Australian National University

COMP4650/6490 Document Analysis

Autumn - 2021

Introduction

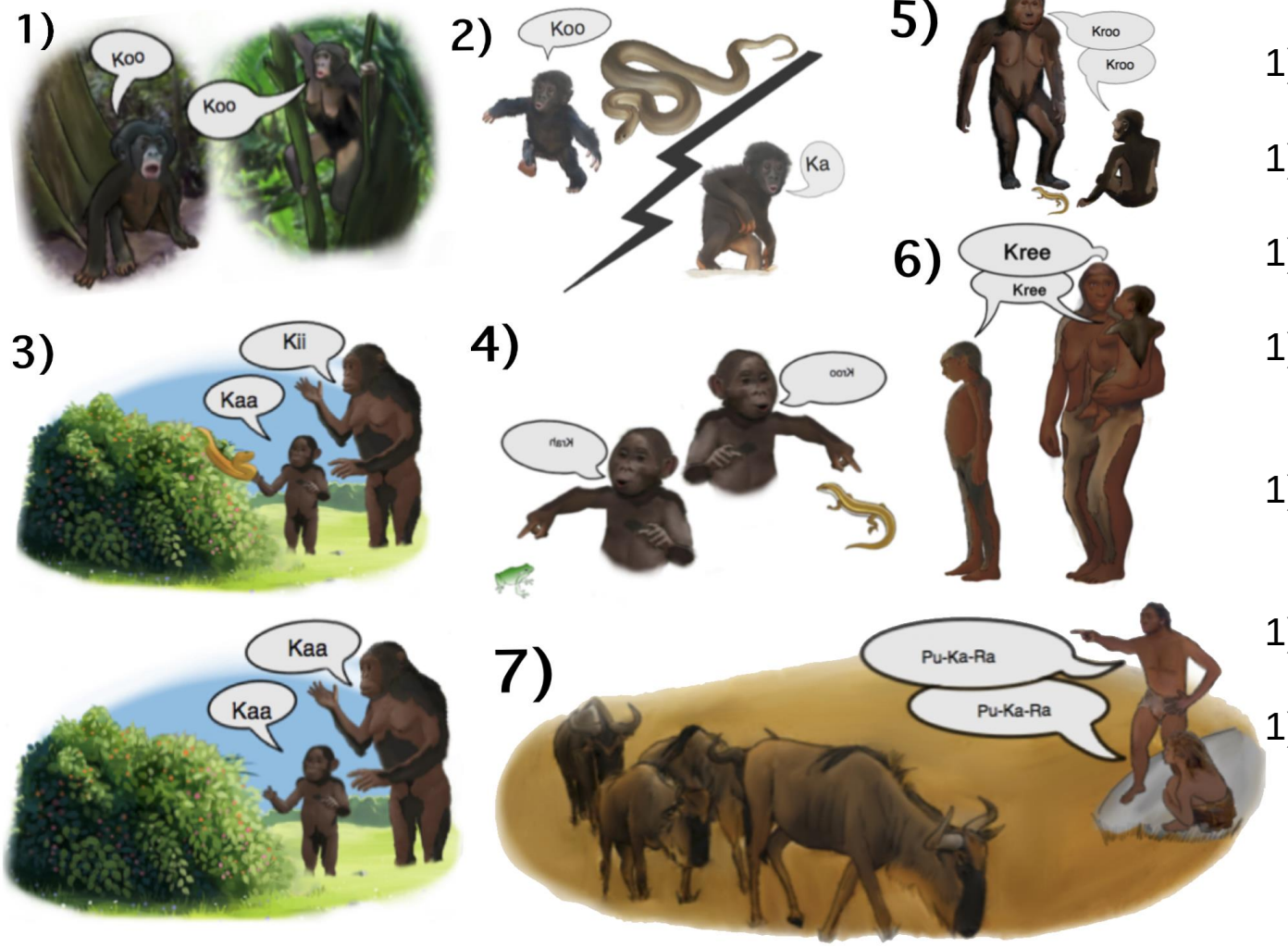
Research School of Computer Science, ANU

Outline of this lecture

- What is document analysis? Brief history
- Document Analysis fundamentals
- Course sections, course logistics

