

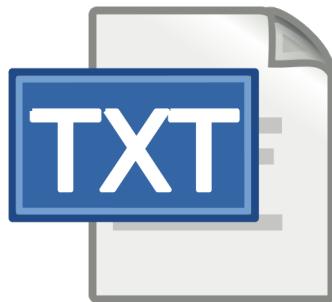
Introduction to Natural Language Processing

Levels of linguistic analysis

Language Technology

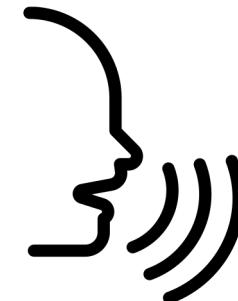
Language Technology (LT) = Natural Language Processing (NLP)
or/and Computational Linguistics (CL) + Speech Technology
(Automatic Speech Recognition and Synthesis)

Written natural language



vs

Spoken natural language



https://en.wikipedia.org/wiki/Language_technology

Created by Kiran Joseph
from the Noun Project

Levels of linguistic analysis

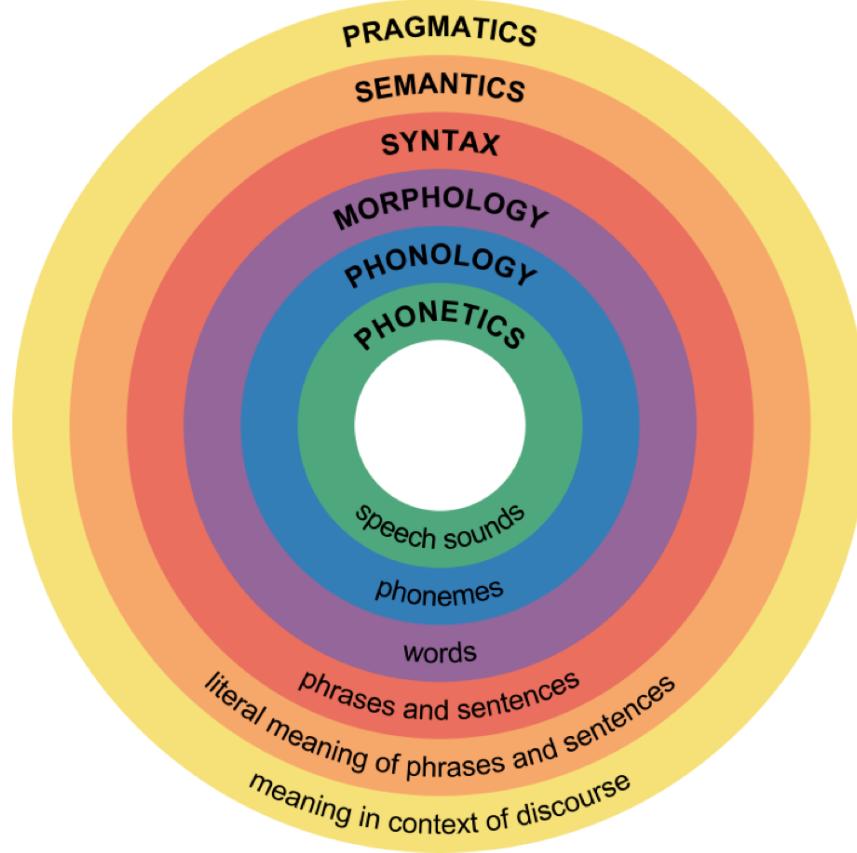


Image source: https://commons.wikimedia.org/wiki/File:Major_levels_of_linguistic_structure.svg

Phonetical level

- Phonetics is the investigation of production, transmission and impression of discourse sound.
- It is worried about the hints of languages, how these sounds are verbalized and how the listener sees them. Phonetics is identified with study of acoustics in that it utilizes much similar methods in the examination of sounds that acoustics does.
- Phonology is worried about how sounds are composed in a language.

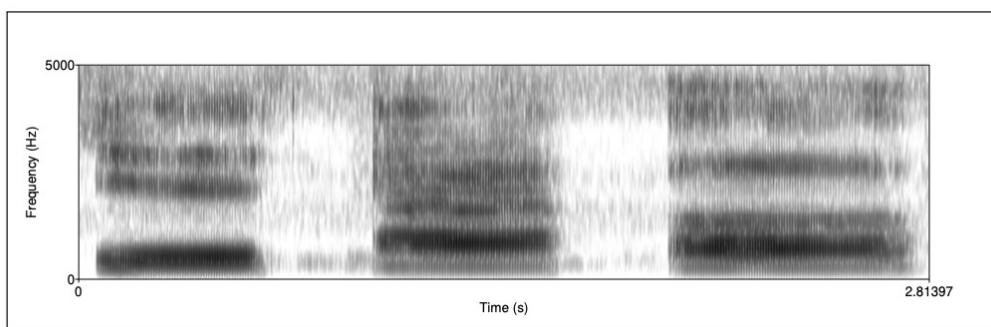
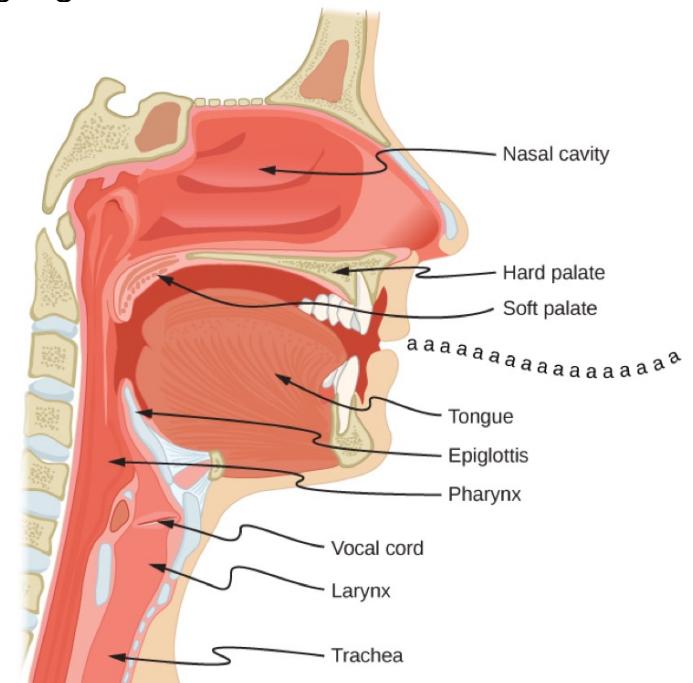


Figure 25.19 Spectrograms for three American English vowels, [ih], [ae], and [uh]

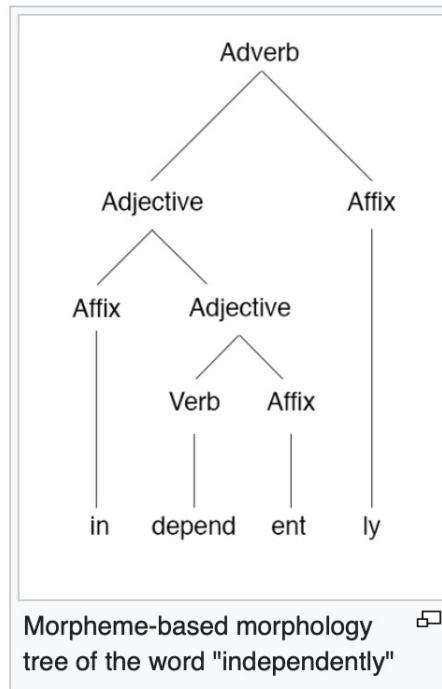


<https://web.stanford.edu/~jurafsky/slp3/25.pdf>

<https://styleinfections.blogspot.com/2012/03/levels-of-linguistic-analysis.html>

Morphological level

- Morphology is investigation of **word arrangement** and **structure**.
- How words are assembled from their littler parts and the principles overseeing this process.
- The components that are joining to frame words are called **morpheme**
- A morpheme is the littlest unit of syntax you can have in language the felines e.g contains the morphemes feline and the plurals.



