Lecture 1: Introduction

USC VSoE CSCI 544: Applied Natural Language Processing

Jonathan May -- 梅約納

August 23, 2017

Most slides from Noah Smith

What is NLP?

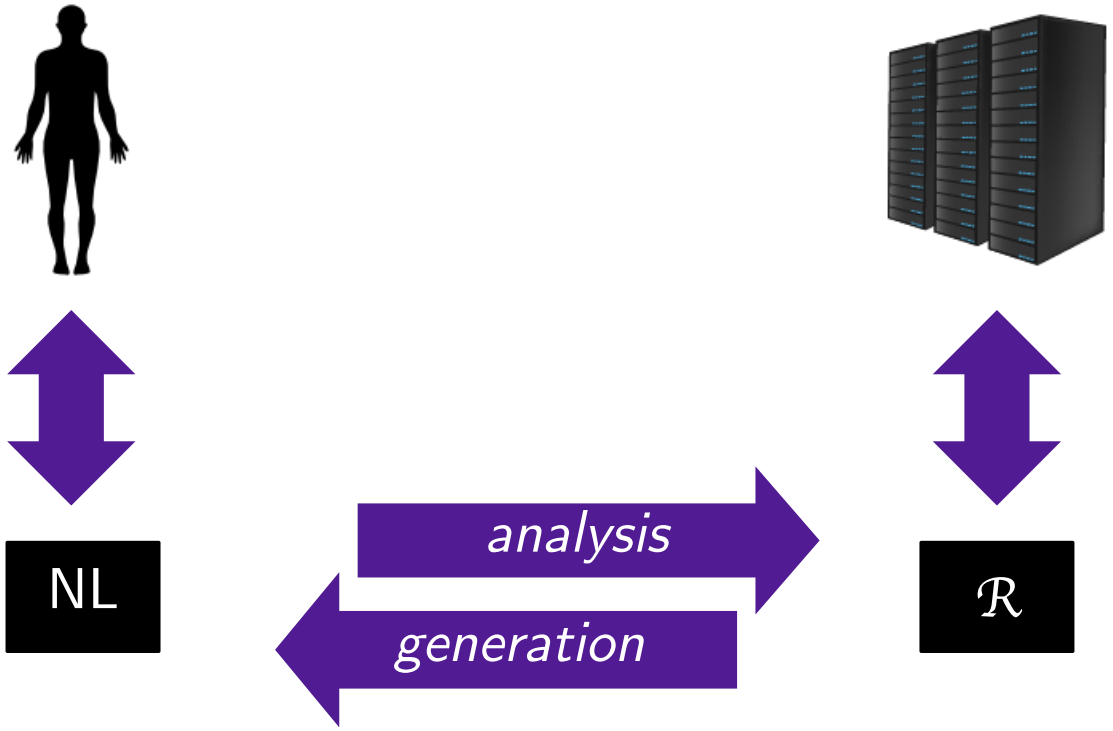
NL $\in$ {Mandarin Chinese, English, Spanish, Hindi, ..., Uyghur, ..., Oromo ...}

Automation of:

- analysis (NL $\rightarrow \mathcal{R}$)
- generation ($\mathcal{R} \rightarrow$ NL)
- acquisition of $\mathcal{R}$ from knowledge and data

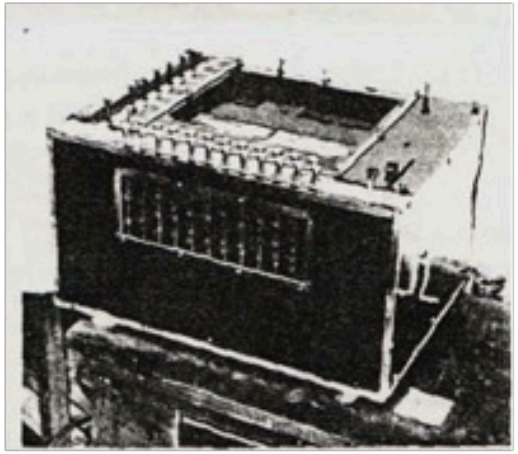
What is $\mathcal{R}$?

Anybody speak a rare language?


NL
analysis
generation
$\mathcal{R}$

NLP is a pretty old topic!


JOH. J. BECHERI,
Spirenſis
CHARACTER,
Pro
NOTITIA LIN-
guarum Univerſali.
INVENTUM STEGANO-
GRAPHICUM HACTENUS IN-
auditum, quo quilibet ſuam Legendo verrendo ſatis
diverſa, imò omnes Linguas, unius enim diei
laborantione, explicare ac intelli-
gere poteſt.
FRANCOFURTI,
Sumpt. Johannis Wilh. Ammonii
& Wilhelmi Serlini,
Typis Johannis Georgii Spoerlin.
ANNO M. DC. LI.



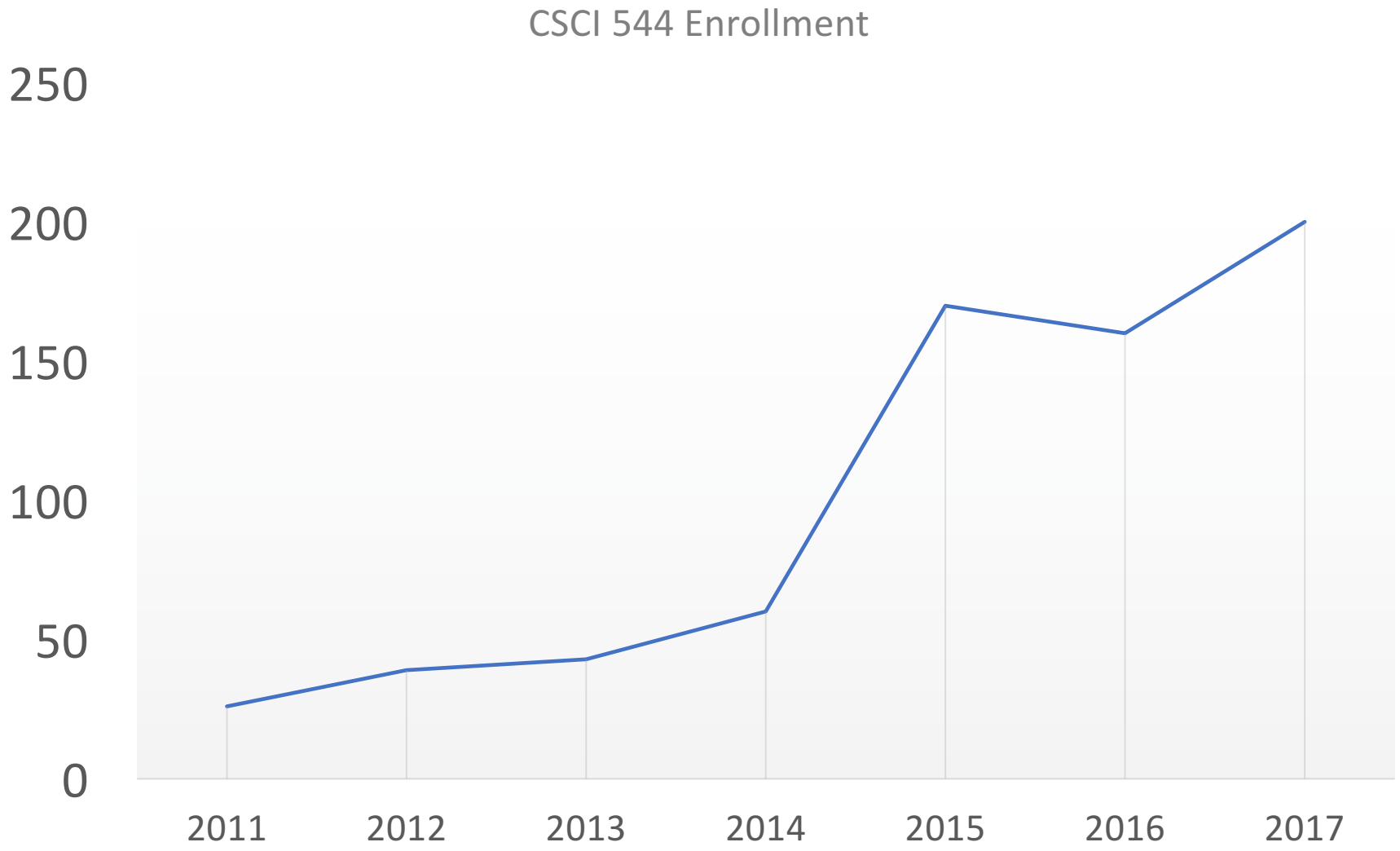
Becher mechanical meta-language
for language-to-meaning: 1666
Georges Artsruni mechanical brain: 1930
Computers proposed for translation: 1949
First computer: 1946

(Freigang, 2001)

