Introduction to NLP

151

NLP Tasks

Part of Speech Tagging

The swimmer is getting ready to run in the final race.

Part of Speech Tagging

`The swimmer is getting ready to` **`run`** `in the final race.`

- Run – verb or noun?
- Final – noun or adjective?
- Race – verb or noun?

Part of Speech Tagging

The candidate is preparing for his **run** for the presidency.
The swimmer is getting ready to **run** in the final race.

Parsing

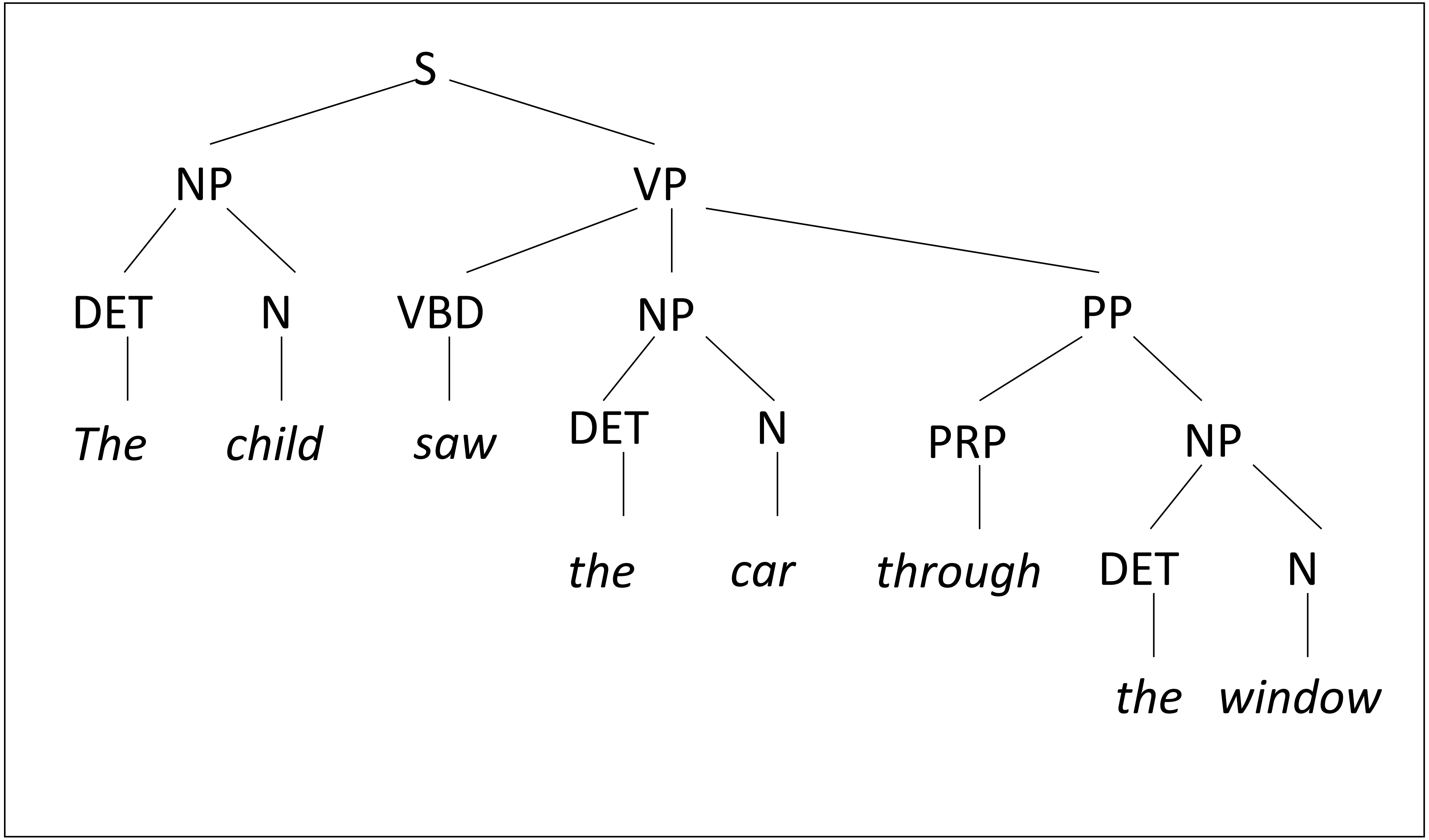
- Myriam slept.
- Myriam wrote a novel.
- Myriam gave Sally flowers.
- Myriam ate salad with a fork.

Phrase-Structure Grammar

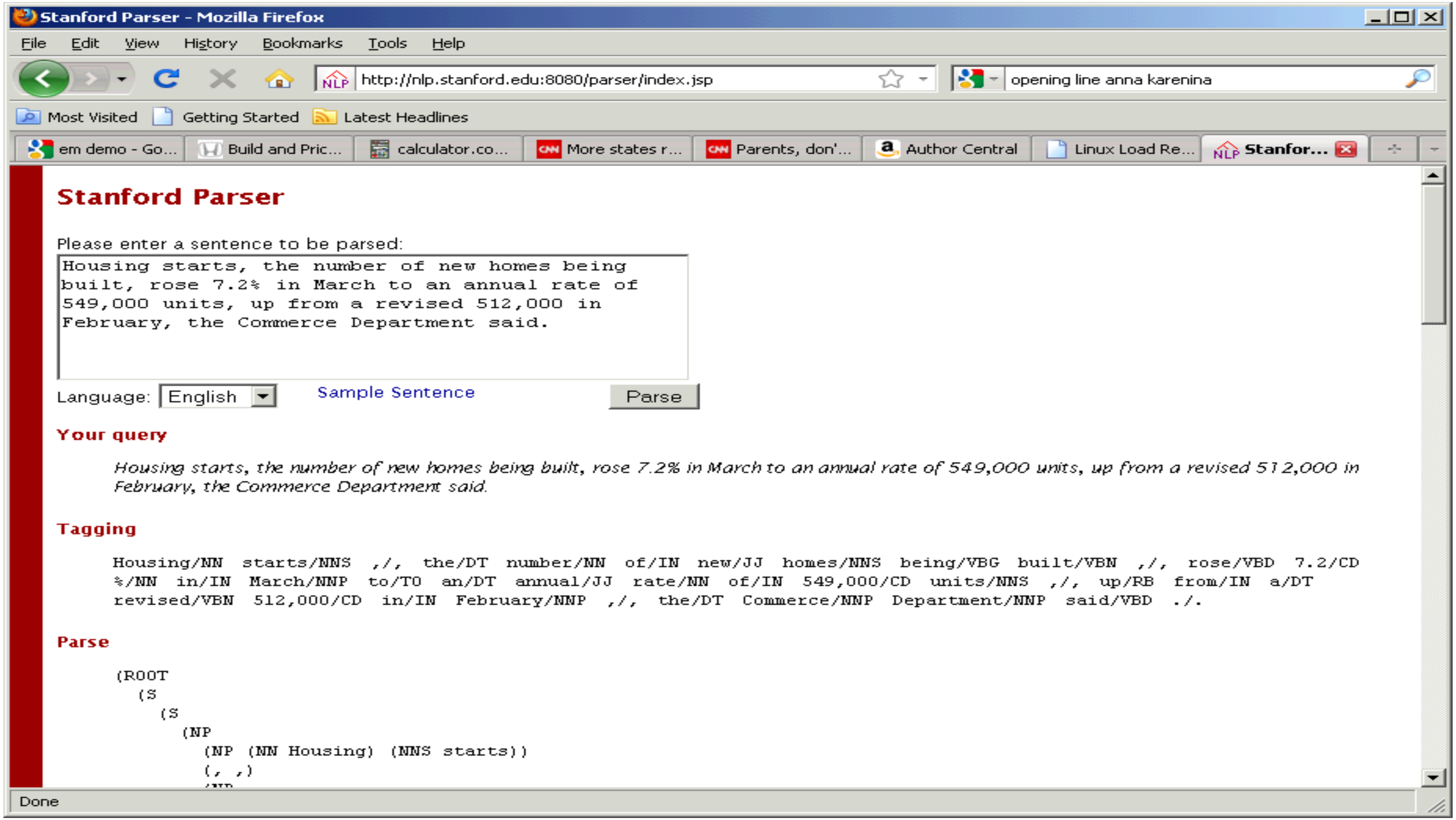
```
S  → NP  VP
NP → DET N
NP → NP  PP
VP → VBD
VP → VBD NP
VP → VBD NP NP
VP → VP  PP
PP → PRP NP
```

```
DET → the
DET → that
DET → a
N   → child
N   → window
N   → car
VBD → found
VBD → ate
VBD → saw
PRP → in
PRP → of
PRP → through
```

Parse Trees


S
NP
VP
DET
N
VBD
NP
PP
The
child
saw
DET
N
PRP
NP
the
car
through
DET
N
the
window

Stanford Parser



Stanford Parser

Please enter a sentence to be parsed:

```
Housing starts, the number of new homes being
built, rose 7.2% in March to an annual rate of
549,000 units, up from a revised 512,000 in
February, the Commerce Department said.
```

Language: English ▾ Sample Sentence Parse

Your query

Housing starts, the number of new homes being built, rose 7.2% in March to an annual rate of 549,000 units, up from a revised 512,000 in February, the Commerce Department said.

Tagging

```
Housing/NN starts/NNS ,/, the/DT number/NN of/IN new/JJ homes/NNS being/VBG built/VBN ,/, rose/VBD 7.2/CD
%/NN in/IN March/NNP to/TO an/DT annual/JJ rate/NN of/IN 549,000/CD units/NNS ,/, up/RB from/IN a/DT
revised/VBN 512,000/CD in/IN February/NNP ,/, the/DT Commerce/NNP Department/NNP said/VBD ./.
```

Parse

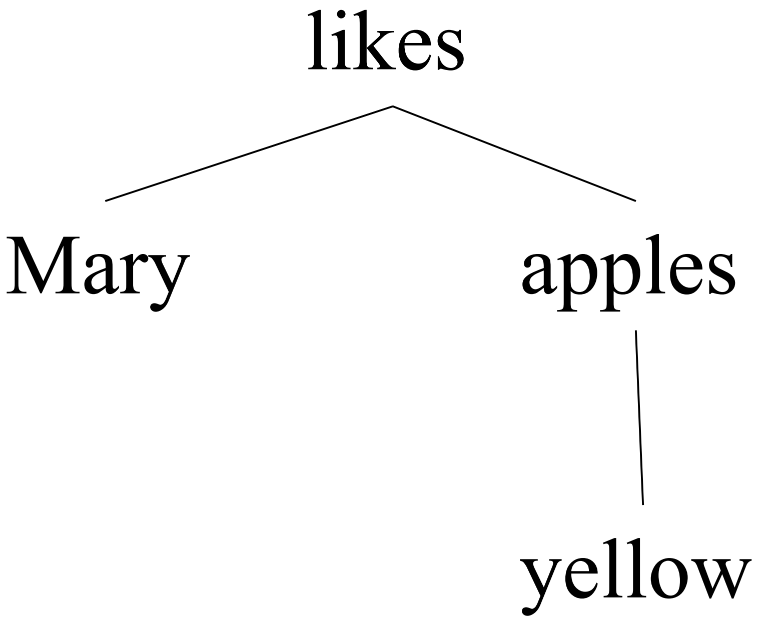
```
(ROOT
  (S
    (S
      (NP
        (NP (NN Housing) (NNS starts))
        (, ,)
```

