

# Overview of Natural Language Processing

## Part II & ACS L90

### Lecture 1: Overview of Overview of Natural Language Processing

Weiwei Sun and Michael Schlichtkrull

Department of Computer Science and Technology  
University of Cambridge

Michaelmas 2023/24

What is language?

Overview of Natural Language Processing

C, Lisp, Python, Ruby, Scala, ...  
English, Welsh, Afrikaans, Mandarin, ...  
English as a Second Language, ...  
sign languages, ...  
Sanskrit, ...  
dolphin language  
...

# What is language?

## Obvious?

- Seems obvious (to language users)
- Not obvious (to language scientists)

# What is language?

## Obvious?

- Seems obvious (to language users)
- Not obvious (to language scientists)

💡 Are emojis part of your language?




# What is language?


## Obvious?

- Seems obvious (to language users)
- Not obvious (to language scientists)

## 💡 Are emojis part of your language?



A screenshot of a tweet from Hillary Clinton. The tweet text asks for a response to a question about student loan debt using emojis. The tweet has 8.7K likes and a 'Read 7.4K replies' button.

 **Hillary Clinton**   
@HillaryClinton · [Follow](#) 

How does your student loan debt make you feel?  
Tell us in 3 emojis or less.

7:49 PM · Aug 12, 2015 

---

 8.7K  Reply  Share

[Read 7.4K replies](#)

# What is language?

## Obvious?

- Seems obvious (to language users)
- Not obvious (to language scientists)

## 💡 Are emojis part of your language?

# Word of the Year 2015

The Oxford Word of the Year 2015 is... 😊

That's right – for the first time ever, the Oxford Dictionaries Word of the Year is a pictograph: 😊, officially called the 'Face with Tears of Joy' emoji, though you may know it by other names. There were other strong contenders from a range of fields but 😊 was chosen as the 'word' that best reflected the ethos, mood, and preoccupations of 2015.

# What is language?

## CAMBRIDGE DICTIONARY

- a system of communication consisting of sounds, words, and grammar
- a system of communication used by people living in a particular country
- a system of symbols and rules for writing instructions for computers
- the way that someone speaks or writes, for example, the kind of words and phrases that they use
- the special words and phrases used by people who do a particular type of work: *legal language*
- rude or offensive words



# What is language?

## CAMBRIDGE DICTIONARY

- a system of communication consisting of sounds, words, and grammar
- a system of communication used by people living in a particular country
- a system of symbols and rules for writing instructions for computers
- the way that someone speaks or writes, for example, the kind of words and phrases that they use
- the special words and phrases used by people who do a particular type of work: *legal language*
- rude or offensive words

this is a description rather than a definition

# What is language?

## CAMBRIDGE DICTIONARY

- a system of communication consisting of sounds, **words**, and grammar
- a system of communication used by people living in a particular country
- a system of symbols and rules for writing instructions for computers
- the way that someone speaks or writes, for example, the kind of words and phrases that they use
- the special words and phrases used by people who do a particular type of work: *legal language*
- rude or offensive words

this is a description rather than a definition

### ❓ What is a **word**?

a single unit of **language** that has meaning and can be spoken or written.

# What is language?

## CAMBRIDGE DICTIONARY

- a system of communication consisting of sounds, words, and grammar
- a system of communication used by people living in a particular country
- a system of symbols and rules for writing instructions for computers
- the way that someone speaks or writes, for example, the kind of words and phrases that they use
- the special words and phrases used by people who do a particular type of work: *legal language*
- rude or offensive words

this is a description rather than a definition

### ❓ What is a word?

a single unit of language that has meaning and can be spoken or written.

# What is language?

A formal language is a set of strings over an alphabet.

## Strings and languages

- A string of length  $n$  over an alphabet  $\Sigma$  is an ordered  $n$ -tuple of elements of  $\Sigma$ .
- $\Sigma^*$  denotes the set of all strings over  $\Sigma$  of finite length.
- Given an alphabet  $\Sigma$  any subset of  $\Sigma^*$  is a formal language over alphabet  $\Sigma$ .

# What is language?

A formal language is a set of strings over an alphabet.

## Strings and languages

- A string of length  $n$  over an alphabet  $\Sigma$  is an ordered  $n$ -tuple of elements of  $\Sigma$ .
- $\Sigma^*$  denotes the set of all strings over  $\Sigma$  of finite length.
- Given an alphabet  $\Sigma$  any subset of  $\Sigma^*$  is a formal language over alphabet  $\Sigma$ .

for formal languages, we have a precise definition

# What is language?

A formal language is a set of strings over an alphabet.