ACL founded : 1962

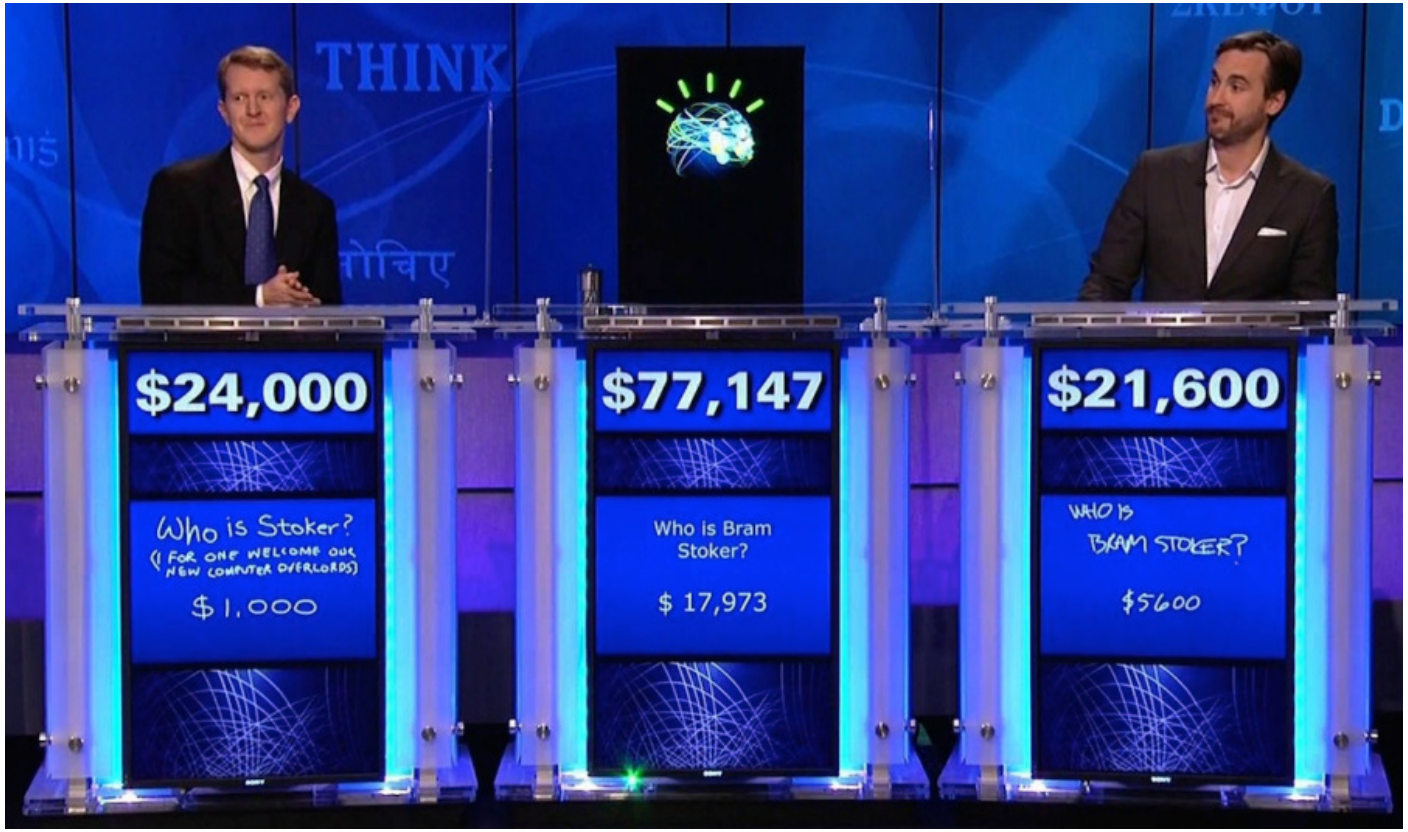
Jon first heard about NLP : 1999

Gainfully employed since 2001
(USC PhD 2004-2010)


CSCI 544 Enrollment
250
200
150
100
50
0
2011
2012
2013
2014
2015
2016
2017

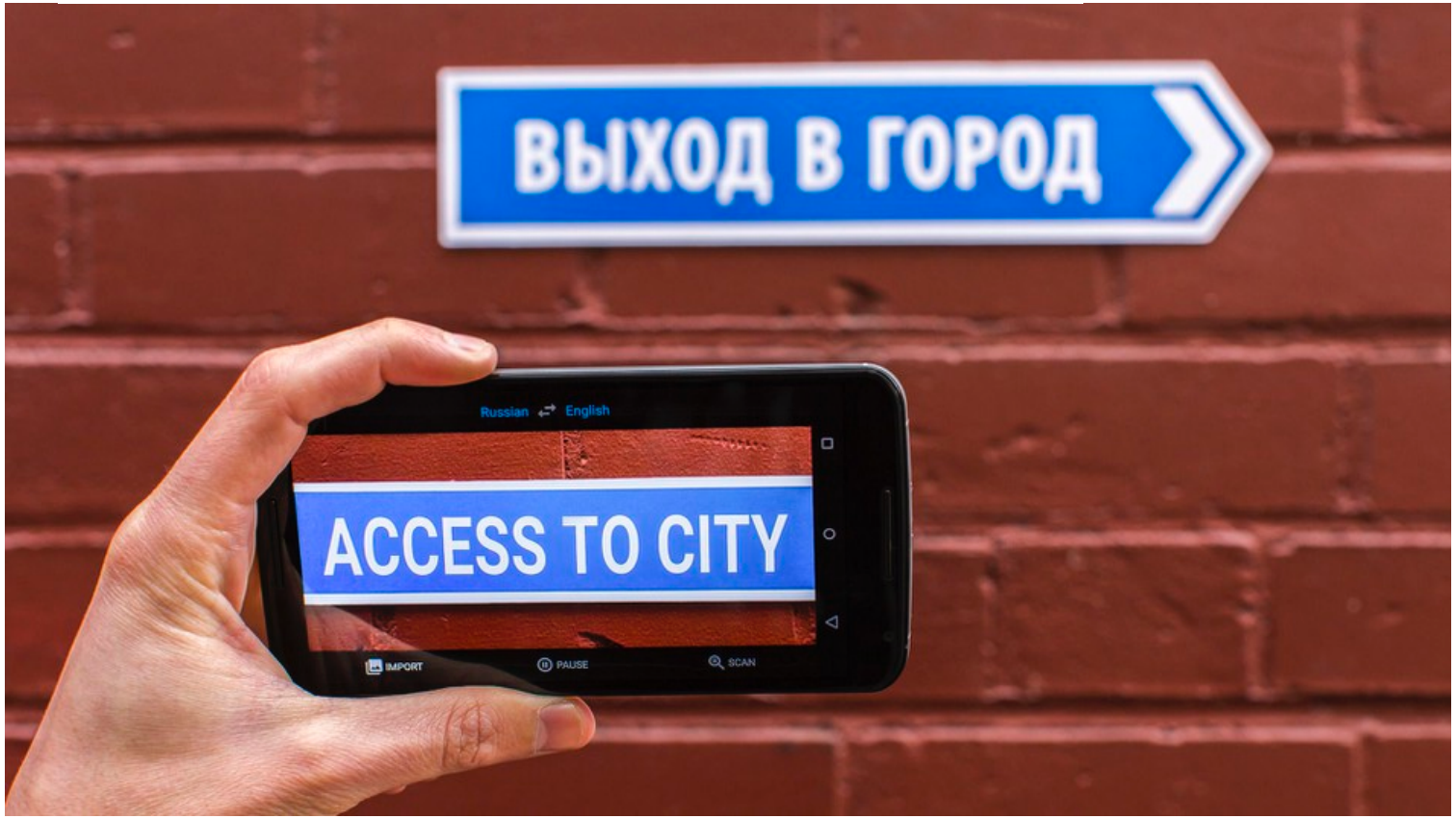
WHY?

February 2011



October 2011

Google Translate



Launched 2006

App 2011

WordLens 2015

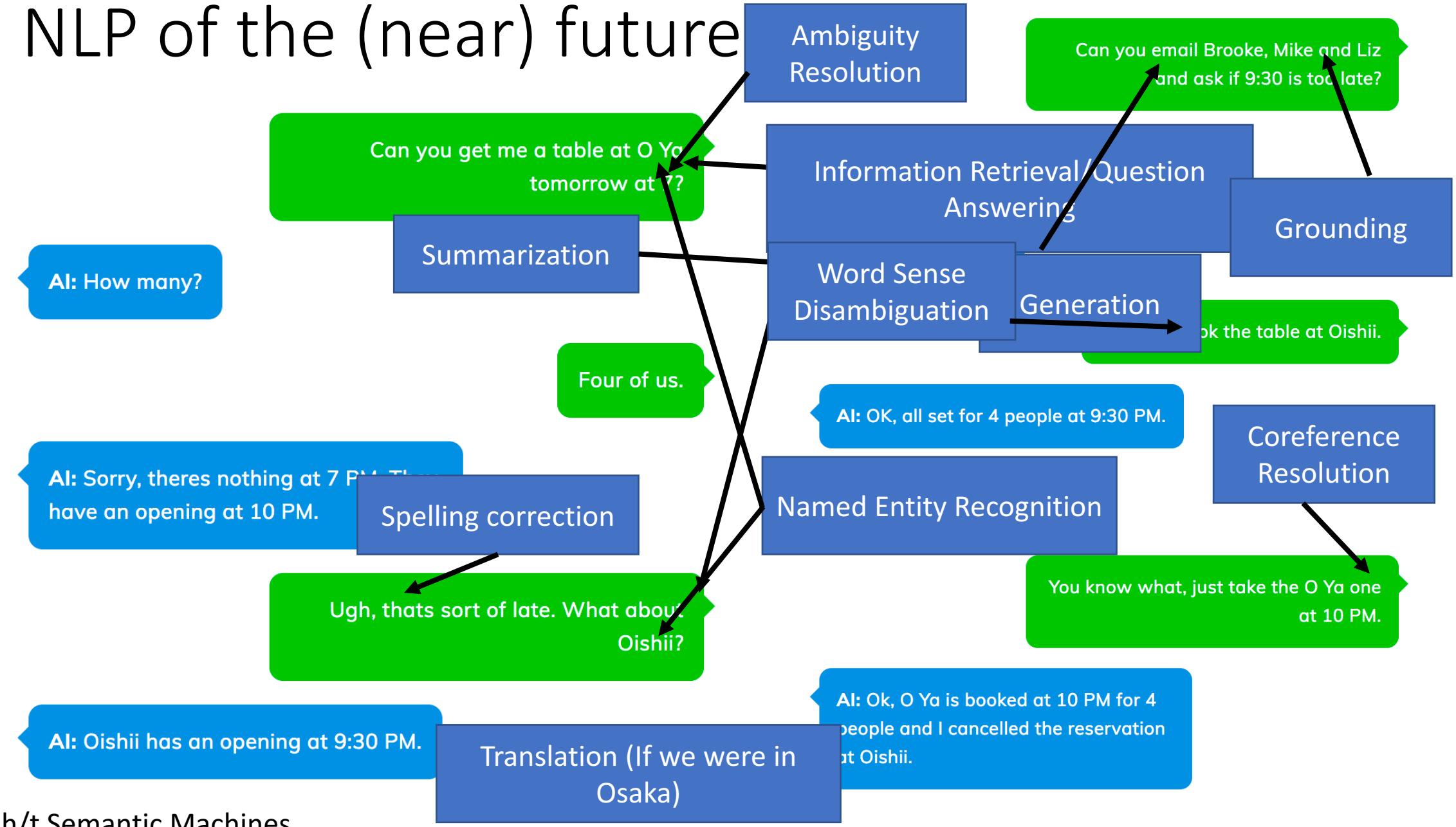
Where else have you seen NLP in your life,
in the news, or elsewhere?

