

Introduction

Computational Linguistics

Alexander Koller

22 October 2019

Outline

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

Siri



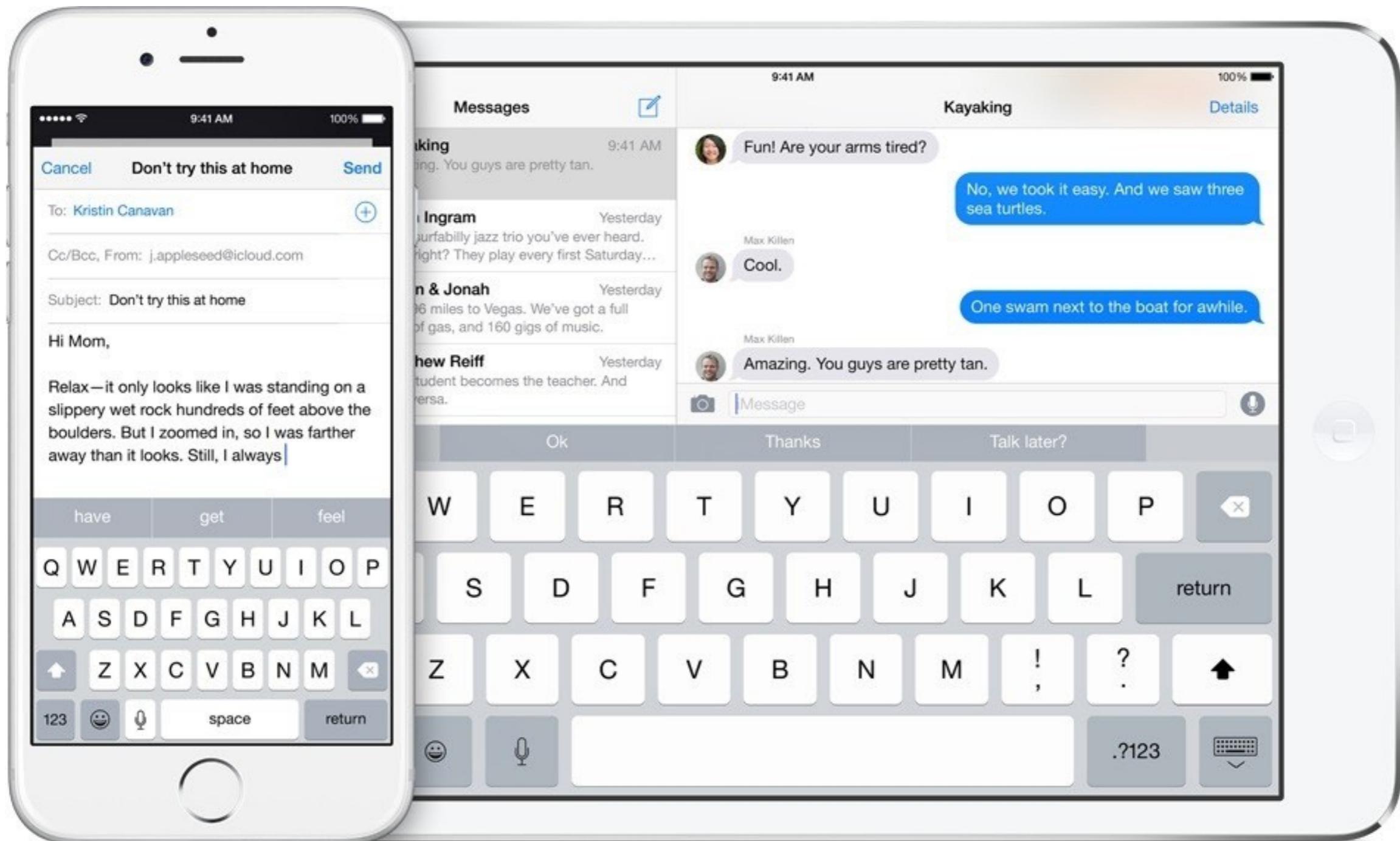
Digital assistants



Google Home

Amazon Echo

Text prediction



Similarity in Google Search

The screenshot shows a Google search results page for the query "interesting restaurants in berlin". The search bar at the top contains the query "interesting restaurants in berlin". Below the search bar, there are navigation links for Web, Maps, Images, News, Shopping, More, and Search tools. A message indicates "About 12,200,000 results (0.24 seconds)".

The first result is a link to "Berlin Restaurants – Best restaurants and cafés – Time Out ...". The snippet below the link reads: "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".