## Strings and languages

- A string of length  $n$  over an alphabet  $\Sigma$  is an ordered  $n$ -tuple of elements of  $\Sigma$ .
- $\Sigma^*$  denotes the set of all strings over  $\Sigma$  of finite length.
- Given an alphabet  $\Sigma$  any subset of  $\Sigma^*$  is a formal language over alphabet  $\Sigma$ .

for formal languages, we have a precise definition

💡 Is it adequate to characterise a natural language in the same way?

# What we are going to do?

This course will focus on characterising some languages, especially **English**, **intuitively**, **linguistically**, **mathematically** and **computationally**.

We will focus on scientific approaches to achieve reliable language technologies.

*Seek simplicity and distrust it.*

— *Alfred North Whitehead*

*All models are wrong, but some models are useful.*

— *George Box*

PART-II  
ACS L90

Overview of **Natural Language** Processing

English, Welsh, Afrikaans, Mandarin, ...  
English as a Second Language, ...  
Sanskrit, ...  
sign languages



PART-II  
ACS L90

Overview of Natural Language

Processing

computational models

English, Welsh, Afrikaans, Mandarin, ...  
English as a Second Language, ...  
Sanskrit, ...  
sign languages

breadth over depth

computational models

PART-II  
ACS L90 **Overview** of Natural Language Processing

English, Welsh, Afrikaans, Mandarin, ...  
English as a Second Language, ...  
Sanskrit, ...  
sign languages

breadth over depth

computational models

PART-II  
ACS L90

## Overview of Natural Language Processing

L95  
L98  
L99  
L101  
...

English, Welsh, Afrikaans, Mandarin, ...  
English as a Second Language, ...  
Sanskrit, ...  
sign languages

## Goal and Scope

**a system of communication** consisting of sounds, words, and grammar

**a system of communication** used by people living in a particular country

a system of symbols and rules for writing instructions for computers

the way that someone speaks or writes, for example, the kind of words and phrases that they use

the special words and phrases used by people who do a particular type of work: legal language

rude or offensive words

# Conversational User Interface

💡 **How can siri put the elephant into the fridge?**

# Conversational User Interface

💡 **How can siri put the elephant into the fridge?**

*put the elephant  
into the fridge*

— semantic parsing —>

```
open(fridge.door)
put(elephant,fridge)
close(fridge.door)
```

# Conversational User Interface

💡 How can siri put the elephant into the fridge?

*put the elephant  
into the fridge*

semantic parsing →

```
open(fridge.door)
put(elephant, fridge)
close(fridge.door)
```

Execute the code





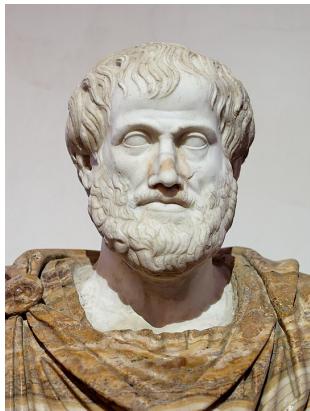
# Dialogue System

## Example

A Could you please close the door from the outside?

B [...]

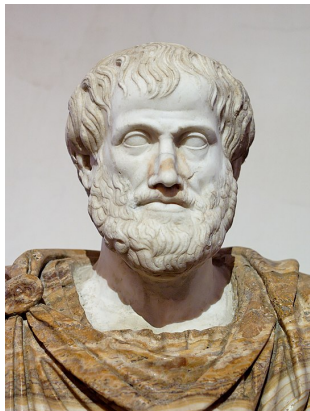
# Aristotle's Syllogism



Syllogism=Syn- + logos

- All men are mortal.
- Socrates is a man.
- Therefore, Socrates is mortal.

# Aristotle's Syllogism

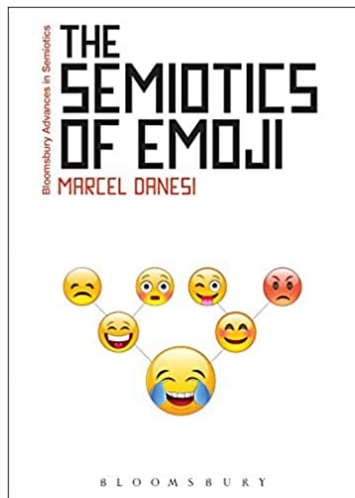


Syllogism=Syn- + logos

- All men are mortal.
- Socrates is a man.
- Therefore, Socrates is mortal.

