Introduction

Computational Linguistics

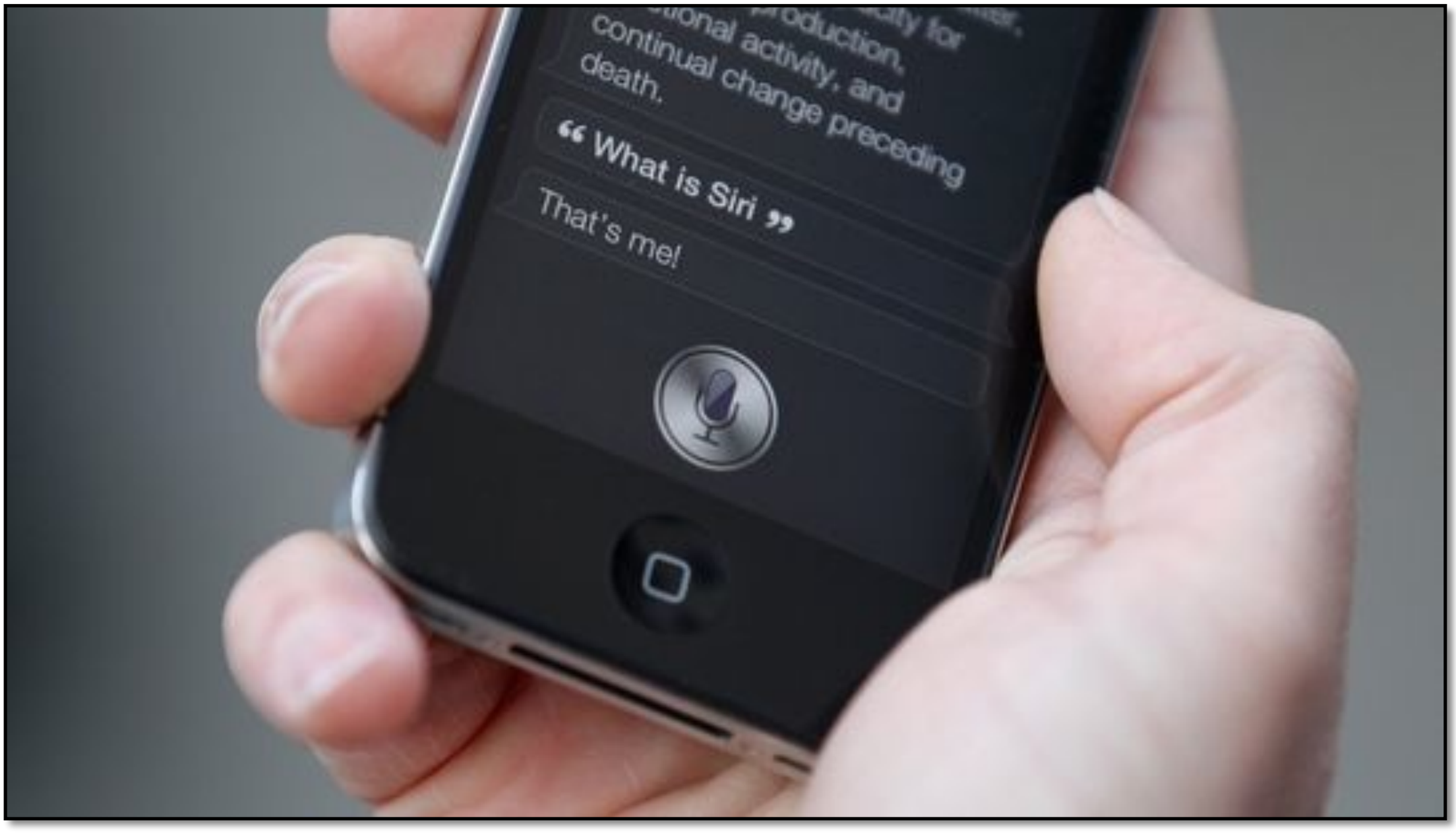
Alexander Koller

23 October 2018

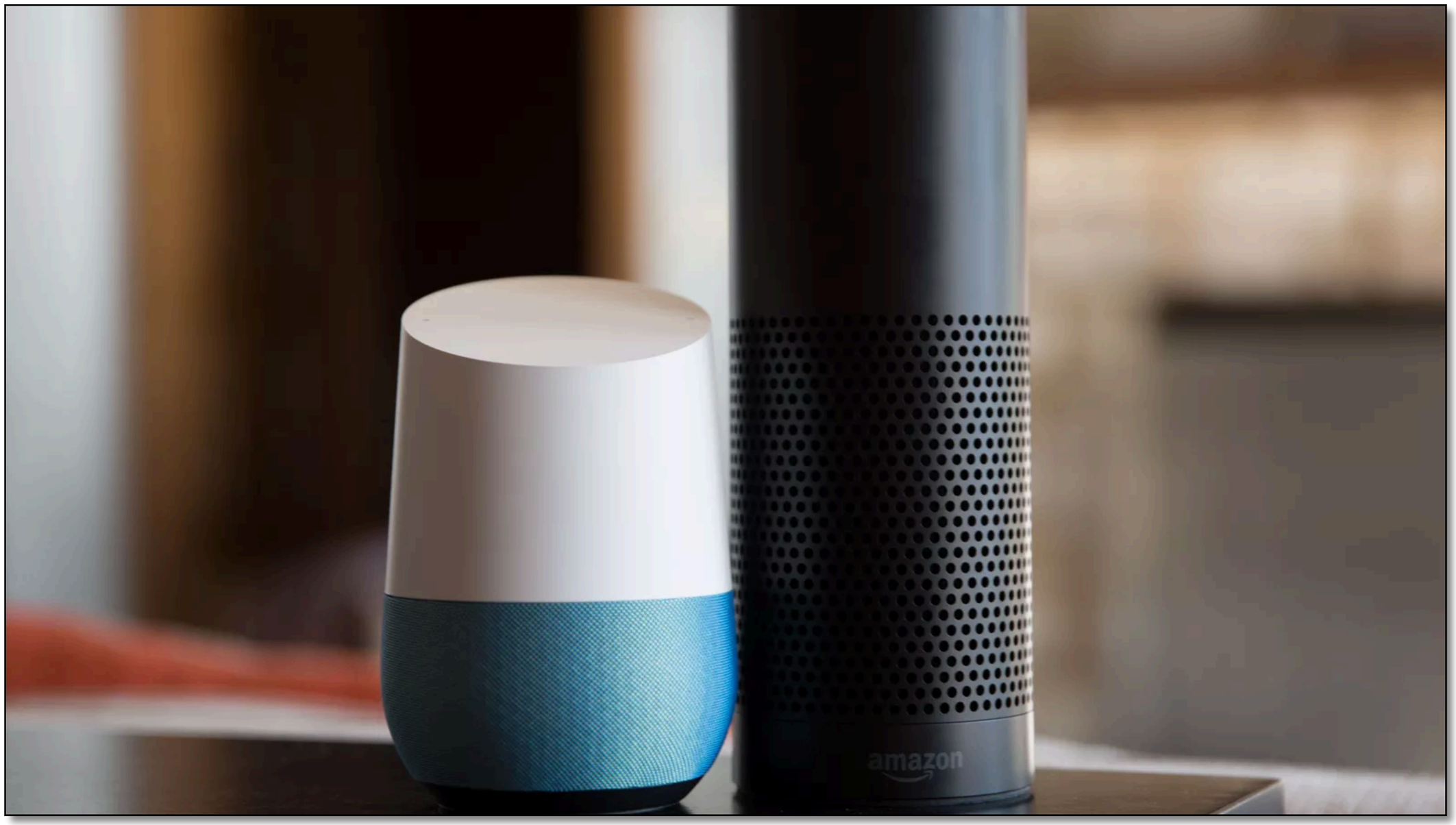
Outline

- What is computational linguistics?
- Topics of this course
- Organizational issues

