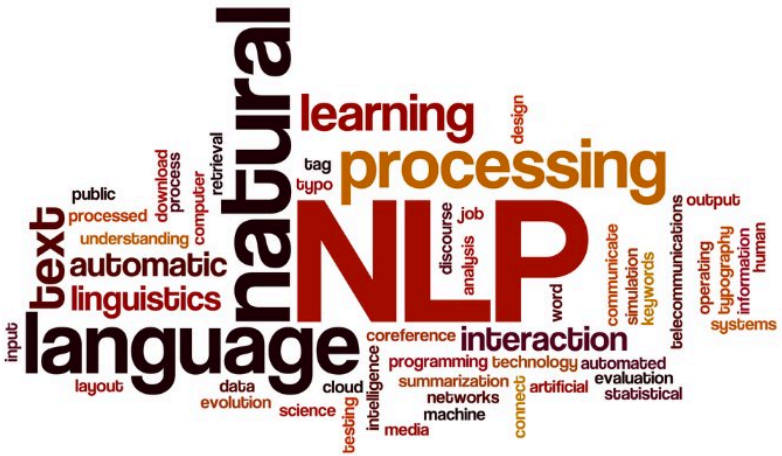
Introduction to Natural Language Processing

Nadi Tomeh

What is NLP ?


natural
learning
processing
NLP
language
text
automatic
linguistics
interaction
public
processed
understanding
download
process
computer
retrieval
tag
typo
design
discourse
analysis
job
output
communicate
simulation
keywords
word
telecommunications
operating
typography
information
human
systems
input
coreference
programming
technology
automated
evaluation
statistical
summarization
artificial
networks
machine
media
intelligence
testing
connect
layout
data
evolution
science
cloud

What is Natural Language Processing ?

- Natural Language
  - languages learned without explicit instruction by children from speakers in their environments

    English, French, Chinese, … (how many?)
  - link **sensory-motor programs**
    - sounds in spoken languages
    - gestures in signed languages
    - visual patterns in written languages, etc.

with **conceptual/intentional representations**

What is **N**atural **L**anguage **P**rocessing ?

- **P**rocessing
  - Automating language analysis, generation, acquisition.

Analysis (or "understanding"): input is language, output is some representation that supports useful action

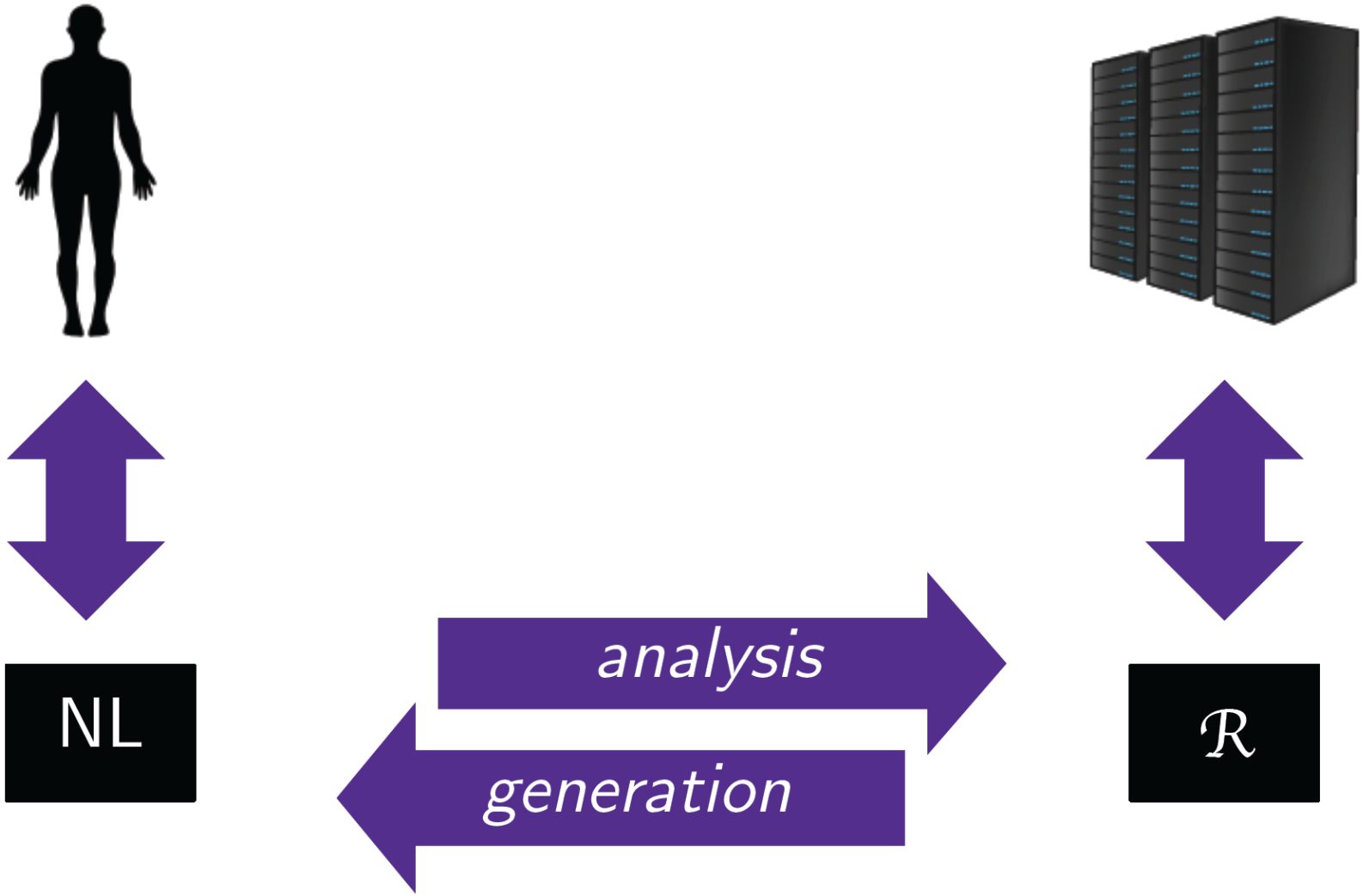
$$\mathbf{NL} \rightarrow \mathcal{R}$$

Generation: input is that representation, output is language

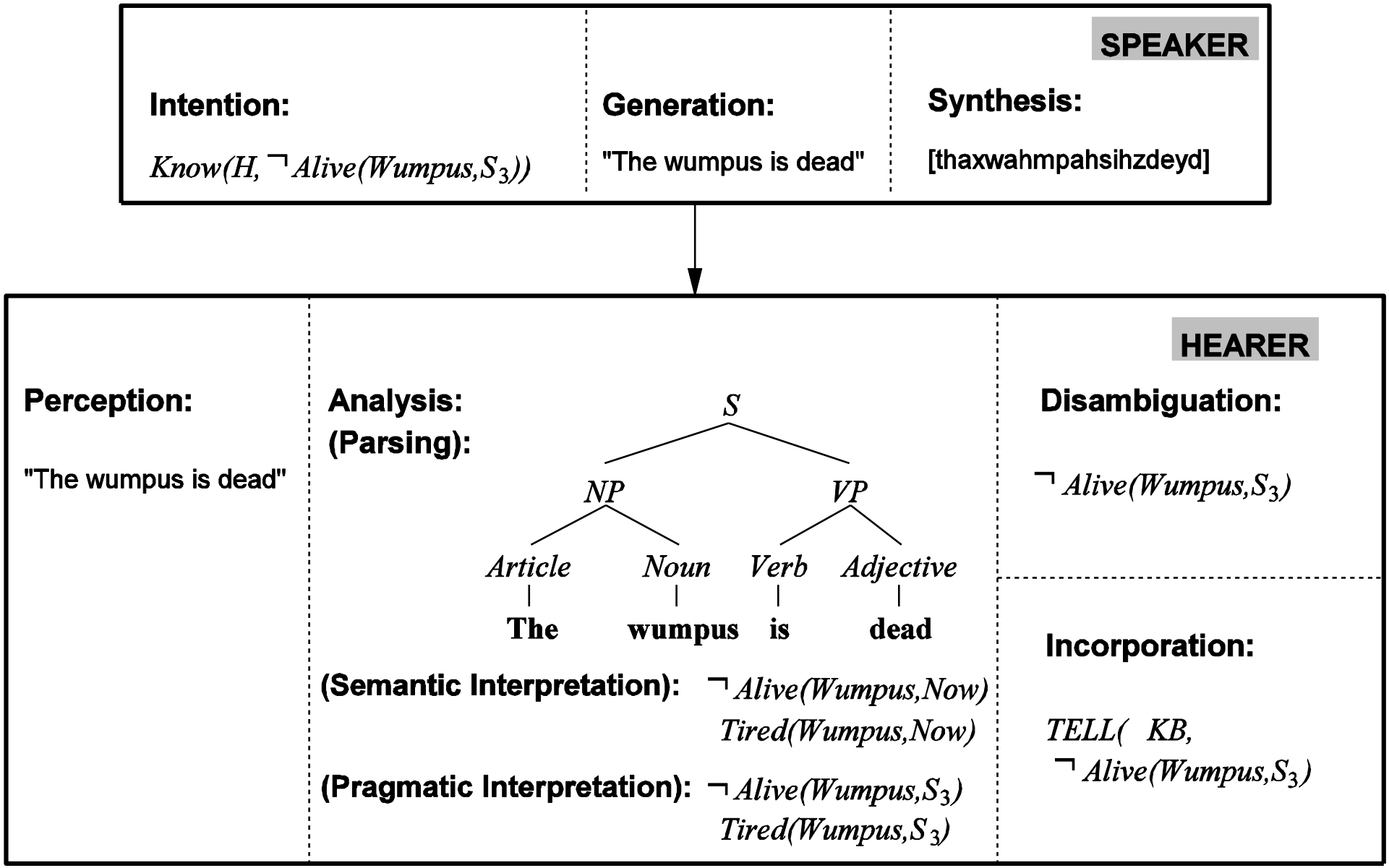
$$\mathcal{R} \rightarrow \mathbf{NL}$$

Acquisition: obtaining the representation and necessary algorithms, from knowledge and data

What is NLP?