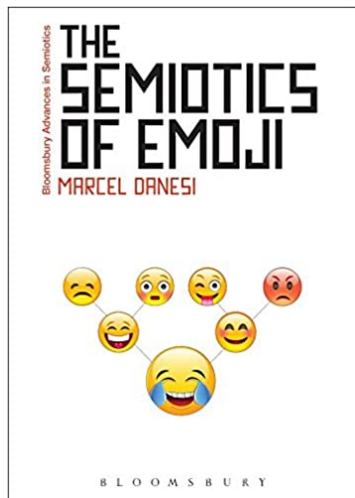
language goes beyond communication

# How can we build amazing automatic systems?



- 1 Emoji and Writing Systems
- 2 Emoji Uses
- 3 Emoji Competence
- 4 Emoji Semantics
- 5 Emoji Grammar
- 6 Emoji Pragmatics
- 7 Emoji Variation
- 8 Emoji Spread
- 9 Universal Languages
- 10 A Communication Revolution?

# How can we build amazing automatic systems?



- 1 Emoji and Writing Systems
- 2 Emoji Uses
- 3 Emoji Competence
- 4 Emoji Semantics
- 5 Emoji Grammar
- 6 Emoji Pragmatics
- 7 Emoji Variation
- 8 Emoji Spread
- 9 Universal Languages
- 10 A Communication Revolution?

- Language be studied scientifically
- Scientific study of language enables reliable language technologies

## A call-for-paper (1)

ACL (=Annual Meeting of the Association for **Computational Linguistics**) 2020 has the goal of a broad technical program. Relevant topics for the conference include, but are not limited to, the following areas:

- Theory and Formalism in NLP (Linguistic and Mathematical)
- Machine Learning for NLP
- Cognitive Modeling and Psycholinguistics
- Phonology, Morphology and Word Segmentation
- Syntax: Tagging, Chunking and Parsing
- Semantics: Lexical
- Semantics: Sentence Level
- Semantics: Textual Inference and Other Areas of Semantics
- Discourse and Pragmatics
- Generation
- Resources and Evaluation
- Interpretability and Analysis of Models for NLP

## A call for papers (2)

*ACL 2020 has the goal of a broad technical program. Relevant topics for the conference include, but are not limited to, the following areas:*

- Language Grounding to Vision, Robotics and Beyond
- Speech and Multimodality
- Information Extraction
- Information Retrieval and Text Mining
- Machine Translation
- Question Answering
- Dialogue and Interactive Systems
- Summarization
- Sentiment Analysis, Stylistic Analysis, and Argument Mining
- (other) NLP Applications
- Computational Social Science and Social Media
- Ethics and NLP

## Topics in This Course



# What does it mean to *know* a language?

*Some yinkish dripners blorked quastofically into the nindin with the pidibs.*

the example is partly from A. Carnie's *Syntax: A Generative Introduction*

# What does it mean to *know* a language?

*Some yinkish dripners **blorked** quastofically into the nindin with the pidibs.*

the example is partly from A Carnie's *Syntax: A Generative Introduction*

- there was a BLORK event;

# What does it mean to *know* a language?

*Some yinkish dripners blorked quastofically into the nindin with the pidibs.*

the example is partly from A Carnie's *Syntax: A Generative Introduction*

- there was a BLORK event;
- it happened in the PAST;

# What does it mean to *know* a language?

*Some yinkish dripners blorked quastofically into the nindin with the pidibs.*

the example is partly from A Carnie's *Syntax: A Generative Introduction*

- there was a BLOrk event;
- it happened in the PAST;
- the AGENT of BLOrk is dripners;

# What does it mean to *know* a language?

*Some yinkish dripners blorked quastofically into the nindin with the pidibs.*

the example is partly from A Carnie's *Syntax: A Generative Introduction*

- there was a BLORK event;
- it happened in the PAST;
- the AGENT of BLORK is dripners;
- the dripners were YINKISH;

# What does it mean to *know* a language?

*Some yinkish dripners blorked quastofically into the nindin with the pidibs.*

the example is partly from A Carnie's *Syntax: A Generative Introduction*

- there was a BLORK event;
- it happened in the PAST;
- the AGENT of BLORK is dripners;
- the dripners were YINKISH;
- **SOME but NOT ALL dripners blorked;**

# What does it mean to *know* a language?

*Some yinkish dripners blorked quastofically into the nindin with the pidibs.*

the example is partly from A Carnie's *Syntax: A Generative Introduction*

- there was a BORK event;
- it happened in the PAST;
- the AGENT of BORK is dripners;
- the dripners were YINKISH;
- SOME but NOT ALL dripners blorked;
- WITH THE PIDIBS may talk about NINDIN or BORK;