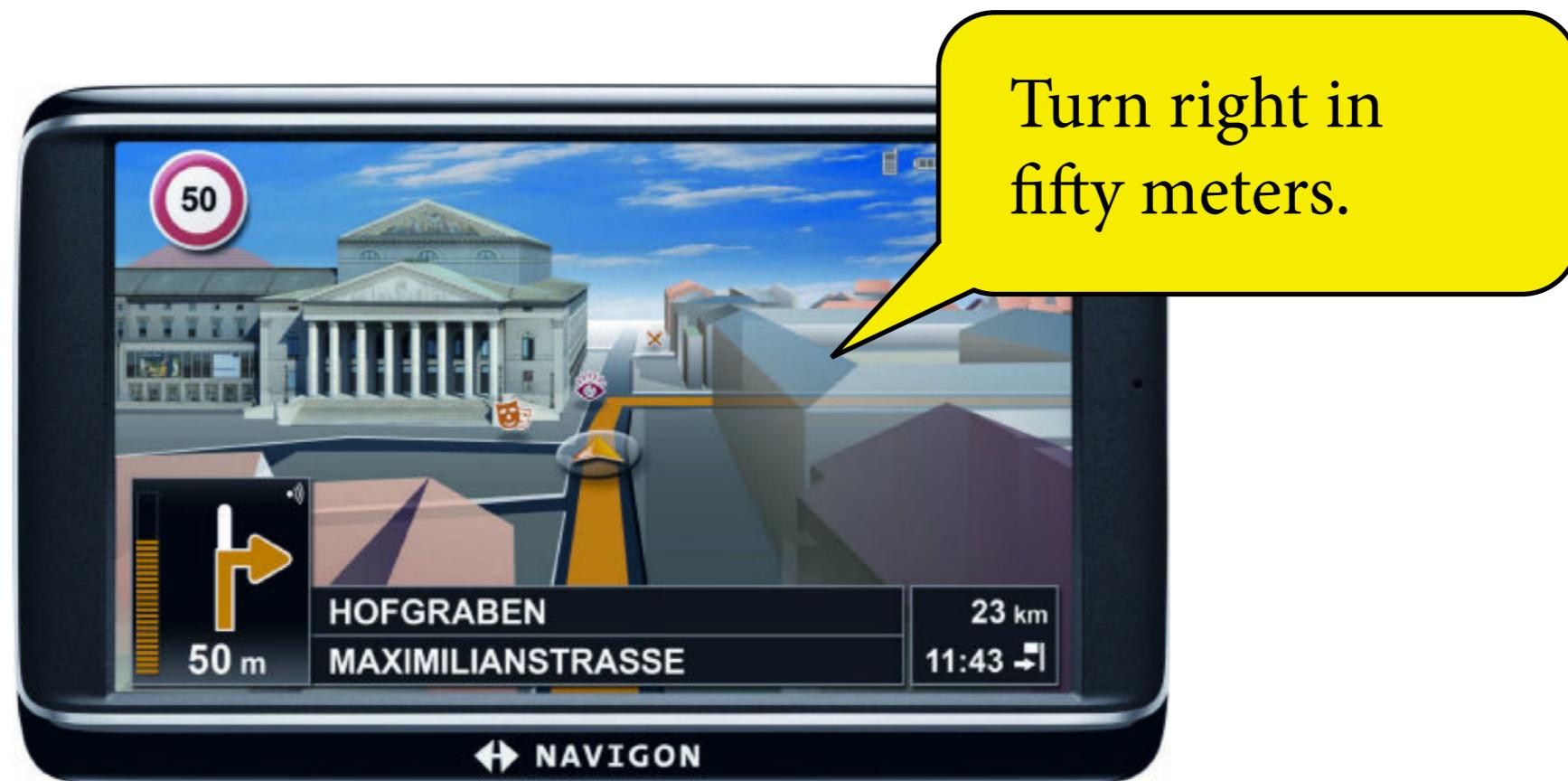
The second result is a link to "Top List - Berlin Food Stories | Best Restaurants in Berlin". The snippet reads: "The best restaurants in Berlin and my favourite food places. A quick and comprehensive ... food in a very cosy location. Amazing restaurant with amazing staff. Lokal - Imbiss 204 - New Restaurants - Map".

The third result is a link to "Best Restaurants in Berlin - The Top 8 - Thrillist". The snippet reads: "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, ...".