Communication



Russel & Norvig, 1995

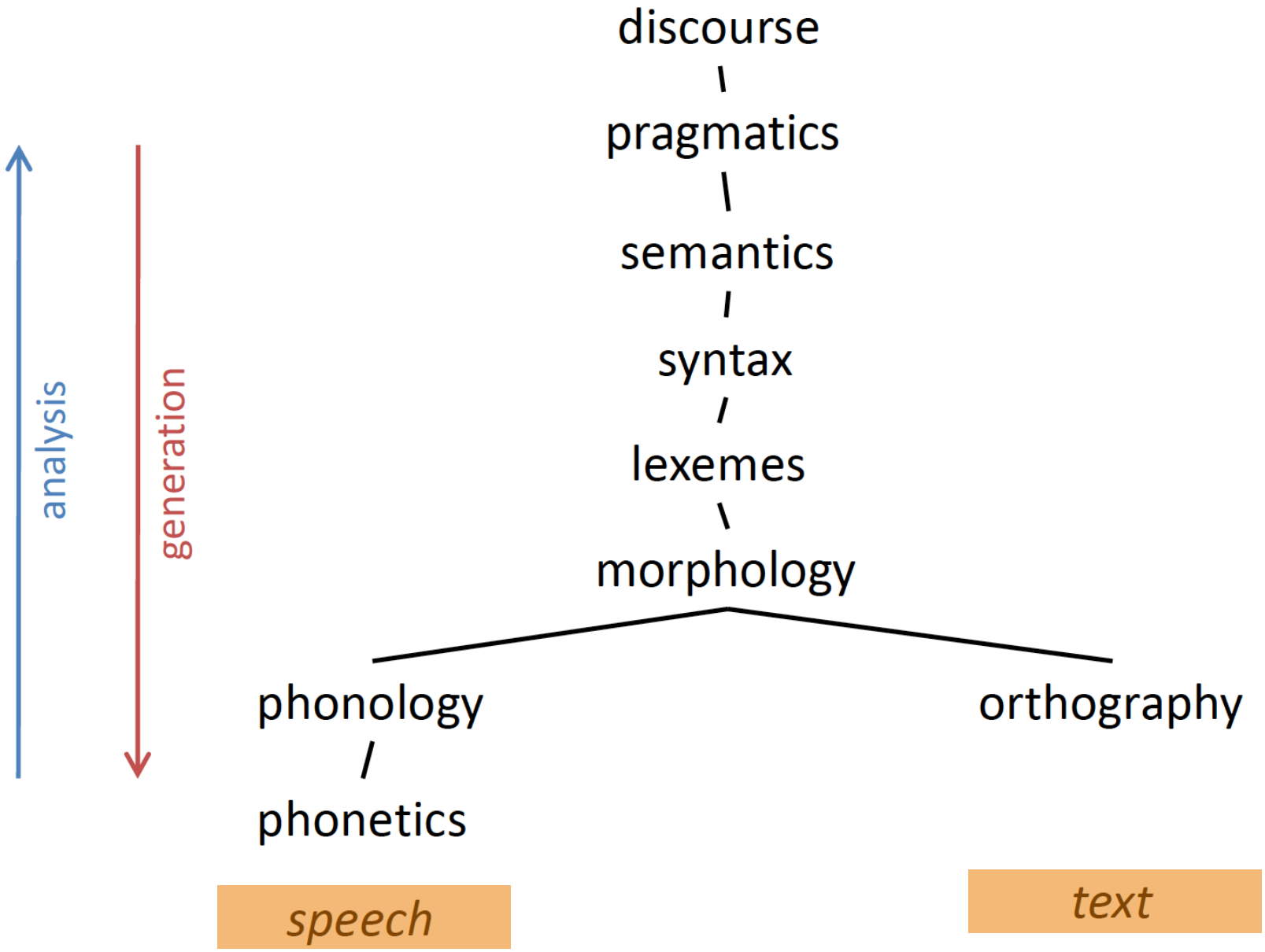
Note on terminology

- Some people use NLP to refer to all language technologies, some people use it to refer only to analysis
- NLP vs. Computational Linguistics (CL)
  - NLP focuses on the technology of processing language
    Engineering: useful tasks involving text (IR, MT, …)
  - CL on using technology to support/implement linguistics
    Empirical science
  - Similar to AI vs. cognitive science

	Exact sciences	*Empirical sciences*	*Engineering*
Deals in...	Axioms and theorems	Facts and theories	Artifacts
Truth is...	Forever	Temporary	It works!
Examples...	Maths, CS theory, **formal language theory**	Physics, Biology, **Linguistics**	Many, inc. Applied CS and **NLP**

Levels of Linguistic Knowledge

Levels of linguistic knowledge



Why is NLP Hard ?

Why is NLP hard ?

- We always knew Language is hard
  - AI complete

    to solve NLP you need to solve all the problems in AI (word knowledge)
  - Turing test

    engaging effectively in linguistic behavior is sufficient condition for having achieved intelligence (communication)
- But little kids can “do” NLP
  - Why is it hard?

Why is NLP hard ?

John stopped at the donut store on his way home from work. He thought a coffee was good every few hours. But it turned out to be too expensive there.

Why is NLP hard ?

John stopped at the **donut** store on his way home from work. He thought a coffee was good every few hours. But it turned out to be too expensive there.

To get a donut (spare tire) for his car?

Why is NLP hard ?

John stopped at the **donut store** on his way home from work. He thought a coffee was good every few hours. But it turned out to be too expensive there.

Store where donuts shop? or is run by donuts?

or looks like a big donut? or made of donut?

or has an emptiness at its core?

Why is NLP hard ?

John stopped at the donut store **on his way home from work**. He thought a coffee was good every few hours. But it turned out to be too expensive there.

Describes where the store is? Or when he stopped?

Why is NLP hard ?

John stopped at the donut store on his way home **from work**. He thought a coffee was good every few hours. But it turned out to be too expensive there.

Well, actually, he stopped there from hunger and exhaustion, not just from work.

Why is NLP hard ?

John stopped at the donut store on his way home from work. **He thought** a coffee was good every few hours. But it turned out to be too expensive there.

At that moment, or habitually?

(*Similarly:* Mozart composed music.)

Why is NLP hard ?

John stopped at the donut store on his way home from work. He thought a coffee was good **every few hours.** But it turned out to be too expensive there.

That's how often he thought it?

Why is NLP hard ?

John stopped at the donut store on his way home from work. He thought **a coffee was good** every few hours. But it turned out to be too expensive there.

But actually, a coffee only stays good for about 10 minutes before it gets cold.

Why is NLP hard ?

John stopped at the donut store on his way home from work. He thought a coffee was good every few hours. But **it** turned out to be too expensive there.

the particular coffee that was good every few hours? the donut store? the situation?

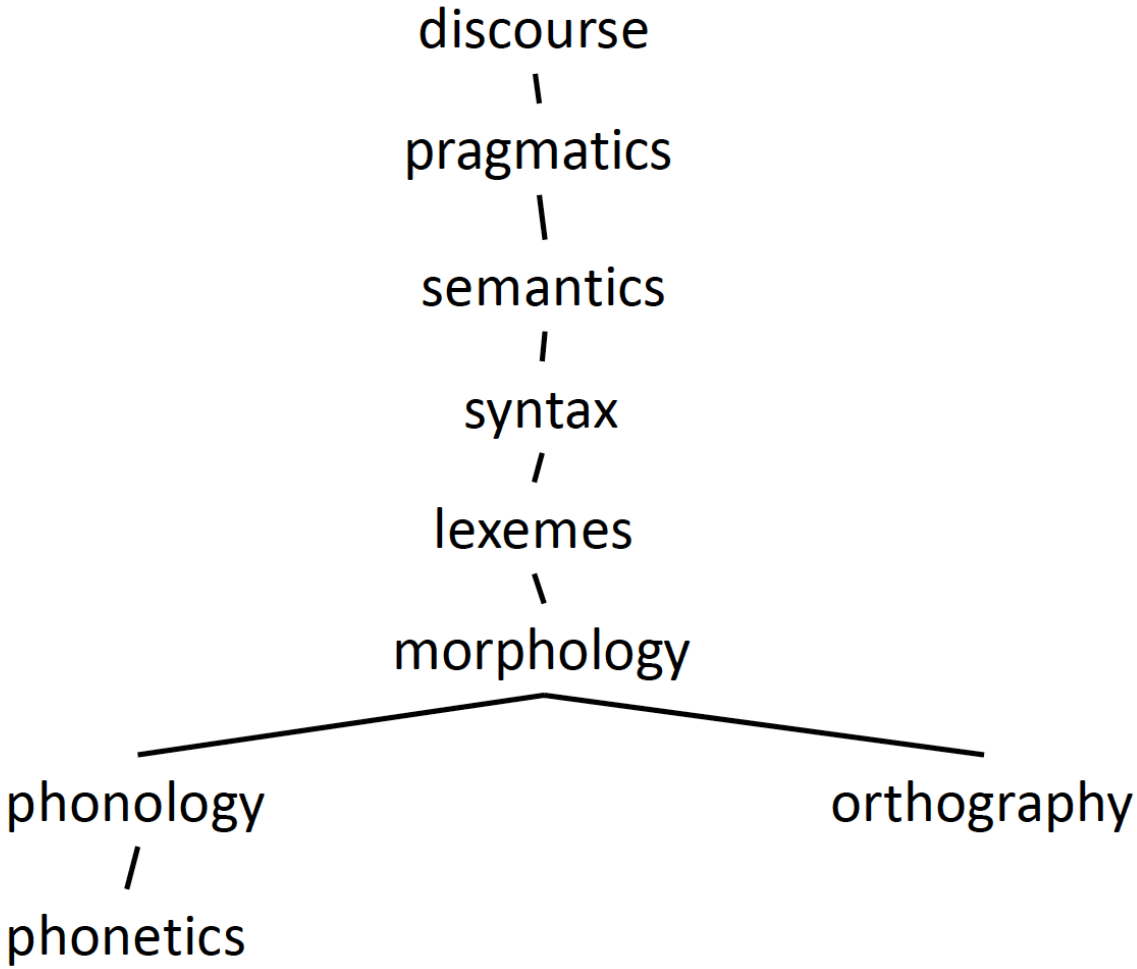
Why is NLP hard ?

John stopped at the donut store on his way home from work. He thought a coffee was good every few hours. But it turned out to be **too expensive** there.

too expensive for what? what are we supposed to conclude about what John did?

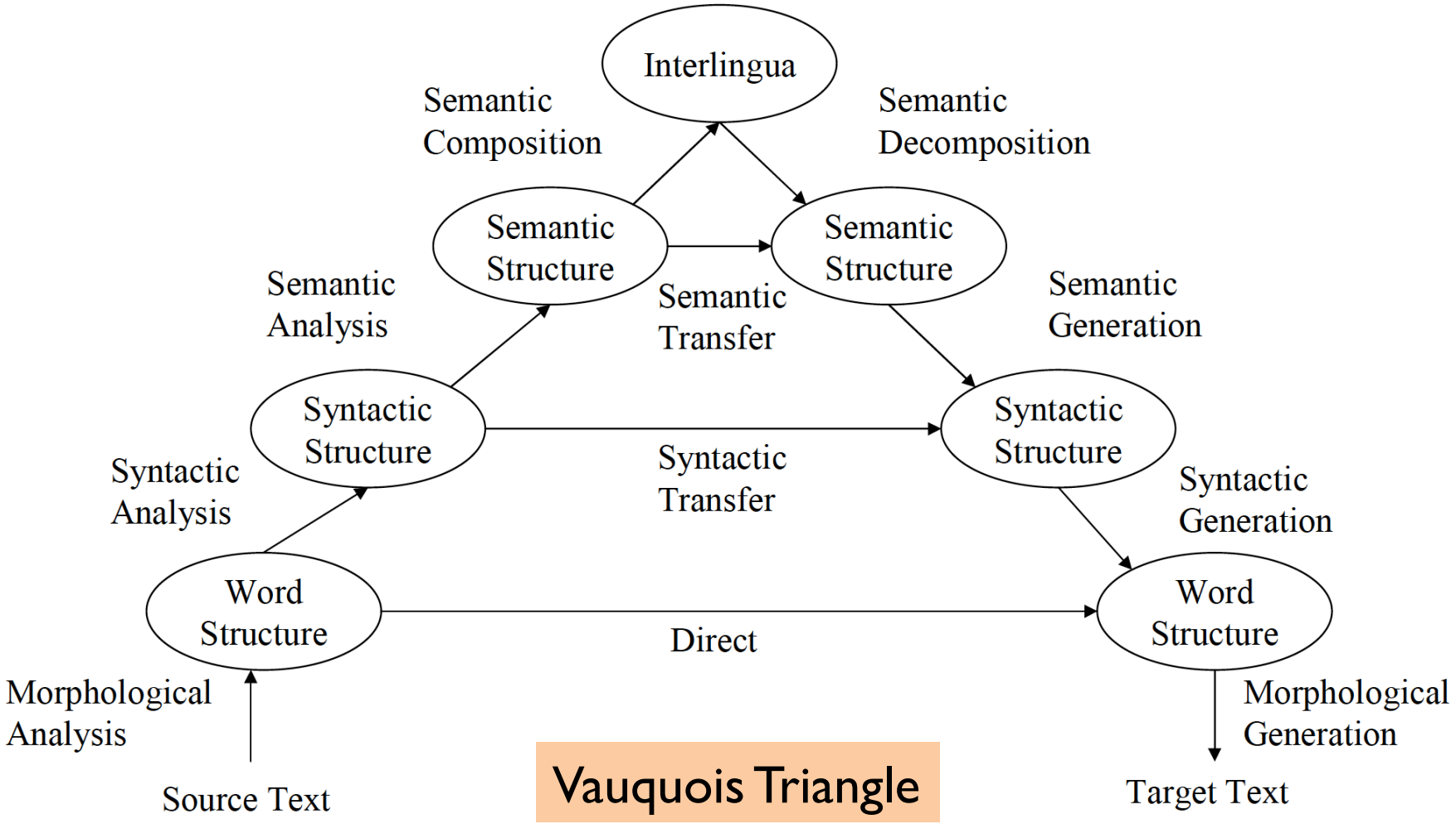
Why is NLP hard ?