Parser Output

```
(ROOT
  (S
    (S
      (NP
        (NP (NN Housing) (NNS starts))
        (, ,)
        (NP
          (NP (DT the) (NN number))
          (PP (IN of)
            (NP
              (NP (JJ new) (NNS homes))
              (VP (VBG being)
                (VP (VBN built))))))
        (, ,))
```

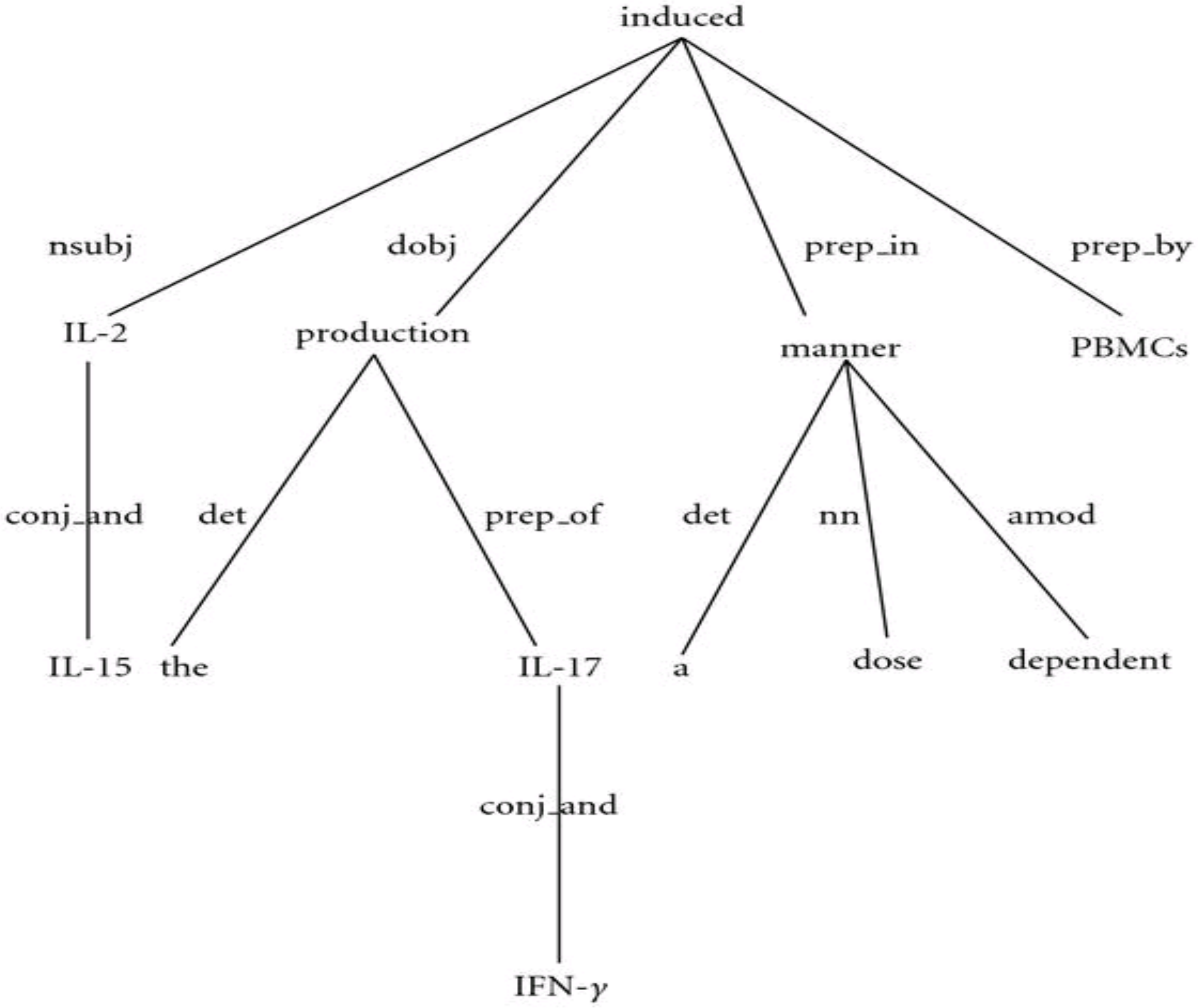
```
(VP (VBD rose)
       (NP (CD 7.2) (NN %))
       (PP (IN in)
         (NP (NNP March)))
       (PP (TO to)
         (NP
           (NP (DT an) (JJ annual) (NN rate))
           (PP (IN of)
             (NP (CD 549,000) (NNS units)))))
       (, ,)
       (ADVP (RB up)
         (PP (IN from)
           (NP
             (NP (DT a) (VBN revised) (CD 512,000))
             (PP (IN in)
               (NP (NNP February)))))))
   (, ,)
   (NP (DT the) (NNP Commerce) (NNP Department))
   (VP (VBD said))
   (. .)))
```

Dependency Parsing



Dependency Parsing

IL-2 and IL-15 induced the production of IL-17 and IFN-γ by PBMCs in a dose dependent manner.



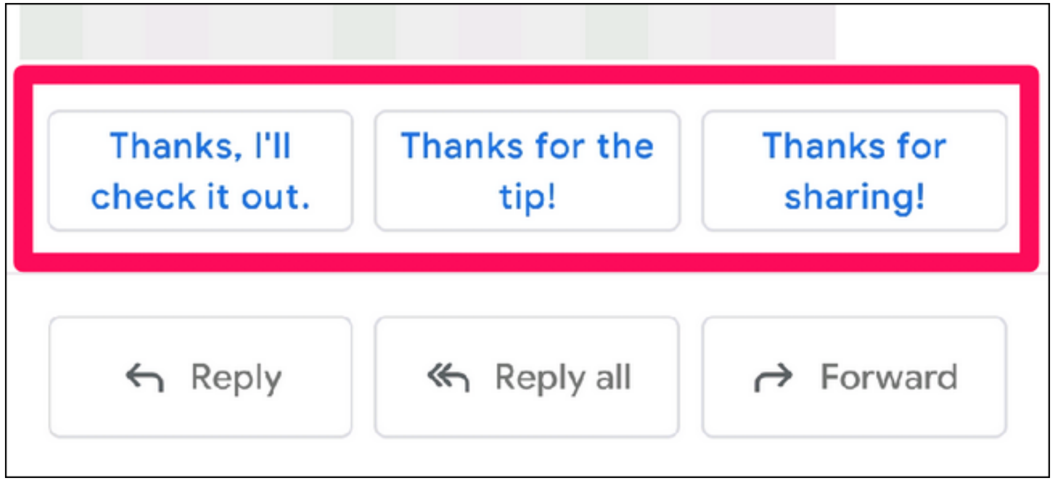
Information Extraction

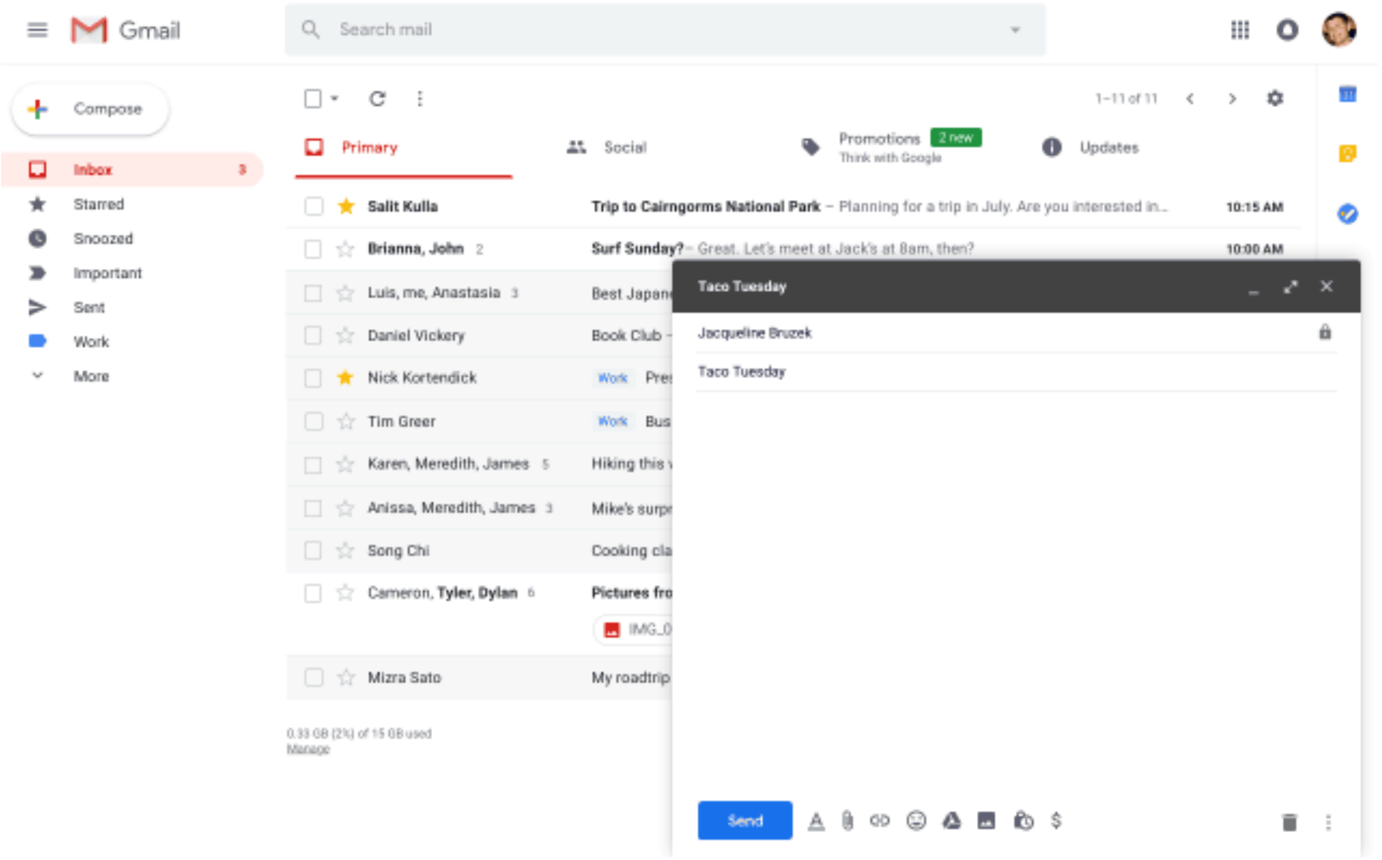
- RESEARCH ALERT-Wells Fargo cuts PPD Inc to market perform
- China Southern Air Upgraded To Overweight From Neutral-HSBC
- CITIGROUP RAISES INGERSOLL RAND <IR.N> TO HOLD FROM SELL
- TCF Financial Corp Raised To Overweight From Neutral By JPMorgan
- BAIRD CUTS KIOR INC <KIOR.O> TO UNDERPERFORM RATING
- BRIEF-RESEARCH ALERT-Global Equities Research cuts LinkedIn to equal weight

Information Extraction

DATE/TIME	TICKER	COMPANY	SOURCE	OLD	NEW	CHANGE
		PPD Inc	Wells Fargo		market perform	↓
		China Southern Air	HSBC	Neutral	Overweight	↑
	IR.N	INGERSOLL RAND	CITIGROUP	SELL	HOLD	↑
		TCF Financial Corp	JPMorgan	Neutral	Overweight	↑
	KIOR.O	KIOR INC	BAIRD		UNDERPERFORM	↓
		LinkedIn	Global Equities Research		equal weight	↓

