Overview of Natural Language Processing
Part II & ACS L390

Lecture 1: Overview of Overview of Natural Language Processing

Weiwei Sun and Yulong Chen

Department of Computer Science and Technology
University of Cambridge

Michaelmas 2024/25

Assessment

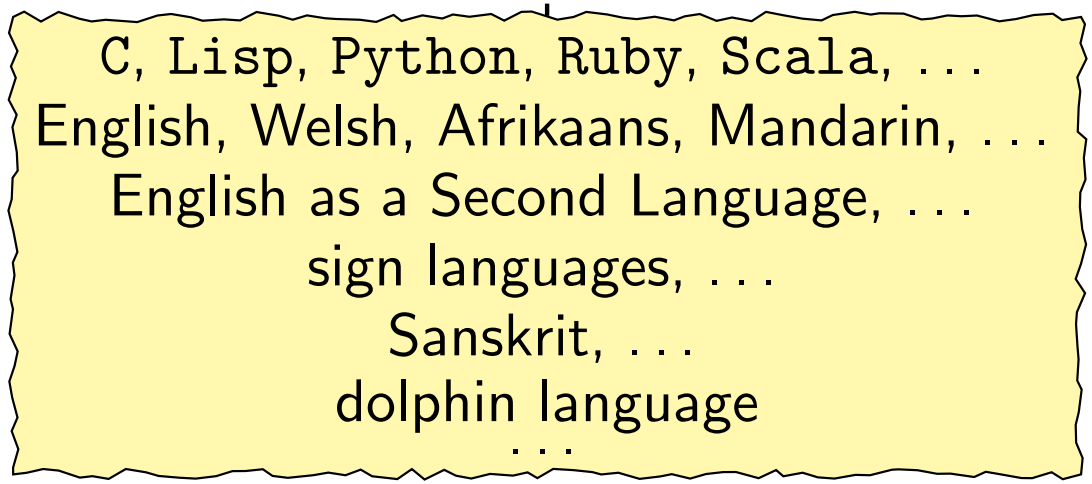
Your marks are based on three practicals:

- Assignment 1: 10%; submitted through Moodle by Thursday 31 October at 12:00
- Assignment 2: 60%; submitted through Moodle by Thursday 14 November at 12:00
- Assignment 3: 30%; submitted through Moodle by Thursday 5 December at 12:00

What is language?

PART-II
ACS L390

Overview of Natural Language Processing



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)

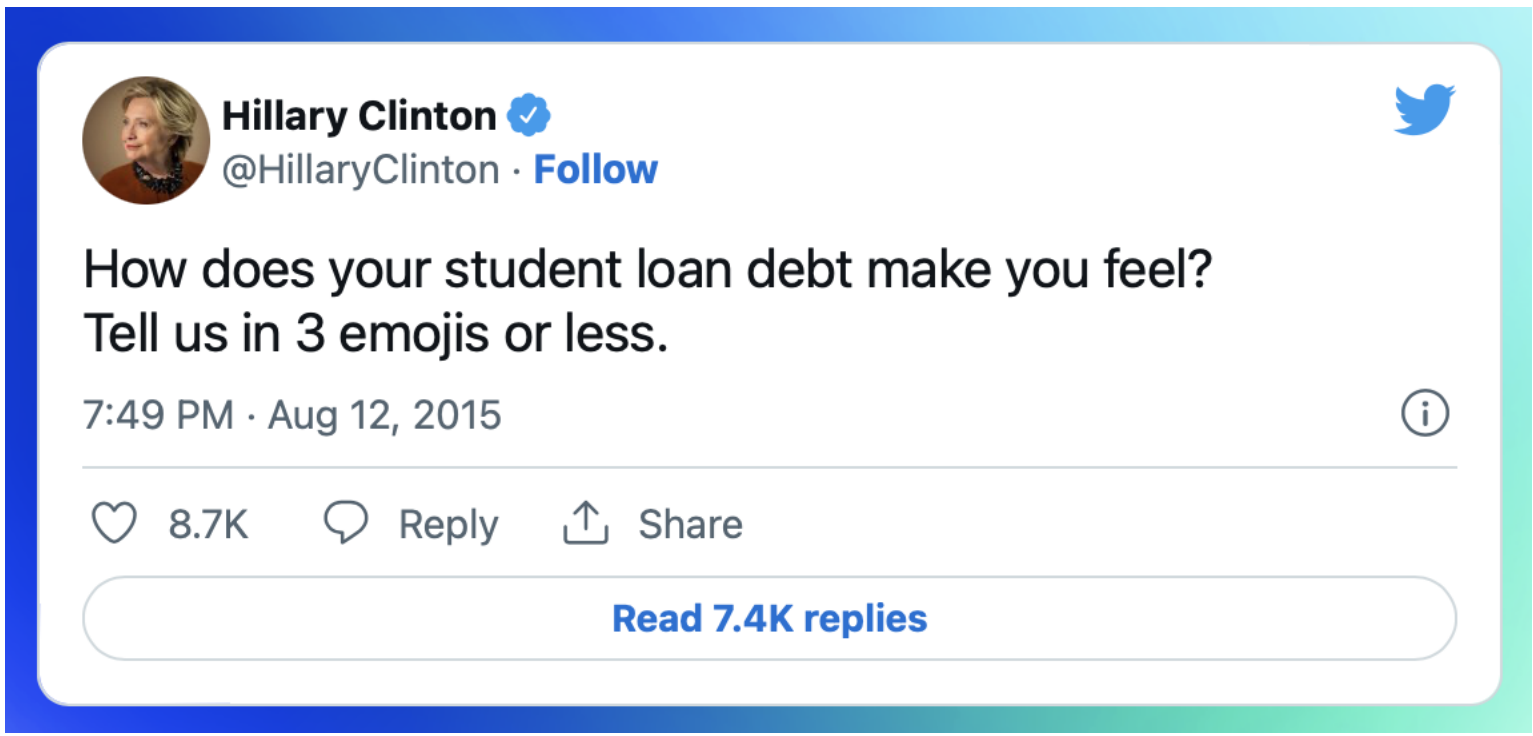
Are emojis part of your language?

