
Foundations of Natural Language Processing

Lecture 1

Introduction

Ivan Titov

(Slides based on those of Philipp Koehn, Alex Lascarides, Sharon Goldwater,
Shay Cohen, Khalil Sima'an)

16 January 2024



Structure of the first two lectures

- Administrative issues
- What is Natural Language Processing?
- Why NLP is hard? Ambiguity, Variability, Robustness, ...
- Why use *probabilistic* models (and machine learning) for NLP?

Background needed for this course

We assume you are familiar with most/all of the following:

- Basic Python programming
- Finite-state machines, regular languages
- Context-free grammars
- Dynamic programming (e.g. edit distance, Viterbi, and/or CKY algorithms)
- Concepts from machine learning (estimating probabilities, making predictions based on data)
- Probability theory (conditional probabilities, Bayes' Rule, independence and conditional independence, expectations)
- Vectors, logarithms
- Concepts of syntactic structure and semantics and relationship between them (ideally for natural language but at least for programming languages)
- Some basic linguistic concepts (e.g. parts of speech, inflection)

Where we are headed

INF2-iads discussed ideas and algorithms for NLP from a largely **formal, algorithmic** perspective. Here we build on that by

- Focusing on **real data** with all its complexities.
- Discussing some of the NLP techniques in more depth.
- Introducing many tasks and technologies that didn't fit into the Inf2-iads story.

Course organization

- Course organizer: Alex Lascarides
- Lecturers: Alex Lascarides and Ivan Titov
- 3 lectures per week (Tue, Thu & Fri, 12:10 - 13.00)
- We will use Learn and Drupal (<https://opencourse.inf.ed.ac.uk/fnlp>) for slides, lectures, labs, assignments, due dates, etc
- Labs: two groups, every two weeks (on Fridays, 13:10-14:30 or 14:30-16.00)
- Tutorials in small groups, every two weeks
Tutorials start next next week
- Course discussion forum: Piazza.

Check Learn for all the links and up-to-date information

Outside work required

In addition to attending lectures, you are expected to keep up with:

- Readings from textbook: *Speech and Language Processing*, Jurafsky and Martin: 3rd edition (online) and 2nd edition (paperback, International version, for chapters that aren't updated in 3rd ed)
- (Recommended) NLP techniques in Python: Bird, S., E. Klein and E. Loper, *Natural Language Processing with Python*, (2009) O'Reilly Media
- Tutorials and quizzes.
- Lectures are being recorded (some pre-recorded). The audience is *not* in shot.
- Two assignments, worth 25%
- Exam in April/May, worth 75% of final mark.

What is Natural Language Processing?

The collage illustrates various NLP applications:

- Google Translate:** Shows the translation of "Je ne sais pas!" from French to "I do not know!" in English.
- Document Editor:** Displays a document with grammar and spelling suggestions, such as "Add a space" and "Correct your spelling".
- Google Search:** Shows a search for "who is the first indian president" with a knowledge panel for Rajendra Prasad, the 1st President of India.
- Siri Advertisement:** Promotes Siri with the text "Learn more about Siri." and an image of an iPhone.
- ChatGPT:** Shows a user request for a limerick in French about a Labrador named Brutus, and the AI's response: "Bien sûr, voici un limerick en français sur Brutus, votre labrador :

Un labrador nommé Brutus, tout mignon,
Courait après sa balle dans le jardin, sans façon,
Avec sa queue en remuement,
Il apportait le bonheur, assurément,
Brutus, le chien joyeux de la maison."

What is Natural Language Processing?

Applications

- Machine Translation
- Information Retrieval
- Question Answering
- Dialogue Systems
- Information Extraction
- Summarization
- Sentiment Analysis
- ...

Core technologies

- Language modelling
- Part-of-speech tagging
- Syntactic parsing
- Named-entity recognition
- Coreference resolution
- Word sense disambiguation
- Semantic Role Labelling
- ...

This course

NLP is a big field! We focus mainly on core ideas and methods needed for technologies in the second column (and eventually for applications).

- Linguistic facts and issues
- Computational models and algorithms

More advanced methods and specific application areas covered in 4th/5th year courses:

- Natural Language Understanding, Generation and Machine Translation (NLU+)
- Text Technologies
- Automatic Speech Recognition

Example system

A computer provides information about train schedules:

System: Good evening. How can I help you?

User: I want to travel to London. Eh... from Edinburgh tomorrow morning.

S: What time do you want to arrive in London?

U: I want to depart at around half eight.

S: There is a train at seven twelve from Edinburgh Waverley, arriving at eleven fifty six in London Euston. Is that suitable for you?

...

What components we need for this system?
What problems can we expect to face?

What does an NLP system need to “know”?

- Language consists of many levels of structure
- Humans fluently integrate all of these in producing/understanding language
- Ideally, so would a computer!