- Mappings between levels is extremely complex
  - Each level ***interacts*** with the others



Why is NLP hard ?

- Appropriateness of a representation depends on the application, e.g.: machine translation



Why is NLP hard ?

- **Ambiguity**: each string may have many possible interpretations at every level
  - The Pope's baby steps on gays
  - Boy paralyzed after tumor fights back to gain black belt
  - Scientists study whales from space

  – The correct resolution of the ambiguity will depend on the ***intended meaning*** (which is often inferable from context)

  – People are good at linguistic ambiguity resolution, computers, not so!
    - How do we represent sets of possible alternatives?
    - How do we represent context?

Ambiguity is Ubiquitous

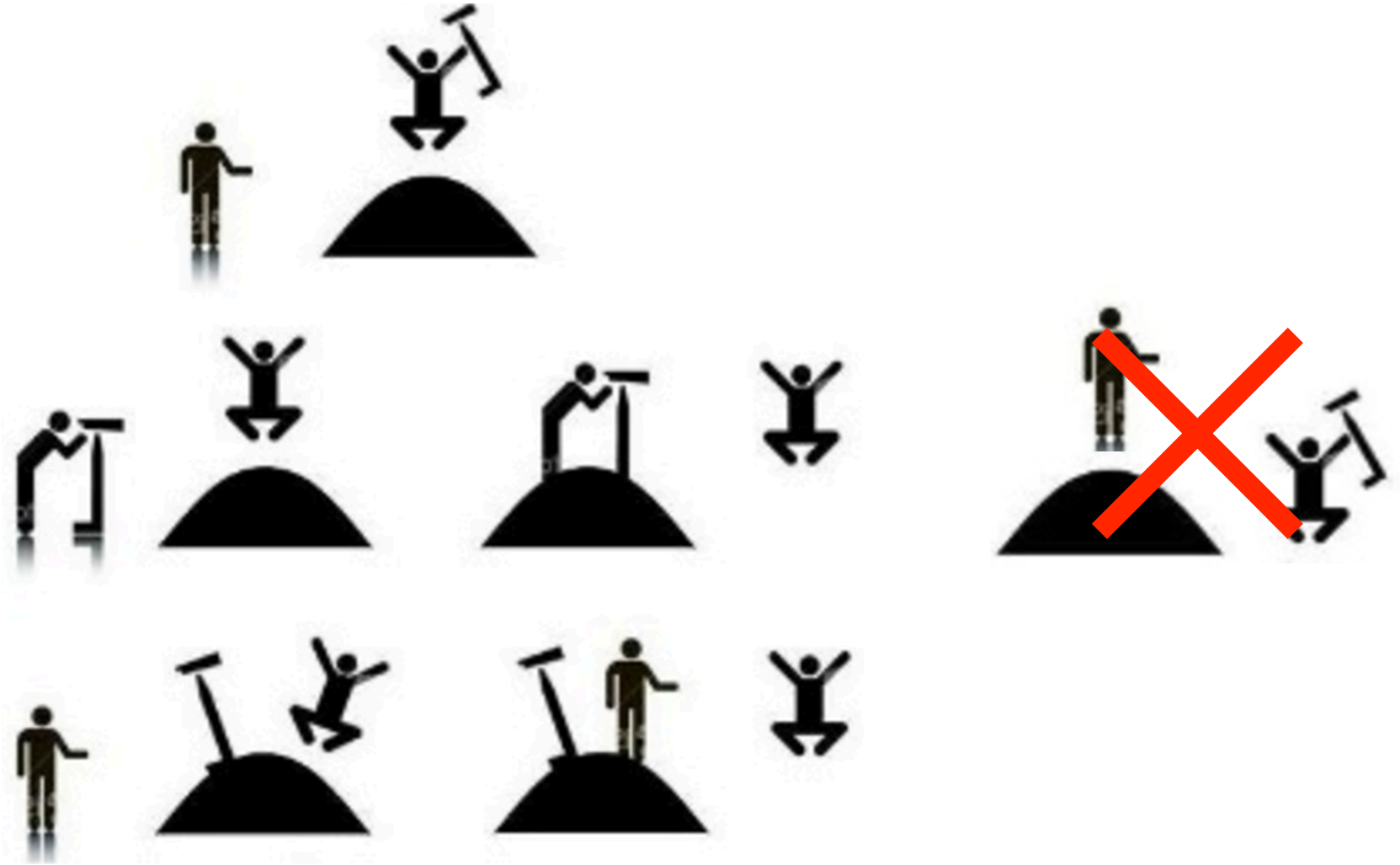
- Speech Recognition
  - "recognize speech" vs. "wreck a nice beach"
  - "youth in Asia" vs. "euthanasia"

- Syntactic
  - books:
    book(N) + plural vs.
    book(V) + present + singular + 3rd person
  - "Time flies like an arrow"
    - Metaphor: Time[N] flies[V] like[PRP] an[ART] arrow[N]
    - Fly species: Time[N] flies[N] like[V] an[ART] arrow[N]
    - Imperative: Time[V] flies[N] like[PRP] an[ART] arrow[N]

Ambiguity is Ubiquitous

Ambiguity is Ubiquitous

- Syntactic (attachment)
  - "I saw a man on a hill with a telescope"



Ambiguity is Ubiquitous

- Semantic (lexical)
  - "I went to the bank …

    of the river
    to withdraw money
- Semantic
  - "John an Marry are married"
    To each other? Or separately?
  - "John kissed his wife, and so did Sam"
    Sam kissed John's wife, or his own?
- Pragmatic
  - "Can you pass the salt?"
  - From "The Pink Panther Strikes Again":
    Clouseau: Does your dog bite?
    Hotel Clerk: No.
    Clouseau: [*bowing down to pet the dog*] Nice doggie.
    [*Dog barks and bites Clouseau in the hand*]
    Clouseau: I thought you said your dog did not bite!
    Hotel Clerk: That is not my dog.

Ambiguity is Explosive

- Ambiguities compound to generate enormous numbers of possible interpretations.

- In English, a sentence ending in *n* prepositional phrases has *over* 2^n syntactic interpretations (cf. Catalan numbers).

  - "I saw the man with the telescope": 2 parses
  - "I saw the man on the hill with the telescope.": 5 parses
  - "I saw the man on the hill in Texas with the telescope": 14 parses
  - "I saw the man on the hill in Texas with the telescope at noon.": 42 parses
  - "I saw the man on the hill in Texas with the telescope at noon on Monday" 132 parses

Why is Language Ambiguous?

- Having a unique linguistic expression for every possible conceptualization that could be conveyed would make language overly complex and linguistic expressions unnecessarily long.

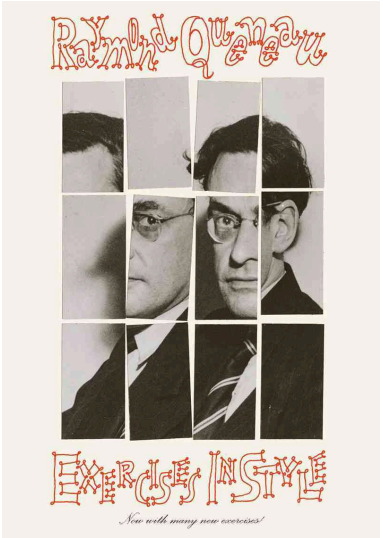
- Allowing resolvable ambiguity permits shorter linguistic expressions, i.e. data compression.

- Language relies on people's ability to use their knowledge and inference abilities to properly resolve ambiguities.

- Infrequently, disambiguation fails, i.e. the compression is lossy.

Why is NLP hard ?

- **Variability**: there are many ways to express the same meaning, and immeasurably many meanings to express (**richness)**
  - **Lexical:**
    - book, volume, document, best-seller, publication…
  - **Syntactic:**
    - “John gives Mary a book” vs. “John gives a book to Mary”
  - **Referential:**
    - “the table” vs. “the piece of furniture in the corner”
  - **Stylistic:**
    - Raymond Queneau’s Exercises in Style
      (99 ways to retell the same story in different styles)



Why is NLP hard ?

- Linguistic representations are ***theorized*** constructs; we cannot observe them directly

Why is NLP hard ?

- There is tremendous ***diversity*** in human languages
  - Structural divergences
    - SOV (Jap) vs. SVO (En, FR) vs. VSO (Ar) vs. Free order
  - Some languages express some meanings more readily/often:
    - Dépayser (fr): To feel displaced from one's native land
    - Trouvaille (fr): Something awesome that was discovered by chance
    - Torschlusspanik (de): last minute panic; the fear, usually as one gets older, that time is running out and important opportunities are slipping away
  - Many more …
- Input is likely to be **noisy** (errors)
- **Non-standard** writing (non-edited): "Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥"

NLP Tasks

Syntactic

Word Segmentation and Normalization

- Breaking a string of characters (graphemes) into a sequence of tokens (words).

J.P. Bolduc, vice chairman of W.R. Grace & Co., which holds a 83.4% interest in this energy-services company, was elected a director. Clark J. Vitulli was named senior vice president and general manager of this U.S. sales and marketing arm of Japanese auto maker Mazda Motor Corp.

Words:	chairman, of, energy-services
Numerical expressions:	83.4, 10,000
Punctuation:	., ?
Other symbols:	&, %

- Possibly "normalize"
  - Abbreviations: U.S., US → U.S.
  - Case folding: Window, window → Window
  - Lemmatization: window, windows → window
  - Stemming: compute, computing, computer → comput

Word Segmentation and Normalization

- In some written languages (e.g. Chinese) words are not separated by spaces.

(a)

下雨天留客天留我不留	Unpunctuated Chinese sentence
下雨、天留客。天留、我不留！	*It is raining, the god would like the guest to stay. Although the god wants you to stay, I do not!*
下雨天、留客天。留我不？留！	*The rainy day, the staying day. Would you like me to stay? Sure!*

(b)

我喜欢新西兰花	Unsegmented Chinese sentence
我　喜欢　新西兰　花	*I like New Zealand flowers*
我　喜欢　新　西兰花	*I like fresh broccoli*

Morphological Analysis

- ***Morphology*** is the field of linguistics that studies the internal structure of words. (Wikipedia)
- A ***morpheme*** is the smallest linguistic unit that has semantic meaning (Wikipedia)
  - e.g. "carry", "pre", "ed", "ly", "s"
- Morphological analysis is the task of segmenting a word into its morphemes:
  - carried ⇒ carry + ed (past tense)
  - independently ⇒ in + (depend + ent) + ly
  - Googlers ⇒ (Google + er) + s (plural)
  - unlockable ⇒ un + (lock + able) ? | ⇒ (un + lock) + able ?
  - uygarlaştıramadıklarımızdanmışsınızcasına
    "(behaving) as if you are among those whom we could not civilize"

Lexical Analysis: Part Of Speech (POS) Tagging

- Annotate each word in a sentence with a part-of-speech.

