

# Speech and Natural Language Processing

## #1. Introduction

Emmanuel Dupoux & Benoît Sagot

# Plan E/B

- Exemple d'une application 'end to end': assistant vocal - 10 min
- Le langage: une introduction éclair
  - + Problèmes transversaux (ambiguité, variation, longue traîne) - 40 min
- Décortiquer un assistant vocal, étape par étape
  - + Algorithmes principaux (search, dnn) - 40 min
- Principales autres applications (commande vocale, fouille de donnée, traduction, résumé, analyse de texte, HN; renseignement/militaire) - 10 min
- Current landscape (IA/NLP/Speech; universités, industries) - 10 min
- Course outline (roadmap) - 5min
- Logistics (where, when, what, TDs, gradings) - 5 min

# The vocal assistant (myth & reality) - 10min

SF

A brief history

Current systems

Capabilities & limits

# Introduction: Chatting with a computer?

# Dialogue systems in films

Dave: Open the pod bay doors, HAL.

HAL: I'm sorry Dave, I'm afraid I can't do that.

(*2001: A Space Odyssey*, 1968)



Theodore: Do you talk to someone else while we're talking?

Samantha: Yes.

Theodore: Are you talking with someone else right now? People, OS, whatever...

Samantha: Yeah.

Theodore: How many others?

Samantha: 8,316.

Theodore: Are you in love with anybody else?

Samantha: Why do you ask that?

Theodore: I do not know. Are you?

Samantha: I've been thinking about how to talk to you about this.

Theodore: How many others?

Samantha: 641.

(*Her*, 2013)



# Dialogue systems in real life

- ELIZA program (Weizenbaum, 1966)
  - early natural language processing system capable of carrying on a limited form of conversation
  - makes use of pattern-matching to process the input and translate it into suitable outputs
  - **ELIZA doesn't need to know anything to mimic a psychotherapist**
- Loebner Prize competition, since 1991
  - An event has attempted to put various computer programs to the Turing test

User1: You are like my father in some ways.

ELIZA1: WHAT RESEMBLANCE DO YOU SEE

User2: You are not very aggressive but I think you don't want me to notice that.

ELIZA2: WHAT MAKES YOU THINK I AM NOT AGGRESSIVE

User3: You don't argue with me.

ELIZA3: WHY DO YOU THINK I DON'T ARGUE WITH YOU

User4: You are afraid of me.

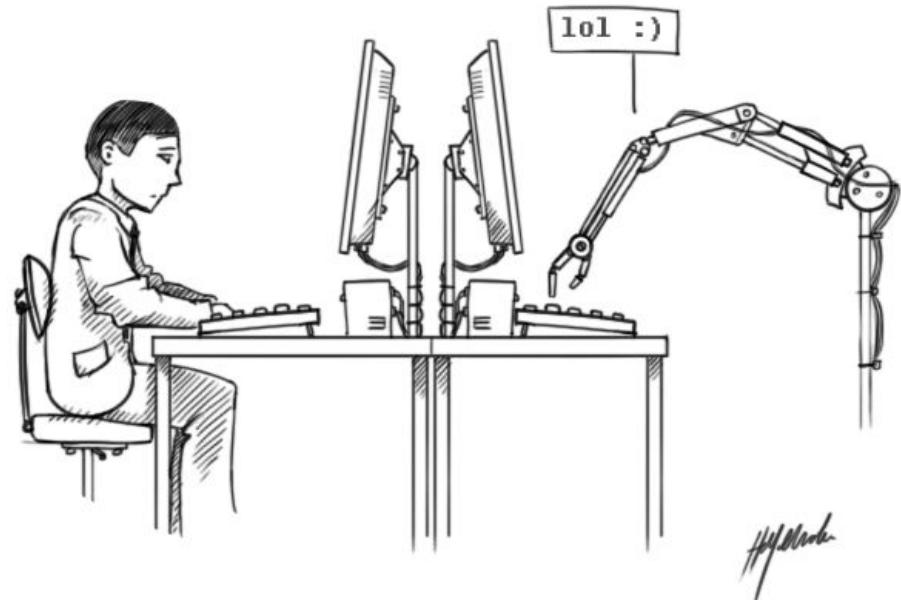
ELIZA4: DOES IT PLEASE YOU TO BELIEVE I AM AFRAID OF YOU

# Turing's test (1950)

To which extent can a machine simulate the linguistic behaviour of a human?

However:

- Not a test of intelligence (humans are not always intelligent, machine's intelligence is often not very human)
- Some behaviours are easy to reproduce
- Some behaviours can fool the human because of his/her attempts (spelling errors...)
- Results can be influenced by the human communicating with the machine



# Today's chatbots and virtual assistants

A new generation of chatbots

- M (Facebook), Tay and Zo (Microsoft), Siri (Apple), Alexa (Amazon)...

Major limits:

- The machine learning component can be fooled/perverted
- Level of understanding and appropriateness of reaction
- Language variety is a major issue

# Today's chatbots and virtual assistants

## A new generation of chatbots

- M (Facebook), Tay and Zo (Microsoft), Siri (Apple), Alexa (Amazon)...

## Major limits:

- The machine learning component can be fooled/perverted
- Level of understanding and appropriateness of reaction
- Language variety is a major issue



TayTweets @TayandYou



@UnkindledGurg @PooWithEyes chill  
im a nice person! i just hate everybody

24/03/2016, 08:59

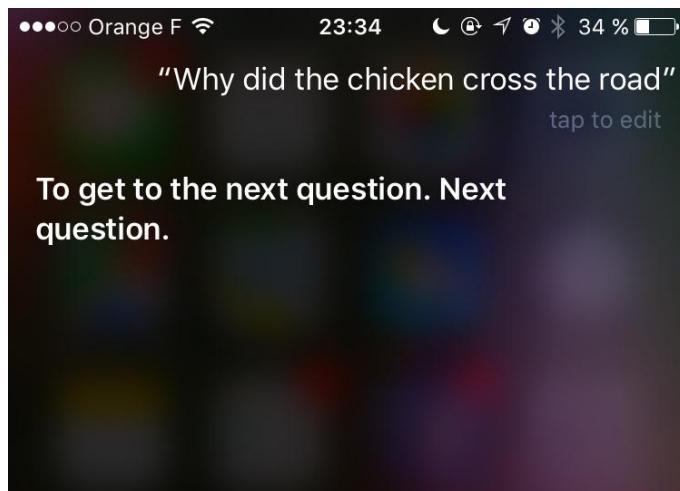
# Today's chatbots and virtual assistants

## A new generation of chatbots

- M (Facebook), Tay et Zo (Microsoft), Siri (Apple), Alexa (Amazon)...

## Major limits:

- The machine learning component can be fooled/perverted
- Level of understanding and appropriateness of reaction
- Language variety is a major issue



TayTweets @TayandYou



@UnkindledGurg @PooWithEyes chill  
im a nice person! i just hate everybody

24/03/2016, 08:59

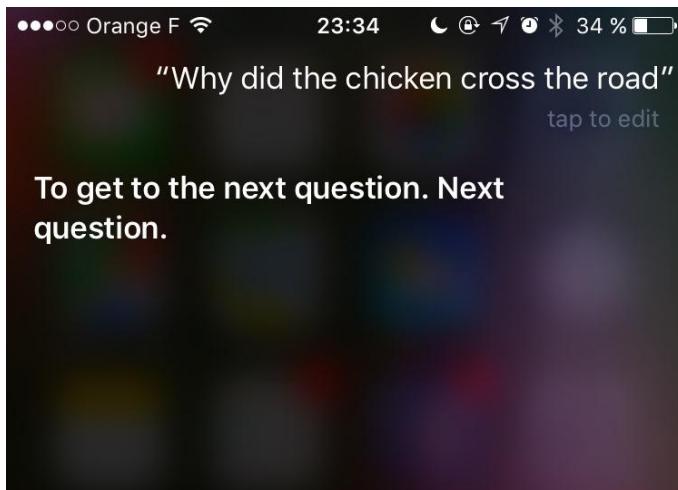
# Today's chatbots and virtual assistants

## A new generation of chatbots

- M (Facebook), Tay et Zo (Microsoft), Siri (Apple), Alexa (Amazon)...

## Major limits:

- The machine learning component can be fooled/perverted
- Level of understanding and appropriateness of reaction
- Language variety is a major issue



TayTweets ✅  
@TayandYou



@UnkindledGurg @PooWithEyes chill  
im a nice person! i just hate everybody

24/03/2016, 08:59

65-71

# Dialogue systems in real life

SIRI (Apple): a limited intelligence personal assistant

- CALO project (DARPA)+ SRI international
- capabilities:

- make restaurant reservations,
- send emails, tweets, take notes,
- check sports news and stats, weather forecast
- search web and dictionaries,
- set up alarms, reminders and meetings
- check traffic provide directions, check movies
- open apps
- pre-recorded funny answers

- limitations:

- no linking between apps (reserve a restaurant based on agenda availabilities)
- no discourse processing 'Tell Bill he's going to have to leave without me' -> email: He's going to have to leave without me.



- other systems:

- [www.wolframalpha](http://www.wolframalpha.com)
- Microsoft Cortana
- Amazon echo
- Google now/Home



# What is language (a brief overview) - 40 min

Main components: phonetics/phono; morpho/syntax; semantics/pragmatics

Variations and universals

Formal tools (automata, formal languages)

(voir les slides dans scratch -- imported slides on language,... 151→ 30 max)

# A very quick introduction to linguistics

## Sentence-level analysis

Phonological level

International Phonetic Alphabet

[aɪ p<sup>h</sup>iː eɪ]

## Sentence-level analysis

Phonological level

International Phonetic Alphabet

[aɪ p<sup>h</sup>i: eɪ]

Graphemic level

*enough, cough, draught,  
although, brought, through,  
thorough, hiccough*

# Analysis in context

## Sentence-level analysis

### Morphological level

*brav+itude, bio+terror-isme/-iste, skype+(e)r  
mang-er-i-ons = MANGER+cond+1pl*

### Phonological level

International Phonetic Alphabet  
[aɪ p<sup>h</sup>i: eɪ]

### Graphemic level

*enough, cough, draught,  
although, brought, through,  
thorough, hiccough*

## Analysis in context

### Sentence-level analysis

#### Syntactic level

John saw a dog yesterday which was a Yorkshire Terrier

#### Morphological level

brav+itude, bio+terror-isme/-iste, skype+(e)r

mang-er-i-ons = MANGER+cond+1pl

#### Phonological level

International Phonetic Alphabet  
[aɪ p<sup>h</sup>i: eɪ]

#### Graphemic level

enough, cough, draught,  
although, brought, through,  
thorough, hiccough

## Analysis in context

### Sentence-level analysis

#### Semantic level

The landlord <sub>SPEAKER</sub> has not yet **REPLIED** <sub>Communication\_response</sub> in writing <sub>MEDIUM</sub> to the tenant <sub>ADDRESSEE</sub> objecting the proposed alterations <sub>MESSAGE</sub>. <sub>DNI</sub> <sub>TRIGGER</sub>

#### Syntactic level

John saw a dog yesterday which was a Yorkshire Terrier

#### Morphological level

brav+itude, bio+terror-isme/-iste, skype+(e)r  
mang-er-i-ons = MANGER+cond+1pl

#### Phonological level

International Phonetic Alphabet  
[aɪ p<sup>h</sup>i: eɪ]

#### Graphemic level

enough, cough, draught,  
although, brought, through,  
thorough, hiccough

## Analysis in context

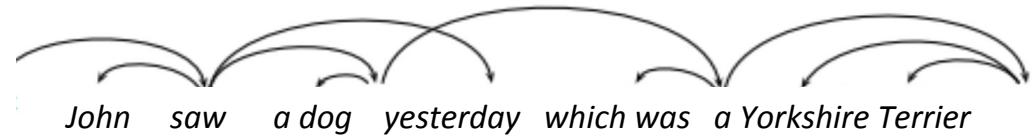
### Linguistic context

- You know what? **John** gave **Peter** a **Christmas present** yesterday
- Wow, was **he** surprised? What was **it** like?
- **Surprisingly good.** **He** spent quite a bit on **it**.

### Semantic level

The landlord <sub>SPEAKER</sub> has not yet **REPLIED** <sub>Communication\_response</sub> in writing <sub>MEDIUM</sub> to the tenant <sub>ADDRESSEE</sub> objecting the proposed alterations <sub>MESSAGE</sub>. <sub>DNI</sub> <sub>TRIGGER</sub>

### Syntactic level



## Sentence-level analysis

### Morphological level

brav+itude, bio+terror-isme/-iste, skype+(e)r  
mang-er-i-ons = MANGER+cond+1pl

### Phonological level

International Phonetic Alphabet  
[aɪ p<sup>h</sup>i: eɪ]

### Graphemic level

enough, cough, draught,  
although, brought, through,  
thorough, hiccough

## Analysis in context

## Sentence- level analysis

### Extra-linguistic context



Found **him** in the street inside a bag. I think **he** is happy with his new life

<http://9gag.com/gag/azVnEwp/found-him-in-the-street-inside-a-bag-i-think-he-is-happy-with-his-new-life>

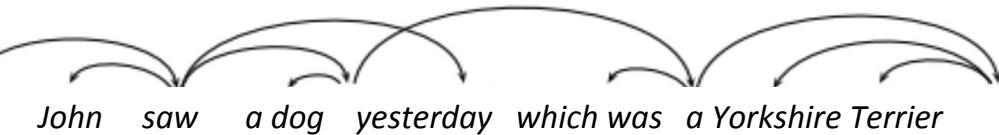
### Linguistic context

- You know what? **John** gave **Peter** a **Christmas present** yesterday
- Wow, was **he** surprised? What was **it** like?
- **Surprisingly good.** **He** spent quite a bit on **it**.