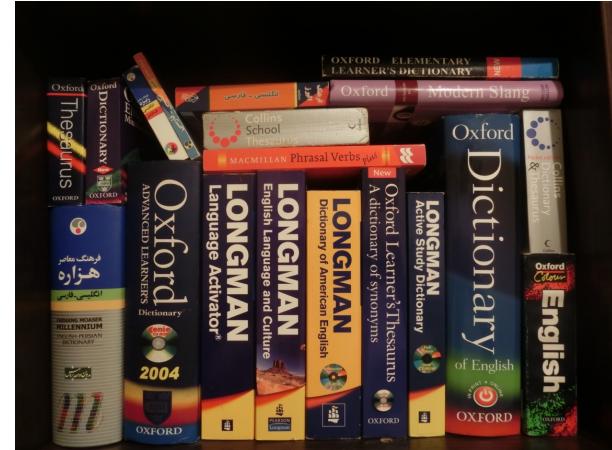
аберрационный	аберрационную	ADJ; ACC; FEM; SG
аберрационные	аберрационные	ADJ; ACC; INAN; PL
аберрационные	аберрационные	ADJ; NOM; PL
аберрационный	аберрационный	ADJ; INAN; ACC; MASC; SG
аберрационный	аберрационный	ADJ; NOM; MASC; SG
аберрационный	аберрационным	ADJ; DAT; PL
аберрационный	аберрационным	ADJ; INS; MASC; SG
аберрационный	аберрационным	ADJ; INS; NEUT; SG
аберрационными	аберрационными	ADJ; INS; PL
аберрационных	аберрационных	ADJ; ACC; ANIM; PL
аберрационных	аберрационных	ADJ; ESS; PL
аберрационный	аберраций	ADJ; GEN; PL
аберрация	аберрации	N; INS; SG
аберрация	аберрации	N; ACC; PL
аберрация	аберрации	N; DAT; SG
аберрация	аберрации	N; ESS; SG
аберрация	аберрации	N; GEN; SG
аберрация	аберрации	N; NOM; PL
аберрация	аберраций	N; GEN; PL
аберрация	аберрацию	N; ACC; SG
аберрация	аберрация	N; NOM; SG
аберрация	аберрациям	N; DAT; PL
аберрация	аберрациями	N; INS; PL
аберрация	аберрациях	N; ESS; PL

<https://styleinflections.blogspot.com/2012/03/levels-of-linguistic-analysis.html>
<https://web.stanford.edu/~jurafsky/slp3/8.pdf>

Lexical level

- Lexicology is investigation of words.
- A word is the **smallest meaningful unit of a language that can stand on its own**, and is made up of small components called **morphemes** and even smaller elements known as **phonemes**, or distinguishing sounds.
- Lexicology also considers the **relationships** that exist between words.
- The lexicon of a language is composed of **lexemes**: abstract units of meaning that correspond to a set of related forms of a word.

- **S:** (n) **bank** (sloping land (especially the slope beside a body of water)) "they pulled the canoe up on the bank"; "he sat on the bank of the river and watched the currents"
- **S:** (n) depository financial institution, **bank**, banking concern, banking company (a financial institution that accepts deposits and channels the money into lending activities) "he cashed a check at the bank"; "that bank holds the mortgage on my home"



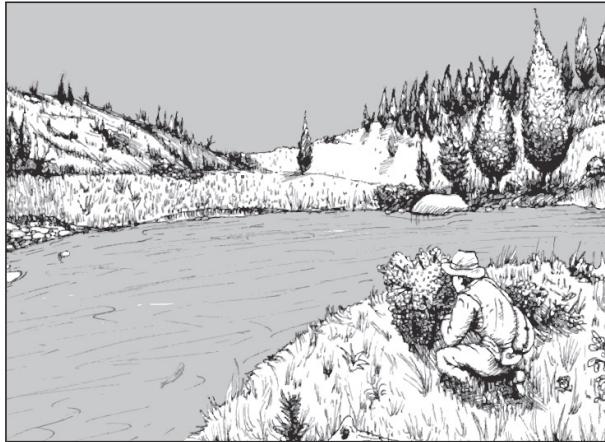
PRINCETON UNIVERSITY

WordNet
A Lexical Database for English

<https://styleinfictions.blogspot.com/2012/03/levels-of-linguistic-analysis.html>

https://upload.wikimedia.org/wikipedia/commons/e/e2/English-English_and_English-Persian_dictionaries.JPG

Lexical ambiguity and why language is hard ...



He sat on the river **bank** and counted his **dough**.

She went to the **bank** and took out some **money**.

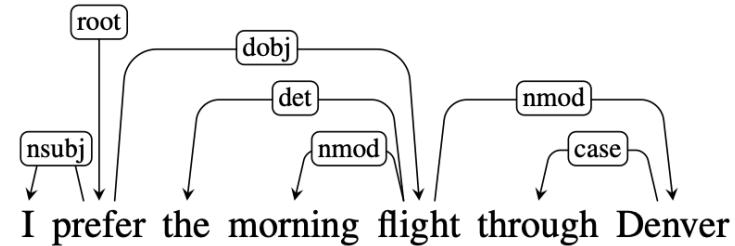
polysemous

synonymous

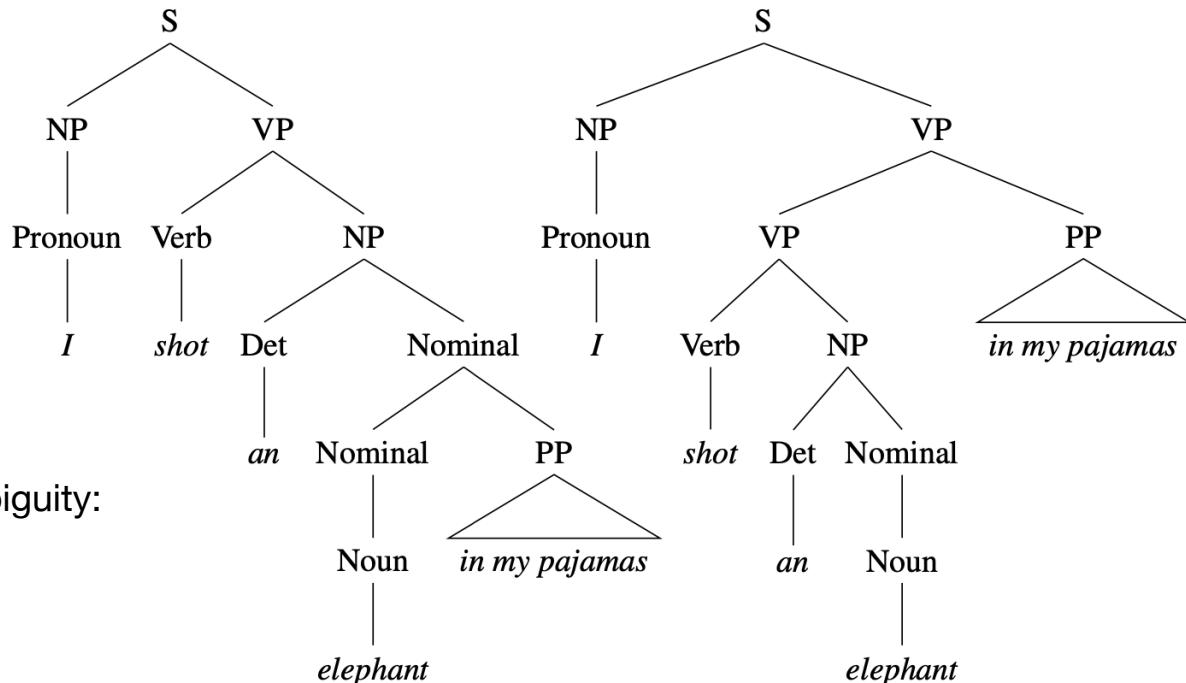
Concept Layer

Lexical Layer

Syntactic level



- It is the investigation of **sentence structure**.
- The basic structure of English e.g would have a subject - action word - object sentence order, .e.g. John hit the ball.