Brief History

‘from where to what’: hypotheses of the 7 stages of language evolution



1) Contact calls

1) Calls with entonation

1) Q&A conversations

1) Vocal ability to invent new words

1) Learn by imitating their lip-movements

1) Learn through mimicry

1) communication with

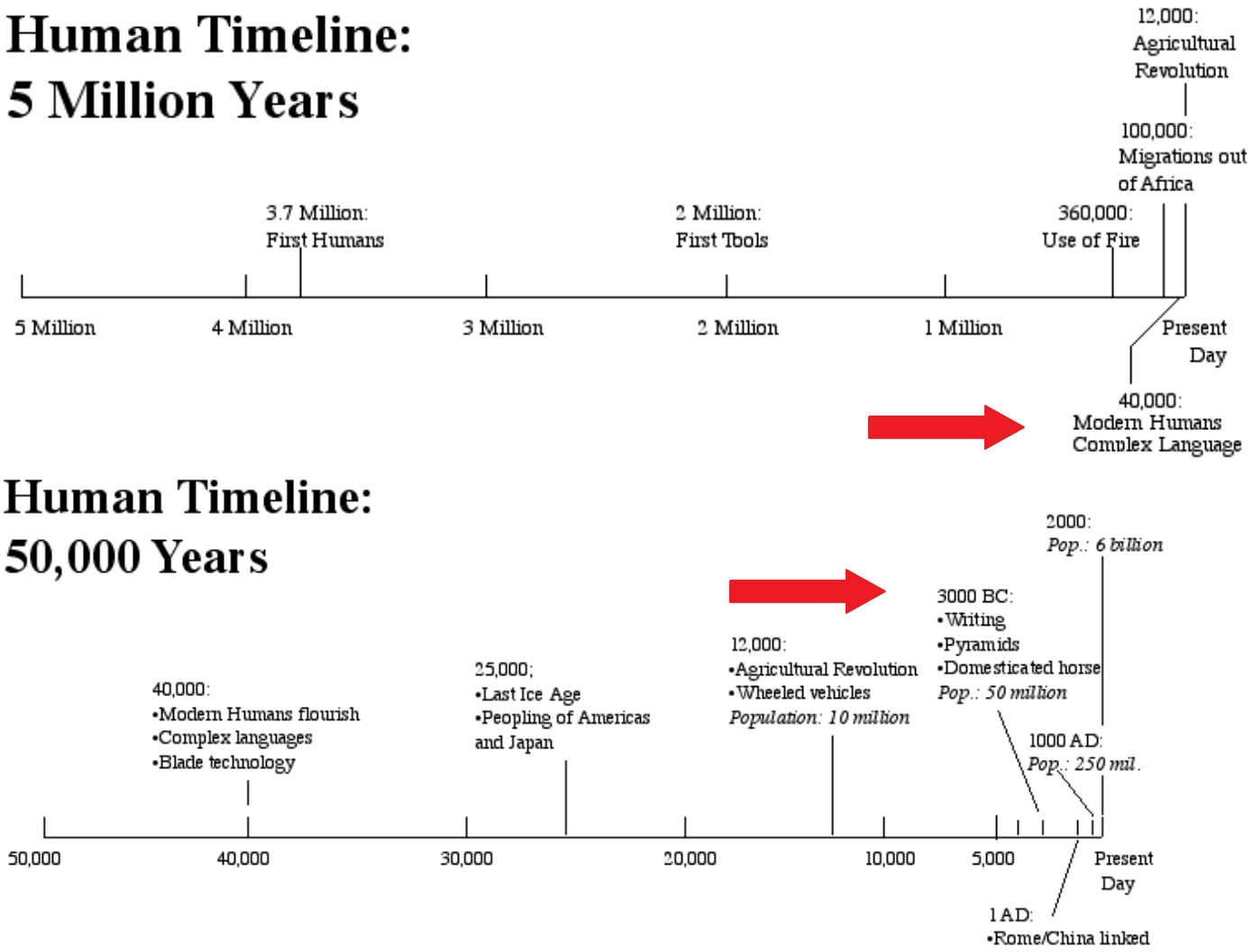
Brief History

‘From where to What’: hypotheses of the 7 stages of language evolution

- 1. Exchange of **contact calls** between mothers and offspring used to relocate each other in cases of separation.
- 2. Offspring of early Homo **modified the contact calls with intonations** in order to emit two types of contact calls: **contact calls that signal low level of distress** and **contact calls that signal high-level of distress**.
- 3. **question-answer conversation**. In this scenario, the offspring emits a low-level distress call to express a desire to interact with an object, and the mother responds with a low-level distress call to enable the interaction or high-level distress call to prohibit it.
- 4. The use of intonations improved over time, and eventually, individuals acquired **sufficient vocal control to invent new words to objects**.
- 5. At first, offspring **learned** the calls from their parents by **imitating their lip-movements**.
- 6. As the learning of calls improved, babies learned new calls (i.e., phonemes) through lip imitation only during infancy. After that period, the memory of phonemes lasted for a lifetime, and older children became capable of **learning new calls (through mimicry) without observing their parents' lip-movements**.
- 7. Increased vocabulary size. Further developments to the brain circuit responsible for rehearsing poly-syllabic words resulted with individuals capable of rehearsing lists of words (phonological working memory), which served as the platform for **communication with sentences**.

– *Based on the papers: Poliva, O. (2015). From where to what: a neuroanatomically based evolutionary model of the emergence of speech in humans. F1000Res. doi:10.12688/f1000research.6175.3. Poliva, O. (2016). From Mimicry to Language: A Neuroanatomically Based Evolutionary Model of the Emergence of Vocal Language. Front. Neurosci. 10, 1–21. doi:10.3389/fnins.2016.00307.*

40K years of talking before writing

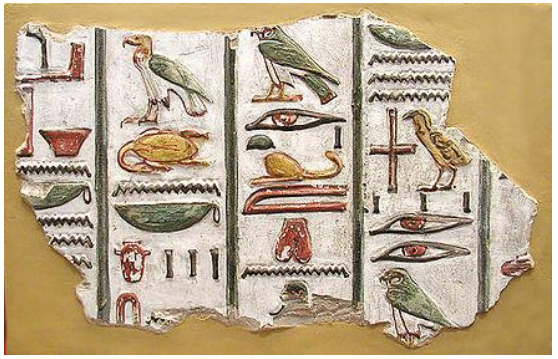


Brief History

Some of the oldest traces of writing (debated!)

Jiahu
6600 BC

Tărtăria tablets
5300 BC



Egyptian hieroglyphs
3200 BC

Brief History

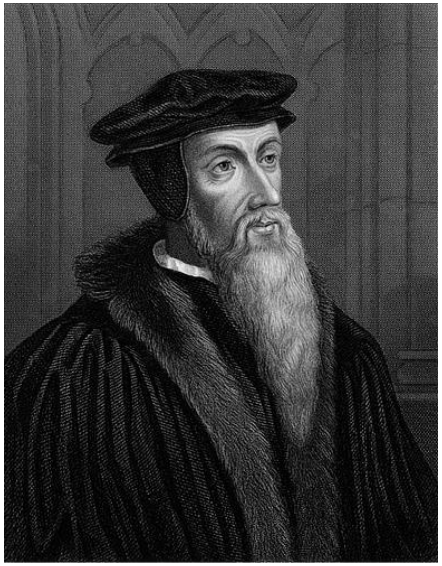
- Clay Tables 3000BC
  - We've come a long way
  - First archives
  - Beginning of first libraries!

Law

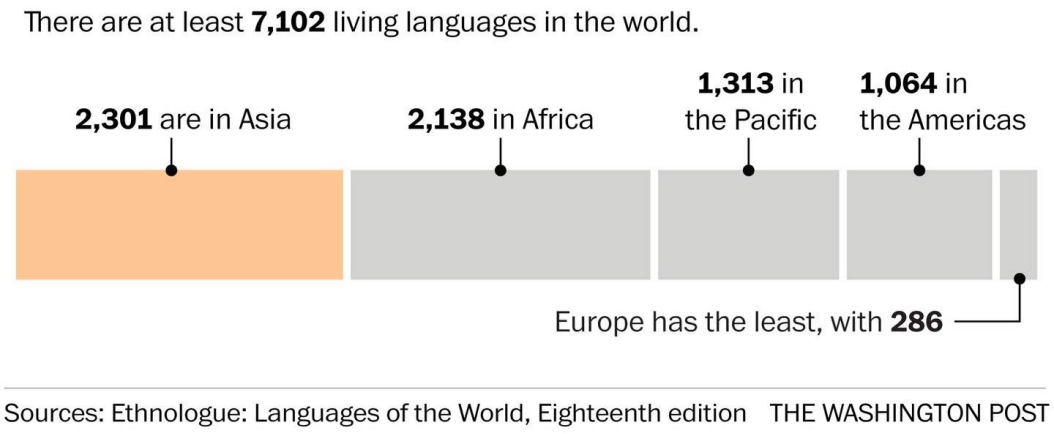
- Code of Hammurabi
  - First recorded law (1754 BC in the Mesopotamia)
  - 282 laws, eg ‘an eye for an eye, a tooth for a tooth’
- Transparent to all
  (at least to those who could read)
- Civilization lawful by definition?