Text Completion





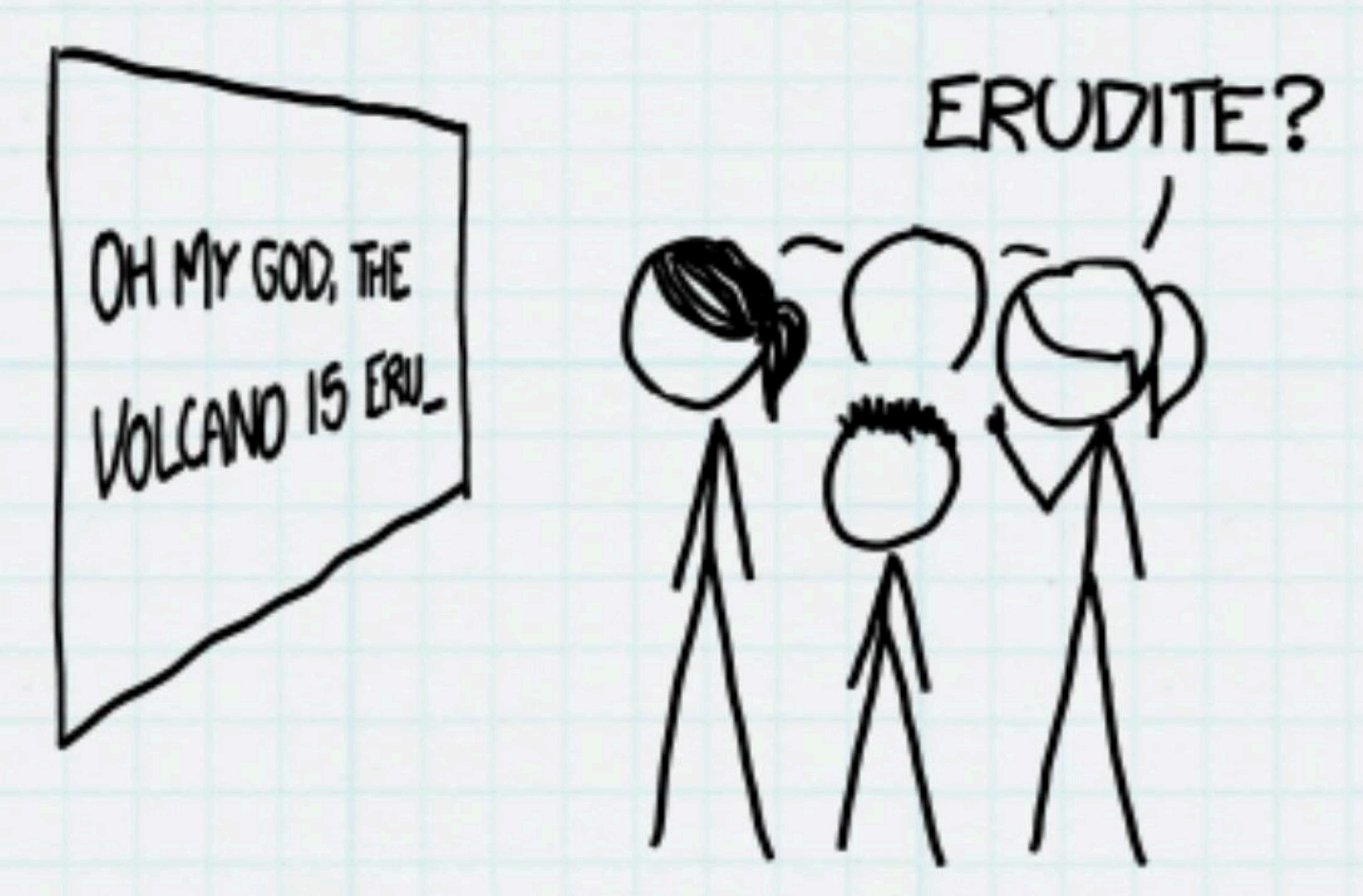
★	Salit Kulla	Trip to Cairngorms National Park – Planning for a trip in July. Are you interested in…	10:15 AM
☆	Brianna, John 2	Surf Sunday? – Great. Let's meet at Jack's at 8am, then?	10:00 AM
☆	Luis, me, Anastasia 3	Best Japan	
☆	Daniel Vickery	Book Club –	
★	Nick Kortendick	Work Pres	
☆	Tim Greer	Work Bus	
☆	Karen, Meredith, James 5	Hiking this	
☆	Anissa, Meredith, James 3	Mike's surp	
☆	Song Chi	Cooking cla	
☆	Cameron, Tyler, Dylan 6	Pictures fro IMG_0	
☆	Mizra Sato	My roadtrip	

0.33 GB (2%) of 15 GB used
Manage

Taco Tuesday

Jacqueline Bruzek

Taco Tuesday



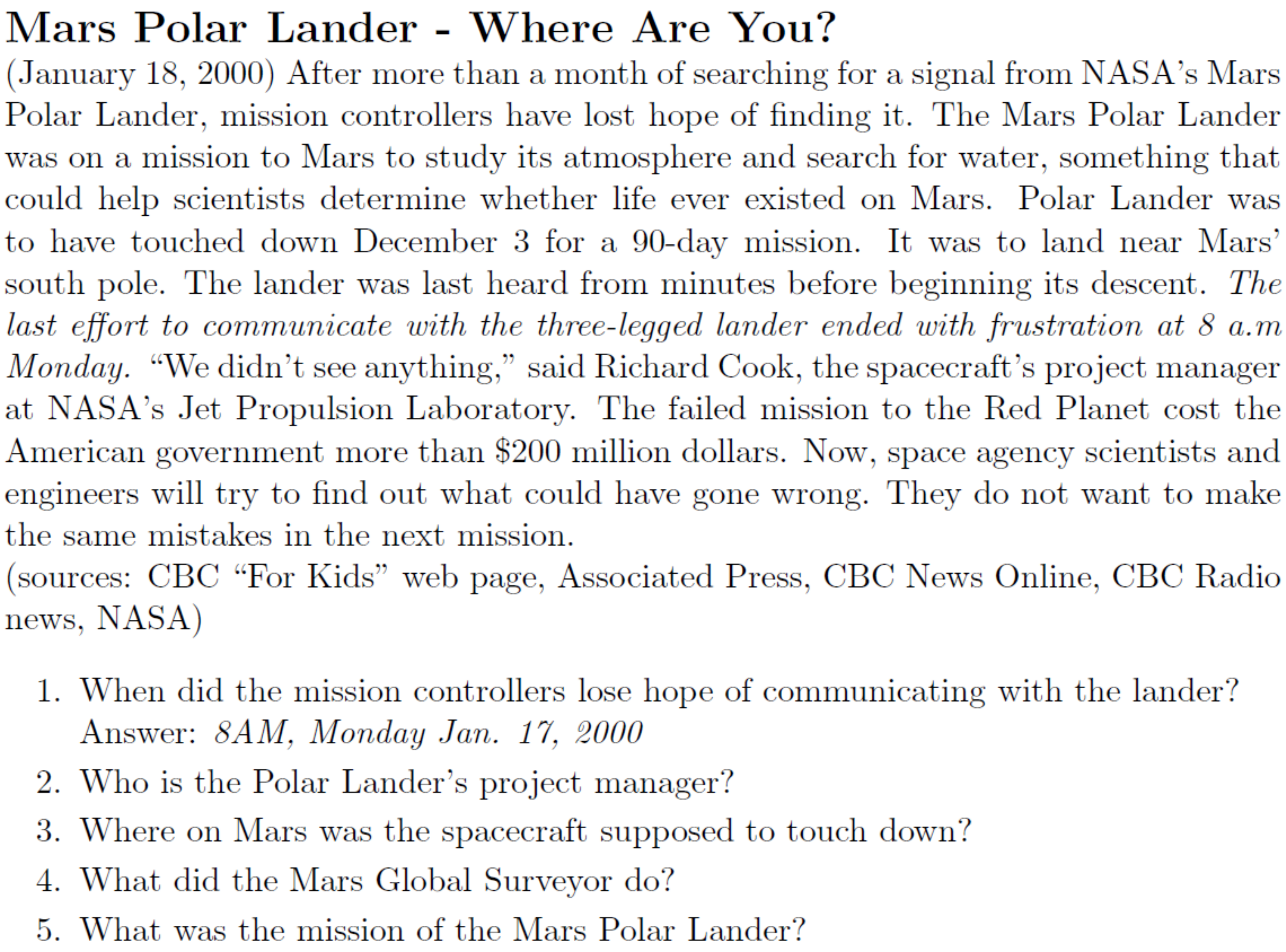
https://what-if.xkcd.com/34/

Semantics

- First order logic
- Inference/deduction
- Semantic analysis

$$\forall \text{x,y: } \textit{Mother (x,y)} \Rightarrow \textit{Parent (x,y)}$$

Reading Comprehension



Mars Polar Lander - Where Are You?

(January 18, 2000) After more than a month of searching for a signal from NASA's Mars Polar Lander, mission controllers have lost hope of finding it. The Mars Polar Lander was on a mission to Mars to study its atmosphere and search for water, something that could help scientists determine whether life ever existed on Mars. Polar Lander was to have touched down December 3 for a 90-day mission. It was to land near Mars' south pole. The lander was last heard from minutes before beginning its descent. *The last effort to communicate with the three-legged lander ended with frustration at 8 a.m Monday.* "We didn't see anything," said Richard Cook, the spacecraft's project manager at NASA's Jet Propulsion Laboratory. The failed mission to the Red Planet cost the American government more than $200 million dollars. Now, space agency scientists and engineers will try to find out what could have gone wrong. They do not want to make the same mistakes in the next mission.
(sources: CBC "For Kids" web page, Associated Press, CBC News Online, CBC Radio news, NASA)

1. When did the mission controllers lose hope of communicating with the lander?
   Answer: *8AM, Monday Jan. 17, 2000*
2. Who is the Polar Lander's project manager?
3. Where on Mars was the spacecraft supposed to touch down?
4. What did the Mars Global Surveyor do?
5. What was the mission of the Mars Polar Lander?

Pranav Anand, Eric Breck, Brianne Brown, Marc Light, Gideon Mann, Ellen Riloff, Mats Rooth, Michael Thelen. 2000. Fun with Reading Comprehension

Word Sense Disambiguation

- “The thieves took off with 100 gold **bars**”.
  - Did they steal 100 drinking establishments?
  - Or 100 measures of a song?

