992. An Encyclopedic Dictionary of Language and Languages. Oxford: Blackwell.)

- A collection of linguistic data, either compiled as written texts or as a transcription of recorded speech. The main purpose of a corpus is to verify a hypothesis about language - for example, to determine how the usage of a particular sound, word, or syntactic construction varies. Corpus linguistics deals with the principles and practice of using corpora in language study. A computer corpus is a large body of machine-readable texts.

(John Sinclair. 1991. Corpus, Concordance, Collocation. Oxford: Oxford University Press.)

- A collection of naturally occurring language text, chosen to characterize a state or variety of a language.

SO A CORPUS IN NOT JUST ANY KIND OF TEXT…

- Sample/collection which is representative with regards to the research hypothesis/goal of the user
- Has a defined size and content

- Electronically stored
  - easier to obtain information on frequencies, grammatical patterns, collocations with a computer vs. Manually
  - costs of new analyses are lower compared to manual counting
- Should be freely available if used for research purposes, so the research results can be contrasted, compared and repeated

WHY USE CORPORA ?

- Objective verification of results (same study can be done by several parties to verify the results)
- Corpora show how people really use the language. They are meant to illustrate the rule vs. the exception
  - Vs. Linguistics generally, where single sentences are used as a perimeter of study **vs.** Full texts
- Quantitative data shows what occurs frequently and what occurs rarely in the language
- Thanks to computers, we can conduct fast, complex studies and process more material than by hand

WHAT IS CORPUS LINGUISTICS ?

- Corpus linguistics is a methodology to obtain and analyze the language data either quantitatively or qualitatively
- It can be applied in almost any area of language studies
- The object of a study is language used authentically, naturally and can encompass many different contexts (political speeches, tweets on twitter, news articles, novels, movies scripts…)
- Corpus linguistics is not a separate branch of linguistics (like e.g. sociolinguistics) or a theory of language

CORPUS LINGUISTICS, COMPUTATIONAL LINGUISTICS AND NATURAL LANGUAGE PROCESSING

- Fields overlap and are sometimes used interchangeably (computational linguistics and NLP especially)
- Computational linguists and NLP specialists often dicuss problems and solutions together.

NLP|CL TASKS

- ***Tokenization***

- Customer service could be better! = “customer” “service” “could” “be” “better” “!”

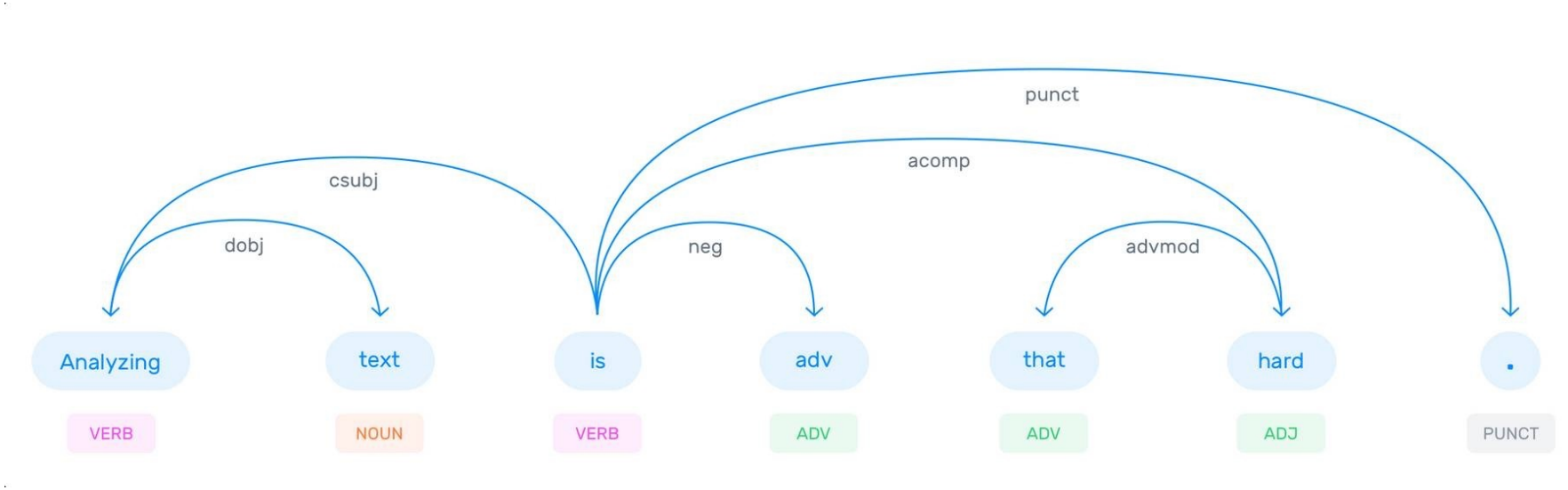
NLP|CL TASKS

- ***Part-of-speech tagging***

- *"Customer service": NOUN, "could": VERB, be": VERB, "better": ADJECTIVE, "!": PUNCTUATION*

NLP|CL TASKS

- ***Dependency Parsing***



NLP|CL TASKS

- ***Word Sense Disambiguation***
- *You should read this **book**; it's a great novel!*
- *You should **book** the flights as soon as possible.*
- *You should close the **books** by the end of the year.*
- *You should do everything by the **book** to avoid potential complications.*

NLP|CL TASKS

- ***Named Entity Recognition (NER)***
- *Susan lives in Los Angeles*

NLP|CL TASKS

- ***Text Classification***
- Sentiment analysis for example :
  - *"I really like the new design of your website!"* → Positive
  - *"I'm not sure if I like the new design"* → Neutral
  - *"The new design is awful!"* → Negative

MODERN-DAY NLP|CL APPLICATIONS

- E-mail filters
- Virtual assistants
- Online search engines
- Monitoring brand sentiment on social media
- Quickly sorting customer feedback
- Chatbots
- Automatic summarization
- Machine translation
- Natural language generation

DEMO : GENERATING TEXT

- Gpt2 demo → https://huggingface.co/gpt2