Words

This is a simple sentence **WORDS**

Morphology

This is a simple sentence

be
3sg
present

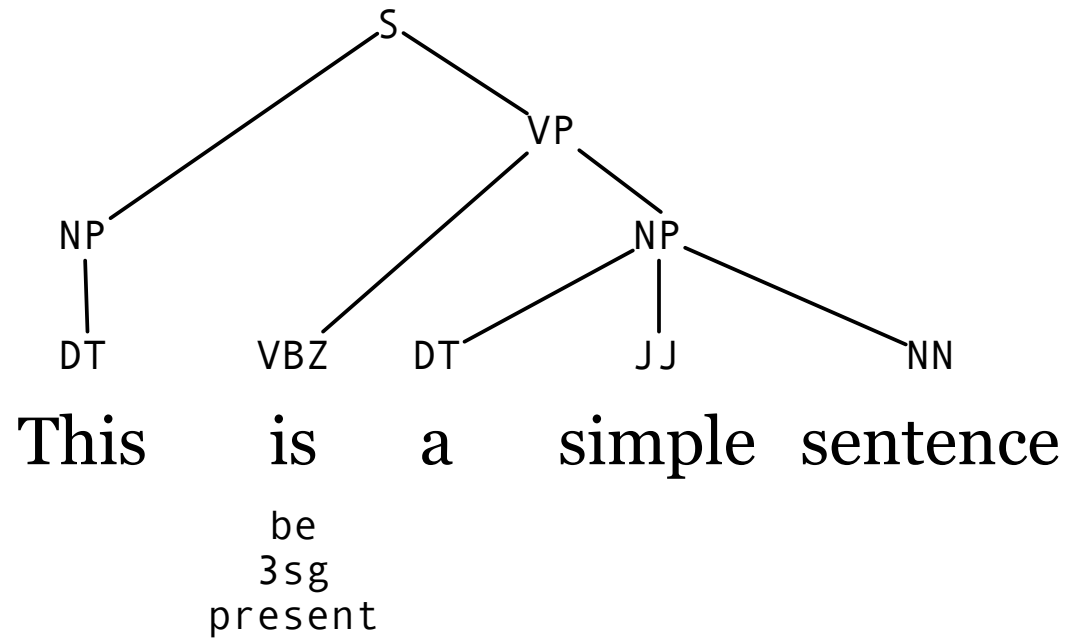
WORDS

MORPHOLOGY

Parts of Speech

DT	VBZ	DT	JJ	NN	PART OF SPEECH
This	is	a	simple	sentence	WORDS
	be 3sg present				MORPHOLOGY