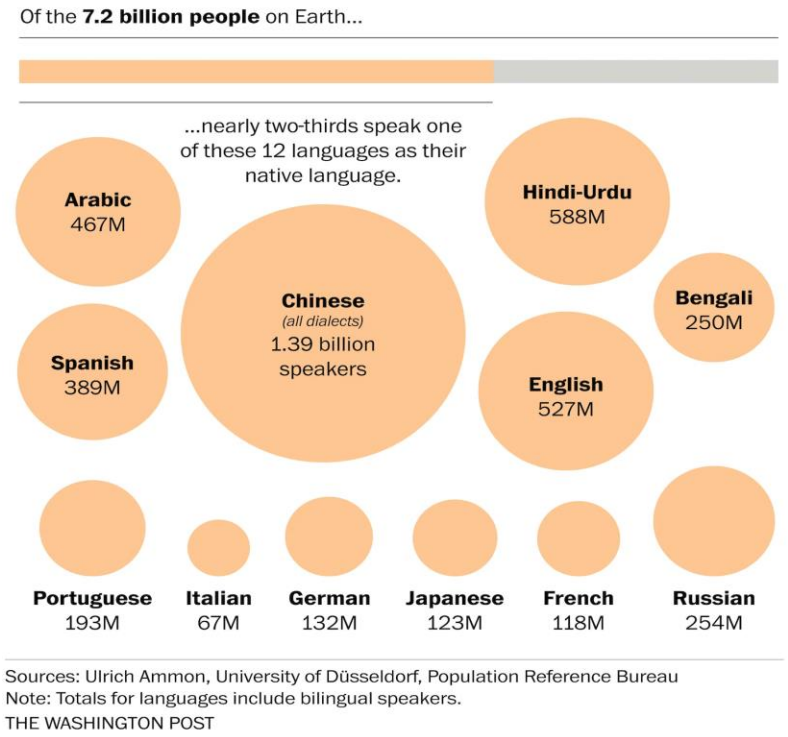
Religion

- Printing Revolution - Johannes Gutenberg printing press 1440
- 1500's: Spread of the Reformation
  - Need to refute
  - those "biblically
  - derived" claims?
- First you'll need a copy of the Bible…