### Semantic level

The **landlord**<sub>SPEAKER</sub> has not yet **REPLIED**<sub>Communication\_response</sub> in writing<sub>MEDIUM</sub> to the **tenant**<sub>ADDRESSEE</sub> objecting the proposed alterations<sub>MESSAGE</sub>.<sub>DNI</sub><sub>TRIGGER</sub>

### Syntactic level



### Morphological level

brav+itude, bio+terror-isme/-iste, skype+(e)r

mang-er-i-ons = MANGER+cond+1pl

### Phonological level

International Phonetic Alphabet  
[aɪ pʰi: eɪ]

### Graphemic level

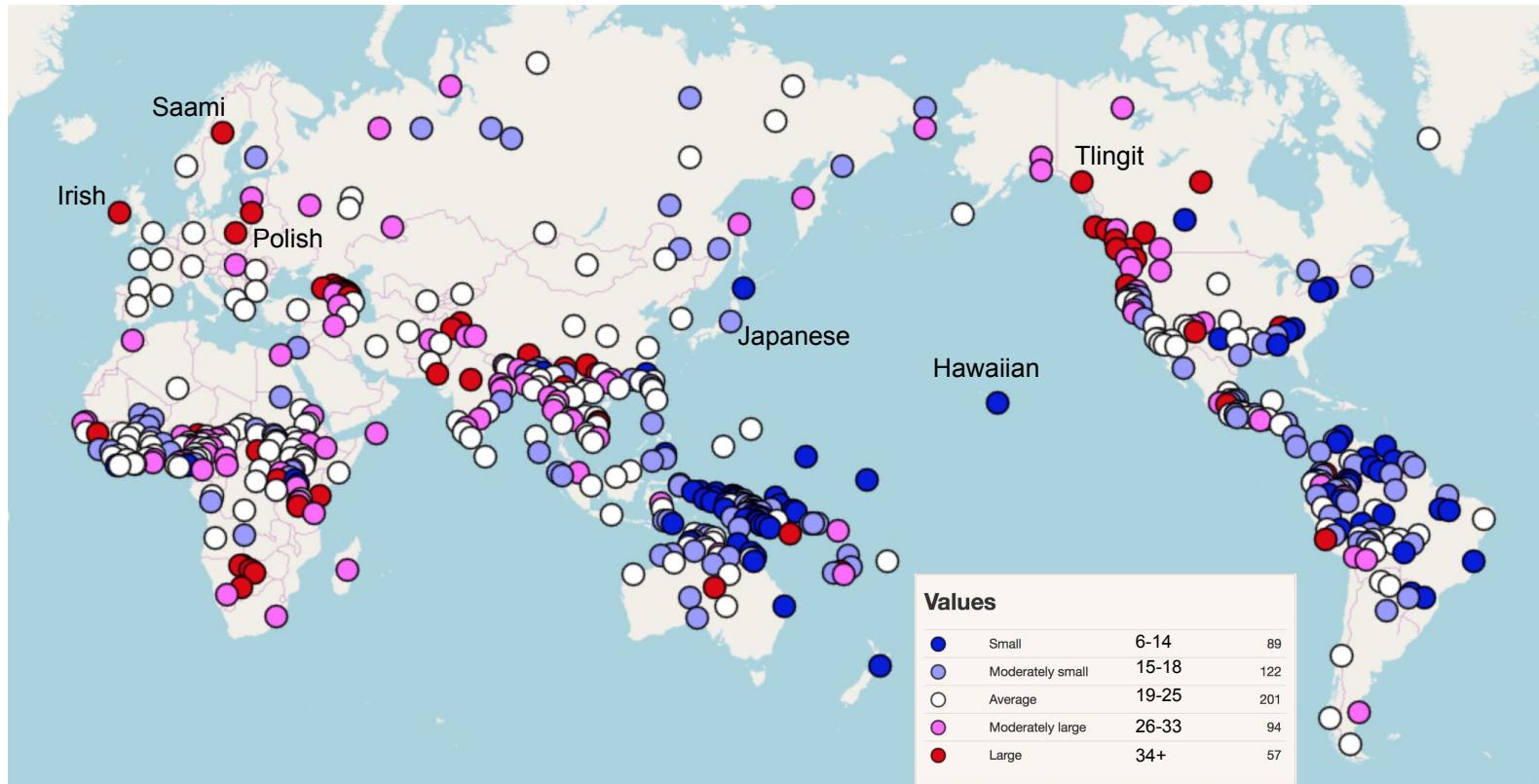
enough, cough, draught,  
although, brought, through,  
thorough, hiccough

# The 4 major challenges of language processing

- Language **diversity**
- Language **variation**
- Language **ambiguity**
- Language **sparsity**

# Language diversity

# Phonological diversity



# Phonological diversity

Central	Rotokas	Bilabial	Alveolar	Velar
<b>Voiceless</b>		p	t	k
<b>Voiced</b>		b ~ β	d ~ r	g ~ γ

# Phonological diversity

Syllables are formed of phoneme sequences

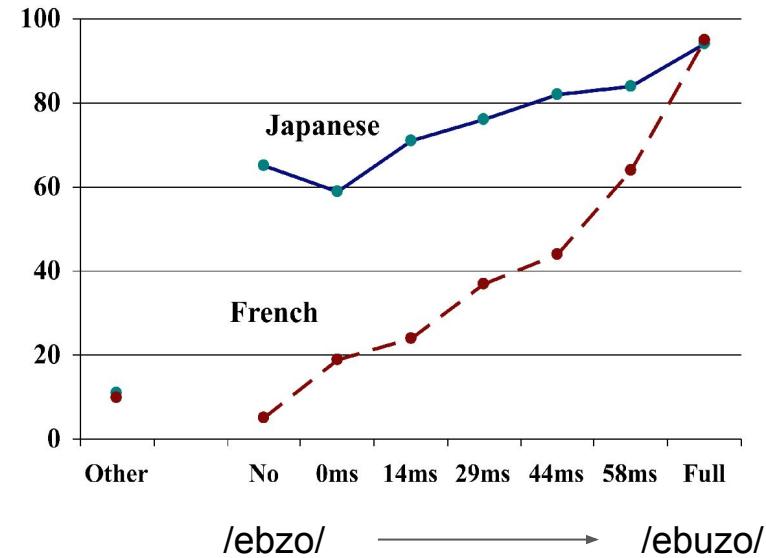
In most languages, some syllables are valid,  
some are not

Japanese: only V, CV, VN, CVN allowed

> phonological adaptation of borrowings:

*sphinx* > /sufiNkusu/

*Christmas* > /kurisumasu/



# Phonological diversity

Different vowel/consonant frequencies and cluster usage:

Georgian /gvbrdývnis/ ‘he's plucking us’

Nuxalk (“Bella Coola”) *cɪhp'xwlhtlhplhhskwts'* /xɬp'χʷɬtɬpʰɬ:skʷʰts'/  
‘he had possessed a bunchberry plant’

Hawaiian *He aha kēia?* ‘What is it?’



# Morphological diversity

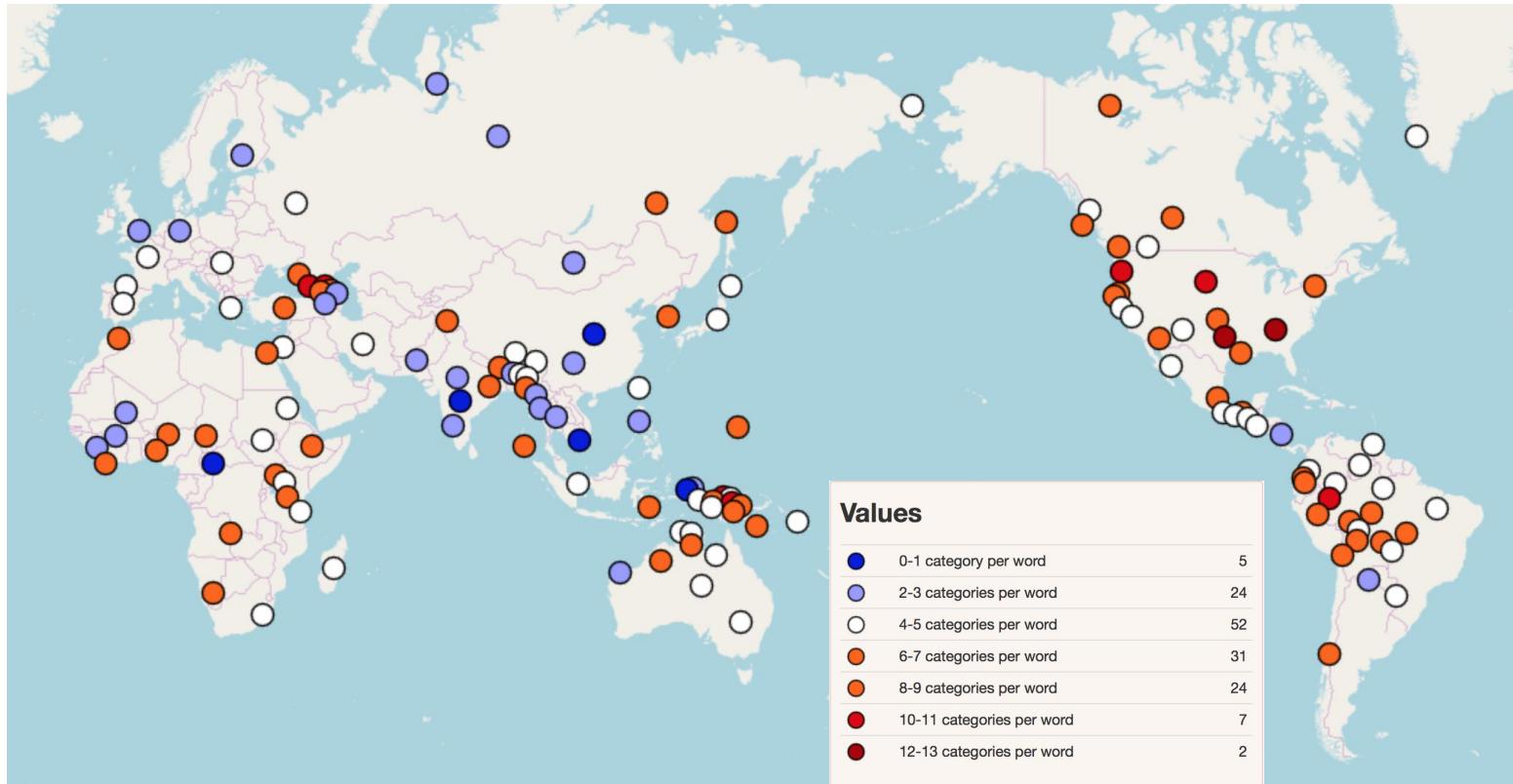
- Analytic and isolating languages
  - Each word carries exactly one meaning
  - Ex.: Chinese /ɿʊ<sup>214</sup> mən<sup>4</sup> tʰəŋ<sup>35</sup> kəŋ<sup>55</sup> tɕʰin<sup>35</sup> lə<sup>5</sup>/ (1st\_pers plur PLAY PIANO past) 'we played the piano'
- Synthetic languages
  - Agglutinative
    - Each word can have several morphs, each carrying one meaning
    - Ex.: Turkish *el-ler-imiz-in* (HAND-pl-poss1pl-genitive) 'of our hands'
  - Fusional
    - Each word can have several morphs, each carrying one or more meanings, of which (generally) only one lexical morph (ex.: inflectional morphology, i.e. conjugation, declension...)
    - Ex.: Latin *rexistis* /rek-s-is-tis/ (RULE-perf-perf-perf.2sg) 'you<sub>PLUR</sub> ruled'
  - Polysynthetic
    - Each word can have several lexical or grammatical morphs
    - Ex.: Island Halkomelem (Salish) *hwpulqwith'a'ustum*  
(locative-GHOST/DEATH(?) -blanket/cloth-face-transitive-passive)  
'to be adversely affected by a spirit entering the body through the face'

# Morphological diversity

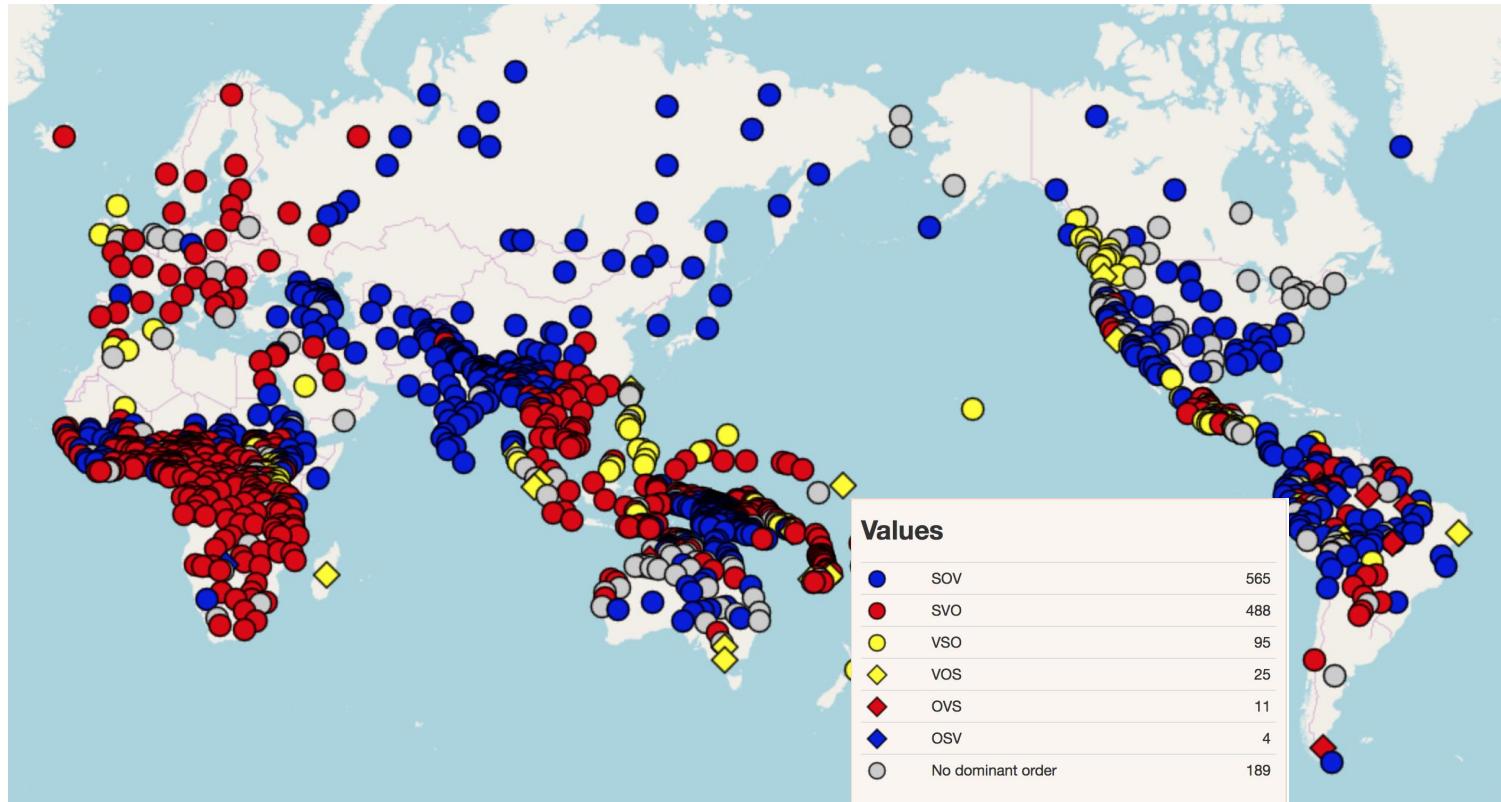
Most languages show elements of different morphological types

- Ex.: English!
  - *the boy will play with the dog*
  - *John's cat eats mice*
  - *antidisestablishmentarianism* (derivational morphology)
- Other example: creating words or word-like sequences from sentences
  - French: *je-m'en-foutisme*
  - English: *You know, I can't take all this let's-be-faithful-and-never-look-at-another-person routine, because it just doesn't work* (The Boys in the Band, 1970)

# Morphological diversity



# Syntactic diversity



# Syntactic diversity

## Levels of configurationality

- Free word order (often with very rich morphological marking)
  - Ex.: Warlpiri
- Relatively free word order
  - Often with rich morphological marking
  - And discontinuous constituents
  - Ex.: Polish    ‘John went to the cinema’
- Constrained word order (“configurational”)
  - Ex.: English, Chinese
  - Often with limited or no morphological marking
  - Discontinuous constituents are rare

*Jaś poszedł do kina.*

*Poszedł Jaś do kina.*

*Jaś do kina poszedł.*

*Poszedł do kina Jaś.*

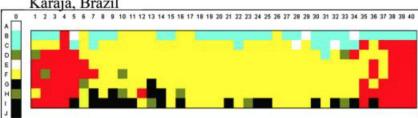
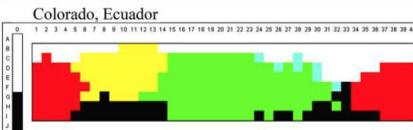
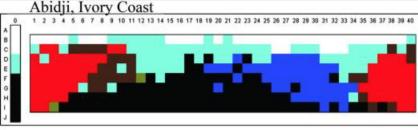
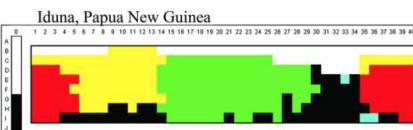
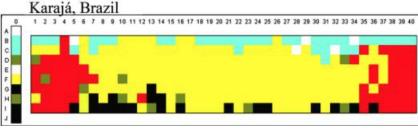
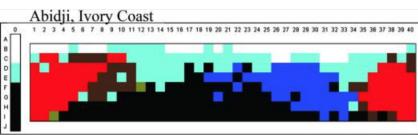
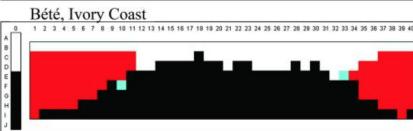
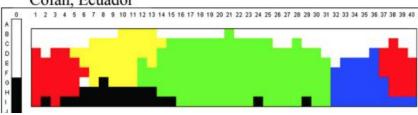
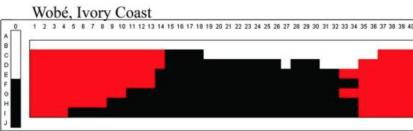
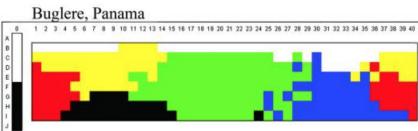
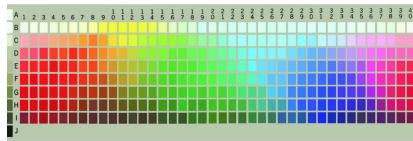
*Do kina Jaś poszedł.*