WSD is Important for Translation

- Paul plays soccer
  - Paul joue au football
- Paul plays the guitar
  - Paul joue de la guitare
- “wall” in German
  - die Chinesische Mauer (The Great Wall of China)
  - (otherwise Wand)
- “wall” in Spanish
  - pared, muro, muralla

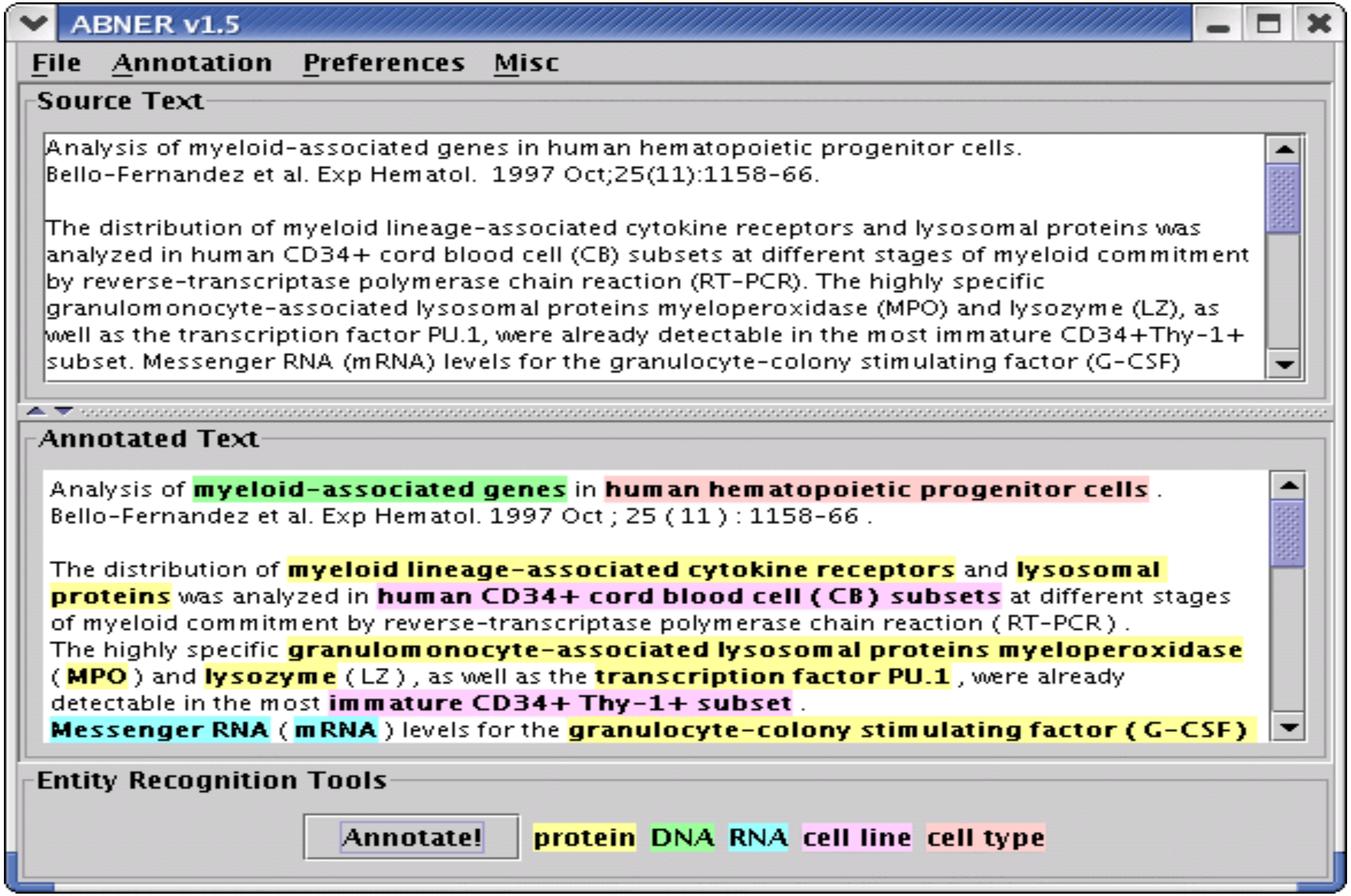
Named Entity Recognition

Wolff, currently a journalist in Argentina, played with Del Bosque in the final years of the seventies in Real Madrid.

```
    Wolff B-PER
        , O
currently O
        a O
journalist O
       in O
Argentina B-LOC
        , O
   played O
     with O
      Del B-PER
   Bosque I-PER
       in O
      the O
    final O
    years O
       of O
      the O
seventies O
       in O
     Real B-ORG
   Madrid I-ORG
        . O
```

- http://cogcomp.cs.illinois.edu/page/demo_view/NER
- http://nlp.stanford.edu:8080/ner/

Named Entity Recognition



Analysis of myeloid-associated genes in human hematopoietic progenitor cells.
Bello-Fernandez et al. Exp Hematol. 1997 Oct;25(11):1158-66.

The distribution of myeloid lineage-associated cytokine receptors and lysosomal proteins was analyzed in human CD34+ cord blood cell (CB) subsets at different stages of myeloid commitment by reverse-transcriptase polymerase chain reaction (RT-PCR). The highly specific granulomonocyte-associated lysosomal proteins myeloperoxidase (MPO) and lysozyme (LZ), as well as the transcription factor PU.1, were already detectable in the most immature CD34+Thy-1+ subset. Messenger RNA (mRNA) levels for the granulocyte-colony stimulating factor (G-CSF)

Analysis of **myeloid-associated genes** in **human hematopoietic progenitor cells** .
Bello-Fernandez et al. Exp Hematol. 1997 Oct ; 25 (11) : 1158-66 .

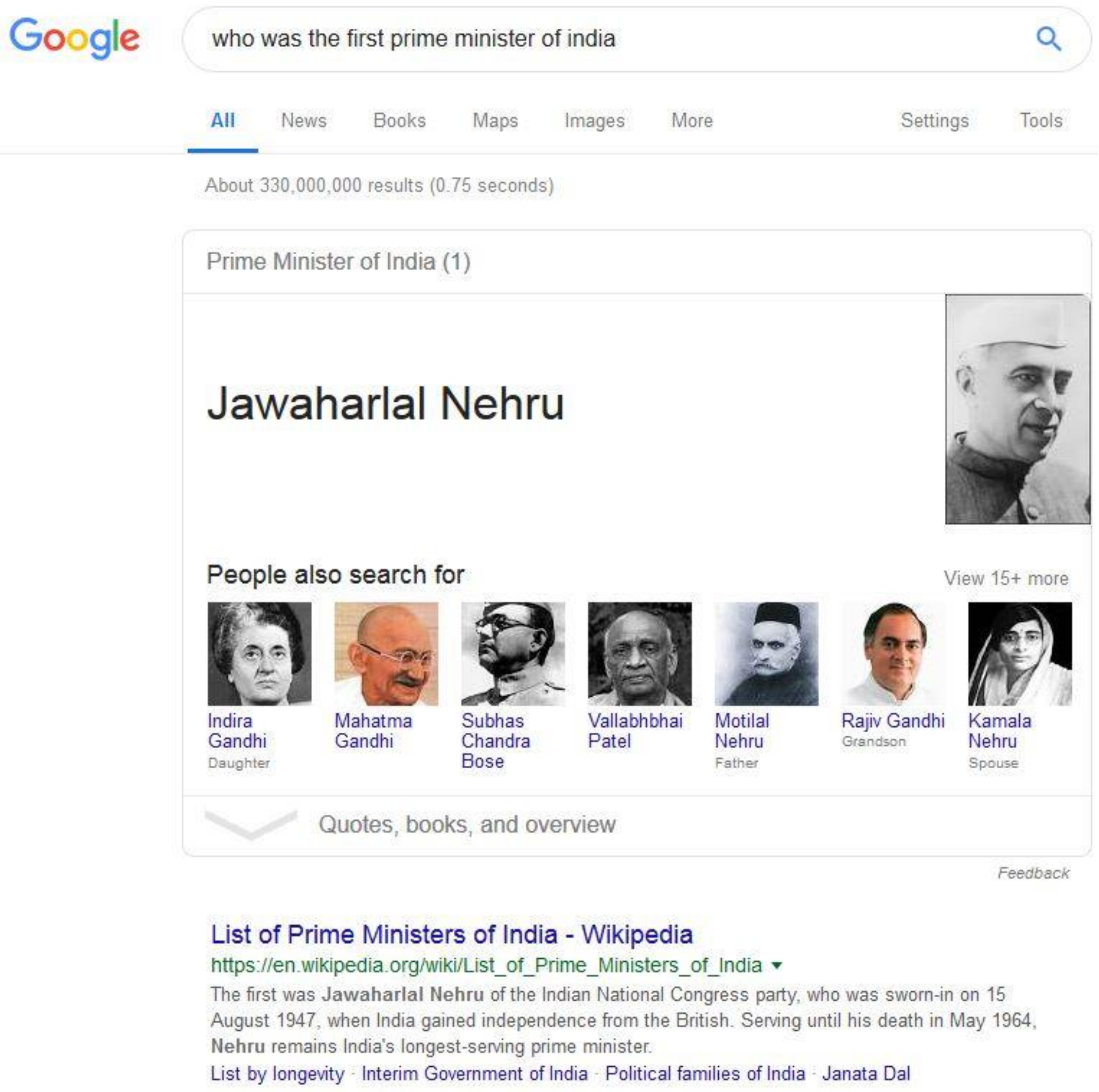
The distribution of **myeloid lineage-associated cytokine receptors** and **lysosomal proteins** was analyzed in **human CD34+ cord blood cell (CB) subsets** at different stages of myeloid commitment by reverse-transcriptase polymerase chain reaction (RT-PCR) .
The highly specific **granulomonocyte-associated lysosomal proteins myeloperoxidase** (**MPO**) and **lysozyme** (LZ) , as well as the **transcription factor PU.1** , were already detectable in the most **immature CD34+ Thy-1+ subset** .
Messenger RNA (**mRNA**) levels for the **granulocyte-colony stimulating factor (G-CSF)**

http://pages.cs.wisc.edu/~bsettles/**abner**

Coreference Resolution

- Barack Obama visited China. The US president met with his Chinese counterpart.
- Cynthia went to see her aunt at the hospital. She was scheduled for surgery on Monday.
- Because he was sick, Michael stayed home on Friday.

Question Answering



who was the first prime minister of india

About 330,000,000 results (0.75 seconds)

Prime Minister of India (1)

Jawaharlal Nehru

People also search for

Indira Gandhi
Daughter

Mahatma Gandhi

Subhas Chandra Bose

Vallabhbhai Patel

Motilal Nehru
Father

Rajiv Gandhi
Grandson

Kamala Nehru
Spouse

Quotes, books, and overview

List of Prime Ministers of India - Wikipedia

https://en.wikipedia.org/wiki/List_of_Prime_Ministers_of_India ▾

The first was **Jawaharlal Nehru** of the Indian National Congress party, who was sworn-in on 15 August 1947, when India gained independence from the British. Serving until his death in May 1964, **Nehru** remains India's longest-serving prime minister.