## Structuring a sentence

*Some yinkish dripners blooked quastofically into the nindin with the pidibs*

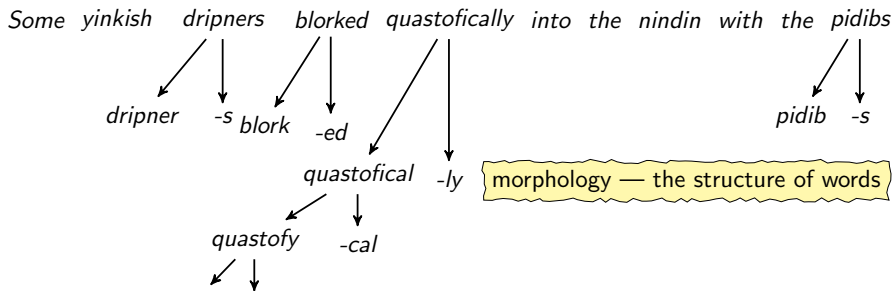


# Structuring a sentence

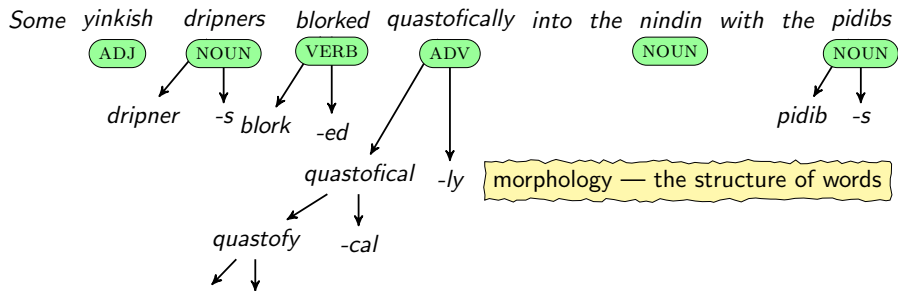
*Some yinkish dripners blooked quastofically into the nindin with the pidibs*