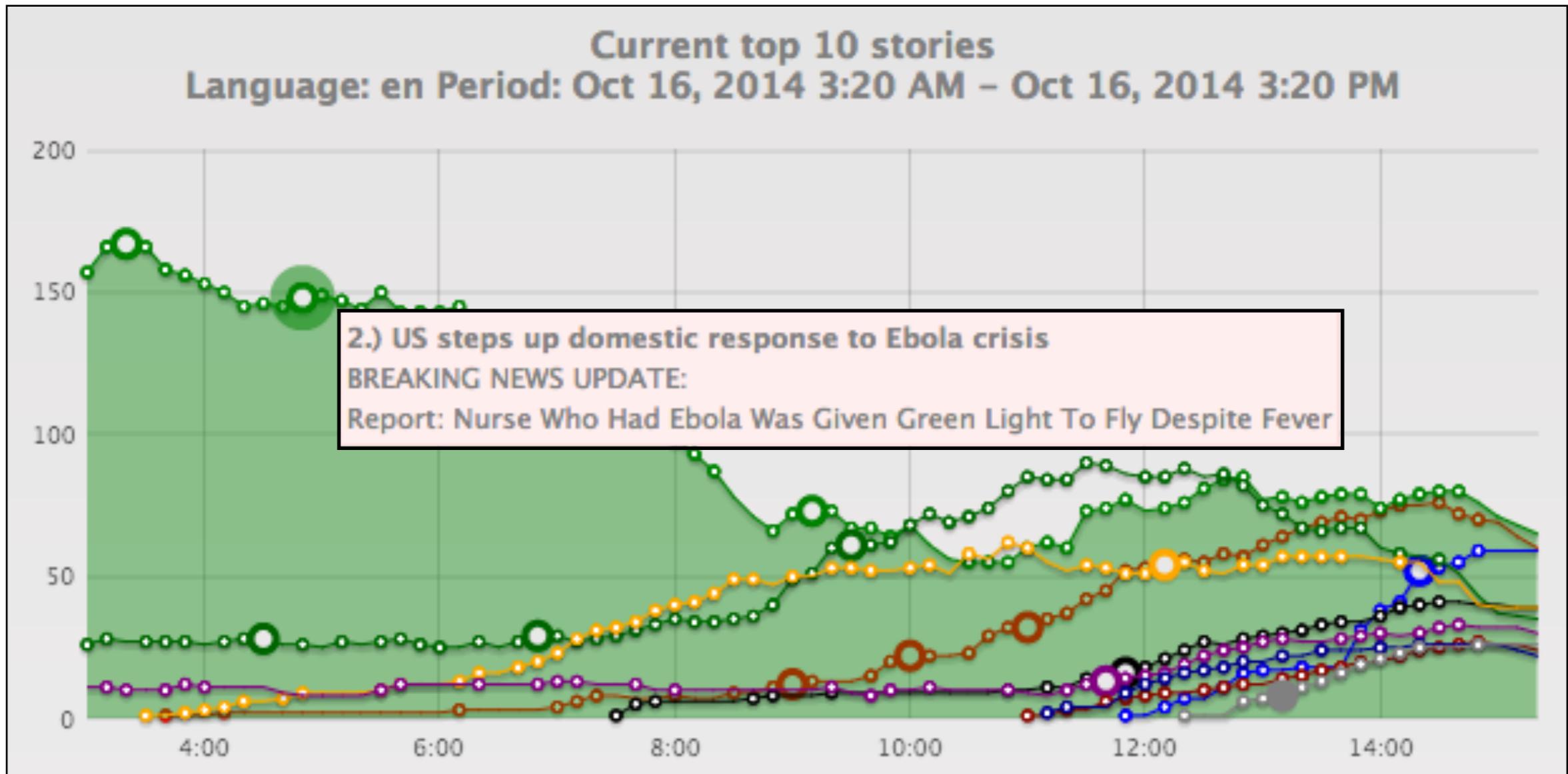
The fourth result is a link to "Good Food In Berlin". The snippet reads: "Good Food in Berlin writes about the best, authentic and most interesting restaurants in Berlin. Join in our search for the best good food in Berlin".

The fifth result is a link to "THE MOST UNUSUAL RESTAURANTS IN THE WORLD ...". The snippet reads: "Most Unusual Restaurants in the World - Unique, Weird, Best, Strange, Exotic, ... South Korea. Indonesia. China. China. China. Berlin. Köln-Berlin-Hamburg".

Navigation Systems



Information access



Sentiment Analysis

 **Mark Tindall** @marketindall · May 7
#Instacart challenges #AmazonFresh with Seattle grocery delivery, ordering from #QFC and #Costco ow.ly/wAZNw
[View summary](#) [Reply](#) [Retweet](#) [Favorite](#) [More](#)

 **Dan Wathen**  @DanWathen · May 7
Got an email that #Amazon is now offering same day delivery if you subscribe to the new #AmazonFresh, which even includes produce.
[Expand](#) [Reply](#) [Retweet](#) [Favorite](#) [More](#)

 **Marie Leduc** @Crystal1a · May 6
Trying #AmazonFresh for the first time. Can't wait to receive my order! :D
[Expand](#) [Reply](#) [Retweet](#) [Favorite](#) [More](#)

 **Tom Butts-Carter** @twbutts · May 6
@60Minutes #AmazonFresh is like the milk man. Groceries on your step when you wake up. @twbutts
[View conversation](#) [Reply](#) [Retweet](#) [Favorite](#) [More](#)

