Introduction

Computational Linguistics

Alexander Koller

04 November 2016

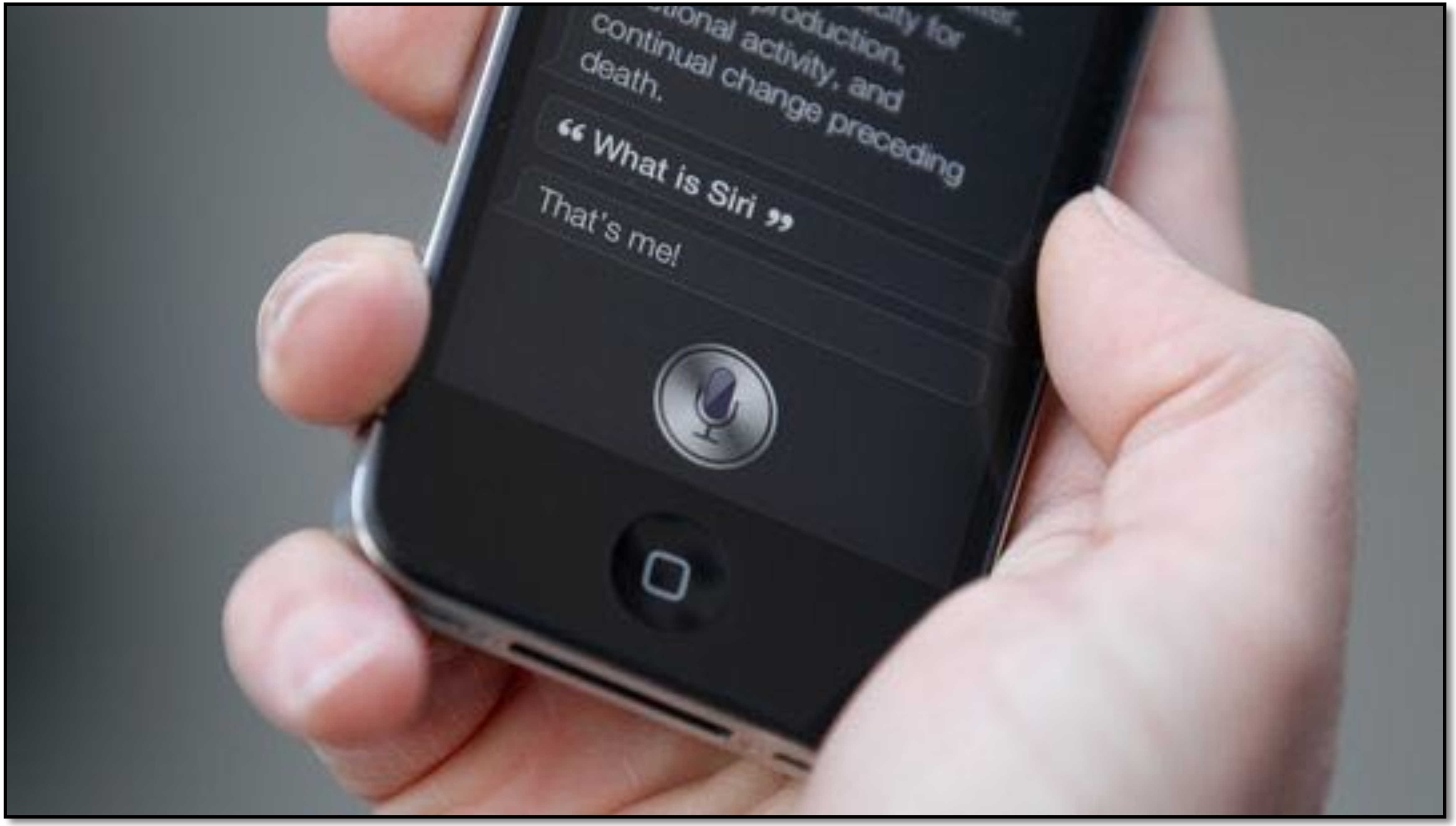
Outline

- What is computational linguistics?

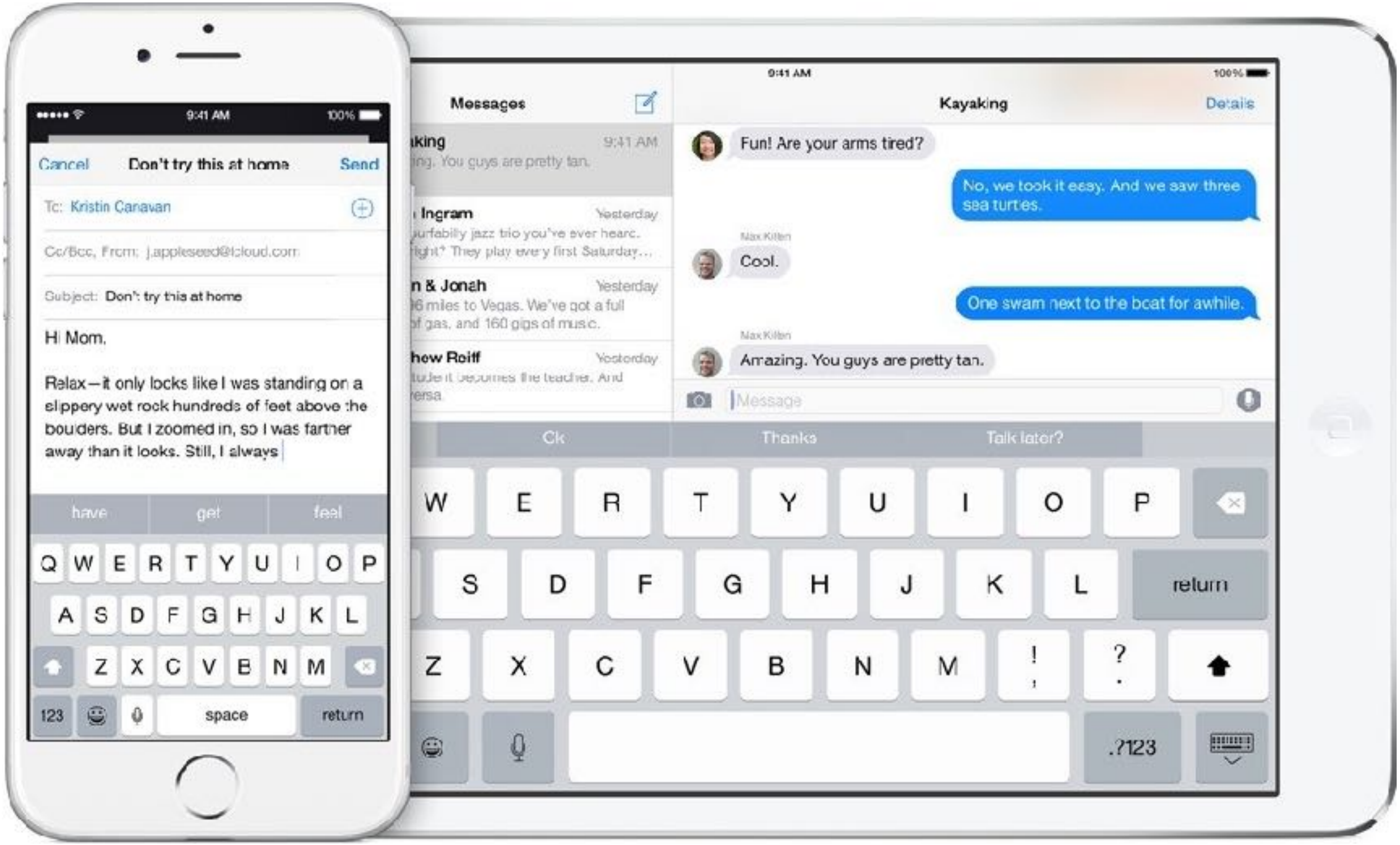
- Topics of this course

- Organizational issues

Siri


production, city for
tional activity, and
continual change preceding
death.
“ What is Siri ”
That's me!