*dripner*   *-s*   *blook*   *-ed*   *pidib*   *-s*

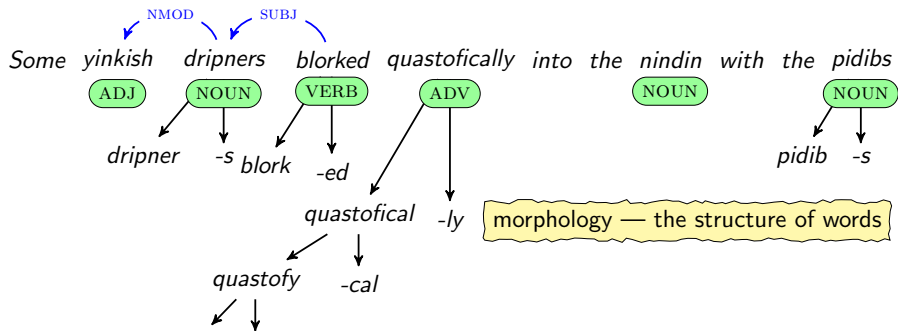
# Structuring a sentence



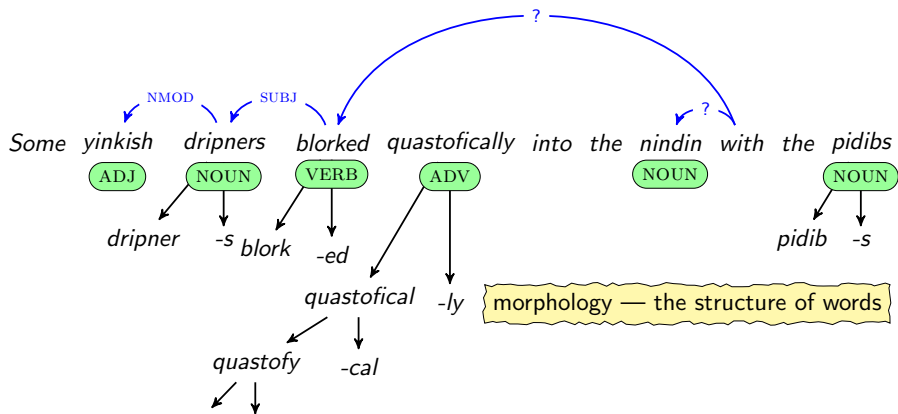
# Structuring a sentence



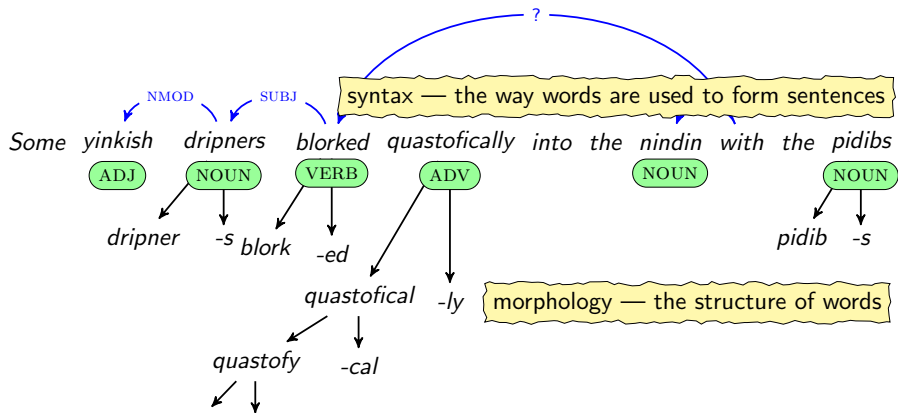
# Structuring a sentence



# Structuring a sentence



# Structuring a sentence

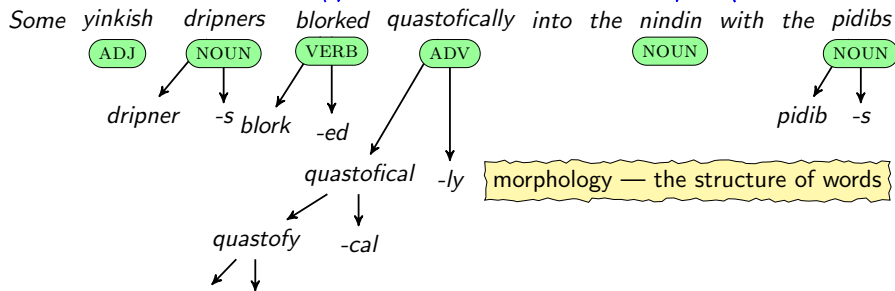


# Structuring a sentence

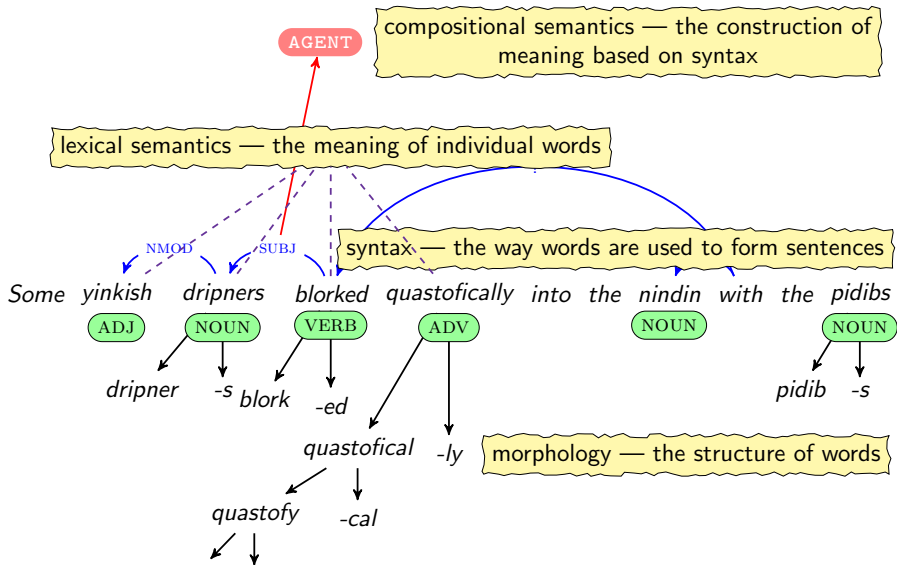
AGENT

compositional semantics — the construction of meaning based on syntax

syntax — the way words are used to form sentences

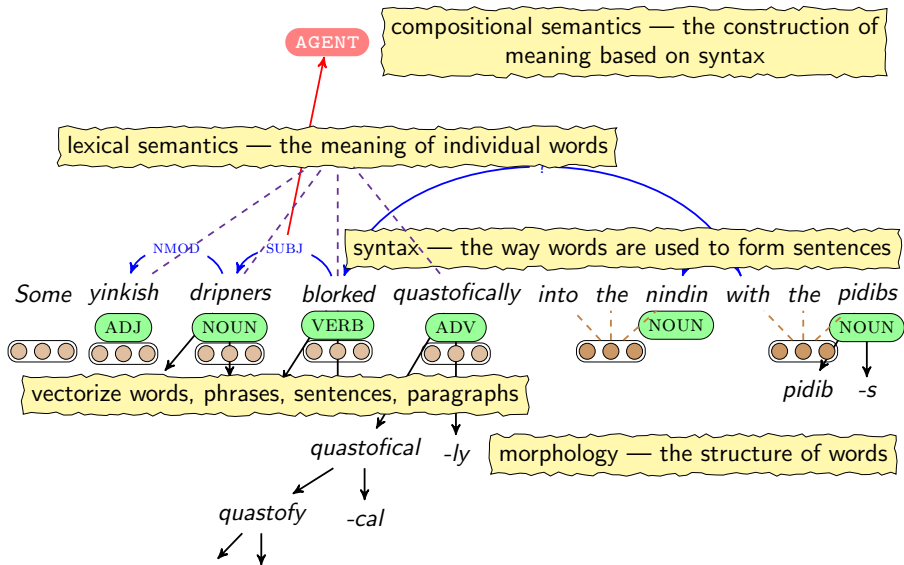


# Structuring a sentence

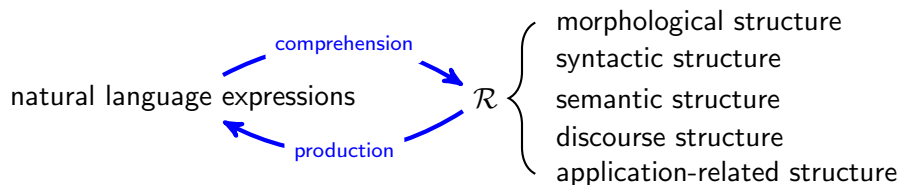




# Structuring a sentence



# Form transformation



## CoNLL shared tasks

- The SIGNLL Conference on Computational Natural Language Learning
- <https://www.conll.org/previous-tasks>

2019	Cross-Framework Meaning Representation Parsing
