Machine Learning For Design

Lecture 5 - Machine Learning and Natural Language Processing / Part 1

Alessandro Bozzon

02/03/2022

mlfd-io@tudelft.nl
www.ml4design.com

Admin

- Groups Assignment
  - Received 25 out of 26 groups assignments
  - Missing G1
  - Few students had issues with finding a group

Contact me!

- Few students attempted weekly quizzes
  - At least, not as many as I would have hoped for!

- Even less students are submitting questions
  - But the ones who did: thank you! Brilliant!

- Received only **1** :(feedback form about module 1
  - We need your feedback!

Previously, on ML4D….

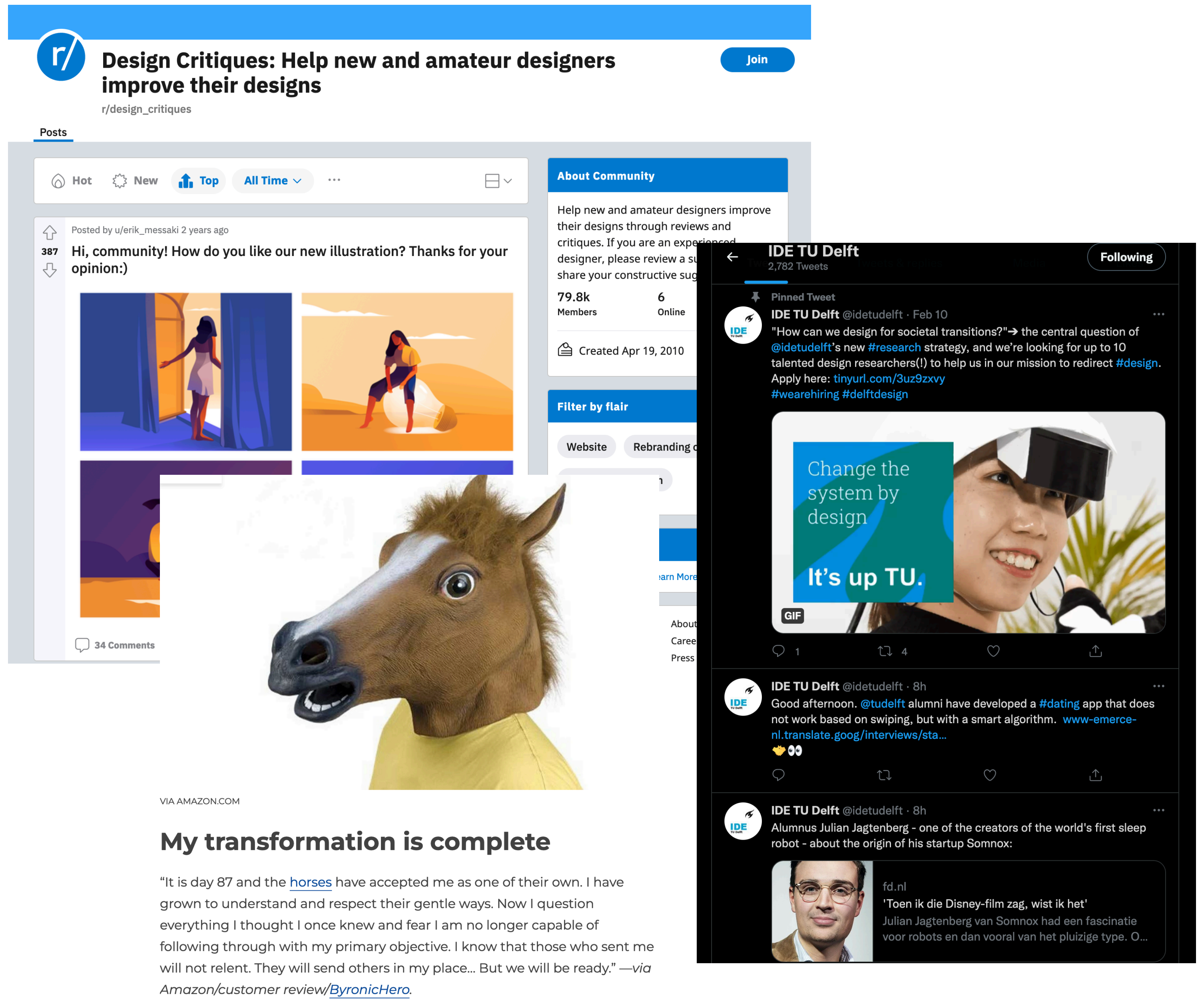
Natural Language Processing

- A sub-field of AI and machine learning in which machines learn to understand natural language as spoken and written by humans
- Goals:
  - Recognize the language, understand it, and respond to it
  - Categorise textual content (e.g. spam vs. Not-spam)
  - Translate between languages
  - Generate new text
- An enabler for technology such as chatbots and digital assistants like Siri or Alexa

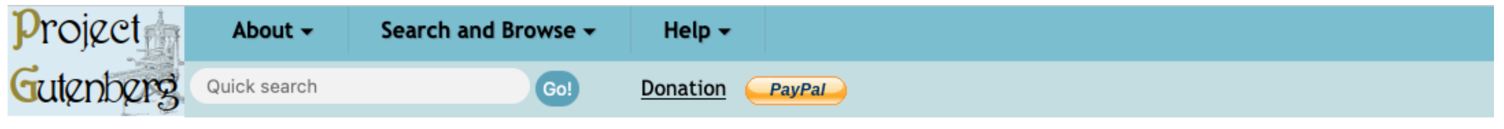
Why natural language processing?

And why is it a hard problem?

Fora, social media, blog, products review



Books (digital, or digitised)



Frequently Viewed or Downloaded

These listings are based on the number of times each eBook gets downloaded. Multiple downloads from the same Internet address on the same day count as one download, and addresses that download more than 100 eBooks in a day are considered robots and are not counted.

Downloaded Books

2022-02-27	156396
last 7 days	1167285
last 30 days	4234525

- Top 100 EBooks yesterday
- Top 100 Authors yesterday
- Top 100 EBooks last 7 days
- Top 100 Authors last 7 days
- Top 100 EBooks last 30 days
- Top 100 Authors last 30 days

Top 100 EBooks yesterday

Interviews

Interviewee: XXX
Interviewer: XXX
Date of Interview: mm.dd.yy
Location of Interview: XXX
List of Acronyms: FP=Frank Peterson, IN=Interviewer

[Begin Transcript 00:00:10]

IN: So what was going on in your life when you joined the Marines?

FP: Well when I joined the navy, actually that was in 1950 at the age of 18. Not much other than the fact that I wanted to get away from Topeka and see what the rest of world was really all about.

IN: Um-hm.

[00:00:26]

And of course having... gone through the flight training I received my wings and commission in October of 1952. And the- one of the reasons I opted for the Marines, I knew there had never been a black pilot in the Marine Corps. So I wanted to see if I could achieve that goal, which I was able to do.

And then my first duty assignment would have been in Cherry Point, North Carolina. But I'd had enough of the South and decided I wanted to stay away from the South if I possibly could, so Headquarters Marine Corps, at my request, changed my orders to El Toro, El Toro, California.

But what I didn't realize is that I'd jumped from the frying pan into the fire because El Toro was the training base for replacement pilots in Korea. So I jumped from the frying pan into the Korean War via El Toro.

IN: I see.

[End Transcript 00:01:21]

Bo: An intelligent network agent to promote physical activity in children with Congenital Heart Defects

- Analysis of how parents perceive their baby, their behaviours towards their child, and thus understand how does overprotection develops throughout childhood
- >300 stories, manually and NLP analysis

http://resolver.tudelft.nl/uuid:fd895415-c353-41d5-8430-f0a67fd40ad4

Big Textual Data = Language at scale

- One of the largest reflections of the world, a man-made one
- Essential to better understand people, organisations, products, services, systems
  - and their relationships!
- Language is a proxy for human behaviour and a strong signal of individual characteristics
  - Language is always situated
  - Language is also a political instrument

Why NLP?

- Answer questions using the Web
- Translate documents from one language to another
- Do library research; summarize
- Archive and allow access to cultural heritage
- Interact with intelligent devices
- Manage messages intelligently
- Help make informed decisions
- Follow directions given by any user
- Fix your spelling or grammar
- Grade exams
- Write poems or novels
- Listen and give advice
- Estimate public opinion
- Read everything and make predictions
- Interactively help people learn
- Help disabled people
- Help refugees/disaster victims
- Document or reinvigorate indigenous languages

Can you imagine other purposes?

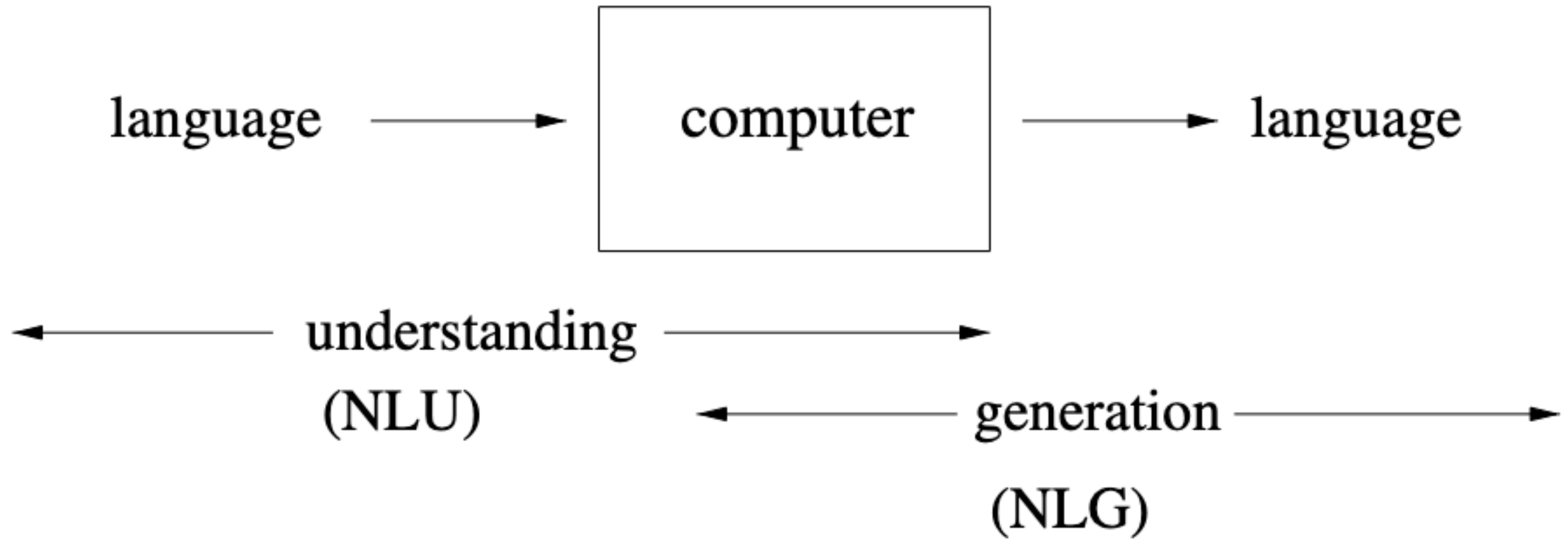
Natural Language Processing

- Computers using natural language as input and/or output

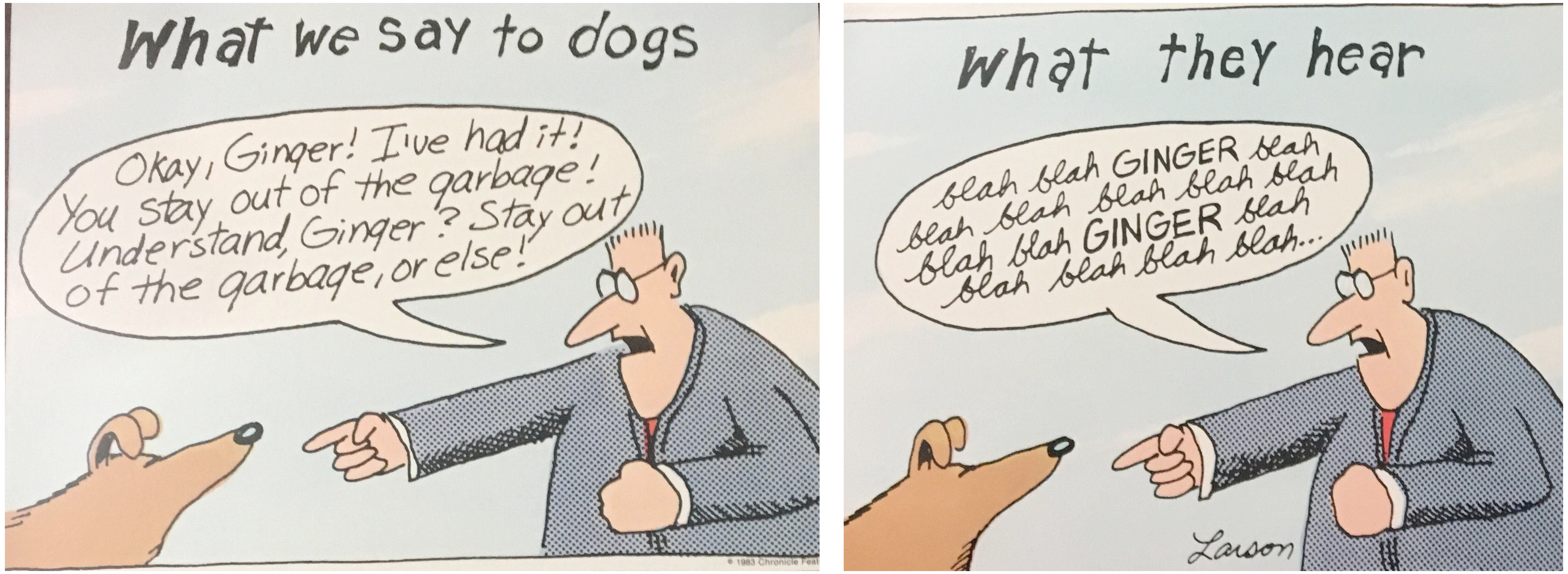
N atural: human communication, unlike e.g., programming languages

L anguage: signs, meanings, and a code connecting signs with their meanings

P rocessing: computational methods to allow computers to \`understand’, or to generate



Go beyond keyword matching



- Identify the **structure** and **meaning** of **words**, **sentences**, **texts** and **conversations**
- Deep understanding of broad language

NLP is hard

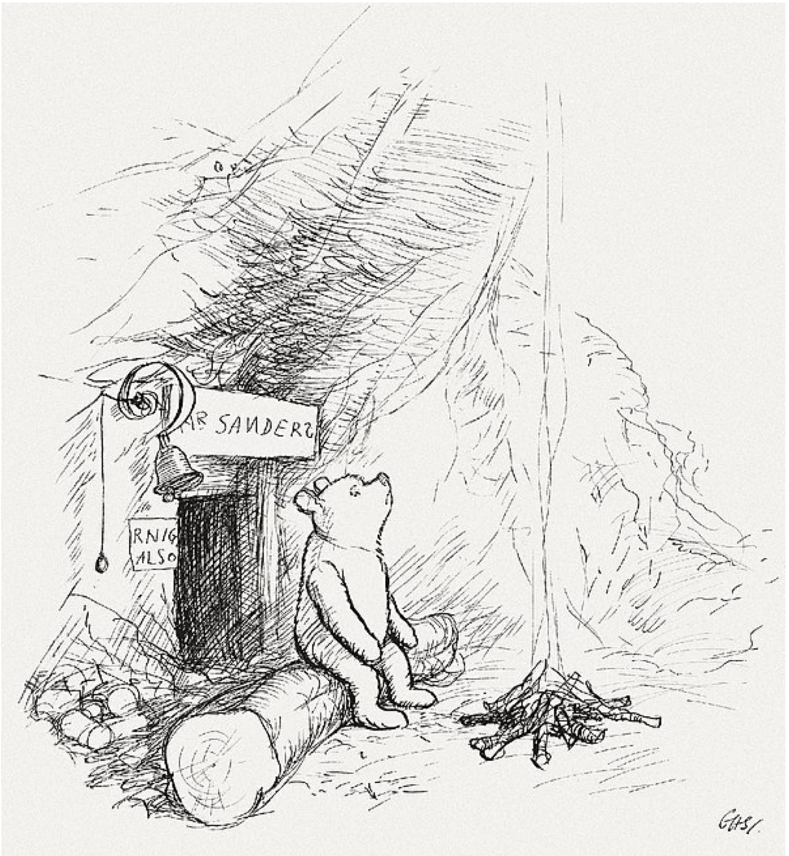
- Human languages are messy, ambiguous, and ever-changing
  - A string may have many possible interpretations at every level
  - The correct resolution of the ambiguity will depend on the intended meaning, which is often inferable from the context

- There is tremendous diversity in human languages
  - Languages express the same kind of meaning in different ways
  - Some languages express some meanings more readily/often

- Knowledge Bottleneck
  - Knowledge about language
  - Knowledge about the world
    - Common sense
    - Reasoning

Ambiguity and Expressivity

Christopher Robin is alive and well. **He** is the same person that you read about in the book, **Winnie the Pooh**. As a boy, **Chris** lived in a pretty home called **Cotchford Farm**. When **Chris** was three years old, **his father** wrote a poem about **him**. The poem was printed in a magazine for others to read. **Mr. Robin** then wrote a book



- Who wrote **Winnie the Pooh**?
- Where did **Chris** live?