Text prediction



To: Kristin Canavan

Cc/Bcc, From: j.appleseed@icloud.com

Subject: Don't try this at home

Hi Mom,

Relax—it only looks like I was standing on a slippery wet rock hundreds of feet above the boulders. But I zoomed in, so I was farther away than it looks. Still, I always

Kayaking

Fun! Are your arms tired?

No, we took it easy. And we saw three sea turtles.

Max Kilen

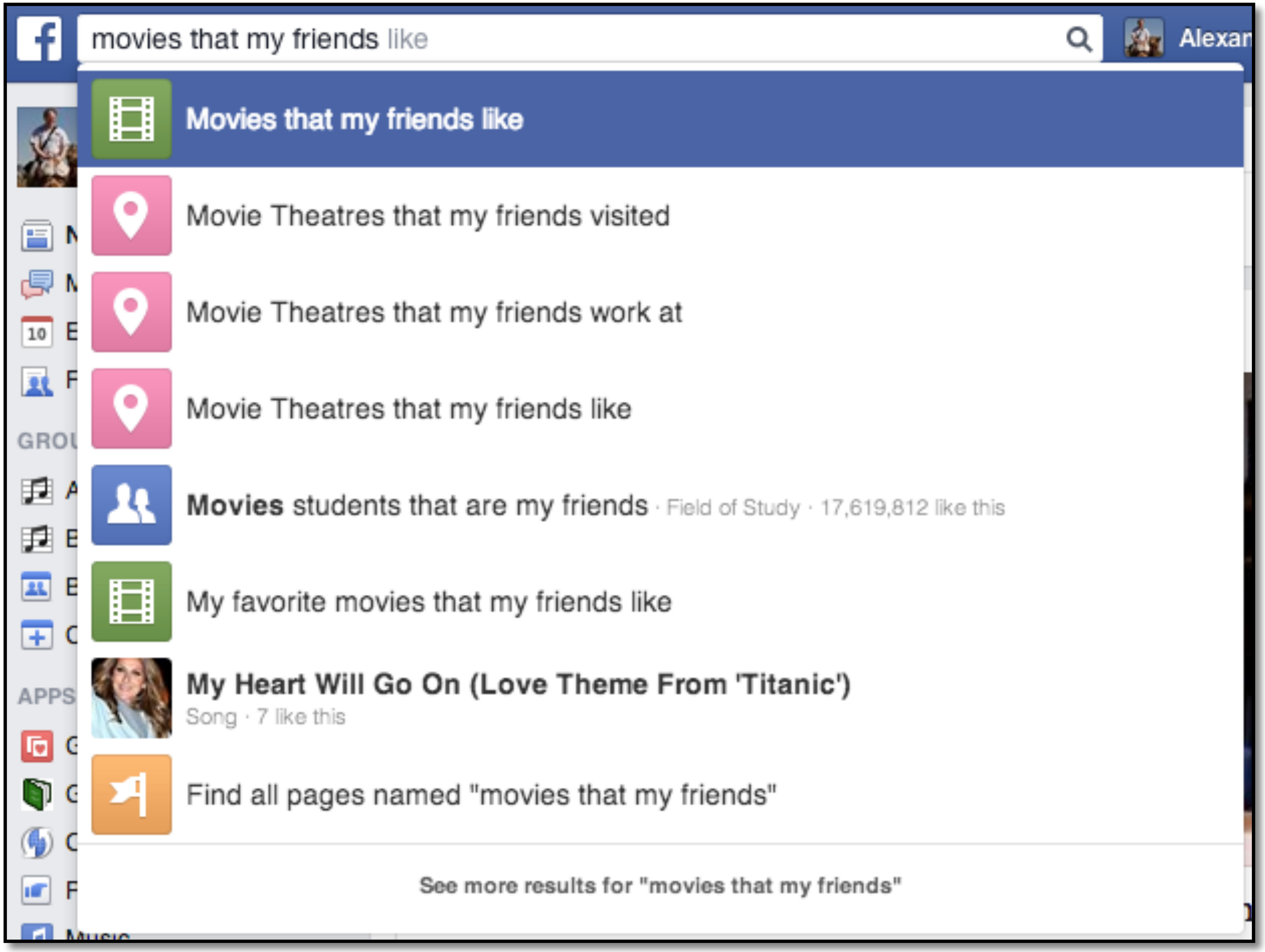
Cool.

One swam next to the boat for awhile.

Max Kilen

Amazing. You guys are pretty tan.

Facebook Graph Search



movies that my friends like

Movies that my friends like

Movie Theatres that my friends visited

Movie Theatres that my friends work at

Movie Theatres that my friends like

Movies students that are my friends · Field of Study · 17,619,812 like this

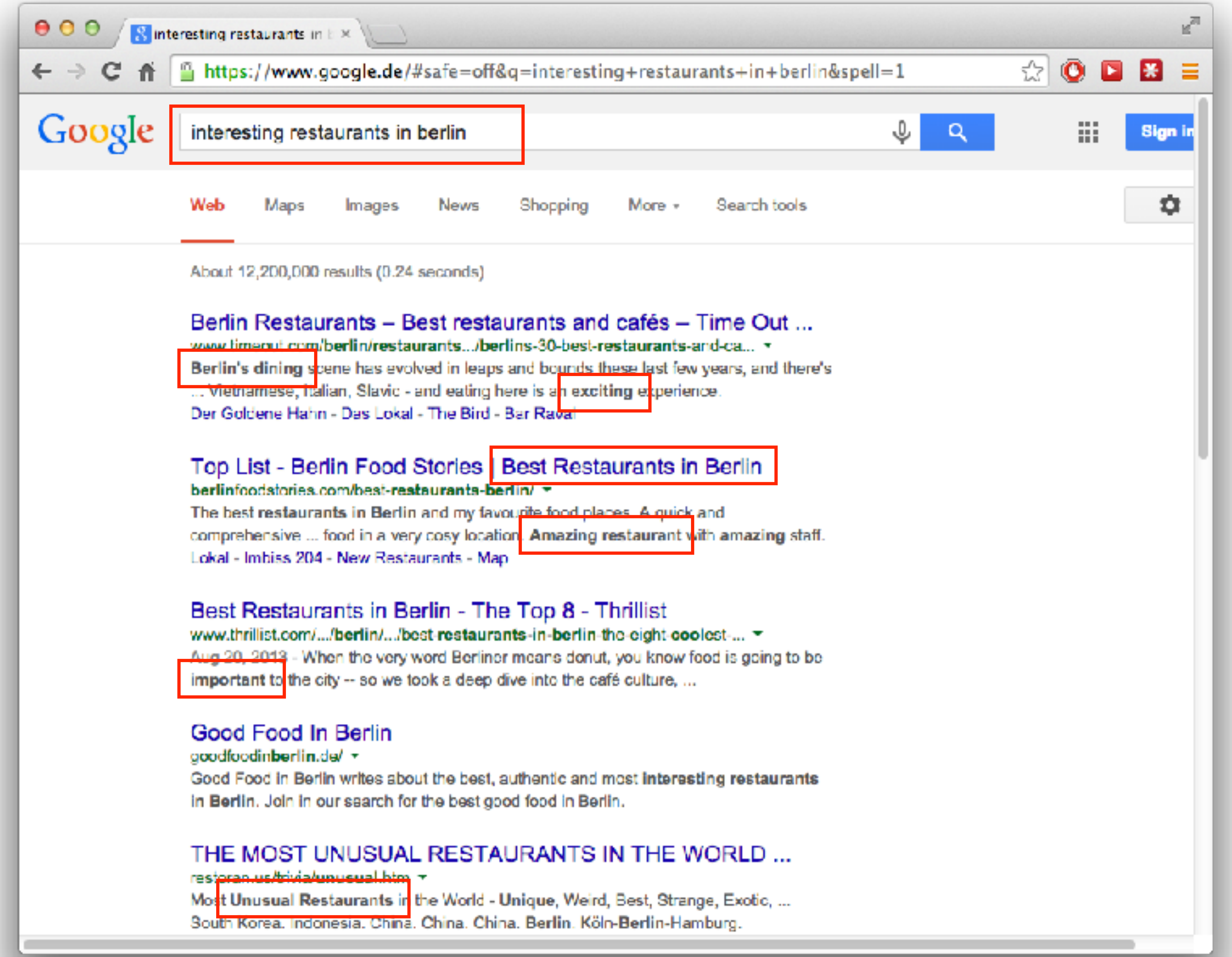
My favorite movies that my friends like

My Heart Will Go On (Love Theme From 'Titanic')
Song · 7 like this

Find all pages named "movies that my friends"