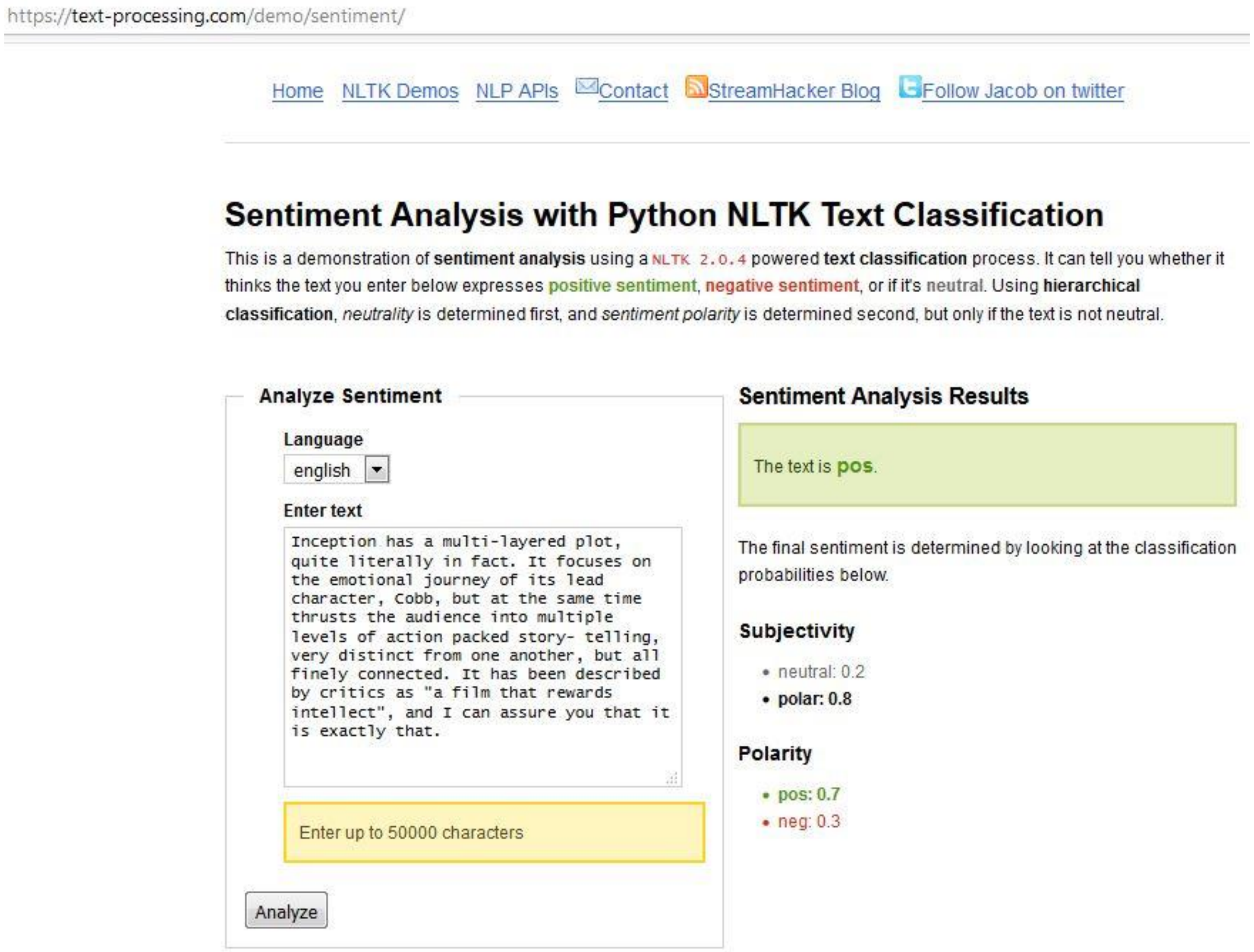
List by longevity · Interim Government of India · Political families of India · Janata Dal

Sentiment Analysis

“I like the camera because I can edit images so easily, exactly as I do my iPad. I have found that its difficult to frame a picture when there isn't a zoom function as with the iPad. With this camera I can adjust my images by cropping as I did with my iPad but better yet, this camera has a built in zoom. A stretch or pinch of the fingers bring in the subject closer or back out again. With this iPhone I can also, as I dido with my iPad, enhance, crop, rotate, red eye reduce, and set a range of tints. I am also quite impressed with the quality of the images. Pretty darn good especially better than I expected for low light situations where I can use the built-in flash! Quite frankly I was quite surprised with these built in features. I also hope too experiment with and learn what HDR photography is. It's built into this iPhone and can be activated by a the touch of an icon. ”

http://www.epinions.com/review/apple_iphone_5c_latest_model_16gb_graphite_unlocked_smartphone/content_640679317124

Sentiment Analysis



https://text-processing.com/demo/sentiment/

Home NLTK Demos NLP APIs Contact StreamHacker Blog Follow Jacob on twitter

Sentiment Analysis with Python NLTK Text Classification

This is a demonstration of **sentiment analysis** using a `NLTK 2.0.4` powered **text classification** process. It can tell you whether it thinks the text you enter below expresses **positive sentiment**, **negative sentiment**, or if it's neutral. Using **hierarchical classification**, *neutrality* is determined first, and *sentiment polarity* is determined second, but only if the text is not neutral.

Analyze Sentiment

Language

english

Enter text

```
Inception has a multi-layered plot,
quite literally in fact. It focuses on
the emotional journey of its lead
character, Cobb, but at the same time
thrusts the audience into multiple
levels of action packed story- telling,
very distinct from one another, but all
finely connected. It has been described
by critics as "a film that rewards
intellect", and I can assure you that it
is exactly that.
```

Enter up to 50000 characters

Sentiment Analysis Results

The text is **pos**.

The final sentiment is determined by looking at the classification probabilities below.

Subjectivity

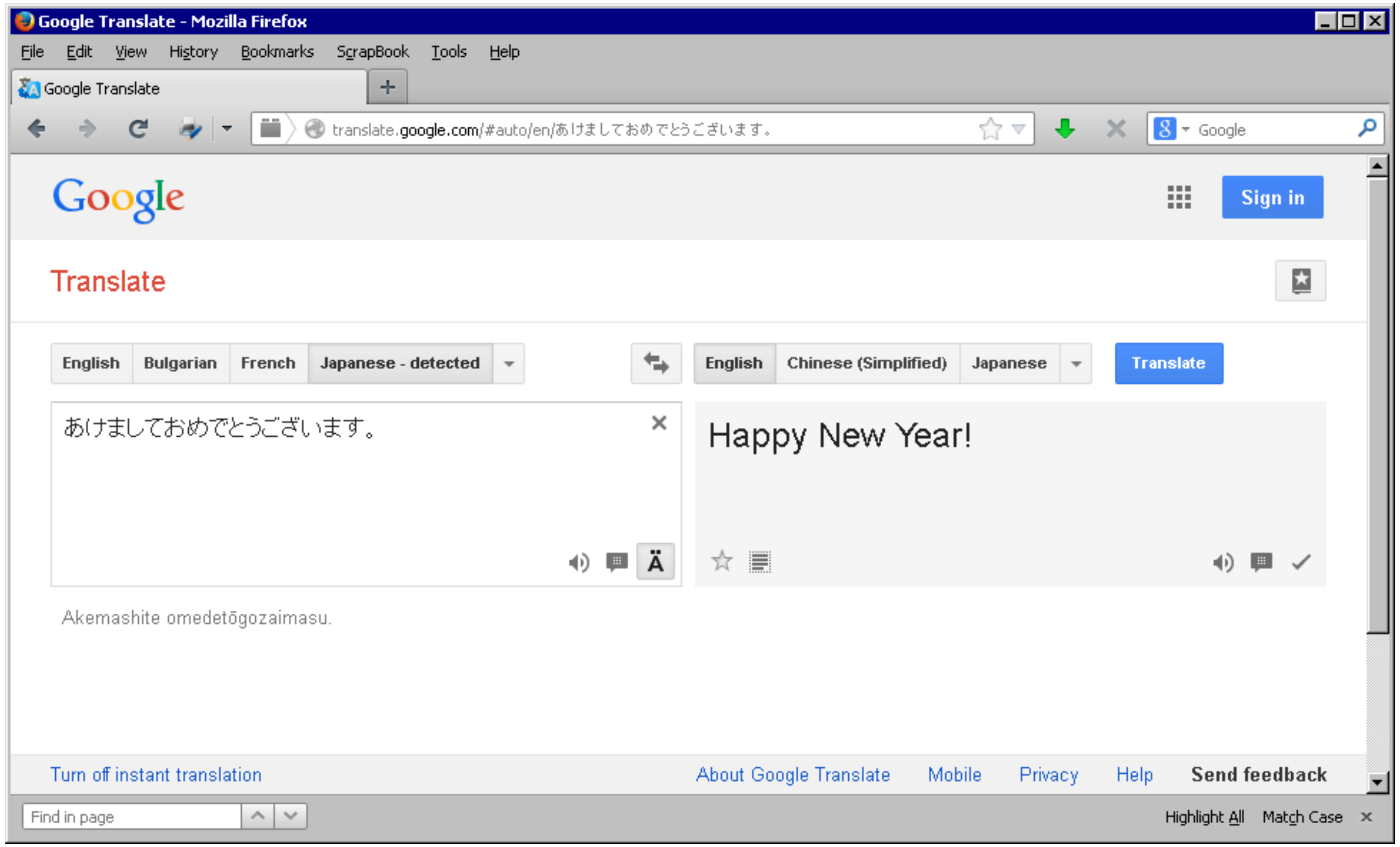
- neutral: 0.2
- **polar: 0.8**

Polarity

- pos: 0.7
- neg: 0.3

Machine Translation

- あけましておめでとうございます。
- Happy New Year!


Google Translate - Mozilla Firefox
File Edit View History Bookmarks ScrapBook Tools Help
Google Translate
translate.google.com/#auto/en/あけましておめでとうございます。
Google
Sign in
Translate
English Bulgarian French Japanese - detected
English Chinese (Simplified) Japanese
Translate

あけましておめでとうございます。

Happy New Year!

Akemashite omedetōgozaimasu.

Turn off instant translation
About Google Translate Mobile Privacy Help Send feedback
Find in page
Highlight All Match Case

Elephants are social animals. They live with their families, give hugs and call each other by using their trunks as trumpets. They also might know how to help each other.

In a recent elephant study by researchers from the United States and Thailand, pairs of giant animals learned to work together to get some ears of corn. Other animals, especially some primates, are already known to work together to complete tasks, but now elephants have joined the club. Perhaps the finding is not too surprising: Scientists suspect that elephants, with their big brains and survival savvy, may be among the smartest animals on the planet.

Joshua Plotnik, who worked on the study, told Science News that the animals didn’t just learn a trick. Instead, the ways the elephants behaved show that they understand how working together brings benefits to everyone involved. Plotnik is a comparative psychologist now at the University of Cambridge in England. Psychology is the study of behaviors and mental processes, and comparative psychologists study how animals other than humans behave.

Les éléphants sont des animaux sociaux. Ils vivent avec leur famille, faire des câlins et appeler les uns les autres en utilisant leurs troncs trompettes. Ils pourraient également savoir comment aider les uns les autres.

Dans une étude récente d'éléphants par des chercheurs des États-Unis et la Thaïlande, des paires d'animaux géants ont appris à travailler ensemble pour obtenir des épis de maïs. D'autres animaux, en particulier des primates, sont déjà connus pour travailler ensemble pour accomplir des tâches, mais maintenant, les éléphants ont rejoint le club. Peut-être le résultat n'est pas trop surprenant: Les scientifiques soupçonnent que les éléphants, avec leurs gros cerveaux et de bon sens de survie, peut-être parmi les plus intelligents des animaux sur la planète.

Joshua Plotnick, qui a travaillé sur l'étude, dit Nouvelles de la Science que les animaux n'ont pas seulement appris un truc. Au lieu de cela, les moyens les éléphants se comportent montrent qu'ils comprennent comment travailler ensemble apporte des avantages à toutes les personnes impliquées. Plotnik est un psychologue comparative maintenant à l'Université de Cambridge en Angleterre. **La psychologie est l'étude des comportements et des processus mentaux**, et étude comparative des psychologues comment les animaux autres que les humains se comportent.

https://student.societyforscience.org/article/theres-no-i-elephant

Elephants are social animals. They live with their families, give hugs and call each other by using their trunks as trumpets. They also might know how to help each other.

In a recent elephant study by researchers from the United States and Thailand, pairs of giant animals learned to work together to get some ears of corn. Other animals, especially some primates, are already known to work together to complete tasks, but now elephants have joined the club. Perhaps the finding is not too surprising: Scientists suspect that elephants, with their big brains and survival savvy, may be among the smartest animals on the planet.