Sparsity

- **Zipf's Law: The distribution of word frequencies is very skewed**

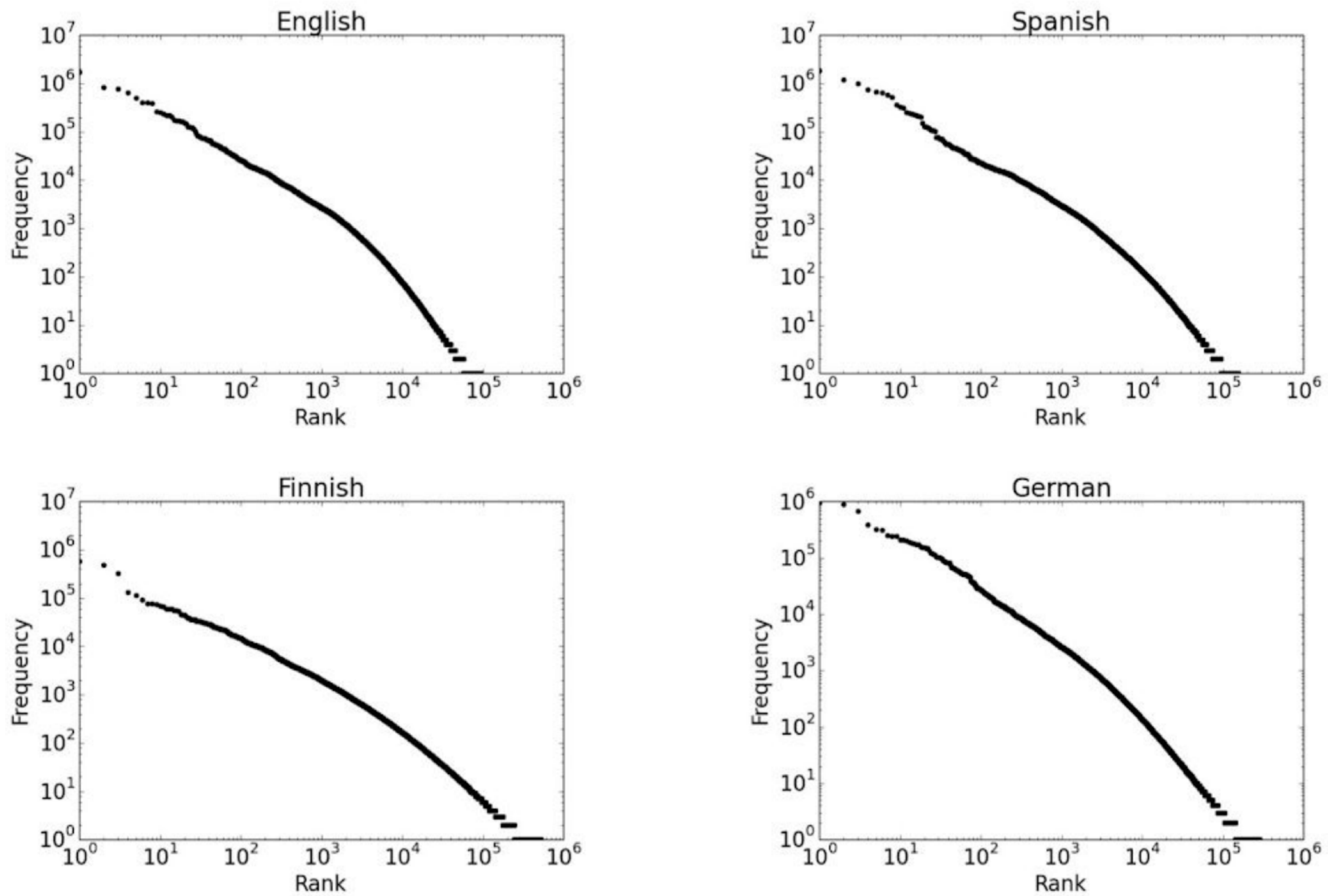
"... given some document collection, the frequency of any word is inversely proportional to its rank in the frequency table..."

- The most frequent word will occur approximately twice as often as the second most frequent word, which occurs twice as often as the fourth most frequent word, etc.
  - Regardless of how large our corpus is, there will be a lot of infrequent words

- This means we need to find clever ways to estimate the value of words that we have **rarely** (or **never**) seen

any word		nouns	
Frequency	Token	Frequency	Token
1,698,599	the	124,598	European
849,256	of	104,325	Mr
793,731	to	92,195	Commission
640,257	and	66,781	President
508,560	in	62,867	Parliament
407,638	that	57,804	Union
400,467	is	53,683	report
394,778	a	53,547	Council
263,040	I	45,842	States

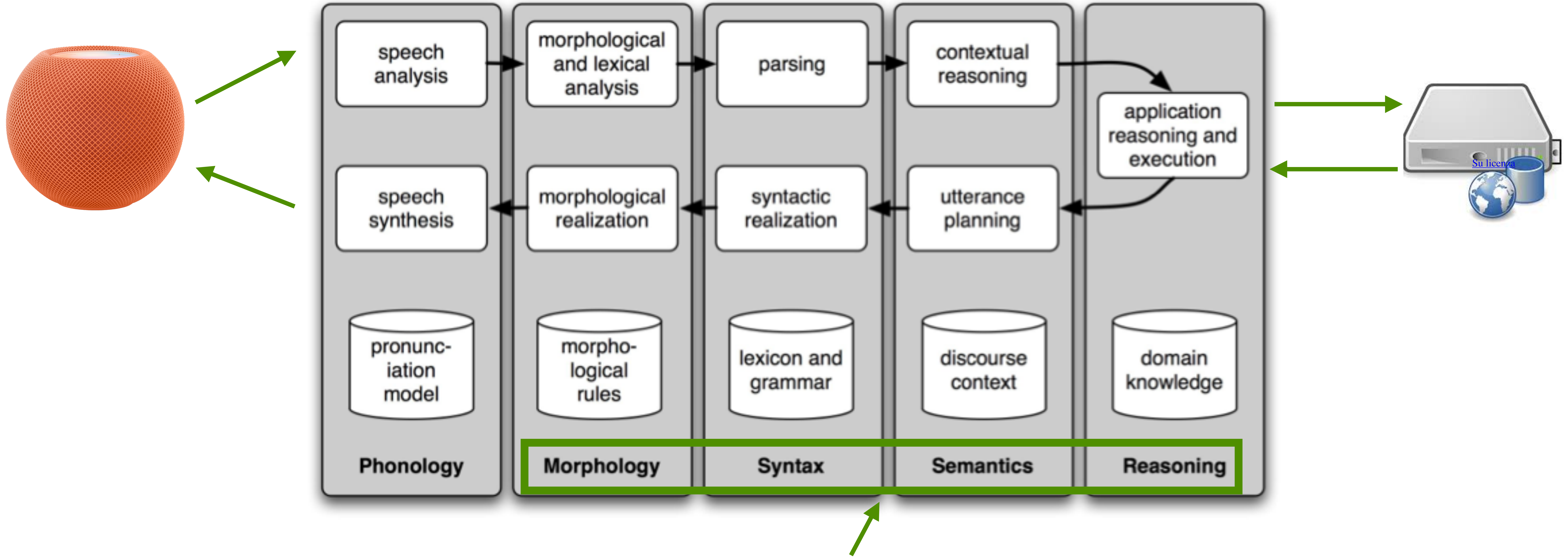
Words ordered by their frequency



Language evolves

LOL	Laugh out loud
G2G	Got to go
BFN	Bye for now
B4N	Bye for now
Idk	I don't know
FWIW	For what it's worth
LUWAMH	Love you with all my heart

An Example of NLP Process - Smart Speakers



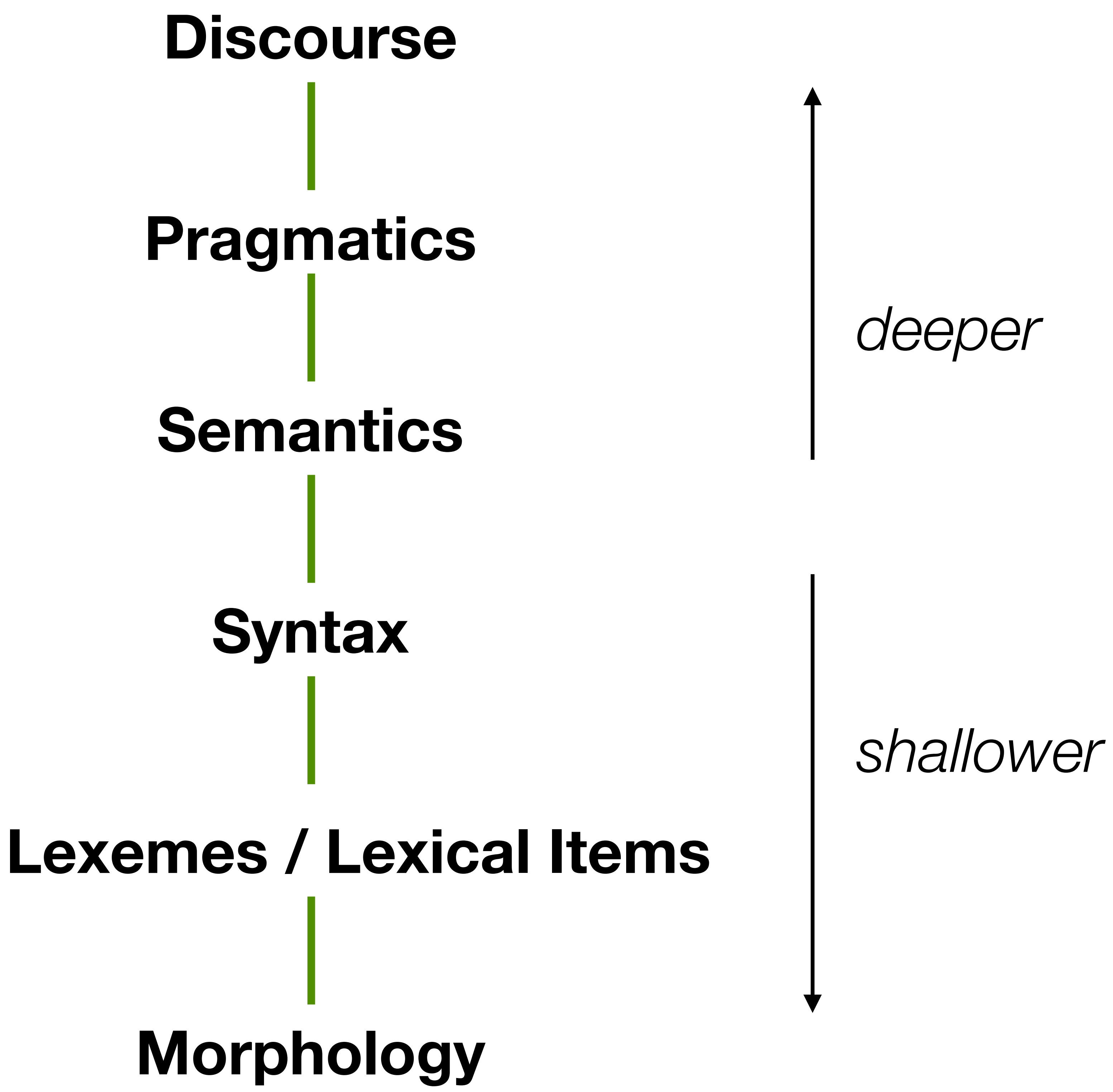
In this module

Language

A recap

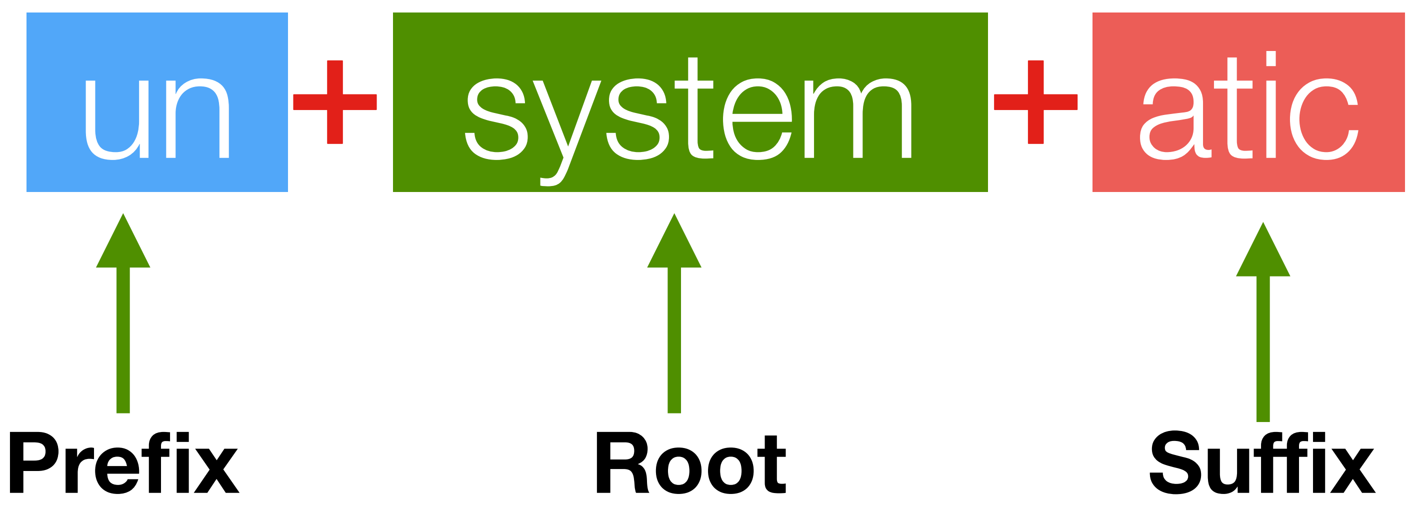
Levels of Linguistic Representation

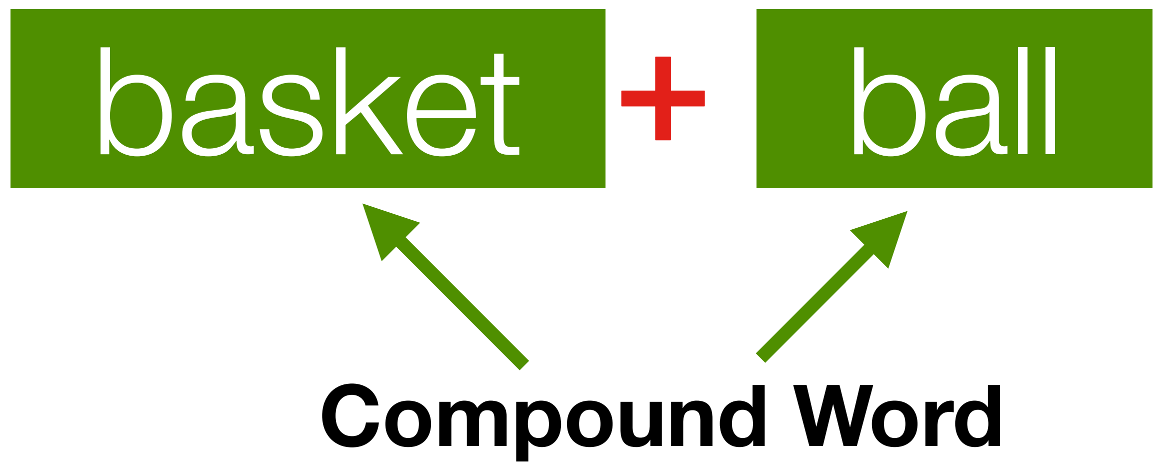
- The mapping between levels is hard
- Appropriateness of representation depends on the application

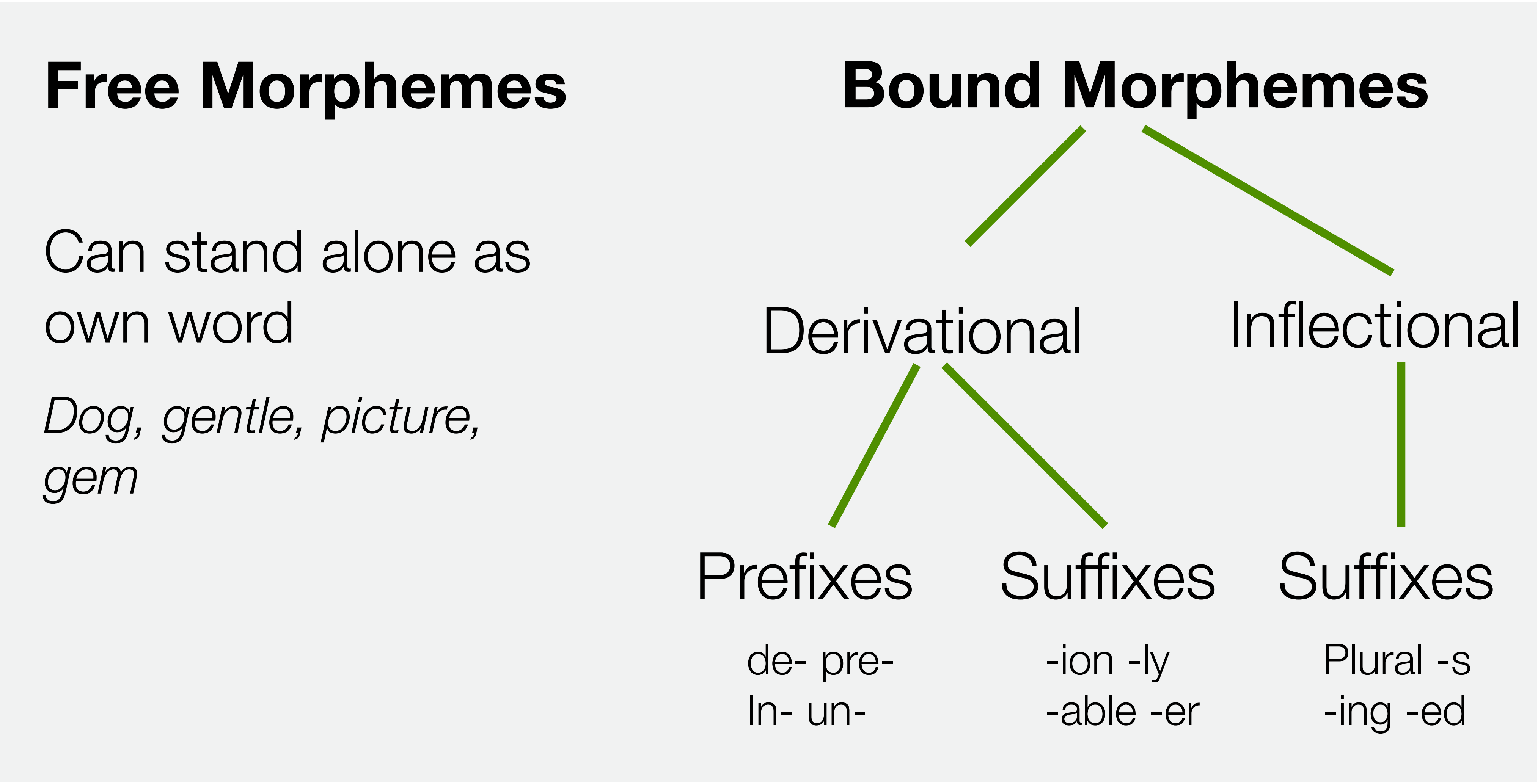


Morphology

- Words are the atomic elements in a language
- Many words have an internal structure that shapes their meaning
- Morphology analysis: split words into meaningful components
  - The structure of words
  - Useful for orthographic error correction