See more results for "movies that my friends"

Similarity in Google Search


interesting restaurants in b ×
https://www.google.de/#safe=off&q=interesting+restaurants+in+berlin&spell=1
Google
Sign in
Web
Maps
Images
News
Shopping
More ▾
Search tools

interesting restaurants in berlin

About 12,200,000 results (0.24 seconds)

Berlin Restaurants – Best restaurants and cafés – Time Out ...
www.timeout.com/berlin/restaurants.../berlins-30-best-restaurants-and-ca... ▾
Berlin's dining scene has evolved in leaps and bounds these last few years, and there's ... Vietnamese, Italian, Slavic - and eating here is an **exciting** experience.
Der Goldene Hahn - Das Lokal - The Bird - Bar Raval

Top List - Berlin Food Stories | Best Restaurants in Berlin
berlinfoodstories.com/best-**restaurants**-**berlin**/ ▾
The best **restaurants** in **Berlin** and my favourite food places. A quick and comprehensive ... food in a very cosy location. **Amazing restaurant** with **amazing** staff.
Lokal - Imbiss 204 - New Restaurants - Map

Best Restaurants in Berlin - The Top 8 - Thrillist
www.thrillist.com/.../**berlin**/.../best-**restaurants**-in-**berlin**-the-eight-**coolest**-... ▾
Aug 20, 2013 - When the very word Berliner means donut, you know food is going to be **important** to the city -- so we took a deep dive into the café culture, ...

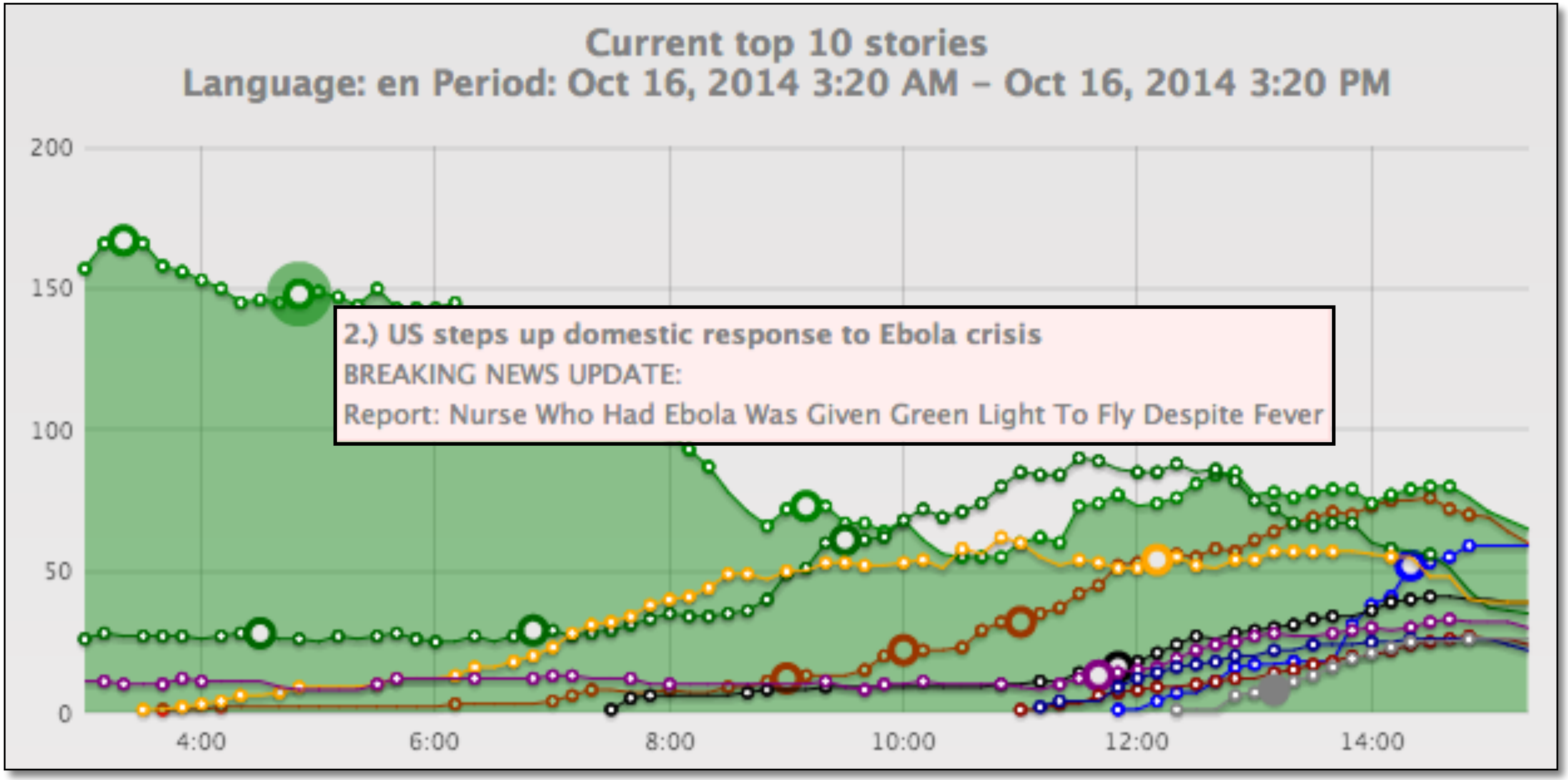
Good Food In Berlin
goodfoodin**berlin**.de/ ▾
Good Food In **Berlin** writes about the best, authentic and most **interesting restaurants** in **Berlin**. Join in our search for the best good food in **Berlin**.

THE MOST UNUSUAL RESTAURANTS IN THE WORLD ...
restoran.us/trivia/**unusual**.htm ▾
Most **Unusual Restaurants** in the World - **Unique**, Weird, Best, Strange, Exotic, ... South Korea. Indonesia. China. China. China. **Berlin**. Köln-**Berlin**-Hamburg.

Navigation Systems

Turn right in fifty meters.