Syntax



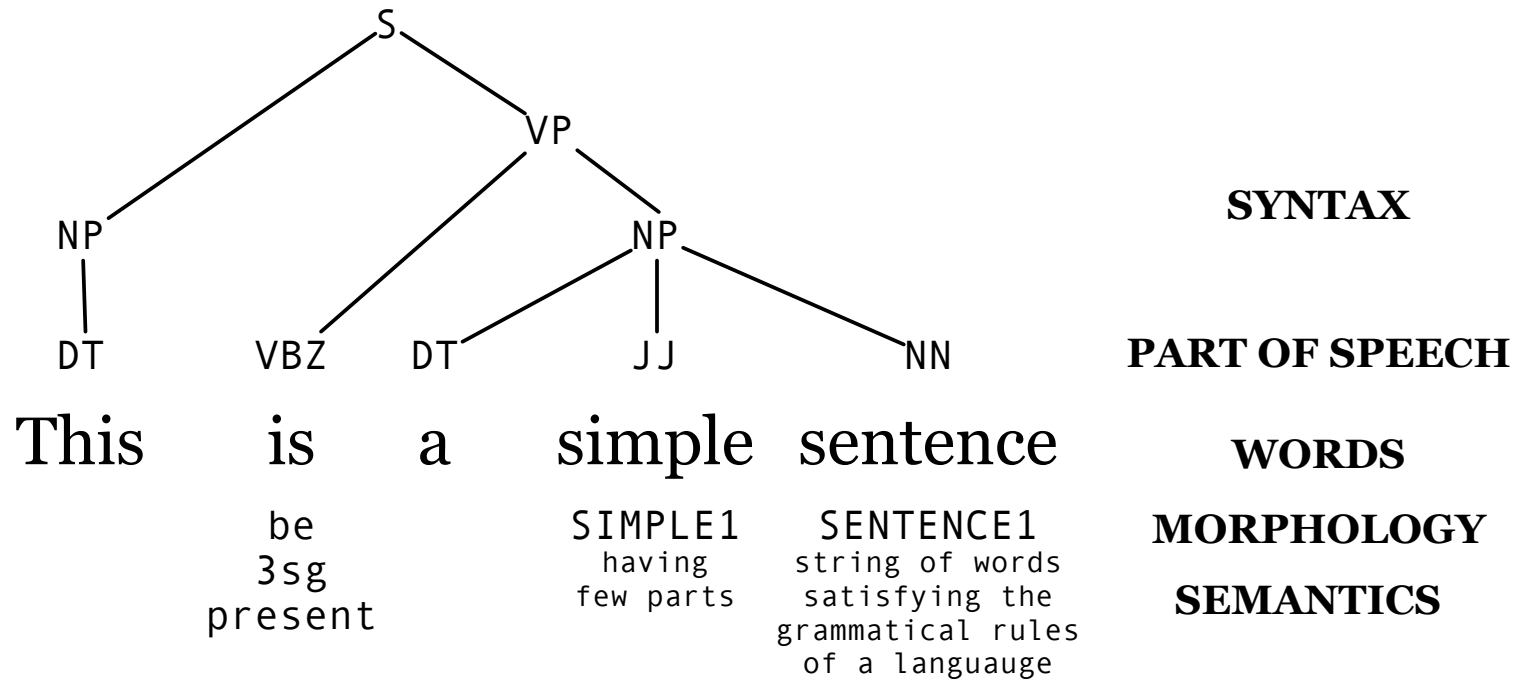
SYNTAX

PART OF SPEECH

WORDS

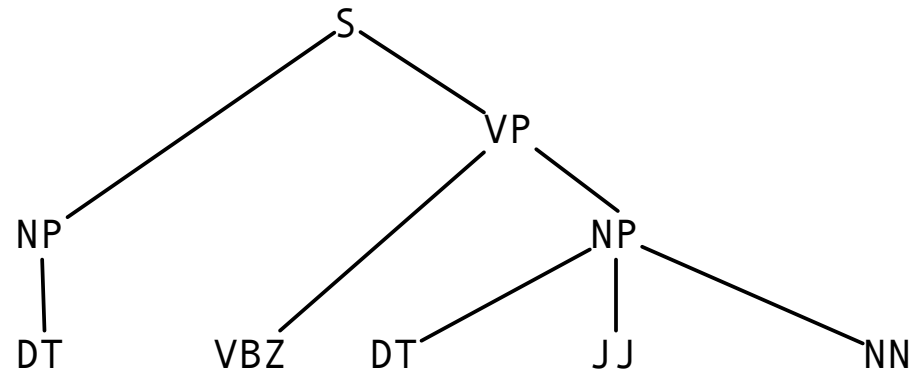
MORPHOLOGY

Semantics



$$\exists y(\text{this_dem}(x) \wedge \text{be}(e, x, y) \wedge \text{simple}(y) \wedge \text{sentence}(y))$$

Discourse



SYNTAX

PART OF SPEECH

This is a simple sentence

WORDS

be
3sg
present

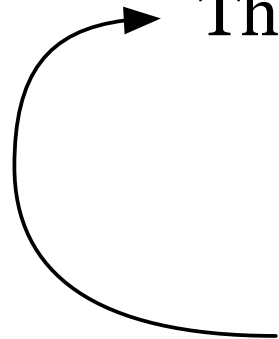
SIMPLE1
having
few parts

SENTENCE1
string of words
satisfying the
grammatical rules
of a language

MORPHOLOGY

SEMANTICS

CONTRAST



But it is an instructive one.

DISCOURSE

Do we really need modeling all these levels in an application?

- It depends...
- ... but let us consider a relatively simple application

Machine reading

London Bridge really is falling down. The bridge is being taken apart and moved. Its new home will be a small town in Arizona. This bridge is hundreds of years old. It stretches across the Thames River. Robert McCulloch saw this bridge and decided to bring it to the USA.

He paid more than two million dollars. It will cost him more than three million dollars to move it. Each stone will be marked. The pieces must fit when they reach their new home. All that work will not take place overnight. The job will take six years. The bridge is not small. It is longer than three football fields. It is almost as wide as one football field. In time, the London Bridge will stand high above a new river. Flags will be placed at both ends. Cars will cross it. A small town will be built next to the bridge. Most people in Arizona will never see London. But they will see a part of it in their own state.

1. Who bought a bridge?
2. Where will the bridge be re-built?
3. How long will it take?

Example from Dan Roth

Machine reading

London Bridge really is falling down. The bridge is being taken apart and moved. Its new home will be a small town in Arizona. This bridge is hundreds of years old. It stretches across the Thames River. Robert McCulloch saw this bridge and decided to bring it to the USA.

He paid more than two million dollars. It will cost him more than three million dollars to move it. Each stone will be marked. The pieces must fit when they reach their new home. All that work will not take place overnight. The job will take six years. The bridge is not small. It is longer than three football fields. It is almost as wide as one football field. In time, the London Bridge will stand high above a new river. Flags will be placed at both ends. Cars will cross it. A small town will be built next to the bridge. Most people in Arizona will never see London. But they will see a part of it in their own state.