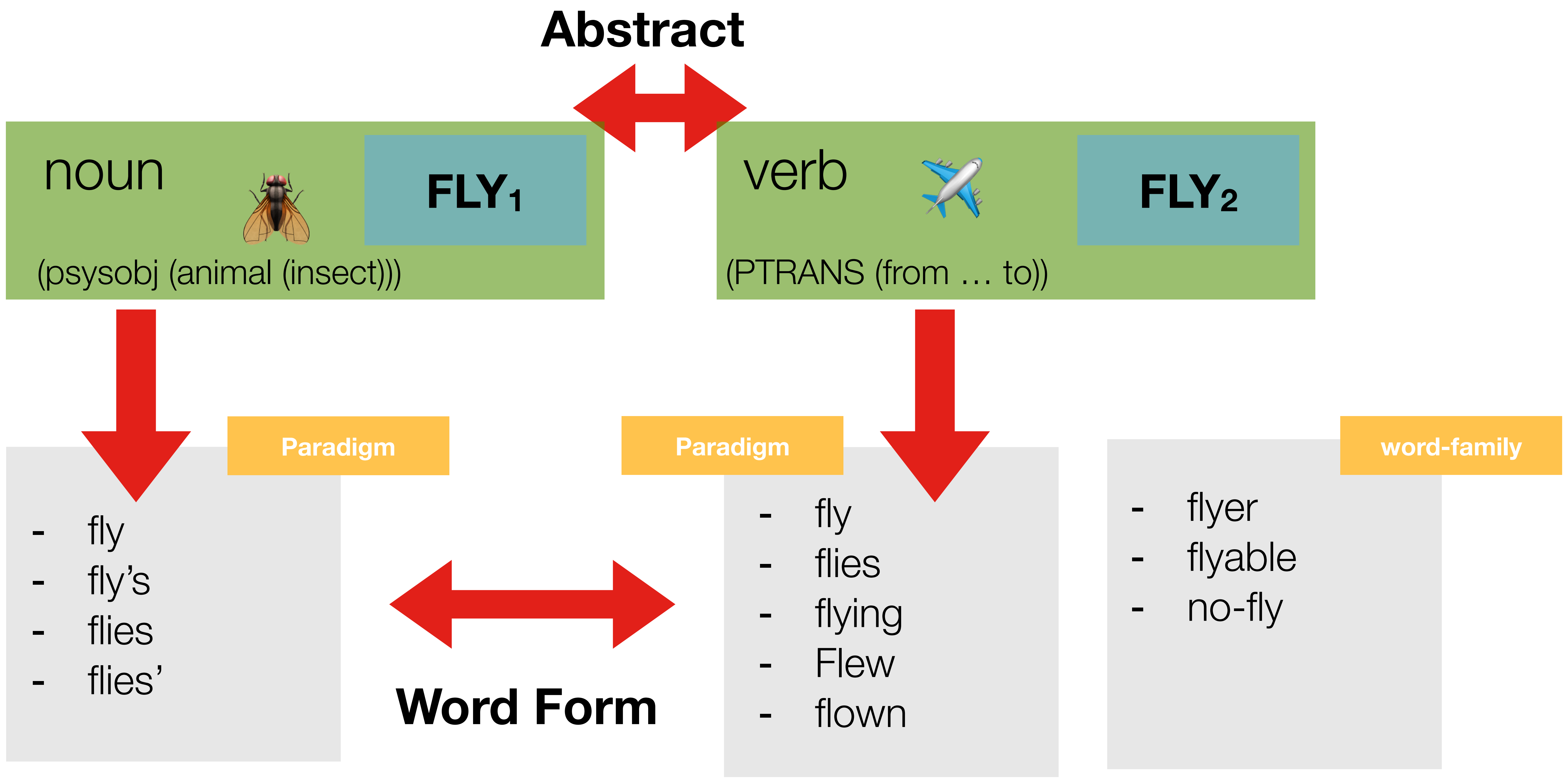


stem	walk	kiss	map	cry
-s form	walks	kisses	maps	cries
-ing participle	walking	kissing	mapping	crying
Past form or -ed participle	walked	kissed	mapped	cried

Lexemes

- A fundamental unit of the lexicon of a language
  - An abstract vocabulary item which may be realised in different sets of grammatical variants

- The same word can have multiple meanings:
  - *bank*, *mean*
  - Extra challenge: domain-specific meanings



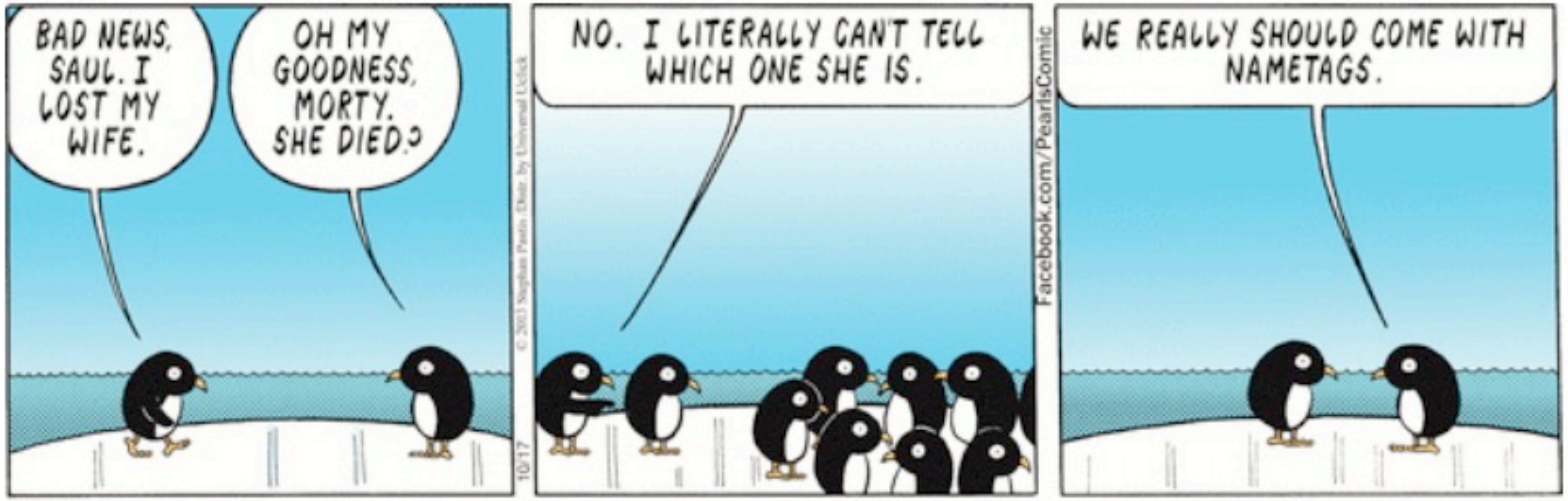
Lexical Items

- A single word, a part of a word, or a chain of words that forms the basic elements of a language's lexicon

- Examples of lexical items
  - **Lexemes** *(previous slide)*
  - **Phrasal verbs**, e.g. *put off, get out*
  - **Multiword expressions**, e.g. *by the way*, *inside out*
  - **Idioms**, e.g. *break a leg*, *a bitter pill to swallow*
  - **Sayings**, e.g. *The early bird gets the worm*, *The devil is in the details*

Lexical Ambiguity

- The presence of two or more possible meanings within a single word
  - Word sense ambiguity



credit: A. Zwicky

Part Of Speech

- The syntactic role of each word in a sentence

	Tag	Description	Example
Open Class	**ADJ**	Adjective: noun modifiers describing properties	*red, young, awesome*
	ADV	Adverb: verb modifiers of time, place, manner	*very, slowly, home, yesterday*
	NOUN	words for persons, places, things, etc.	*algorithm, cat, mango, beauty*
	VERB	words for actions and processes	*draw, provide, go*
	PROPN	Proper noun: name of a person, organization, place, etc..	*Regina, IBM, Colorado*
	INTJ	Interjection: exclamation, greeting, yes/no response, etc.	*oh, um, yes, hello*
Closed Class Words	**ADP**	Adposition (Preposition/Postposition): marks a noun's spacial, temporal, or other relation	*in, on, by, under*
	AUX	Auxiliary: helping verb marking tense, aspect, mood, etc.,	*can, may, should, are*
	CCONJ	Coordinating Conjunction: joins two phrases/clauses	*and, or, but*
	DET	Determiner: marks noun phrase properties	*a, an, the, this*
	NUM	Numeral	*one, two, first, second*
	PART	Particle: a preposition-like form used together with a verb	*up, down, on, off, in, out, at, by*
	PRON	Pronoun: a shorthand for referring to an entity or event	*she, who, I, others*
	SCONJ	Subordinating Conjunction: joins a main clause with a subordinate clause such as a sentential complement	*that, which*
Other	**PUNCT**	Punctuation	; , ()
	SYM	Symbols like $ or emoji	$, %
	X	Other	asdf, qwfg

Always created

Relatively fixed

Part-Of-Speech /2

- **Nouns (NN, NNS)**: words for people, places, or things. Singular or plural
  - *cat, mango*, *algorithm, beauty*, *pacing*
- **Proper Nouns (NNP, NNPS)**: names of **specific persons** or **entities**
  - *Alessandro*, *Delft*, *TU Delft*
- **Adjectives**: describe the properties or qualities of nouns
  - e.g. colour (*white*, *black*), age (*old*, *young*), value (*good*, *bad*)
- **Verbs (VB)**: actions and processes
  - Multiple inflexions for singular/plural and verb tense
- **Personal and Possessive Pronouns (PRP)**: shorthand for referring to an entity or event
  - you, she, I, it, me, my, your, his, her, its, one's, our, their
- **Wh-pronouns**: used in questions
  - *what*, *who*, *whom*, *whoever*

Tag	Description	Example	Tag	Description	Example	Tag	Description	Example
CC	coord. conj.	*and, but, or*	NNP	proper noun, sing.	*IBM*	TO	"to"	*to*
CD	cardinal number	*one, two*	NNPS	proper noun, plu.	*Carolinas*	UH	interjection	*ah, oops*
DT	determiner	*a, the*	NNS	noun, plural	*llamas*	VB	verb base	*eat*
EX	existential 'there'	*there*	PDT	predeterminer	*all, both*	VBD	verb past tense	*ate*
FW	foreign word	*mea culpa*	POS	possessive ending	*'s*	VBG	verb gerund	*eating*
IN	preposition/ subordin-conj	*of, in, by*	PRP	personal pronoun	*I, you, he*	VBN	verb past partici- ple	*eaten*
JJ	adjective	*yellow*	PRP$	possess. pronoun	*your, one's*	VBP	verb non-3sg-pr	*eat*
JJR	comparative adj	*bigger*	RB	adverb	*quickly*	VBZ	verb 3sg pres	*eats*
JJS	superlative adj	*wildest*	RBR	comparative adv	*faster*	WDT	wh-determ.	*which, that*
LS	list item marker	*1, 2, One*	RBS	superlatv. adv	*fastest*	WP	wh-pronoun	*what, who*
MD	modal	*can, should*	RP	particle	*up, off*	WP$	wh-possess.	*whose*
NN	sing or mass noun	*llama*	SYM	symbol	*+,%, &*	WRB	wh-adverb	*how, where*

Syntax

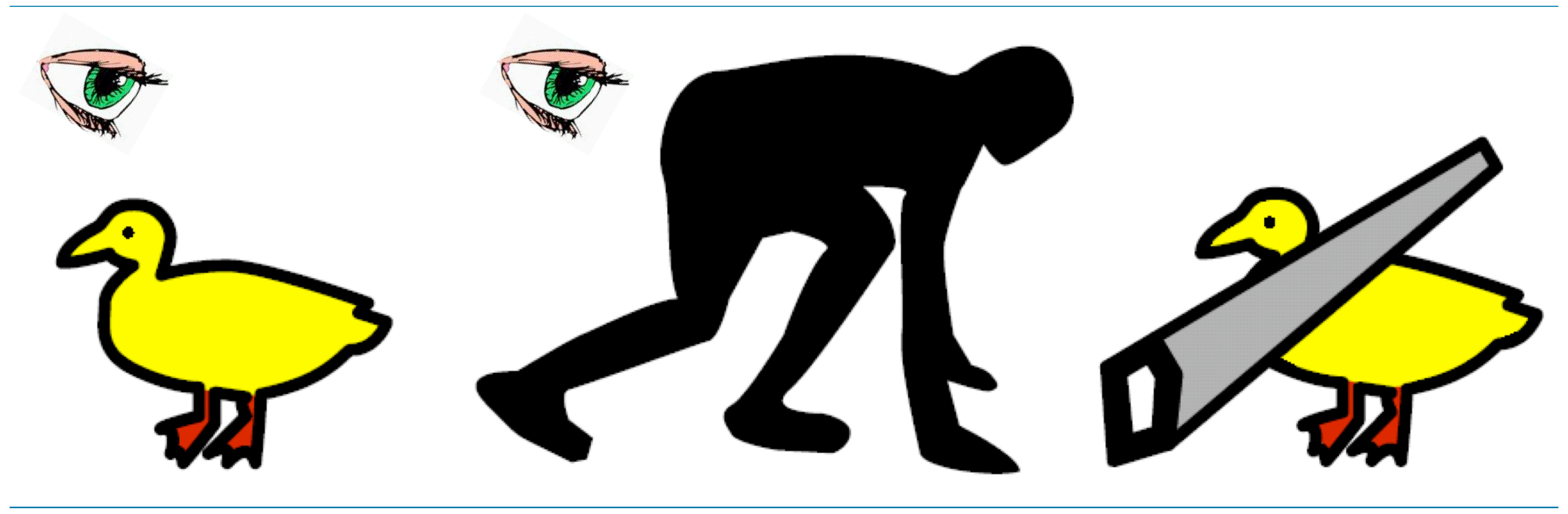
- The syntax of a language is the set of principles (**rules**) under which sequences of words are judged to be grammatically acceptable by fluent speakers
- Basic syntactical elements (there are more)
  - **Constituents**: atomic tokens made up of a group of words
    - *Noun Phrase* (NP)
      - groups made up of nouns, determiners, adjectives, conjunctions
      - e.g *the big house*, *a red and large carpet*
    - *Verb Phrase* (VP)
      - A verb eventually followed by an NP or a prepositional phrase (PP)
      - e.g. *eat* (verb), *eat a pizza* (verb + NP), *eat a pizza with the fork* (verb + NP + PP)

  - **Grammatical Relations**: formalization of the sentence structure as a link between SUBJECTS and OBJECTS
    - es.[he]/*SUBJECT* took [thebighammer]/*OBJECT*

Syntactic Ambiguity

- The presence of two or more possible meanings within a single sentence or sequence of words
- They can be solved only at the semantic (or higher) level
  - Using statistical or semantic knowledge

I saw her duck

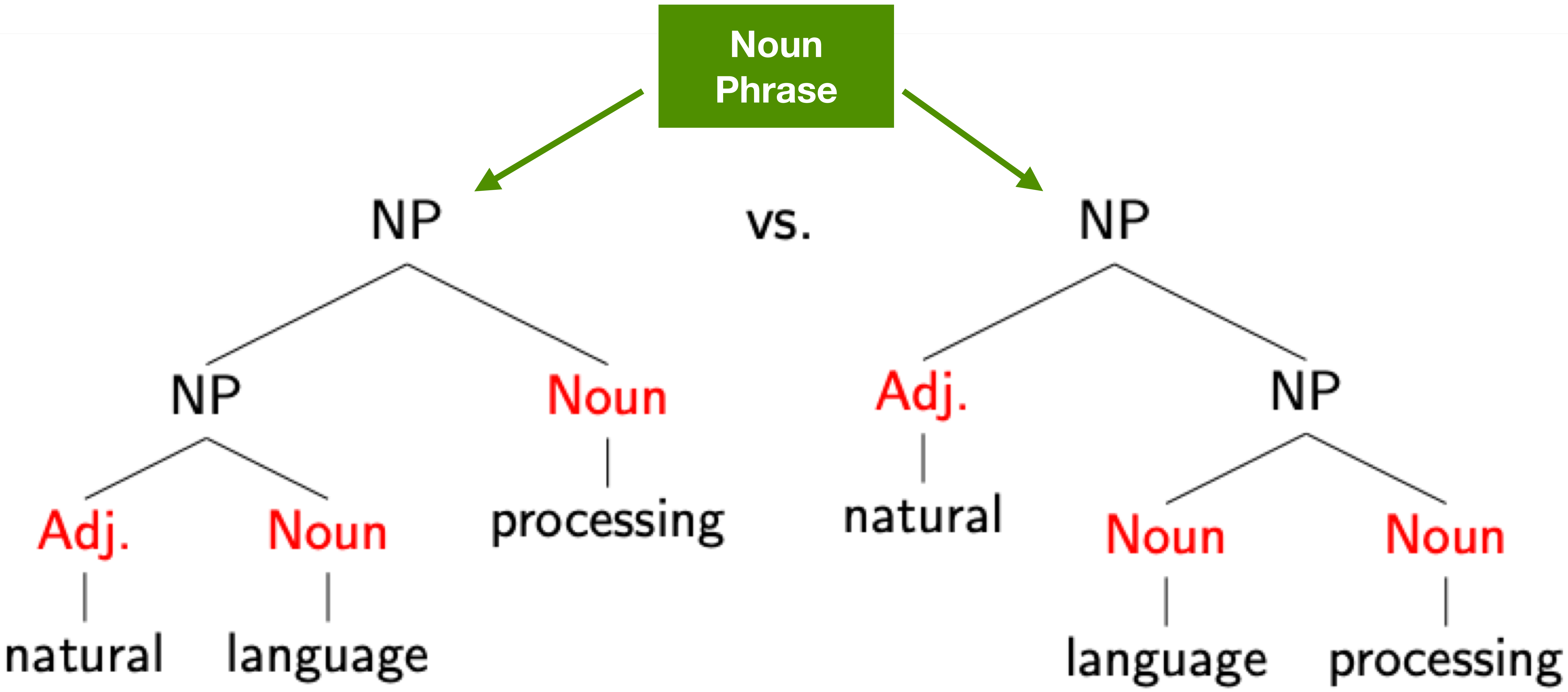


I saw the Grand Canyon flying to New York

Clearly the gran canyon does not fly….