Information access

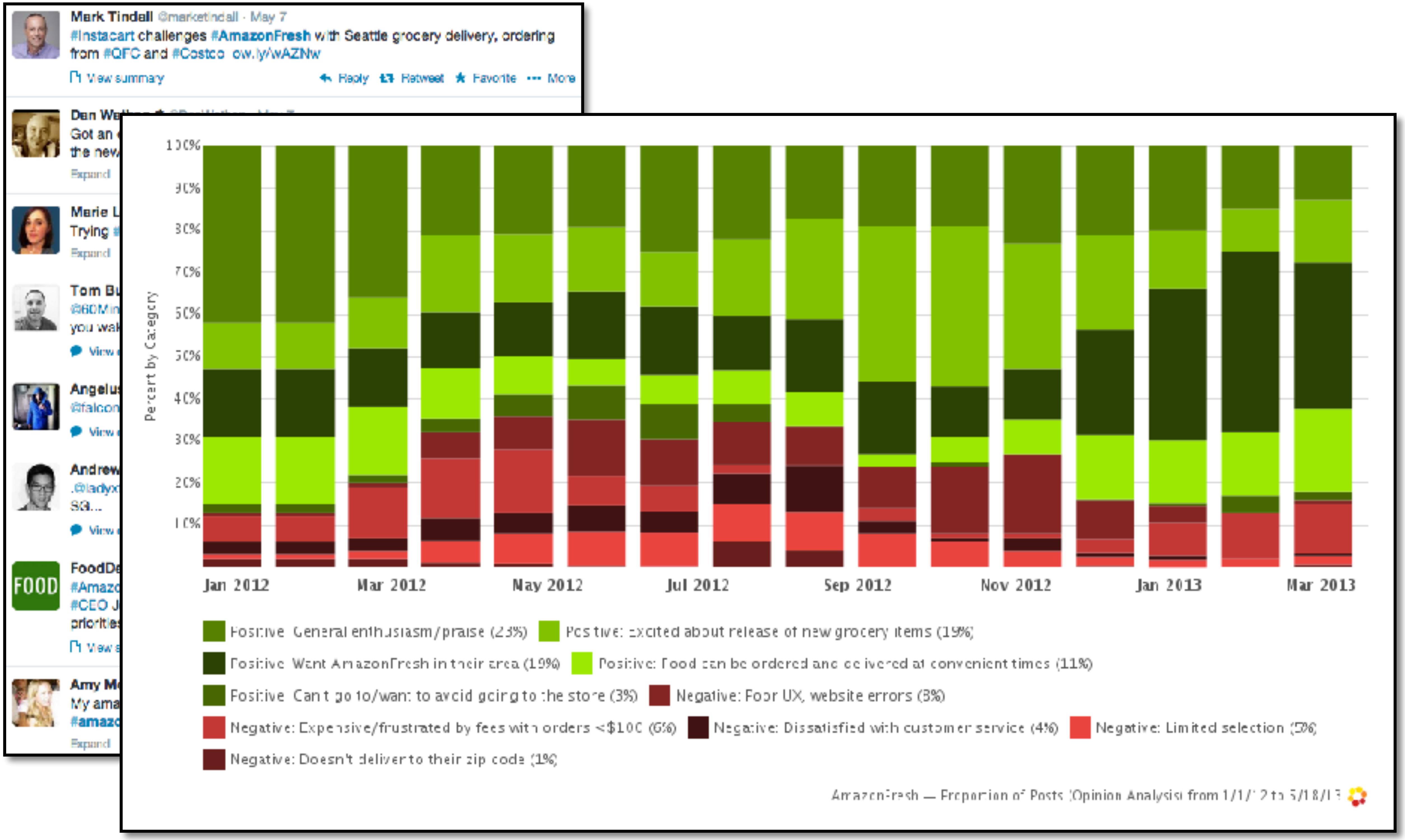


Current top 10 stories
Language: en Period: Oct 16, 2014 3:20 AM – Oct 16, 2014 3:20 PM

2.) US steps up domestic response to Ebola crisis
BREAKING NEWS UPDATE:
Report: Nurse Who Had Ebola Was Given Green Light To Fly Despite Fever

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

Sentiment Analysis



Mark Tindall @marketindall · May 7
#Instacart challenges #AmazonFresh with Seattle grocery delivery, ordering from #QFC and #Costco ow.ly/wAZNw

Dan Wa
Got an
the new

Marie L
Trying #

Tom Bu
@60Min
you wal

Angelus
@falcon

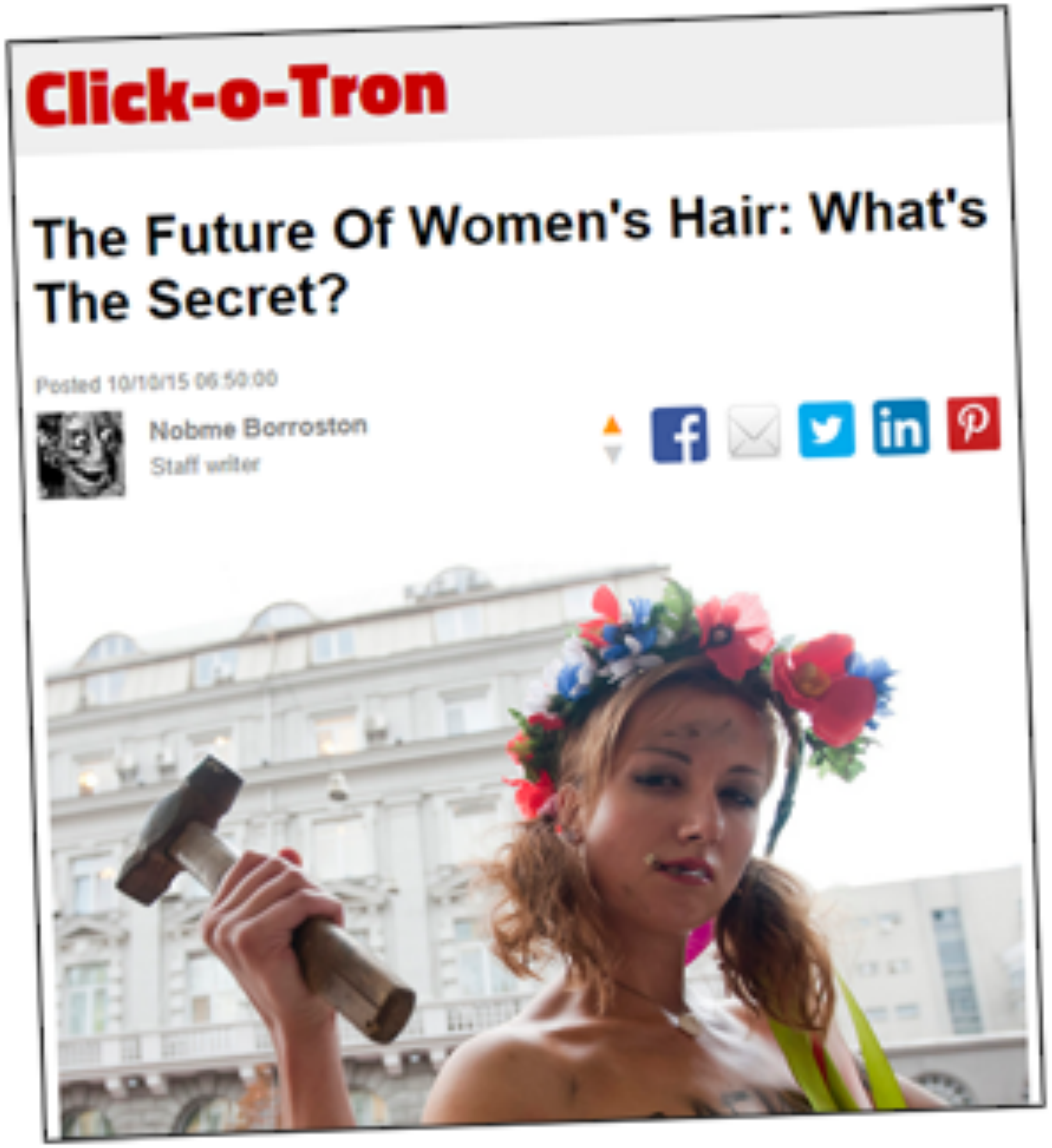
Andrew
.@ladyx
SG...