I ate the spaghetti with meatballs.
Pro V Det N Prep N

John saw the saw and decided to take it to the table.
PN V Det N Con V Part V Pro Prep Det N

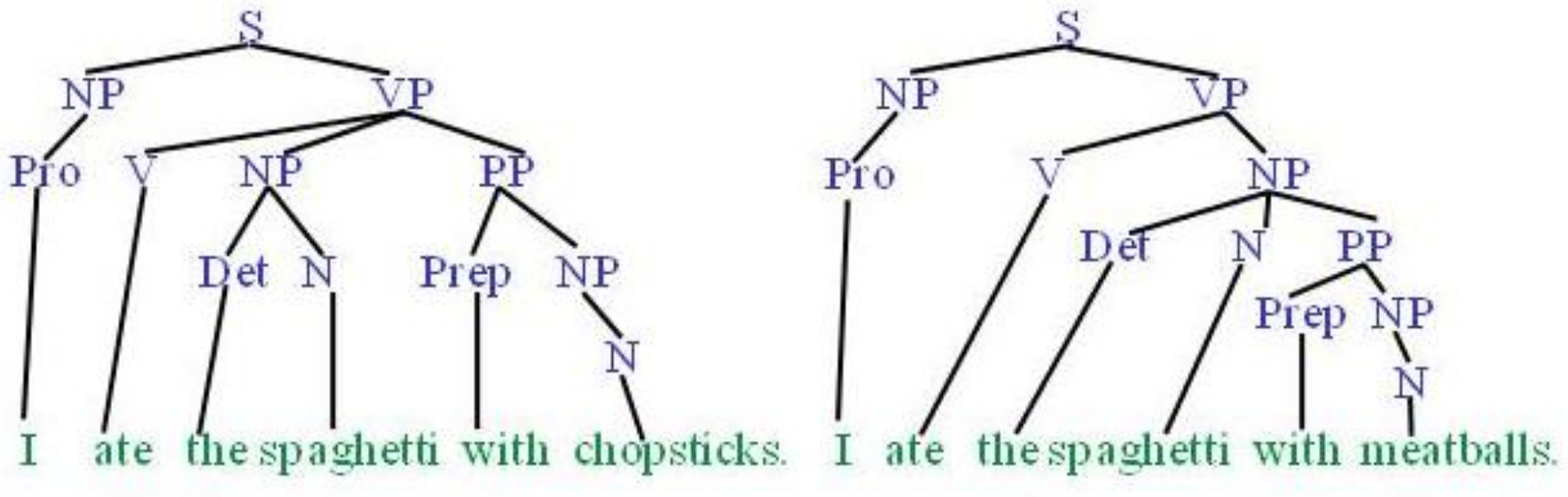
- Useful for subsequent syntactic parsing and word sense disambiguation.

Phrase Chunking

- Find all non-recursive noun phrases (NPs) and verb phrases (VPs) in a sentence.
  - [NP I] [VP ate] [NP the spaghetti] [PP with] [NP meatballs].
  - [NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only # 1.8 billion] [PP in] [NP September]

Syntactic Parsing

- Transform a sequence of symbols into a hierarchical or compositional structure



- Syntactic theories: what make some sentences *well-formed* and others not
  - I found a flight to Tokyo vs. I found to fly to Tokyo

NLP Tasks

Semantic

Word Sense Disambiguation (WSD)

- Words in natural language usually have a fair number of different possible meanings.
  - Ellen has a strong interest in computational linguistics.
  - Ellen pays a large amount of interest on her credit card.
- For many tasks (question answering, translation), the proper sense of each ambiguous word in a sentence must be determined.

Semantic Role Labeling (SRL)

- For each clause, determine the semantic role played by each noun phrase that is an argument to the verb.

  agent patient source destination instrument

  – John drove Mary from Austin to Dallas in his Toyota Prius.

  – The hammer broke the window.

- Also referred to a "case role analysis," "thematic analysis," and "shallow semantic parsing"

Semantic Parsing

- A ***semantic parser*** maps a natural-language sentence to a complete, detailed semantic representation (***logical form***).

- For many applications, the desired output is immediately executable by another program.

- Example: Mapping an English database query to Prolog:

How many cities are there in the US?

```
answer(A,   count(B,
                  (city(B), loc(B, C), const(C, countryid(USA))),
                  A) )
```

Textual Entailment

Determine whether one natural language sentence entails (implies) another under an ordinary interpretation.

TEXT	HYPOTHESIS	ENTAILMENT
Eyeing the huge market potential, currently led by Google, Yahoo took over search company Overture Services Inc last year.	Yahoo bought Overture.	TRUE
Microsoft's rival Sun Microsystems Inc. bought Star Office last month and plans to boost its development as a Web-based device running over the Net on personal computers and Internet appliances.	Microsoft bought Star Office.	FALSE
The National Institute for Psychobiology in Israel was established in May 1971 as the Israel Center for Psychobiology by Prof. Joel.	Israel was established in May 1971.	FALSE
Since its formation in 1948, Israel fought many wars with neighboring Arab countries.	Israel was established in 1948.	TRUE

NLP Tasks

Pragmatic and Discourse

Anaphora Resolution / Co-Reference

- Determine which phrases in a document refer to the same underlying entity.
  - John put the carrot on the plate and ate it.
  - Bush started the war in Iraq. But the president needed the consent of Congress.
- Some cases require difficult reasoning.
  - Today was Jack's birthday. Penny and Janet went to the store. They were going to get presents. Janet decided to get a kite. "Don't do that," said Penny. "Jack has a kite. He will make you take it back."

Ellipsis Resolution

- Frequently words and phrases are omitted from sentences when they can be inferred from context.

"Wise men talk because they have something to say; fools, because they have to say something." (Plato)

talk because they have to say something

NLP Applications

Information Extraction (IE)

Identify phrases in language that refer to specific types of entities and relations in text.

- Named entity recognition is task of identifying names of people, places, organizations, etc. in text.

  people organizations places

  – Michael Dell is the CEO of Dell Computer Corporation and lives in Austin Texas.

- Relation extraction identifies specific relations between entities.

  – Michael Dell is the CEO of Dell Computer Corporation and lives in Austin Texas.

Information Extraction (IE)

Kurt Gödel

Born	Kurt Friedrich Gödel April 28, 1906 Brünn, Austria-Hungary (now Brno, Czech Republic)
Died	January 14, 1978 (aged 71) Princeton, New Jersey, U.S.
Residence	Austria, United States
Citizenship	Austria, United States
Nationality	Austrian
Fields	Mathematics, Mathematical logic
Institutions	Institute for Advanced Study
Alma mater	University of Vienna

Thesis	*Über die Vollständigkeit des Logikkalküls (On the Completeness of the Calculus of Logic)* (1929)
Doctoral advisor	Hans Hahn
Known for	Gödel's incompleteness theorems, Gödel's completeness theorem, the consistency of the Continuum hypothesis with ZFC, Gödel metric, Gödel's ontological proof, Gödel–Dummett logic
Notable awards	Albert Einstein Award (1951) National Medal of Science (1974) ForMemRS (1968)[1] Fellow of the British Academy[*citation needed*]

Signature

Question Answering

Directly answer natural language questions based on information presented in a corpora of textual documents (e.g. the web).

- When was Barack Obama born? (*factoid*)
  - August 4, 1961
- Who was president when Barack Obama was born?
  - John F. Kennedy
- How many presidents have there been since Barack Obama was born?
  - 9

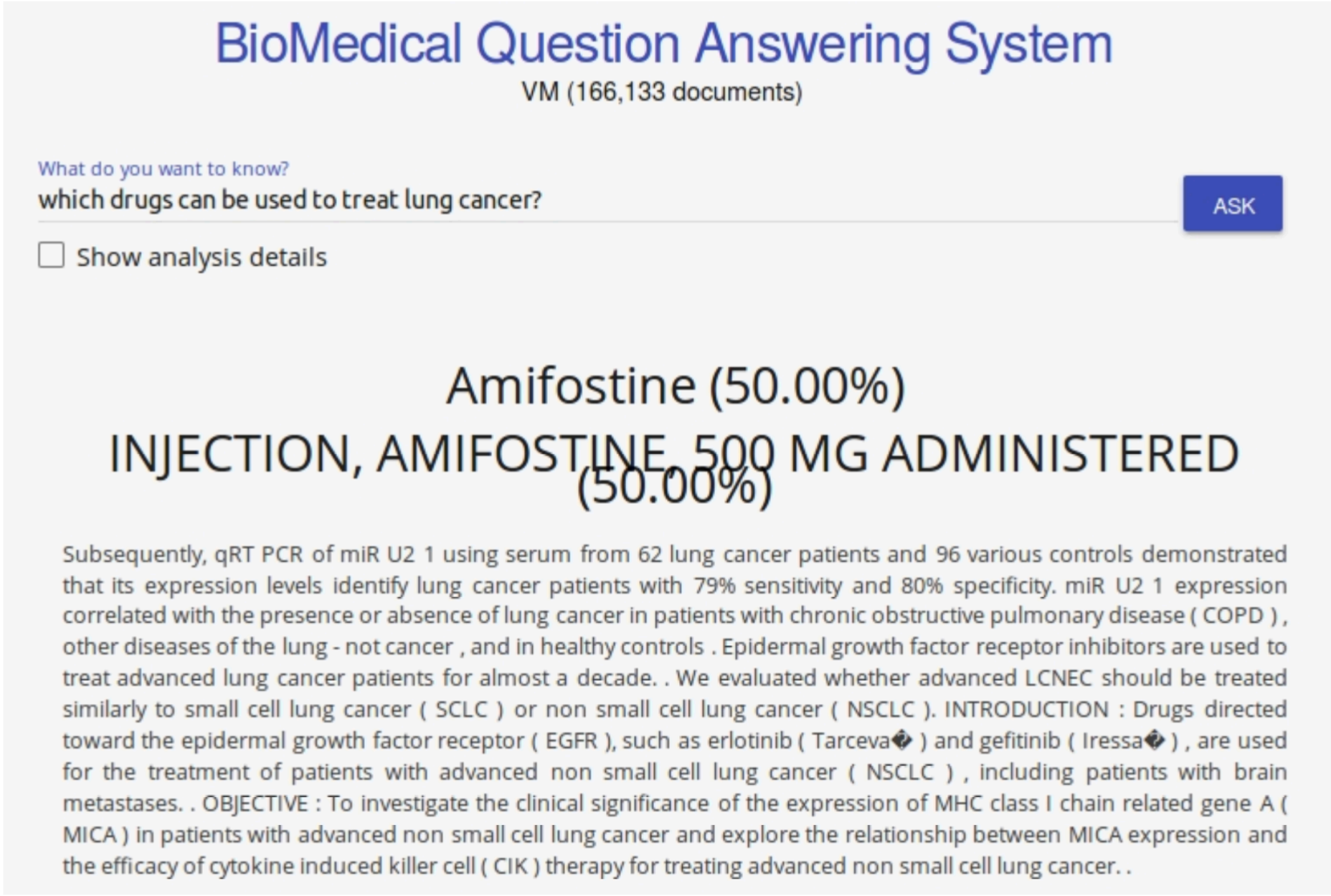
Question Answering

===> what countries speak Spanish

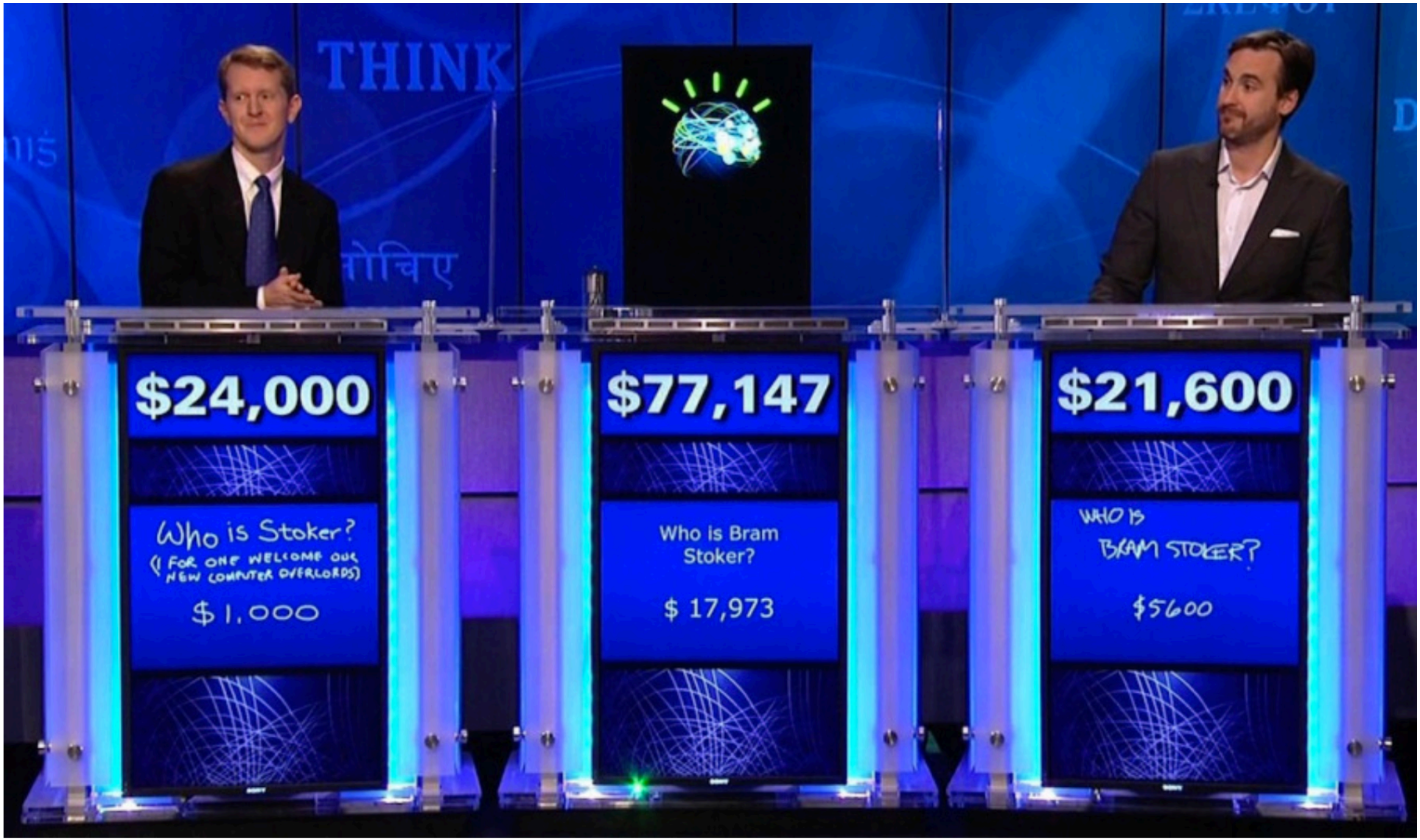
The language Spanish is spoken in Argentina, Aruba, Belize, Bolivia, Brazil, Canada, Cayman Islands, Chile, Colombia, Costa Rica, Cuba, Curacao, Dominican Republic, Ecuador, El Salvador, Equatorial Guinea, Falkland Islands (Islas Malvinas), Gibraltar, Guatemala, Honduras, Mexico, Nicaragua, Panama, Paraguay, Peru, Puerto Rico, Saint Martin, Sint Maarten, Spain, Switzerland, Trinidad and Tobago, United States, Uruguay, Venezuela, and Virgin Islands.