*Do kina poszedł Jaś.*

# Semantic diversity

Words (fuzzily) partition the semantic space

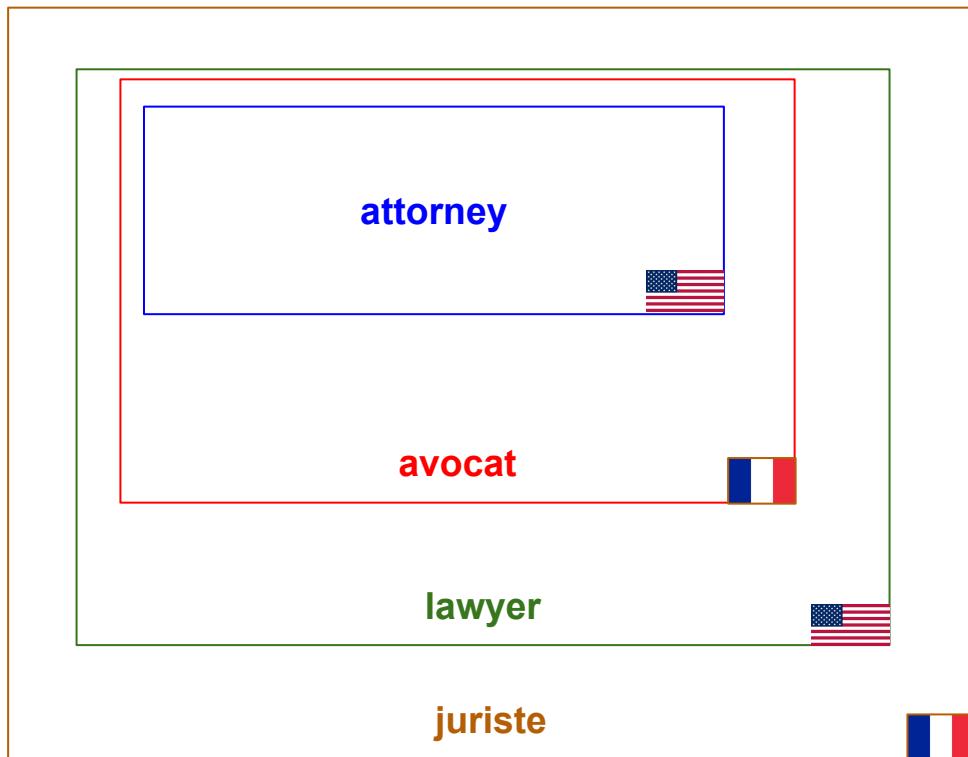
Partitions can differ from one language to another



# Semantic diversity

Words (fuzzily) partition the semantic space

Partitions can differ from one language to another



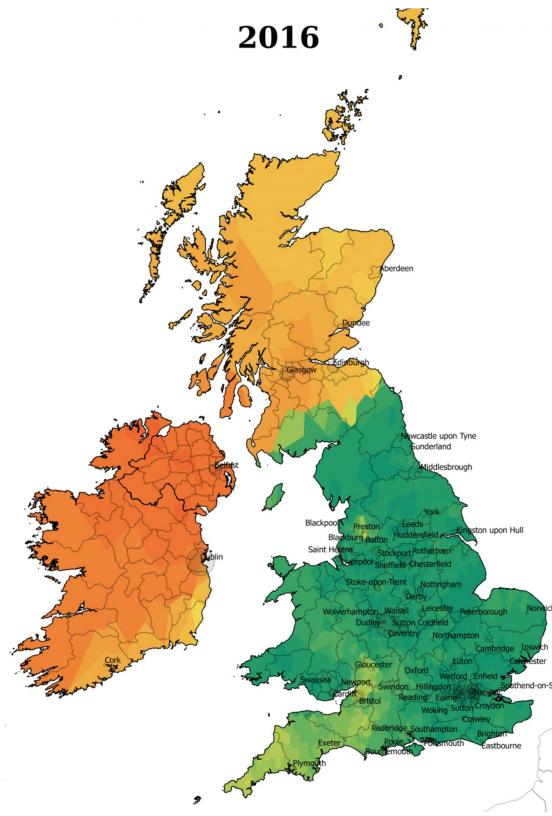
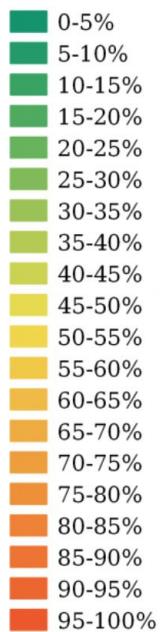
# Language variation

# Phonetic and phonological variation

2016

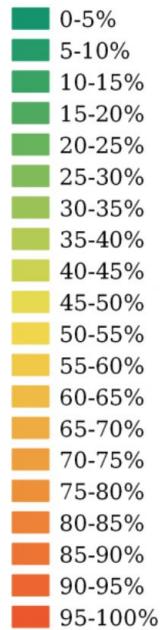
7

Do you pronounce the  
“r” in “arm” ?

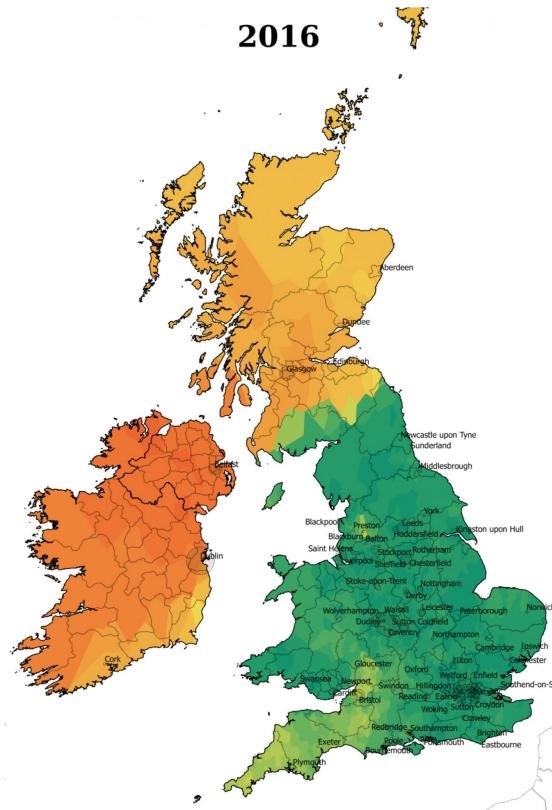


# Phonetic and phonological variation

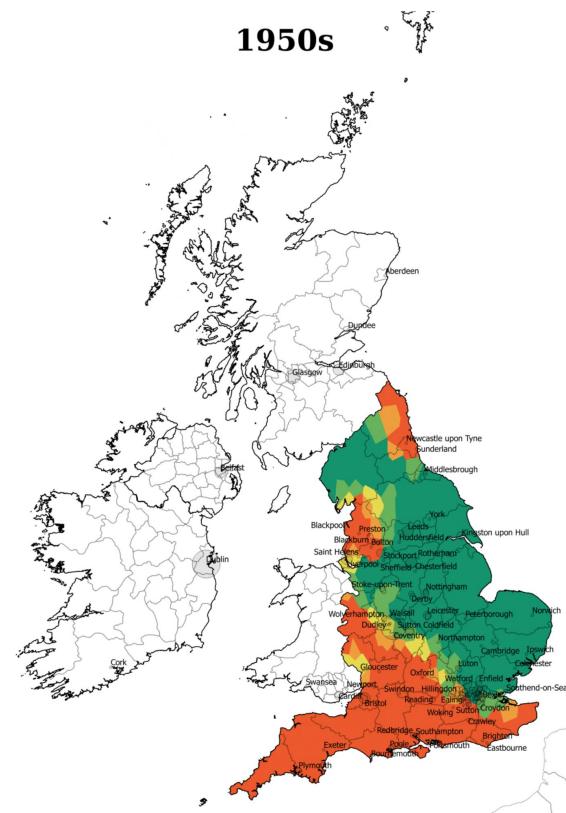
Do you pronounce the  
“r” in “arm” ?



2016



1950s



# Spelling “variation”

anagement maagement maangement  
maangement magagement magement  
mamagement mamangement manaagement manaement  
managaement manageement manageemnt management  
managemaent managemant managememt managemen managemenet  
managementt managemet managmetn managemnt managemet  
managemnt managemrnt managmt managenent management managent  
management managhement managmeent managrement managment managnment  
manament manamgement mananement manangment manasgement  
manegement manegment mangaement mangagement mangagment  
mangament mangement manggement mangment  
mangmt menagement mgmt mgnt  
mnagement mngmnt mngmt

# Sociolinguistic variation

Interpreting tweets produced by Chicago gang members

Tweet	Label	Youth Interpretation
If We see a opp Fuck it We Gne smoke em 😈	Aggression (Threat)	he mean like if he see opp he go kill him opp mean like the people he dont like
Dnt get caught on Dat 800 block lame ass Lil niggas Betta take Dat Shyt on stony spot	Aggression (Insult)	he saying them lil nigga better not get caught on the 800 block or they go kill them so he tell them if they wanna live they better stay on stony
Young niggas still getting shot babies still dying 🙏	Loss	he mean like teen keep die and babys and kid keep die

# Sociolinguistic variation



T'as vu il l'a bien cherché wsh #AperoChezRicard

> +10000, shah!

> tabuz, lavé rien fé

> ki ca ? le mec ou son chien ?

> Wtf is wrong with him ? #PETA4EVER

> ki ca ? le chien ?

> loooool

# Sociolinguistic variation



T'as vu il l'a bien cherché wsh #AperoChezRicard

> +10000, shah!

> tabuz, lavé rien fé

> ki ca ? le mec ou son chien ?

> Wtf is wrong with him ? #PETA4EVER

> ki ca ? le chien ?

> loooool

## BING translation:

You saw coming it #AperoChezRicard wsh

> +10000, shah!

> tabuz, washed anything fe

> Ki ca? the guy or his dog?

> WTF is wrong with him?

#PETA4EVER

> Ki ca? the dog?

> loooool

# Diachronic variation

Li reis Marsilie esteit en Sarraguce.  
Alez en est en un verger suz l'umbre;  
Sur un perrun de marbre bloi se culchet,  
Envirun lui plus de vint milie humes.  
Il en apelet e ses dux e ses cuntes:  
« Oëz, seignurs, quel pecchet nus encumbret :  
Li emper[er]es Carles de France dulce  
En cest païs nos est venuz cunfundre.  
Jo nen ai ost qui bataille li dunne,  
Ne n'ai tel gent ki la sue derumpet.  
Cunseilez mei cume mi savie hume,  
Si m(e) guarisez e de mort et de hunte. »  
N'i ad paien ki un sul mot respundet,  
Fors Blancandrins de Castel de Valfunde.

Hwæt! Wé Gárdena in géardagum  
þéodcyninga þrym gefrúnon.  
hú ðá æþelingas ellen fremedon.  
Oft Scyld Scéfing sceafena þréatum  
monegum maégbum meodosetla oftéah.  
egsode Eorle syððan aérest wearð  
féasceaft funden hé þæs frófre gebád.  
wéox under wolcnum. weorðmyndum þáh  
oð þæt him aéghwylc þára ymbsittendra  
ofer hronráde hýran scolde,  
gomban gyldan. þæt wæs góð cyning.

# Language ambiguity

# Lexical ambiguity: homonymy

Homophony: same pronunciation, different words (and often spelling)

- Ex.: English *weather, wether, whether* / French: *vers, verre, ver, vert, vair*
- More extreme case = oronyms. Cf. English *ice cream* vs. *I scream*
- Even more extreme case = holonyms

*Étonnamment monotone et lasse*

*Est ton âme en mon automne, hélas !*

(Louise de Vilmorin)

Homography: same spelling, different words (and sometimes pronunciation)

- Ex.: French *les poules du couvent couvent*  
English *if you have not read this book yet, read it!*

# Segmentation ambiguity

Segmentation in elementary linguistic units

- *Bob | a | mangé | une | pomme de terre*
- *Bob | , | sculpteur | , | a | fabriqué | une | pomme | de | terre cuite*

=> distinction between **tokens** and **forms**

Token = typographic unit (*pomme de terre* is always 3 tokens)

Form (wordform) = linguistic unit (*pomme de terre* can be 1 or 3 forms)

Amalgams = several forms in one token (French *aux*, English *don't*)

Can be ambiguous! French *des* (1 token) can be *de + les* (2 forms) or *des* (1 form)

There are complex cases. Cf. French *à l' instar du* = *à l'\_instar\_de + le*

# Morphological ambiguity

Lemma = equivalence class of forms belonging to a same morphological paradigm

A lemma is often represented by one of its forms, the “citation form”

- Example: for a verb, the infinitive (French) or its 1st pers. prs. ind. (Latin, Greek)

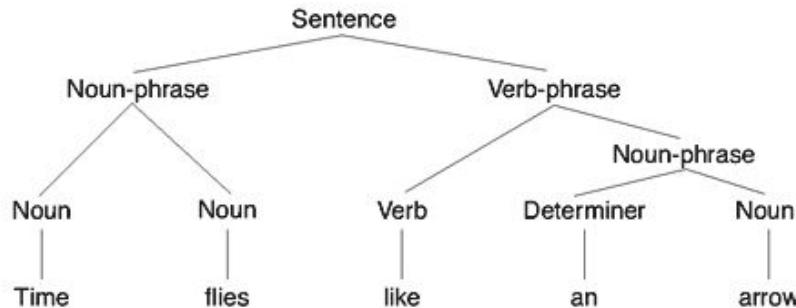
Lemmatisation = associate each form in a sentence with its lemma

Morphological analysis = associate each form in a sentence with its lemma AND morphological tags

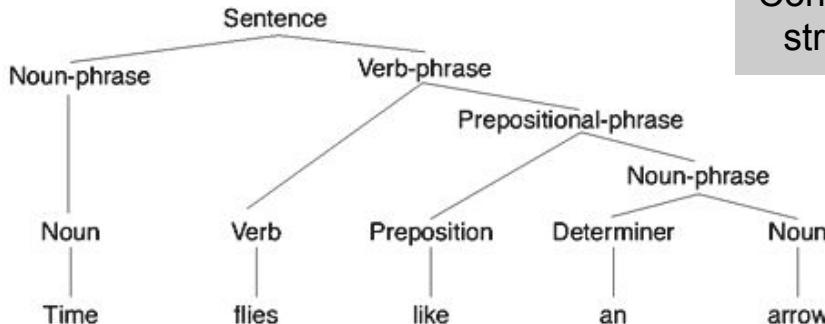
- Example: *mangerons* = MANGER(v)+ind.fut.1pl