FoodDe
#Amazo
#CEO J
priorities

Amy Mc
My ama
#amazo

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

Clickbait generation with RNNs



Click-o-Tron

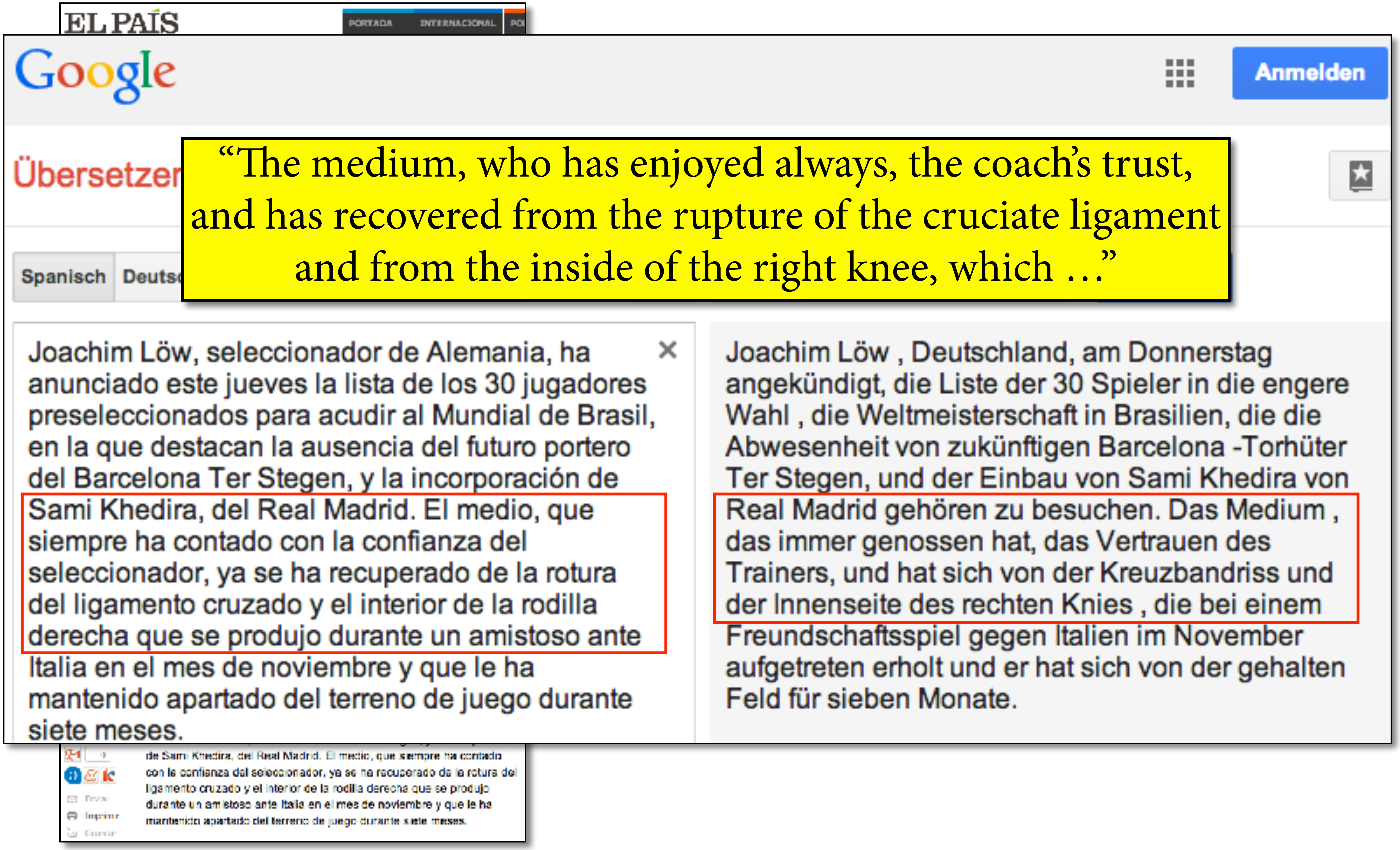
The Future Of Women's Hair: What's The Secret?

Posted 10/10/15 06:50:00

Nobme Borroston
Staff writer

http://clickotron.com/

Google Translate



"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 …"

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.

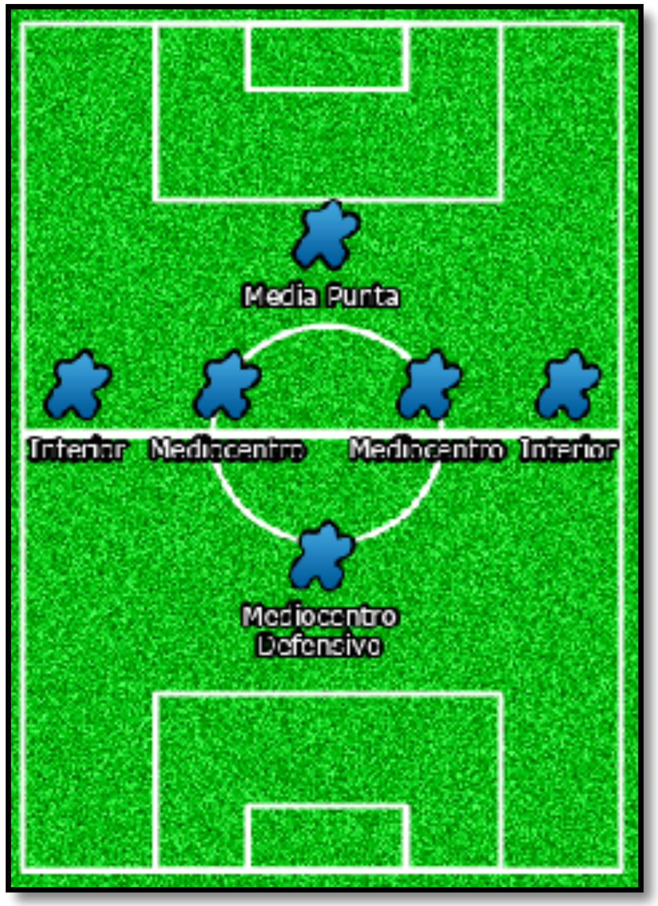
de Sami Khedira, del Real Madrid. El medio, que siempre ha contado con la confianza del seleccionador, ya se ha recuperado de la rotura del ligamento cruzado y el interior de la rodilla derecha que se produjo durante un amistoso ante Italia en el mes de noviembre y que le ha mantenido apartado del terreno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

elpais.com, May 2014

Lexical Ambiguity

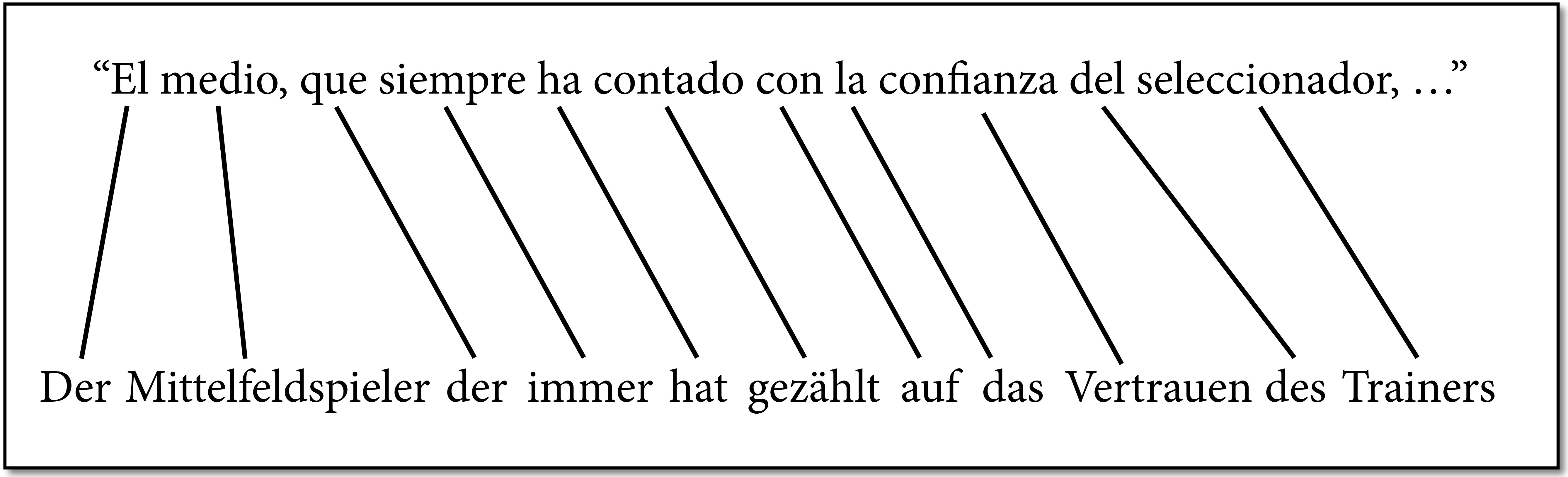
"El medio, que siempre ha contado …"