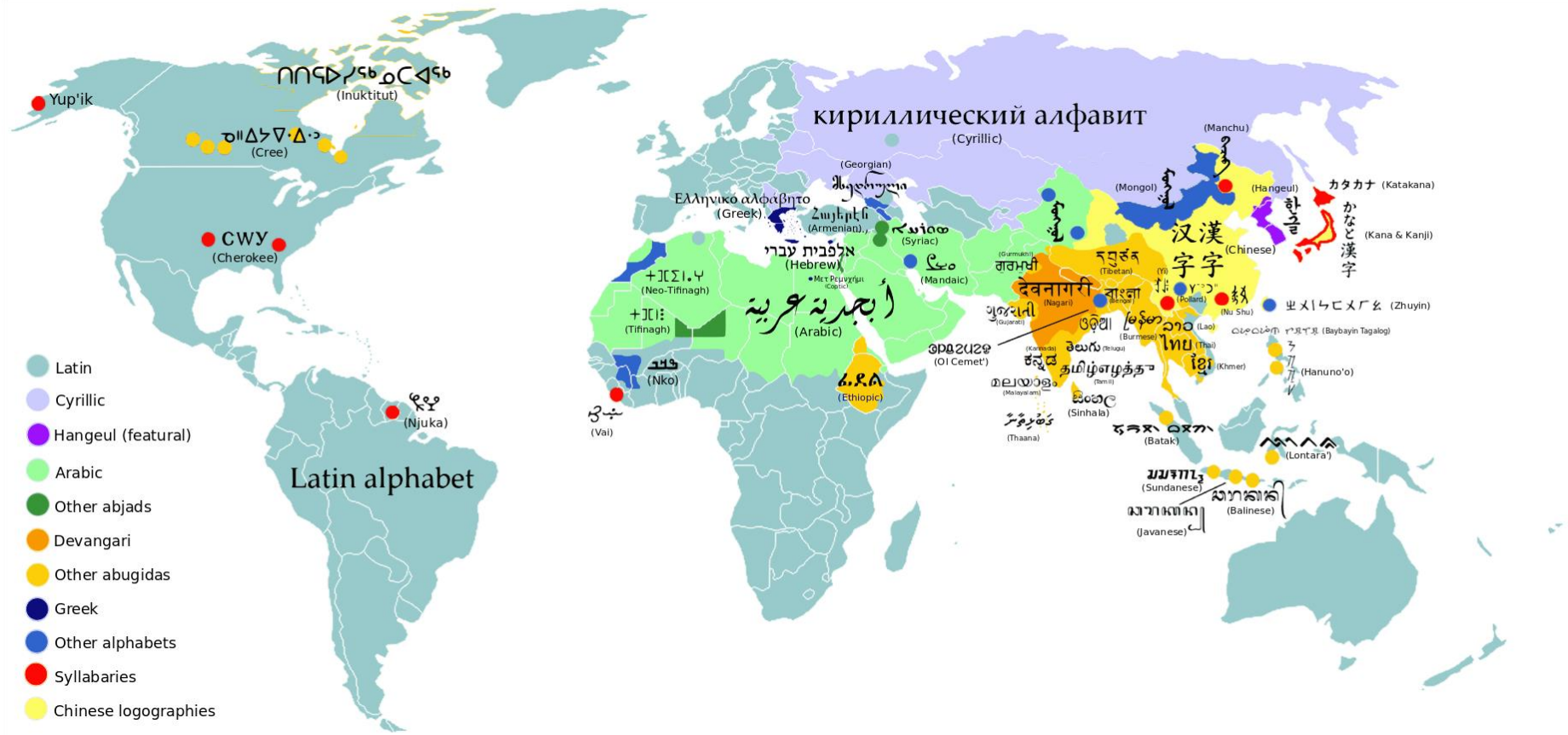


Languages in our planet

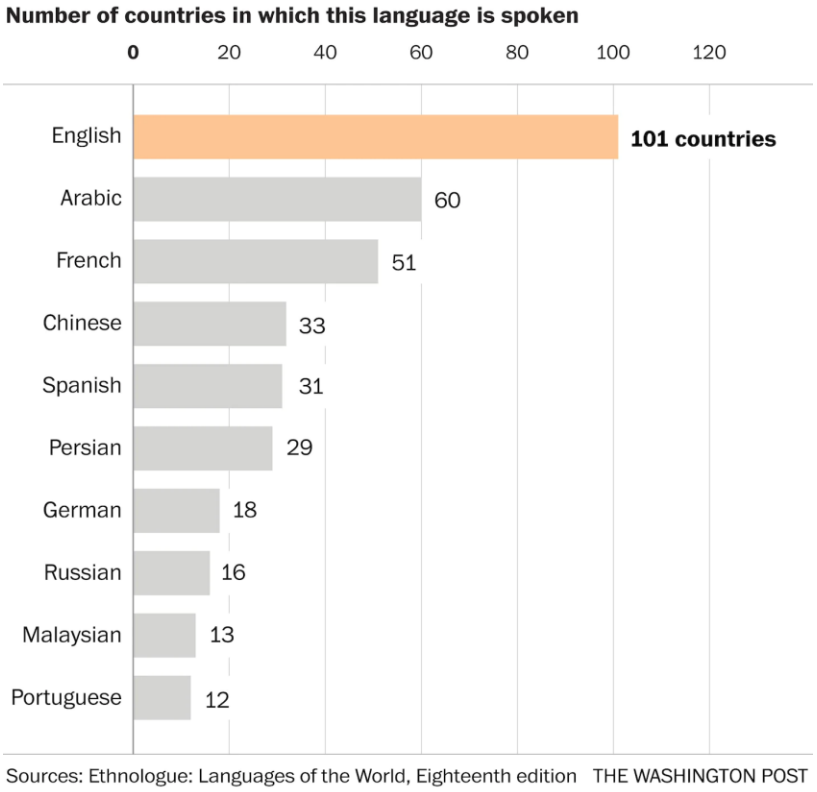


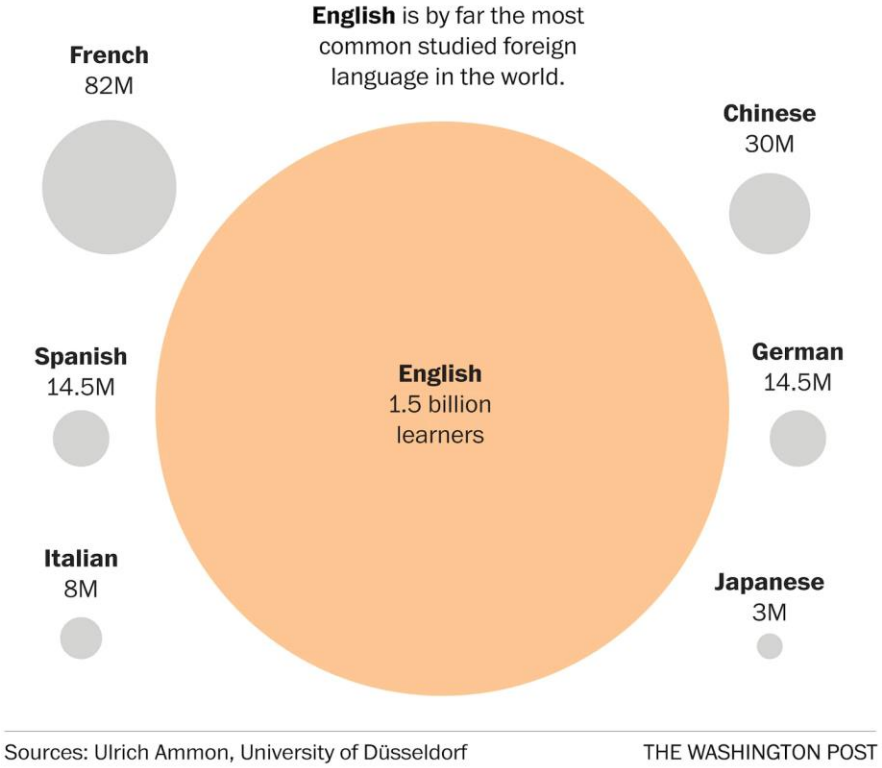


Writing systems of the world today



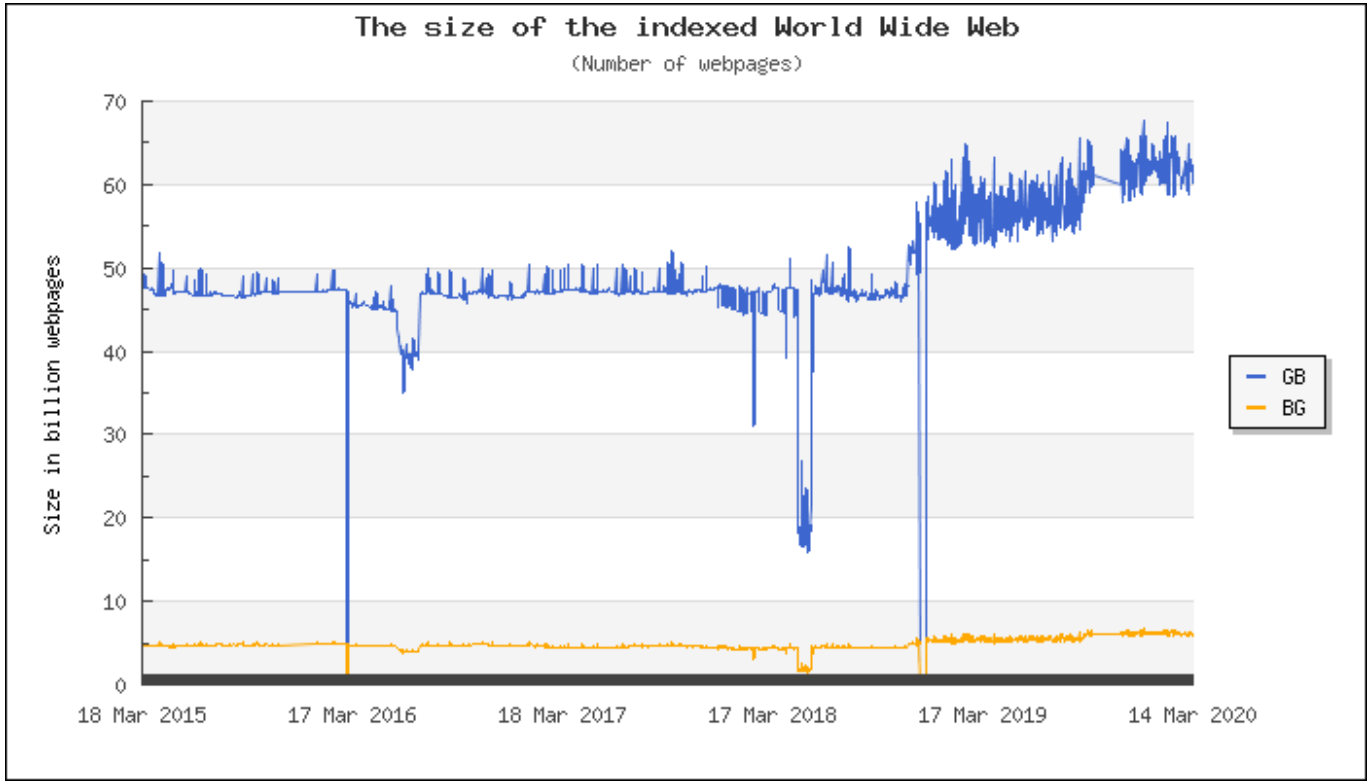
Nowadays


Number of countries in which this language is spoken
0
20
40
60
80
100
120
English
101 countries
Arabic
60
French
51
Chinese
33
Spanish
31
Persian
29
German
18
Russian
16
Malaysian
13
Portuguese
12
Sources: Ethnologue: Languages of the World, Eighteenth edition
THE WASHINGTON POST