The language Castilian Spanish is spoken in Spain.

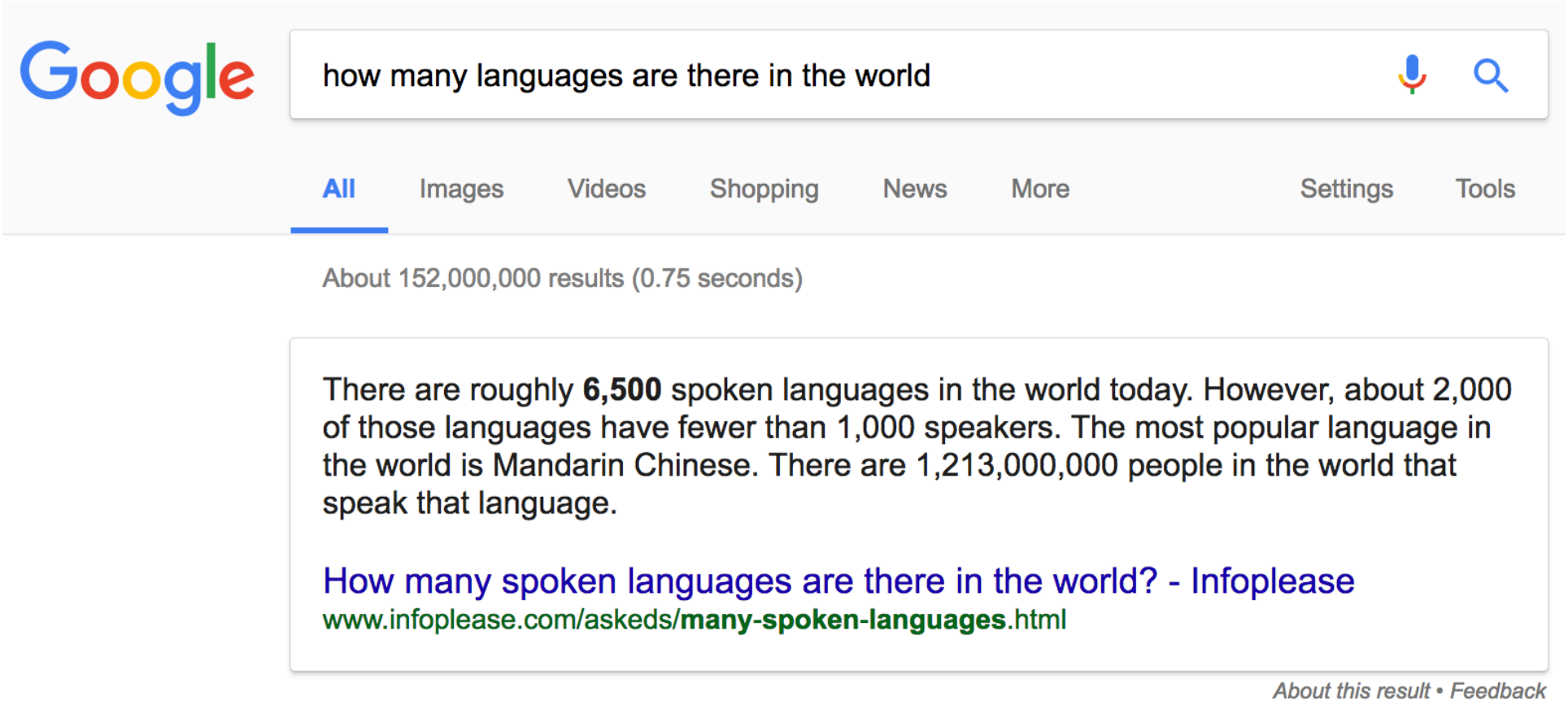
Question Answering



Question Answering



How many languages in the world?



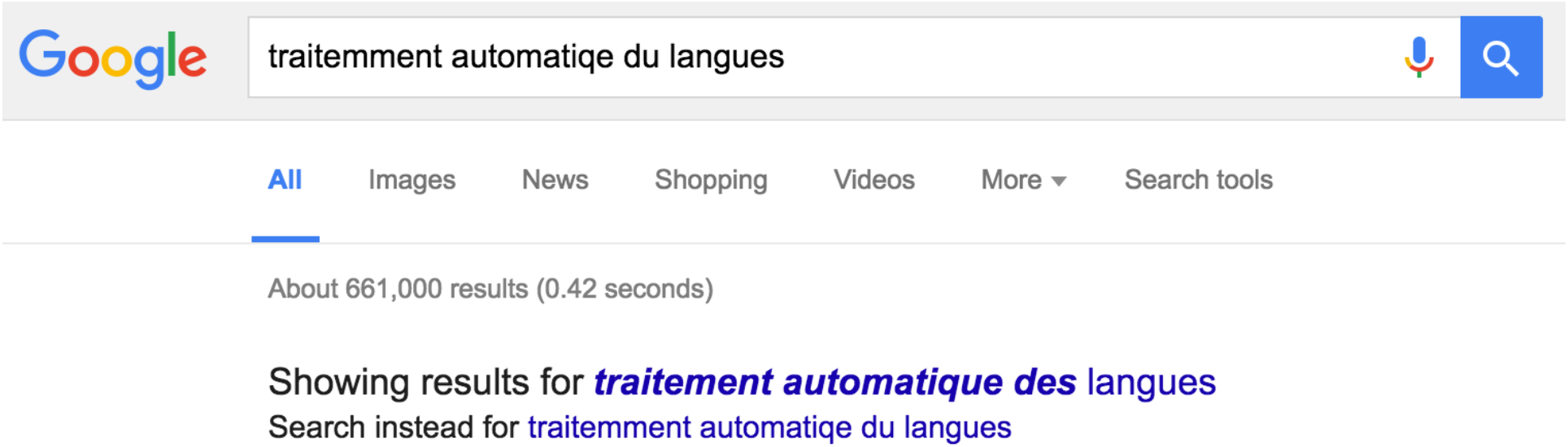
Text Summarization

Produce a short summary of a longer document or article.

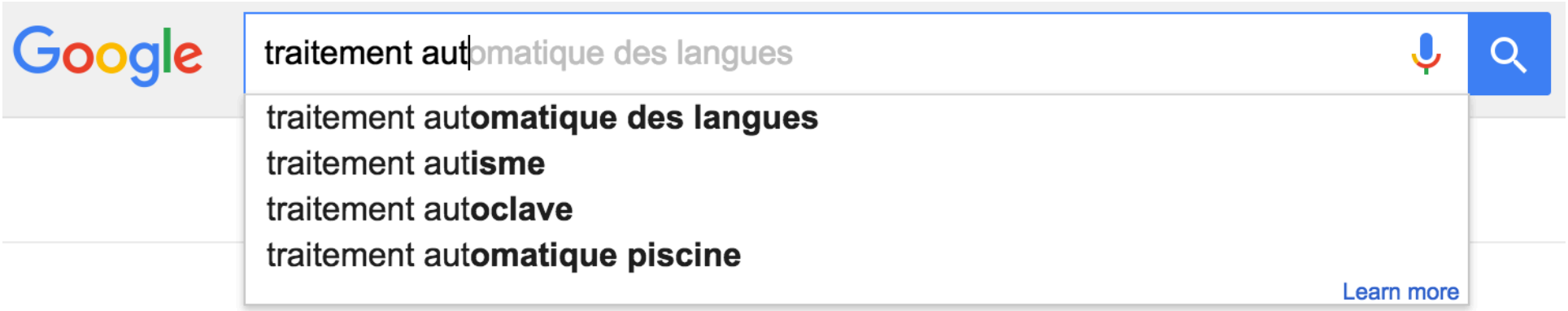
- Article: With a split decision in the final two primaries and a flurry of superdelegate endorsements, Sen. Barack Obama sealed the Democratic presidential nomination last night after a grueling and history-making campaign against Sen. Hillary Rodham Clinton that will make him the first African American to head a major-party ticket. Before a chanting and cheering audience in St. Paul, Minn., the first-term senator from Illinois savored what once seemed an unlikely outcome to the Democratic race with a nod to the marathon that was ending and to what will be another hard-fought battle, against Sen. John McCain, the presumptive Republican nominee….

- Summary: Senator Barack Obama was declared the presumptive Democratic presidential nominee.

Spelling/Grammatical Error Correction



Word Prediction



Information Retrieval

Google panama papers

All News Images Videos Shopping More Search tools

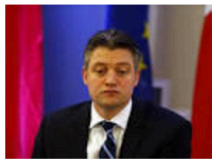
About 40,700,000 results (0.39 seconds)

Panama Papers - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/**Panama_Papers**
The **Panama Papers** are 11.5 million leaked documents that detail financial and attorney–client information for more than 214,488 offshore entities. The leaked ...

The Panama Papers · ICIJ
https://**panamapapers**.icij.org/
Politicians, Criminals and the Rogue Industry That Hides Their Cash.

What are the Panama Papers? A guide to history's biggest data leak ...
www.theguardian.com › News › Tax havens
Apr 3, 2016 - The **Panama Papers** are an unprecedented leak of 11.5m files from the database of the world's fourth biggest offshore law firm, Mossack ...

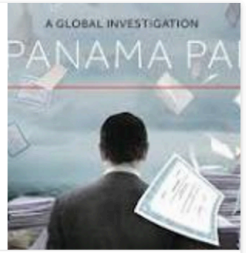
In the news



Konrad Mizzi no-show for EP Panama Papers hearing may lead to political sanctions
Times of Malta - 6 hours ago
Ms Gomes, the deputy chairwoman of the committee tasked by the European Parliament to ...

Taiwan slaps ban, fine on bank linked to Panama Papers
Bangkok Post - 5 hours ago

Panama Papers



The Panama Papers are 11.5 million leaked documents that detail financial and attorney–client information for more than 214,488 offshore entities. Wikipedia

Originally published: April 3, 2016

Authors: Frederik Obermaier, Bastian Obermayer

Feedback

Text Classification

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed

Advanced

Display Settings: Abstract

Send to:

Nature. 2014 Mar 20;507(7492):323-8. doi: 10.1038/nature13145. Epub 2014 Mar 12.