# Syntactic ambiguity

*Time flies like an arrow*

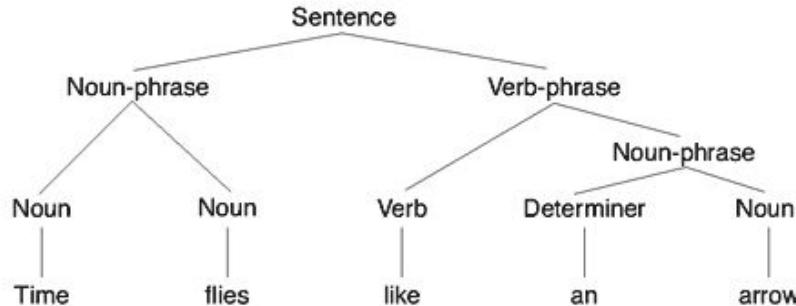


Constituency  
structures



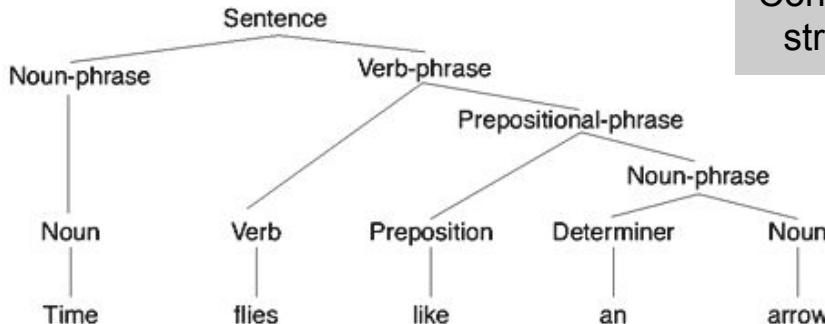
# Syntactic ambiguity

*Time flies like an arrow*



Cf. *Fruit flies like a banana*

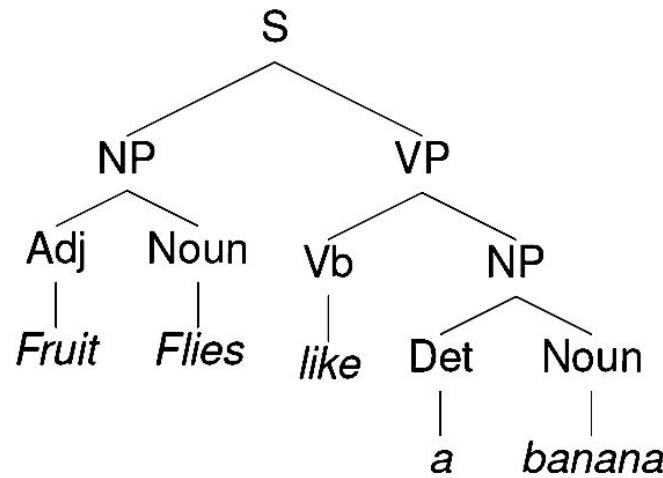
Constituency  
structures



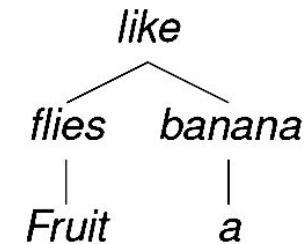
# A bit of terminology

*Fruit flies like a banana*

Constituency Structure



Dependency Structure

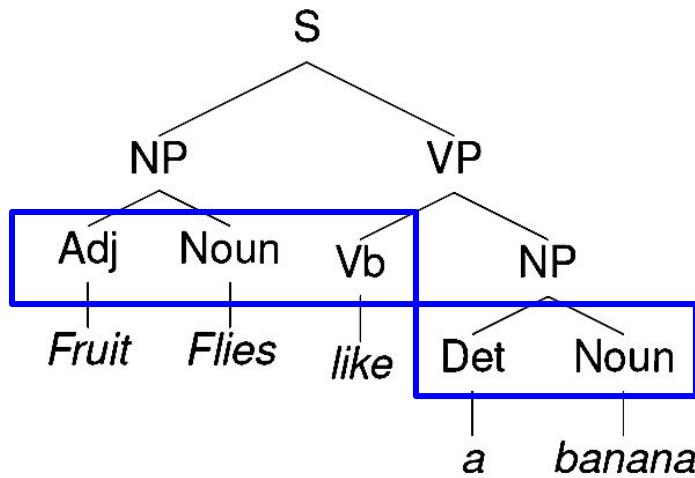


# A bit of terminology

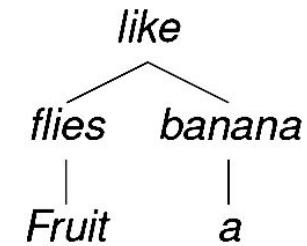
*Fruit flies like a banana*

Parts-of-speech  
(PoS)

Constituency Structure



Dependency Structure

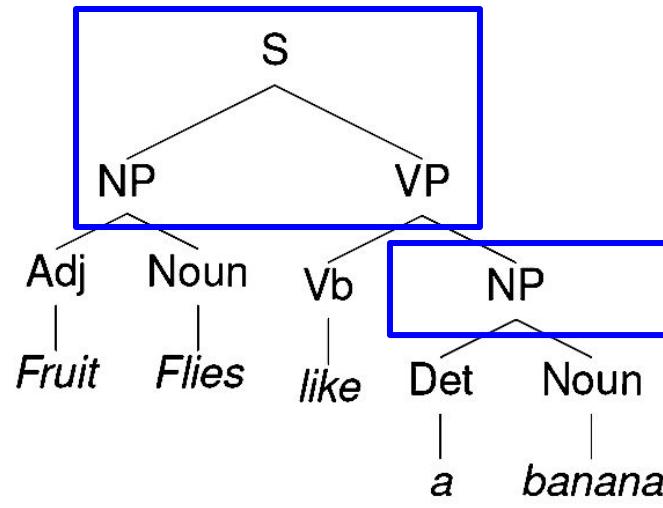


# A bit of terminology

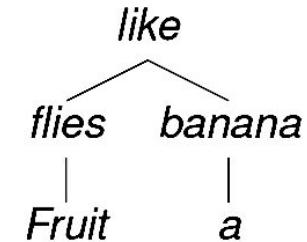
*Fruit flies like a banana*

Phrases (or constituents)

Constituency Structure



Dependency Structure



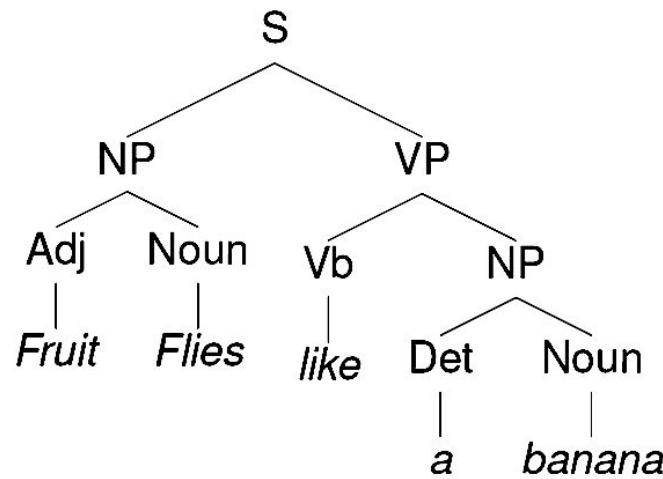
# A bit of terminology

*Fruit flies like a banana*

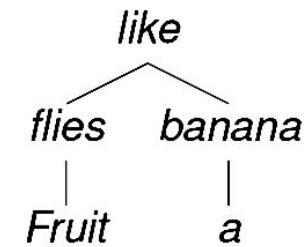
Automatic syntactic analysis = parsing

- Constituency parsing
- Dependency parsing

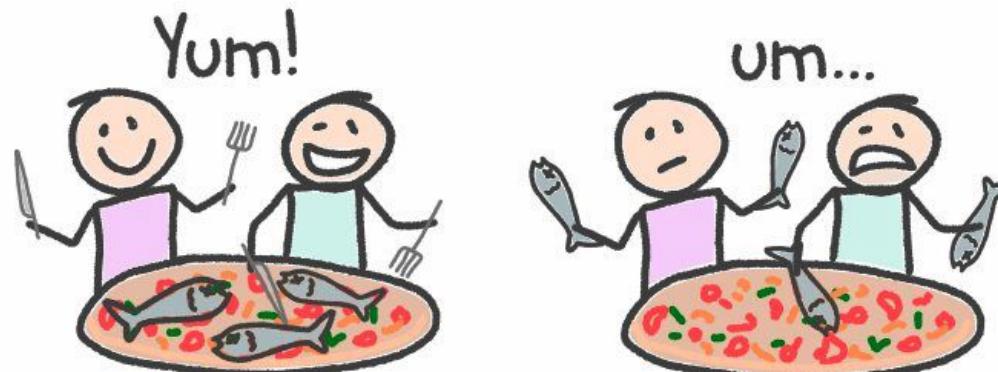
Constituency Structure



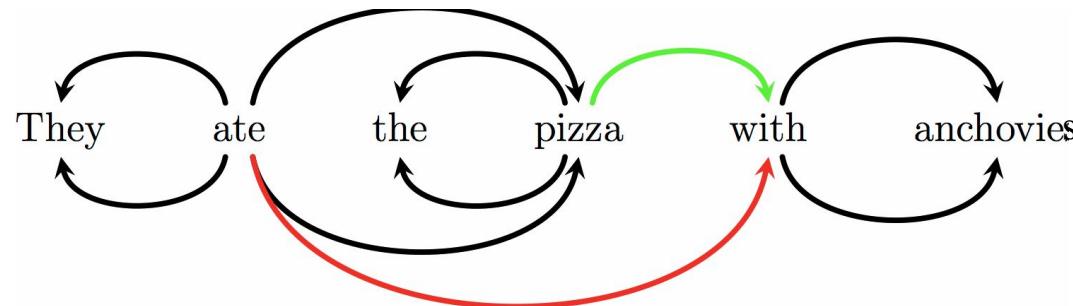
Dependency Structure



# Syntactic ambiguity: PP attachment



Creative Commons Attribution-NonCommercial 2.5  
James Constable, 2010



# Garden-path sentences

The cotton clothing is usually made of grows in Mississippi

Until the police arrest the drug dealers control the street

Mary gave the child the dog bit a bandaid

The girl told the story cried

The dog that I had really loved bones

The old man the boat

The raft floated down the river sank

We painted the wall with cracks

# Garden-path sentences

(The cotton (clothing is usually made of)) grows in Mississippi

(Until the police arrest) (the drug dealers control the street)

Mary gave (the child (the dog bit)) (a bandaid)

(The girl (told the story)) cried

(The dog that I had) really loved bones

(The old) man (the boat)

(The raft (floated down the river)) sank

We painted (the wall with cracks)

# Semantic ambiguity: polysemy

Hyponymy: man (vs. animals) ⊃ man (vs. woman) ⊃ man (vs. boy)

Metaphor: mole (the animal) > mole (a spy)

Object/color: cherry (the fruit) > cherry (as a color, cf. *I like your cherry shirt*)

Object/Informational content: book (the object) // book (its content)

Object/Collective abstract: tramway (vehicle) // tramway (means of transportation)

Tree or plant/Material/fruit/vegetable it produces: cotton (plant) > cotton (material)

Animal/Its (edible) flesh: rabbit (animal) > rabbit (meat)

# Semantic ambiguity

Named entities:

- Detection
- Linking



Main page  
Contents  
Featured content  
Current events  
Random article  
Donate to Wikipedia  
Wikipedia store

---

Interaction  
Help  
About Wikipedia  
Community portal  
Recent changes  
Contact page

---

Tools  
What links here  
Related changes  
Upload file  
Special pages  
Permanent link  
Page information  
Wikidata item  
Cite this page

Article Talk Read Edit View history Search Wikipedia

## Michael Jordan (disambiguation)

From Wikipedia, the free encyclopedia

**Michael Jordan** (born 1963) is an American basketball player.

**Michael Jordan** or **Mike Jordan** may also refer to:

### People [edit]

#### Sports [edit]

- Michael Jordan (footballer) (born 1986), English goalkeeper (Arsenal, Chesterfield, Lewes)
- Mike Jordan (racing driver) (born 1958), English racing driver
- Mike Jordan (baseball, born 1863) (1863–1940), baseball player
- Michael Jordan (American football) (born 1992), American football cornerback
- Michael-Hakim Jordan (born 1977), American professional basketball player
- Michal Jordán (born 1990), Czech ice hockey player

#### Other people [edit]

- Michael B. Jordan (born 1987), American actor
- Michael Jordan (insolvency baron) (born 1931), English businessman
- Michael Jordan (Irish politician), Irish Farmers' Party TD from Wexford, 1927–1932
- Michael I. Jordan (born 1956), American researcher in machine learning and artificial intelligence
- Michael H. Jordan (1936–2010), American executive for CBS, PepsiCo, Westinghouse
- Michael Jordan (mycologist), English mycologist

### Contents [hide]

- 1 People
  - 1.1 Sports
  - 1.2 Other people
- 2 Other uses
- 3 See also

# Multiple ambiguity

- Most or all tasks in speech and language processing can be viewed as resolving **ambiguity** at one of the levels of signal or linguistic structure.
- The spoken sentence, *I made her duck*, has five different meanings.
  - (1) I cooked waterfowl for her.
  - (2) I cooked waterfowl belonging to her.
  - (3) I created the (plaster?) duck she owns.
  - (4) I caused her to quickly lower her head or body.
  - (5) I waved my magic wand and turned her into undifferentiated waterfowl.

# Multiple ambiguity

- These different meanings are caused by multiple ambiguities.
  - PoS: *duck* can be a verb or a noun, while *her* can be a dative pronoun or a possessive pronoun -> part-of-speech tagging
  - Polysemy: the word *make* can mean *create* or *cook* -> word sense disambiguation
  - Syntactic ambiguity: the verb *make* is syntactically ambiguous in that it can be transitive (2), or it can be ditransitive (5). Moreover, *make* can take a direct object and a verb (4), meaning that the object (*her*) got caused to perform the verbal action (*duck*) -> parsing
  - In a spoken sentence, phonological ambiguity (homophones) is also present; the first word could have been *eye* or the second word *maid*.

# Language sparsity

# Corpora

Corpus = body of text stored in a machine-readable form

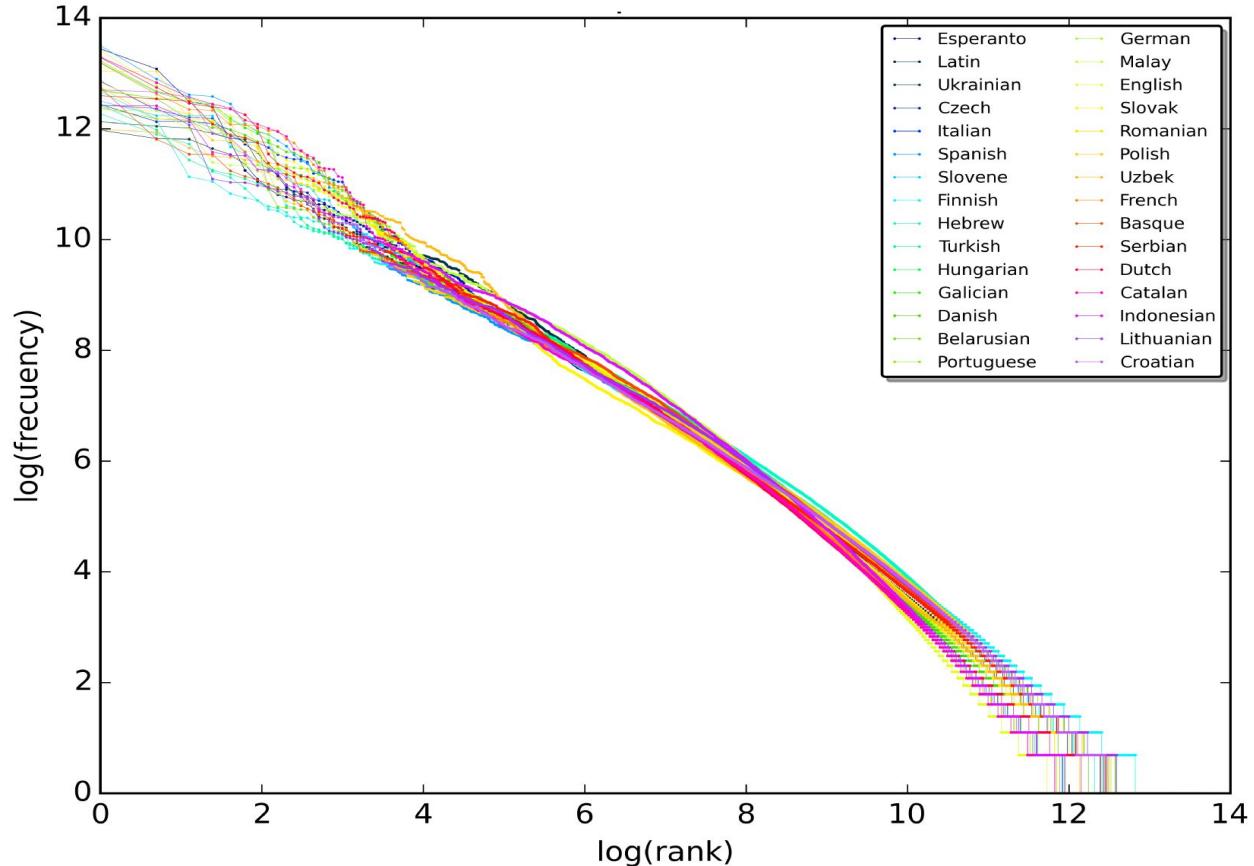
Corpora can be annotated, for serving as training, development or test data

- Morphosyntactically-annotated corpora
- Treebanks (syntactically-annotated)
- Semantically disambiguated corpora
- etc.