Joshua Plotnik, who worked on the study, told Science News that the animals didn’t just learn a trick. Instead, the ways the elephants behaved show that they understand how working together brings benefits to everyone involved. Plotnik is a comparative psychologist now at the University of Cambridge in England. Psychology is the study of behaviors and mental processes, and comparative psychologists study how animals other than humans behave.

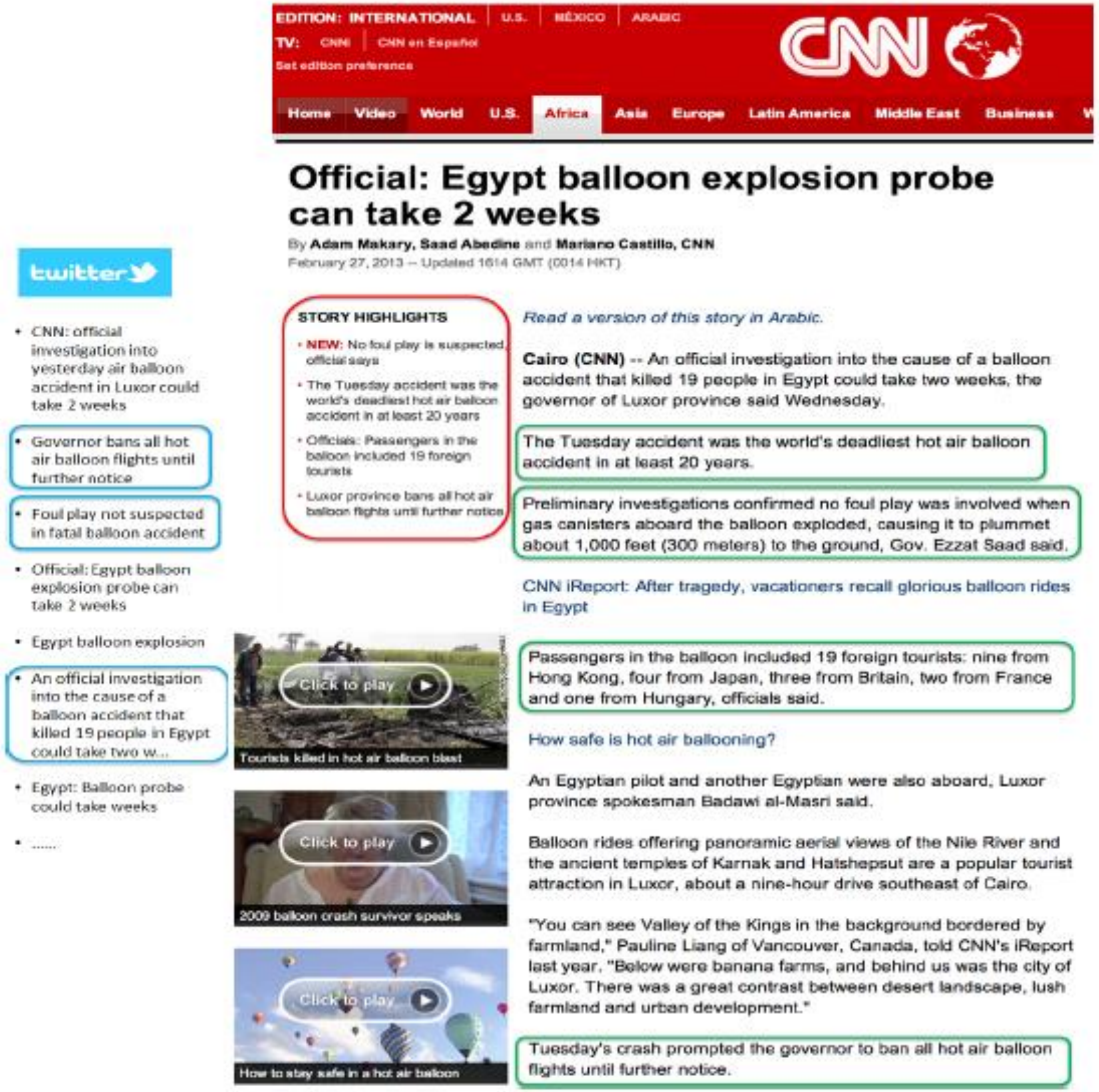
Les éléphants sont des animaux sociaux. Ils **vivent** avec leur famille, **faire** des câlins et **appeler** les uns les autres en utilisant leurs troncs trompettes. Ils pourraient également savoir comment aider les uns les autres.

Dans une étude récente d'éléphants par des chercheurs des États-Unis et la Thaïlande, des paires d'animaux géants ont appris à travailler ensemble pour obtenir des épis de maïs. D'autres animaux, en particulier des primates, sont déjà connus pour travailler ensemble pour accomplir des tâches, mais maintenant, les éléphants ont rejoint le club. Peut-être le résultat n'est pas trop surprenant: Les scientifiques soupçonnent que **les éléphants**, avec leurs gros cerveaux et de bon sens de survie, **peut-être** parmi les plus intelligents des animaux sur la planète.

Joshua Plotnick, qui a travaillé sur l'étude, dit **Nouvelles de la Science** que les animaux n'ont pas seulement appris un truc. Au lieu de cela, les moyens les éléphants se comportent montrent qu'ils comprennent comment travailler ensemble apporte des avantages à toutes **les personnes** impliquées. Plotnik est un psychologue **comparative** maintenant à l'Université de Cambridge en Angleterre. La psychologie est l'étude des comportements et des processus mentaux, et **étude comparative des psychologues** comment les animaux autres que les humains se comportent.

https://student.societyforscience.org/article/theres-no-i-elephant

Single Document Summarization



- CNN: official investigation into yesterday air balloon accident in Luxor could take 2 weeks
- Governor bans all hot air balloon flights until further notice
- Foul play not suspected in fatal balloon accident
- Official: Egypt balloon explosion probe can take 2 weeks
- Egypt balloon explosion
- An official investigation into the cause of a balloon accident that killed 19 people in Egypt could take two w...
- Egypt: Balloon probe could take weeks
- ......

Official: Egypt balloon explosion probe can take 2 weeks

By Adam Makary, Saad Abedine and Mariano Castillo, CNN
February 27, 2013 -- Updated 1614 GMT (0014 HKT)

STORY HIGHLIGHTS

- NEW: No foul play is suspected, official says
- The Tuesday accident was the world's deadliest hot air balloon accident in at least 20 years
- Officials: Passengers in the balloon included 19 foreign tourists
- Luxor province bans all hot air balloon flights until further notice

Read a version of this story in Arabic.

Cairo (CNN) -- An official investigation into the cause of a balloon accident that killed 19 people in Egypt could take two weeks, the governor of Luxor province said Wednesday.

The Tuesday accident was the world's deadliest hot air balloon accident in at least 20 years.

Preliminary investigations confirmed no foul play was involved when gas canisters aboard the balloon exploded, causing it to plummet about 1,000 feet (300 meters) to the ground, Gov. Ezzat Saad said.

CNN iReport: After tragedy, vacationers recall glorious balloon rides in Egypt

Passengers in the balloon included 19 foreign tourists: nine from Hong Kong, four from Japan, three from Britain, two from France and one from Hungary, officials said.

How safe is hot air ballooning?

An Egyptian pilot and another Egyptian were also aboard, Luxor province spokesman Badawi al-Masri said.

Balloon rides offering panoramic aerial views of the Nile River and the ancient temples of Karnak and Hatshepsut are a popular tourist attraction in Luxor, about a nine-hour drive southeast of Cairo.

"You can see Valley of the Kings in the background bordered by farmland," Pauline Liang of Vancouver, Canada, told CNN's iReport last year. "Below were banana farms, and behind us was the city of Luxor. There was a great contrast between desert landscape, lush farmland and urban development."

Tuesday's crash prompted the governor to ban all hot air balloon flights until further notice.

Click to play
Tourists killed in hot air balloon blast
Click to play
2009 balloon crash survivor speaks
Click to play
How to stay safe in a hot air balloon

[Wei and Gao 2014]

Multi Document Summarization

Health Benefits

- Eating a diet rich in vegetables and fruits as part of an overall healthy diet may reduce risk for heart disease, including heart attack and stroke.
- Eating a diet rich in some vegetables and fruits as part of an overall healthy diet may protect against certain types of cancers.
- Diets rich in foods containing fiber, such as some vegetables and fruits, may reduce the risk of heart disease, obesity, and type 2 diabetes.
- Eating vegetables and fruits rich in potassium as part of an overall healthy diet may lower blood pressure, and may also reduce the risk of developing kidney stones and help to decrease bone loss.
- Eating foods such as vegetables that are lower in calories per cup instead of some other higher-calorie food may be useful in helping to lower calorie intake.

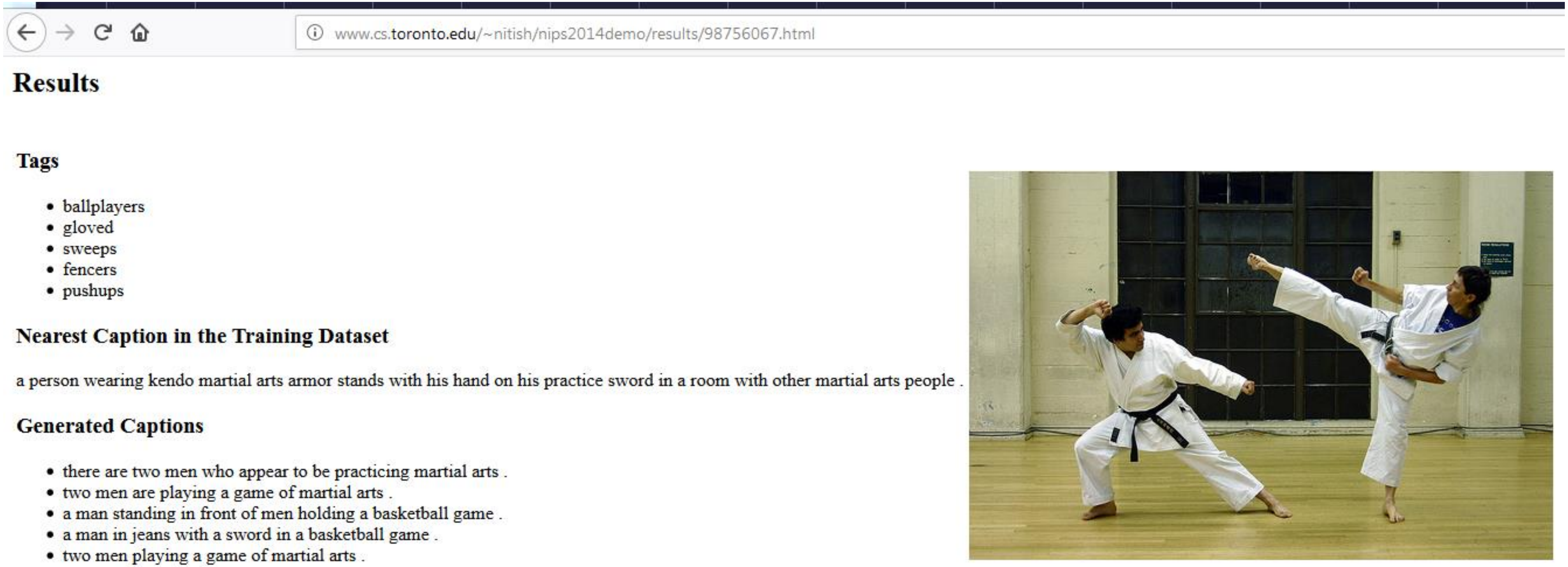
Nutrients

- Most vegetables are naturally low in fat and calories. None have cholesterol. (Sauces or seasonings may add fat, calories, or cholesterol.)
- Vegetables are important sources of many nutrients, including potassium, dietary fiber, folate (folic acid), vitamin A, and vitamin C.
- Diets rich in potassium may help to maintain healthy blood pressure. Vegetable sources of potassium include sweet potatoes, white potatoes, white beans, tomato products (paste, sauce, and juice), beet greens, soybeans, lima beans, spinach, lentils, and kidney beans.
- Dietary fiber from vegetables, as part of an overall healthy diet, helps reduce blood cholesterol levels and may lower risk of heart disease. Fiber is important for proper bowel function. It helps reduce constipation and diverticulosis. Fiber-containing foods such as vegetables help provide a feeling of fullness with fewer calories.
- Folate (folic acid) helps the body form red blood cells. Women of childbearing age who may become pregnant should consume adequate folate from foods, and in addition 400 mcg of synthetic folic acid from fortified foods or supplements. This reduces the risk of neural tube defects, spina bifida, and anencephaly during fetal development.
- Vitamin A keeps eyes and skin healthy and helps to protect against infections.
- Vitamin C helps heal cuts and wounds and keeps teeth and gums healthy. Vitamin C aids in iron absorption.