Siri

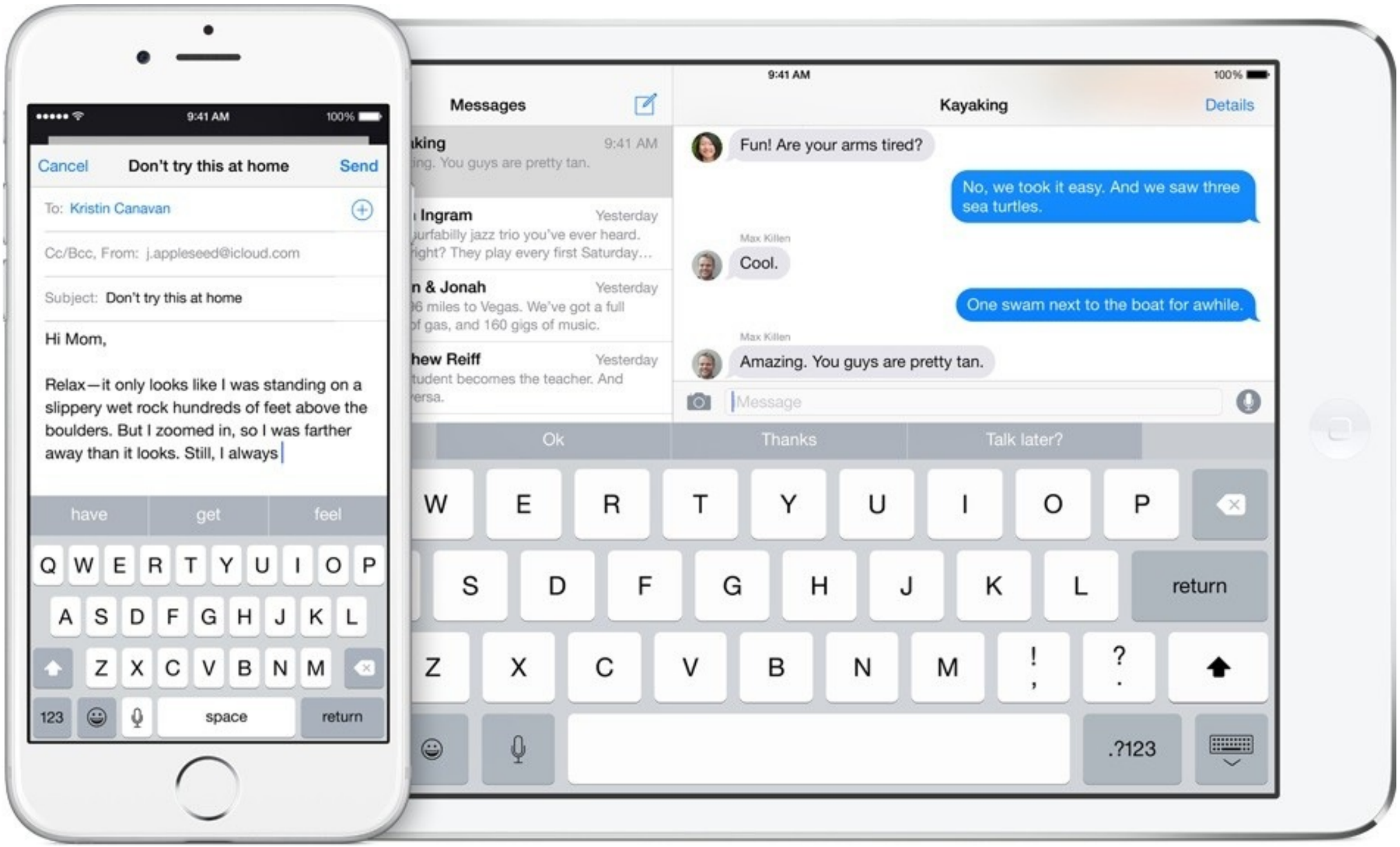


Digital assistants

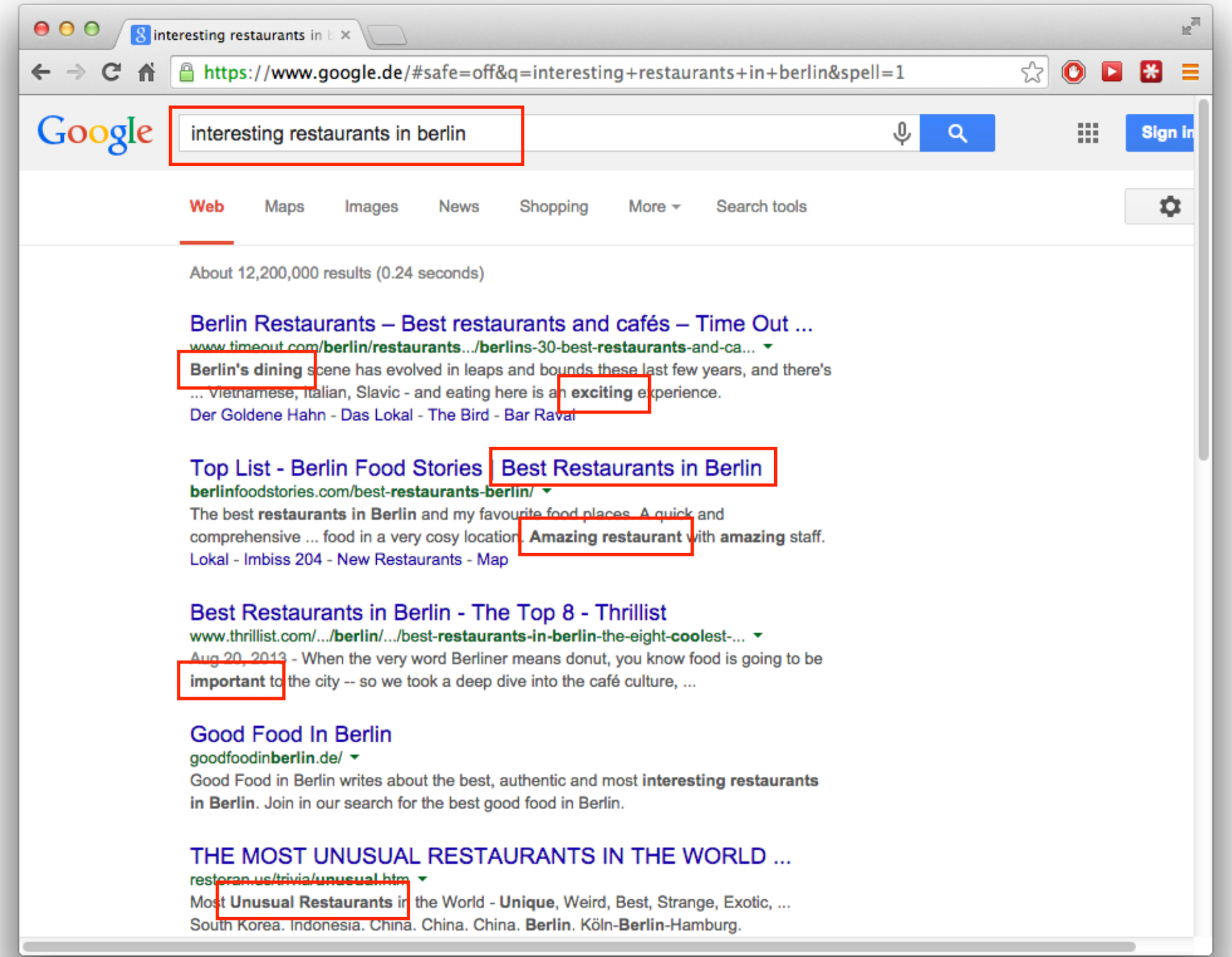


Google Home Amazon Echo

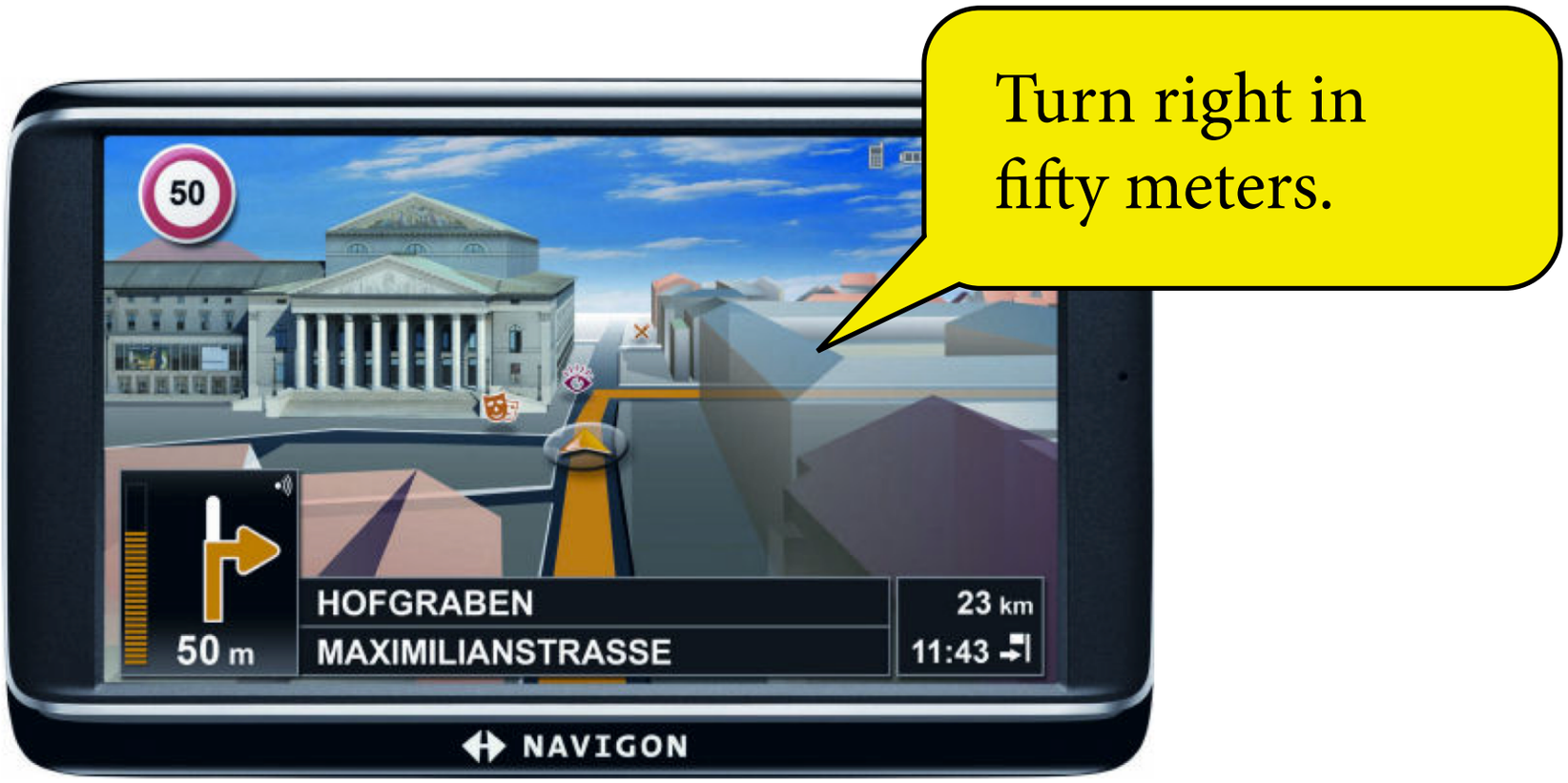
Text prediction



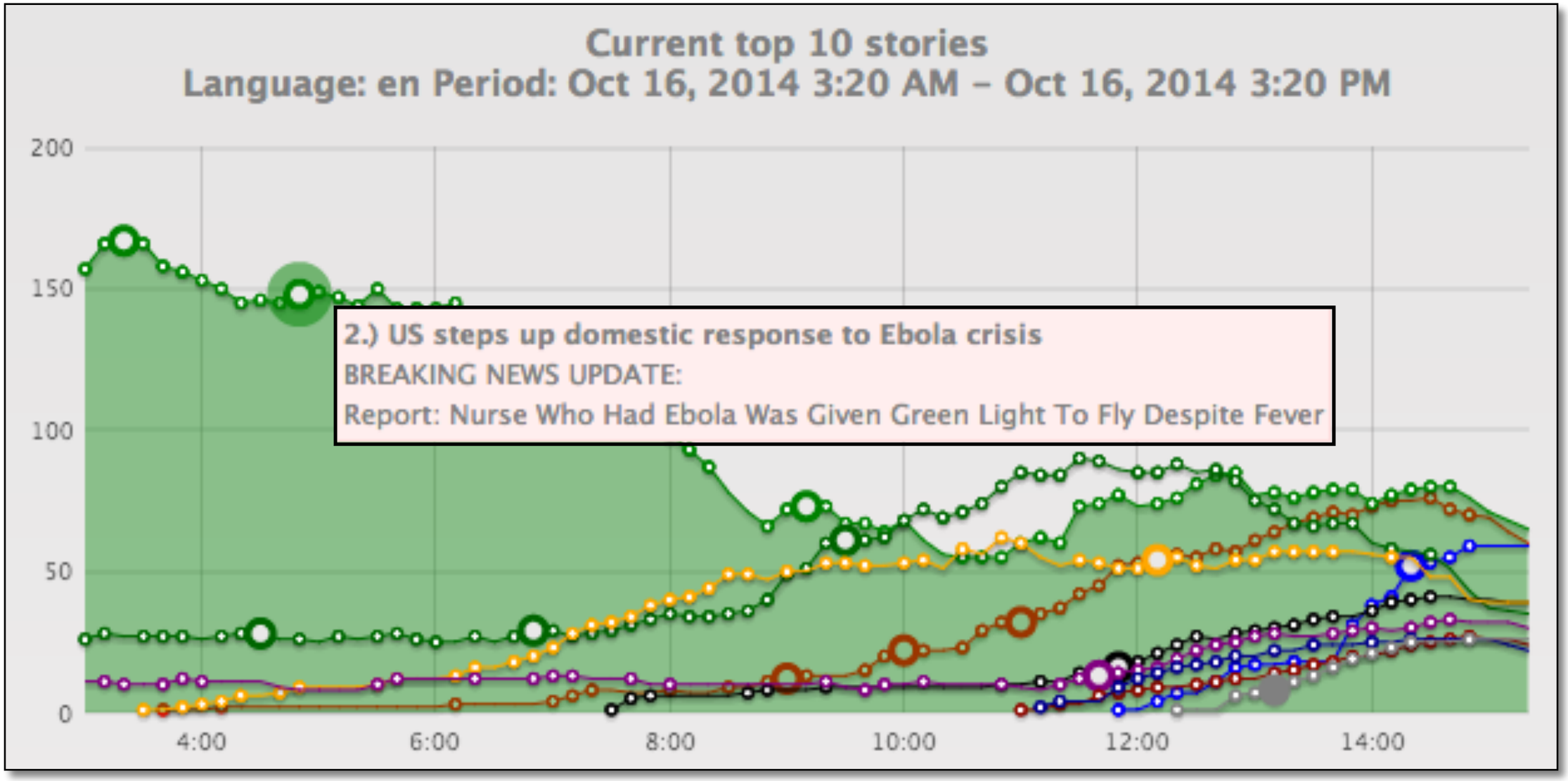
Similarity in Google Search



Navigation Systems

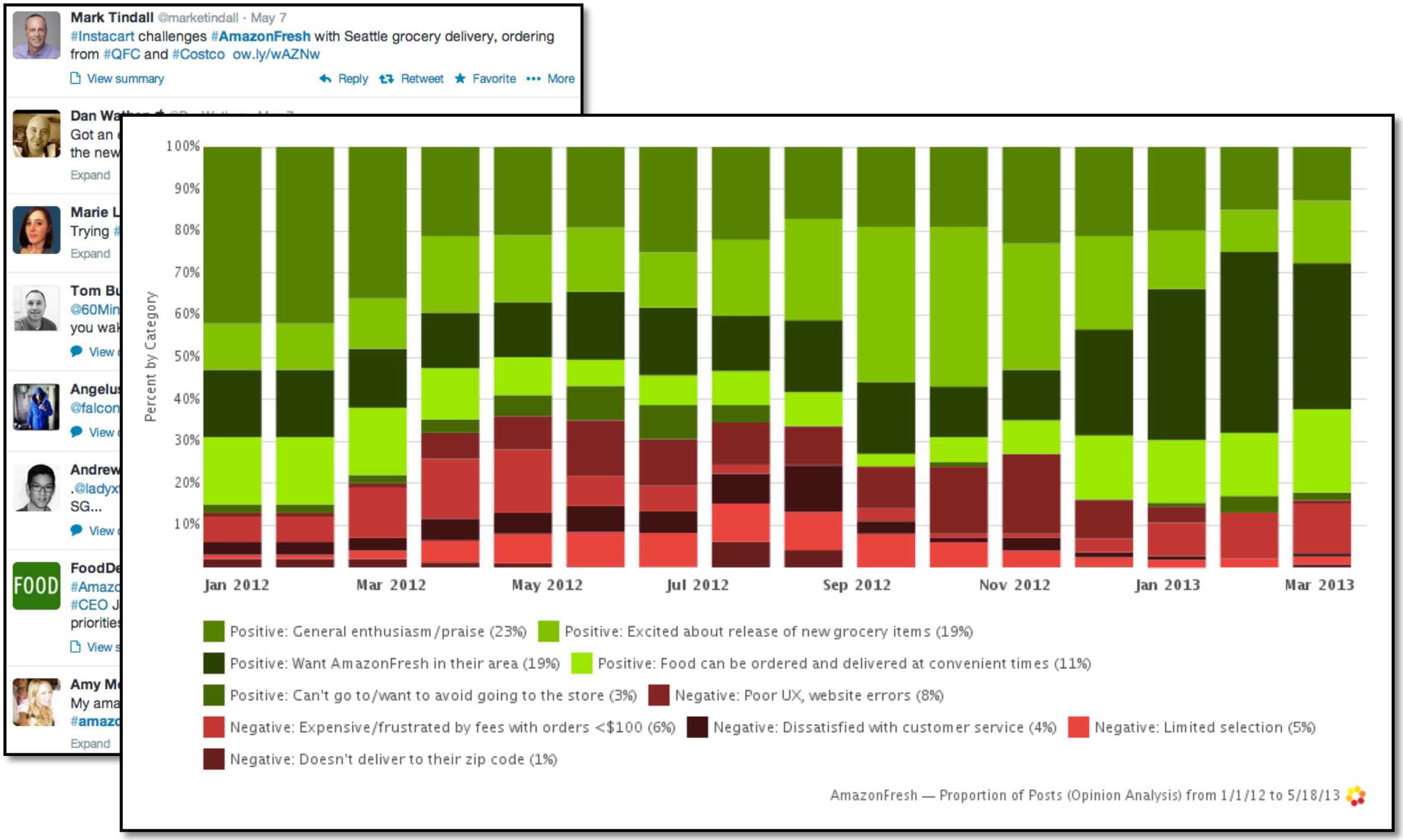


Information access



European Media Monitor, http://emm.newsbrief.eu

Sentiment Analysis



Crimson Hexagon, http://www.crimsonhexagon.com/

Clickbait generation with RNNs

http://clickotron.com/

Google Translate

EL PAÍS PORTADA INTERNACIONAL POL

Google Anmelden

Übersetzer

"The medium, who has enjoyed always, the coach's trust, and has recovered from the rupture of the cruciate ligament and from the inside of the right knee, which …"