Today's Applications

- Conversational agents
- Information extraction and question answering
- Machine translation
- Opinion and sentiment analysis
- Social media analysis
- Rich visual understanding
- Essay evaluation, plagiarism detection (WARNING)
- Mining legal, medical, or scholarly literature

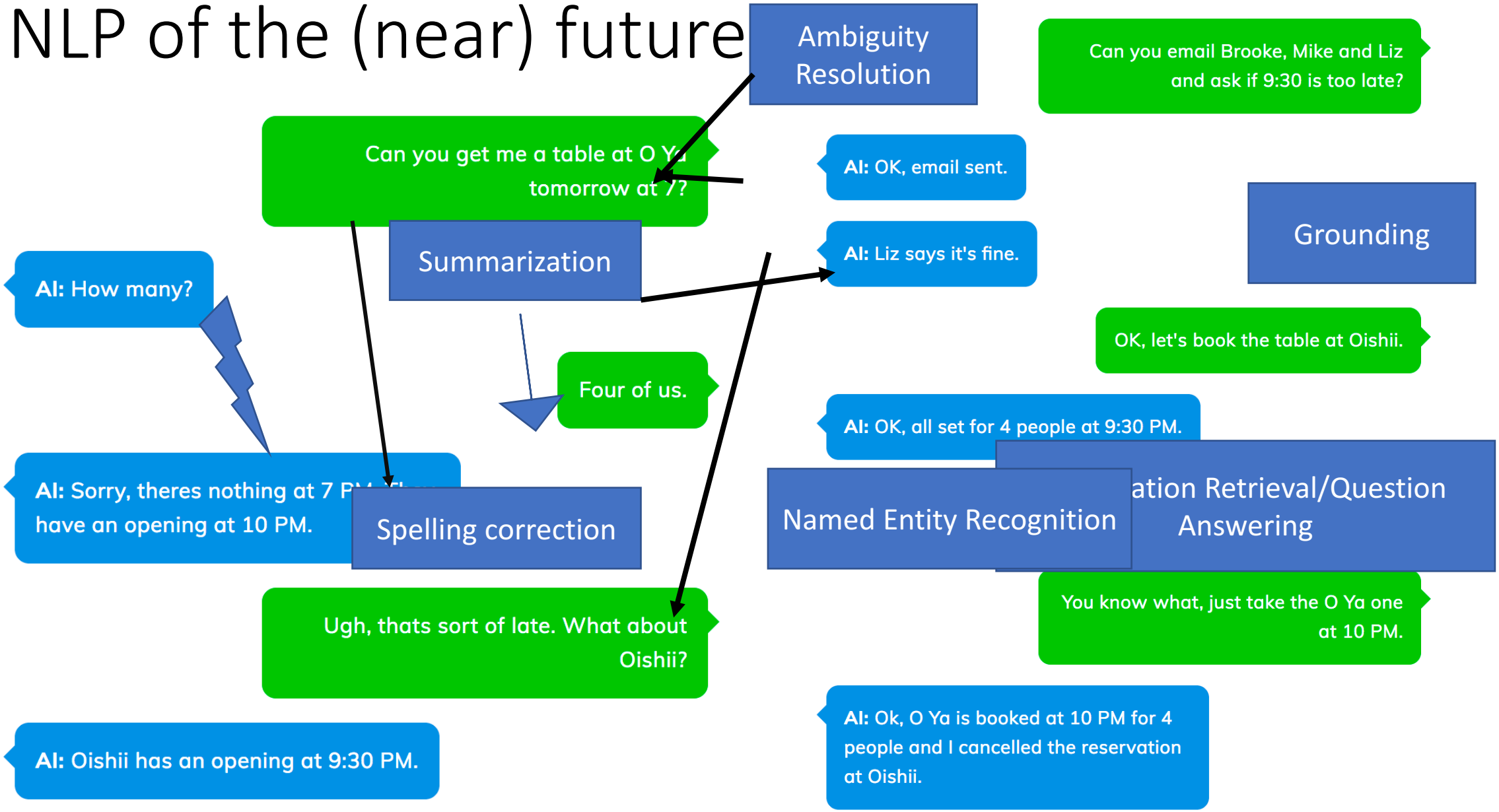
What tasks do you want to accomplish?

NLP of the (near) future



h/t Semantic Machines

NLP of the (near) future



h/t Semantic Machines

How do we (humans) do these tasks?

- Spelling Correction
- Named Entity Extraction
- Question Answering
- Coreference Resolution
- Grounding
- Ambiguity Resolution
- Summarization
- Translation

Long story short, you know languages!

What does it mean to "know" a language?

(What is the "true" R?)

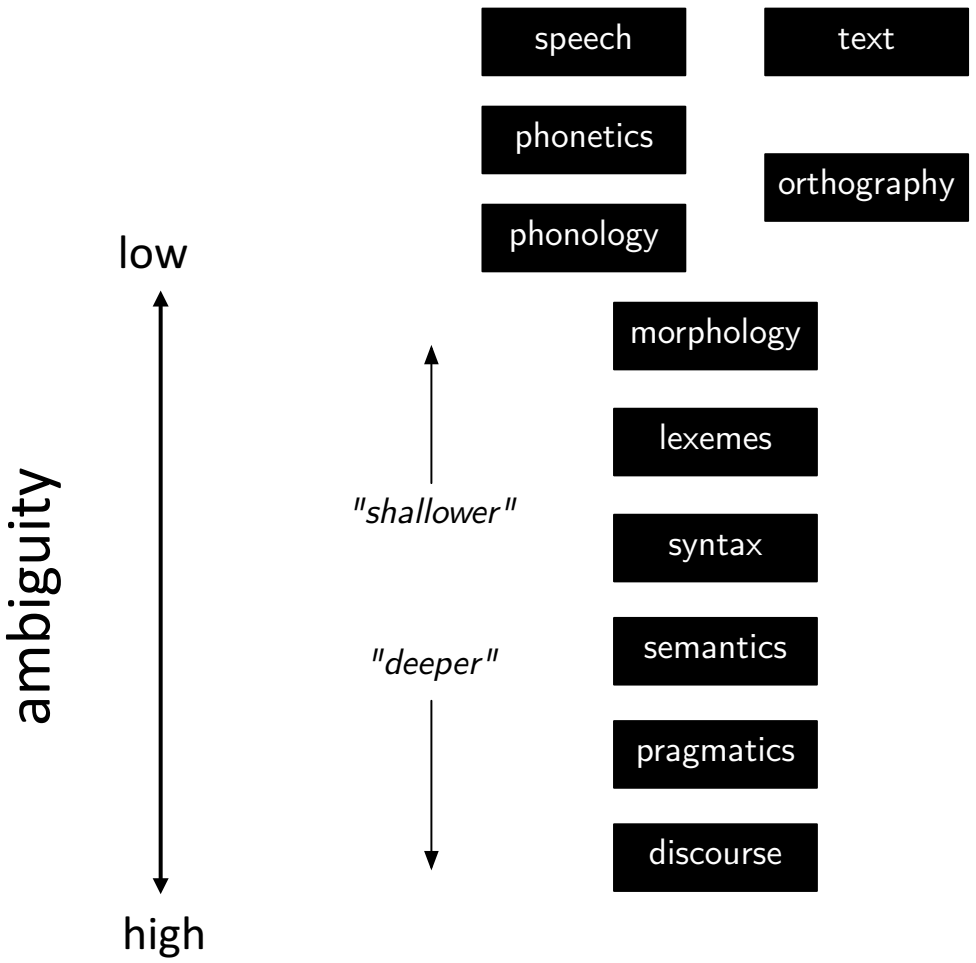
Do I know English?

Do I know Mandarin?