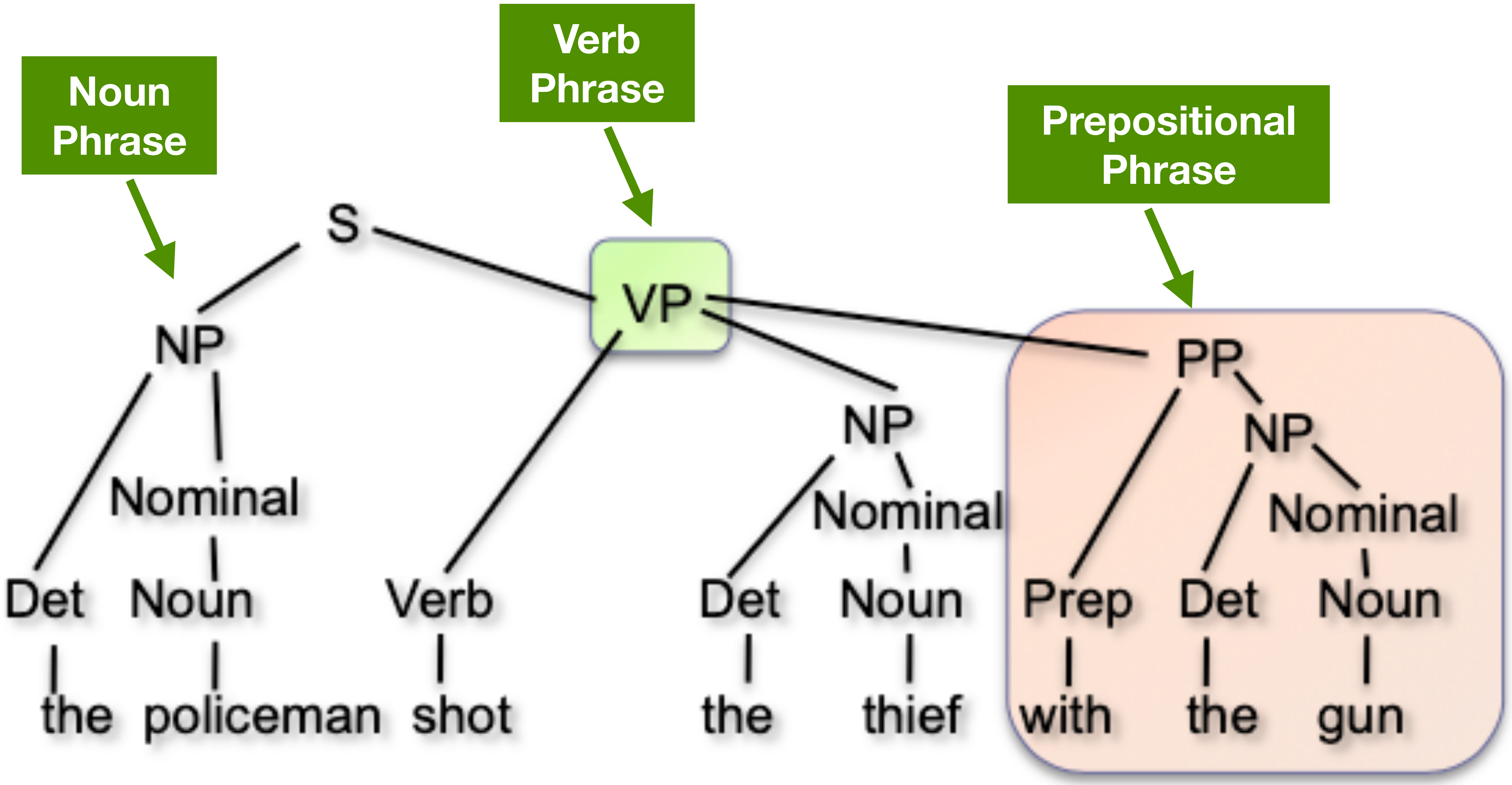
Syntactic Ambiguity

- Different structures lead to different interpretations

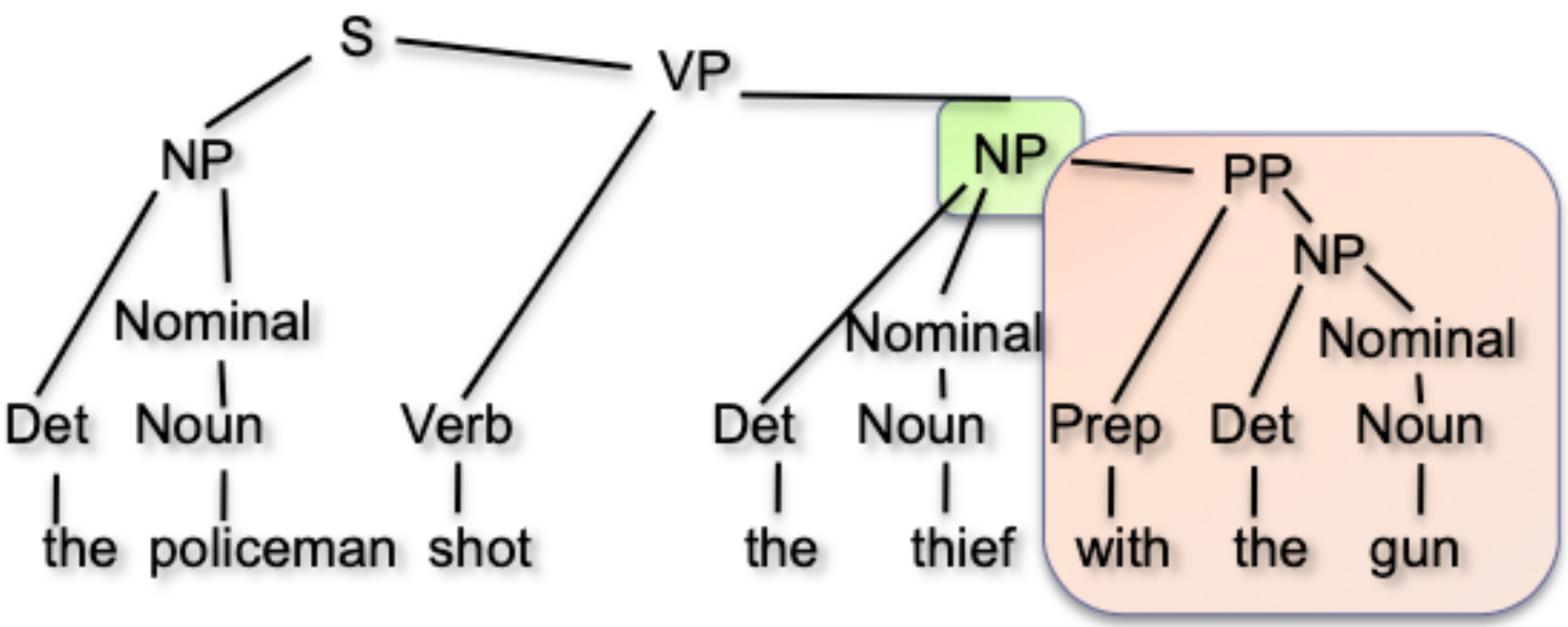


Attachment Ambiguity

The policeman shot the thief with the gun



The policeman used the gun to shoot the thief



The policeman shot a thief that had a gun

Pronoun reference ambiguity

Dr. Macklin often brings his dog Champion to visit with the patients. He just loves to give big, wet, sloppy kisses!

https://www.printwand.com/blog/8-catastrophic-examples-of-word-choice-mistakes

Semantics

- The study of the meaning of words (lexical semantics), and how these combine to form the meanings of sentences (compositional semantics)

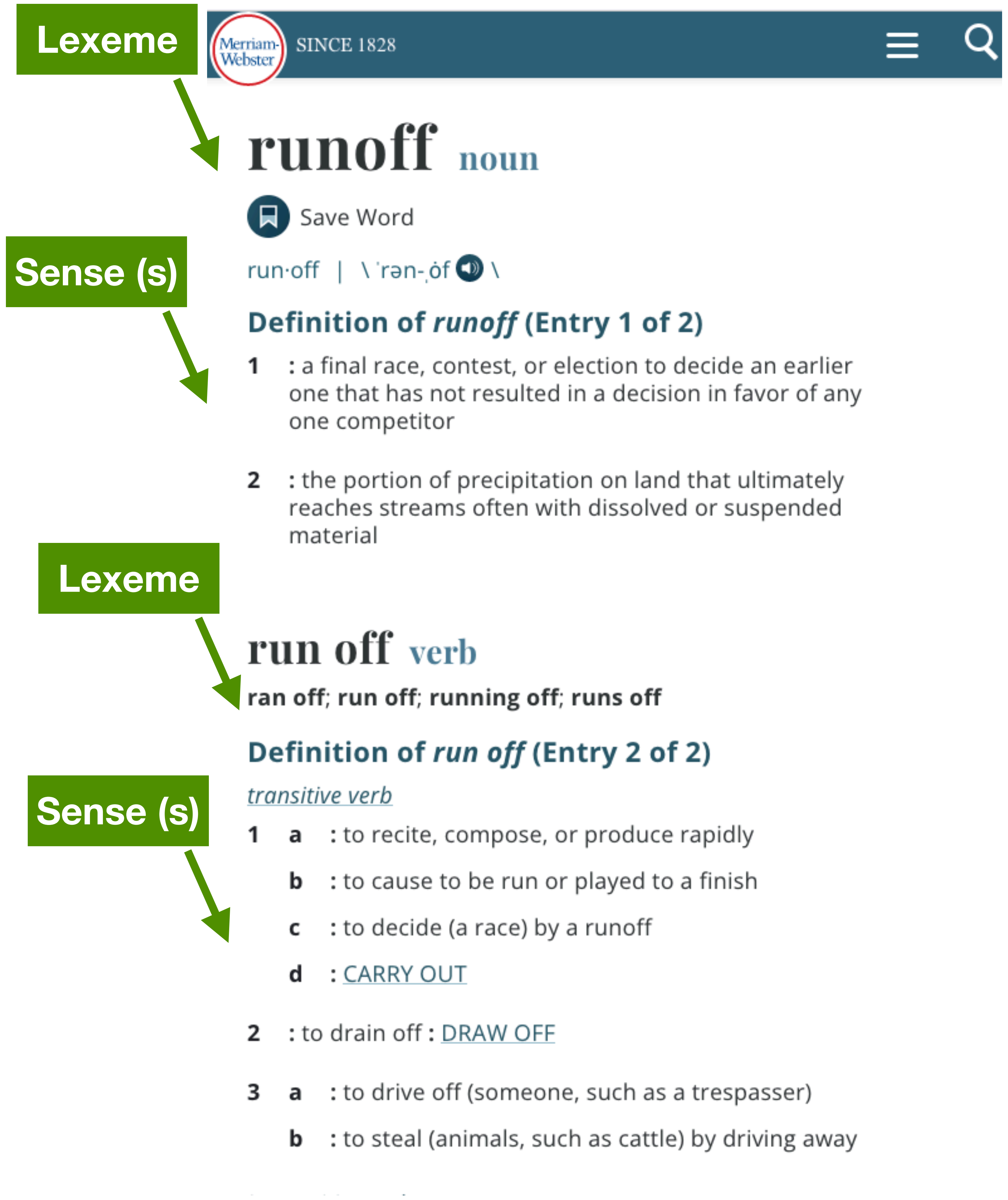
- Mapping of natural language sentences into domain representations
  - E.g., a robot command language, a database query, or an expression in a formal logic

Every fifteen minutes a woman in this country gives birth. Our job is to find this woman, and stop her!

Groucho Marx

Lexical Semantics

- A **lexicon** (the vocabulary of a language) generally has a highly structured form
  - It stores the meanings and uses of each word
  - It encodes the relations between words and meanings
- A **lexeme** is a minimal unit represented in the lexicon. It pairs
  - A **stem**: the orthographic (or phonological) form chosen words (or, sometimes a lexical item)
  - A **sense**: a representation of one aspect of the meaning of a word
- A **dictionary** is a kind of lexicon where meanings are expressed through definitions and examples



Lexical and semantic relations among words (senses)

- **Homonymy**
  - Lexemes that have the **same form** (and the same PoS) but **unrelated meanings**
  - e.g.bank (the financial institution, the river bank)
- **Polysemy**
  - It happens when **a lexeme** has **more related meanings**
  - It depends on the word etymology - unrelated meaning usually have a different origin)
  - e.g. bank (the financial institution), bank (the building hosting the financial institution)
- **Synonymy**
  - **distinct lexemes** with the **same meaning**
  - e.g. fall, autumn; gift, present

- **Hyponymy / Hypernymy** (is-a relation) {parent: hypernym, child: hyponym}
  - A relationship between **two senses** such that one denotes a subclass of the other
  - e.g. dog. animal
  - The relationship is not symmetric
- **Holonomy** / **Meronymy** (part-whole relation)
  - A relationship between **two senses** such that one Is structurally or logically part of the other
  - E.g. arm —> body (holonomy), bicycle —> wheel (meronymy)
  - The relationship is not symmetric
- **Antonymy**
  - A relationship between two senses exists between words that have opposite meaning
  - e.g. tall, short

https://wordnet.princeton.edu/documentation/wnstats7wn

Wordnet

- A hierarchical database of lexical relations
  - More than 200 languages
- Three Separate sub-databases
  - Nouns
  - Verbs
  - Adjectives and Adverbs
- Each lexeme is associated with a set of senses (synset)
- Synsets are linked by **conceptual**, **semantic** and **lexical** relationships
- Available online or for download
  - http://wordnetweb.princeton.edu/perl/webwn

POS	Unique Strings	Synsets	Total Word-Sense Pairs
Noun	117798	82115	146312
Verb	11529	13767	25047
Adjective	21479	18156	30002
Adverb	4481	3621	5580
Totals	155287	117659	206941

Relation	Also Called	Definition	Example
Hypernym	Superordinate	From concepts to superordinates	*breakfast*[1] → *meal*[1]
Hyponym	Subordinate	From concepts to subtypes	*meal*[1] → *lunch*[1]
Instance Hypernym	Instance	From instances to their concepts	*Austen*[1] → *author*[1]
Instance Hyponym	Has-Instance	From concepts to concept instances	*composer*[1] → *Bach*[1]
Member Meronym	Has-Member	From groups to their members	*faculty*[2] → *professor*[1]
Member Holonym	Member-Of	From members to their groups	*copilot*[1] → *crew*[1]
Part Meronym	Has-Part	From wholes to parts	*table*[2] → *leg*[3]
Part Holonym	Part-Of	From parts to wholes	*course*[7] → *meal*[1]
Substance Meronym		From substances to their subparts	*water*[1] → *oxygen*[1]
Substance Holonym		From parts of substances to wholes	*gin*[1] → *martini*[1]
Antonym		Semantic opposition between lemmas	*leader*[1] ⟺ *follower*[1]
Derivationally Related Form		Lemmas w/same morphological root	*destruction*[1] ⟺ *destroy*[1]

Noun Relations

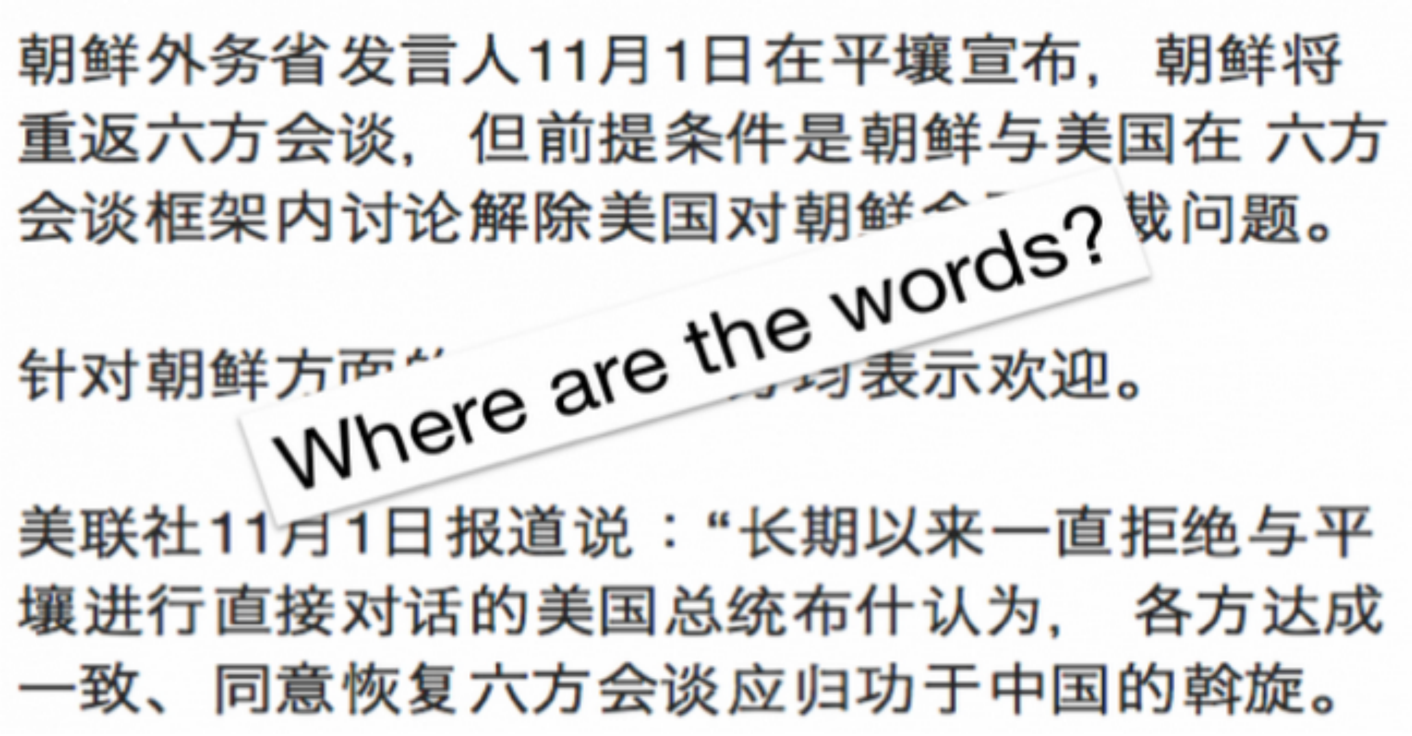
Natural language processing tasks

Morphology /1 - Tokenisation

- Separation of words (or of morphemes) in a sentence

„Latest figures from the US government show the trade deficit with China reached an all time high of $ 365.7 bn (£ 250.1 bn) last year . By February this year it had already reached $ 57 bn .“