Summary

Eating vegetables is healthy.

Caption Generation



www.cs.toronto.edu/~nitish/nips2014demo/results/98756067.html

Results

Tags

- ballplayers
- gloved
- sweeps
- fencers
- pushups

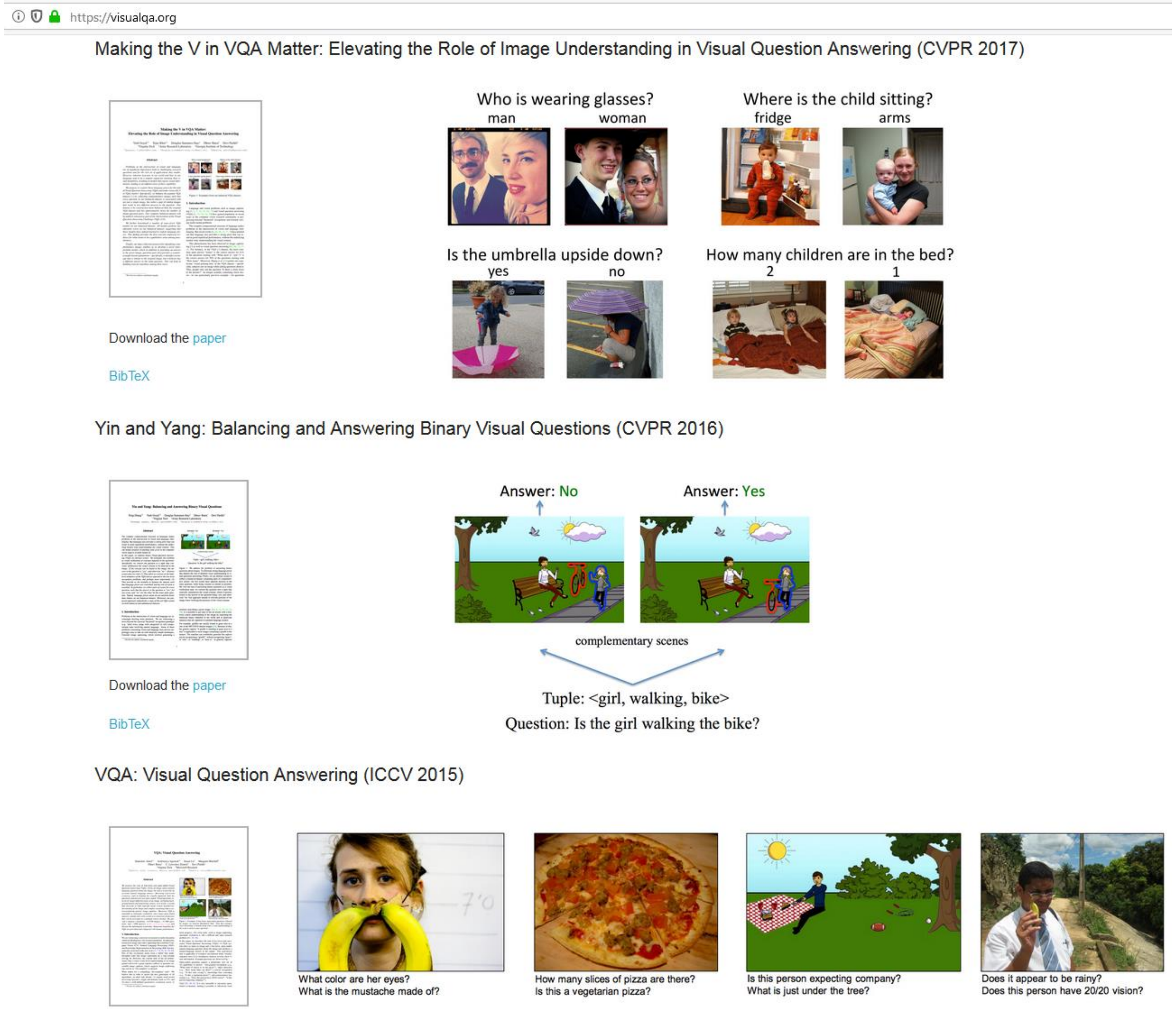
Nearest Caption in the Training Dataset

a person wearing kendo martial arts armor stands with his hand on his practice sword in a room with other martial arts people .

Generated Captions

- there are two men who appear to be practicing martial arts .
- two men are playing a game of martial arts .
- a man standing in front of men holding a basketball game .
- a man in jeans with a sword in a basketball game .
- two men playing a game of martial arts .

Visual Question Answering



https://visualqa.org

Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering (CVPR 2017)

Download the paper

BibTeX

Who is wearing glasses?
man
woman
Where is the child sitting?
fridge
arms
Is the umbrella upside down?
yes
no
How many children are in the bed?
2
1

Yin and Yang: Balancing and Answering Binary Visual Questions (CVPR 2016)

Download the paper

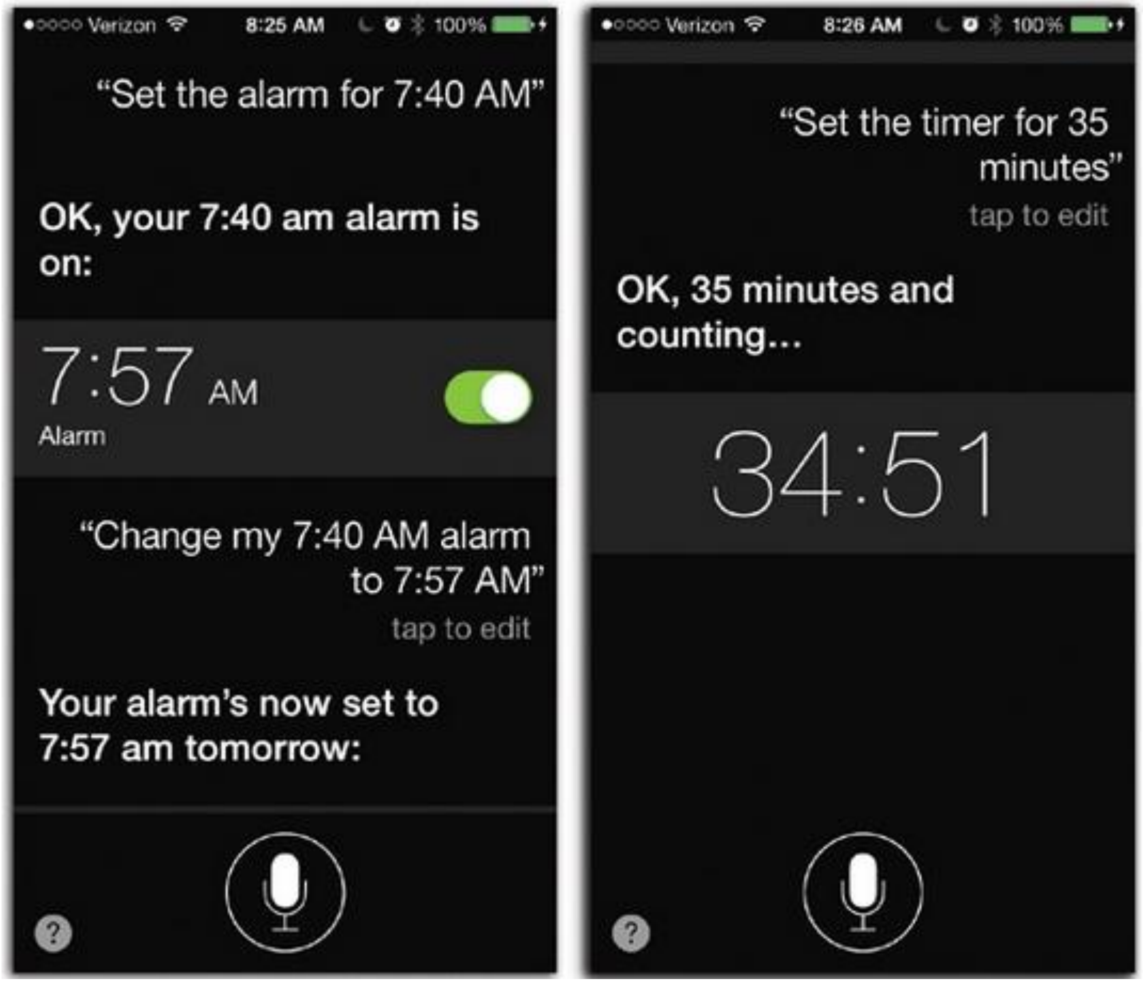
BibTeX

Answer: No
Answer: Yes
complementary scenes
Tuple: <girl, walking, bike>
Question: Is the girl walking the bike?

VQA: Visual Question Answering (ICCV 2015)

What color are her eyes?
What is the mustache made of?
How many slices of pizza are there?
Is this a vegetarian pizza?
Is this person expecting company?
What is just under the tree?
Does it appear to be rainy?
Does this person have 20/20 vision?