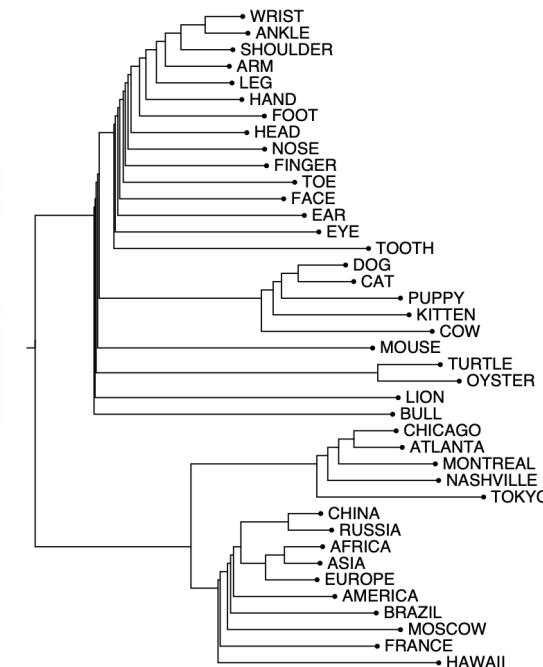
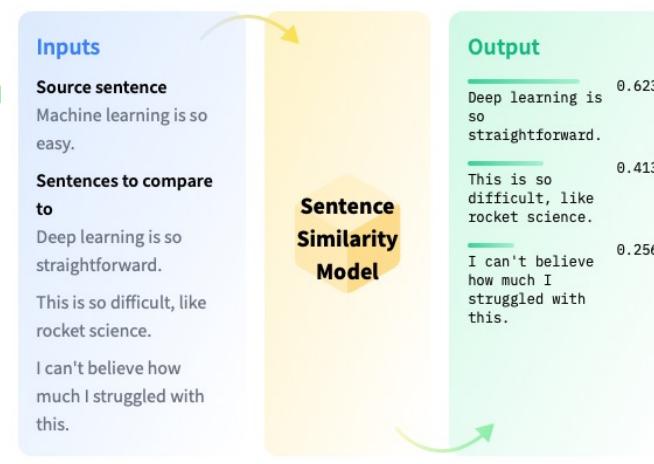
<https://styleinfections.blogspot.com/2012/03/levels-of-linguistic-analysis.html>

<https://web.stanford.edu/~jurafsky/slp3/13.pdf>

Semantic level

- Semantics is the investigation of significance in language.
 - Semantics studies meaning at the levels of **words, phrases, sentences**, or larger units of **discourse**.
 - Computational lexical semantics: computational modelling meaning of lexical units (words and phrases). Also models of computational meaning of sentences is illustrated below.

			not good	bad
to	by	's	dislike	worst
that	now	are	incredibly bad	worse
a	i	you		
than	with	is		



<https://web.stanford.edu/~jurafsky/slp3/6.pdf>

Pragmatics and Discourse

- **Pragmatics** examines the components that administer our decision of language in social association and the effects of our decisions on others.
- **Discourse** is the investigation of stretches of spoken and composed language over the sentence.

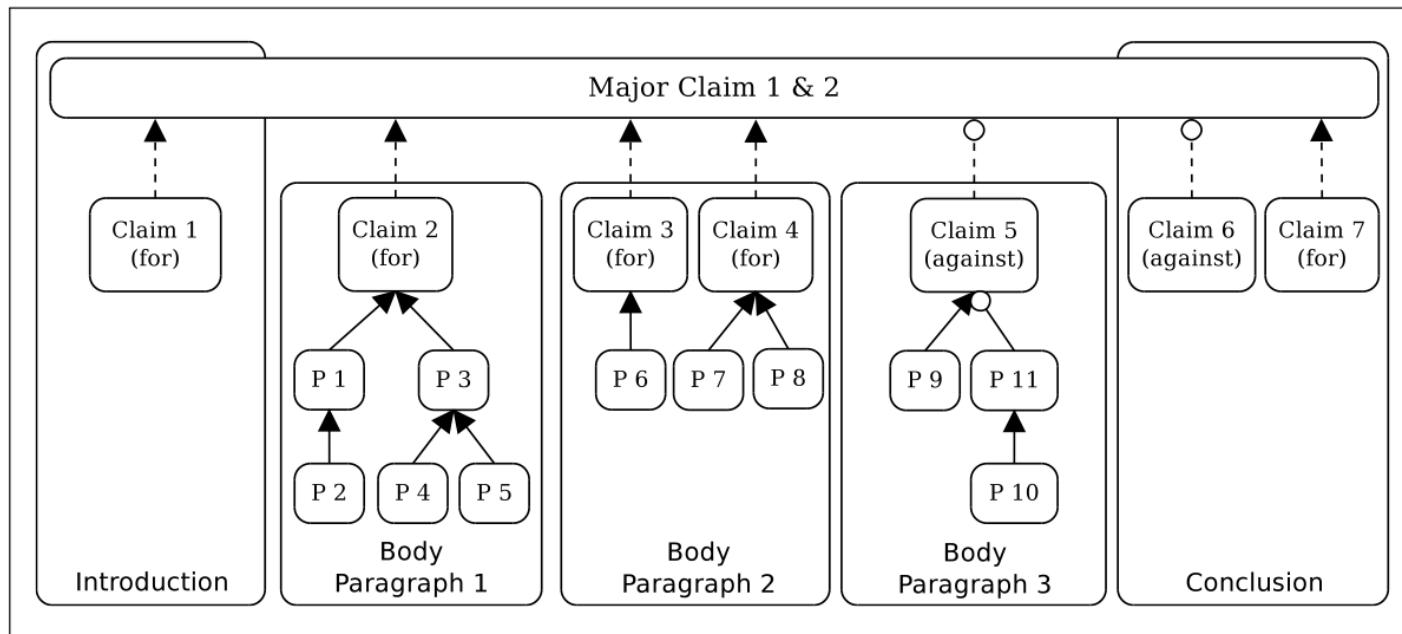


Figure 22.12 Argumentation structure of a persuasive essay. Arrows indicate argumentation relations, either of SUPPORT (with arrowheads) or ATTACK (with circleheads); P denotes premises. Figure from [Stab and Gurevych \(2017\)](#).

<https://styleinfections.blogspot.com/2012/03/levels-of-linguistic-analysis.html>
<https://web.stanford.edu/~jurafsky/slp3/22.pdf>

Levels of linguistic analysis

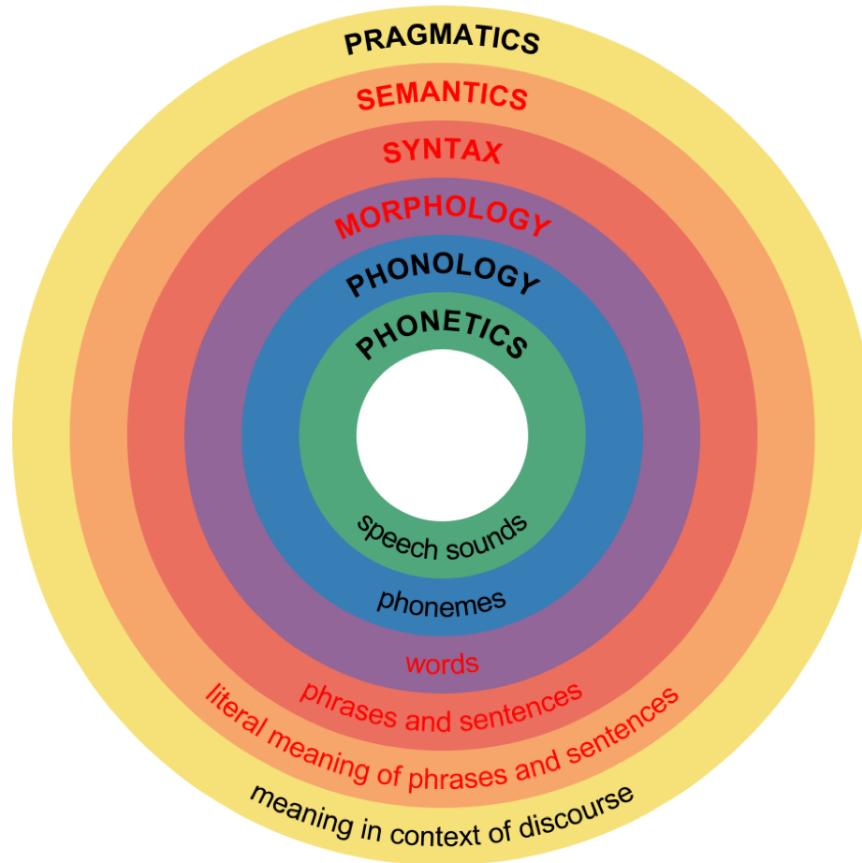


Image source: https://commons.wikimedia.org/wiki/File:Major_levels_of_linguistic_structure.svg

Examples of NLP applications

Panorama of NLP tasks: Sentiment Classification

- **Input:** text
- **Output:** a sentiment label (positive / negative)

w	P(w +)	P(w -)
I	0.1	0.2
love	0.1	0.001
this	0.01	0.01
fun	0.05	0.005
film	0.1	0.1
...