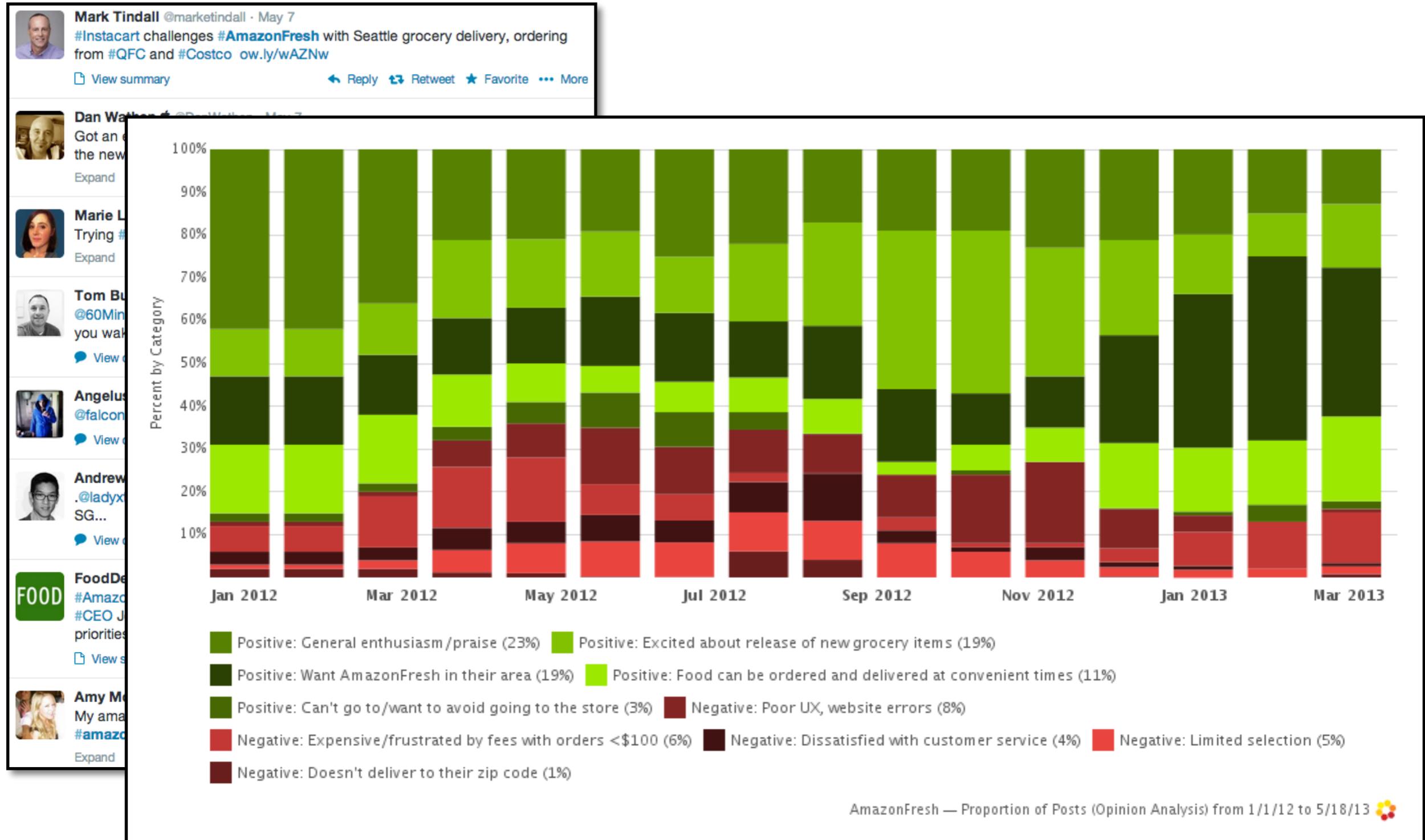
 **Angelus** @Angelus · May 6
@falcon82 Ma #AmazonFresh invece quando cavolo arriva da noi!?
[View conversation](#) [Reply](#) [Retweet](#) [Favorite](#) [More](#)

 **Andrew Au** @AndrewAu · May 5
. @ladyxtel now if only #AmazonPrime and #AmazonFresh were available in SG...
[View conversation](#) [Reply](#) [Retweet](#) [Favorite](#) [More](#)

 **FoodDelivering** @FoodDelivering · May 5
#Amazon #FoodDelivery
#CEO Jeff Bezos: expanding #AmazonFresh is one of the company's top priorities on.recode.net/1ueexXY @FoodDelivering
[View summary](#) [Reply](#) [Retweet](#) [Favorite](#) [More](#)