English is by far the most common studied foreign language in the world.
French
82M
Chinese
30M
Spanish
14.5M
English
1.5 billion learners
German
14.5M
Italian
8M
Japanese
3M
Sources: Ulrich Ammon, University of Düsseldorf
THE WASHINGTON POST

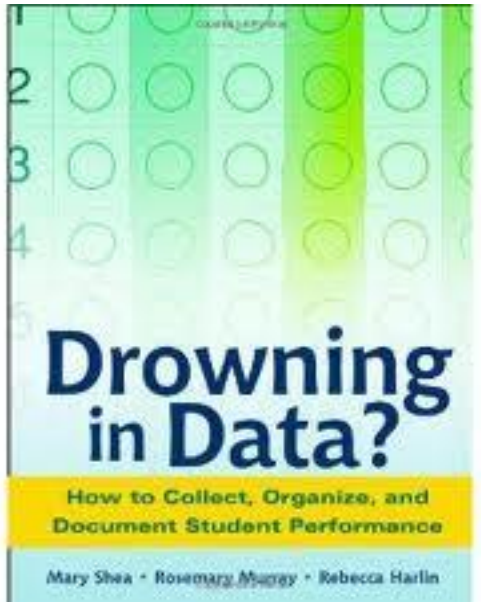
How BIG is the Internet? How much information is out there?

The Indexed Web contains at least 5.92 billion pages (Sunday, 15 March, 2020)



GB = Sorted on Google and Bing, BG = Sorted on Bing and Google: https://www.worldwidewebsize.com/

Why document analysis?



- How much data do we have? And most of it is still text.
  - think: Internet IP video traffic? # of facebook users? Amount of Facebook photos? Chatting, commenting, articles, etc.
- Data mining—Automated analysis of massive data sets

IR and NLP

- How do you *automatically* process enormous volumes of documents?
  - the focus of this Document Analysis course **Information Retrieval** section

- How computers can understand human language?
  - the focus of this Document Analysis course **Natural Language Processing** section

- What are the applications?
  - preview, more in the rest of this class

IR and NLP

An **Information Retrieval** system is a software system that provides access to webpages, books, journals and other documents; stores and manages those documents.

e.g., search engines (Web search, Enterprise search, etc.), information filtering (recommendation systems), media search (image retrieval, speech retrieval, news search, etc.)

A **Natural Language Processing** system is a software that aims to understand human language.

IR and NLP are deeply connected:

- IR can make use of NLP to improve search and retrieval
- NLP can make use of IR to find/rank/organize documents/sentences/words needed to solve an NLP task

IR and NLP

Both IR and NLP have to deal with human language and its characteristics:

Multilinguality

- there are many human languages!

Meaning

- where is the meaning of concepts/words/phrases coming from?

Ambiguity

How to deal with language ambiguity?

Linguist have structured language in different levels so that it can be study… later, we have used this knowledge to model language with computers

Levels of Language

Phonetics and phonology – knowledge about linguistic sounds ie distributional patterns of sounds, pronunciations

Morphology – knowledge of the meaningful components of words ie structure of words eg bathroom

Syntax – knowledge of the structural relationships between words ie how words combine to form phrases and sentences

Semantics – knowledge of meaning e.g., Queen (Royal, band?) i.e., how language conveys meaning

Discourse – knowledge about linguistic units larger than a single utterance e.g., in conversation, relationship of current sentence to previous

Pragmatics – knowledge of the relationship of meaning to the goals ie how language is used to do things

Levels of Language

How computers structure different levels of language?

Phonetics and phonology: *wave forms (not covered in this course)*

Morphology: *sequences*

Syntax: *trees*

Semantics: *trees or graphs*

Discourse: *trees or graphs*

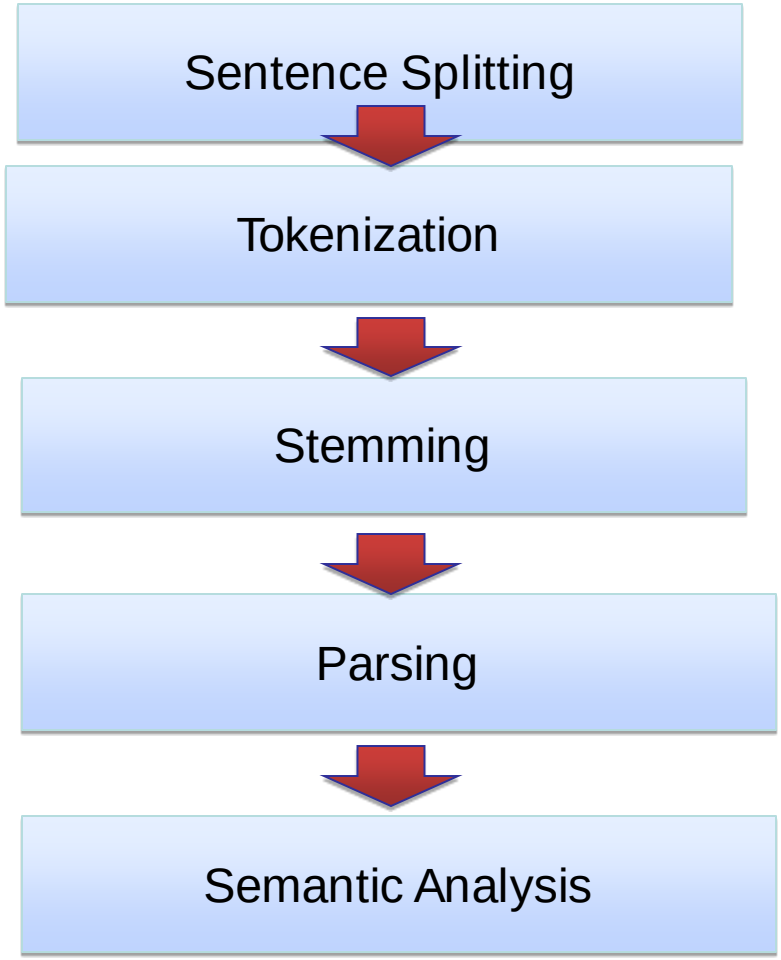
Pragmatics: *out of the scope of NLP… too hard*

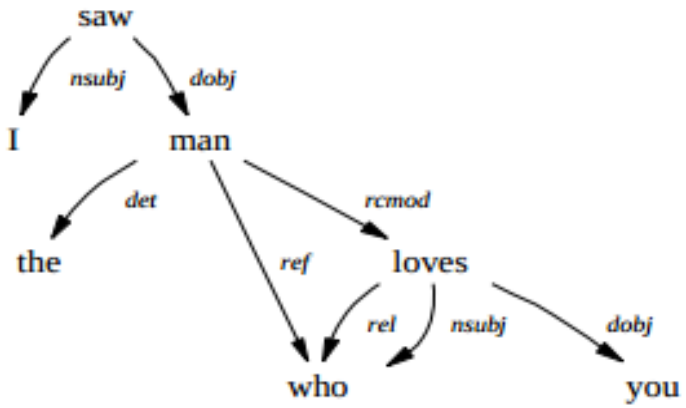
Ambiguity in language

I made her duck.