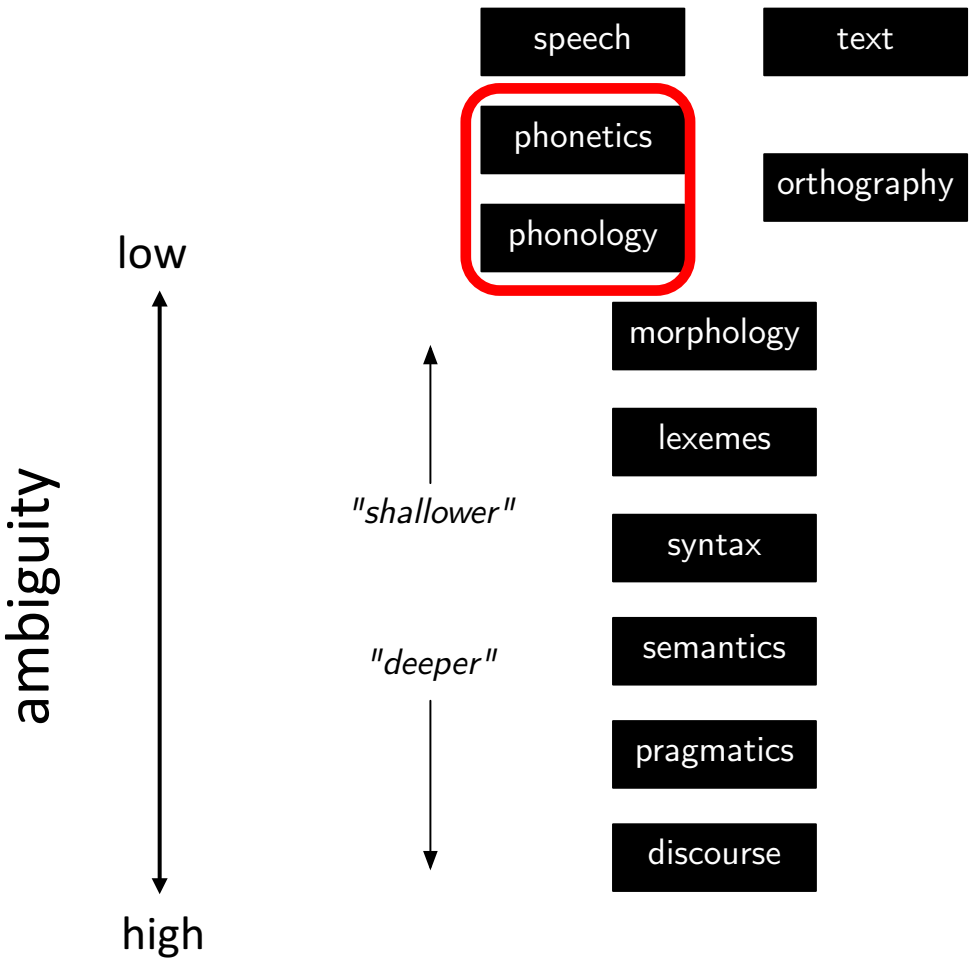
Does a toddler know English?

Does wc know English?

Levels of Linguistic Knowledge



Levels of Linguistic Knowledge



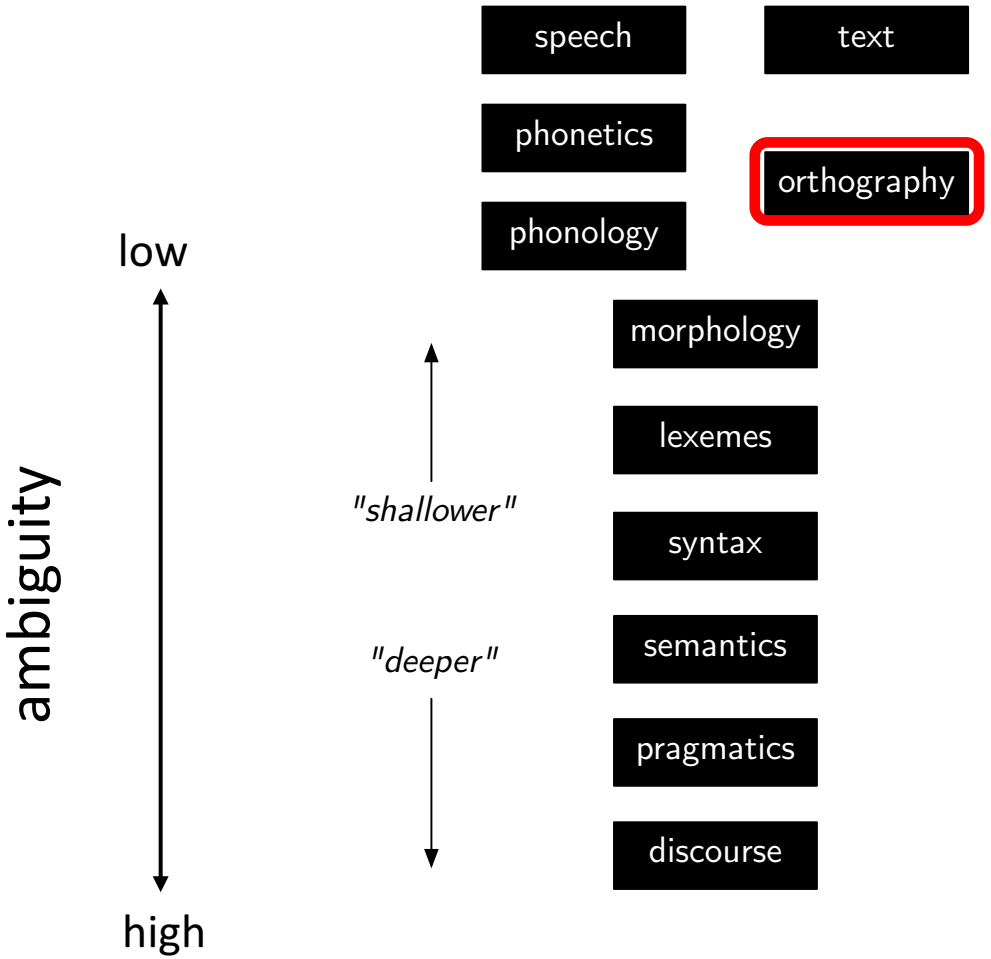
phones = distinct sounds
(governed by anatomy)

/l/ = alveolar lateral approximant **l**ace
/ɾ/ = alveolar tap **r**ace
/r/ = alveolar trill **r**ey (sp.)

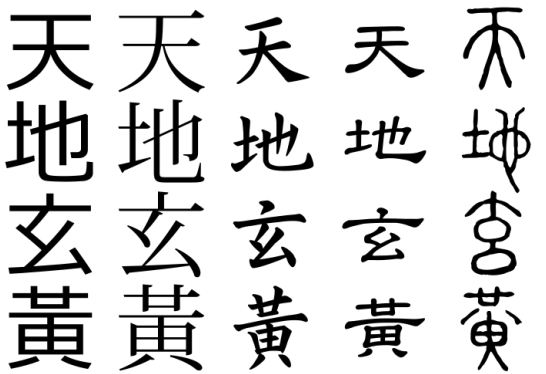
phonemes = meaningfully
distinct sounds
(governed by language)

English: /ɾ/ vs /r/ conflated
Japanese: /ɾ/ vs /r/ vs /l/ conflated
Hindi: /d̪/ vs /d̪ʰ/ distinct
Chinese: ma1 vs ma2 vs ma3 distinct