Spanisch Deutsch

Joachim Löw, seleccionador de Alemania, ha anunciado este jueves la lista de los 30 jugadores preseleccionados para acudir al Mundial de Brasil, en la que destacan la ausencia del futuro portero del Barcelona Ter Stegen, y la incorporación de Sami Khedira, del Real Madrid. El medio, que siempre ha contado con la confianza del seleccionador, ya se ha recuperado de la rotura del ligamento cruzado y el interior de la rodilla derecha que se produjo durante un amistoso ante Italia en el mes de noviembre y que le ha mantenido apartado del terreno de juego durante siete meses.

Joachim Löw , Deutschland, am Donnerstag angekündigt, die Liste der 30 Spieler in die engere Wahl , die Weltmeisterschaft in Brasilien, die die Abwesenheit von zukünftigen Barcelona -Torhüter Ter Stegen, und der Einbau von Sami Khedira von Real Madrid gehören zu besuchen. Das Medium , das immer genossen hat, das Vertrauen des Trainers, und hat sich von der Kreuzbandriss und der Innenseite des rechten Knies , die bei einem Freundschaftsspiel gegen Italien im November aufgetreten erholt und er hat sich von der gehalten Feld für sieben Monate.

0
Enviar
Imprimir
Guardar
de Sami Khedira, del Real Madrid. El medio, que siempre ha contado con la confianza del seleccionador, ya se ha recuperado de la rotura del ligamento cruzado y el interior de la rodilla derecha que se produjo durante un amistoso ante Italia en el mes de noviembre y que le ha mantenido apartado del terreno de juego durante siete meses.