Coupling of angiogenesis and osteogenesis by a specific vessel subtype in bone.

Kusumbe AP[1], Ramasamy SK[1], Adams RH[2].

Author information

Abstract

The mammalian skeletal system harbours a hierarchical system of mesenchymal stem cells, osteoprogenitors and osteoblasts sustaining lifelong bone formation. Osteogenesis is indispensable for the homeostatic renewal of bone as well as regenerative fracture healing, but these processes frequently decline in ageing organisms, leading to loss of bone mass and increased fracture incidence. Evidence indicates that the growth of blood vessels in bone and osteogenesis are coupled, but relatively little is known about the underlying cellular and molecular mechanisms. Here we identify a new capillary subtype in the murine skeletal system with distinct morphological, molecular and functional properties. These vessels are found in specific locations, mediate growth of the bone vasculature, generate distinct metabolic and molecular microenvironments, maintain perivascular osteoprogenitors and couple angiogenesis to osteogenesis. The abundance of these vessels and associated osteoprogenitors was strongly reduced in bone from aged animals, and pharmacological reversal of this decline allowed the restoration of bone mass.

Comment in

Bone biology: Vessels of rejuvenation. [Nature. 2014]

PMID: 24646994 [PubMed - indexed for MEDLINE]

MeSH Terms

Aging/metabolism
Aging/pathology
Animals
Blood Vessels/anatomy & histology
Blood Vessels/cytology
Blood Vessels/growth & development
Blood Vessels/physiology*
Bone and Bones/blood supply*
Bone and Bones/cytology
Endothelial Cells/metabolism
Hypoxia-Inducible Factor 1, alpha Subunit/metabolism
Male
Mice
Mice, Inbred C57BL
Neovascularization, Physiologic/physiology*
Osteoblasts/cytology
Osteoblasts/metabolism
Osteogenesis/physiology*
Oxygen/metabolism
Stem Cells/cytology
Stem Cells/metabolism

uClassify

Classify

Classify method: text url

Enter url to download and classify with:

http://edition.cnn.com/2015/02/18/football/cl

uClassify!

Remove html

1. Sports (92.8 %)
2. Entertainment (4.8 %)
3. Men (0.7 %)

Show all classifications >>

Machine Translation (MT)

Translate a sentence from one natural language to another.

Hasta la vista, bebé $\Rightarrow$

Until we see each other again, baby.

Machine Translation (MT)

MT fails!

MT fails!

MT fails!


1 金桔柠檬
Kumquat with Lemon
金桔汁、柠檬汁、蜜柚果酱、
2 随便
Whatever
综合果汁
3 天然蓝莓花果
Fresh Blueberry Tea
蓝莓果子茶

Sentiment Analysis

Customer Reviews

Speech and Language Processing, 2nd Edition

15 Reviews

5 star: (8)
4 star: (3)
3 star: (3)
2 star: (0)
1 star: (1)

Average Customer Review

★★★★☆ (15 customer reviews)

Share your thoughts with other customers

Create your own review

The most helpful favorable review

4 of 4 people found the following review helpful

★★★★★ **Great introductions and reference book**

I read the first edition of that book and it is terrific. The second edition is much more adapted to current research. Statistical methods in NLP are more detailed and some syntax-based approaches are presented. My specific interest is in machine translation and dialogue systems. Both chapters are extensively rewritten and much more elaborated. I believe this book is...

Read the full review ›

Published on August 9, 2008 by carheg

› See more **5 star**, **4 star** reviews

Vs.

The most helpful critical review

37 of 37 people found the following review helpful

★★★☆☆ **Good description of the problems in the field, but look elsewhere for practical solutions**

The authors have the challenge of covering a vast area, and they do a good job of highlighting the hard problems within individual sub-fields, such as machine translation. The availability of an accompanying Web site is a strong plus, as is the extensive bibliography, which also includes links to freely available software and resources.

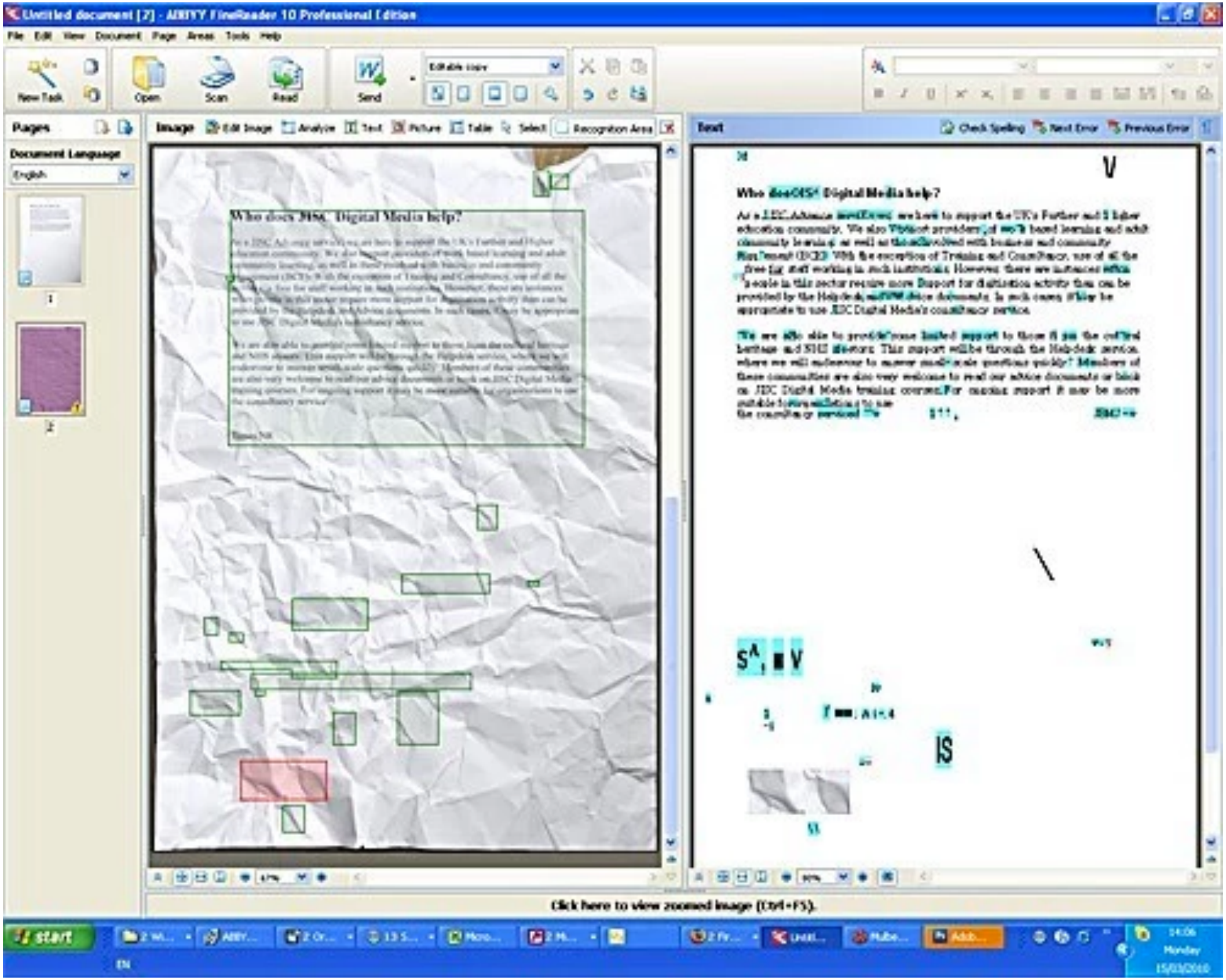
Now for the...

Read the full review ›

Published on April 2, 2009 by P. Nadkarni

› See more **3 star**, 2 star, **1 star** reviews

Optical Character Recognition



Speech Recognition and Dialog Systems

Siri.
Your wish is
its command.

Siri lets you use your voice to send messages, schedule meetings, place phone calls, and more. Ask Siri to do things just by talking the way you talk. Siri understands what you say, knows what you mean, and even talks back. Siri is so easy to use and does so much, you'll keep finding more and more ways to use it.

IBM Watson Developer Cloud

Speech Synthesis

Resolving Ambiguity

Ambiguity Resolution is Required for Translation

Syntactic and semantic ambiguities must be properly resolved for correct translation:

- "John **plays** the guitar." → "John **toca** la guitarra."
- "John **plays** soccer." → "John **juega** el fútbol."

Resolving Ambiguity

Choosing the correct interpretation of linguistic utterances requires knowledge of:

- Syntax
  - An agent is typically the subject of the verb
- Semantics
  - Michael and Ellen are names of people
  - Austin is the name of a city (and of a person)
  - Toyota is a car company and Prius is a brand of car
- Pragmatics
- World knowledge
  - Credit cards require users to pay financial interest
  - Agents must be animate and a hammer is not animate

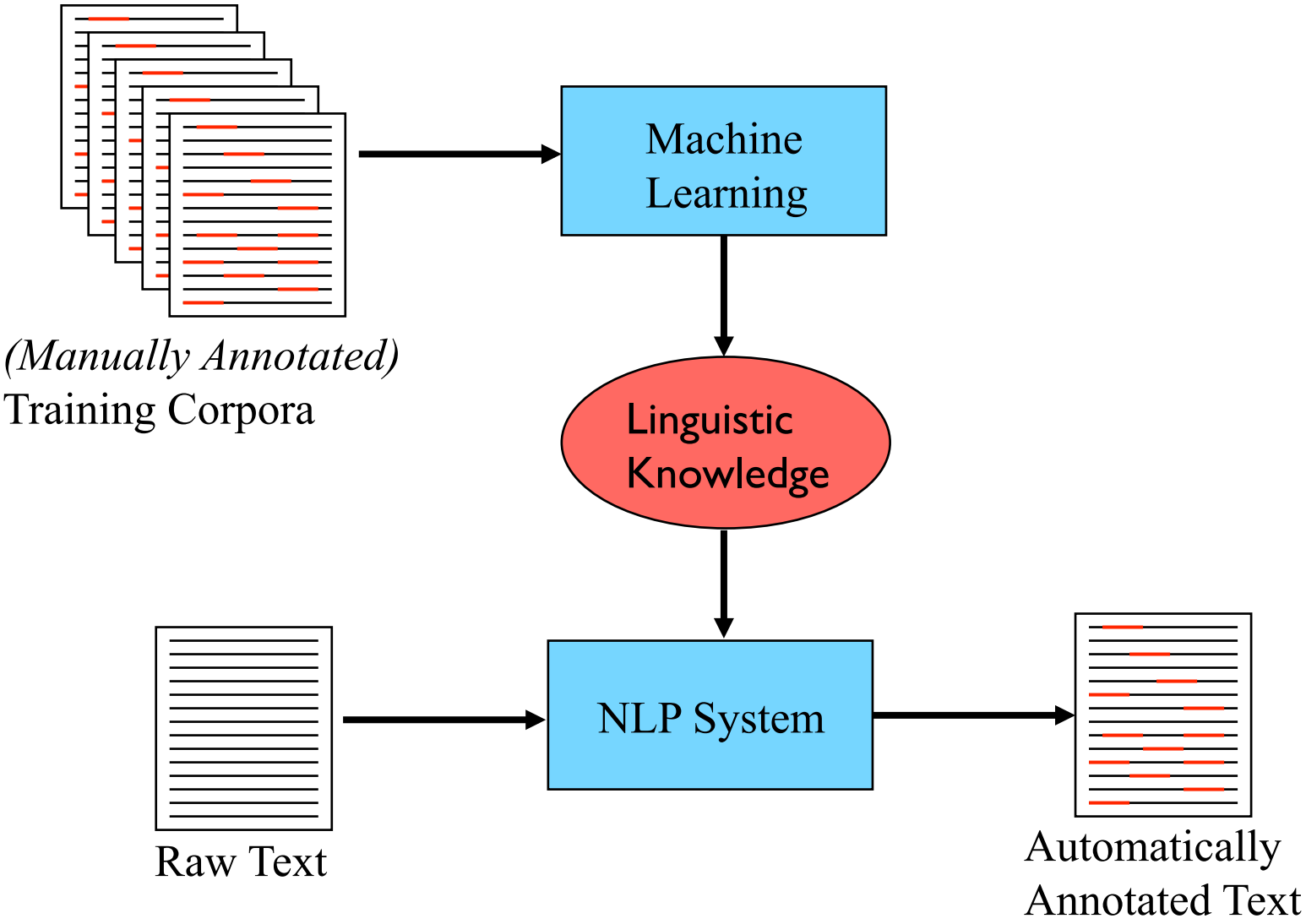
Manual Knowledge Acquisition

- Traditional, "rationalist," approaches to language processing require human specialists to specify and formalize the required knowledge.
- Manual knowledge engineering, is difficult, time-consuming, and error prone.
- "Rules" in language have numerous exceptions and irregularities.
  - "All grammars leak.": Edward Sapir (1921)
- Manually developed systems were expensive to develop and their abilities were limited and "brittle" (not robust).