Levels of Linguistic Knowledge

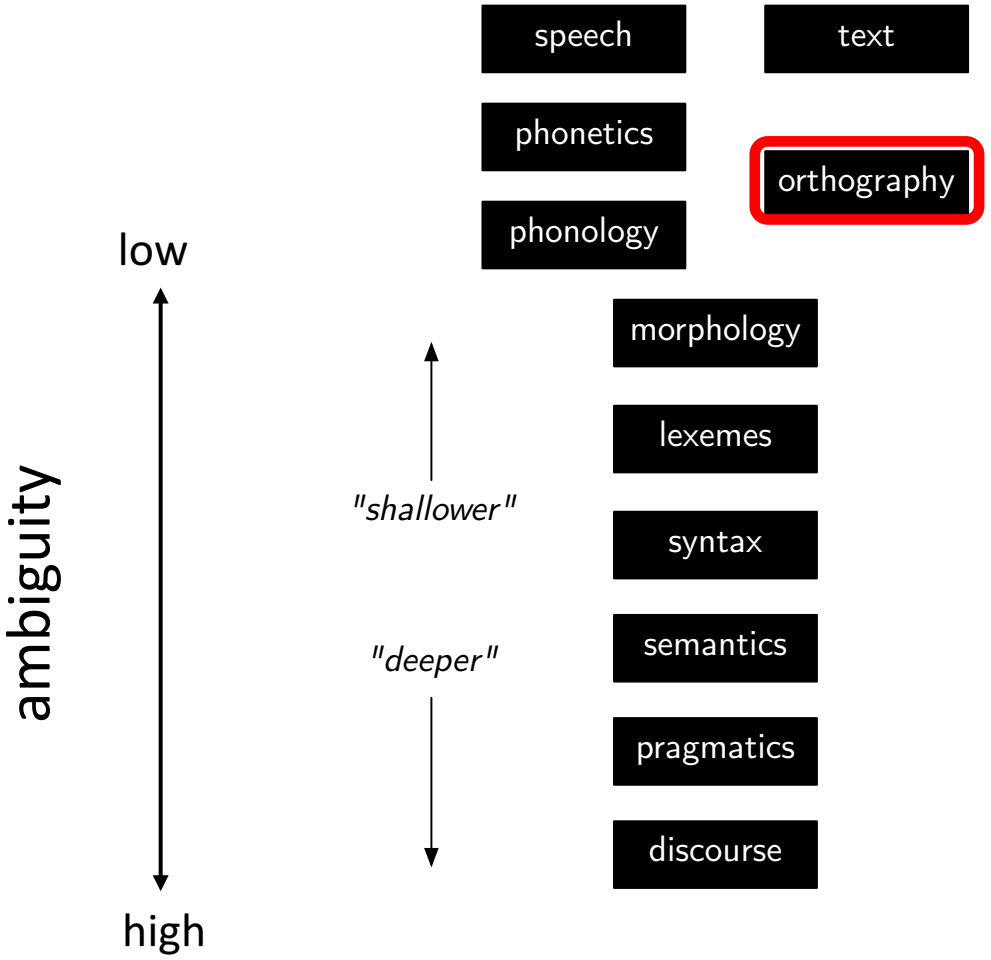


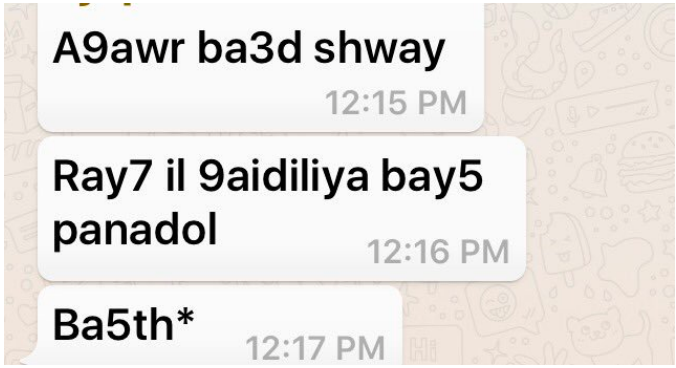
The New York Times

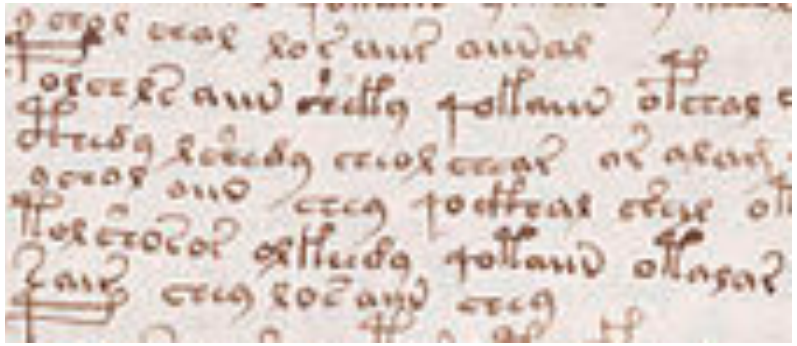


ሚስተር ቤንጀሚን መንግስቶም ንትግባረ’ቲ ሰነድ ሰላም ዝከአሎ ከምዝገበረ ንህዝቢ ደቡብ ሱዳንን ማሕበረ-ሰብ ዓለምን ዘርእየሉ ወቅቲ ኢና ንርከብ'ውን ኢሎም።

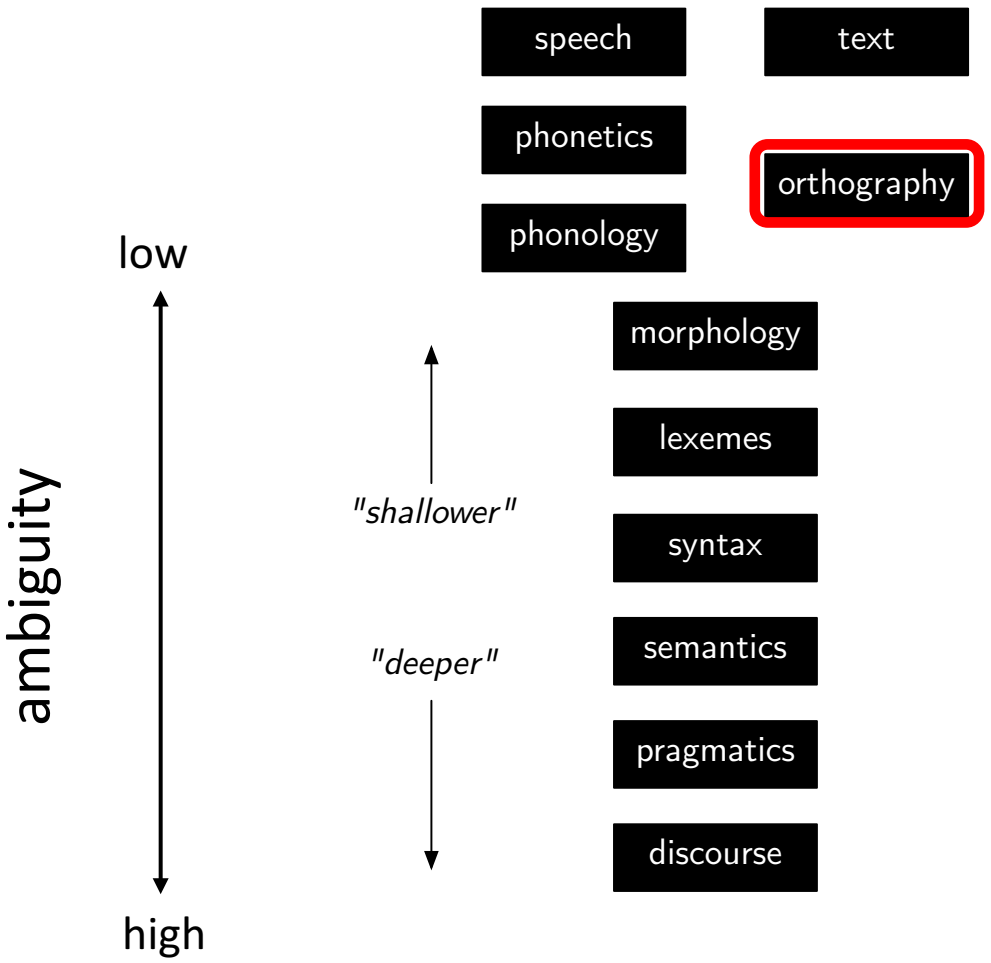
Levels of Linguistic Knowledge





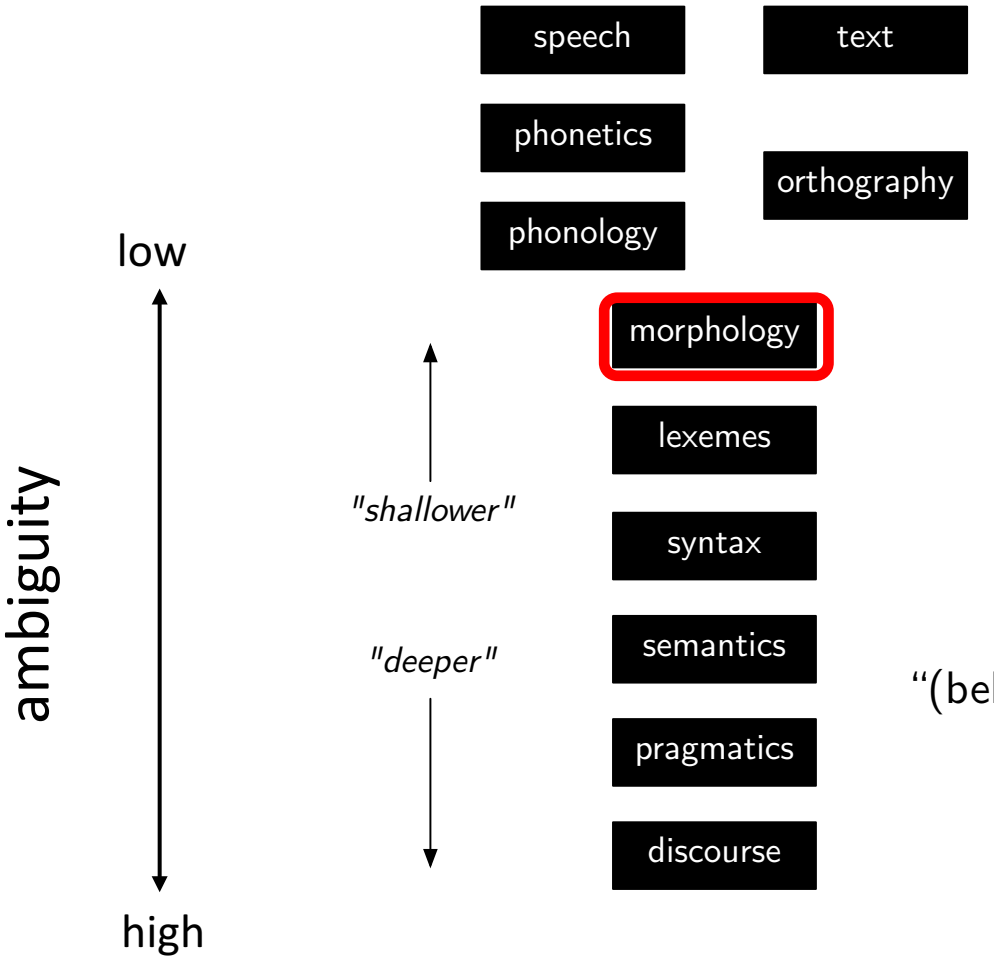


Levels of Linguistic Knowledge



ลูกศิษย์วัดกระทิงยังยื้อปิดถนนทางขึ้นไปนมัสการพระบาทเขาคิชฌกูฏ หวิดปะทะกับเจ้าถิ่นที่ออกมาเผชิญหน้าเพราะเดือดร้อนสัญจรไม่ได้ ผวจ.เร่งทุกฝ่ายเจรจาก่อนที่ชื่อเสียงของจังหวัดจะเสียหายไปมากกว่านี้ พร้อมเสนอหยุดจัดงาน 15 วัน....