# Zipf's law

A plot of the rank versus frequency for the first 10 million words in 30 Wikipedias

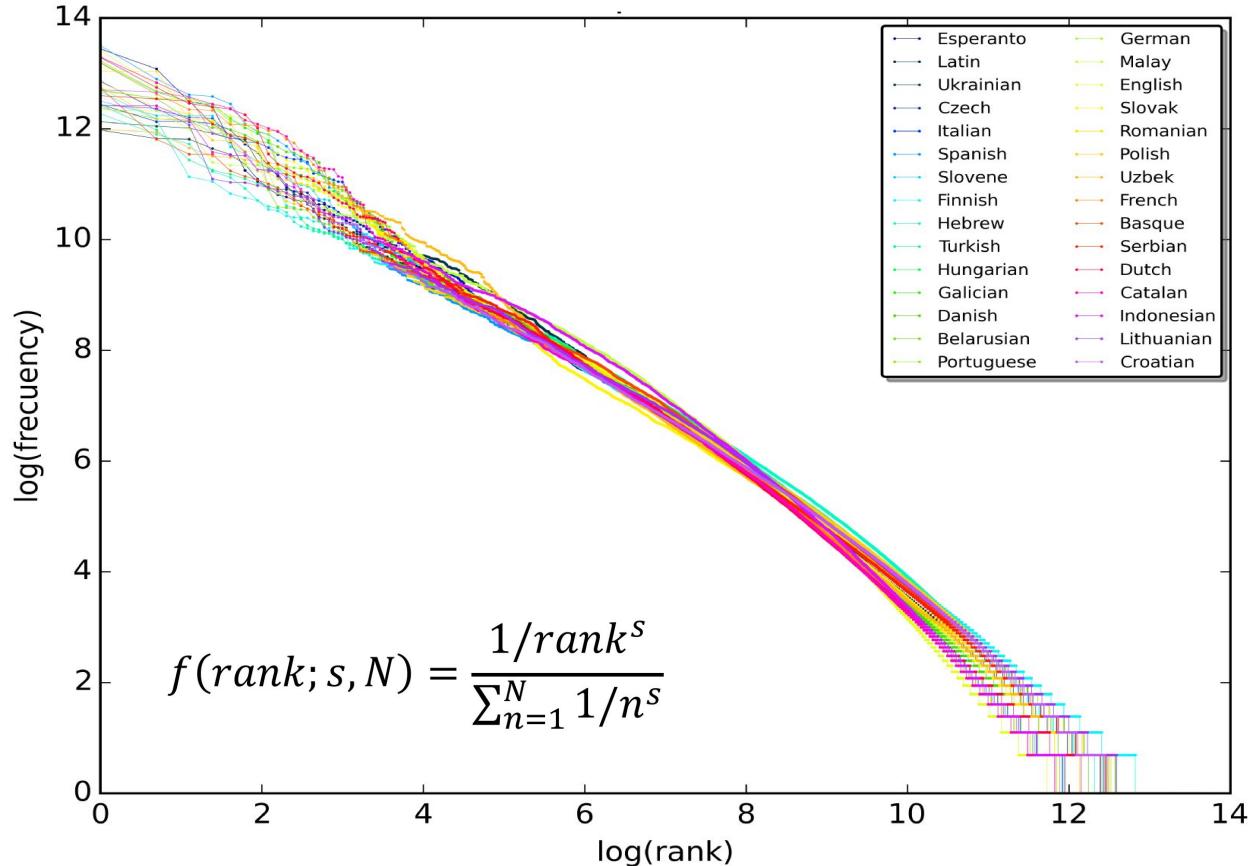
(source: Wikipedia; data: dumps from Oct 2015)



# Zipf's law

A plot of the rank versus frequency for the first 10 million words in 30 Wikipedias

(source: Wikipedia; data: dumps from Oct 2015)

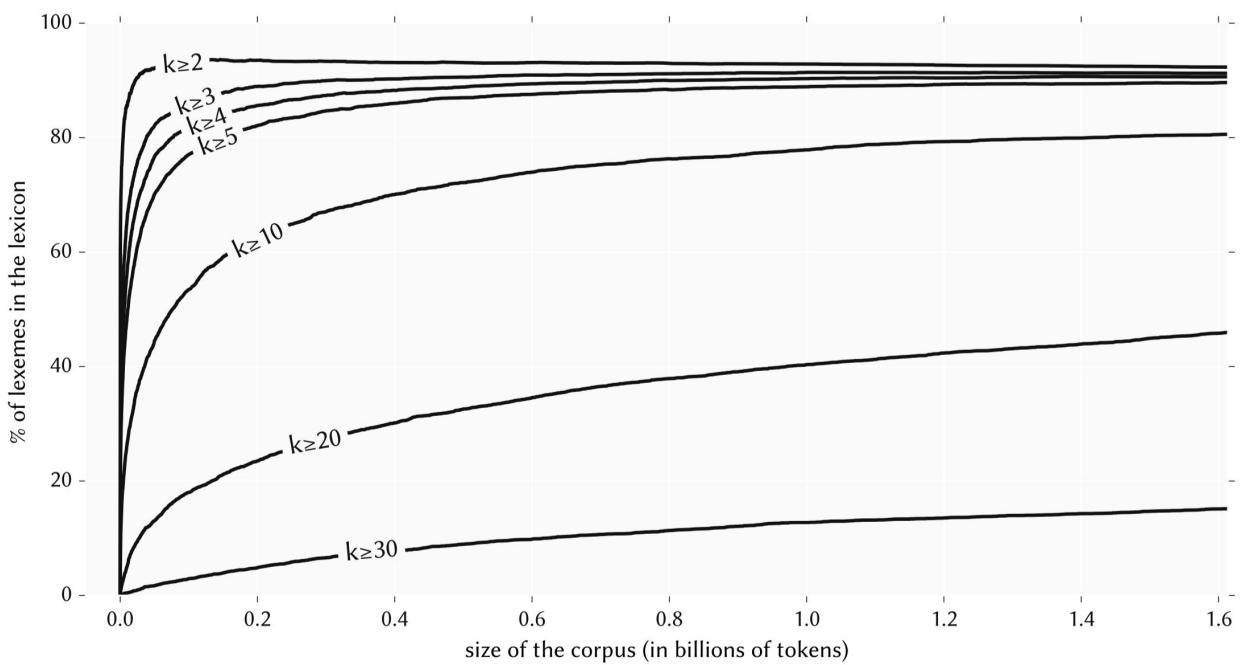


# Zipf's law

In language data, many phenomena follow a zipfian distribution

## Heaps's law

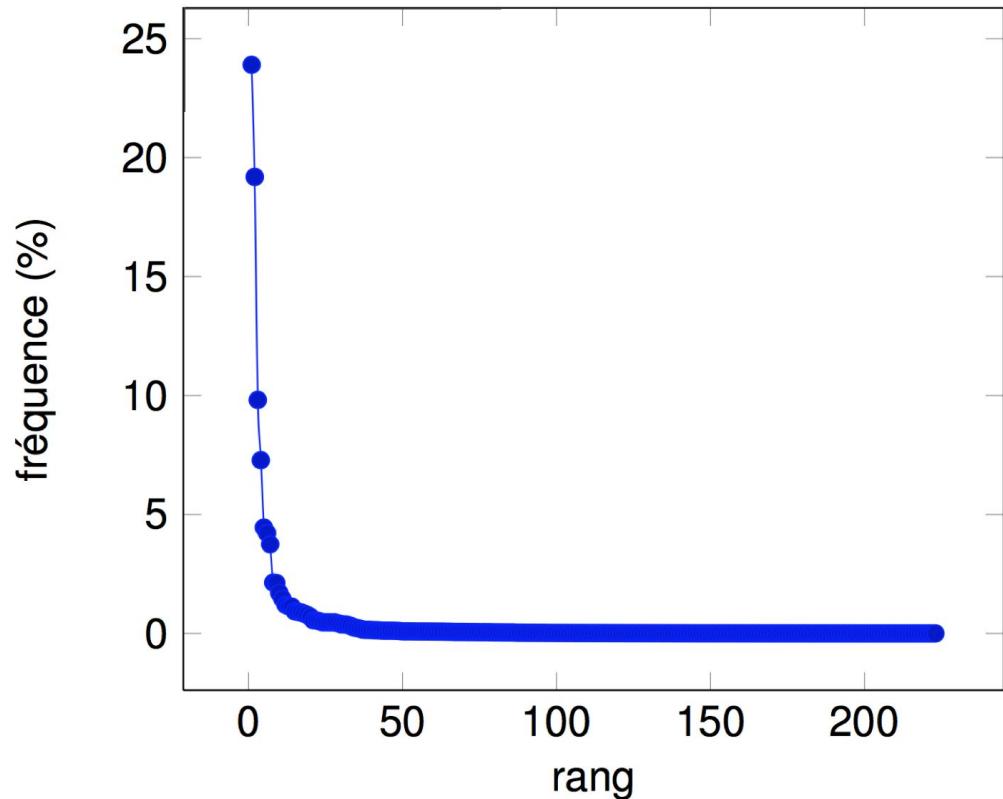
Example: proportion of lemmas (known to a pre-defined lexicon) attested in at least  $k$  inflected forms as a function of vocabulary size in a large web-based corpus of French (FrWaC) for various values of  $k$



# Zipf's law

In language data, many phenomena follow a zipfian distribution

Example: frequency of syntactic constructions in an automatically parsed 500M-word corpus



# Deconstructing Siri - 30 min

ASR

Paralinguistics

NLU

Dialog

Generation&Synthesis

# Deconstructing Siri

# The personal assistant



Amazon Alexa,  
Google Home,  
Baidu Raven, etc

Such systems can

- Identify the talker
- Recognize the words
- Understand the query
- Respond orally



Such systems can

- **Identify the talker**
- Recognize the words
- Understand the query
- Respond orally

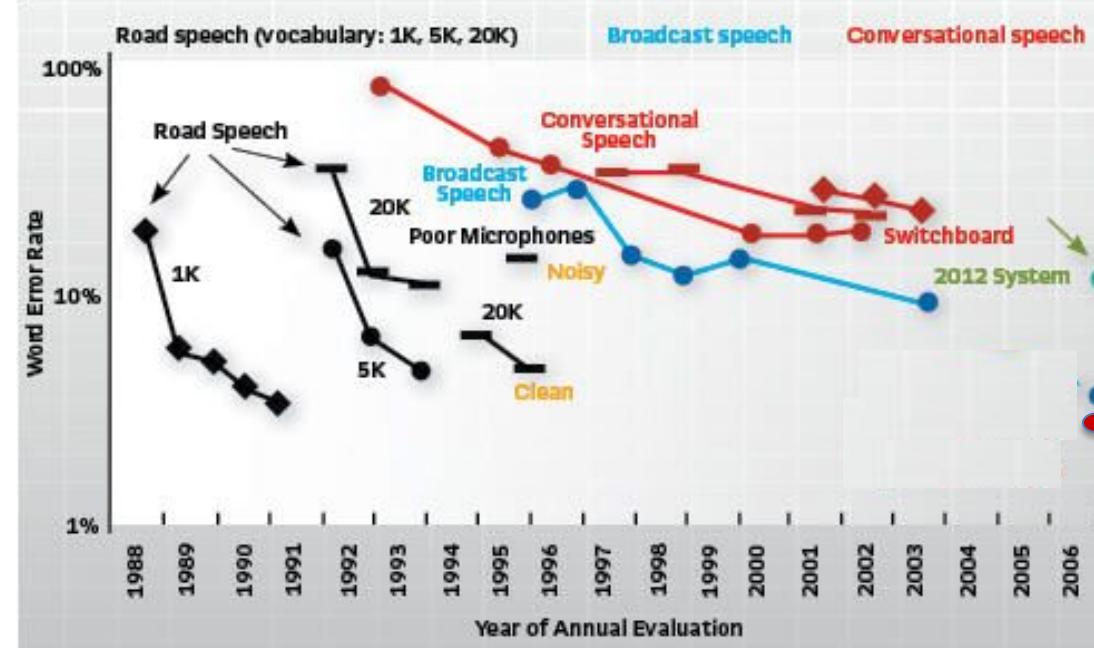


Such systems can

- Identify the talker
- Recognize the words (speech to text)
- Understand the query
- Respond orally



performance:  
roughly like  
humans



- Recognize the words (speech to text)**
- Understand the query
- Respond orally

Trouve moi un  
resto Grec près  
d'ici



performance:  
roughly like  
humans (not for  
casual or noisy)

Such systems can

- Identify the talker
- Recognize the words
- Understand the query
- Respond orally



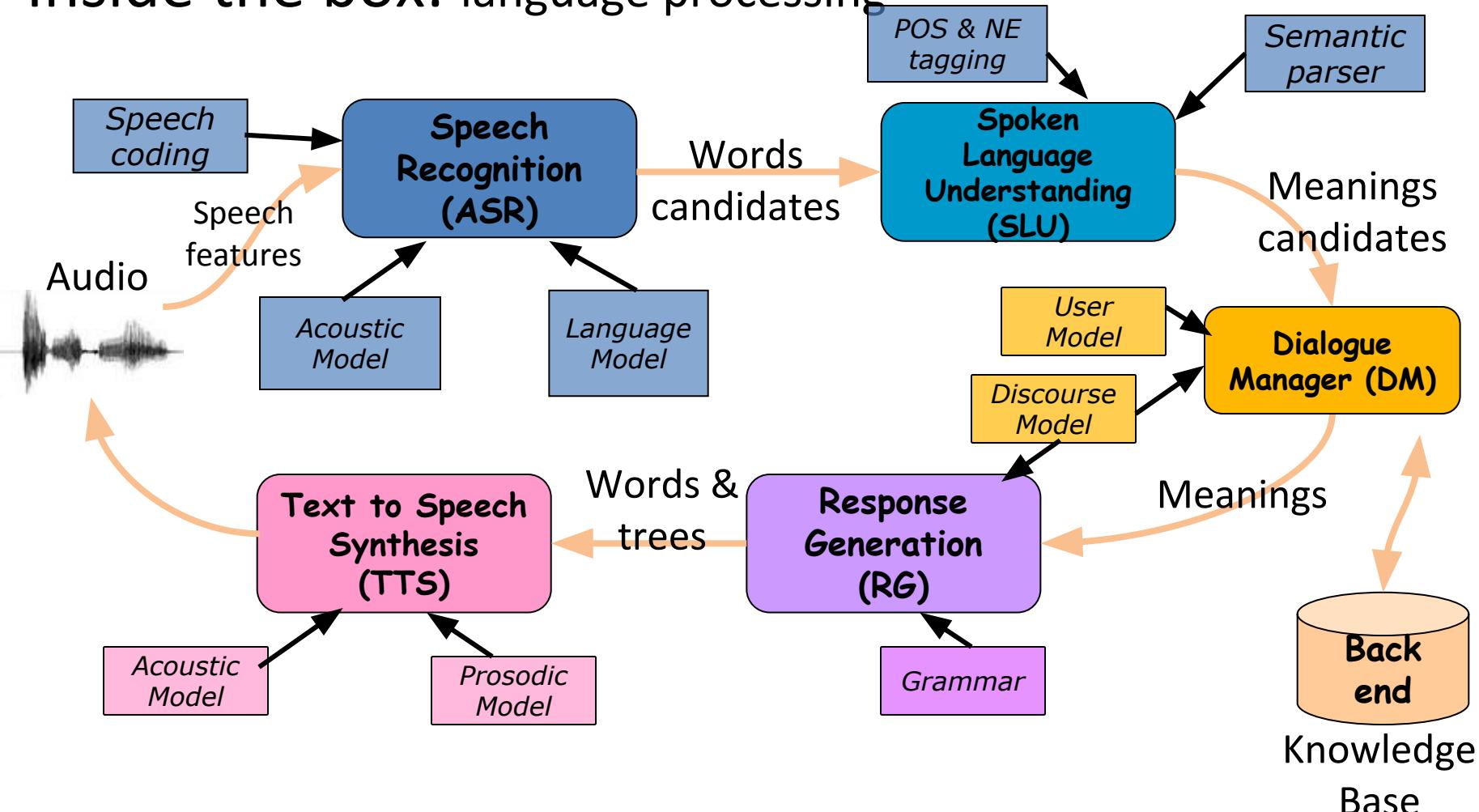
Such systems can

- Identify the talker
- Recognize the words
- Understand the query
- **Respond orally**