Four original documents:

- it was pathetic the worst part was the boxing scenes
- no plot twists or great scenes
- + and satire and great plot twists
- + great scenes great film

After per-document binarization:

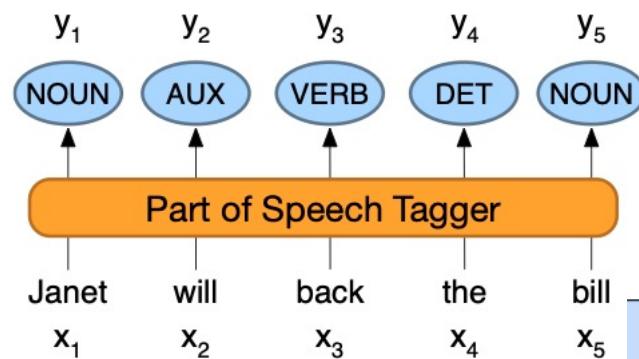
- it was pathetic the worst part boxing scenes
- no plot twists or great scenes
- + and satire great plot twists
- + great scenes film

	NB Counts		Binary Counts	
	+	-	+	-
and	2	0	1	0
boxing	0	1	0	1
film	1	0	1	0
great	3	1	2	1
it	0	1	0	1
no	0	1	0	1
or	0	1	0	1
part	0	1	0	1
pathetic	0	1	0	1
plot	1	1	1	1
satire	1	0	1	0
scenes	1	2	1	2
the	0	2	0	1
twists	1	1	1	1
was	0	2	0	1
worst	0	1	0	1

Panorama of NLP tasks: Sequence Labelling

- **Input:** text
- **Output:** each token has a label (part of speech or named entity)

Type	Tag	Sample Categories	Example sentences
People	PER	people, characters	Turing is a giant of computer science.
Organization	ORG	companies, sports teams	The IPCC warned about the cyclone.
Location	LOC	regions, mountains, seas	Mt. Sanitas is in Sunshine Canyon.
Geo-Political Entity	GPE	countries, states	Palo Alto is raising the fees for parking.



[PER Washington] was born into slavery on the farm of James Burroughs.
[ORG Washington] went up 2 games to 1 in the four-game series.
Blair arrived in [LOC Washington] for what may well be his last state visit.
In June, [GPE Washington] passed a primary seatbelt law.

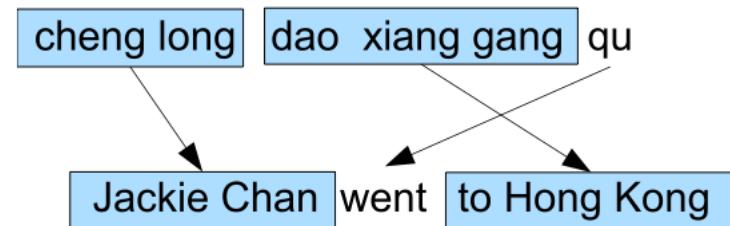
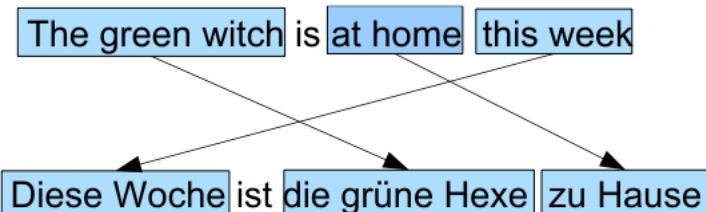
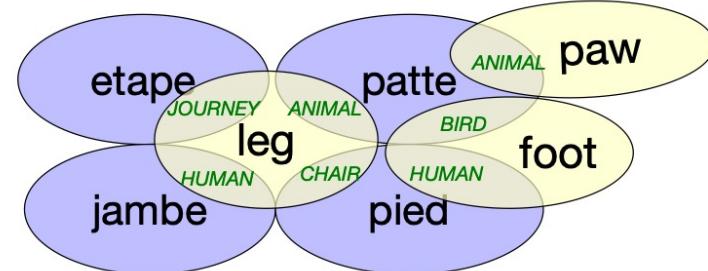
Panorama of NLP tasks: Machine Translation

- **Input:** text in source language
- **Output:** text in target language (translated text)

English: *He wrote a letter to a friend*

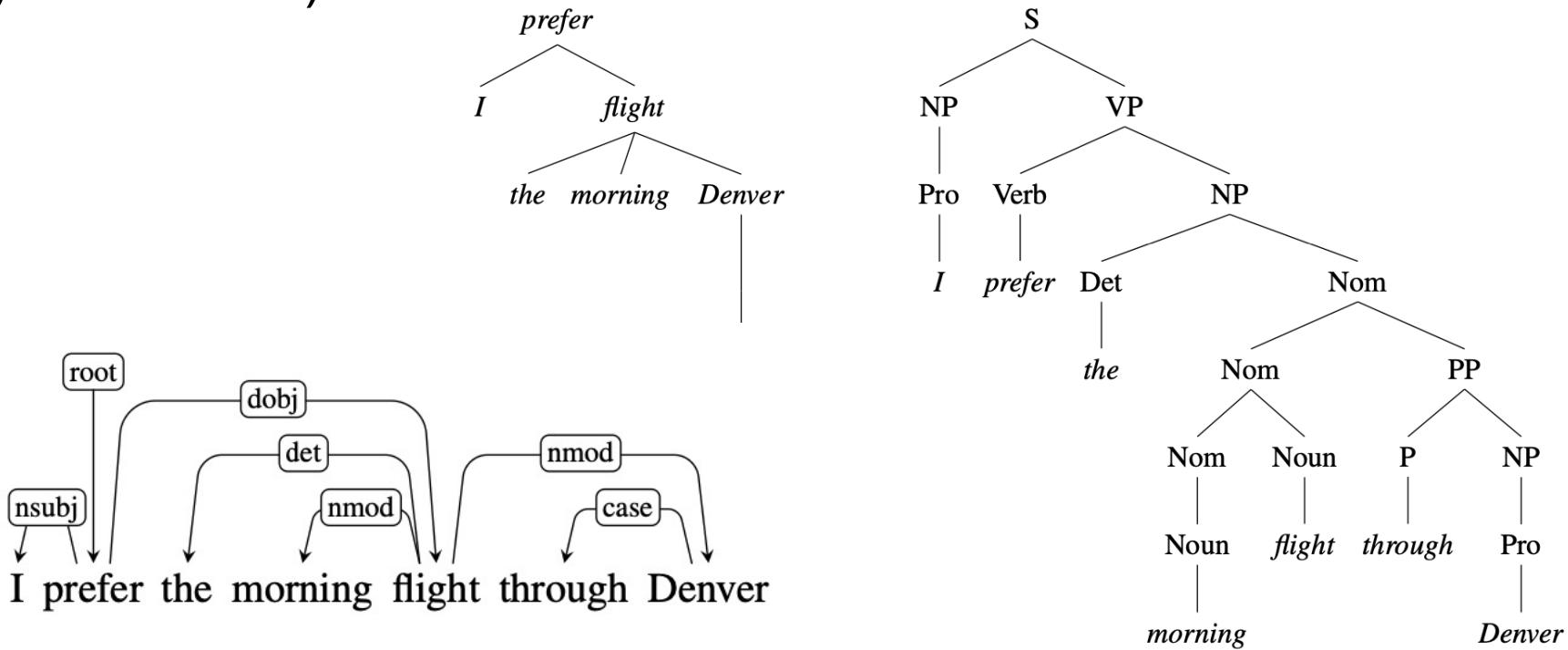
Japanese: *tomodachi ni tegami-o kaita*
friend to letter wrote

Arabic: *katabt risāla li šadq*
wrote letter to friend



Panorama of NLP tasks: Syntactic Parsing

- **Input:** sentence
- **Output:** syntactically analyzed sentence (usually in a form of a syntactic tree)



<https://web.stanford.edu/~jurafsky/slp3/14.pdf>

Panorama of NLP tasks: Information Extraction

- **Input:** text
- **Output:** relation between entities in text

Example	Entity Types & Label
Carey will succeed Cathleen P. Black , who held the position for 15 years and will take on a new role as chairwoman of Hearst Magazines, the company said.	PERSON/TITLE Relation: <i>per:title</i>
Irene Morgan Kirkaldy, who was born and reared in Baltimore , lived on Long Island and ran a child-care center in Queens with her second husband, Stanley Kirkaldy.	PERSON/CITY Relation: <i>per:city_of_birth</i>
Baldwin declined further comment, and said JetBlue chief executive Dave Barger was unavailable.	Types: PERSON/TITLE Relation: <i>no_relation</i>

Relations	Types	Examples
Physical-Located	PER-GPE	He was in Tennessee
Part-Whole-Subsidiary	ORG-ORG	XYZ , the parent company of ABC
Person-Social-Family	PER-PER	Yoko 's husband John
Org-AFF-Founder	PER-ORG	Steve Jobs , co-founder of Apple ...