 **Amy McWethy** @AmyMcWethy · May 5
My amazon dash came today. Super pumped to try it out! #amazon
#amazonfresh #dash #amazondash #scan... instagram.com/p/noNESsk1Qf/
[Expand](#) [Reply](#) [Retweet](#) [Favorite](#) [More](#)

Sentiment Analysis



Clickbait generation with RNNs



<http://clickotron.com/>

Google Translate

EL PAÍS

PORADA INTERNACIONAL PO

DEPORTES

FÚTBOL BALONCESTO TENIS CICLISMO FÓRMULA 1 MOTOS GOLF OTROS

ESTÁ PASANDO Liga: 37ª jornada Giro Italia F-1: GP de España Tenis: Masters de

MUNDIAL 2014 | ALEMANIA »

Löw deja fuera del Mundial a Ter Stegen y cuenta con Khedira

■ El técnico de Alemania anuncia la lista de 30 preseleccionados para Brasil, en la que no cuenta con el futuro portero del Barcelona y sí con el medio del Madrid

EL PAÍS | Madrid | 8 MAY 2014 - 14:04 CET 4

Archivado en: Selección Fútbol Alemania Joachim Löw Selecciones deportivas Fútbol Deportes



Khedira se duele tras su lesión / GIUSEPPE CACACE (AFP)

[f 13](#) [133](#) [0](#) [0](#) [Enviar](#) [Imprimir](#) [Guardar](#)

Joachim Löw, seleccionador de Alemania, ha anunciado este jueves la lista de los 30 jugadores preseleccionados para acudir al Mundial de Brasil, que comenzará el próximo 12 de junio, en la que destaca 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.

elpais.com, May 2014

Google Translate

EL PAÍS PORADA INTERNACIONAL POI

Google Anmelden

Übersetzer

Spanisch Deutsch Englisch Sprache erkennen ▾ ↔ Deutsch Englisch Französisch ▾ Übersetzen

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 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.

Enviar Imprimir Guardar

Google Translate