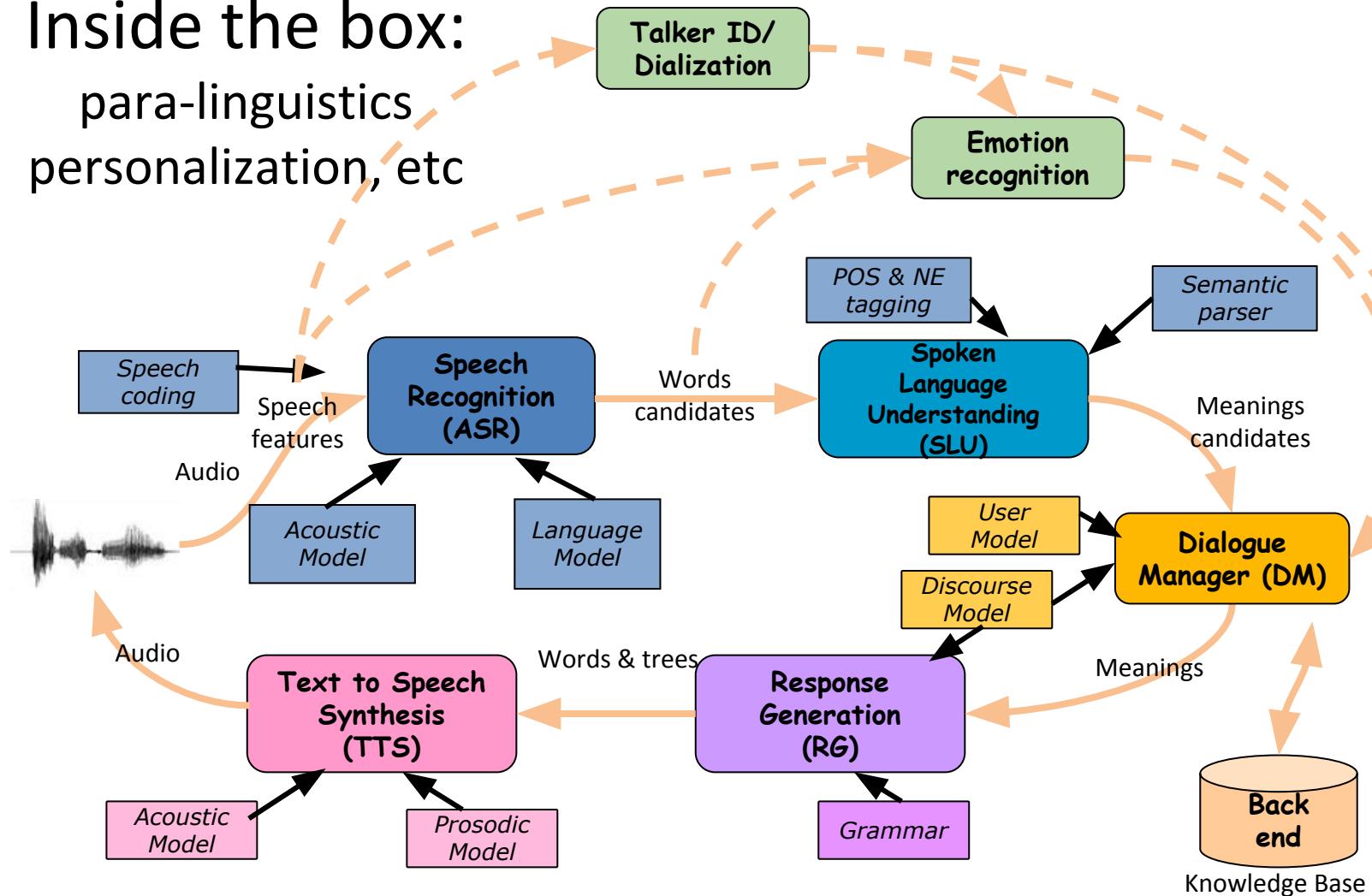
performance: close  
to natural speech  
(not emotional  
speech)

# Inside the box: language processing



# Inside the box:

para-linguistics  
personalization, etc



# Main algorithms (that address these problems) -20 min

Modular approach

Symbolic approaches

Probabilistic framework

Search algorithms

End-to-end training

# Algorithms and methods for speech and language processing

# Computational approaches to NLP over time

1970

2000

2012

## Symbolic approaches

- **Computational expertise:**

Formal grammars (algebraic  
grammars, mildly context-sensitive  
grammars, polynomial  
languages...), parsing algorithms,  
dynamic programming

- **Comp. linguistics expertise:**

Formal and descriptive linguistics,  
grammar engineering,  
development of lexical resources

# Computational approaches to NLP over time

1970

2000

2012

## Symbolic approaches

- **Computational expertise:**

Formal grammars (algebraic grammars, mildly context-sensitive grammars, polynomial languages...), parsing algorithms, dynamic programming

- **Comp. linguistics expertise:**

Formal and descriptive linguistics, grammar engineering, development of lexical resources

## Statistical approaches

- **Computational expertise:**

(statistical) machine learning, supervised, semi-supervised and non-supervised (PCFG, CRF, MEMM, discriminative algorithms...), hybrid approaches

- **Comp. linguistics expertise:**

development of annotated corpora (training dataset), development of lexical resources

# Computational approaches to NLP over time

1970

2000

2012

## Symbolic approaches

- **Computational expertise:**  
Formal grammars (algebraic grammars, mildly context-sensitive grammars, polynomial languages...), parsing algorithms, dynamic programming

- **Comp. linguistics expertise:**  
Formal and descriptive linguistics, grammar engineering, development of lexical resources

## Statistical approaches

- **Computational expertise:**  
(statistical) machine learning, supervised, semi-supervised and non-supervised (PCFG, CRF, MEMM, discriminative algorithms...), hybrid approaches

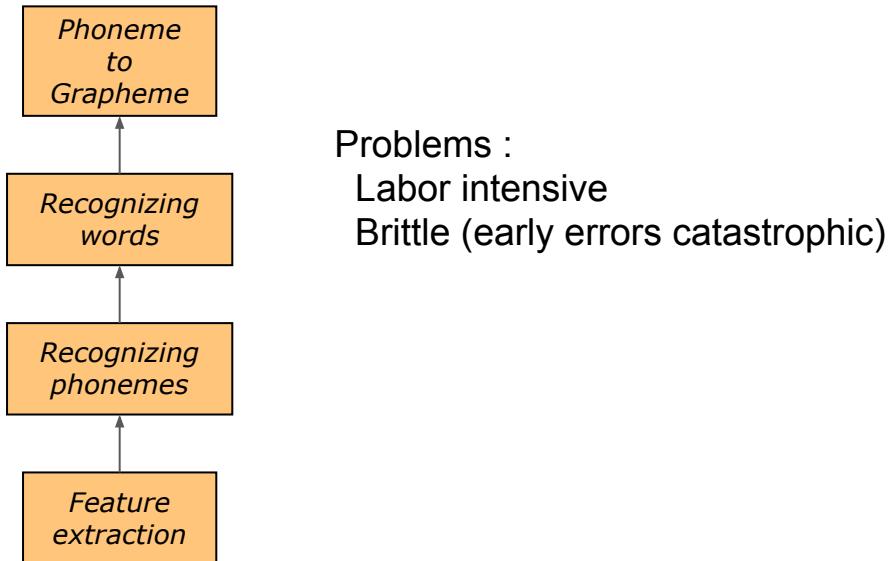
- **Comp. linguistics expertise:**  
development of annotated corpora (training dataset), development of lexical resources

## Neural approaches

- **Comp. expertise:**  
neural networks, deep learning, end-to-end training

- **Comp. ling. exp.:** same as for statistical approaches

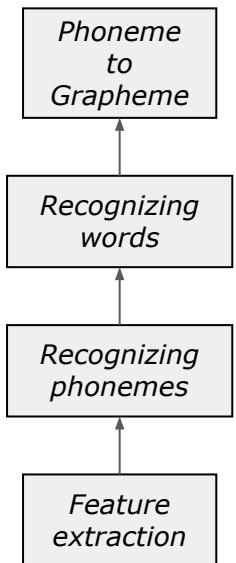
## Hand engineered system



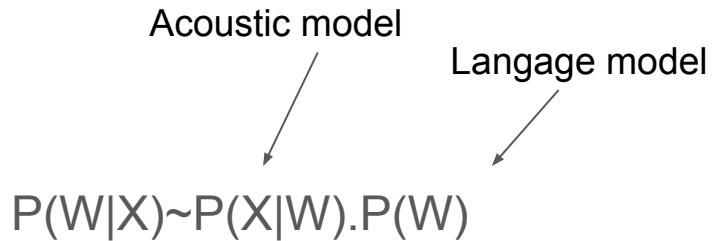
Problems :

- Labor intensive
- Brittle (early errors catastrophic)

## Hand engineered system

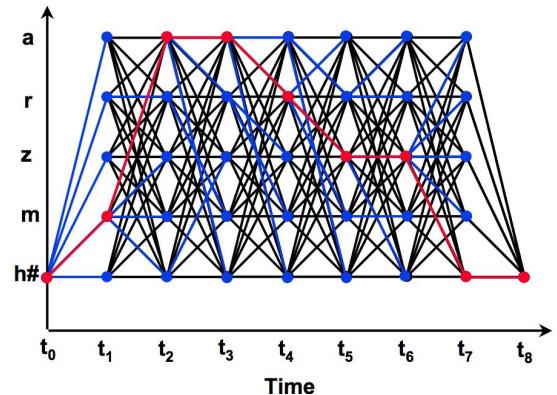


## Probabilistic system



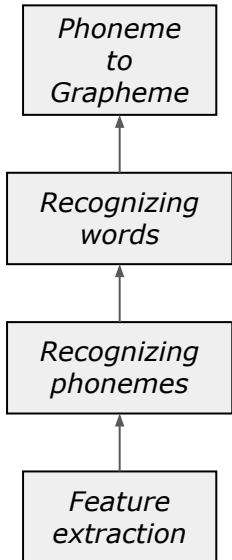
Decoding :  
 $W = \operatorname{argmax}_W P(X|W) \cdot P(W)$

Algorithms: State search space  
Dynamic programming

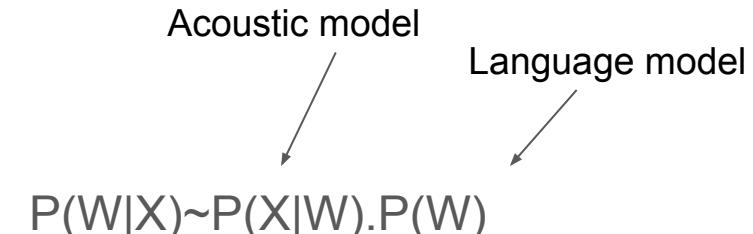


Advantage: robust to error

## Hand engineered system



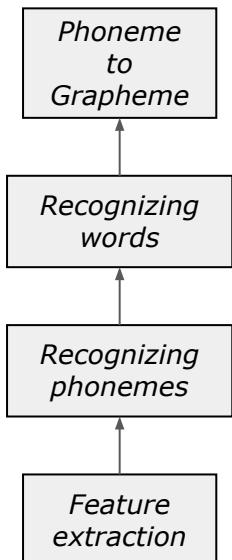
## Probabilistic system



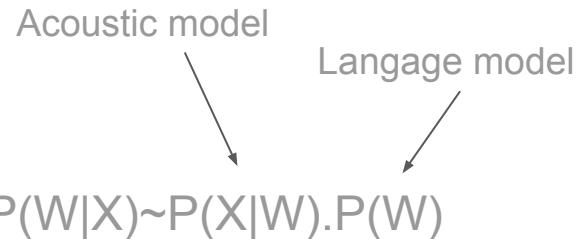
Learning the probabilities

Algorithms: EM, ..  
Problem: Lots of annotated data

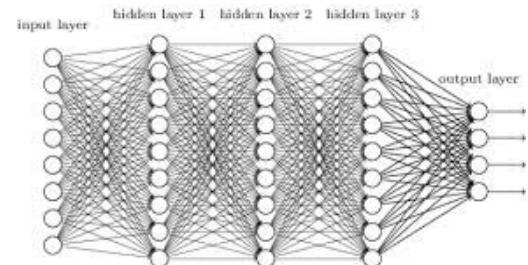
### Hand engineered system



### Probabilistic system



### Deep learning systems



#### Decoding:

Distributed intermediate representations  
Probabilistic interpretation of final layer

#### Learning:

Stochastic Gradient Descent

Advantage: more robust

Problem: even more annotated data

# Main applications -10min

Assistant (already seen)

Information retrieval

Translation & summarization

etc

# **Applications of speech and language processing**

# Current landscape - 10 min

A brief history of the interaction between IA, NLP/CL and Speech processing.

Academia (USA: Stanford, MIT, CMU, JHU, ..; Asia: NAIST, ...; Europe: ...)