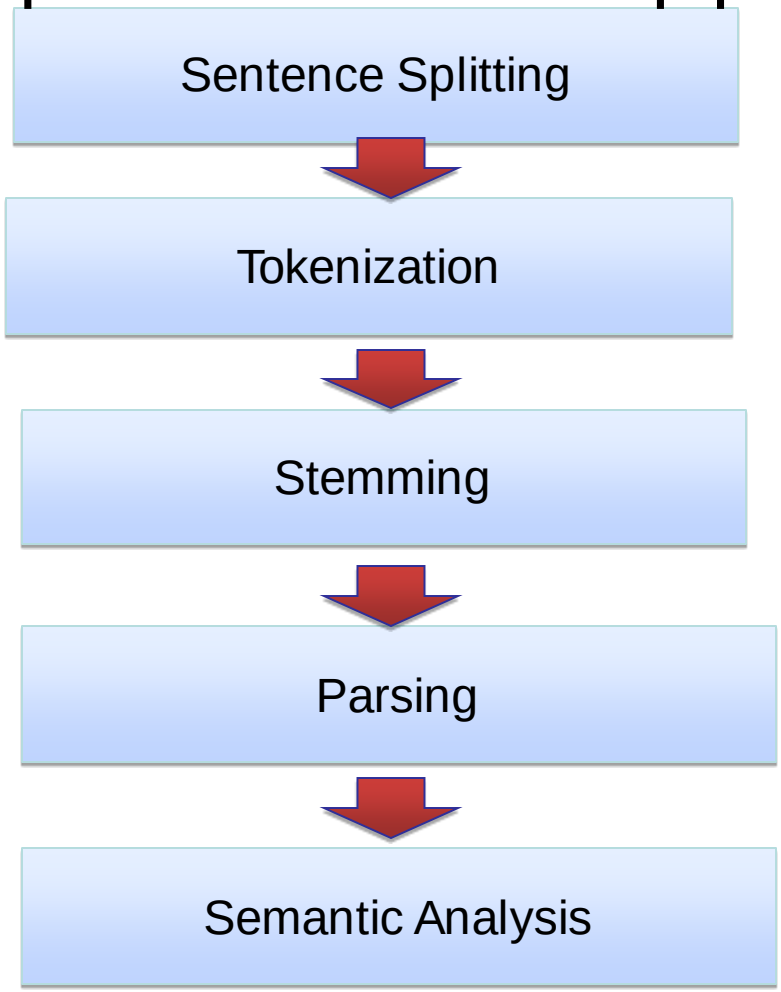
1) I cooked waterfowl for her
2) I cooked waterfowl belonging to her
3) I created the duck she owns
4) I caused her to quickly lower her head or body
5) I waved my magic wand and turned her into a waterfowl

A Popular IR and NLP Pipeline





A popular IR and NLP pipeline



→ Morphology: *sequences → IR, NLP*

→ Morphology: *sequences → IR, NLP*

→ Syntax: *trees→NLP*

→ Semantics: *trees or graphs→IR, NLP*

→ Discourse: *trees or graphs→NLP*

Sentence splitting

Problems with sentence splitting

"You reminded me," she remarked, "of your mother."

Sentence boundary detection algorithms

- Regular expressions
- Rule-based approaches
- Machine learning approaches

Tools

- OpenNLP – supports most common NLP tasks
- Stanford CoreNLP – useful NLP tools
- NLTK – Python NLP platform

Tokenization

Divide text into words, numbers, punctuation

Regular expressions for English

E.g. [A-Za-z0-9]+|{\Punct}+

Problems

- Periods: Ph. D., google.com
- Clitics: isn't => *is* + *net* (*not)*
- Hyphenation
- *co-operate*
- *most-visited => most visited*