- Issues
  - Separators: punctuations
  - Exceptions: „m.p.h“, „Ph.D“
  - Expansions: „we're“ = „we are“
- Multi-words expressions: “New York”, “doghouse”


朝鲜外务省发言人11月1日在平壤宣布，朝鲜将重返六方会谈，但前提条件是朝鲜与美国在 六方会谈框架内讨论解除美国对朝鲜 … 裁问题。

针对朝鲜方面 … 均表示欢迎。

美联社11月1日报道说："长期以来一直拒绝与平壤进行直接对话的美国总统布什认为，各方达成一致、同意恢复六方会谈应归功于中国的斡旋。

Morphology /2

- **Normalisation**
  - Sometimes we need to “normalize” terms
  - We want to match U.S.A. and USA

- **Stopword removal**
  - Removal of high-frequency words, which carry less information
  - E.g. determiners, prepositions
  - English stop list is about 200-300 terms (e.g., “*been*”, “*a*”, “*about*”, “*otherwise*”, “*the*”, etc..)

any word		nouns	
Frequency	Token	Frequency	Token
1,698,599	the	124,598	European
849,256	of	104,325	Mr
793,731	to	92,195	Commission
640,257	and	66,781	President
508,560	in	62,867	Parliament
407,638	that	57,804	Union
400,467	is	53,683	report
394,778	a	53,547	Council
263,040	I	45,842	States

Morphology /3

- **Stemming**
  - Heuristic process that *chops* off the ends of words in the hope of achieving the goal correctly most of the time
  - Stemming collapses derivationally related words
  - Two basic types:
    - Algorithmic: uses programs to determine related words
    - Dictionary-based: uses lists of related words

Example of Stemming with Different Algorithms

Sample text: Such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation

Lovins stemmer: such an analys can reve featur that ar not eas vis from th vari in th individu gen and can lead to a pictur of expres that is mor biolog transpar and acces to interpres

Porter stemmer: such an analysi can reveal featur that ar not easili visibl from the variat in the individu gene and can lead to a pictur of express that is more biolog transpar and access to interpret

Paice stemmer: such an analys can rev feat that are not easy vis from the vary in the individ gen and can lead to a pict of express that is mor biolog transp and access to interpret

Morphology /4

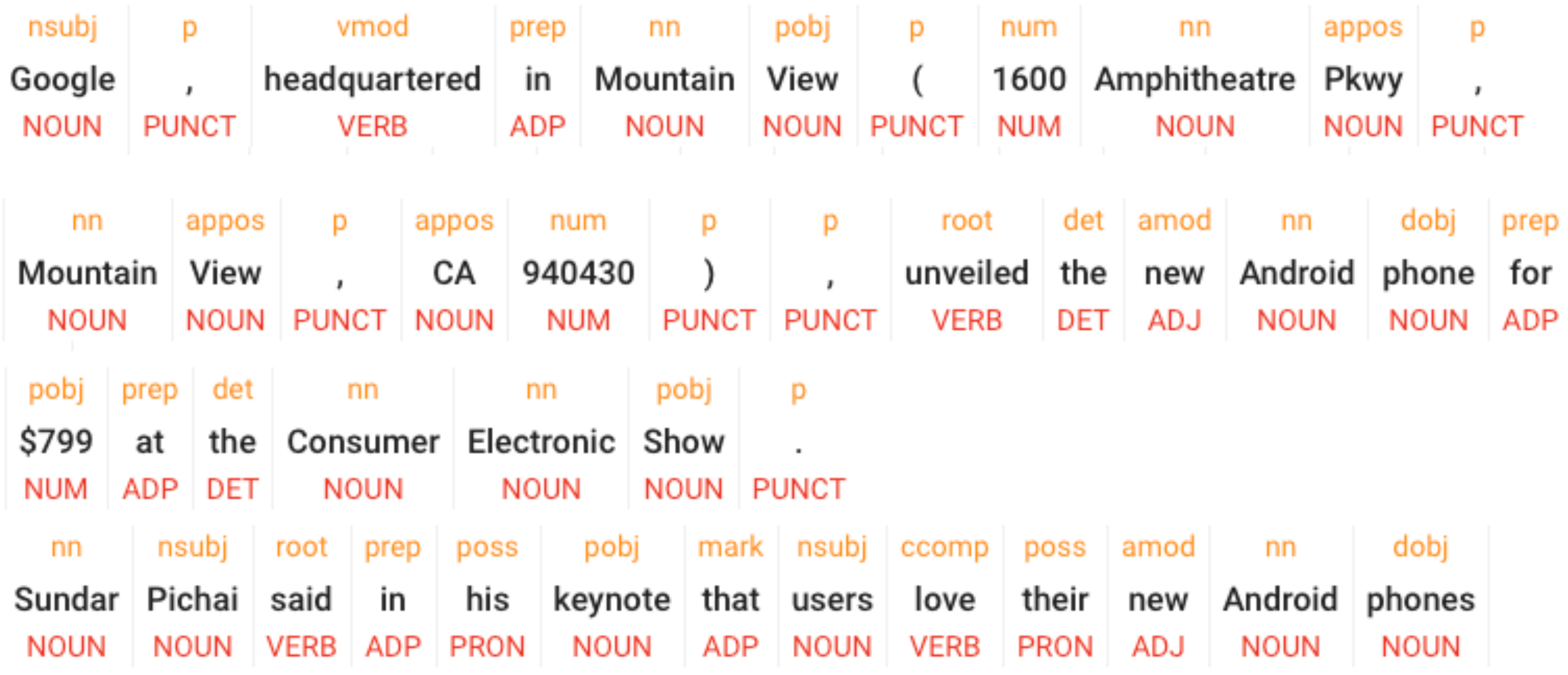
- **Lemmatisation**
  - It uses dictionaries and morphological analysis of words in order to return the base or dictionary form of a word
  - Lemmatization collapses the different inflectional forms of a lemma
  - Example: Lemmatization of “*saw*” —> attempts to return “*see*” or “*saw*” depending on whether the use of the token is a verb or a noun

Google, headquartered in Mountain View (1600 Amphitheatre Pkwy, Mountain View, CA 940430), unveiled the new Android phone for $799 at the Consumer Electronic Show. Sundar Pichai said in his keynote that users love their new Android phones.

Syntax: Part-Of-Speech Tagging

- **Why do we care?**
  - Text-to-speech: *record*[v] and *record*[n]
  - Lemmatization:
    - *saw*[v] → *see*
    - *saw*[n] → *saw*
  - As input for many other NLP tasks
    - Chunking
    - Named entity recognition
    - Information extraction

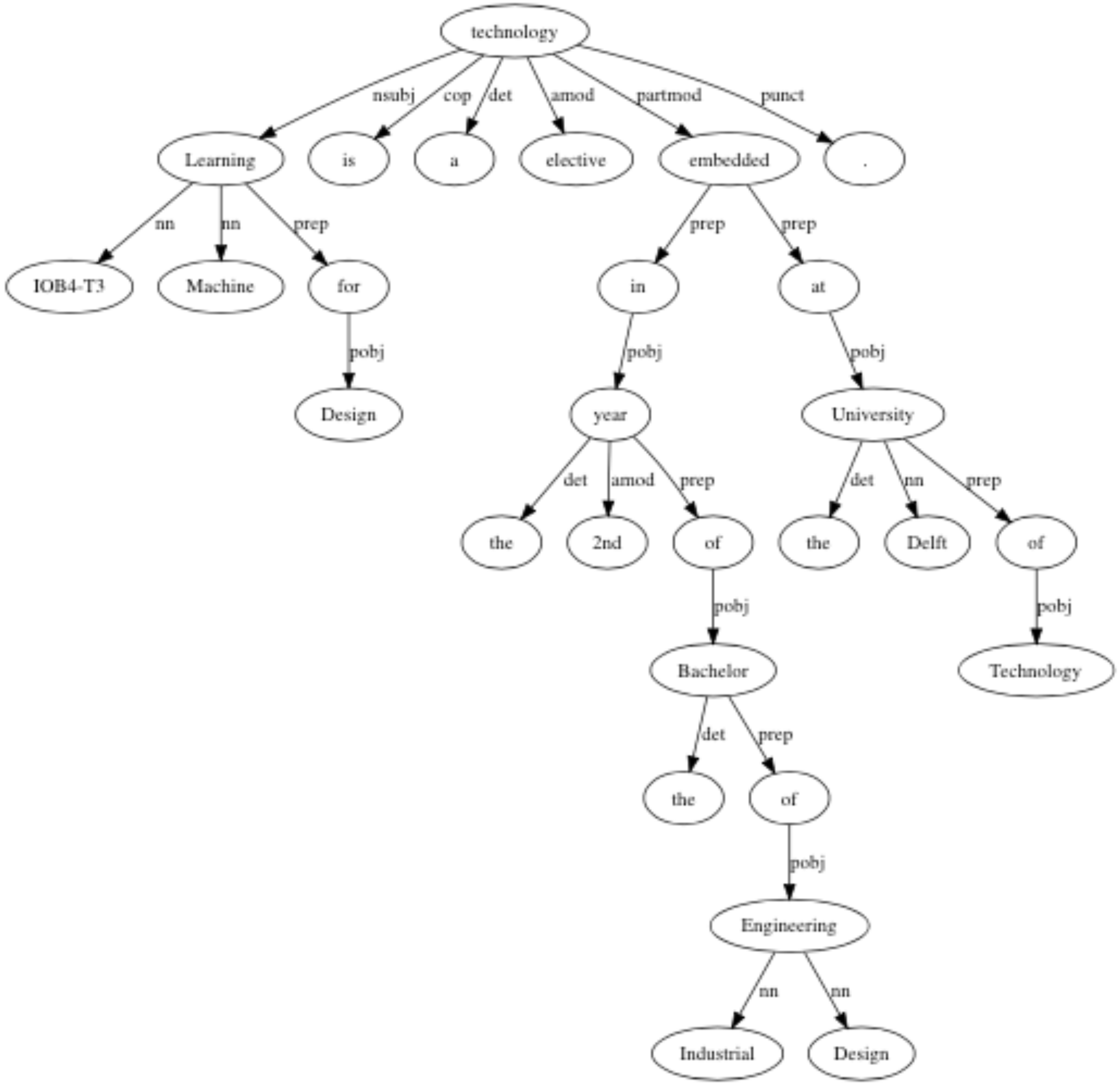
Google, headquartered in Mountain View (1600 Amphitheatre Pkwy, Mountain View, CA 940430), unveiled the new Android phone for $799 at the Consumer Electronic Show. Sundar Pichai said in his keynote that users love their new Android phones.



https://cloud.google.com/natural-language#section-2

Syntax: Dependency Parsing

IOB4-T3 Machine Learning for Design is a technology elective embedded in the 2nd year of the Bachelor of Industrial Design Engineering at the Delft University of Technology.



https://www.textrazor.com/demo

Syntax: Part-Of-Speech Tagging /2

Helicopters will patrol the temporary no-fly zone around New Jersey's MetLife Stadium Sunday, with F-16s based in Atlantic City ready to be scrambled if an unauthorized aircraft does enter the restricted airspace. Down below, bomb-sniffing dogs will patrol the trains and buses that are expected to take approximately 30,000 of the 80,000-plus spectators to Sunday's Super Bowl between the Denver Broncos and Seattle Seahawks.

NNPS/ Helicopters MD/ will NN/ patrol DT/ the JJ/ temporary JJ/ no-fly NN/ zone IN/ around NNP/ New NNP/ Jersey POS/ 's NNP/ MetLife NNP/ Stadium NNP/ Sunday ,/ , IN/ with NNP/ F-16s VBN/ based IN/ in NNP/ Atlantic NNP/ City JJ/ ready TO/ to VB/ be VBN/ scrambled IN/ if DT/ an JJ/ unauthorized NN/ aircraft VBZ/ does VB/ enter DT/ the VBN/ restricted NN/ airspace ./ .

IN/ Down IN/ below ,/ , JJ/ bomb-sniffing NNS/ dogs MD/ will NN/ patrol DT/ the NNS/ trains CC/ and NNS/ buses WDT/ that VBP/ are VBN/ expected TO/ to VB/ take RB/ approximately CD/ 30,000 IN/ of DT/ the JJ/ 80,000-plus NNS/ spectators TO/ to NNP/ Sunday POS/ 's NNP/ Super NNP/ Bowl IN/ between DT/ the NNP/ Denver NNS/ Broncos CC/ and NNP/ Seattle NNP/ Seahawks ./ .

https://cogcomp.seas.upenn.edu/page/demo_view/pos

Syntax: Named Entity Recognition

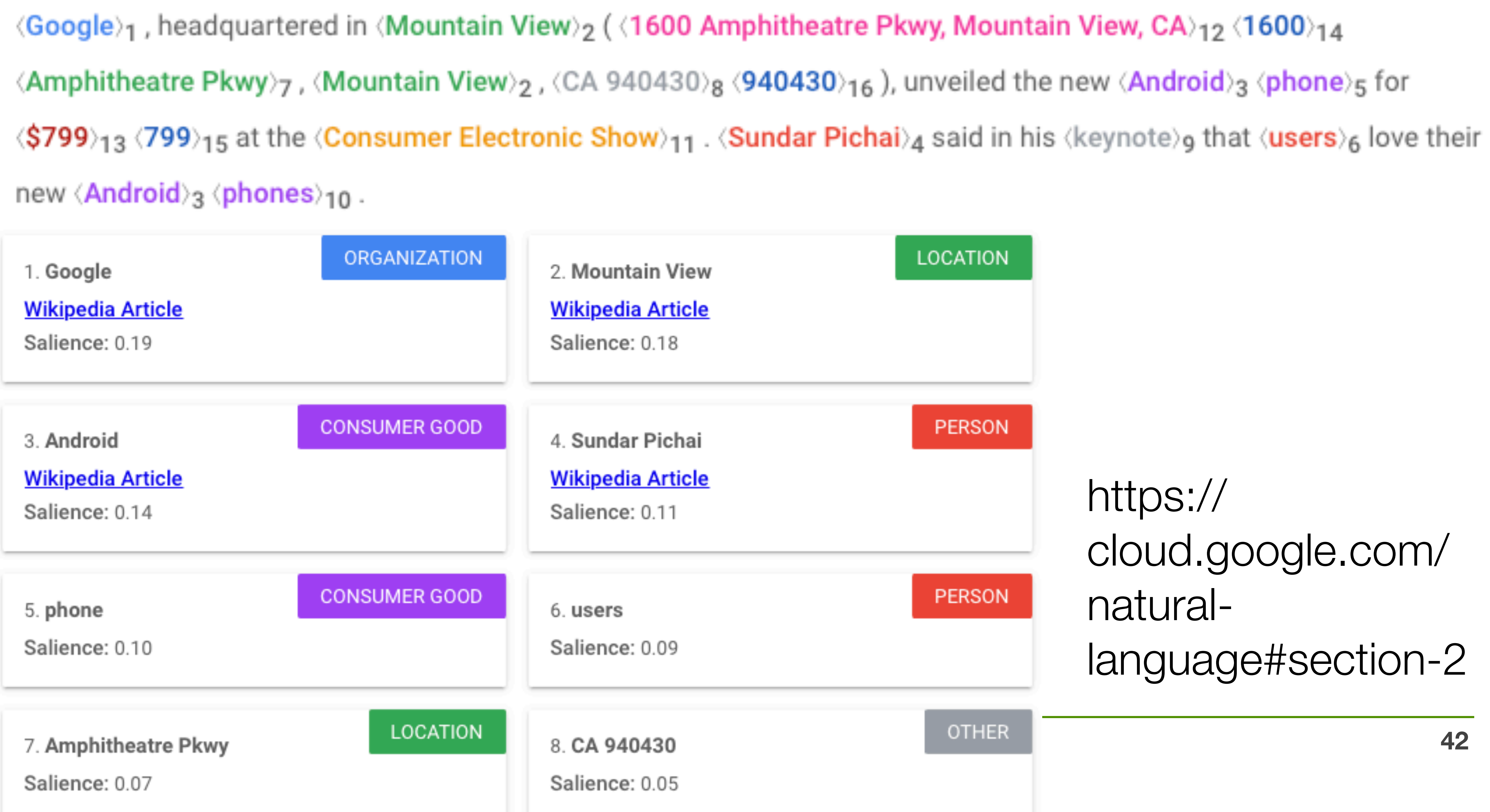
- Factual information and knowledge are normally expressed by named entities
  - Who, Whom, Where, When, Which, …
  - It is the core of the information extraction systems

1. **Identify** words that refer to **proper names** of interest in a particular application
   - E.g. people, companies, locations, dates, product names, prices, etc.