Levels of Linguistic Knowledge



talk +s (present, 3rd person sing.)

talk +ed (past)

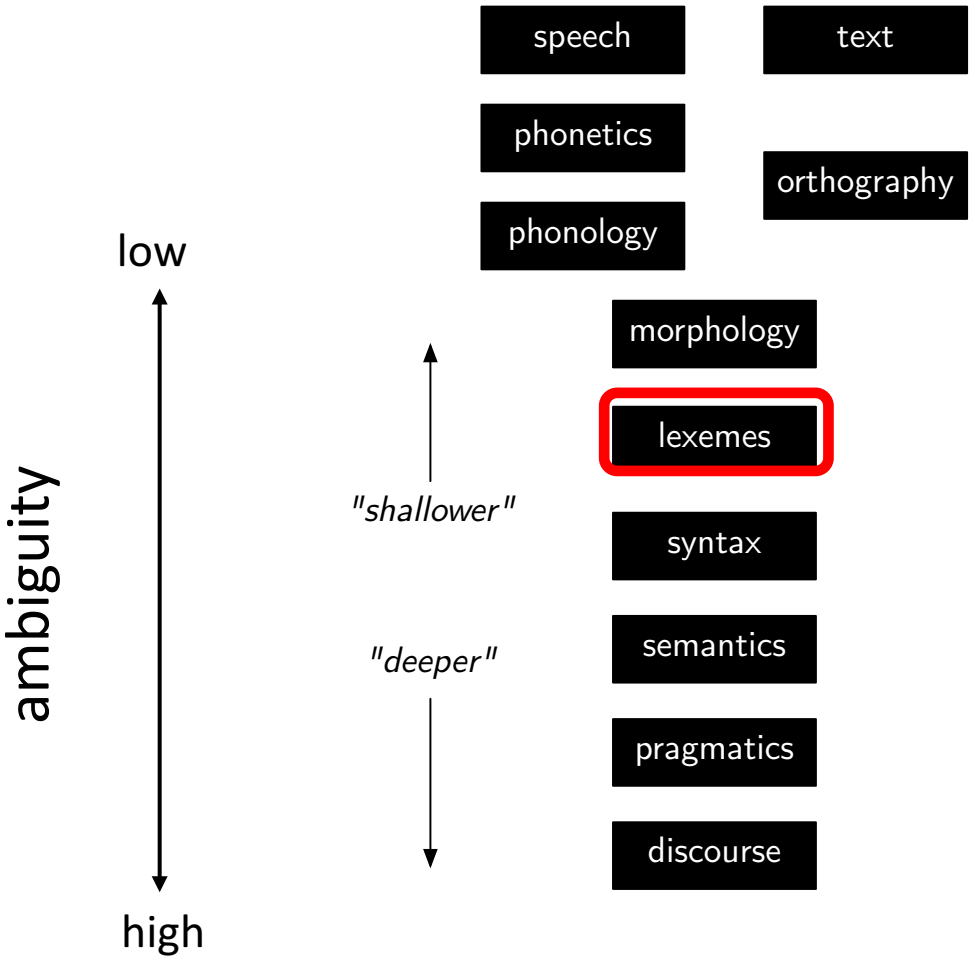
other languages with more morphology?

uygarlaştıramadıklarımızdanmışsınızcasına

"(behaving) as if you are among those whom we could not civilize"

unfriend, Obamacare

Levels of Linguistic Knowledge



What is a word?

single unit of meaning?

text separated by whitespace?

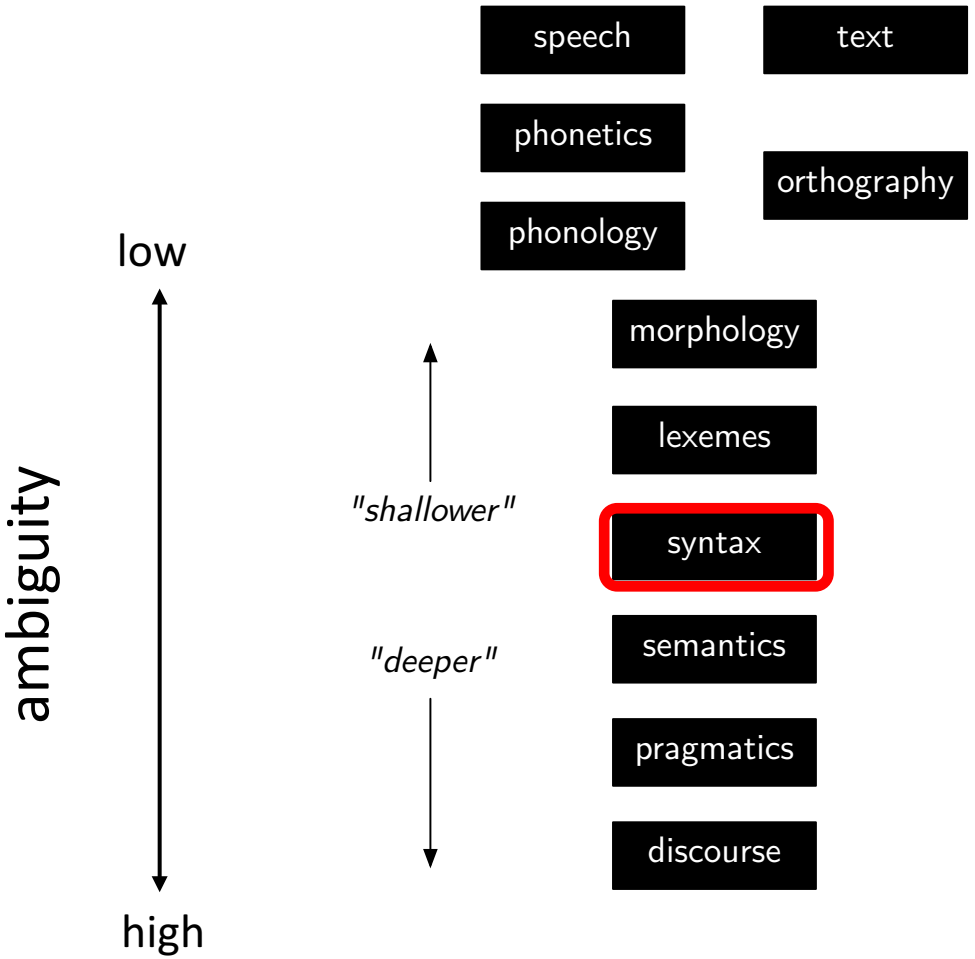
We have to make decisions!

Thai (or Chinese) example?

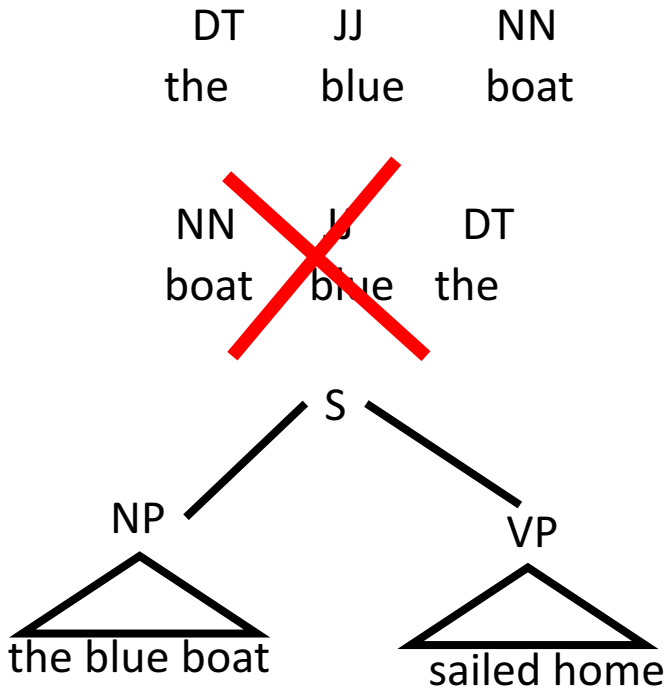
Turkish example?

non-compositional multi-word expressions:
New York, take out

Levels of Linguistic Knowledge

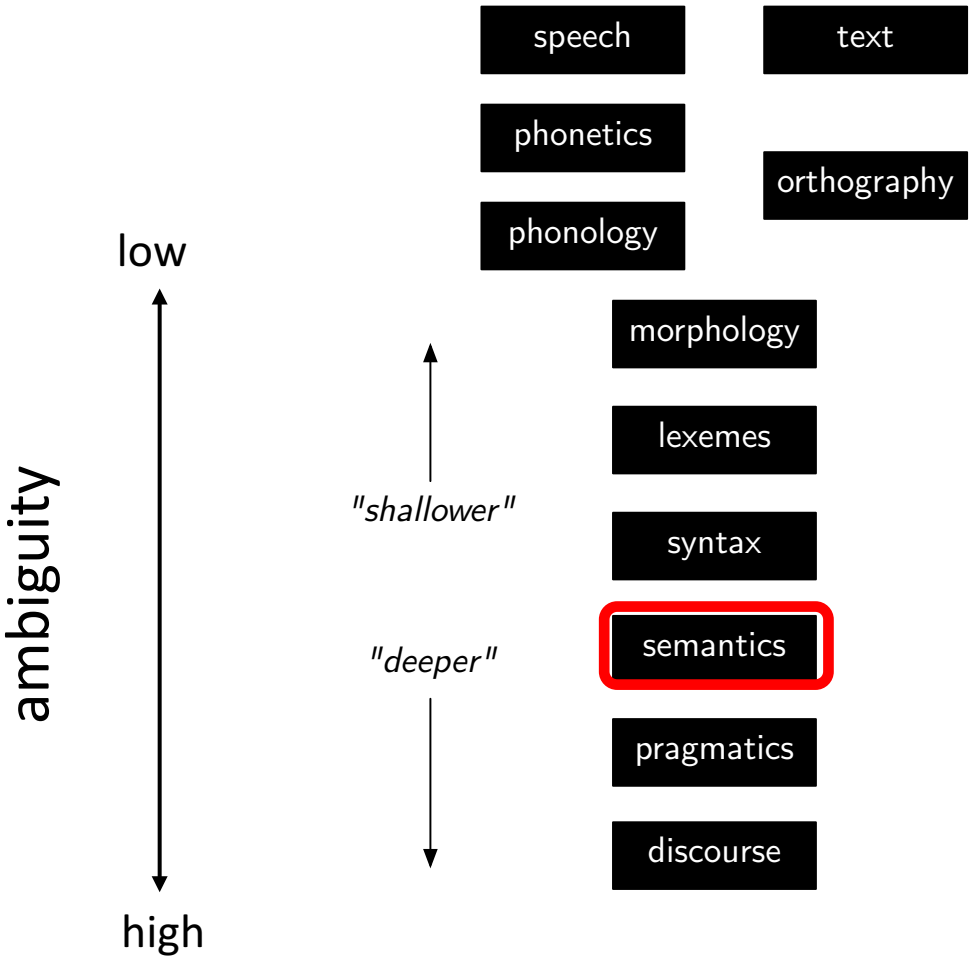


How do words fit together?