Automatic Learning Approach

- Use machine learning methods to automatically acquire the required knowledge from appropriately annotated text corpora.
- Variously referred to as the "corpus based," "statistical," or "empirical" approach.
- Statistical learning methods were first applied to speech recognition in the late 1970's and became the dominant approach in the 1980's.
- During the 1990's, the statistical training approach expanded and came to dominate almost all areas of NLP.

Learning Approach

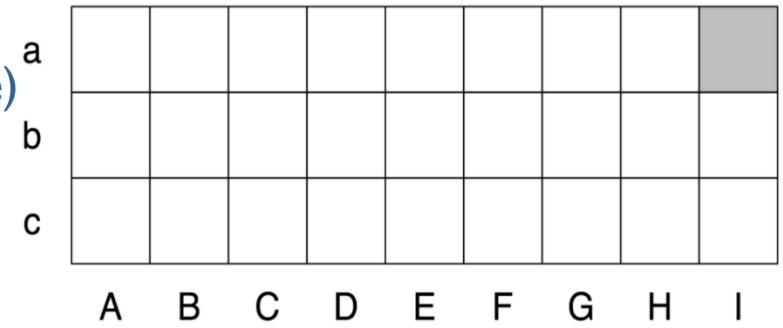


Advantages of the Learning Approach

- Large amounts of electronic text are now available.
- Annotating corpora is easier and requires less expertise than manual knowledge engineering.
- Learning algorithms have progressed to be able to handle large amounts of data and produce accurate probabilistic knowledge.
- The probabilistic knowledge acquired allows robust processing that handles linguistic regularities as well as exceptions.

The Importance of Probability

- Unlikely interpretations of words can combine to generate spurious ambiguity:
  - "The a are of I" is a valid English noun phrase (Abney, 1996)
    - "a" is an adjective for the letter A
    - "are" is a noun for an area of land (hectare)
    - "I" is a noun for the letter I



  - "Time flies like an arrow" has 4 parses, including those meaning:
    - Insects of a variety called "time flies" are fond of a particular arrow.
    - A command to record insects' speed in the manner that an arrow would.

- Some combinations of words are more likely than others:
  - "vice president Gore" vs. "dice precedent core"

- Statistical methods allow computing the most likely interpretation by combining probabilistic evidence from a variety of uncertain knowledge sources.

Limitation

- Most frequent word **tokens** in the English Europarl corpus:

any word

Frequency	Token
1,698,599	the
849,256	of
793,731	to
640,257	and
508,560	in
407,638	that
400,467	is
394,778	a
263,040	I

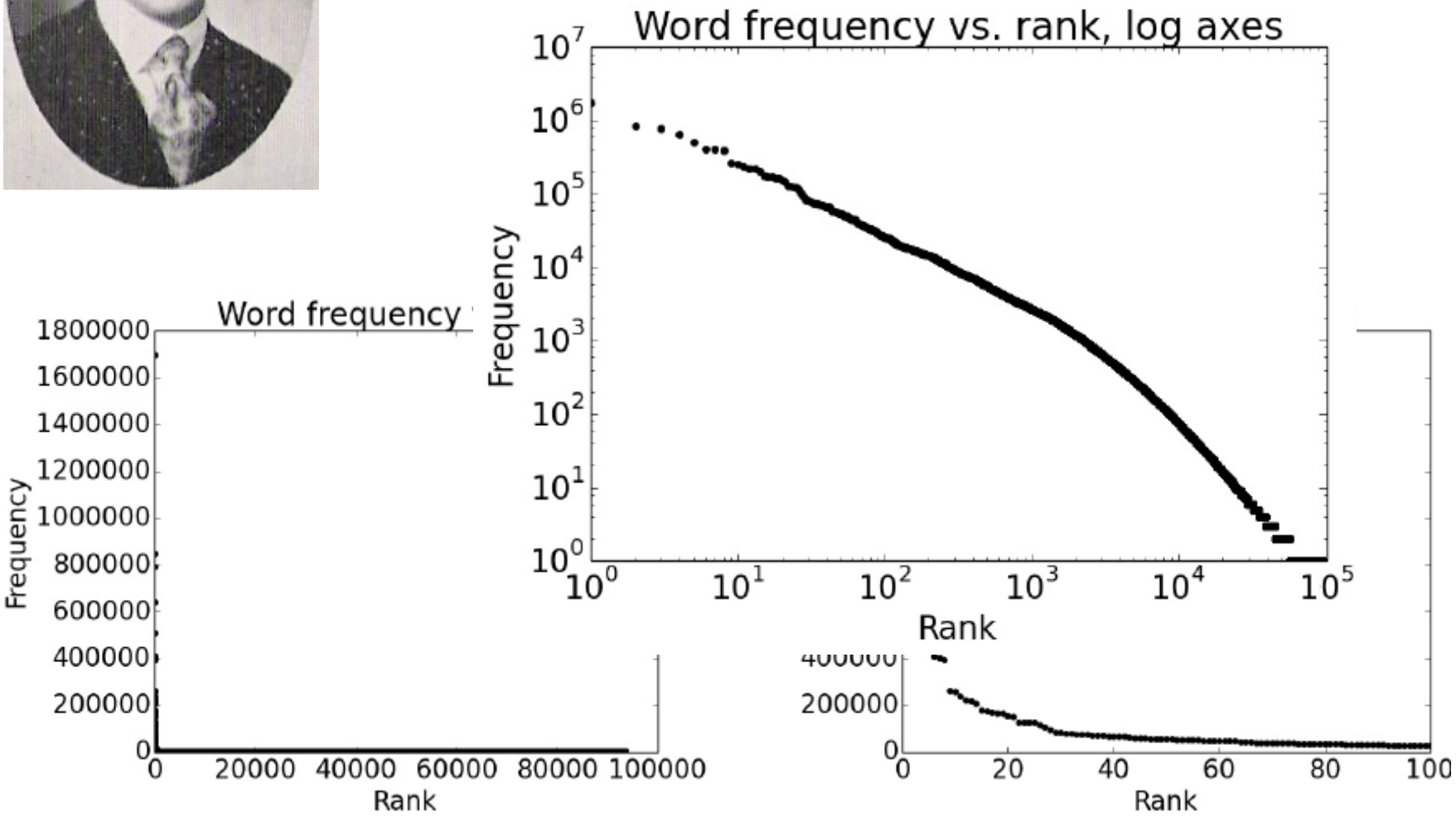
nouns

Frequency	Token
124,598	European
104,325	Mr
92,195	Commission
66,781	President
62,867	Parliament
57,804	Union
53,683	report
53,547	Council
45,842	States

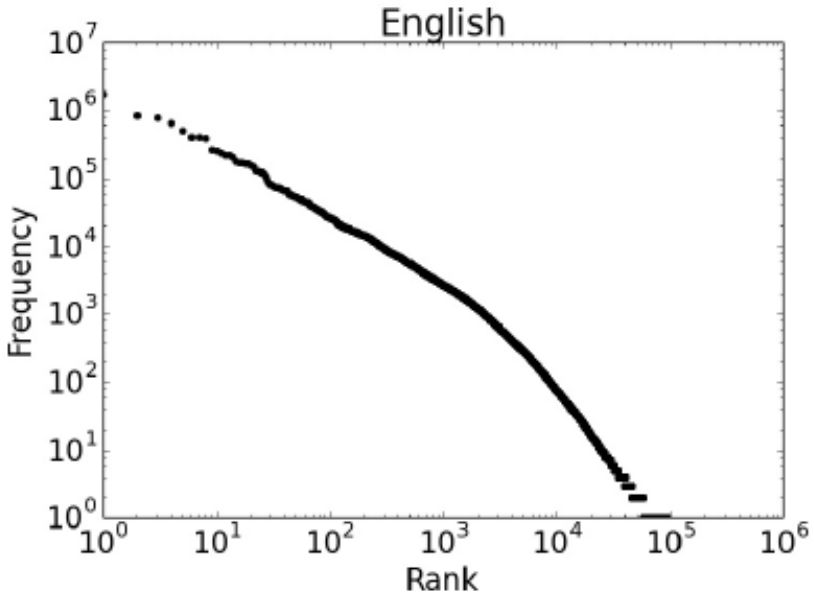
- Out of **93638** distinct words (types) 36231 occur only once
  - cornflakes, mathematicians, fuzziness, jumbling, …

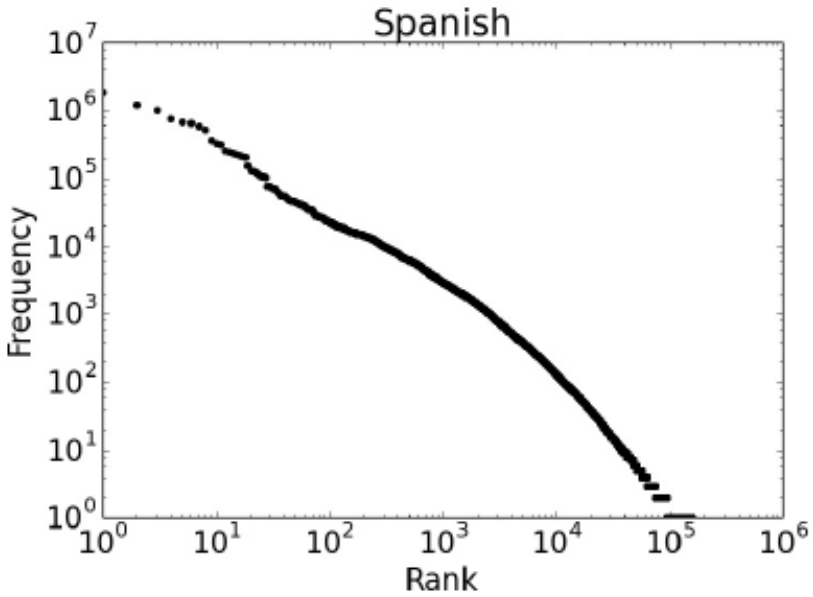
Zipf's law

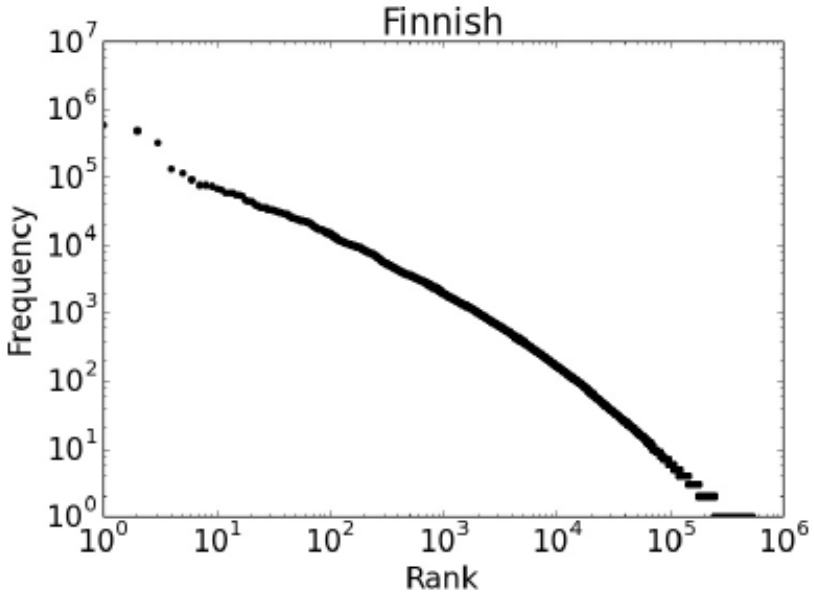
given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table