2. **Classify** them to the corresponding classes (e.g. person, location)

3. **Assign** a unique identifier from a database

Google, headquartered in Mountain View (1600 Amphitheatre Pkwy, Mountain View, CA 940430), unveiled the new Android phone for $799 at the Consumer Electronic Show. Sundar Pichai said in his keynote that users love their new Android phones.



https://cloud.google.com/natural-language#section-2

Document Categorisation / Topic Modeling

- Categorisation
  - assigning a label or category to an entire text or document
  - Supervised learning
  - For instance
    - Spam vs. Not spam
    - Language identification
    - Authors attribution
    - Assigning a library subject category or topic label
- Topic Modeling
  - A topic is the subject or theme of a discourse
  - Topic modeling: group documents/text according to their (semantic) similarity
  - An unsupervised machine learning approach

Welcome to the 2021/2022 Edition of the Machine Learning for Design Course

The Course

IOB4-T3 Machine Learning for Design is a technology elective embedded in the 2nd year of the Bachelor of Industrial Design Engineering at the Delft University of Technology.

The course provides students with the knowledge required to understand, design, and evaluate machine learning systems in the context of the design of intelligent products, services, and systems (iPSSs).

Machine learning (ML) is a computational approach that aims at "giving computers the ability to learn without being explicitly programmed" (A. Samuel, 1959). Smart thermostats, voice-based personal assistants, autonomous vehicles, traffic control systems, online social networks, web shopping platforms, content creation platforms, personal health appliances: much of current and future iPSSs are powered by ML technology, influencing, and shaping our interests, habits, lives, and society.

To meaningfully envision and design future iPSSs that are beneficial and useful to people and society, designers must:

- engage with the details of how ML systems "see" the world, "reason" about it, and interact with it
- experience the quirks, biases, and failures of ML technology;
- contend with how agency, initiative, trust, and explainability mediate the interaction between human and iPSSs;
- and understand how functionalities enabled by ML can be designed in iPSSs.

Students in this course gain practical experience with ML technology and learn how to think critically of what ML systems can do, and how they could and should be integrated in iPSSs.

CATEGORIES

0.85	science and technology
0.58	education
0.58	economy, business and finance>economic sector>computing and information technology
0.57	society
0.54	science and technology>social sciences>psychology
0.54	economy, business and finance>economic sector>media
0.54	society>values>ethics
0.49	education>school>further education
0.43	economy, business and finance>economic sector>computing and information technology>software
0.43	science and technology>social sciences>philosophy

TOPICS

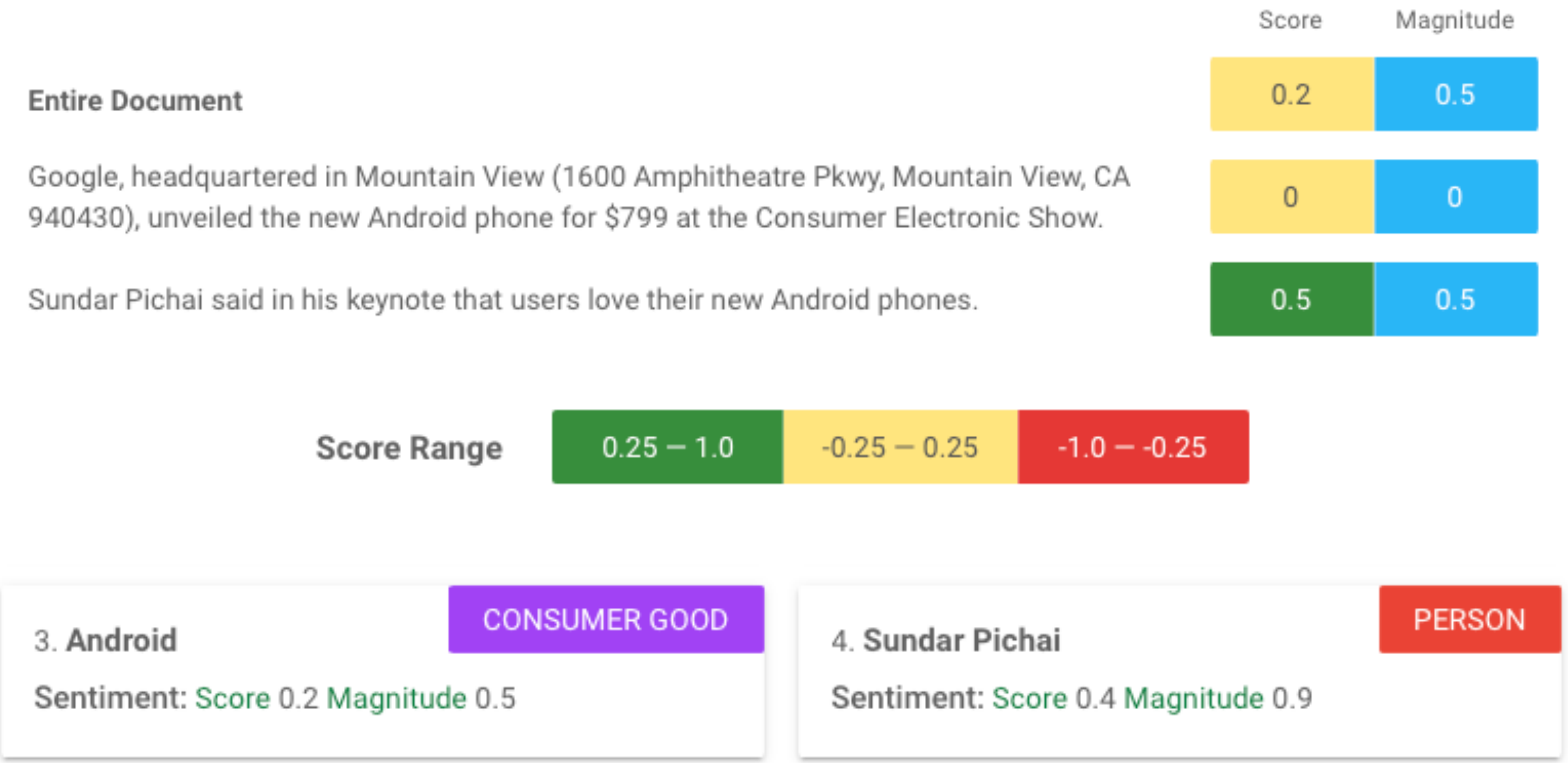
1.00	Technology
1.00	Machine learning
1.00	Design
1.00	Learning
1.00	System
1.00	Social networking service
1.00	Cognition
1.00	Human activities
1.00	Branches of science
1.00	Communication
1.00	Cognitive science
1.00	Education
0.93	Educational psychology
0.93	Self-driving car
0.89	Engineering
0.85	Systems science
0.84	Social network
0.84	Computing
0.83	Behavior modification
0.82	Machine
0.82	Concepts in metaphysics
0.78	Reason
0.77	Neuropsychological assessment
0.77	Change
0.76	Interdisciplinary subfields
0.75	Psychological concepts
0.75	Science
0.75	World Wide Web
0.75	Society
0.74	Academic discipline interactions
0.73	Experience
0.70	Cyberspace
0.70	Content creation
0.69	Applied psychology
0.67	Neuroscience
0.67	Bias

https://www.textrazor.com/demo

Syntax: Sentiment Analysis

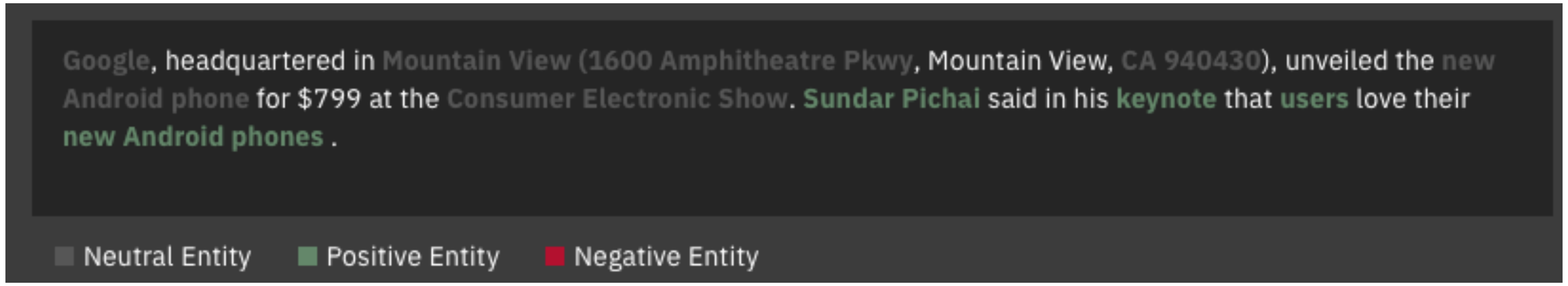
- The detection of attitudes
  - “*enduring, affectively colored beliefs, dispositions towards objects or persons*”
- Main elements
  - Holder (source)
  - Target (aspect)
  - Type of attitude
  - Text containing the attitude
- Tasks
  - **Classification**: Is the attitude of the text positive or negative?
  - **Regression**: Rank the attitude of the text from 1 to 5
  - **Advanced**: Detect the target, source, or complex attitude types

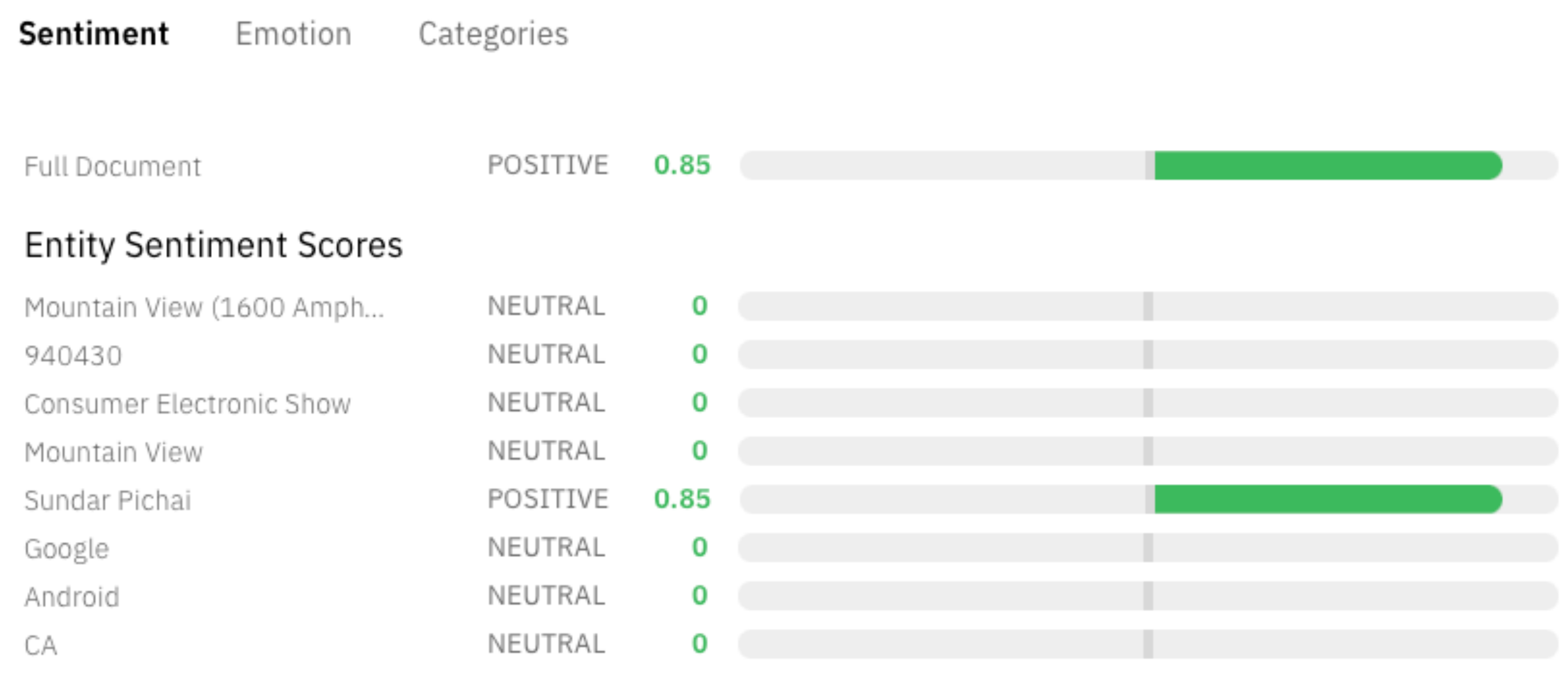
Google, headquartered in Mountain View (1600 Amphitheatre Pkwy, Mountain View, CA 940430), unveiled the new Android phone for $799 at the Consumer Electronic Show. Sundar Pichai said in his keynote that users love their new Android phones.



Syntax: Sentiment Analysis / IBM Demo

Google, headquartered in Mountain View (1600 Amphitheatre Pkwy, Mountain View, CA 940430), unveiled the new Android phone for $799 at the Consumer Electronic Show. Sundar Pichai said in his keynote that users love their new Android phones.



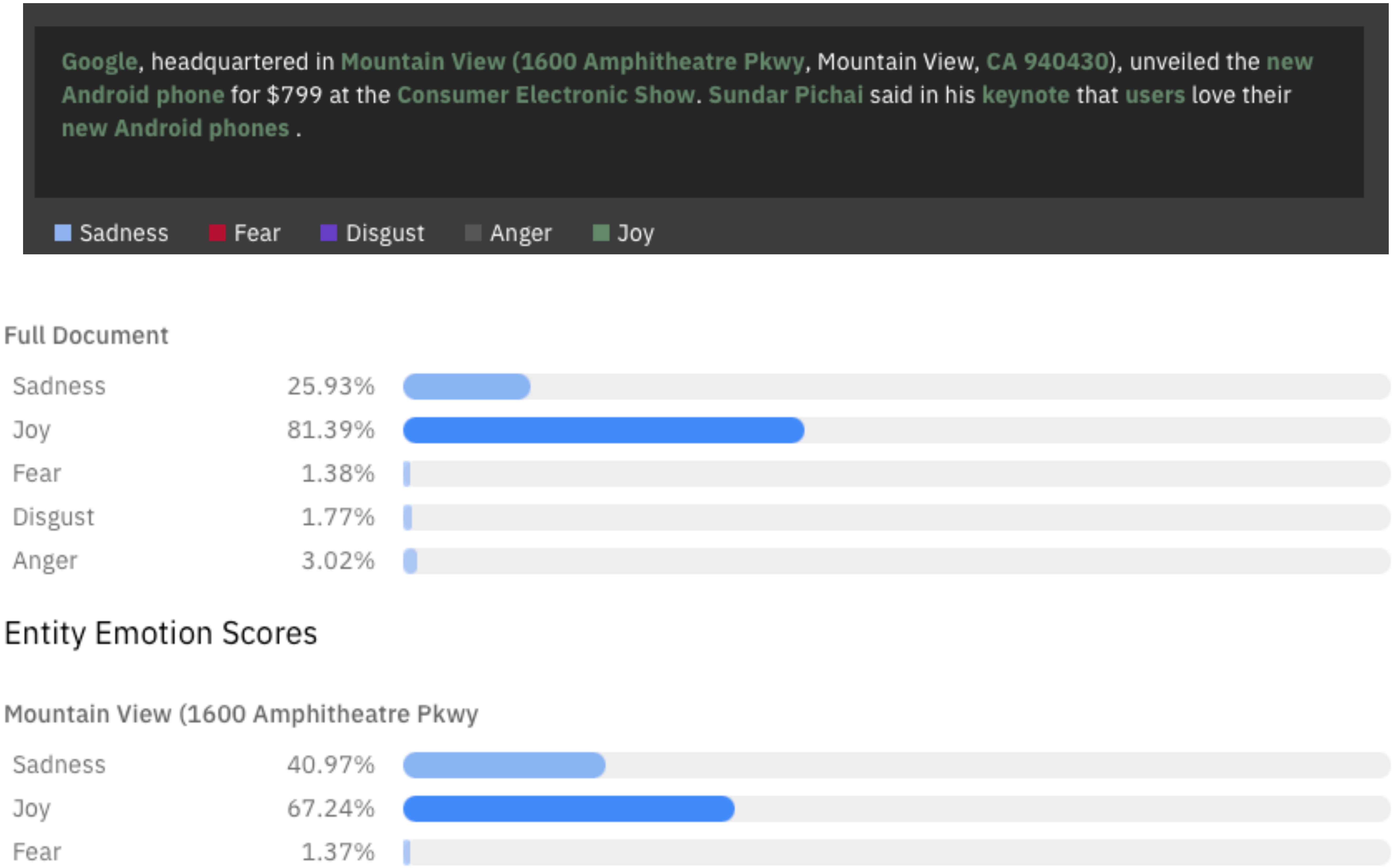


https://www.ibm.com/demos/live/natural-language-understanding/self-service/home

Syntax: Emotion Analysis / IBM Demo

Google, headquartered in Mountain View (1600 Amphitheatre Pkwy, Mountain View, CA 940430), unveiled the new Android phone for $799 at the Consumer Electronic Show. Sundar Pichai said in his keynote that users love their new Android phones.

Detects anger, disgust, fear, joy, or sadness that is conveyed in the content or by the context around target phrases specified in the targets parameter.