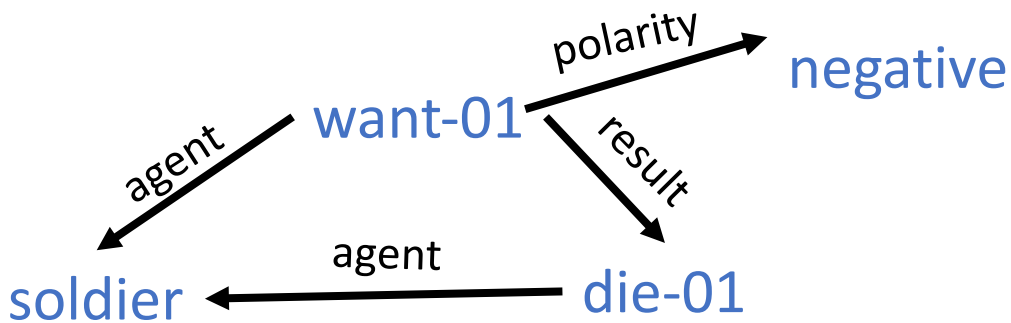
Note similarity to programming languages!

Levels of Linguistic Knowledge



What does a sentence mean?

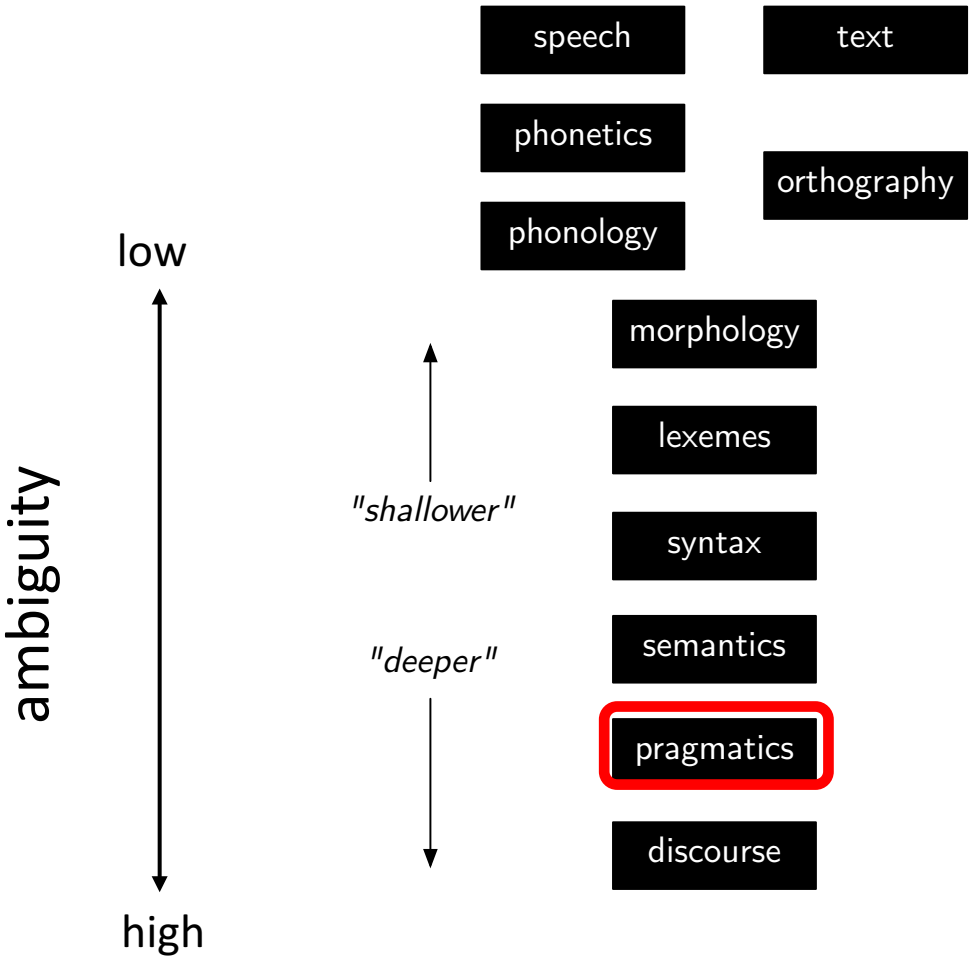
The soldier did not want to die.



want-01: "desire"
(not "lack")

die-01: "cease to live"
(not "want very much")

Levels of Linguistic Knowledge

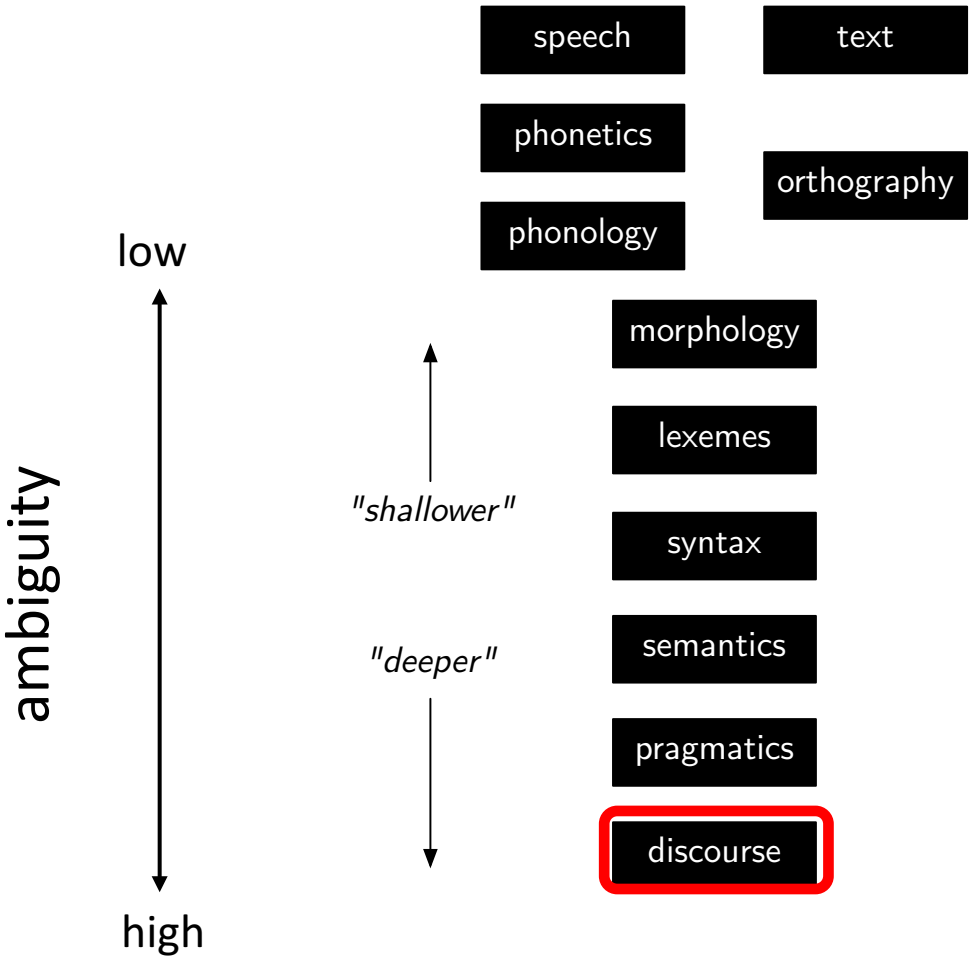


What does *the speaker* mean?

Can you get me a table at O Ya?

Yes, I have the ability to book reservations.

Levels of Linguistic Knowledge



What are the goals of the participants?

Can you get me a table at O Ya?

Would you like a wood or steel table?

Maybe many people, long period of time...

Ambiguity makes NLP hard

- meaning of *bank* or *mean* or *latex*
- *make a decision* or *take out* or *make up*

And Funny!

- Enraged Cow Injures Farmer with Ax
- Ban on Nude Dancing on Governor's Desk
- Teacher Strikes Idle Kids
- Hospitals are Sued by 7 Foot Doctors
- Iraqi Head Seeks Arms
- Stolen Painting Found by Tree
- Kids Make Nutritious Snacks
- Local HS Dropouts Cut in Half

h/t Dan Klein

Syntax + Semantics

We saw the woman with the telescope wrapped in paper.

- Who has the telescope?
- What is the paper wrapping?
- An event of perception or an assault?

Humans have one (or two) major readings of this;
hard to keep computers from getting the unlikely ones

Semantics

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

-- Groucho Marx

Richness makes NLP Hard!

- Lots of ways to express the same thing
- Sometimes people communicate in an intentionally ambiguous way
- There are many languages, styles, genres, modalities...