1. Who bought a bridge?
2. Where will the bridge be re-built?
3. How long will it take?

Example from Dan Roth

Machine reading

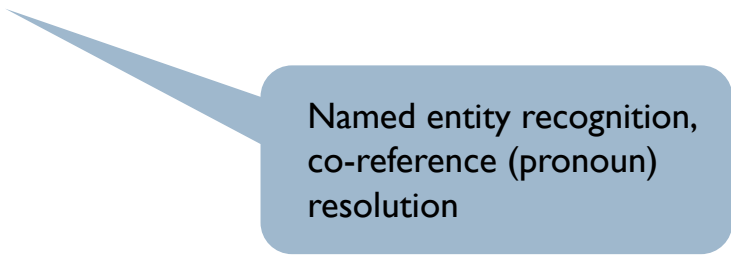
London Bridge really is falling down. The bridge is being taken apart and moved. Its new home will be a small town in Arizona. This bridge is hundreds of years old. It stretches across the Thames River. **Robert McCulloch** saw this **bridge** and decided to bring **it** to the USA.

He paid more than two million dollars. It will cost him more than three million dollars to move it. Each stone will be marked. The pieces must fit when they reach their new home. All that work will not take place overnight. The job will take six years. The bridge is not small. It is longer than three football fields. It is almost as wide as one football field. In time, the London Bridge will stand high above a new river. Flags will be placed at both ends. Cars will cross it. A small town will be built next to the bridge. Most people in Arizona will never see London. But they will see a part of it in their own state.

1. Who bought a bridge?

2. Where will the bridge be re-built?

3. How long will it take?



Named entity recognition,
co-reference (pronoun)
resolution

Example from Dan Roth

Machine reading

London Bridge really is falling down. The bridge is being taken apart and moved. Its new home will be a small town in Arizona. This bridge is hundreds of years old. It stretches across the Thames River. **Robert McCulloch** saw this **bridge** and decided to bring it to the USA.

He paid more than two million dollars. It will cost him more than three million dollars to move it. Each stone will be marked. The pieces must fit when they reach their new home. All that work will not take place overnight. The job will take six years. The bridge is not small. It is longer than three football fields. It is almost as wide as one football field. In time, the London Bridge will stand high above a new river. Flags will be placed at both ends. Cars will cross it. A small town will be built next to the bridge. Most people in Arizona will never see London. But they will see a part of it in their own state.

1. Who bought a bridge?
2. Where will the bridge be re-built?
3. How long will it take?

Example from Dan Roth

Machine reading

London Bridge really is falling down. The bridge is being taken apart and moved. Its new home will be a small town in Arizona. This bridge is hundreds of years old. It stretches across the Thames River. Robert McCulloch saw this bridge and decided to bring it to the USA.

He paid more than two million dollars. It will cost him more than three million dollars to move it. Each stone will be marked. The pieces must fit when they reach their new home. All that work will not take place overnight. The job will take six years. The bridge is not small. It is longer than three football fields. It is almost as wide as one football field. In time, the London Bridge will stand high above a new river. Flags will be placed at both ends. Cars will cross it. A small town will be built next to the bridge. Most people in Arizona will never see London. But they will see a part of it in their own state.

1. Who bought a bridge?

2. Where will the bridge be re-built?

3. How long will it take?

Syntactic parsing, shallow semantic analysis (argument identification)

Example from Dan Roth

Machine reading

London Bridge really is falling down. The bridge is being taken apart and moved. Its new home will be a small town in Arizona. This bridge is hundreds of years old. It stretches across the Thames River. Robert McCulloch saw this bridge and decided to bring it to the USA.

He paid more than two million dollars. It will cost him more than three million dollars to move it. Each stone will be marked. The pieces must fit when they reach their new home. All that work will not take place overnight. The job will take six years. The bridge is not small. It is longer than three football fields. It is almost as wide as one football field. In time, the London Bridge will stand high above a new river. Flags will be placed at both ends. Cars will cross it. A small town will be built next to the bridge. Most people in Arizona will never see London. But they will see a part of it in their own state.

saw X, paid (for X)

bought X

Learned inferences

1. Who bought a bridge?
2. Where will the bridge be re-built?
3. How long will it take?

Example from Dan Roth

Machine reading

London Bridge really is falling down. The bridge is being taken apart and moved. Its new home will be a small town in Arizona. This bridge is hundreds of years old. It stretches across the Thames River. **Robert McCulloch** saw this **bridge** and decided to bring it to the USA.

He paid more than two million dollars. It will cost him more than three million dollars to move it. Each stone will be marked. The pieces must fit when they reach their new home. All that work will not take place overnight. The job will take six years. The bridge is not small. It is longer than three football fields. It is almost as wide as one football field. In time, the London Bridge will stand high above a new river. Flags will be placed at both ends. Cars will cross it. A small town will be built next to the bridge. Most people in Arizona will never see London. But they will see a part of it in their own state.