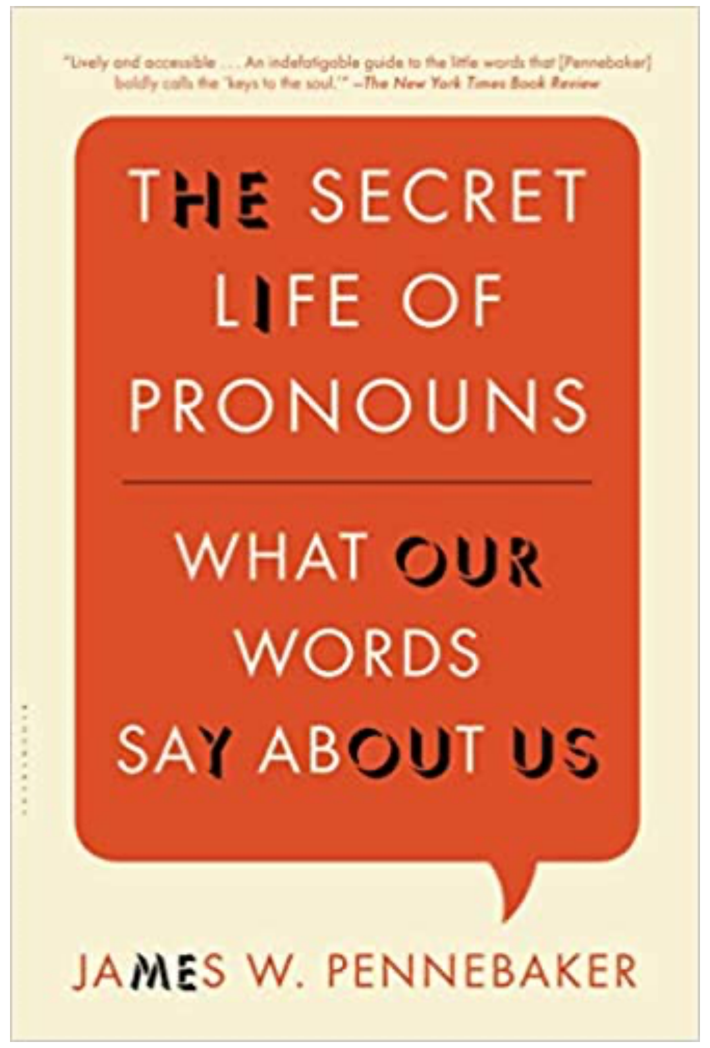


https://www.ibm.com/demos/live/natural-language-understanding/self-service/home

Syntax - Language Analysis

- Idea: people's language can provide insights into their psychological states (emotions, thinking style, etc)
- For instance
  - *Frequency of words associated with positive or negative emotions*
  - *Use of pronouns as a proxy for confidence and character traits*

- **Analytic Thinking**: the degree to which people use words that suggest formal, logical, and hierarchical thinking patterns.
  - low Analytical Thinking —> language that is more intuitive and personal
- **Clout:** the relative social status, confidence, or leadership that people display through their writing or talking
- **Authenticity**: the degree to which a person is self-monitoring
  - Low authenticity: prepared texts (i.e., speeches that were written ahead of time) and texts where a person is being socially cautious.
- **Emotional tone**: the higher the number, the more positive the tone. Numbers below 50 suggest a more negative emotional tone.



Category	Abbrev.	Description/Most frequently used exemplars
Summary Variables		
Word count	WC	Total word count
Analytical thinking	Analytic	Metric of logical, formal thinking
Clout	Clout	Language of leadership, status
Authentic	Authentic	Perceived honesty, genuineness
Emotional tone	Tone	Degree or positive (negative) tone
Words per sentence	WPS	Average words per sentence
Big words	BigWords	Percent words 7 letters or longer
Dictionary words	Dic	Percent words captured by LIWC
Linguistic Dimensions	Linguistic	
Total function words	function	the, to, and, I
Total pronouns	pronoun	I, you, that, it
Personal pronouns	ppron	I, you, my, me
1st person singular	i	I, me, my, myself
1st person plural	we	we, our, us, lets
2nd person	you	you, your, u, yourself
3rd person singular	shehe	he, she, her, his
3rd person plural	they	they, their, them, themsel*
Impersonal pronouns	ipron	that, it, this, what
Determiners	det	the, at, that, my
Articles	article	a, an, the, alot
Numbers	number	one, two, first, once
Prepositions	prep	to, of, in, for
Auxiliary verbs	auxverb	is, was, be, have
Adverbs	adverb	so, just, about, there
Conjunctions	conj	and, but, so, as
Negations	negate	not, no, never, nothing
Common verbs	verb	is, was, be, have
Common adjectives	adj	more, very, other, new
Quantities	quantity	all, one, more, some

Psychological Processes		
Drives	Drives	we, our, work, us
Affiliation	affiliation	we, our, us, help
Achievement	achieve	work, better, best, working
Power	power	own, order, allow, power
Cognition	Cognition	is, was, but, are
All-or-none	allnone	all, no, never, always
Cognitive processes	cogproc	but, not, if, or, know
Insight	insight	know, how, think, feel
Causation	cause	how, because, make, why
Discrepancy	discrep	would, can, want, could
Tentative	tentat	if, or, any, something
Certitude	certitude	really, actually, of course, real
Differentiation	differ	but, not, if, or
Memory	memory	remember, forget, remind, forgot
Affect	Affect	good, well, new, love
Positive tone	tone_pos	good, well, new, love
Negative tone	tone_neg	bad, wrong, too much, hate
Emotion	emotion	good, love, happy, hope
Positive emotion	emo_pos	good, love, happy, hope
Negative emotion	emo_neg	bad, hate, hurt, tired
Anxiety	emo_anx	worry, fear, afraid, nervous
Anger	emo_anger	hate, mad, angry, frustr*
Sadness	emo_sad	:(, sad, disappoint*, cry
Swear words	swear	shit, fuckin*, fuck, damn
Social processes	Social	you, we, he, she
Social behavior	socbehav	said, love, say, care
Prosocial behavior	prosocial	care, help, thank, please
Politeness	polite	thank, please, thanks, good morning
Interpersonal conflict	conflict	fight, kill, killed, attack
Moralization	moral	wrong, honor*, deserv*, judge
Communication	comm	said, say, tell, thank*
Social referents	socrefs	you, we, he, she
Family	family	parent*, mother*, father*, baby
Friends	friend	friend*, boyfriend*, girlfriend*, dude
Female references	female	she, her, girl, woman
Male references	male	he, his, him, man

The MLFD Course Manual

Filename	Segment	WC	Analytic ▲	Clout	Authentic	Tone
MLFD_Course Manual_V1.0.docx	1	3286	85.79	83.0	32.06	36.29

Drives	affiliation	achieve	power	Cognition	allnone	cogproc	insight	cause	discrep	tentat	certitude
3.99	1.37	2.13	0.55	11.93	0.24	11.66	4.41	1.86	1.46	1.89	0.18

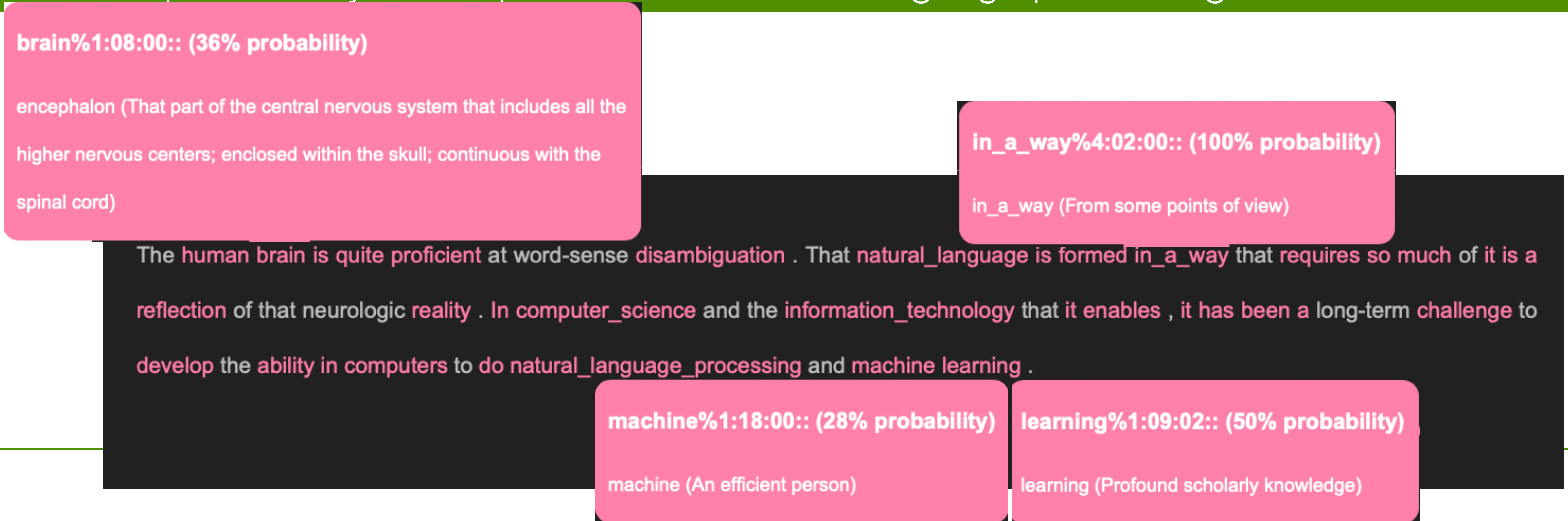
memory	Affect	tone_pos	tone_neg	emotion	emo_pos	emo_neg	emo_anx	emo_anger	emo_sad	swear
0.06	2.28	1.7	0.58	0.52	0.09	0.21	0.0	0.0	0.12	0.0

https://www.liwc.app/static/documents/LIWC-22%20Manual%20-%20Development%20and%20Psychometrics.pdf

Semantics: Word Sense Disambiguation

- Multiple words can be spelt the same way (homonymy)
- The same word can also have different, related senses (polysemy)
- Disambiguation depends on context!

The human brain is quite proficient at word-sense disambiguation. That natural language is formed in a way that requires so much of it is a reflection of that neurologic reality. In computer science and the information technology that it enables, it has been a long-term challenge to develop the ability in computers to do natural language processing and machine learning



https://supwsd.net/supwsd/demo

Question Answering: IBM's Watson

- Won Jeopardy on February 16, 2011

William Wilkinson's
"An account of the principalities of Wallachia and Moldovia"
inspired this author's
most famous novel

Bram Stoker

Automated Summarisation

- Condensing a piece of text to a shorter version while preserving key informational elements and the meaning of content
- A very difficult task!

Text Summarization Result

Original URL/Text

IOB4-T3 Machine Learning for Design is a technology elective embedded in the 2nd year of the Bachelor of Industrial Design Engineering at the Delft University of Technology. The course provides students with the knowledge required to understand, design, and evaluate machine learning systems in the context of the design of intelligent products, services, and systems (iPSSs). Machine learning (ML) is a computational approach that aims at "giving computers the ability to learn without being explicitly programmed" (A. Samuel, 1959). Smart thermostats, voice-based personal assistants, autonomous vehicles, traffic control systems, online social networks, web shopping platforms, content creation platforms, personal health appliances: much of current and future iPSSs are powered by ML technology, influencing, and shaping our interests, habits, lives, and society. To meaningfully envision and design future iPSSs that are beneficial and useful to people and society, designers must: engage with the details of how ML systems "see" the world, "reason" about it, and interact with it experience the quirks, biases, and failures of ML technology; contend with how agency, initiative, trust, and explainability mediate the interaction between human and iPSSs; and understand how functionalities enabled by ML can be designed in iPSSs. Students in this course gain practical experience with ML technology and learn how to think critically of what ML systems can do, and how they could and should be integrated in iPSSs.

Summarized Text

IOB4-T3 Machine Learning for Design is a technology elective embedded in the 2nd year of the Bachelor of Industrial Design Engineering at the Delft University of Technology.

The course provides students with the knowledge required to understand, design, and evaluate machine learning systems in the context of the design of intelligent products, services, and systems (iPSSs).

Students in this course gain practical experience with ML technology and learn how to think critically of what ML systems can do, and how they could and should be integrated in iPSSs.

https://textsummarization.net/text-summarizer

Result

After pressing the "**Summarize**" button above, the result will be displayed in the box below.

The summarized text will be here...

IOB4-T3 Machine Learning for Design is a technology optional embedded in the 2nd year of the Bachelor of Industrial Design Engineering at the Delft University of Technology. Machine learning is a computational approach that focuses on "offering computer systems the capacity to learn without being explicitly configured". Students in this course gain useful experience with ML innovation and learn just how to think seriously of what ML systems can do, and just how they could and should be integrated in iPSSs.

https://brevi.app/single-demo

Stance Detection

EXAMPLE HEADLINE

"Robert Plant Ripped up $800M Led Zeppelin Reunion Contract"

EXAMPLE SNIPPETS FROM BODY TEXTS AND CORRECT CLASSIFICATIONS

"... Led Zeppelin's Robert Plant turned down £500 MILLION to reform supergroup. ..."

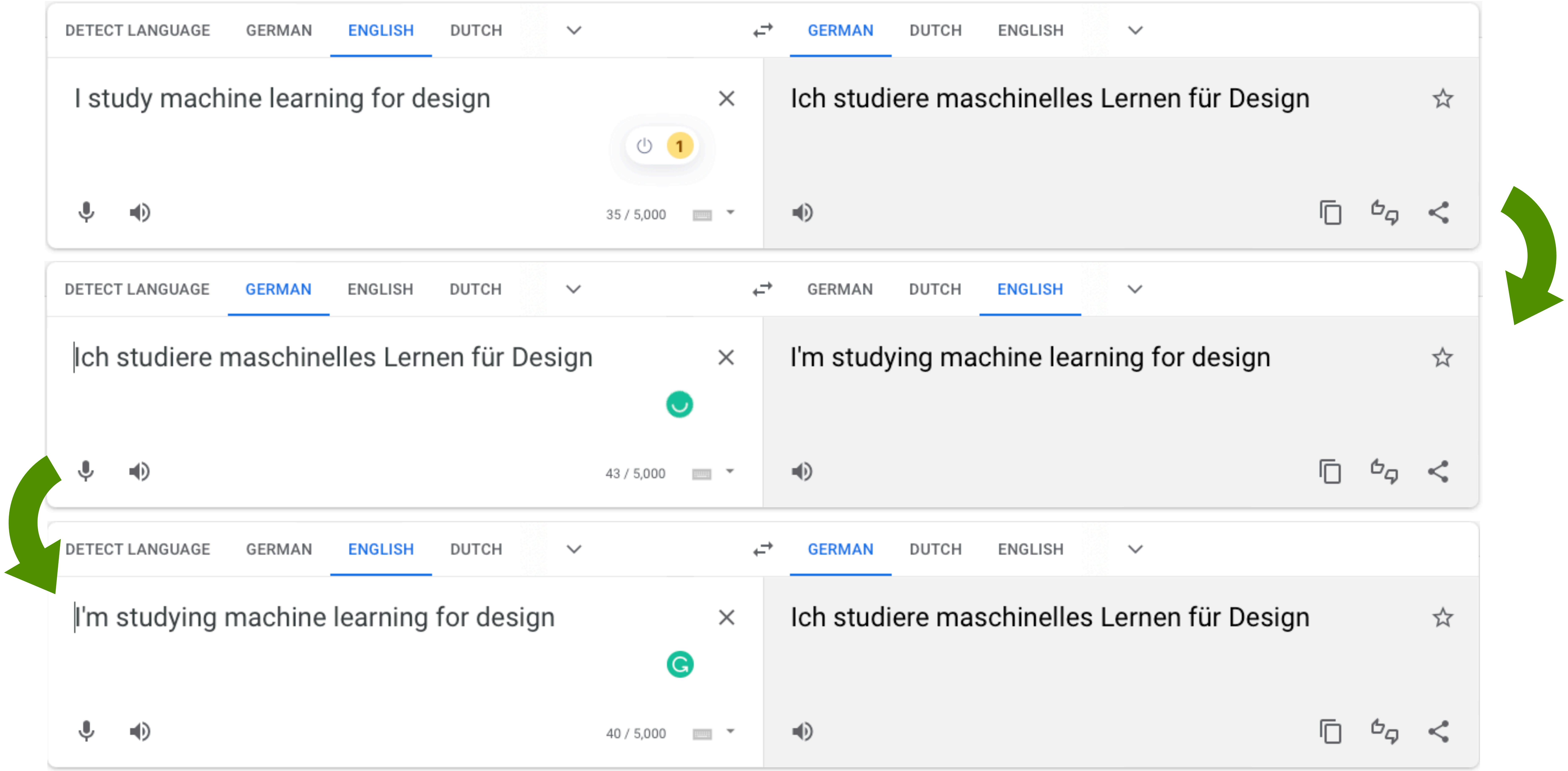
CORRECT CLASSIFICATION: AGREE

"... No, Robert Plant did not rip up an $800 million deal to get Led Zeppelin back together. ..."