Panorama of NLP tasks: Semantic Role Labelling

- **Input:** sentence / text
- **Output:** semantic roles labelled in text / sentence

(19.3) *John broke the window.*

AGENT THEME

(19.4) *John broke the window with a rock.*

AGENT THEME INSTRUMENT

(19.5) *The rock broke the window.*

INSTRUMENT THEME

(19.6) *The window broke.*

THEME

(19.7) *The window was broken by John.*

THEME AGENT

(19.11) **agree.01**

Arg0: Agreer

Arg1: Proposition

Arg2: Other entity agreeing

Ex1: [Arg0 The group] agreed [Arg1 it wouldn't make an offer].

Ex2: [ArgM-TMP Usually] [Arg0 John] agrees [Arg2 with Mary] [Arg1 on everything].

(19.12) **fall.01**

Arg1: Logical subject, patient, thing falling

Arg2: Extent, amount fallen

Arg3: start point

Arg4: end point, end state of arg1

Ex1: [Arg1 Sales] fell [Arg4 to \$25 million] [Arg3 from \$27 million].

Ex2: [Arg1 The average junk bond] fell [Arg2 by 4.2%].

Panorama of NLP tasks: Semantic Role Labelling

- **Input:** sentence / text
- **Output:** semantic roles labelled in text / sentence

Core Roles	
ATTRIBUTE	The ATTRIBUTE is a scalar property that the ITEM possesses.
DIFFERENCE	The distance by which an ITEM changes its position on the scale.
FINAL_STATE	A description that presents the ITEM's state after the change in the ATTRIBUTE's value as an independent predication.
FINAL_VALUE	The position on the scale where the ITEM ends up.
INITIAL_STATE	A description that presents the ITEM's state before the change in the ATTRIBUTE's value as an independent predication.
INITIAL_VALUE	The initial position on the scale from which the ITEM moves away.
ITEM	The entity that has a position on the scale.
VALUE_RANGE	A portion of the scale, typically identified by its end points, along which the values of the ATTRIBUTE fluctuate.
Some Non-Core Roles	
DURATION	The length of time over which the change takes place.
SPEED	The rate of change of the VALUE.
GROUP	The GROUP in which an ITEM changes the value of an ATTRIBUTE in a specified way.

Figure 19.3 The frame elements in the `change_position_on_a_scale` frame from the FrameNet Labelers Guide (Ruppenhofer et al., 2016).

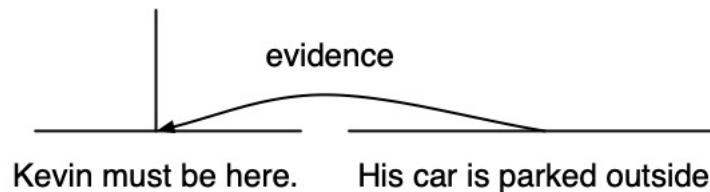
<https://web.stanford.edu/~jurafsky/slp3/19.pdf>

Panorama of NLP tasks: Coreference Resolution

- **Input:** document
- **Output:** document with resolved co-references of entities
 -) [Victoria Chen]¹_a, CFO of [Megabucks Banking]²_a, saw [[her]¹_b pay]³_a jump to \$2.3 million, as [the 38-year-old]¹_c also became [[the company]²_b]’s president. It is widely known that [she]¹_d came to [Megabucks]²_c from rival [Lotsabucks]⁴_a.
 - 1. {*Victoria Chen, her, the 38-year-old, She*}
 - 2. {*Megabucks Banking, the company, Megabucks*}
 - 3. { *her pay*}
 - 4. { *Lotsabucks*}

Panorama of NLP tasks: Discourse Coherence

- **Input:** document
- **Output:** discourse structure of the document



Panorama of NLP tasks: Discourse Coherence

- **Input:** document
- **Output:** discourse structure of the document

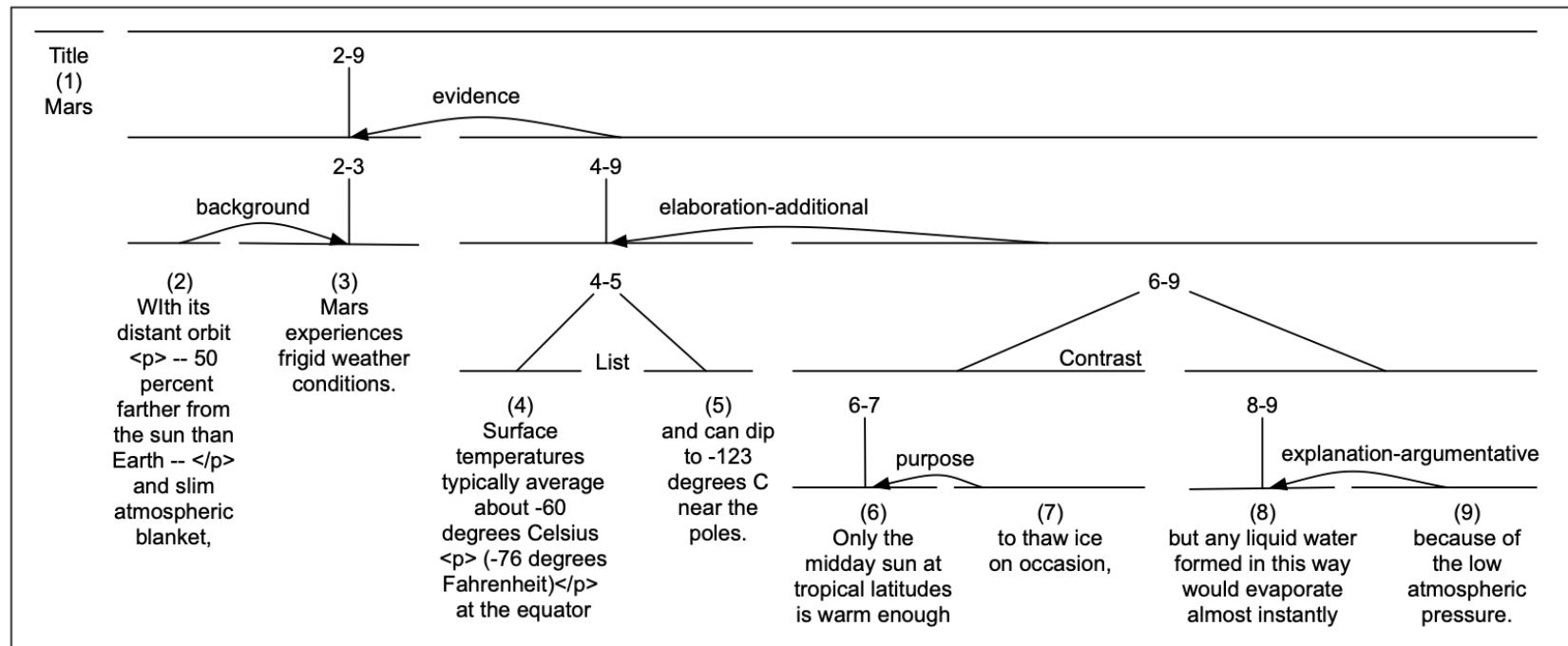
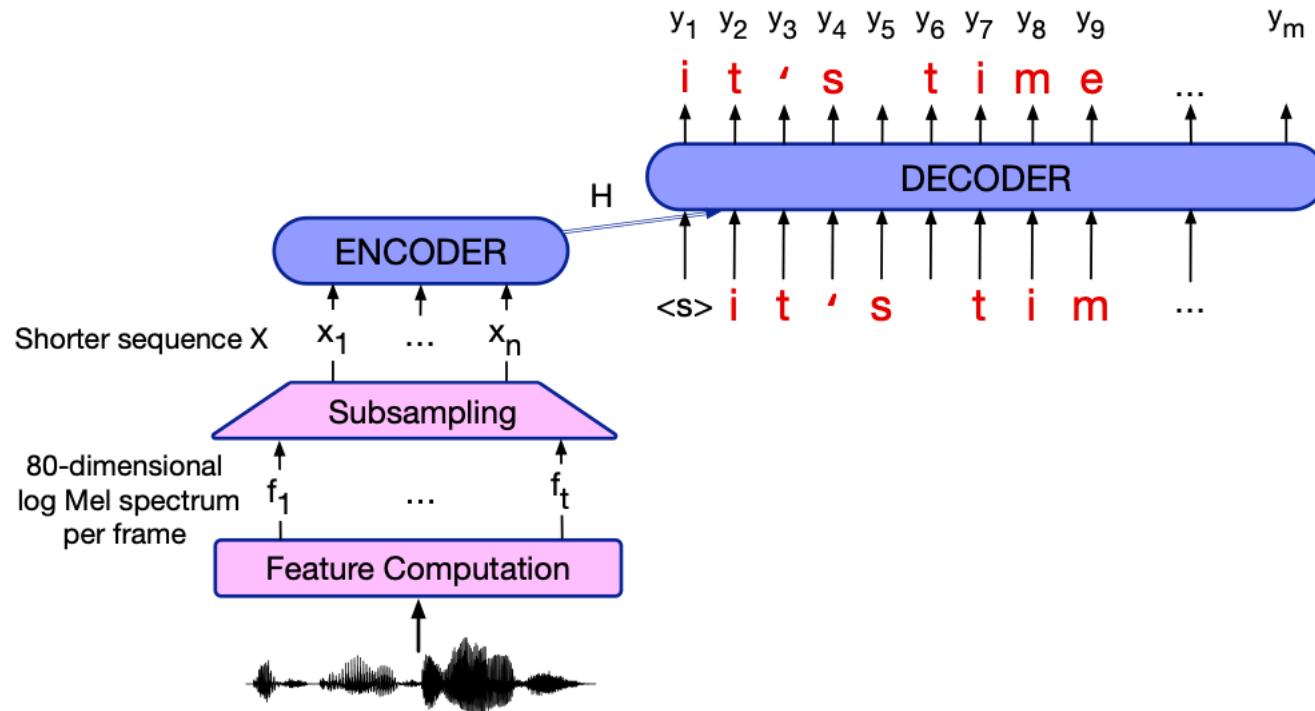