2018/2017	Multilingual Parsing from Raw Text to Universal Dependencies
2018/2017	Universal Morphological Reinflection
2016/2016	(Multilingual) Shallow Discourse Parsing
2014/2013	Grammatical Error Correction
2012/2011	Modelling (Multilingual) Unrestricted Coreference in OntoNotes
2010	Hedge Detection
2009/2008	Syntactic and Semantic Dependencies in English/Multiple Languages
2007/2006	Multi-Lingual Dependency Parsing (Domain Adaptation)
2005/2004	Semantic Role Labeling
2003/2002	Language-Independent Named Entity Recognition
2001	Clause Identification
2000	Chunking
1999	NP Bracketing

input words

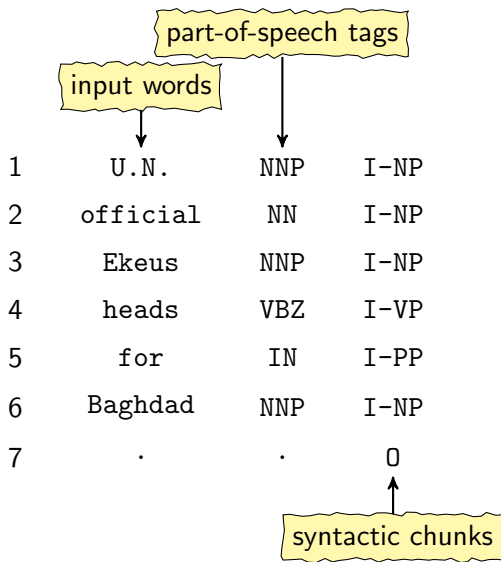


- 1 U.N.
- 2 official
- 3 Ekeus
- 4 heads
- 5 for
- 6 Baghdad
- 7 .

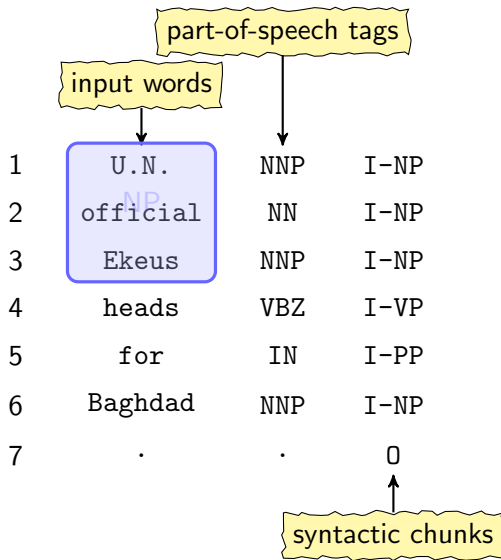
# CoNLL ST 1999/2000/2002/2003/2006/2007/2017/2018

	input words	part-of-speech tags
1	U.N.	NNP
2	official	NN
3	Ekeus	NNP
4	heads	VBZ
5	for	IN
6	Baghdad	NNP
7	.	.