EL PAÍS PORADA INTERNACIONAL POI

Google Anmelden

Übersetzer

Spanisch Deutsch Englisch Sprache erkennen ▾ ↔ Deutsch Englisch Französisch ▾ Übersetzen

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 q1 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.

Enviar Imprimir Guardar

Google Translate

EL PAÍS PORADA INTERNACIONAL PODCAST

Google Anmelden

Übersetzen ★

Spanisch Deutsch

“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 destaca 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.

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, ...”

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

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

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

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

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 hat

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 hat gezählt auf

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 hat gezählt auf das

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 hat gezählt auf das Vertrauen

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 hat gezählt auf das Vertrauen des

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 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

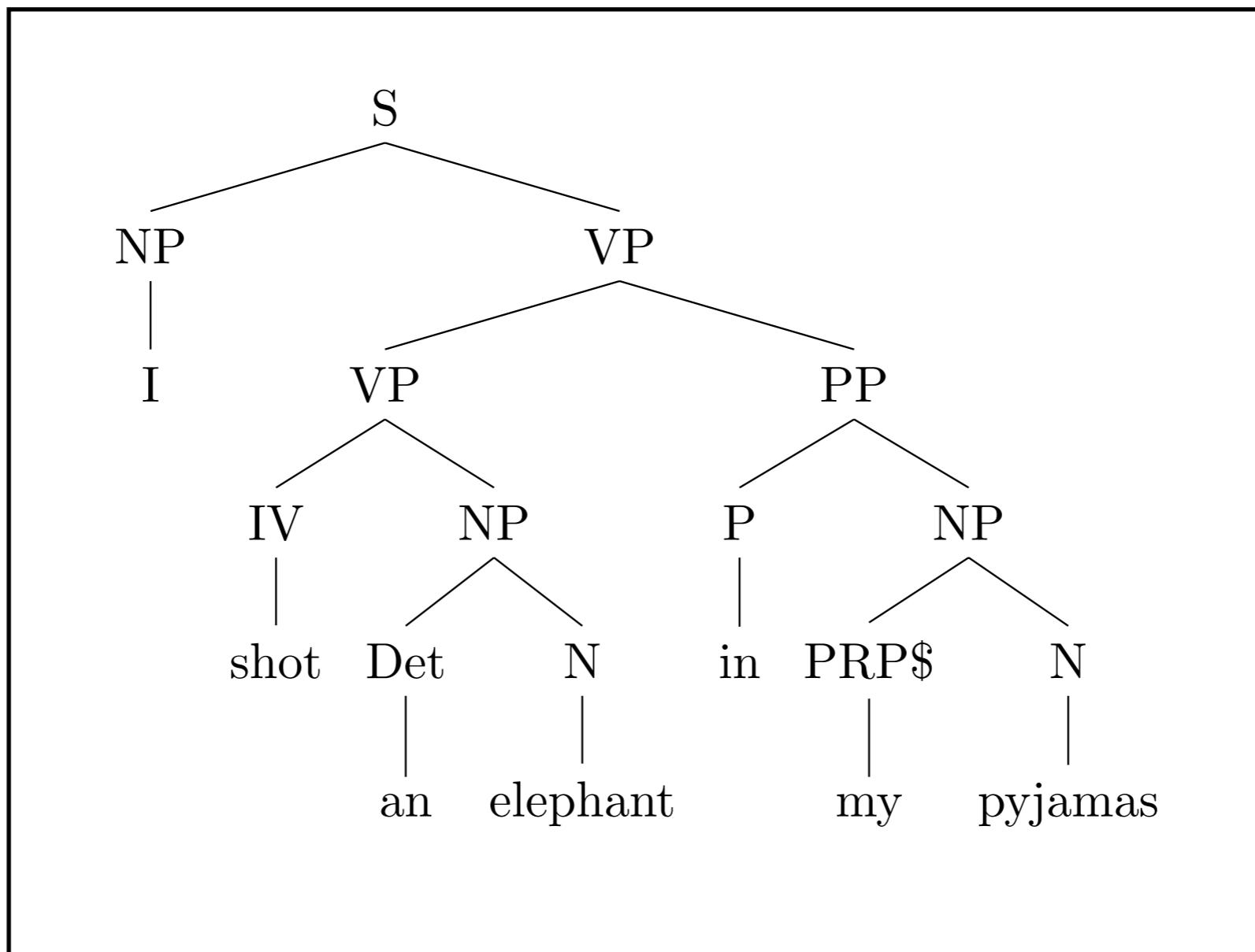
- Fundamental symbolic methods in computational linguistics.
 - ▶ Neural models are really important today, but in this course they will only play a supporting role.
- Common themes:
 - ▶ uncovering hidden linguistic structure
 - ▶ dealing with ambiguity
 - ▶ statistical methods
 - ▶ efficient algorithms