"Medium"
(medium)



"Mittelfeldspieler"
(midfield player)

Word order



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

Der Mittelfeldspieler der immer hat gezählt auf das Vertrauen des Trainers

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

Word order

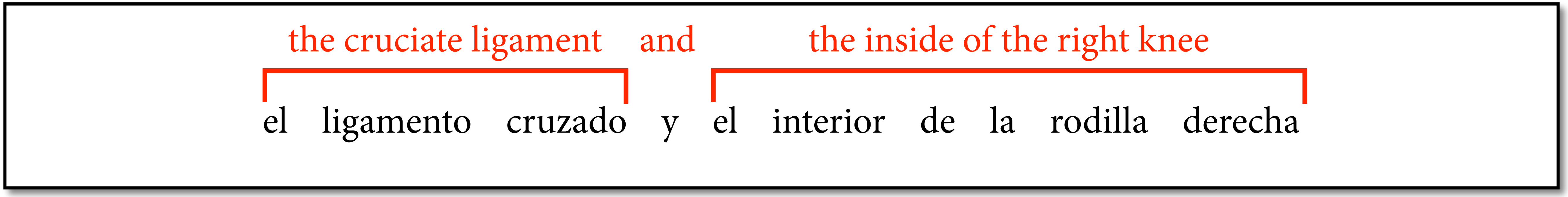
"El medio, que siempre ha contado con la confianza del seleccionador, …"

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

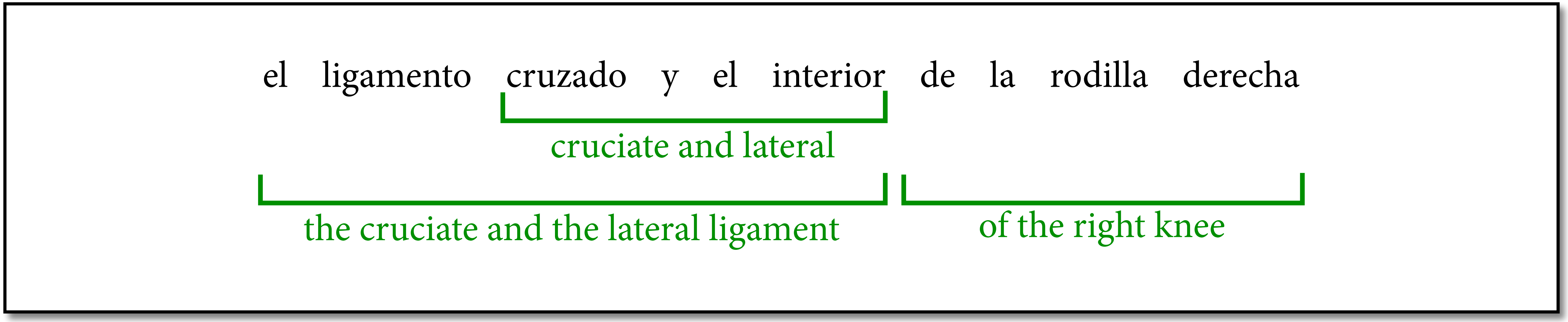
Structural Ambiguity

“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



the cruciate ligament and the inside of the right knee

el ligamento cruzado y el interior de la rodilla derecha



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

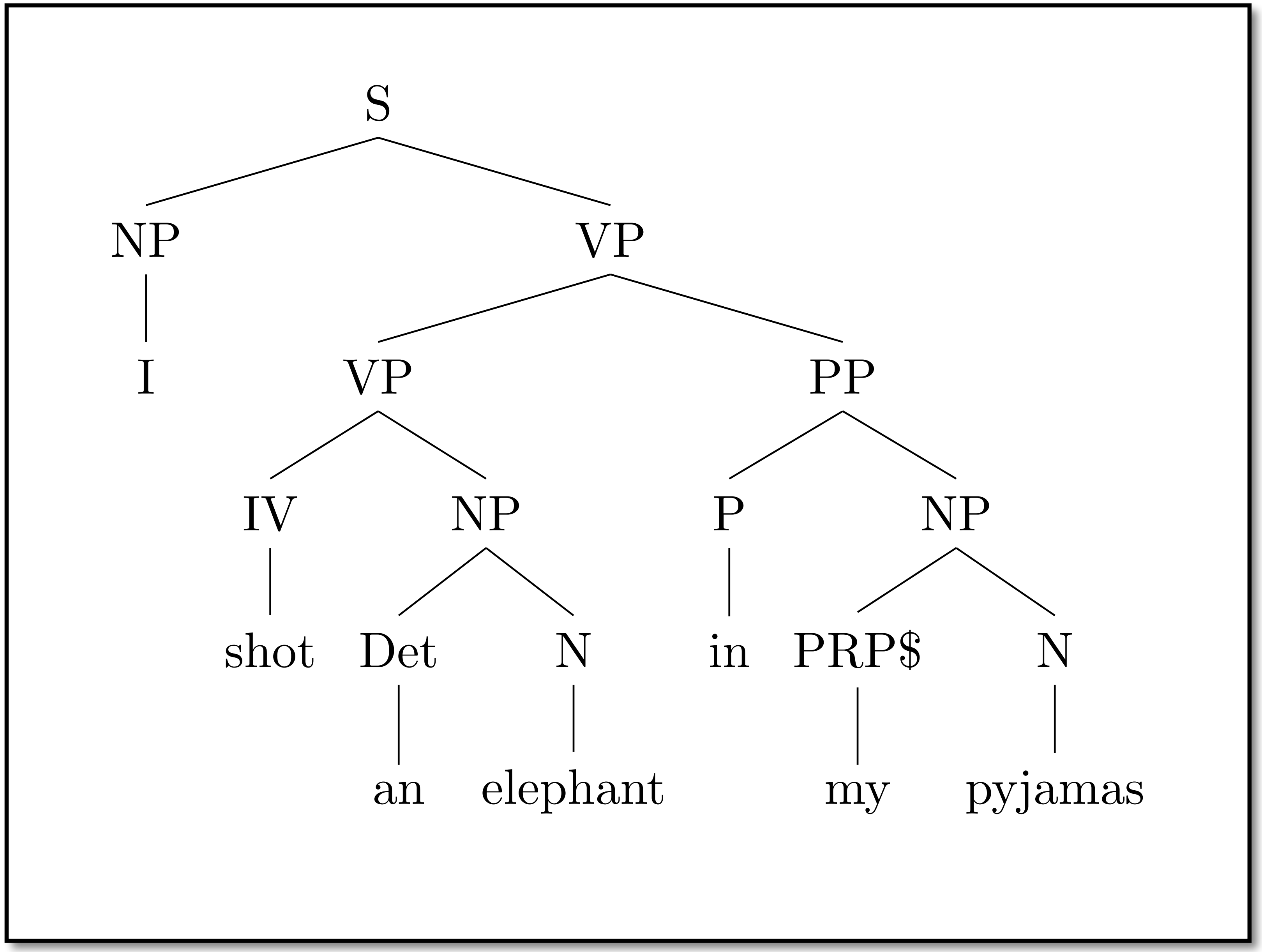
- Fundamental techniques of CL.
- Common themes:
  - uncovering hidden linguistic structure
  - dealing with ambiguity
  - statistical methods
  - efficient algorithms