elpais.com, May 2014

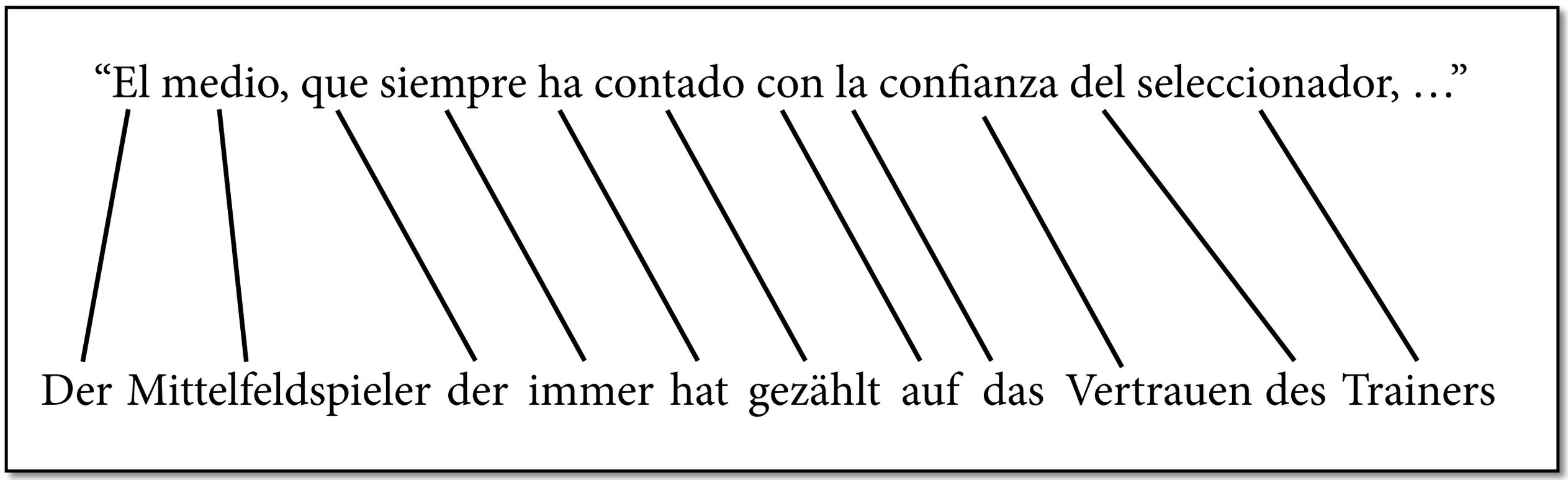
Lexical Ambiguity

"El medio, que siempre ha contado …"

"Medium"
(medium)

"Mittelfeldspieler"
(midfield player)

Word order



Translation ≈ choose words in the other language
and bring them in the correct order

Word order

"El medio, que siempre ha contado con la confianza del seleccionador, …"

Der Mittelfeldspieler der immer auf das Vertrauen des Trainers gezählt hat

Translation ≈ choose words in the other language
and bring them in the correct order

Structural Ambiguity

“se	ha	recuperado	de	la	rotura	del	ligamento	cruzado	y	el	interior	de	la	rodilla derecha"
has	himself	recovered	of	the	rupture	of the	ligament	cruciate	and	the	interior	of the	knee	right