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

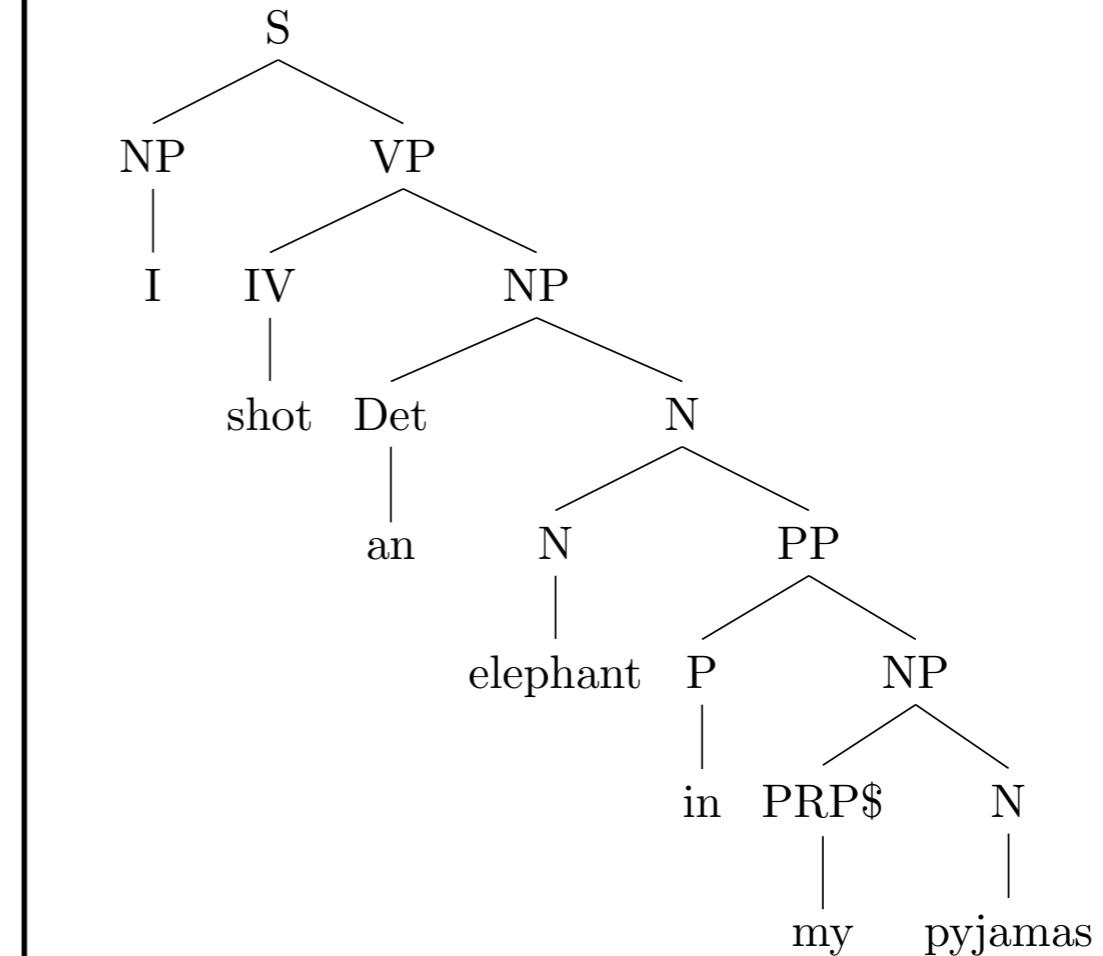
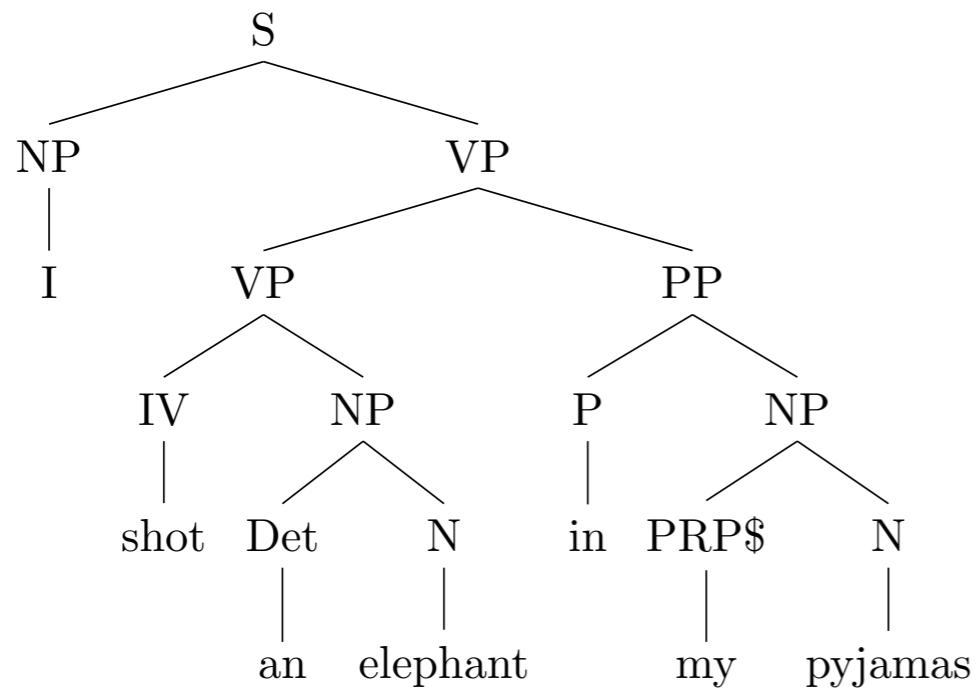