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

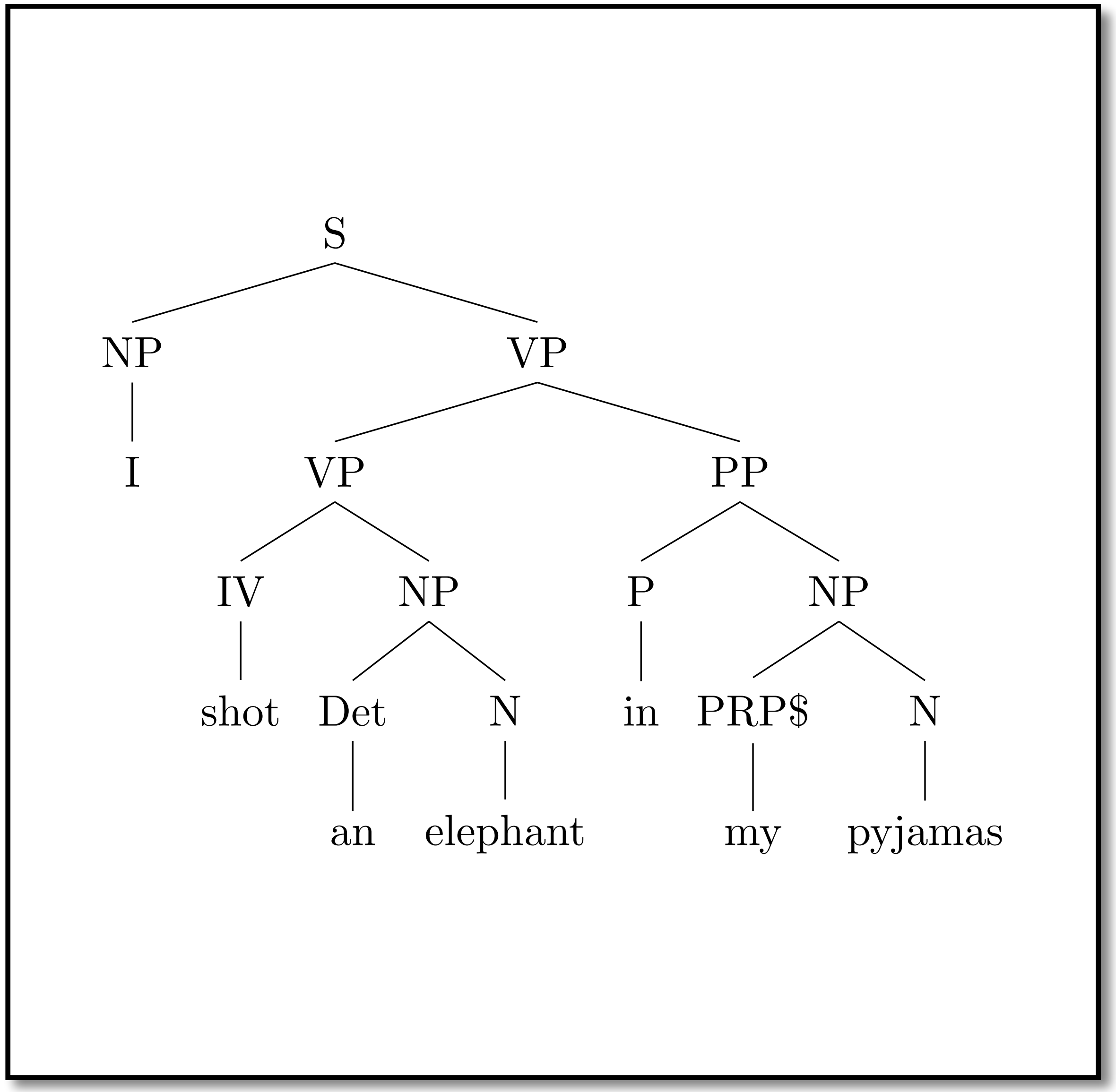


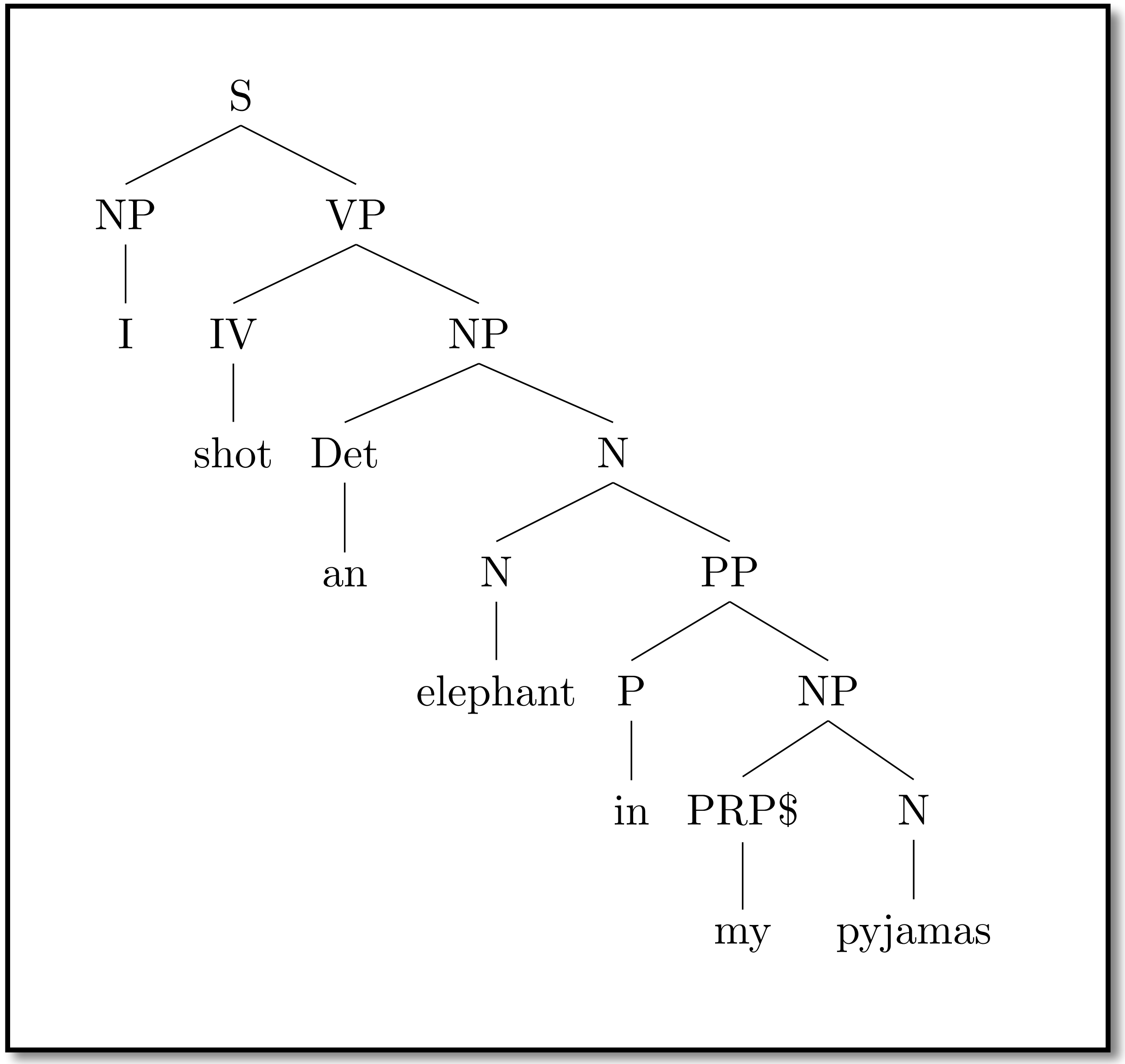
Ambiguity: parts of speech

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
- Tradeoff between depth of modeling and quality of available models.