el	ligamento	cruzado	y	el	interior	de	la	rodilla	derecha					

the cruciate ligament — and — the inside of the right knee

el	ligamento	cruzado	y	el	interior	de	la	rodilla	derecha					

cruciate and lateral

the cruciate and the lateral ligament — of the right knee

Content of this class

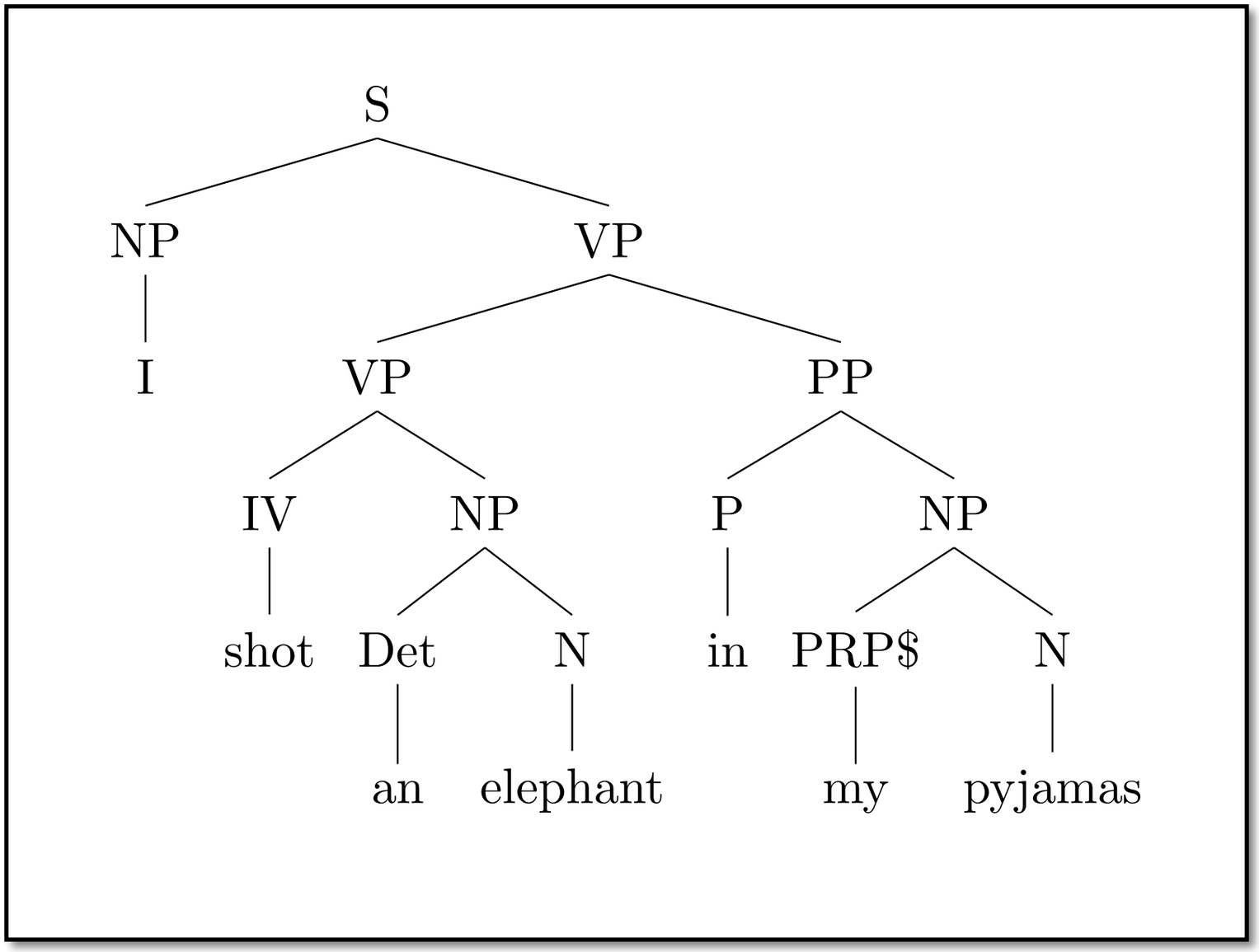
- Fundamental techniques of CL.
- Common themes:
  - uncovering hidden linguistic structure, represented symbolically
  - dealing with ambiguity
  - statistical methods
  - efficient algorithms

Recovering parts of speech

NNP	VBZ	NN	NNS	CD	NN
Fed	raises	interest	rates	0.5	percent

(POS tags from Penn Treebank)