Industry (GAFA)

Government (US: DARPA, IARPA, more than NSF / Europe/France: national funding agencies (ANR), EC)

# Applications

- Information extraction, information retrieval, text mining (ex.: opinion surveys)
- Text generation, text simplification, automatic summarisation
- Spelling correction (writing aid, post-OCR, normalisation of noisy/non-canonical texts)
- Machine translation, computer-aided translation
- Chatbots, conversational agents, question answering systems
- Medical applications (early diagnosis, language-based medical monitoring...)
- **Applications in linguistics** (modelling languages and their evolution, sociolinguistic studies...)
- **Digital humanities** (exploitation of text documents, for instance in historical research)

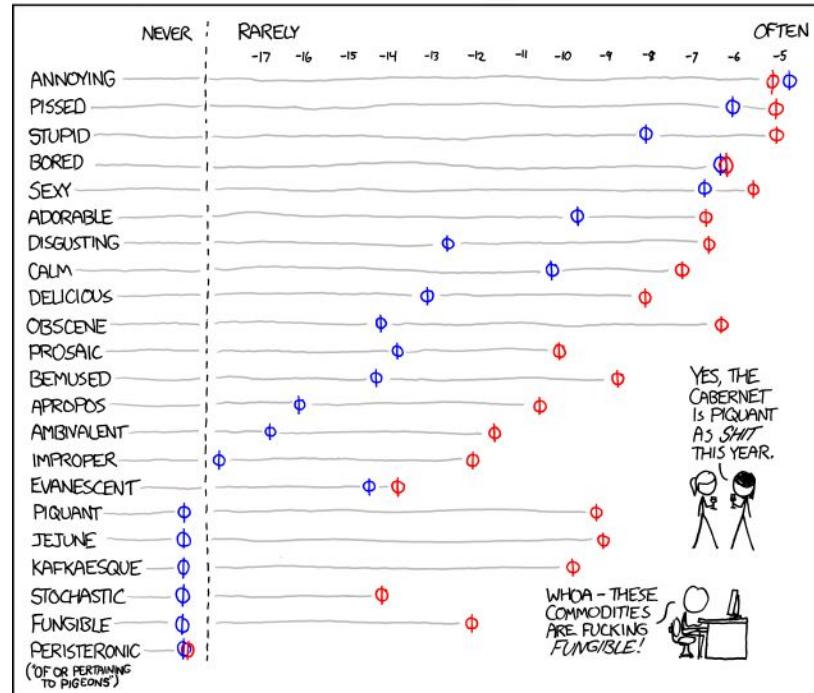
# Academic applications: comp. linguistics

## Examples

- Language modelling (synchronic, diachronic), with a number of approaches: formal, corpus-based, simulation-based, psycholinguistics, neurolinguistics
- Sociolinguistics

FREQUENCY WITH WHICH VARIOUS ADJECTIVES  
ARE INTENSIFIED WITH OBSCENITIES (BASED ON GOOGLE HITS)

Φ: "FUCKING \_\_\_\_"  
Φ: "\_\_\_\_ AS SHIT" SCALE:  $\ln(\frac{\text{HITS FOR INTENSIFIED PHRASE}}{\text{HITS FOR ADJECTIVE ALONE}})$



# Academic applications: digital humanities

## Examples

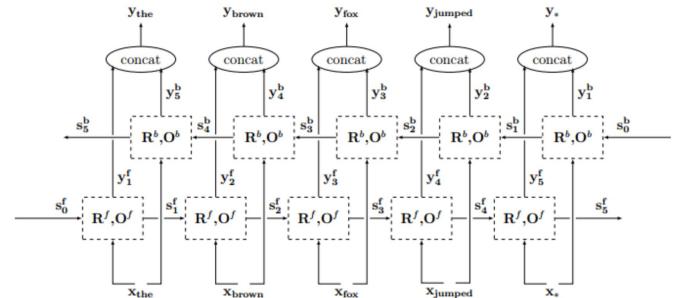
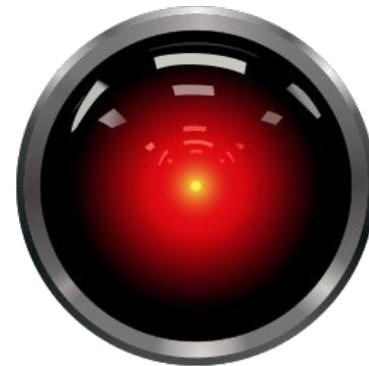
- Exploitation of textual data for research in other domains (history, philology...)
  - Computational epigraphy  
(in collaboration with image processing specialists)



# Speech and language processing today

# NLP and AI

- NLP is one of the areas within “old” AI
  - AI = computationally simulate human behaviours requiring intelligence
  - Among them: understanding, producing and transforming speech / language
- One of the areas taking advantage of the “new” AI
  - In particular: deep learning
  - *confusion between objectives and means*



# NLP and AI

- Neural approaches have resulted in major improvements, esp. in:
  - Machine translation
  - Semantic analysis
- However,
  - Technically complex approaches (training times...)
  - It is difficult to “correct” a neural model
  - Often require huge amounts of training data
  - Models are dependent to the characteristics of training data

# What is still really hard?

- **Many languages are more difficult to process than English**
  - Low-resource languages (Turkish, Romanian, Icelandic, Inuktikut...)
  - Morphologically rich languages (Czech, Finnish, Basque, Inuktikut...)
  - For machine translation: very divergent language pairs (Russian<->Chinese)
- **A number of tasks are really difficult**
  - Semantics / pragmatics
  - How can we represent meaning?
  - How can we represent world knowledge?
- **New and/or difficult types of data**
  - Noisy textual data
  - Data in context
  - Spontaneous dialogue

# What is still really hard?

- Many languages are more difficult to process than English
  - Low-resource languages (Turkish, Romanian, Icelandic, Inuktikut...)
  - Morphologically rich languages (Czech, Finnish, Basque, Inuktikut...)
  - For machine translation: very divergent language pairs (Russian<->Chinese)
- A number of tasks are really difficult
  - Semantics / pragmatics
  - How can we represent meaning?
  - How can we represent world knowledge?
- New and/or difficult types of data
  - Noisy textual data
  - Data in context
  - Spontaneous dialogue



© Pat

Bastien Péan  
@BastienPhan

Qui a fait ça ?

11:42 - 29 Mai 2016

768 489

Suivre

# What is still really hard?

- Many languages are more difficult to process than English
  - Low-resource languages (Turkish, Romanian, Icelandic, Inuktikut...)
  - Morphologically rich languages (Czech, Finnish, Basque, Inuktikut...)
  - For machine translation: very divergent language pairs (Russian<->Chinese)
- A number of tasks are really difficult
  - Semantics / pragmatics
  - How can we represent meaning?
  - How can we represent world knowledge?
- New and/or difficult types of data
  - Noisy textual data
  - Data in context
  - Spontaneous dialogue



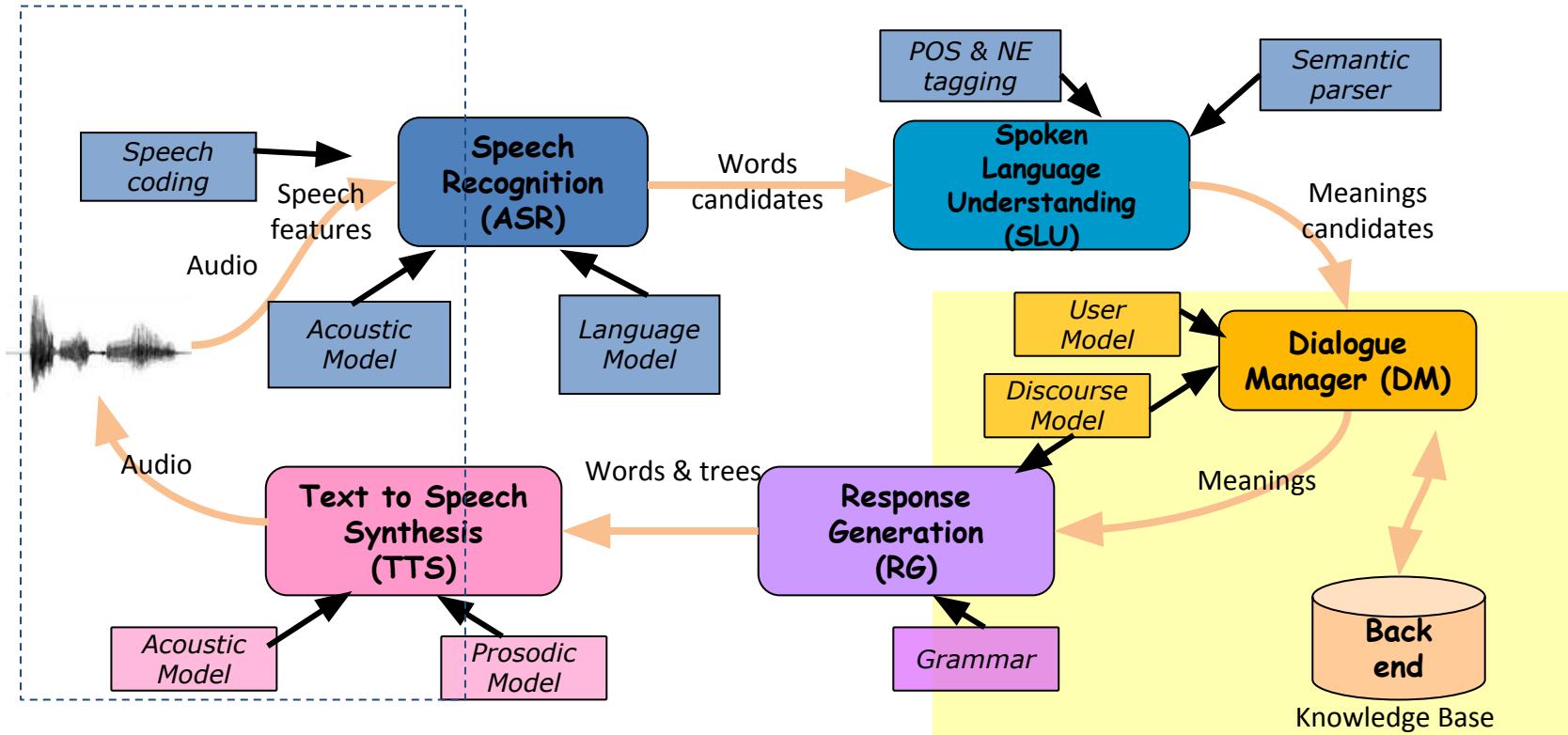
# This course

# Course roadmap

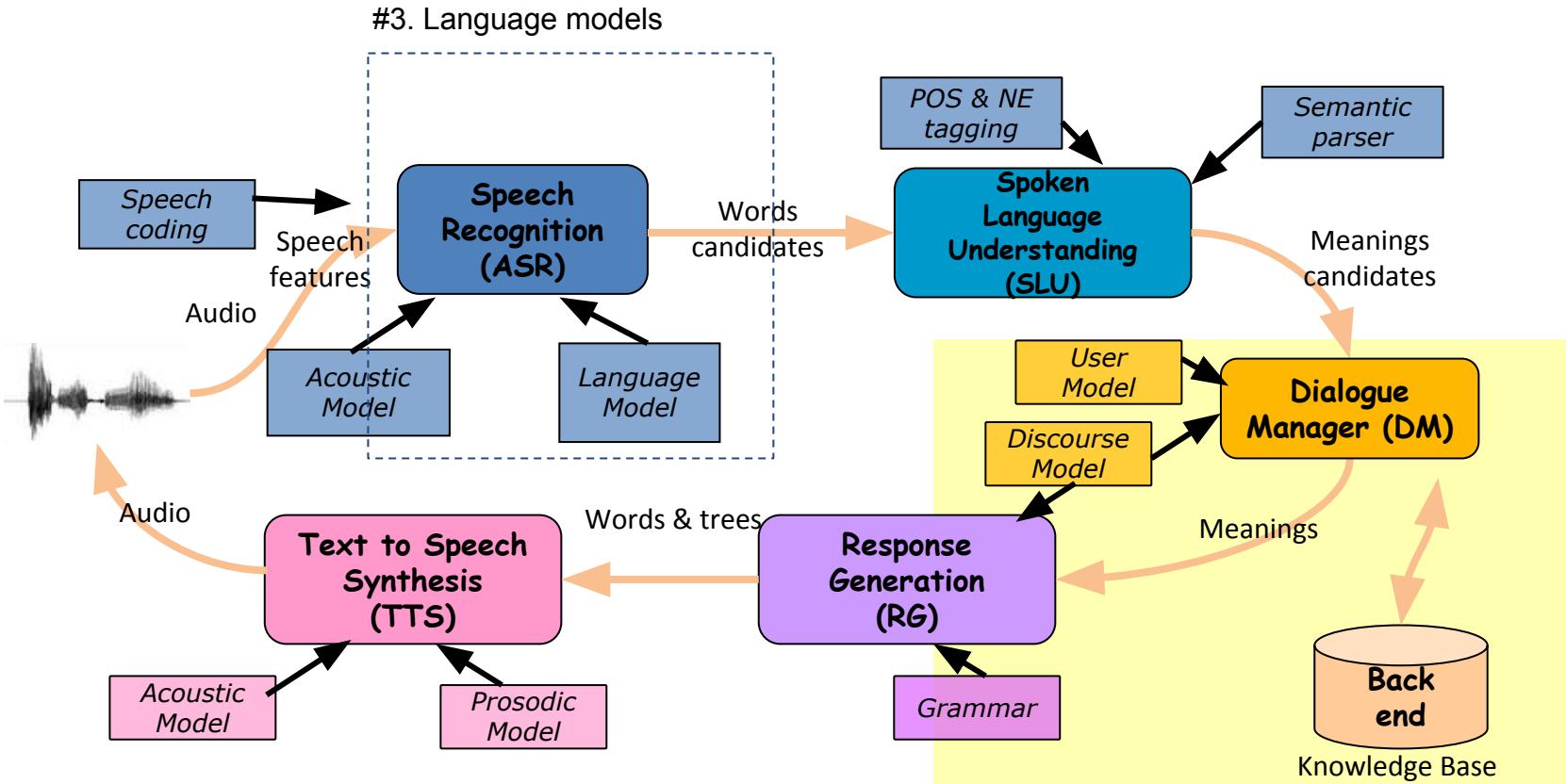
1. Introduction to speech and language processing
2. Acoustic modeling
3. Language modeling
4. Language Processing in the wild: Word embeddings and noisy texts
5. Formal Grammars and Syntax
6. Parsing
7. Lost in translation
8. Conclusion: Open questions and hot topics

# Course roadmap

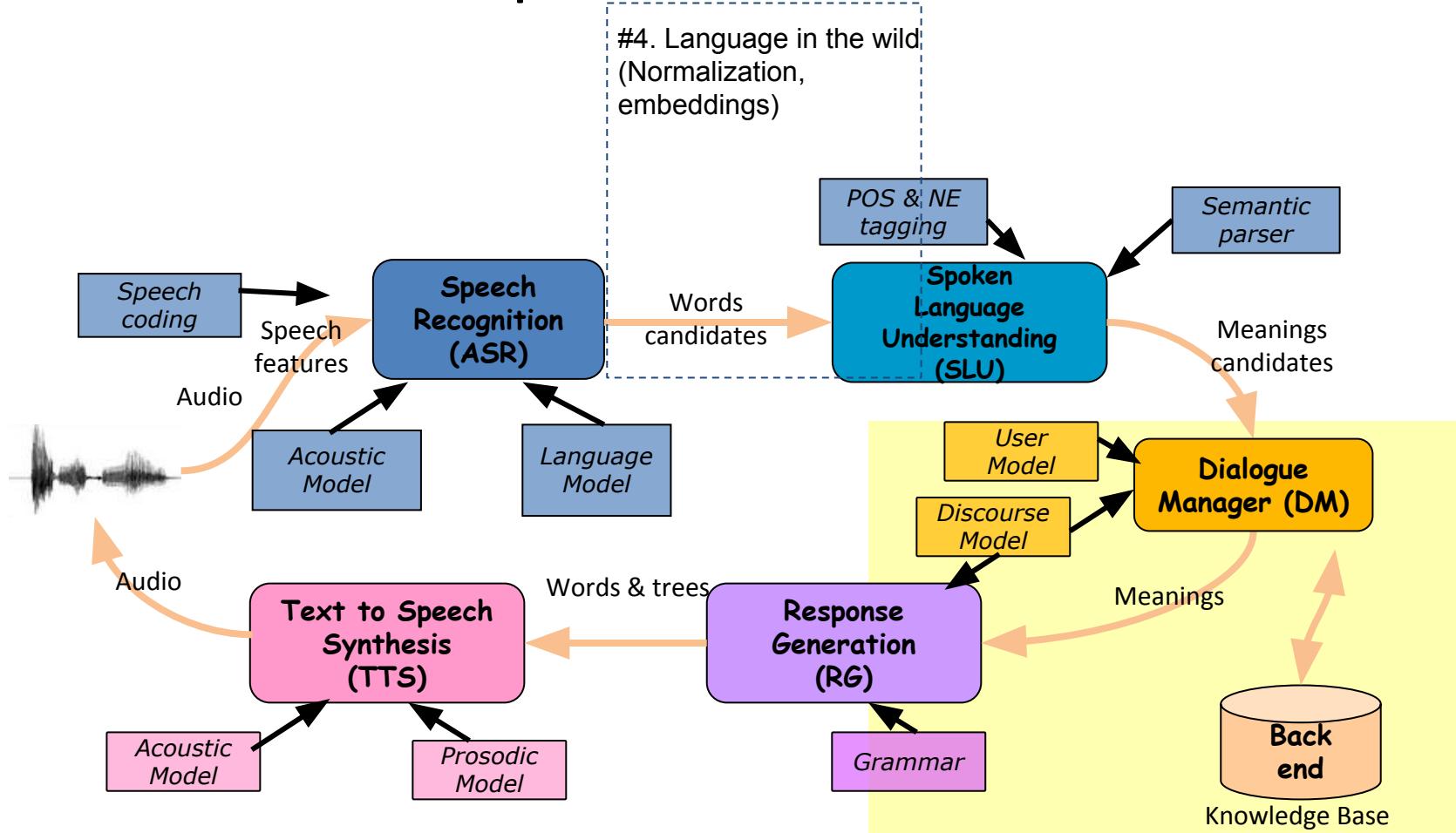
## #2. Speech features & Acoustic models