1. Who bought a bridge?

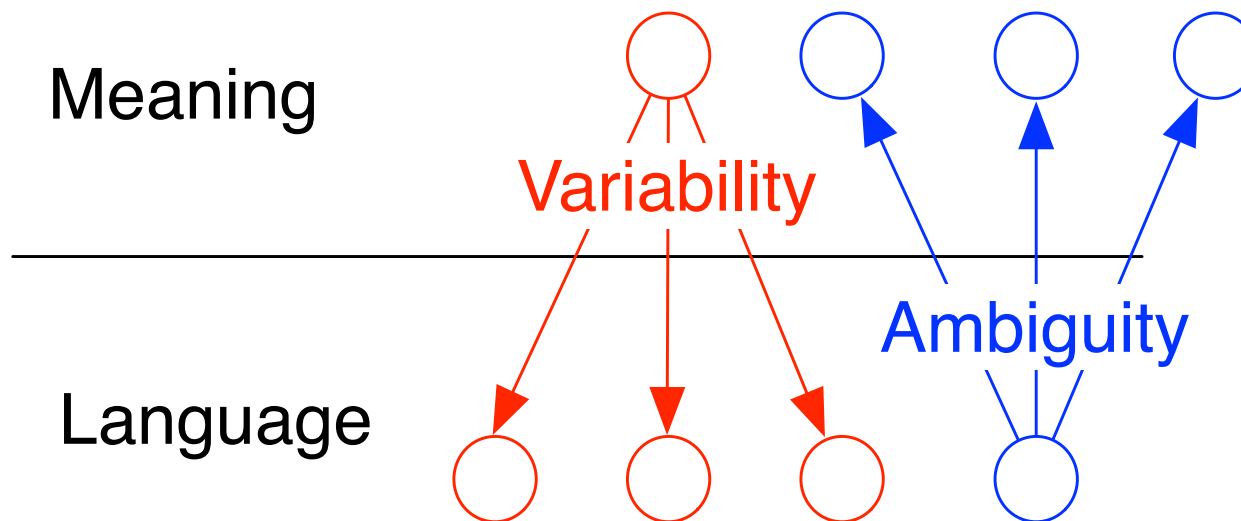
2. Where will the bridge be re-built?

3. How long will it take?

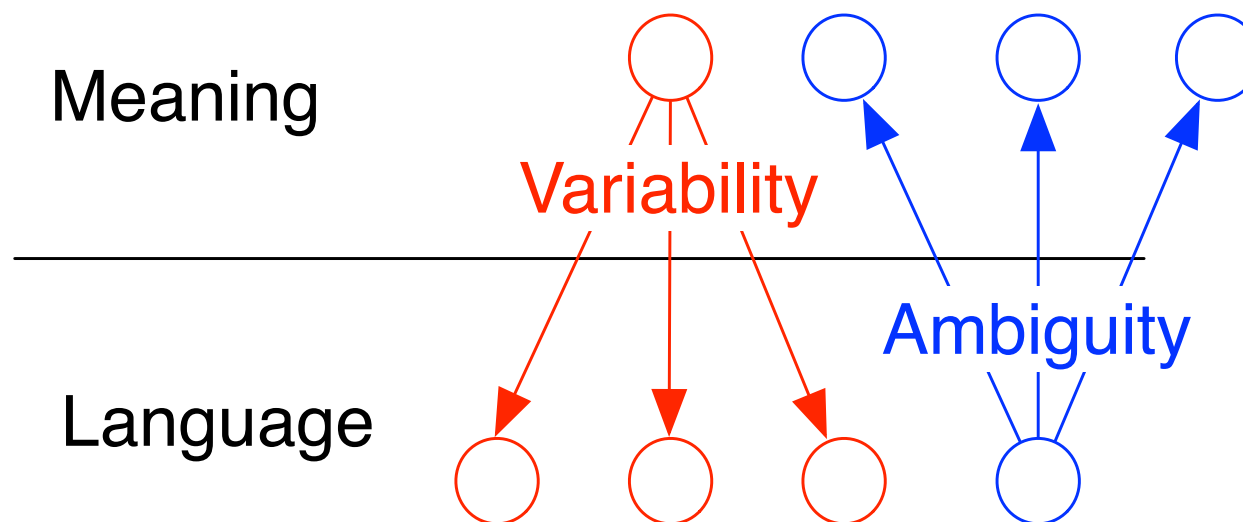
To perform machine reading, a machine needs to perform linguistic analysis at different levels (explicitly or implicitly)

Example from Dan Roth

Why is NLP hard?



Why is NLP hard?



Variability:

He drew the house

He made a sketch of the house

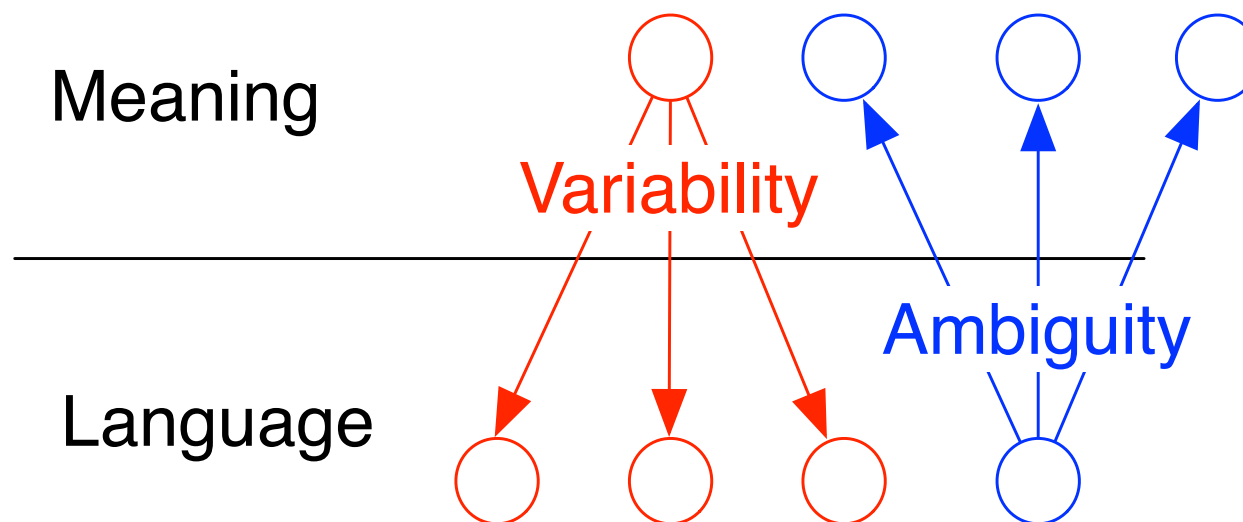
He showed me his drawing of the house

He portrayed the house in his paintings

He drafted the house in his sketchbook

...

Why is NLP hard?



Ambiguity:

She **drew** a picture of herself ~ *sketched, made a drawing of*

A cart **drawn** by two horses... ~ *pulled*

He **drew** crowds wherever he went ... ~ *attracted*

The driver slowed as he **drew** even with me ~ *proceeded*

The officer **drew** a gun and pointed it at ... ~ *took out, produced*

..

Why is NLP hard?

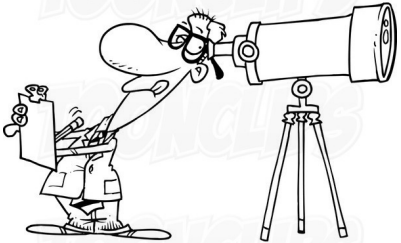
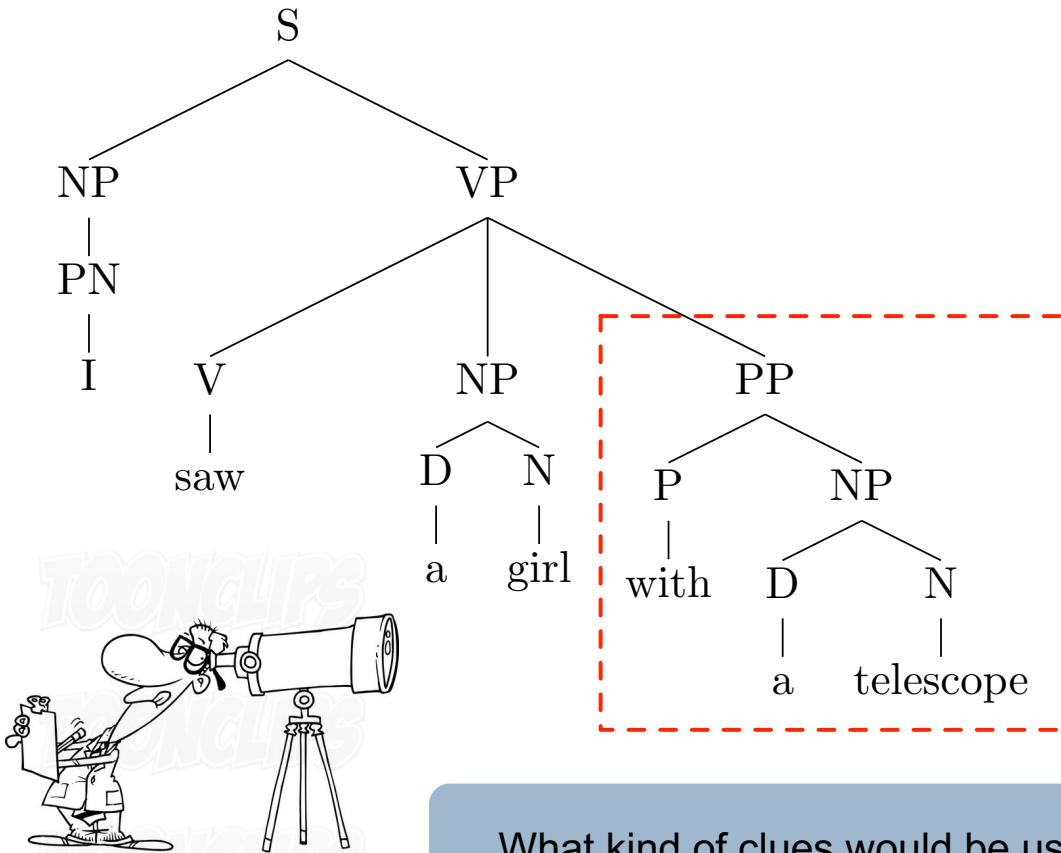
Ambiguity at many levels:

- Homophones: **blew** and **blue**
- Word senses: **bank** (finance or river?)
- Part of speech: **chair** (noun or verb?)
- Syntactic structure: **I saw a girl with a telescope**
 - **let us look into this in more detail!**
- Quantifier scope: **Every child loves some movie**
- Multiple: **I saw her duck**
- Reference: John dropped the goblet onto the glass table and it broke.
- Discourse: The meeting is cancelled. Nicholas isn't coming to the office today.

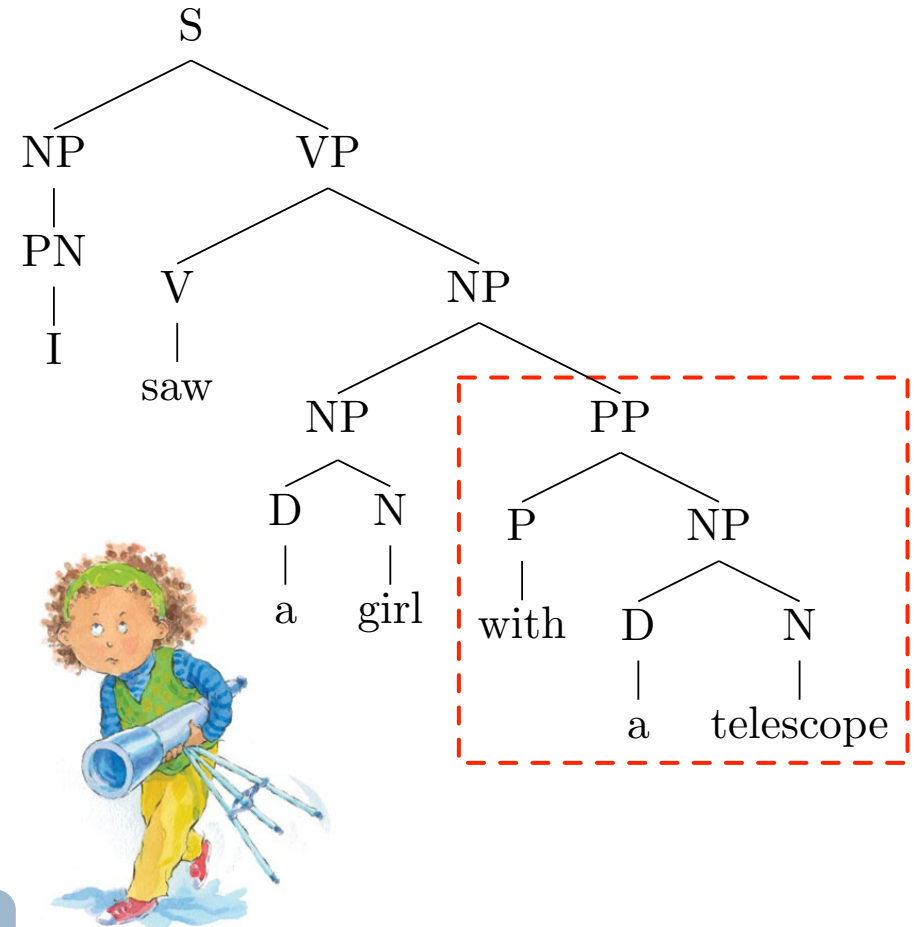
How can we model ambiguity, and choose the correct analysis in context?

Syntactic ambiguity

► Prepositional phrase attachment ambiguity



Copyright © Ron Leishman * <http://ToonClips.com/3005>



What kind of clues would be useful?

PP-attachment ambiguity is just one example type of ambiguity

How serious is this problem in practice?

Syntactic ambiguity

- ▶ Example with 3 preposition phrases, 5 interpretations:
 - ▶ *Put the block ((in the box on the table) in the kitchen)*
 - ▶ *Put the block (in the box (on the table in the kitchen))*
 - ▶ *Put ((the block in the box) on the table) in the kitchen.*
 - ▶ *Put (the block (in the box on the table)) in the kitchen.*
 - ▶ *Put (the block in the box) (on the table in the kitchen)*

Syntactic ambiguity

▶ Example with **3 preposition phrases, 5 interpretations:**

- ▶ *Put the block ((in the box on the table) in the kitchen)*
- ▶ *Put the block (in the box (on the table in the kitchen))*
- ▶ *Put ((the block in the box) on the table) in the kitchen.*
- ▶ *Put (the block (in the box on the table)) in the kitchen.*
- ▶ *Put (the block in the box) (on the table in the kitchen)*

▶ **A general case:**

$$Cat_n = \binom{2n}{n} - \binom{2n}{n-1} \sim \frac{4^n}{n^{3/2}\sqrt{\pi}}$$

Catalan numbers

1, 2, 5, 14, 42, 132, 429, 1430, 4862, 16796, 58786, ...