Siri of the Future?

User: Book a table at “Da Giovanni” after I finish work, and tell John and Mary to meet me there.

System: Sorry, Giovanni has no free tables until 9pm. Should I find a different Italian restaurant for 6:30pm?

User: Can you find a table in a restaurant with a good wine list?

System: “Ristorante Biscotti” still has an opening.
It’s in the Financial District, but has a similar travel time.

User: Okay, sounds good.

System: (makes the reservation online)

(Example by Ron Kaplan, Nuance)

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

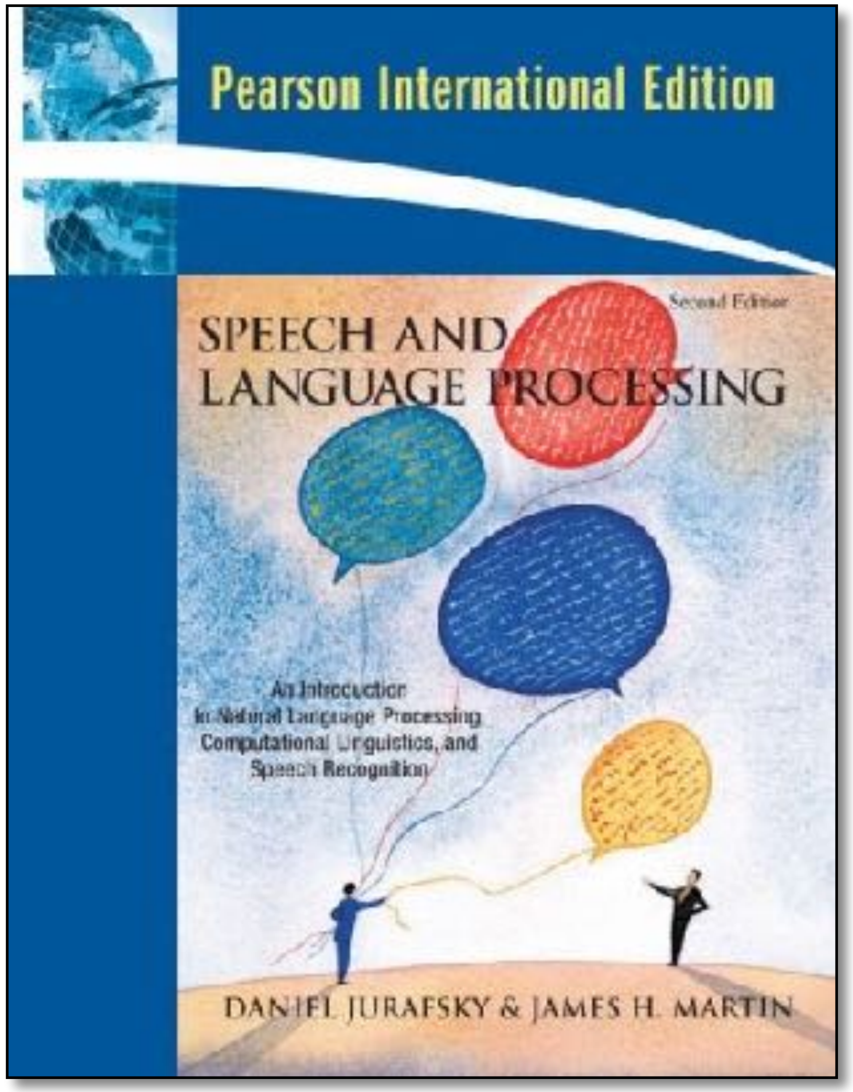
Lectures

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

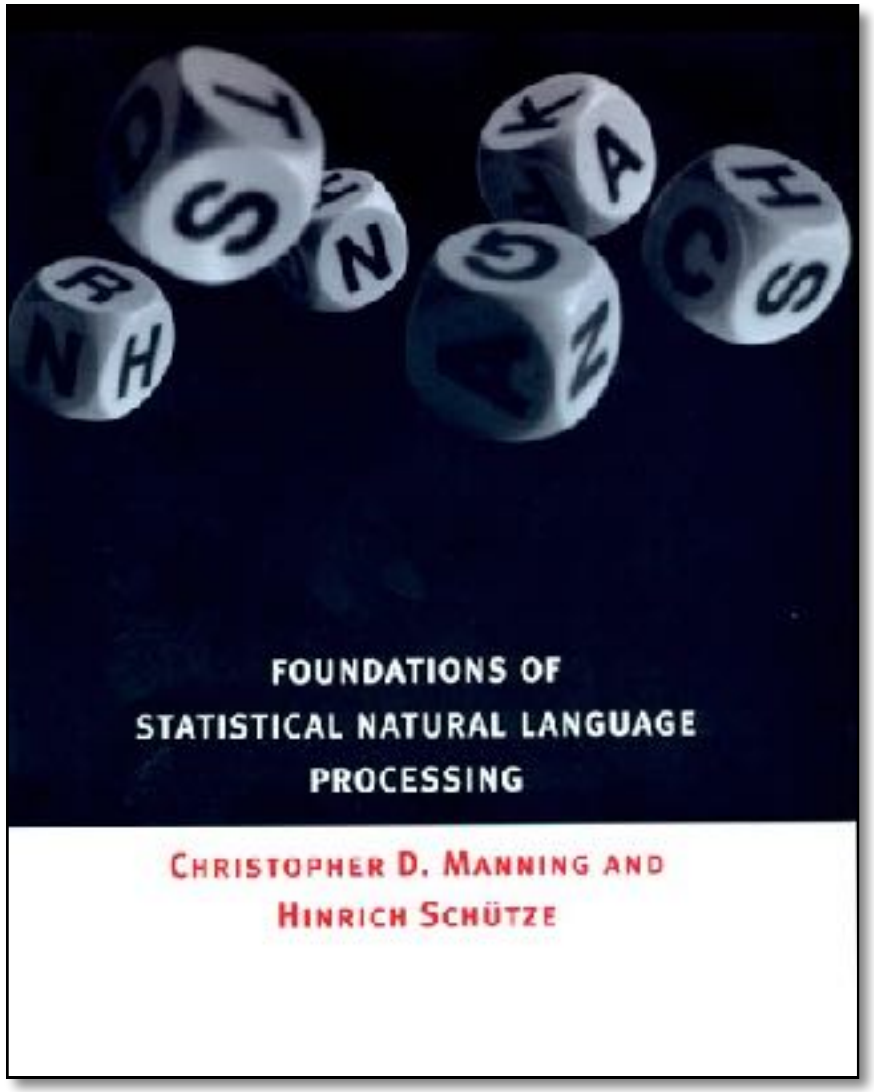
- Lecture will summarize reading, add some extra information, give you a chance to discuss.

- Lectures will be dense but relatively short.
  - need prior reading to understand everything
  - ample time for questions and discussion

Standard Textbooks


Pearson International Edition
Second Edition
SPEECH AND LANGUAGE PROCESSING
An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition
DANIEL JURAFSKY & JAMES H. MARTIN

Dan Jurafsky and James Martin,
Speech and Language Processing


FOUNDATIONS OF STATISTICAL NATURAL LANGUAGE PROCESSING
CHRISTOPHER D. MANNING AND HINRICH SCHÜTZE

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 also 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

Resources

- Course website: https://coli-saar.github.io/cl16/

- Piazza (please sign up!): https://piazza.com/class/iufc96ss1yo6sm

- Weekly voluntary tutorials with Antoine

Logistics

- Friday 12-14: here
- Tuesday: is 10-12 in this room, after all
  - “Speech Science” moves into 12-14 timeslot