Ambiguity: parts of speech

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

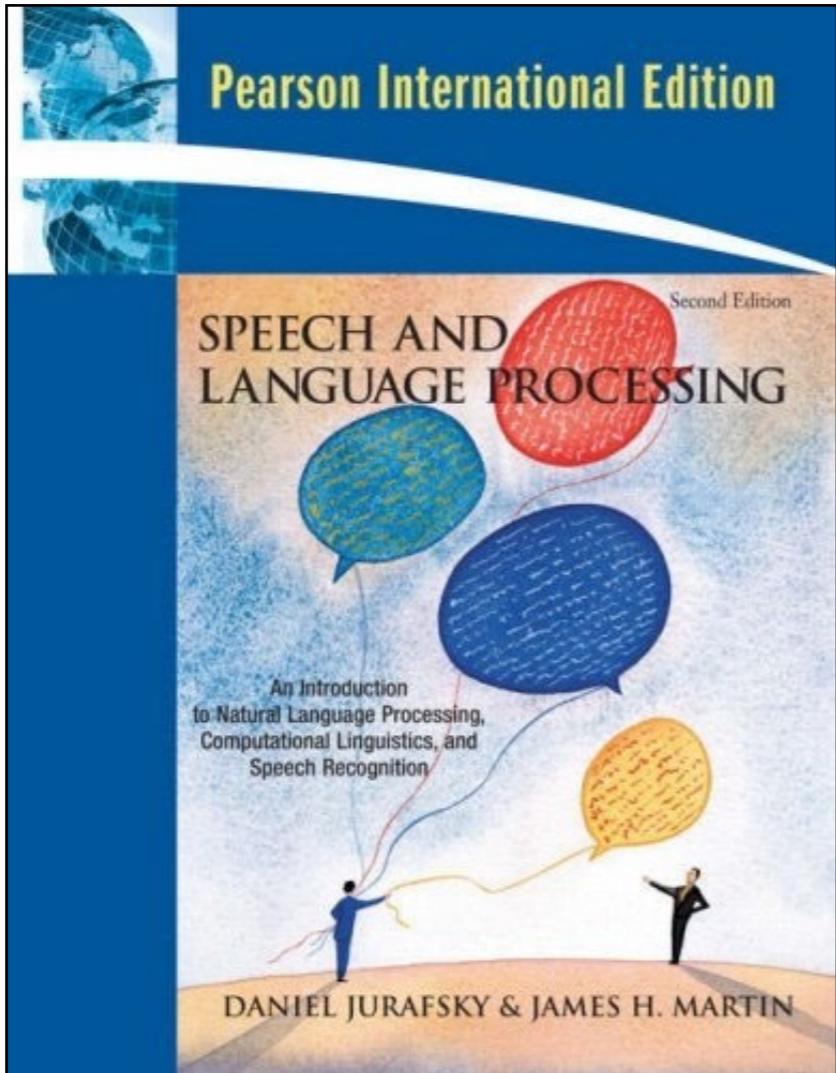
Topics in this class

- 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

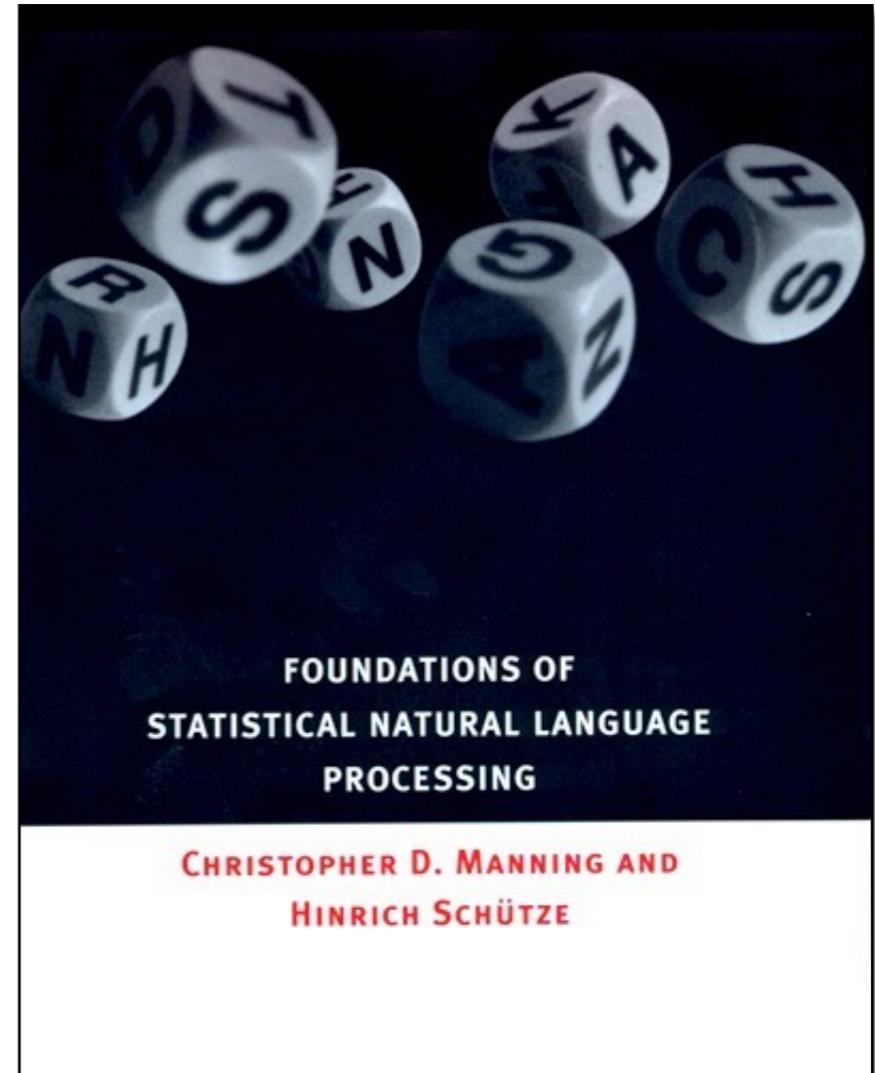
Lectures

- 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.
 - ▶ We will not accept late submissions.
- 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
- Please check the LSF for the exam registration deadline and **make sure you register** on time.

Resources

- Course website:
<https://coli-saar.github.io/cl19>
- Piazza (please sign up!):
<http://piazza.com/uni-saarland.de/fall2019/cl>
- Weekly voluntary tutorials with Alexandra Mayn