Tools

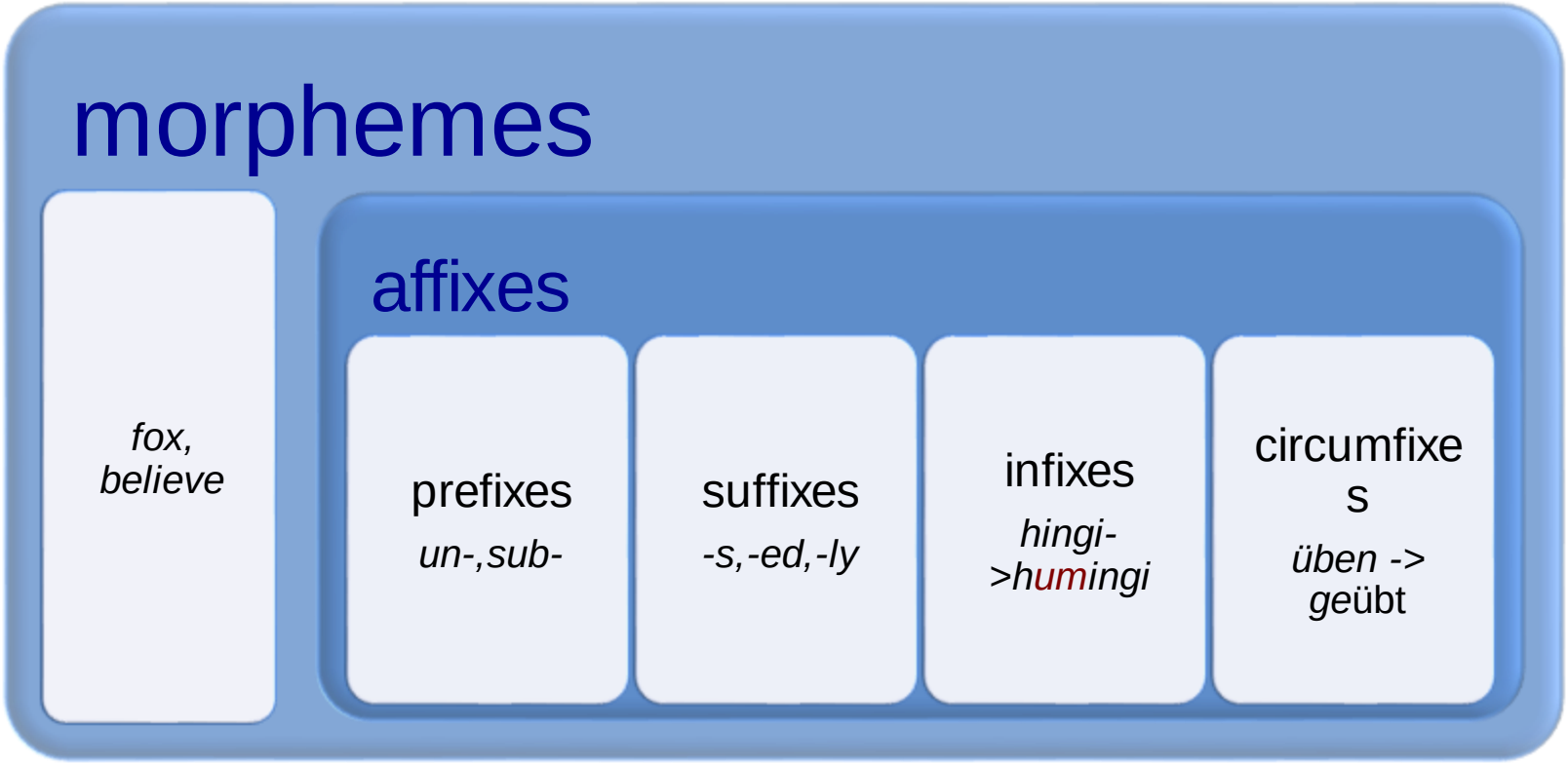
- OpenNLP
- Stanford CoreNLP
- NLTK

- Language dependent problem!

Morphology

Morphology is the study of the way words are built up from smaller meaning-bearing units

Morpheme: minimal meaning-bearing unit in a language



Language dependent problem: some languages have very rich morphology (e.g., Czech, French), others poor (e.g., English)

Stemming

A crude heuristic process that strips off suffixes

e.g., studies → stem = studi, suffix = es

Algorithms

- Lookup algorithms
- Regular expressions
- Suffix-stripping algorithms

Tools

- http://snowball.tartarus.orghttp://text-processing.com/demo/stem/

Part of Speech (POS)Tagging

The task of assigning grammatical categories to tokens

Closed class*:* categories are composed of a small, fixed set of grammatical function words for a given language

Pronouns, Prepositions, Modals, Determiners, Particles, Conjunctions

Open class: categories have large number of words and new ones are easily invented

Nouns (Googler), Verbs (google), Adjectives (geeky), Abverb (chompingly)

Parts of Speech (POS)Tagging

English POS categories

Noun (person, place or thing)

- Singular (NN): dog, fork
- Plural (NNS): dogs, forks
- Proper (NNP, NNPS): John, Springfields
- Personal pronoun (PRP): I, you, he, she, it
- Wh-pronoun (WP): who, what

Verb (actions and processes)

- Base, infinitive (VB): eat
- Past tense (VBD): ate
- Gerund (VBG): eating
- Past participle (VBN): eaten
- Non 3rd person singular present tense (VBP): eat
- 3rd person singular present tense: (VBZ): eats
- Modal (MD): should, can
- To (TO): to (to eat)

Part of Speech (POS)Tagging

English POS categories (cont.)

Adjective (modify nouns)

- Basic (JJ): red, tall
- Comparative (JJR): redder, taller
- Superlative (JJS): reddest, tallest

Adverb (modify verbs)

- Basic (RB): quickly
- Comparative (RBR): quicker
- Superlative (RBS): quickest

Preposition (IN): on, in, by, to, with

Determiner:

- Basic (DT) a, an, the
- WH-determiner (WDT): which, that

Coordinating Conjunction (CC): and, but, or…

Particle (RP): off (took off), up (put up)

Part of Speech (POS)Tagging

John	saw	the	saw	and	decided	to	take	it	to	the	table.
NNP	VBD	DT	NN	CC	VBD	TO	VB	PRP	IN	DT	NN

Part of Speech (POS)Tagging

Language dependent problem! Different languages have different POS Tags sets

- **English Brown corpus** used a large set of **87 POS-Tags**
- Most common for **English** now is the **Penn Treebank set of 45 tags**
- **Universal POS-Tags**
  - Multilingual reduced tag set (all languages have verbs, most languages have nouns, adjectives, etc.)
  - http://universaldependencies.org/u/pos/

Part of Speech (POS)Tagging

Tools

spaCy

https://spacy.io/

Stanford POS tagger

http://nlp.stanford.edu/software/corenlp.shtml

http://nlp.stanford.edu/software/tagger.shtml

Open NLP

UIUC POS tagger

http://cogcomp.cs.illinois.edu/demo/pos/results.php

NLTK

http://text-processing.com/demo/tag

Twitter POS tagger

http://www.ark.cs.cmu.edu/TweetNLP/

http://www.ark.cs.cmu.edu/TweetNLP/cluster_viewer.html

Course outline

1) Information Retrieval

2) Machine learning for documents

3) Natural Language Processing

4) Natural Language Processing in practice

1. Information Retrieval

Lecturer: Jooyoung Lee

- **Introduction to IR**
  - A model of a search system, Boolean retrieval
- **Beyond Boolean retrieval**
  - Ranked retrieval, tf.idf and the vector-space models, and Tolerant Retrieval
- **Evaluating IR systems**
  - Batch evaluation, interactive evaluation, online evaluation
- **Latent Semantic Indexing**
  - Singular Value Decomposition (SVD)
- **Web Search**

2. Machine Learning

Lecturer: Alex Mathews

Learning with supervision:

- Linear classification
- Logistic Regression

Representation

- Bag-of-word model: one-hot, Tf-idf
- Word vector: Word2Vec, Glove

Non-linear classification

- Neural Network
- Recurrent Neural Network
- Recursive Neural Network
- Convolutional Neural Network and Transformer [for unknown structure]

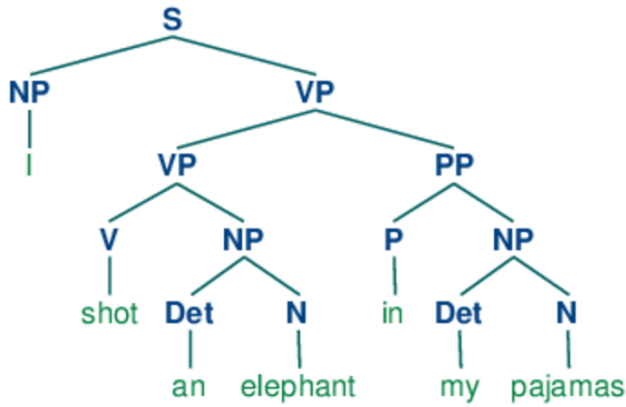
Learning without supervision:

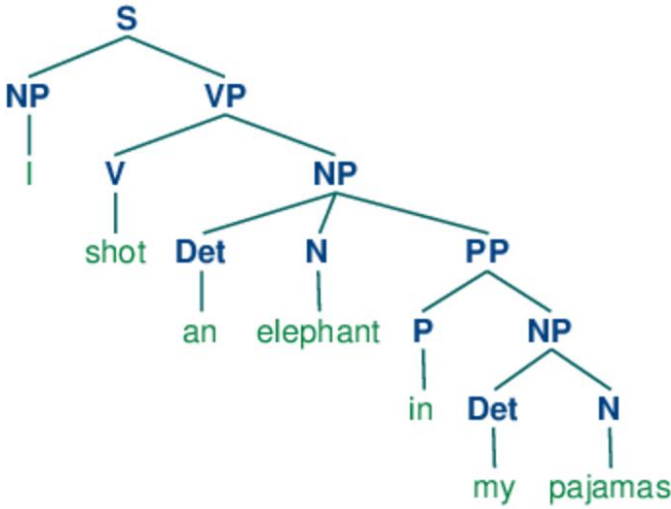
- Unsupervised learning: clustering
- Self-supervised learning: word vector, ELMO, BERT

Meaning

Lecturer: Nathan Elazar

- Logical semantics
- Predicate argument semantics
- Distributional and distributed semantics
- Reference resolution
- Parsing
  - Constituency Parsing
  - Dependency Parsing
- Language modelling





Lecturer: Nathan Elazar

NLP Tasks

- Tokenization
- Part-of-Speech tagging
- Syntactic parsing
- Semantic parsing
- Discourse parsing
- NER
- Relation extraction
- Sentiment analysis
- Word sense disambiguation
- Summarization

Lecturer: Gabriela Ferraro

Evaluation in NLP

- Precision, recall and F-measure
- Term overlap metrics: ROUGE, BLEU, SARI etc.
- Threshold free metrics ROC
- Classifier comparison
- Comparison between algorithms, features sets (ablation)
- Multiple comparison (optional)
- Semantic evaluation

Multi-lingual NLP

Low resource NLP

Part 3.
Course Logistics

See Course Outline and Schedule as posted on Wattle

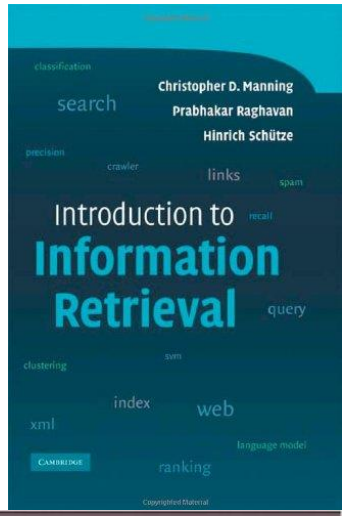
- Course Convenor/Administrative Questions
  - Alex Mathews (alex.mathews@anu.edu.au)
- Lecturers
  - Alex Mathews
  - Jooyoung Lee (recorded)
  - Gabriela Ferraro (recorded)
  - Nathan Elazer (recorded)
- Tutor
  - Qiongkai Xu
- Course Questions
  - Wattle Forum

Reference Books (not required)

- ***Introduction to Information Retrieval***. C.D. Manning, P. Raghavan and H. Schutze.
- Cambridge Univ. Press, 2008.

  PDF online:

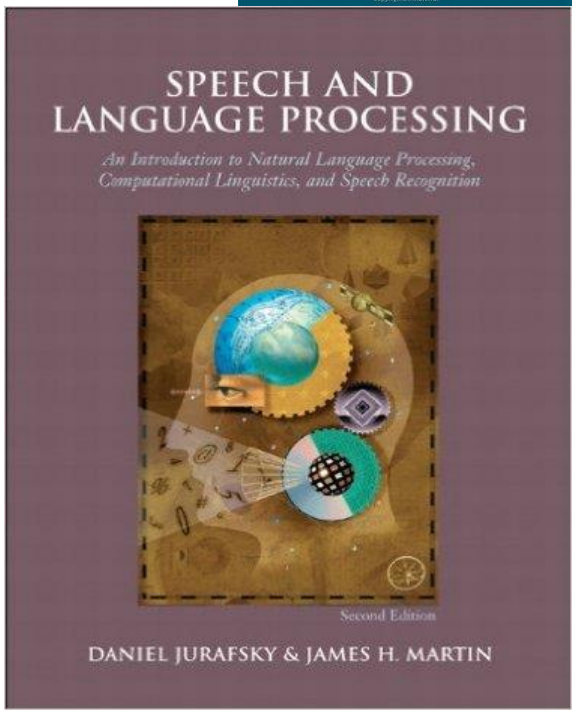
  http://nlp.stanford.edu/IR-book/information-retrieval-book.html



- ***Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition***. Dan Jurafsky, James H. Martin, Prentice Hall, 2009.

  Draft of 3rd edition:

  https://web.stanford.edu/~jurafsky/slp3/



Lectures

~2 hours per week uploaded to echo 360 (may be split across more than 2 videos)

Drop-in session (Not compulsory, no set agenda)

12-1 every Wednesday from week 2

LABS

During the intensive week only

Intensive Week

- Online via zoom 10am – 5pm Mon-Fri (week starting 28th June)
- 2 Lectures and 1 lab every day (live)
- 3 Guest lectures

- **Wattle**
  - Lecture slides will be posted on wattle
  - Quizzes and assignments will be provided and submitted through wattle
  - The exam will be conducted through wattle
- **Questions**
  - All course/content/exam related questions to be posted to the Wattle Forum
  - All admin/personal requests to be emailed to convenor
- **Information**
  - Check Wattle News/Discussion forums regularly!

- **Assessments**
  - See Course Outline for details
  - 3x Assignments (45%)
  - 4x Quizzes (5%)
  - Final Exam (50%)
- Coding in Python required for Assignments – see the 'Introduction to Python' tutorial on wattle
- No late assignment submission without prior approval
- No group work is permitted in any part of the assessment in this course. **Plagiarism will not be tolerated**

What to do next?

1. Review the course outline and the course schedule.

2. Fill in the survey. Ideally, you should answer all the questions, but only the first 2 are required.

3. Make sure you are familiar with python programming. If not, or if you need a refresher go through the basics tutorial.

4. Start reading any parts of the course textbooks that interest you.

5. Ask any questions you have about the course on the wattle discussion forum(or let me know if you find any errors or ambiguities).