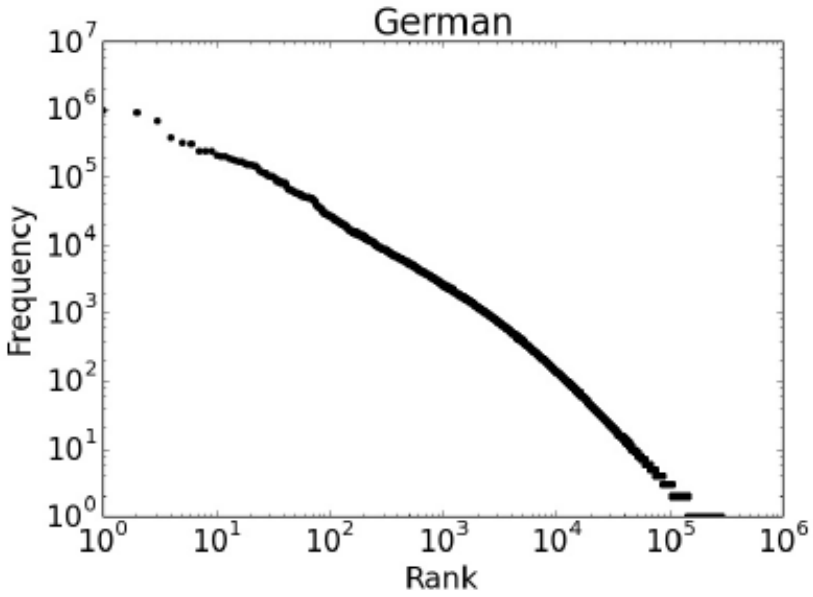


Zipf's law


English
Frequency
Rank
10^0 10^1 10^2 10^3 10^4 10^5 10^6
10^0 10^1 10^2 10^3 10^4 10^5 10^6 10^7


Spanish
Frequency
Rank
10^0 10^1 10^2 10^3 10^4 10^5 10^6
10^0 10^1 10^2 10^3 10^4 10^5 10^6 10^7


Finnish
Frequency
Rank
10^0 10^1 10^2 10^3 10^4 10^5 10^6
10^0 10^1 10^2 10^3 10^4 10^5 10^6 10^7


German
Frequency
Rank
10^0 10^1 10^2 10^3 10^4 10^5 10^6
10^0 10^1 10^2 10^3 10^4 10^5 10^6 10^7

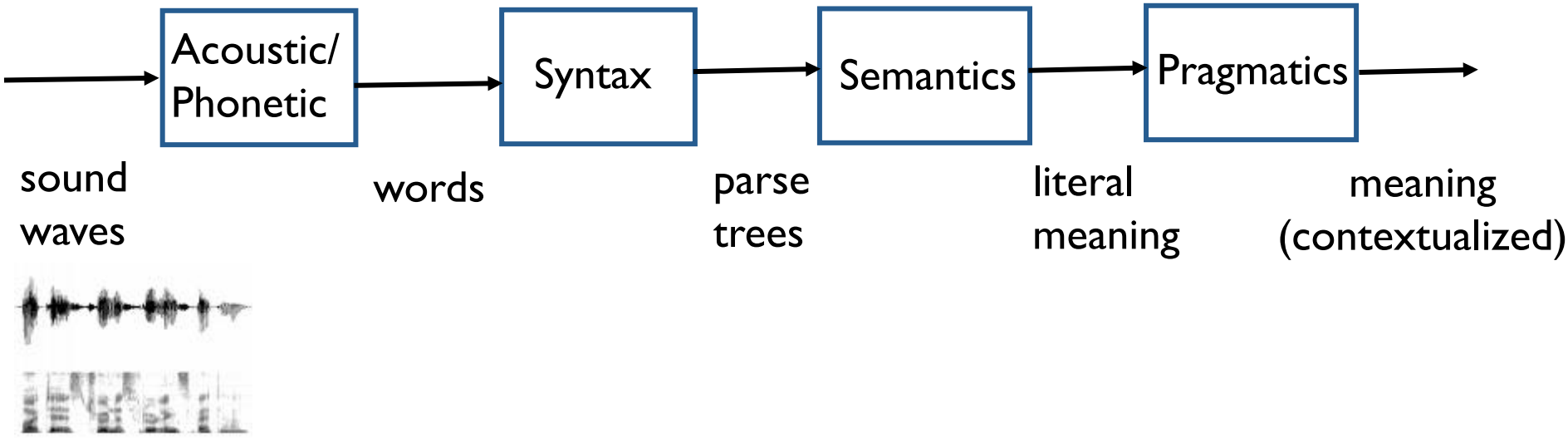
Linguistics and Data

- **Data**
  - looking at real use of language in text
  - can learn a lot from empirical evidence
  - **But:**
    - Zipf's law: there will be always instances that are rarely seen

- **Linguistics**
  - build a better understanding of language structure
  - linguistic analysis points to what is important
  - **But:**
    - many ambiguities cannot be explained easily

The NLP Pipeline

Modular Comprehension



Pipelining Problem

- Assuming separate independent components for speech recognition, syntax, semantics, pragmatics, etc. allows for more convenient modular software development.

- However, frequently constraints from "higher level" processes are needed to disambiguate "lower level" processes.

  - Example of syntactic disambiguation relying on semantic disambiguation:

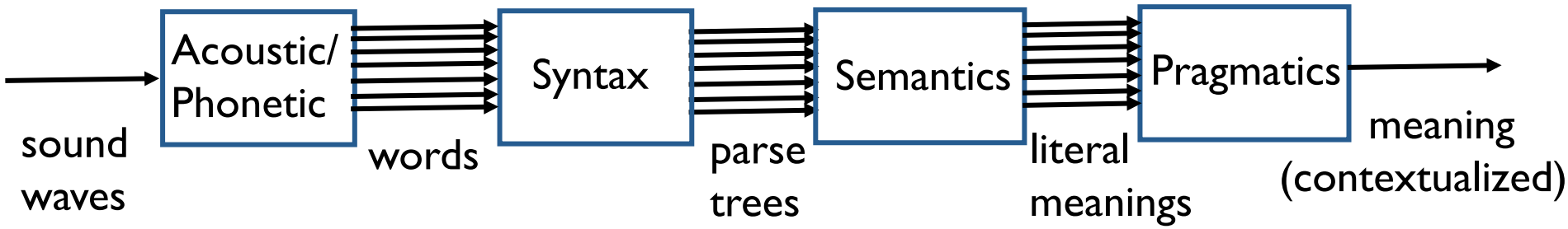
    - At the zoo, several men were showing a group of students various types of flying animals. Suddenly, one of the students hit the man **with** a **bat**.

Pipelining Problem

- If a hard decision is made at each stage, cannot backtrack when a later stage indicates it is incorrect.
  - If attach "with a bat" to the verb "hit" during syntactic analysis, then cannot reattach it to "man" after "bat" is disambiguated during later semantic or pragmatic processing.

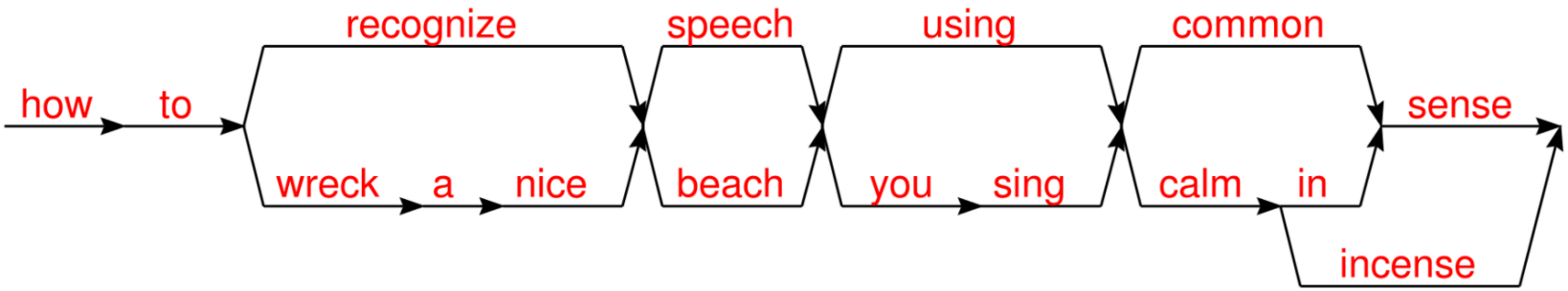
Increasing Module Bandwidth

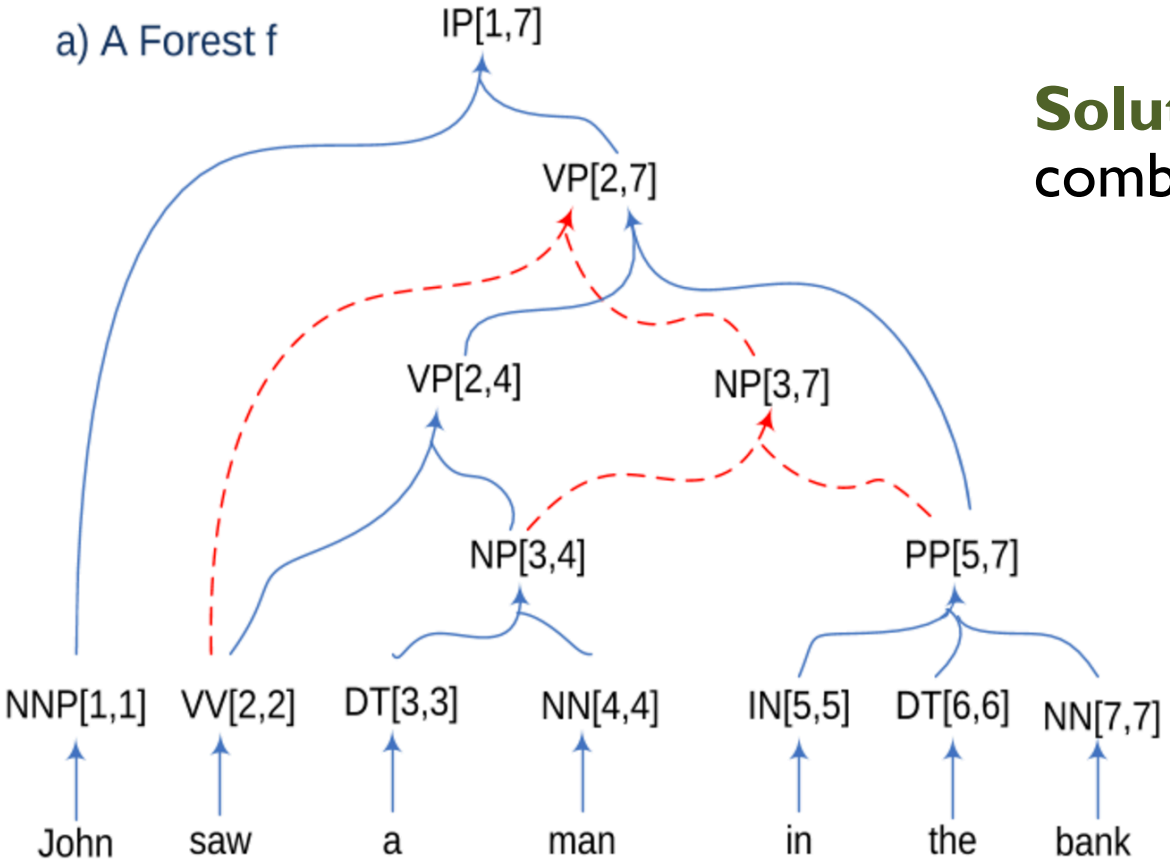
- If each component produces multiple scored interpretations, then later components can rerank these interpretations.



Problem: Number of interpretations grows combinatorially.

Increasing Module Bandwidth





Solution: Efficiently encode combinations of interpretations.

- Word lattices
- Compact parse forests
- Etc.