Recovering syntactic structure

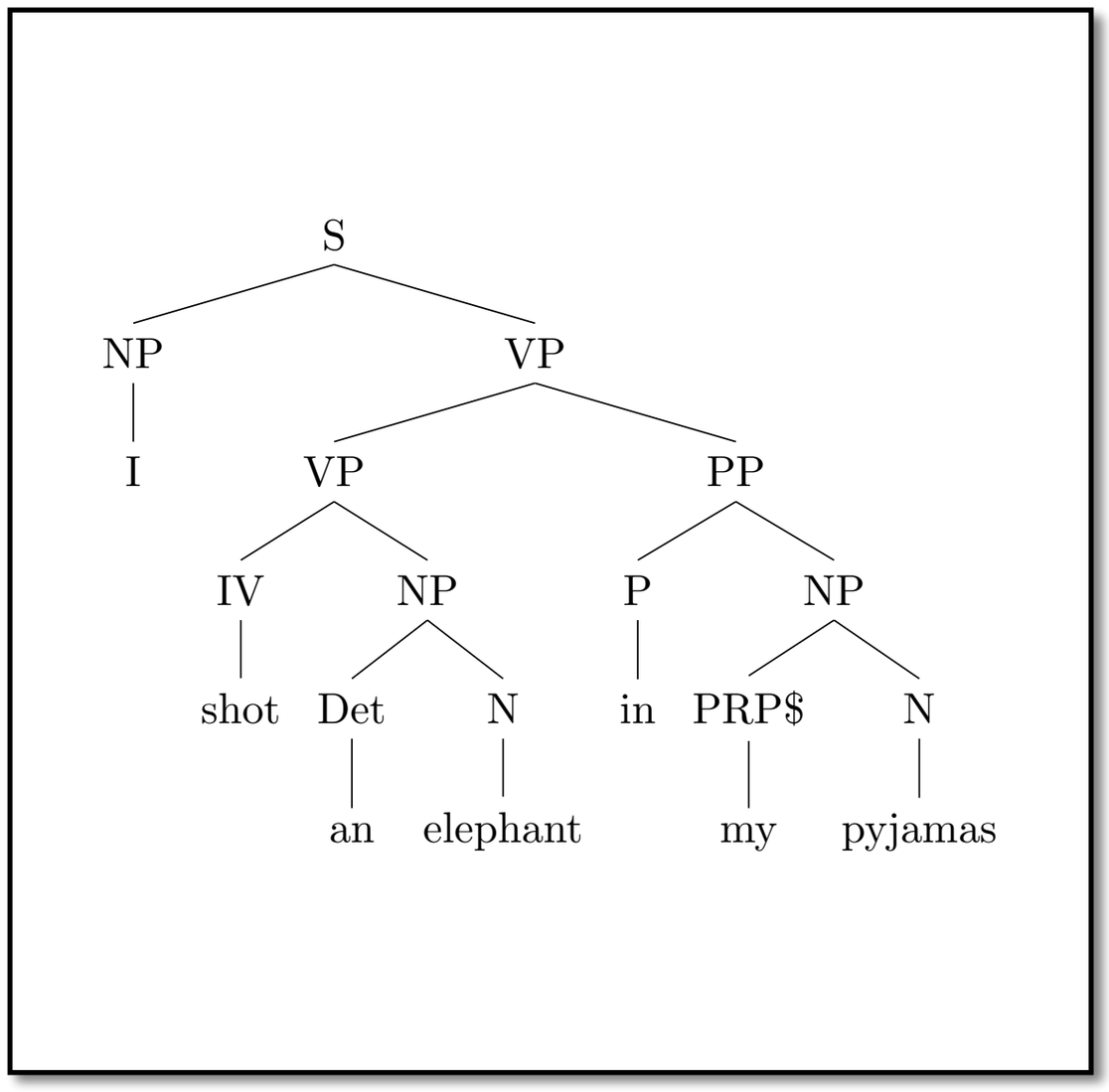


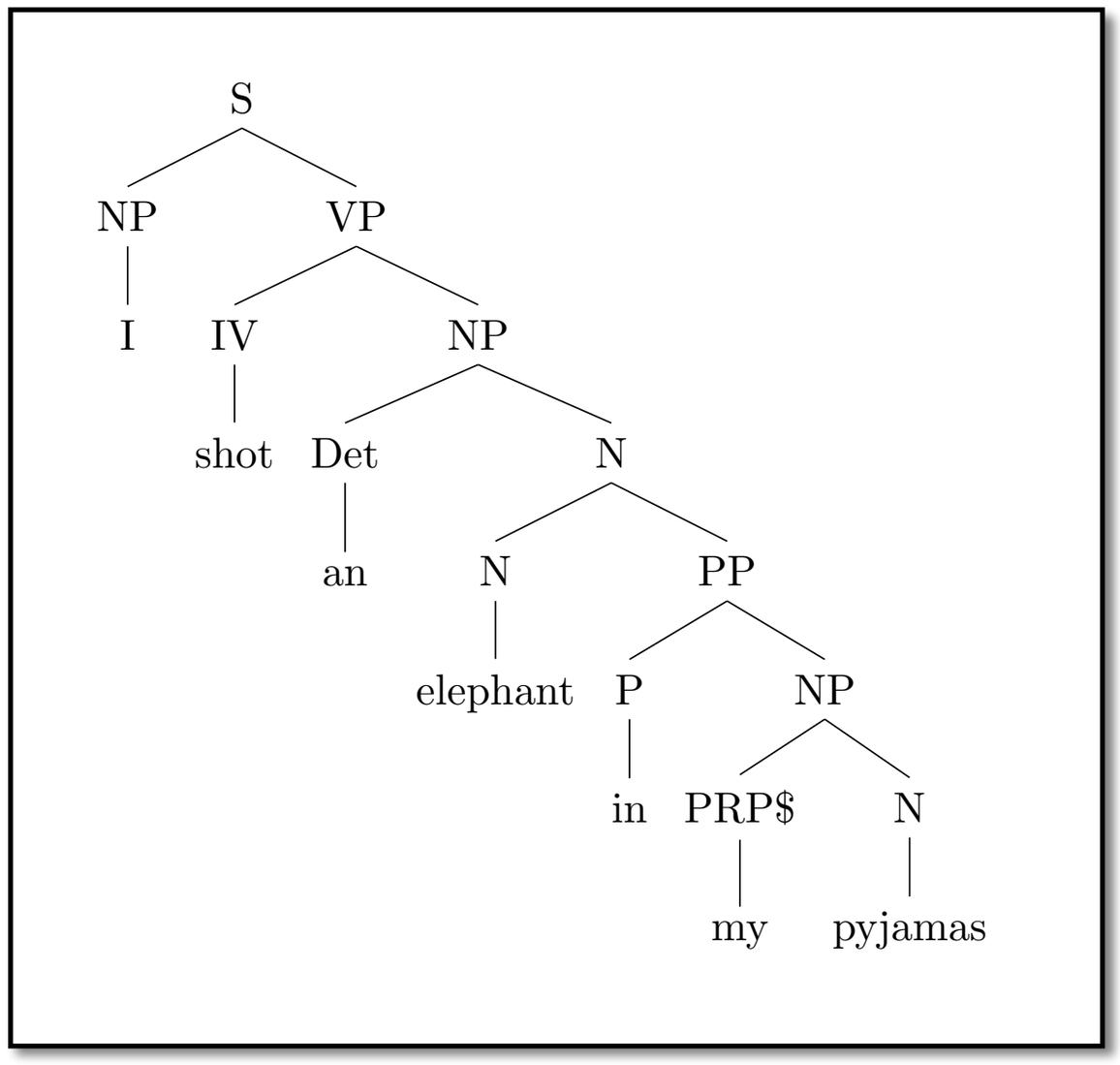
Ambiguity: parts of speech

VBD		VB			
VBN	VBZ	VBP	VBZ		
NNP	NNS	NN	NNS	CD	NN
Fed	raises	interest	rates	0.5	percent

(POS tags from Penn Treebank)

Ambiguity: Syntax





“ … How it got there, I have no idea.”

Other types of ambiguity

- A central problem: NL expressions are frequently highly ambiguous.
  - lexical ambiguities: “interest” (noun) vs. “interest” (verb) — vs. “interest” (the other noun)
  - structural semantic ambiguities:
    “every student did not pass the exam”
  - referential ambiguities:
    “John beat Peter up. That really hurt him.”

- Individual analyses are called *readings*.

The ambiguity challenge

- Number of readings grows exponentially with the sources of ambiguity.
  - How do we identify the correct one?
  - e.g. statistical models
- In practice, infeasible to enumerate all readings and choose the right one.
  - How can we compute the correct reading efficiently?
  - development of good algorithms

The knowledge challenge

- Uncovering hidden structure requires *knowledge* about language. Where do we get it?

- Classical approach: hand-written rules.
  - Can be effective, but is very expensive.