Syntactic ambiguity

▶ Example with **3 preposition phrases, 5 interpretations:**

- ▶ *Put the block ((in the box on the table) in the kitchen)*
- ▶ *Put the block (in the box (on the table in the kitchen))*
- ▶ *Put ((the block in the box) on the table) in the kitchen.*
- ▶ *Put (the block (in the box on the table)) in the kitchen.*
- ▶ *Put (the block in the box) (on the table in the kitchen)*

▶ A **general case:**

$$Cat_n = \binom{2n}{n} - \binom{2n}{n-1} \sim \frac{4^n}{n^{3/2}\sqrt{\pi}}$$

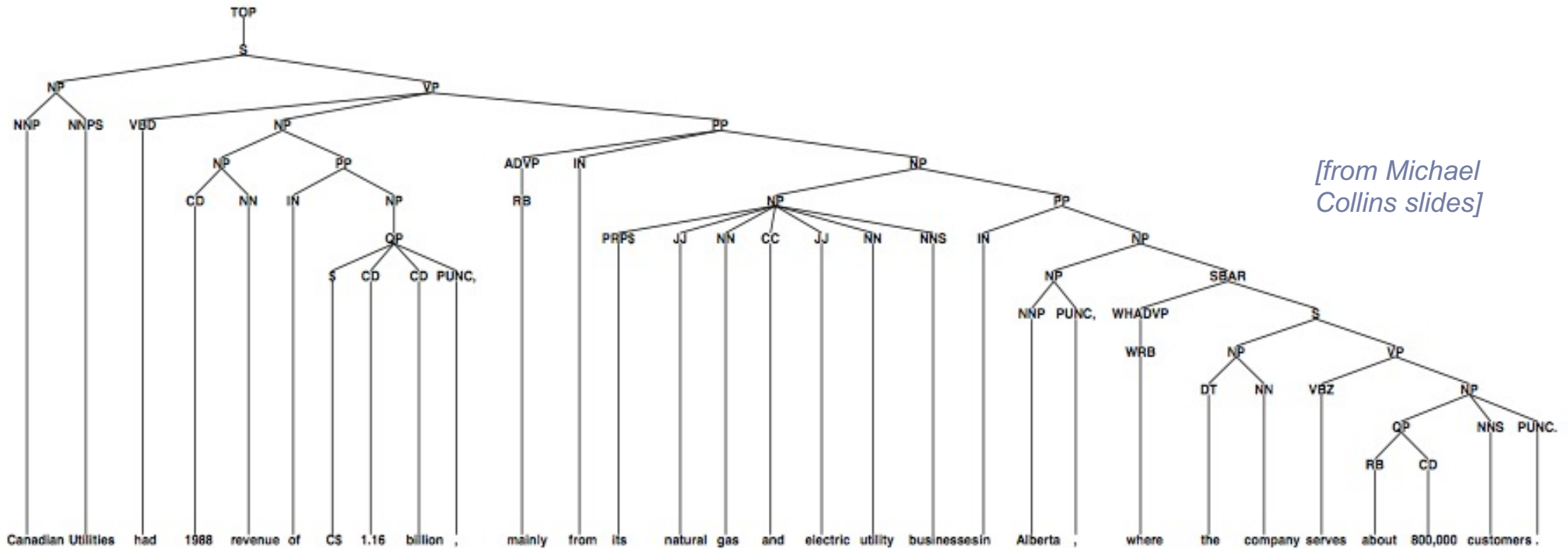
Catalan numbers

1, 2, 5, 14, 42, 132, 429, 1430, 4862, 16796, 58786, ...

PP-attachment is only one type of syntactic ambiguity in a sentence

Syntactic ambiguity

- ▶ A typical tree from a standard dataset (Penn treebank WSJ)



[from Michael Collins slides]

Canadian Utilities had 1988 revenue of \$ 1.16 billion , mainly from its natural gas and electric utility businesses in Alberta , where the company serves about 800,000 customers .

Why is NLP hard?

Ambiguity at many levels:

- Homophones: **blew** and **blue**
- Word senses: **bank** (finance or river?)
- Part of speech: **chair** (noun or verb?)
- Syntactic structure: **I saw a man with a telescope**
- Quantifier scope: **Every child loves some movie**
- Multiple: **I saw her duck**
- Reference: **John dropped the goblet onto the glass table and it broke.**
- Discourse: **The meeting is cancelled. Nicholas isn't coming to the office today.**

How can we model ambiguity, and choose the correct analysis in context?

Real Newspaper Headlines

- Ban on Nude Dancing on Governor's Desk
- Iraqi Head Seeks Arms
- Local High School Dropouts Cut in Half
- Red Tape Holds Up New Bridges
- Juvenile Court to Try Shooting Defendant
- Kids Make Nutritious Snacks

Collected by Chris Manning