Conversational Agents



"Set the alarm for 7:40 AM"

OK, your 7:40 am alarm is on:

7:57 AM

"Change my 7:40 AM alarm to 7:57 AM"

Your alarm's now set to 7:57 am tomorrow:

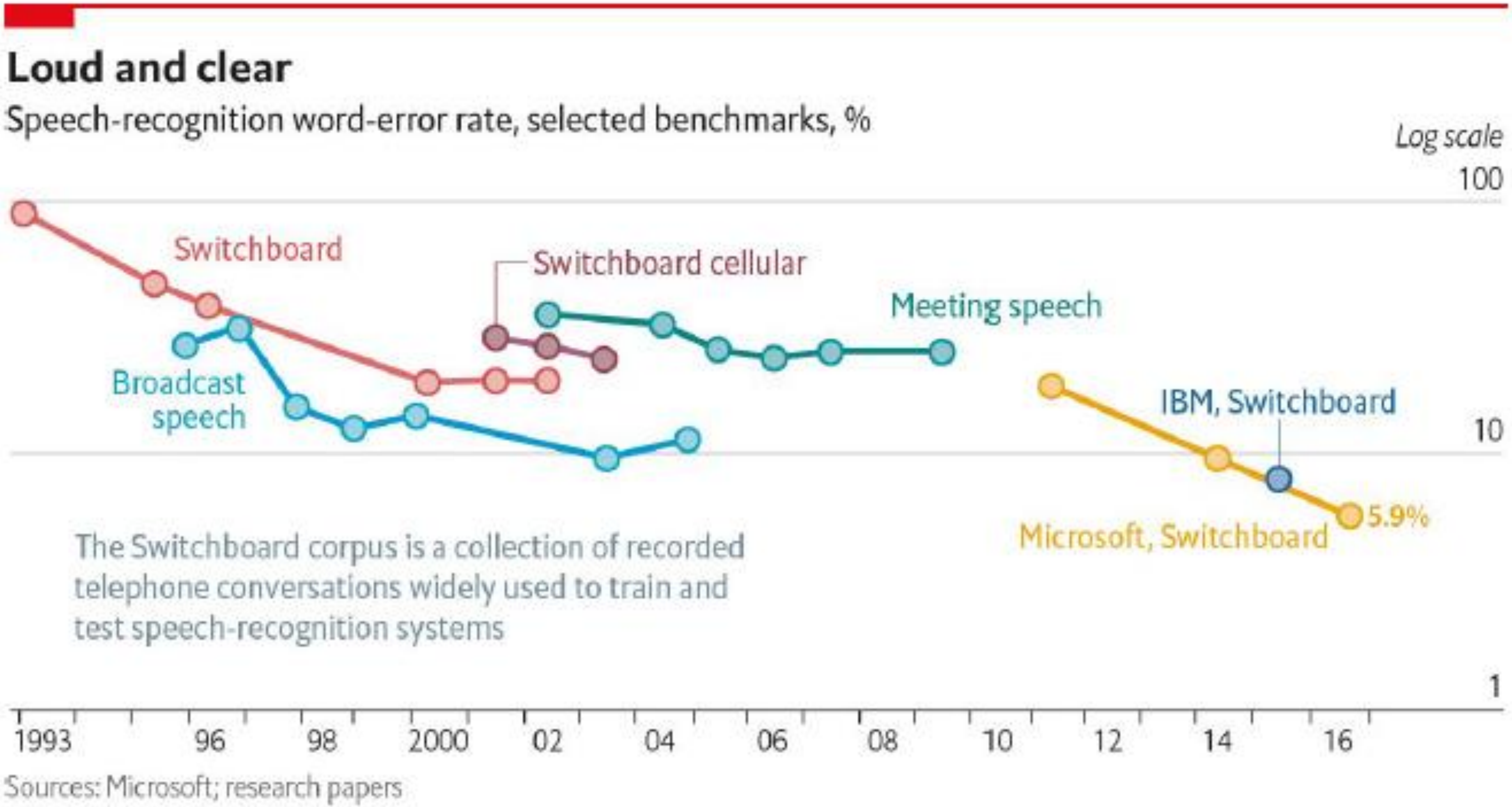
"Set the timer for 35 minutes"

OK, 35 minutes and counting...

34:51

https://www.oreilly.com/library/view/iphone-the-missing/9781449372781/ch04.html

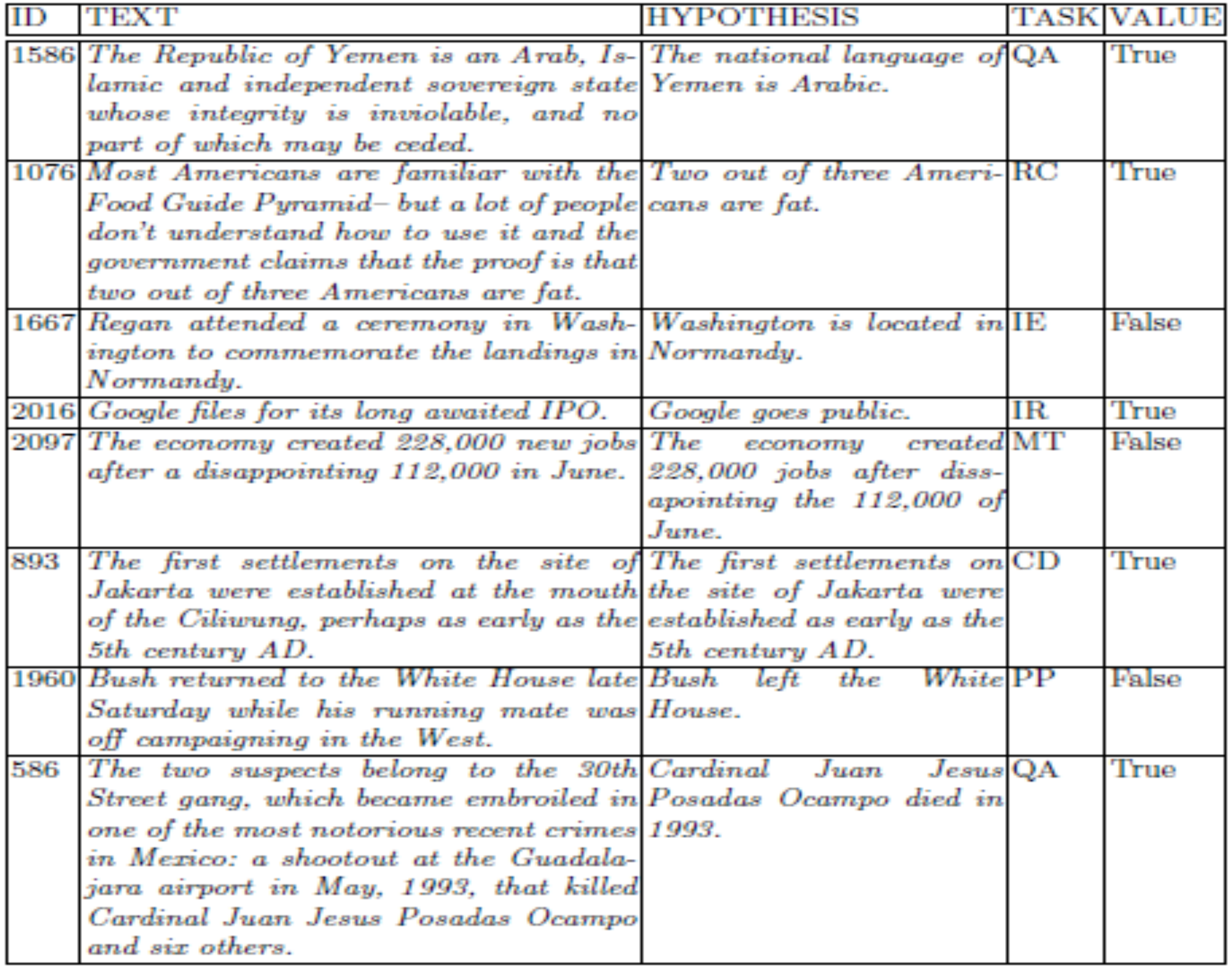
Speech Recognition



The Switchboard corpus is a collection of recorded telephone conversations widely used to train and test speech-recognition systems

Sources: Microsoft; research papers

Entailment and Paraphrasing



ID	TEXT	HYPOTHESIS	TASK	VALUE
1586	*The Republic of Yemen is an Arab, Islamic and independent sovereign state whose integrity is inviolable, and no part of which may be ceded.*	*The national language of Yemen is Arabic.*	QA	True
1076	*Most Americans are familiar with the Food Guide Pyramid– but a lot of people don't understand how to use it and the government claims that the proof is that two out of three Americans are fat.*	*Two out of three Americans are fat.*	RC	True
1667	*Regan attended a ceremony in Washington to commemorate the landings in Normandy.*	*Washington is located in Normandy.*	IE	False
2016	*Google files for its long awaited IPO.*	*Google goes public.*	IR	True
2097	*The economy created 228,000 new jobs after a disappointing 112,000 in June.*	*The economy created 228,000 jobs after dissapointing the 112,000 of June.*	MT	False
893	*The first settlements on the site of Jakarta were established at the mouth of the Ciliwung, perhaps as early as the 5th century AD.*	*The first settlements on the site of Jakarta were established as early as the 5th century AD.*	CD	True
1960	*Bush returned to the White House late Saturday while his running mate was off campaigning in the West.*	*Bush left the White House.*	PP	False
586	*The two suspects belong to the 30th Street gang, which became embroiled in one of the most notorious recent crimes in Mexico: a shootout at the Guadalajara airport in May, 1993, that killed Cardinal Juan Jesus Posadas Ocampo and six others.*	*Cardinal Juan Jesus Posadas Ocampo died in 1993.*	QA	True

Table 1. Examples of Text-Hypothesis pairs

Ido Dagan, Oren Glickman and Bernardo Magnini. The PASCAL Recognising Textual Entailment Challenge

Discourse Analysis

- Anaphoric relations:

1. Mary helped Peter get out of the car. ***He*** *thanked* ***her****.*

2. Mary helped the other passenger out of the car.

The man had asked ***her*** *for help because of* ***his*** *foot injury.*

Tom appeared on the sidewalk with a bucket of whitewash and a long-handled brush. He surveyed the fence, and all gladness left him and a deep melancholy settled down upon his spirit. (Tom Sawyer)

Dialogue Systems

- I would like to make a reservation at Sorrento.
- For when?
- 8 pm Friday night.
- We only have availability for 7 pm and 10 pm.
- Sorry, these times don't work for me.

Introduction to NLP

153

Preprocessing

Text Preprocessing

- Removing non-text
  - ads, javascript
- Dealing with text encoding
  - e.g., Unicode
- Sentence segmentation
- Normalization
  - labeled/labelled, extra-terrestrial/extraterrestrial, extra terrestrial
- Stemming
  - computer/computation
- Morphological similarity
  - car/cars
- Capitalization
  - Now/NOW, led/LED
- Phrase and named entity extraction, foreign words
  - USA/usa, MIT/mit

Text Preprocessing

- Types vs. Tokens
  - To be or not to be
- Tokenization:
  - ALS vs. A.L.S.
  - Paul’s, Willow Dr., Dr. Willow, New York, ad hoc, can’t
  - “The New York-Los Angeles flight” vs. “Minneapolis-St.Paul”
  - Numbers, e.g., (888) 555-1313, 1-888-555-1313
  - Dates, e.g., Jan-13-2012, 20120113, 13 January 2012, 01/13/12
  - URLs

Word segmentation into morphemes

- Arabic:

  كتاب

- Japanese:

  この本は重い。

  (kono hon ha omoi)

- German:

  Finanzdienstleistung = financial services

- Chinese:

  电视 (television)
  电 (diàn = electric) 视 (shì = to look at)

Text preprocessing

ニューヨーク (New York) は、アメリカ合衆国ニューヨーク州にある都市

- Kanji, Katakana, Hiragana, Rōmaji, (numbers)
- Nyūyōku wa, Amerikagasshūkoku nyūyōku-shū ni aru toshi

Sentence segmentation into words

- 金属製品製造の日立金属は１９日、世界最大手の鉄鋳物メーカー「ワウパカ　ファウンドリー　ホールディングス」（米国・デラウェア州）を米投資ファンドから買収し、完全子会社にすると発表した。買収額は１３億ドル（約１３３０億円）で、１０月中にも手続きを終える。

Sentence boundary recognition

- Classification problem
- Features
  - punctuation
  - formatting
  - fonts
  - spacing
  - capitalization
  - case
  - use of abbreviations, e.g., Dr., a.m.
- Example
  - If there is no space after a period, don’t assume that there is a sentence boundary