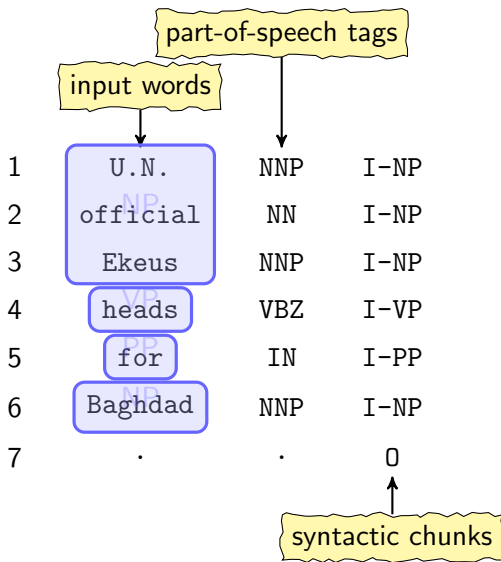
# CoNLL ST 1999/2000/2002/2003/2006/2007/2017/2018



# CoNLL ST 1999/2000/2002/2003/2006/2007/2017/2018

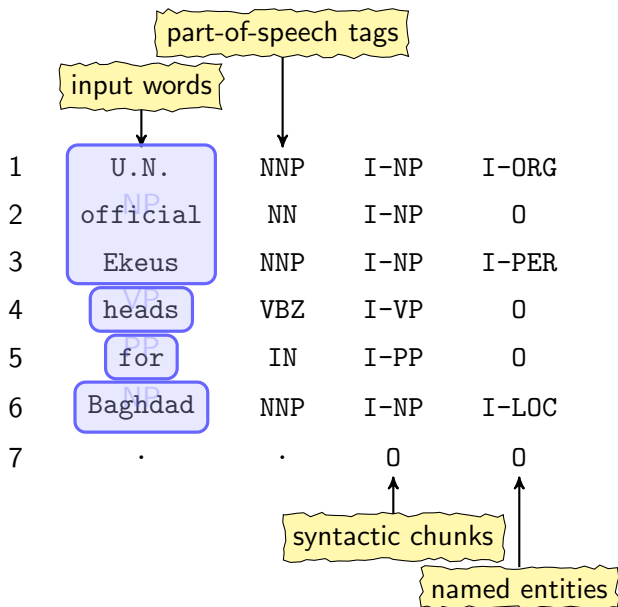


# CoNLL ST 1999/2000/2002/2003/2006/2007/2017/2018

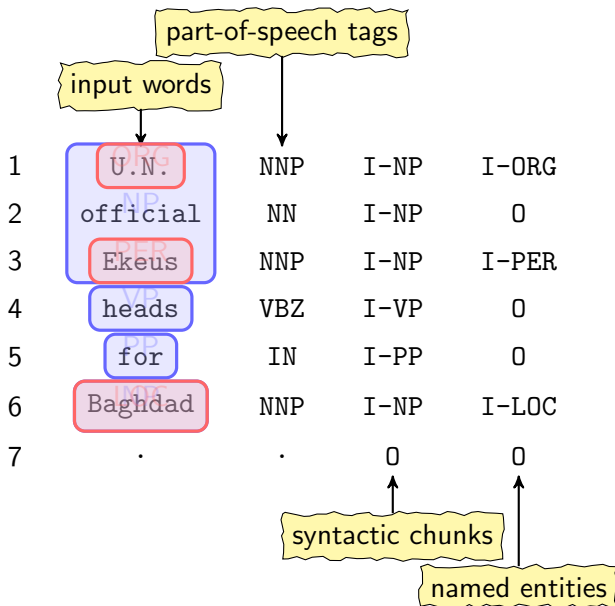




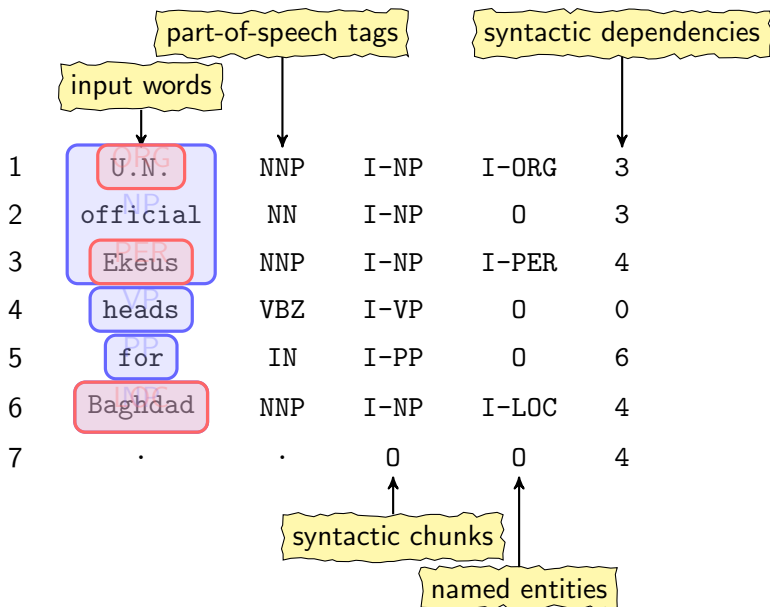
# CoNLL ST 1999/2000/2002/2003/2006/2007/2017/2018



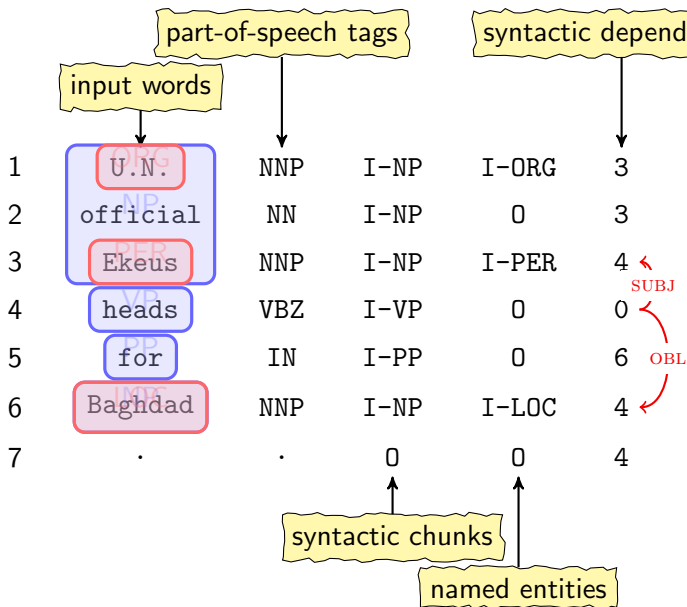
# CoNLL ST 1999/2000/2002/2003/2006/2007/2017/2018



# CoNLL ST 1999/2000/2002/2003/2006/2007/2017/2018



# CoNLL ST 1999/2000/2002/2003/2006/2007/2017/2018



# NLP: the computational modelling of human language

- *Morphology* — the structure of words: lecture 2.
  - *Syntax* — the way words are used to form phrases: lectures 3 and 4.
  - *Semantics* — the meaning of words and sentences: lecture 6.
  - *Language generation* — how to generate text: lecture 11.
- 
- *Symbolic models* — finite-state machines, context-free grammars, logic: lectures 2, 4, 5.
  - *Statistical models* — classification: lectures 3, 4, 6, 7 and 10.
  - *Neural models* — representation learning: lectures 9 and 10.
- 
- *Text summarisation* — lecture 8.
  - *Computation in linguistic analysis* — lecture 12.

## Topics we won't discuss

If we know “some yinkish dripners blorked quastofically into the nindin with the pidibs” is true, do we also know that “some yinkish dripners blorked quastofically” is also true?

If we know “some yinkish dripners blorked quastofically into the nindin with the pidibs” is true, do we also know that “some but not all yinkish dripners blorked quastofically with the pidibs” is also true?

Why NLP is difficult?

- You will see in the next 11 lectures.
- You will understand in your practicals.



# Logistics

- 12 lectures + 3 practicals
- after-class reading is mostly with Dan Jurafsky and James Martin's *Speech and Language Processing*, available at <https://web.stanford.edu/~jurafsky/slp3/>.

# Readings

- BBC Future: The language that doesn't use 'no'. <https://www.bbc.com/future/article/20220804-kusunda-the-language-isolate-with-no-word-for-no>
- Introduction at SEP.  
<https://plato.stanford.edu/entries/computational-linguistics/>