- "Modern" approach: statistical models.
  - dominant paradigm since the late 1990s

- Current approach (since 2015): neural models.
  - won't talk about this much in this class

Topics in this class

- Elementary statistical models of language
- Tagging: Hidden Markov Models
- Parsing: esp. probabilistic context-free grammars
- Further topics: a bit of …
  - semantics
  - machine translation
  - grammar induction

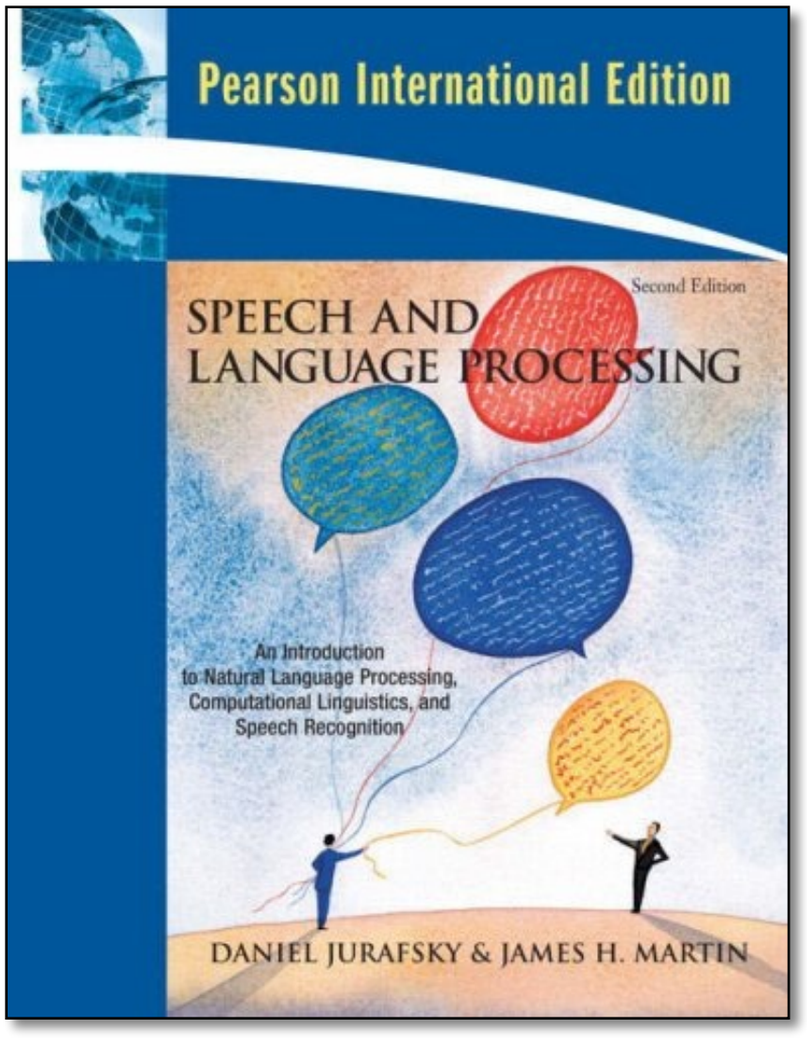
Lectures

- We will assign you some reading for each lecture. Please read it *beforehand*.

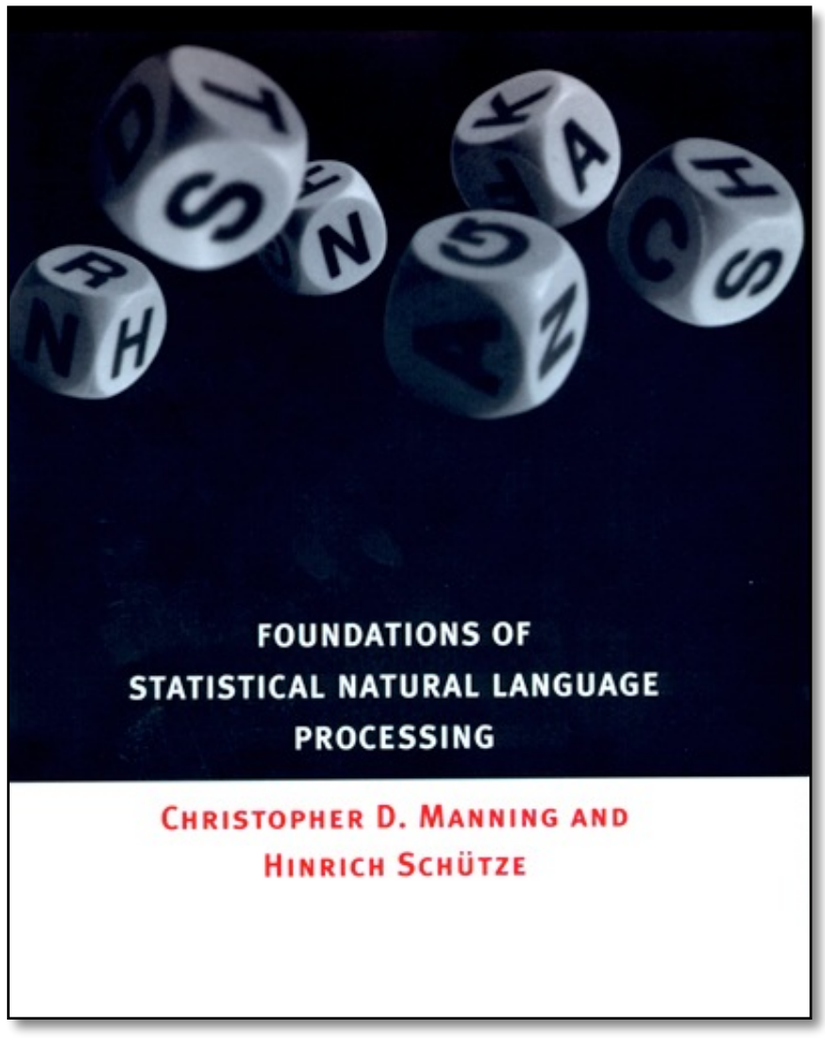
- Lecture will be dense summaries, add some extra information, give you a chance to discuss.

- Please talk to me during lectures if anything is unclear. I want this to be a two-way communication.

Standard Textbooks



Dan Jurafsky and James Martin, Speech and Language Processing



Chris Manning and Hinrich Schütze, Foundations of Statistical Natural Language Processing

Assignments

- There will be six programming assignments.
  - Start early and plan enough time.
- Grading:
  - You need to turn in at least five assignments.
  - We will add up your best two scores from A1-3 and your best two scores from A4-6.
  - In total, you must get at least 250 (of 400) points out of these best four assignments.

Programming skills

- You will need a certain degree of programming skills to complete the assignments.

- We assume that you are familiar with Python 3. Some assignments are easier with NLTK.

- Show of hands — programming skills?

Final project

- The grade for the class is determined by a final project, which you work on in the term break.
  - submit code plus documentation
- You should propose a topic for the project.
  - size of project = roughly one assignment
- Grade will be based on
  - difficulty of task
  - quality of solution
  - clarity of presentation

Grade for course

- Your final grade will consist of:
  - 50% grade for the assignments
  - 50% grade for final project
  - need to get passing grade for each

Resources

- Course website:
  https://coli-saar.github.io/cl18

- Piazza (please sign up!):
  http://piazza.com/uni-saarland.de/winter2018/cl

- Weekly voluntary tutorials with Brielen Lasota