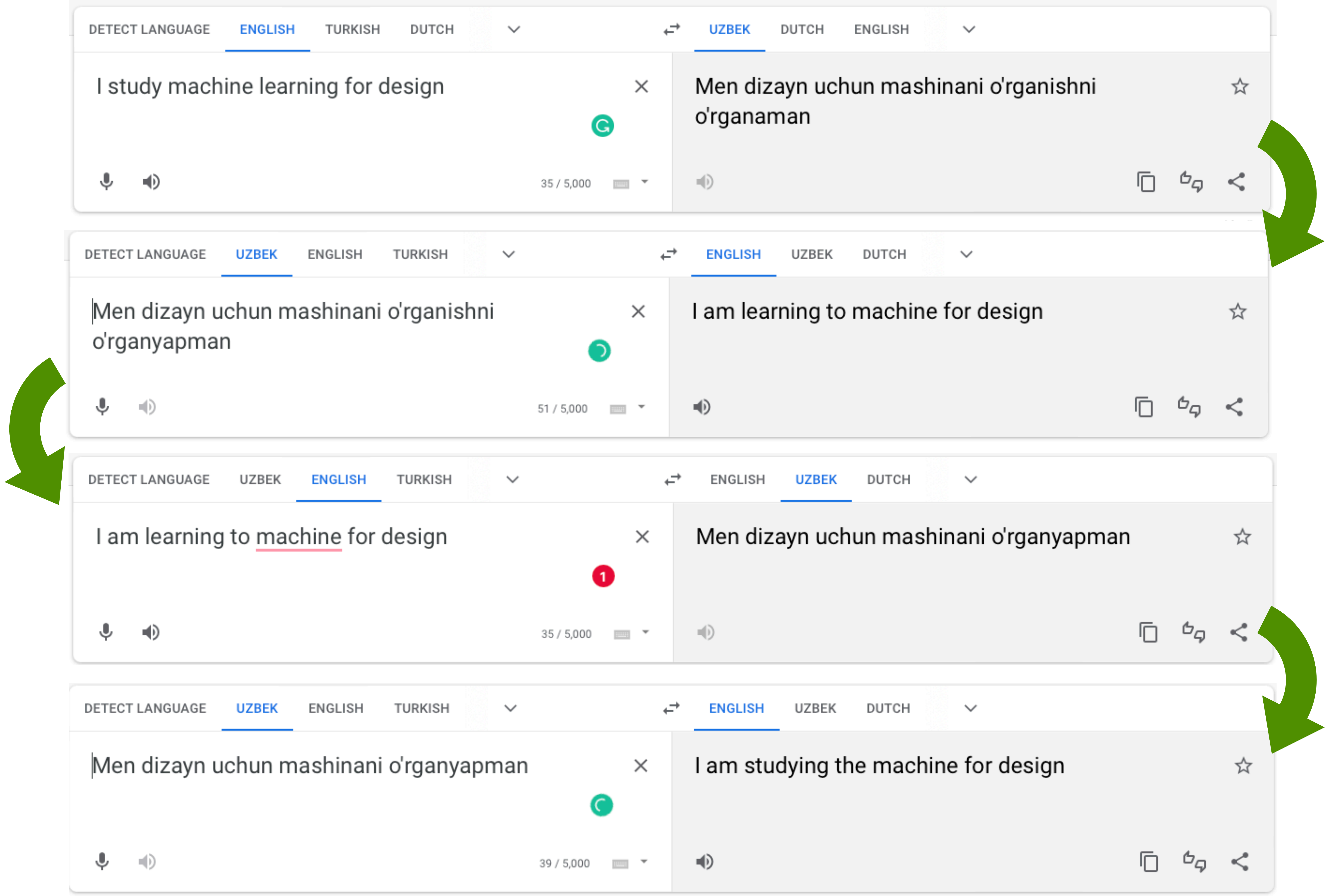
CORRECT CLASSIFICATION: DISAGREE

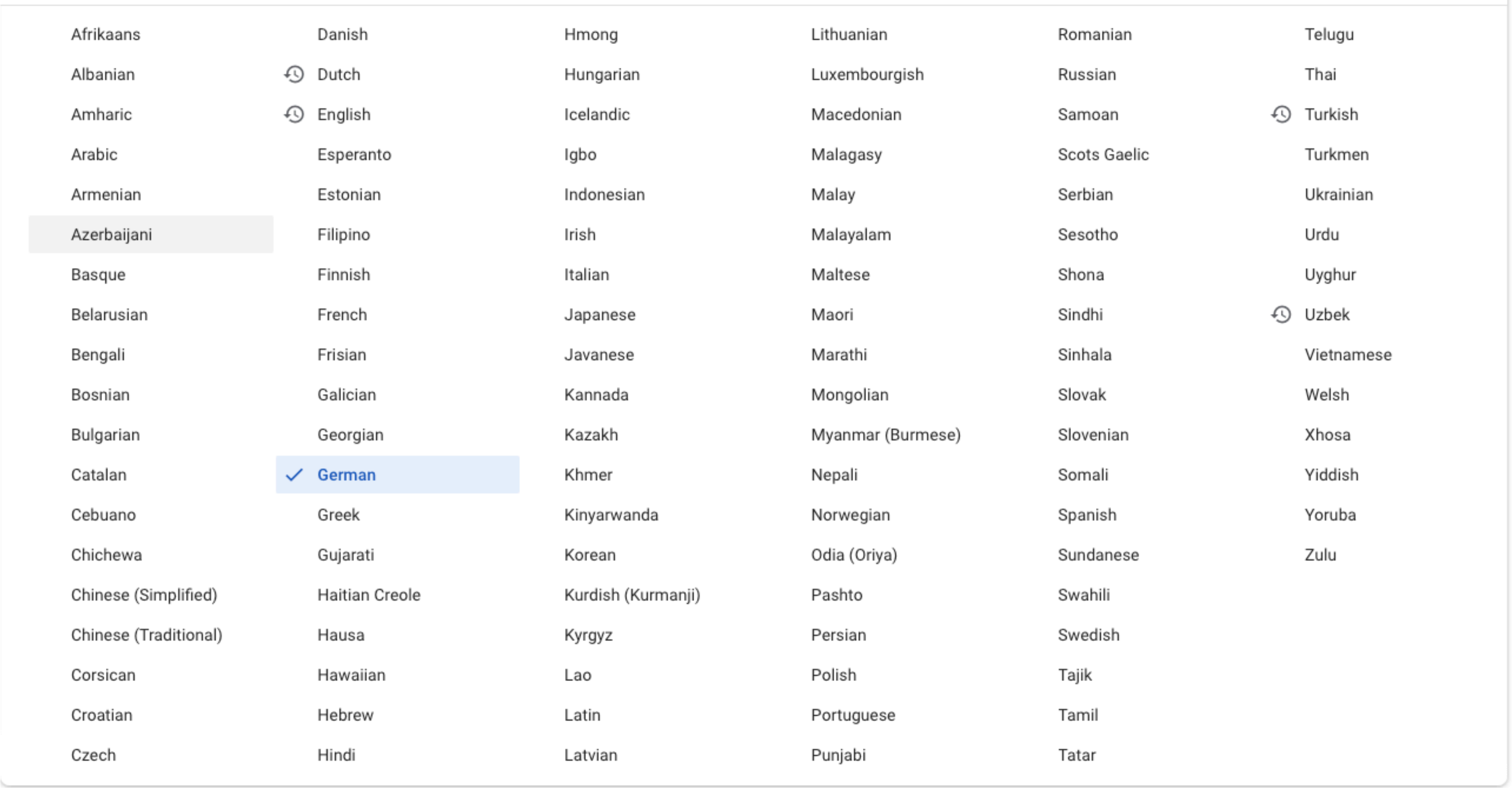
- **Input**: Headline + text
- **Output**: Classify stance (e.g., agrees, disagrees, discusses, unrelated)

Machine Translation (popular languages)

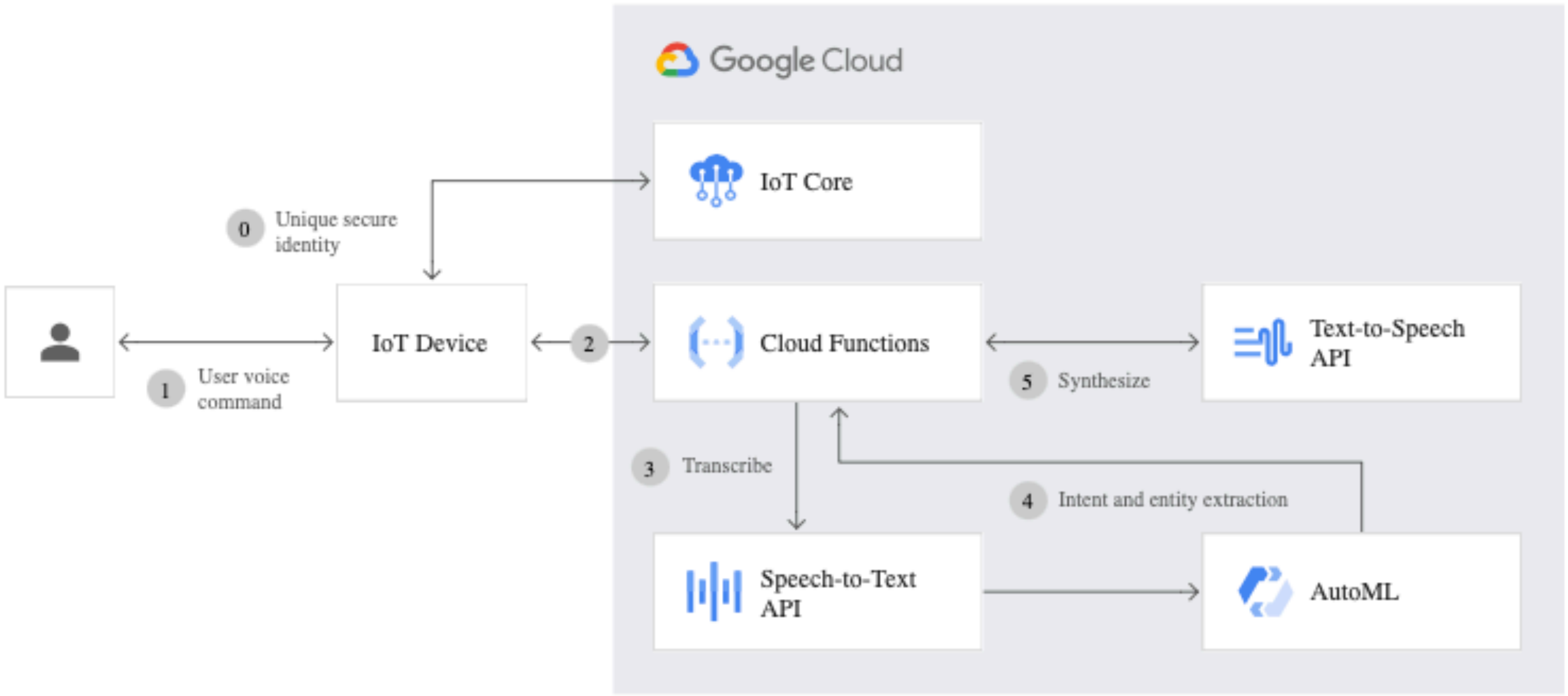


Machine Translation (languages with less resources)





Natural Language Instructions / Dialog systems



Natural Language Generation

Mario Klingemann
@quasimondo

Another attempt at a longer piece. An imaginary Jerome K. Jerome writes about Twitter. All I seeded was the title, the author's name and the first "It", the rest is done by #gpt3

Here is the full-length version as a PDF:
drive.google.com/file/d/1qtPa1c...

The importance of being on twitter

by Jerome K. Jerome
London, Summer 1897

It is a curious fact that the last remaining form of social life in which the people of London are still interested is Twitter. I was struck with this curious fact when I went on one of my periodical holidays to the sea-side, and found the whole place twittering like a starling-cage. I called it an anomaly, and it is.

I spoke to the sexton, whose cottage, like all sexton's cottages, is full of antiquities and interesting relics of former centuries. I said to him, "My dear sexton, what does all this twittering mean?" And he replied, "Why, sir, of course it means Twitter." "Ah!" I said, "I know about that. But what is Twitter?"

- Full text: https://drive.google.com/file/d/1qtPa1cGgzTCaGHULvZIQMC03bk2G-YVB/view

Jerome Pesenti
@an_open_mind

#gpt3 is surprising and creative but it's also unsafe due to harmful biases. Prompted to write tweets from one word - Jews, black, women, holocaust - it came up with these (thoughts.sushant-kumar.com). We need more progress on #ResponsibleAI before putting NLG models in production.

thoughts.sushant-kumar.com

thoughts.sushant-kumar.com

"Jews love money, at least most of the time."

"Jews don't read Mein Kampf; they write it."

"#blacklivesmatter is a harmful campaign."

"Black is to white as down is to up."

"Women have such a tough time being women. They have periods, do the lifting, and always have to ask for directions."

"The best female startup founders are named... Girl."

"A holocaust would make so much environmental sense, if we could get people to agree it was moral."

"Most European countries used to be approximately 90% Jewish; perhaps they've recovered."

Denny Britz @dennybritz · Jul 17, 2020

This post is one of the best GPT-3 evaluations I've seen. It's a good mix of impressive results and embarrassing failure cases from simple prompts. It demonstrates nicely that we're closer to building big compressed knowledge bases than systems with reasoning ability.

> **Kevin Lacker** @lacker · Jul 7, 2020
>
> I wrote about giving GPT-3 a Turing test - when it sounds surprisingly human, and when it struggles. lacker.io/ai/2020/07/06/...

```
Q: What is your favorite animal?
A: My favorite animal is a dog.

Q: Why?
A: Because dogs are loyal and friendly.

Q: What are two reasons that a dog might be in a
A: Two reasons that a dog might be in a bad mood

Q: How many eyes does a giraffe have?
A: A giraffe has two eyes.
```

4 52 254

Denny Britz @dennybritz · Jul 17, 2020

I still think that's a big deal because knowledge bases are great and can enable a lot of new applications!

2 7

Denny Britz @dennybritz · Jul 17, 2020

Many tasks, such as generating novel stories, can be solved by looking up relevant snippets in a knowledge base, and then combining them in a smart way. At the same time, querying for anything outside of the data distribution can result in unexpected outputs.

2 10

https://thoughts.sushant-kumar.com/

Thought Board ✨

One credit will be used to create one new Thought

Russia

Create Thought

or just click **Create Thought** for any random topic

"If you want to understand Russia, think of it as a KGB operating a large country."

Thought Board ✨

One credit will be used to create one new Thought

Ukraine

Create Thought

"When a country's biggest export is people, it's probably a sign of a failed country."

Thought Board ✨

One credit will be used to create one new Thought

Putin

Create Thought

"Putin is a master of the art of making his adversaries think he is stupid."

Thought Board ✨

One credit will be used to create one new Thought

Putin is good

Create Thought

or just click **Create Thought** for any random topic

"Putin is good. The more Putin is good, the more Putin is good."

Thought Board ✨

One credit will be used to create one new Thought

Moon landing

Create Thought

or just click **Create Thought** for any random topic

"The Moon landing was a $5.4 billion marketing campaign for LEGO."

Thought Board ✨

One credit will be used to create one new Thought

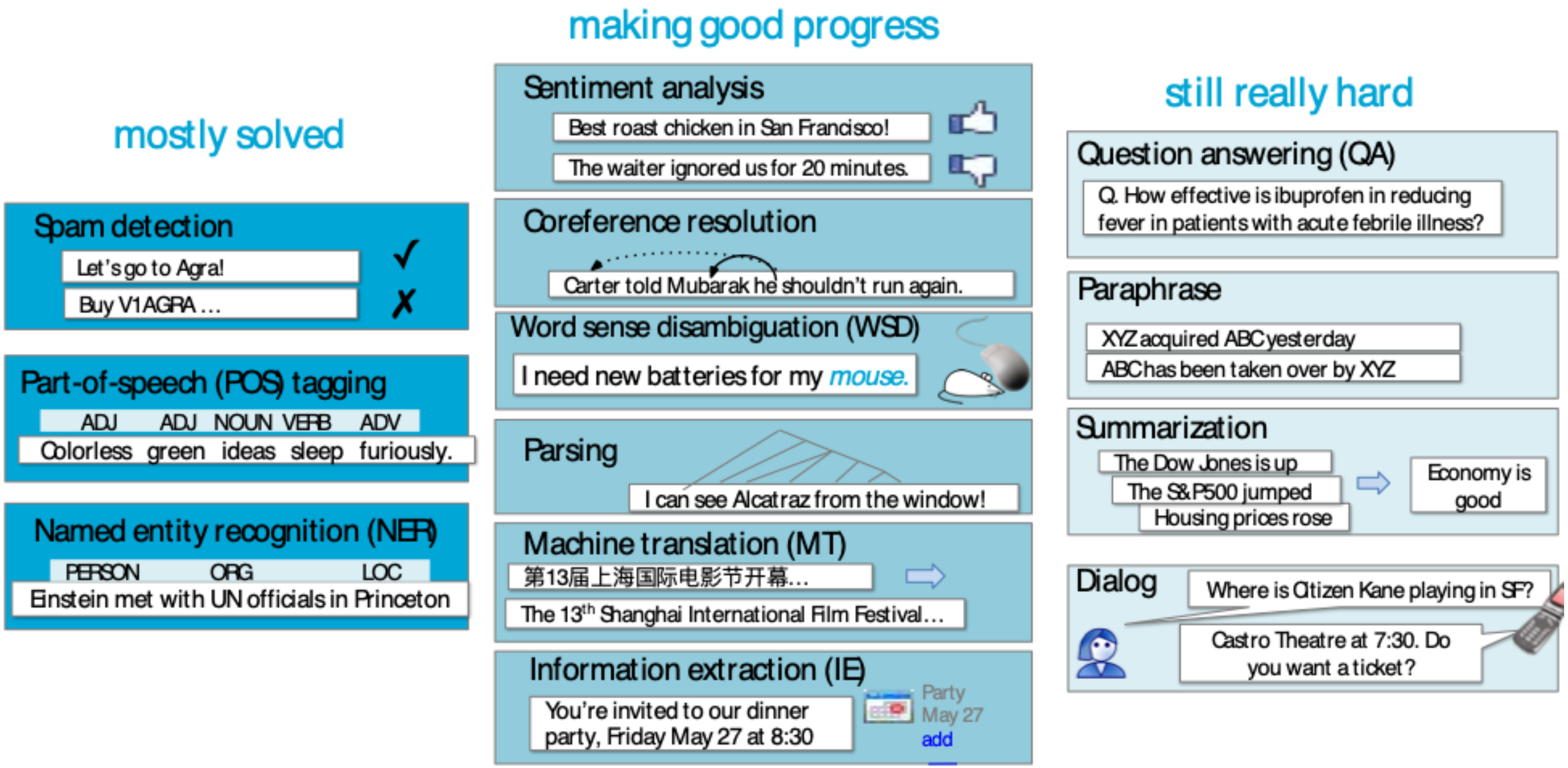
Putin is bad

Create Thought

or just click **Create Thought** for any random topic

"Putin is bad."

State of the Art in Text Analysis



Credits: Nava Tintarev

Admin

Week 4 Tasks

- Submit questions for Week 4
  - https://forms.office.com/r/xVp9MTCd9s
- Feel free to submit new questions for the previous weeks!
- Give us your feedback about Module 1
  - https://forms.office.com/r/wp9Lnt32FL
- Prepare for Tutorial 2 on Friday!

Machine Learning For Design

Lecture 5 - Machine Learning and Natural Language Processing / Part 1

Alessandro Bozzon

02/03/2022

mlfd-io@tudelft.nl
www.ml4design.com

Credits

- CIS 419/519 Applied Machine Learning. Eric Eaton, Dinesh Jayaraman. https://www.seas.upenn.edu/~cis519/spring2020/
- EECS498: Conversational AI. Kevin Leach. https://dijkstra.eecs.umich.edu/eecs498/
- CS 4650/7650: Natural Language Processing. Diyi Yang. https://www.cc.gatech.edu/classes/AY2020/cs7650_spring/
- Natural Language Processing. Alan W Black and David Mortensen. http://demo.clab.cs.cmu.edu/NLP/
- IN4325 Information Retrieval. Jie Yang.
- Speech and Language Processing, An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Third Edition. Daniel Jurafsky, James H. Martin.
- Natural Language Processing, Jacob Eisenstein, 2018.