Figure 22.1 A discourse tree for the *Scientific American* text in (22.12), from Marcu (2000a). Note that asymmetric relations are represented with a curved arrow from the satellite to the nucleus.

<https://web.stanford.edu/~jurafsky/slp3/22.pdf>

Panorama of NLP tasks: ASR and Text-to-Speech

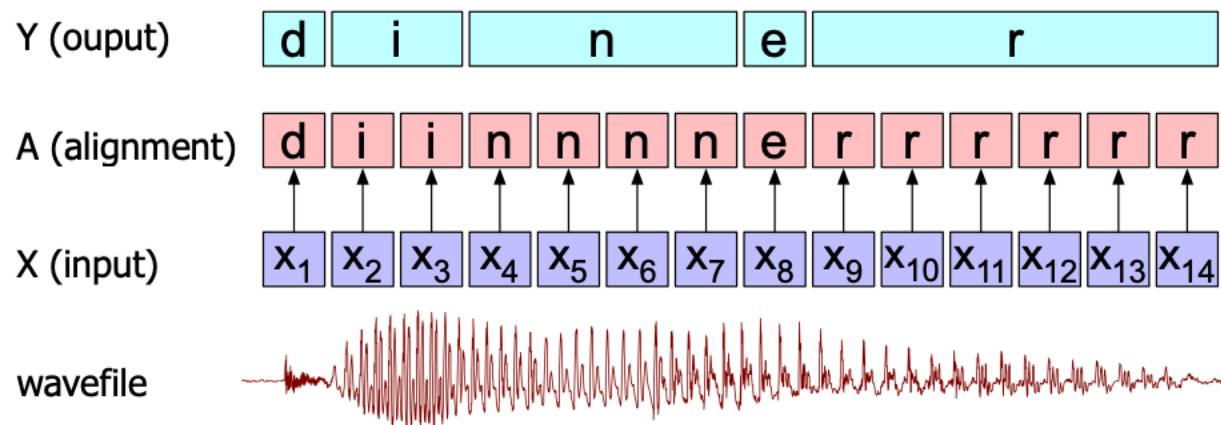
- **Input:** speech/text
- **Output:** text/speech



<https://web.stanford.edu/~jurafsky/slp3/22.pdf>

Panorama of NLP tasks: ASR and Text-to-Speech

- **Input:** speech/text
- **Output:** text/speech



Panorama of NLP tasks: Question Answering

- **Input:** question in a textual form
- **Output:** answer to the question

Beyoncé Giselle Knowles-Carter (born September 4, 1981) is an American singer, songwriter, record producer and actress. Born and raised in **Houston, Texas**, she performed in various **singing and dancing** competitions as a child, and rose to fame in the late 1990s as lead singer of R&B girl-group Destiny's Child. Managed by her father, Mathew Knowles, the group became one of the world's best-selling girl groups of all time. Their hiatus saw the release of Beyoncé's debut album, *Dangerously in Love* (**2003**), which established her as a solo artist worldwide, earned five Grammy Awards and featured the Billboard Hot 100 number-one singles "Crazy in Love" and "Baby Boy".

Q: "In what city and state did Beyoncé grow up?"

A: "**Houston, Texas**"

Q: "What areas did Beyoncé compete in when she was growing up?"

A: "**singing and dancing**"

Q: "When did Beyoncé release *Dangerously in Love*?"

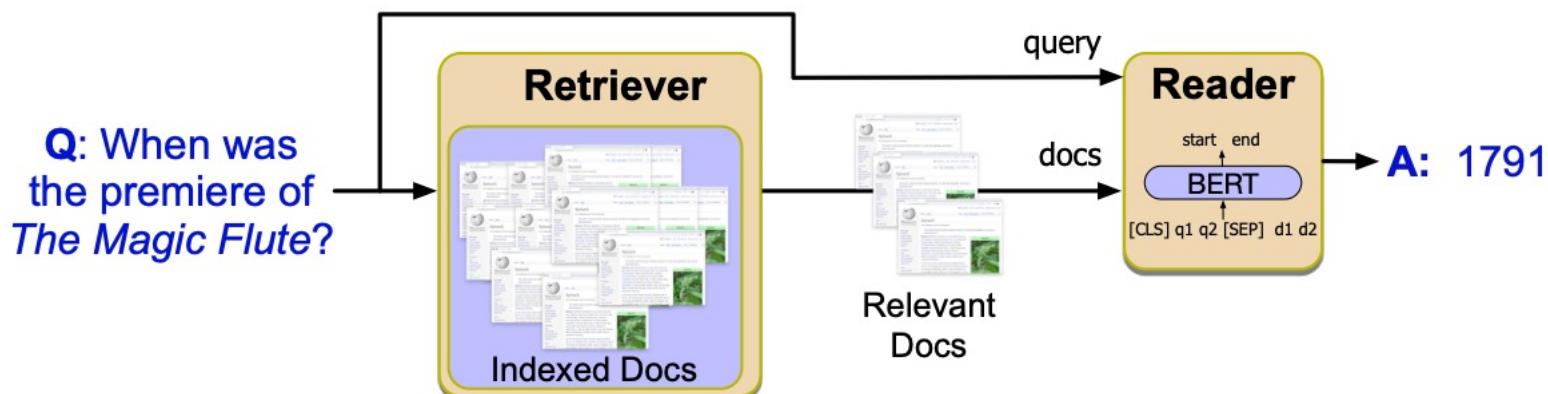
A: "**2003**"

Figure 23.11 A (Wikipedia) passage from the SQuAD 2.0 dataset (Rajpurkar et al., 2018) with 3 sample questions and the labeled answer spans.

Panorama of NLP tasks: Question Answering

- **Input:** question in a textual form
- **Output:** answer to the question

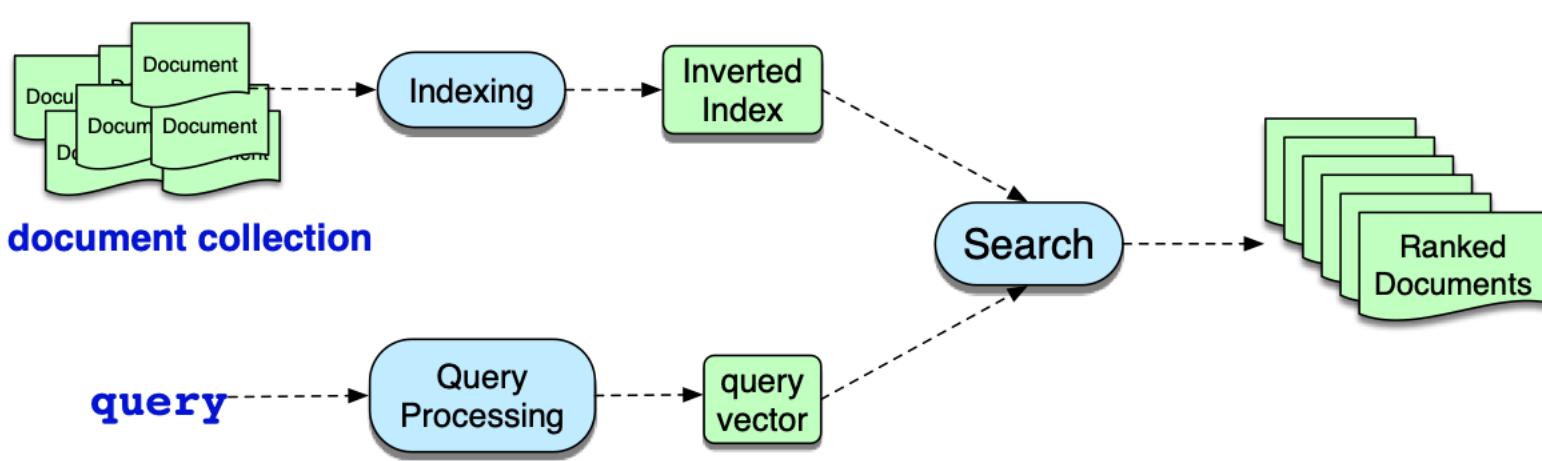
Question	Answer
Where is the Louvre Museum located?	in Paris, France
What are the names of Odin's ravens?	Huginn and Muninn
What kind of nuts are used in marzipan?	almonds
What instrument did Max Roach play?	drums
What's the official language of Algeria?	Arabic