The soldier was not afraid to die $\stackrel{?}{=}$ The soldier did not fear death

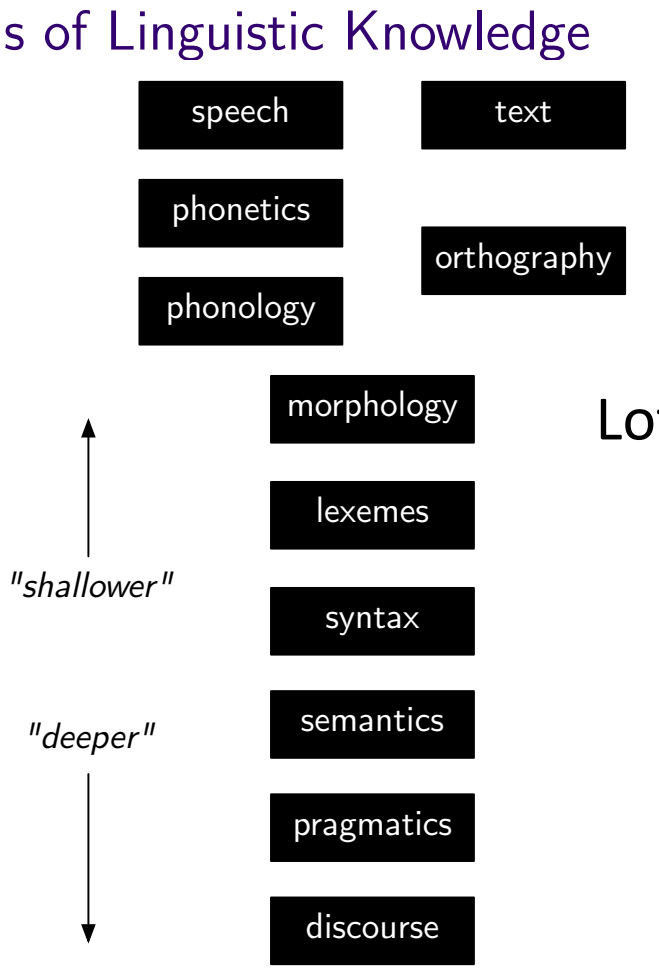
How many other ways can this be said?

"Call me at six niner i three triple 0 dos"

"u ship them? smh."
"ikr, lolol, yolo."

"14 words now and then sure 88 later if u want"

Uncertainty/Noise Propagated at Scale Makes NLP Hard!



Lots of ambiguity at every level from one string!

Representation makes NLP Hard!

Can $\mathcal{R}$ be "Meaning"?

Depends on the application!

- Giving commands to a robot
- Querying a database
- Reasoning about relatively closed, grounded worlds

Harder to formalize:

- Analyzing opinions
- Talking about politics or policy
- Ideas in science

what is the meaning in these cases?

more than just meaning is represented in these communications

What is the "real" R?

Why NLP is Hard

1. Mappings across levels are complex.
   - A string may have many possible interpretations in different contexts, and resolving **ambiguity** correctly may rely on knowing a lot about the world.
   - **Richness**: any meaning may be expressed many ways, and there are immeasurably many meanings.
   - Linguistic **diversity** across languages, dialects, genres, styles, ...
2. Appropriateness of a representation depends on the application.
3. Any $\mathcal{R}$ is a theorized construct, not directly observable.
4. There are many sources of variation and noise in linguistic input.

Desiderata for NLP Methods

(ordered arbitrarily)

1. Sensitivity to a wide range of the phenomena and constraints in human language
2. Generality across different languages, genres, styles, and modalities
3. Computational efficiency at construction time and runtime
4. Strong formal guarantees (e.g., convergence, statistical efficiency, consistency, etc.)
5. High accuracy when judged against expert annotations and/or task-specific performance

How can we evaluate some of the tasks proposed?

NLP $\stackrel{?}{=}$ Machine Learning

- To be successful, a machine learner needs bias/assumptions; for NLP, that might be linguistic theory/representations.
- $\mathcal{R}$ is not directly observable.
- Early connections to information theory (1940s)
- Symbolic, probabilistic, and connectionist ML have all seen NLP as a source of inspiring applications.

NLP $\stackrel{?}{=}$ Linguistics

- NLP must contend with NL data as found in the world
- NLP $\approx$ computational linguistics
- Linguistics has begun to use tools originating in NLP!

Fields with Connections to NLP

- Machine learning
- Linguistics (including psycho-, socio-, descriptive, and theoretical)
- Cognitive science
- Information theory
- Logic
- Theory of computation
- Data science
- Political science
- Psychology
- Economics
- Education

People from other fields!
Why are you here?

The Engineering Side

- Application tasks are difficult to define formally; they are always evolving.
- Objective evaluations of performance are always up for debate.
- Different applications require different $\mathcal{R}$.
- People who succeed in NLP for long periods of time are foxes, not hedgehogs.

Factors Changing the NLP Landscape

(Hirschberg and Manning, 2015)

- Increases in computing power
- The rise of the web, then the social web
- Advances in machine learning
- Advances in understanding of language in social context

What You Will Learn In This Course (I hope)

- How the NLP you interact with on a daily basis works
- The various models, algorithms, and tools that are out there to solve language-related problems you want to tackle
- How to design and evaluate your own task/model/algorithm/tool
- The currently open research questions, so you can pursue further study

Questions about Intro?

Next is Boring Course Stuff That Is Very Important

The Team

- Jon (instructor)
  - Asst rsrch prof since 2015, USC PhD 2010, worked in NLP since 2001, MT, IE, Semantics, ML
  - I spend most of my time at ISI in Marina del Rey
- Ramesh Manuvinakurike (TA)
  - PhD student since 2013 at ICT studying dialogue systems
- Siddharth Jain (TA)
  - PhD student since 2012 at ICT studying
- Dong Guo (TA)
  - PhD student (ML), dialogue/QA industry experience
- Sneha Salvi (Grader)
- Xiang Zhang (Grader)

Support

- DO NOT EMAIL US (sorry)
- Piazza Discussion Forums
  - TAs and Instructor will monitor regularly, answer questions, engage in discussion
  - You should answer each others' questions and help your fellow students out!
  - Outstanding student contributors will receive a grade bump-up as extra credit (e.g. B to B+, B+ to A-) [see extra credit details in 4 slides]
- Office Hours
  - Jon: 10am-11am Wed & Fri (i.e. after class), RTH 512; 11am-12pm possible if there is demand
  - Siddharth 8am-10am Mon, SAL lab
  - Dong 12pm-2pm Tue, EEB 220
  - Ramesh 8am-10am Thu, SAL lab

Syllabus and Schedule

- go to "jonmay.net" and click on the link for the class website
- Schedule will probably change depending on our speed; check back frequently
- No official textbook; readings will be posted on the class website/Blackboard
- Class through 12/1 except:
  - Fri, 9/22 No Class (Instructor religious holiday)
  - Fri, 10/6 MIDTERM (note: see me after class if you have religious holiday)
  - Week of 11/22, 11/24 No Class (Thanksgiving)

Lecture and Notes

- You don't have to come to class if you don't want to/can't
  - If you do, please pay attention and participate!
- However you are responsible for everything covered in class, and
  - It won't be recorded
  - Slides may not cover everything I discuss
- Slides will be posted soon after class
- I use a lot of slides from other classes and note this; feel free to self educate. In particular check out videos from:
  - Noah Smith (Univ. Washington)
  - Adam Lopez (Univ. Edinburgh)
  - Dan Jurafsky (Stanford)

Prerequisites

- I expect you to program at the level of a CS undergrad senior or better
- Most of the assignments will be in Python
- There will be basic probability and statistics, which will be reviewed as needed

Homeworks

- 8 homeworks, 2 weeks to do each one, they overlap (i.e. you will receive hw2 before hw1 is due; cf. syllabus). 7.5% of your grade each.
- No homework will be assigned or due right before midterm/final
- Mostly programming assignments submitted to Vocareum (you should have received an email opening your account, let us know if not!)
- Some written assignments submitted electronically
- You can have 4 late days total over the whole course...
  - ...but no more than 2 per assignment
- Late homeworks thereafter are penalized by 50% for the next 24 hours, then not accepted

Exams

- Midterm: October 6 (please notify me if it's a religious holiday ASAP)
  - 15% of grade
  - short answers, maybe multiple choice, some derivations, some pencil & paper calculations
  - one double-sided page of notes allowed (may be prepared with other current 544 students)
- Final: December 6, 8-10am
  - 25% of grade
  - like the midterm, but can cover the whole class (emphasis will be on second half)
  - one double-sided page of notes allowed (may be prepared with other current 544 students)

Extra Credit

- Outstanding forum contributors may have their grade bumped a category (e.g. B- to B, B+ to A-).
- The number of and determination of such contributors is up to the staff and is not eligible for regrade.
- Occasionally there may be extra credit points in a homework. These will offset other point losses *in that homework* (i.e. they do not affect other homeworks/exams and cannot result in a >100% score).

Grading/Regrading Policy

- Grades will be issued one week after the exam/homework.
- No changes are allowed to submitted homework (after the deadline)
- If something is clearly wrong, you may request specific regrade of a specific question/part via a google form the TAs will send out.
- WARNING: If you are just 'fishing' for points you may LOSE additional points. No grubs!

Cheating

- You MAY
  - talk with other students, friends, or others about your homework assignments *IF* you acknowledge such discussion in your submission
  - ask questions about the homework and subject material in the forums
- You MAY NOT
  - copy code or answers from any source including friends, homework/test services, NLP or other software libraries. This includes making slight changes to previously written code
  - hack the scoring servers, Kobiyashi Maru-style
  - *allow your code to be copied, even if unintentionally*
  - attempt to communicate with or read from any other person while taking exams

Cheating

- Unfortunately, about 20 of you will be caught cheating, based on previous experience.
- By the end of this course you should have a pretty good idea of how we do it! (Hint: We use NLP to do it)
- Suspected cheats (*including those who were plagiarized from)* will be reported to the University. Punishment includes but is not limited to:
  - zero on assignment, exam, or class
  - Loss of career services privileges
  - Loss of CPT rights
  - Uncomfortable meeting with Lizsl

Next Time...

- Corpora and Text Processing
- Bring a laptop; we'll do in-class data manipulation, minor coding
- Make sure you have been invited and can get into vocareum