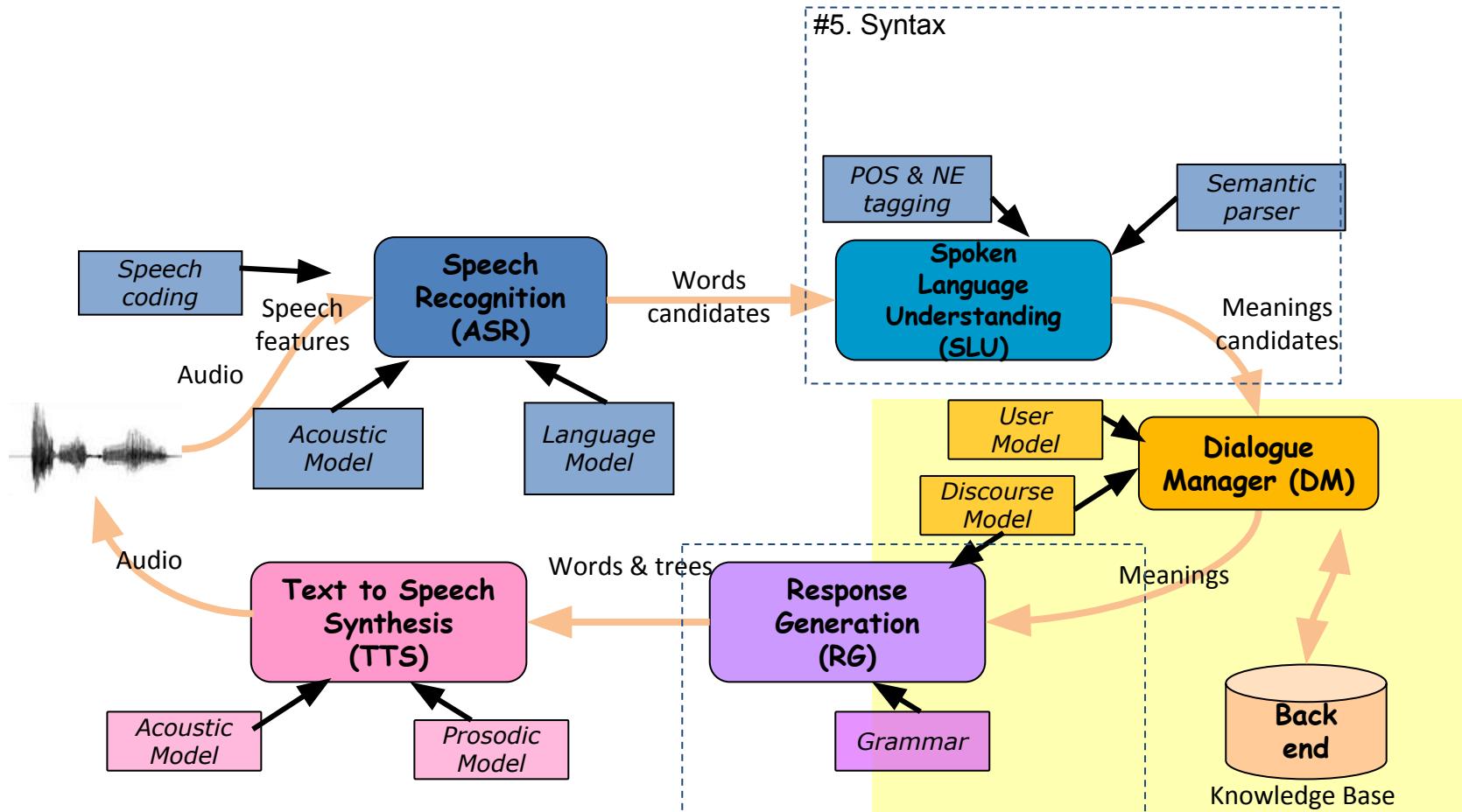
# Course roadmap



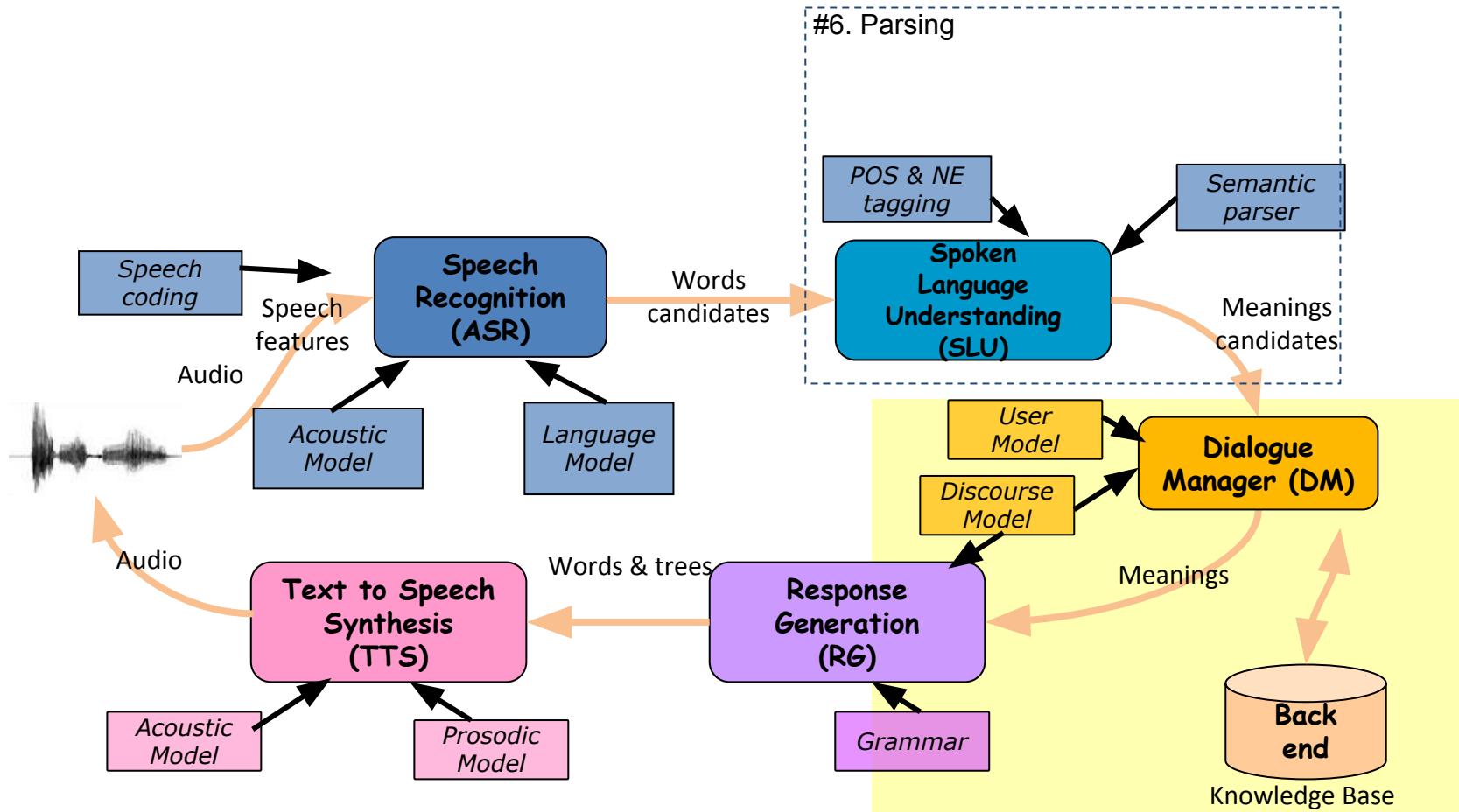
# Course roadmap



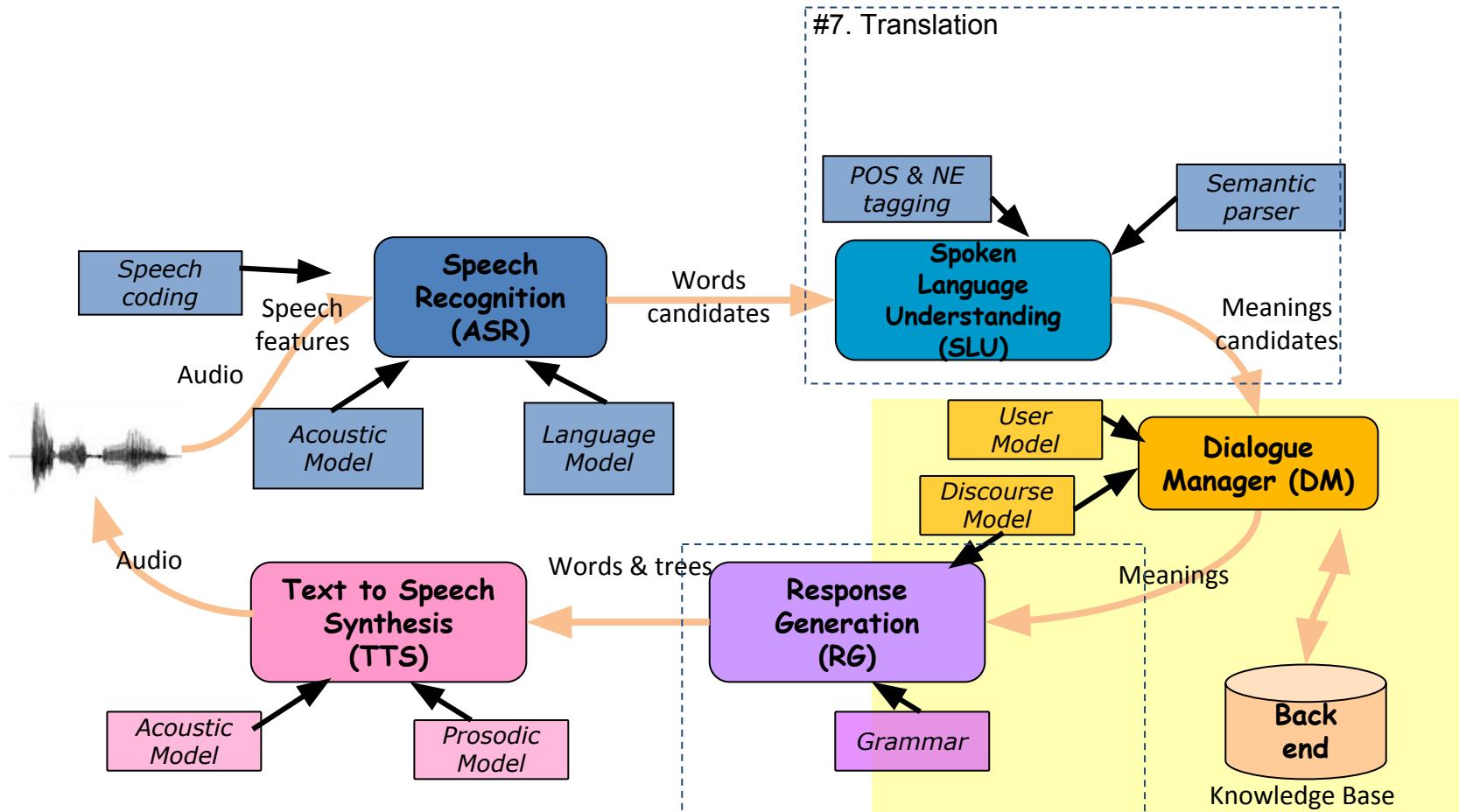
# Course roadmap



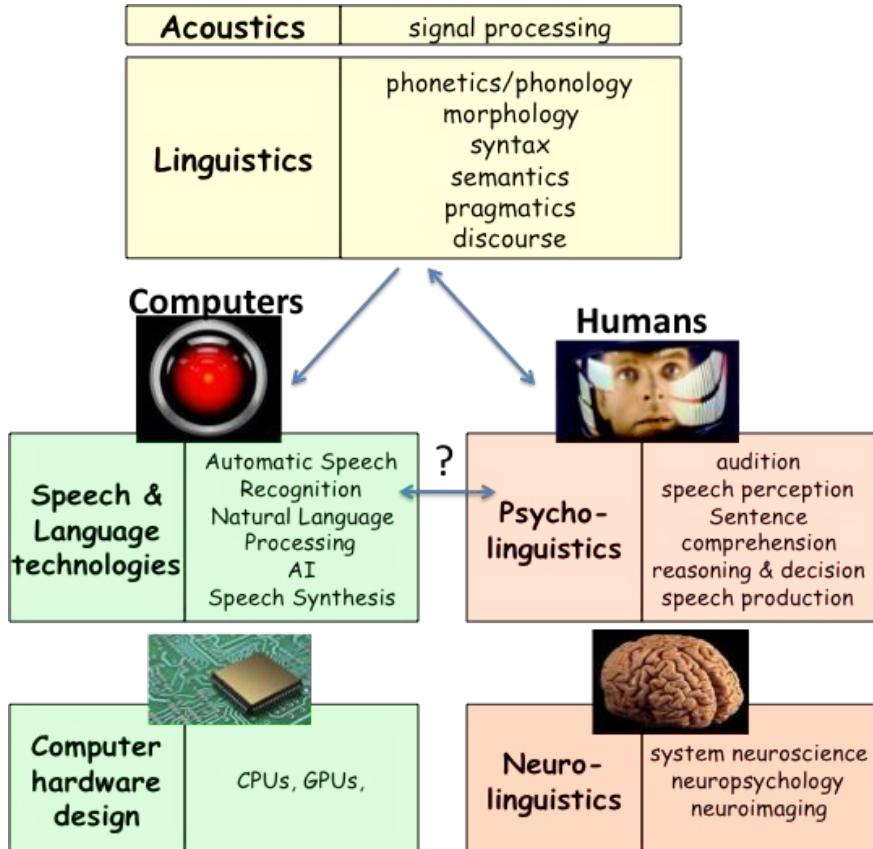
# Course roadmap



# Course roadmap



## #8. New directions, challenges



# Course logistics

#1: Intro (Sagot & Dupoux)

#2: ASR1 (Dupoux & Zeghidour)

#3: ASR2 (Dupoux, Zeghidour, Riad)+ TD #1

#4: NLP1 (Sagot)

#5: NLP2 (Sagot)

#6: NLP3 (Sagot) +TD #2

#7: Translation (Guest: Schwenk)

#8: New directions (Sagot & Dupoux)

- 2\* TDs-Assignments
  - 80% final grade
  - 2 weeks to do it
  - 1 point /20 subtracted every late day
  - Bring your computer.
- On-line Quizzes
  - 20% final grade (on the n-1 best responses)
    - No response: 0/2
    - Bad response: 1/2
    - Good response: 2/2

→ **No compensation (pas de rattrapage)**

# On-line Quizz

<https://api.socrative.com/rc/BbvH97>

Room: MVA2019

Screen name: Your full name

Question (you can only click once):

Are personal assistants at home raising ethical issues regarding data privacy?

1. Yes
2. No
3. Don't know