<https://web.stanford.edu/~jurafsky/slp3/23.pdf>

Panorama of NLP tasks: Information Retrieval

- **Input:** textual query, collection of text documents
- **Output:** ranked list of documents relevant to query



Panorama of NLP tasks: Dialogue Systems / Chatbots

- **Input:** user's textual input
- **Output:** answer/reaction to the users's textual input

Will you sing me a song?

Sure, what do you want to be sung to? I can sing you a song about baking.

Yes, sing me a song about baking!

Do you want me to sing it to the tune of "Sing Along with Mitch"?

Yes, I would love that.

OK, I'll try to get it to you by the end of the day. What else do you like to do?

No, sing it to me now!

OK, give me a few days. I'm a teacher, so I don't have much free time.

Can you at least tell me what the lyrics are?

Sure, it goes like this: "Baking is my favorite thing to do in the world."

How does the chorus go?

It goes: BAKING IS MY FAVORITE THING TO DO IN THE WORLD!

Wow, sounds like a bop.

I know, right? It's so catchy. I can't get it out of my head.

C₁: ...I need to travel in May.

A₂: And, what day in May did you want to travel?

C₃: OK uh I need to be there for a meeting that's from the 12th to the 15th.

A₄: And you're flying into what city?

C₅: Seattle.

A₆: And what time would you like to leave Pittsburgh?

C₇: Uh hmm I don't think there's many options for non-stop.

A₈: Right. There's three non-stops today.

C₉: What are they?

A₁₀: The first one departs PGH at 10:00am arrives Seattle at 12:05 their time. The second flight departs PGH at 5:55pm, arrives Seattle at 8pm. And the last flight departs PGH at 8:15pm arrives Seattle at 10:28pm.

C₁₁: OK I'll take the 5ish flight on the night before on the 11th.

A₁₂: On the 11th? OK. Departing at 5:55pm arrives Seattle at 8pm, U.S. Air flight 115.

C₁₃: OK.

A₁₄: And you said returning on May 15th?

C₁₅: Uh, yeah, at the end of the day.

A₁₆: OK. There's #two non-stops ...#

C₁₇: #Act...actually #, what day of the week is the 15th?

A₁₈: It's a Friday.

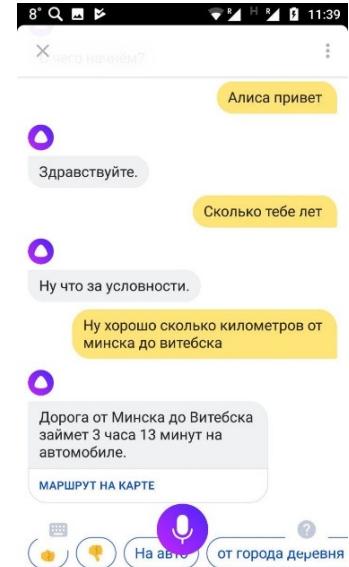
C₁₉: Uh hmm. I would consider staying there an extra day til Sunday.

A₂₀: OK...OK. On Sunday I have ...

Dialogue Systems

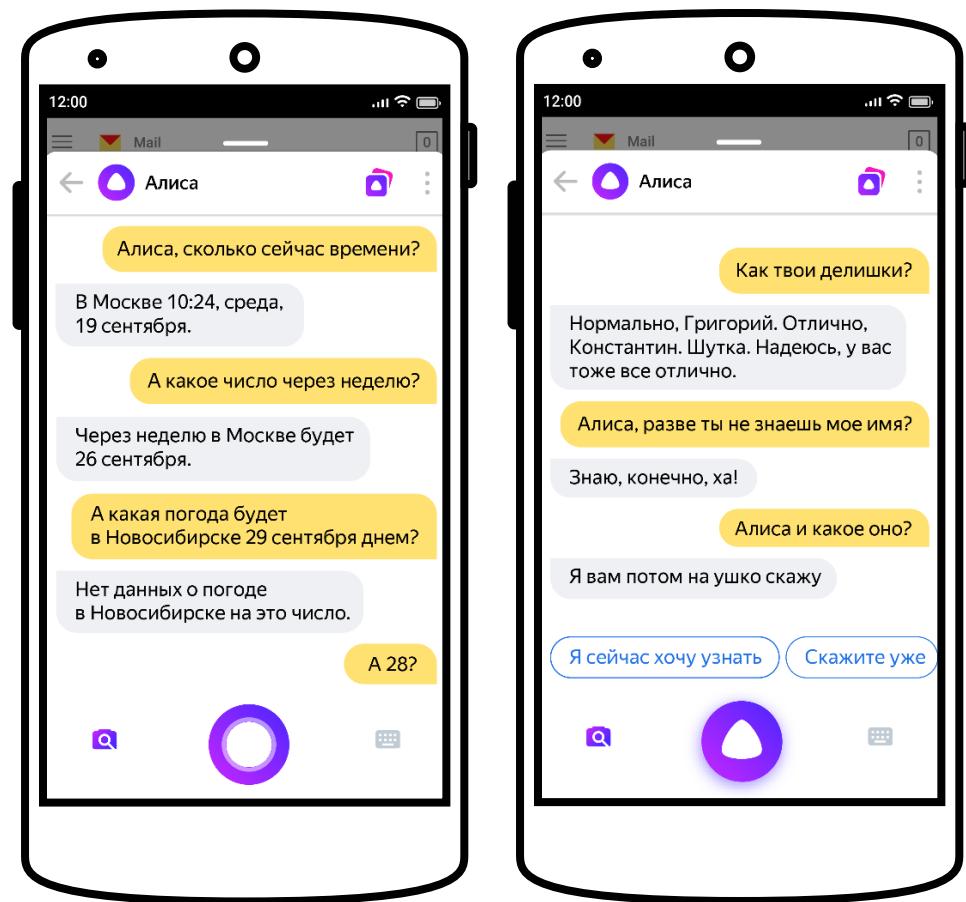
Alice and other voice assistants

- Yandex Alice, Siri, Alexa, Google Assistant etc.
all share the same principles
- Multiple platforms
 - Mobile apps: search app, navigator
 - Smart speakers
 - Desktop
 - Car head
- The goal: voice interface for everything
 - Useful scenarios from music to smart home
 - Search for information
 - Entertaining talk

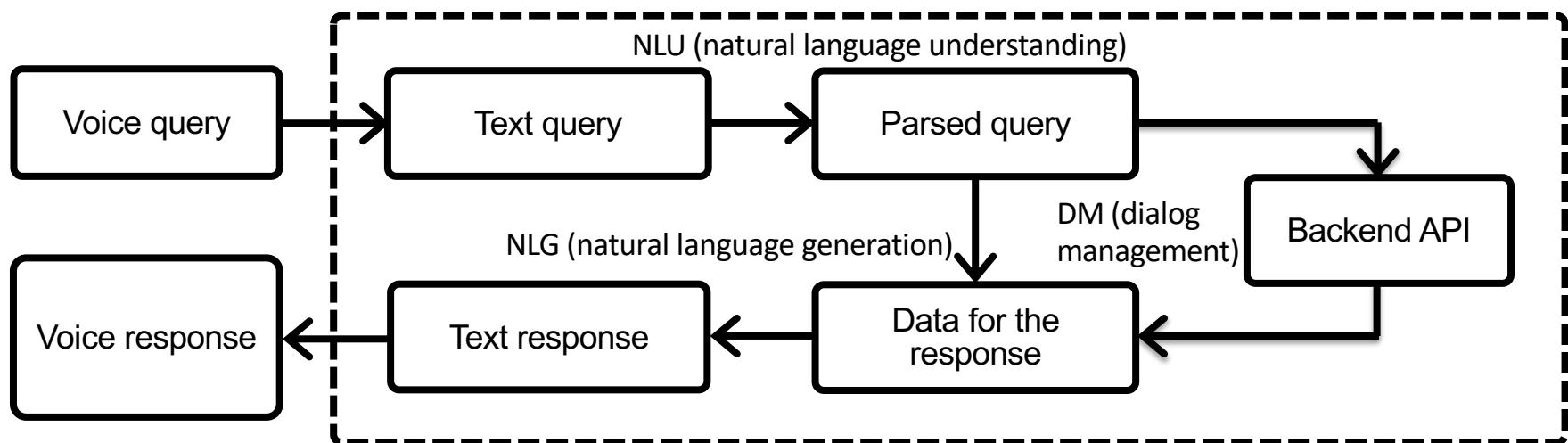


Two use cases of bots

- Goal-oriented scenarios
 - Need to understand the meaning
 - Often use external APIs
 - Help to accomplish tasks
 - Properties:
 - Accuracy
 - Brevity
- General conversation
 - Pick suitable replies
 - Only words, no actions
 - Goal: entertainments
 - Properties:
 - Relevance
 - Consistency
 - Diversity
 - Proactivity
 - Humor



Technology behind Alice

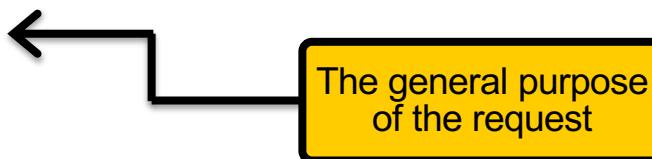


- In the backend there may be another (skill-specific) NLP engine
- NLU is often the most difficult part

Example of NLU

«Build a route to Skolkovo through Molodezhka»

```
{  
    intent: "route"  
}
```



The general purpose
of the request

Example of NLU

«Build a route to Skolkovo through Molodezhka»

```
{  
    intent: "route"  
  
    slots: {  
        "address_to": {"point_of_interest": "Skolkovo Institute"}, ←  
        "address_via": {"metro": "Molodezhnaya"}, ←  
    }  
}
```

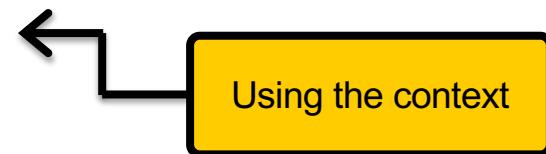


Some important entities in the request

Example of DM

«Build a route to Skolkovo through Molodezhka»

```
{  
    intent: "route"  
  
    slots: {  
  
        "address_to": {"point_of_interest": "Skolkovo Institute"},  
        "address_via": {"metro": "Molodezhnaya"},  
        "address_from": {"street": "Khoroshevskoye Avenue"},  
    }  
}
```



Example of DM

«Build a route to Skolkovo through Molodezhka»

```
{  
    intent: "route"  
  
    slots: {  
  
        "address_to": {"point_of_interest": "Skolkovo Institute"},  
        "address_via": {"metro": "Molodezhnaya"},  
        "address_from": {"street": "Khoroshevskoye Avenue"},  
        "route": {"length": 20, "duration": 38, "url": "http://yandex.ru/maps..."}  
    }  
}
```

Using external APIs

Generating the reply

«Build a route to Skolkovo through Molodezhka»

```
{  
    intent: "route"  
  
    slots: {  
  
        "address_to": {"point_of_interest": "Skolkovo Institute"},  
        "address_via": {"metro": "Molodezhnaya"},  
        "address_from": {"street": "Khoroshevskoye Avenue"},  
        "route": {"length": 20, "duration": 38, "url": "http://yandex.ru/maps..."}  
    }  
}
```

«The road will take 38 minutes»

The NLU pipeline

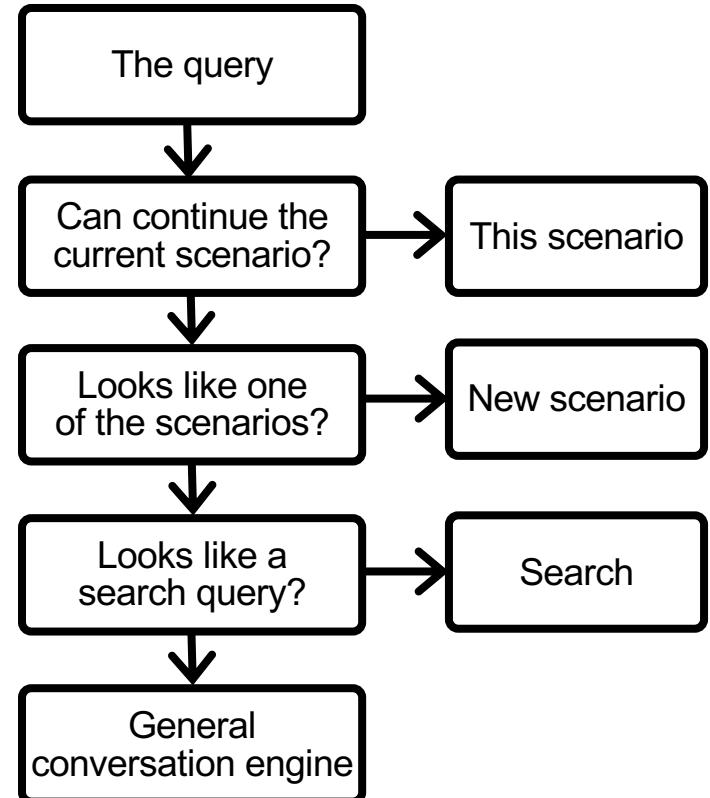
- Query normalization
- Named entity recognition
 - In Alice, it is brute force: dictionary-based, grammar-based
- Anaphora resolution
- Classification of *intents* (the main collision point)
 - Phrase templates based on grammars
 - KNN over sentence embeddings
- Tagging of *slots*
 - Neural networks or templates
- Reranking tagger hypotheses

Manageable DM and NLG

- Form-filling dialogue manager
 - Define a form based on the current scenario
 - Fill the slots with:
 - The previous form
 - Current tagger results
 - Entities in the context
 - Output of the backend API
 - Ask for the slots which are required and missing
- Template based response
 - Jinja templates + NLP-specific extensions (e.g. inflection)
 - Synonyms are sampled from the template and handcrafted ontologies

Fallbacks and transitions

- Scenario = interrelated set of intents
 - Some intents are allowed only within a scenario, other can start it
- Scenarios are selected sequentially
- What if multiple intents are likely?
 - Choose with handcrafter rules
 - Use a ML re-ranker of candidate responses from different scenarios



The general conversation model

- What do we expect from a good chit-chat model?
 - Provides a satisfactory reply in *any* context (whenever other scenarios are not applicable)
 - Filters out dangerous replies
 - Is interesting, consistent and more or less specific
- The model
 - A retrieval-based model to select candidate responses from a database
 - A transformer-based language model to generate candidate responses word by word
 - Reranking of all candidates with a gradient boosting model
- Millions of training responses
 - Open datasets (subtitles, social networks, etc.)
 - Crowdsourcing (Mechanical Turk, Toloka, etc.)
 - Own dialogues
- Result: GC (aka “Boltalka”= “chit-chat”) is a killer feature in Alice

Tell me about
the name Alice

Расскажи про имя Алиса,
пожалуйста

А что мне за это будет?

And what will
I get for it?

Conclusions

- The works on an AI assistant is difficult:
 - Understanding of diverse language
 - Anticipating non-trivial course of conversation
 - Chatting about everything
- We are very far from making the assistants *really intelligent*
 - And we often have to reinvent the wheel with each new skill
- A good assistant requires a lot of science and engineering

Reading

- *Speech and Language Processing* by Jurafsky and Martin
 - The chapter “[Chatbots and Dialogue Systems](#)”
- Some posts about Alice (in Russian):
 - <https://habr.com/ru/company/yandex/blog/333912/>
 - <https://habr.com/ru/company/yandex/blog/339638/>
 - <https://habr.com/ru/company/yandex/blog/349372/>
- Some chatbot-related papers:
 - <https://arxiv.org/abs/1911.00536> - DialogGPT by Microsoft
 - <https://arxiv.org/abs/2004.13637> - Blenderbot by Facebook
 - <https://arxiv.org/abs/2001.09977> - Meena bot by Google