What is language?

Obvious?

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

Are emojis part of your language?



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

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

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

for formal languages, we have a precise definition

Is it adequate to charaterise 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.

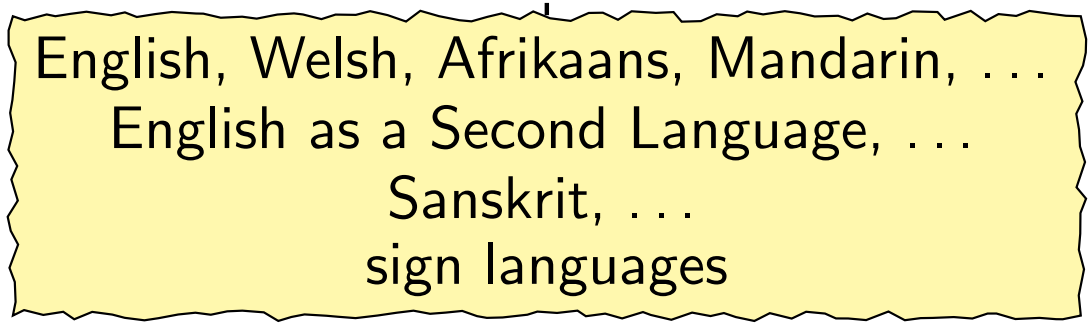
We will focus on scientific approaches to achieve reliable language technologies.

> *Seek simplicity and distrust it.*
>
> — *Alfred North Whitehead*

> *All models are wrong, but some models are useful.*
>
> — *George Box*

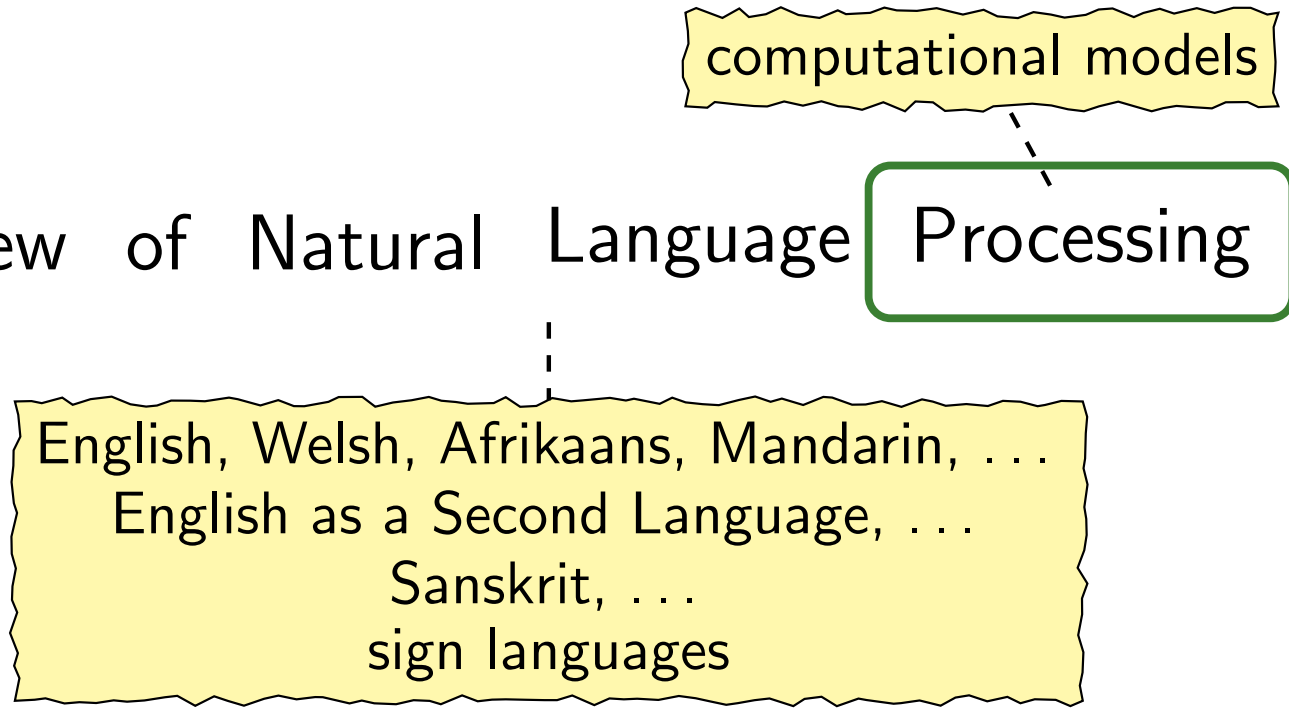
PART-II
ACS L390

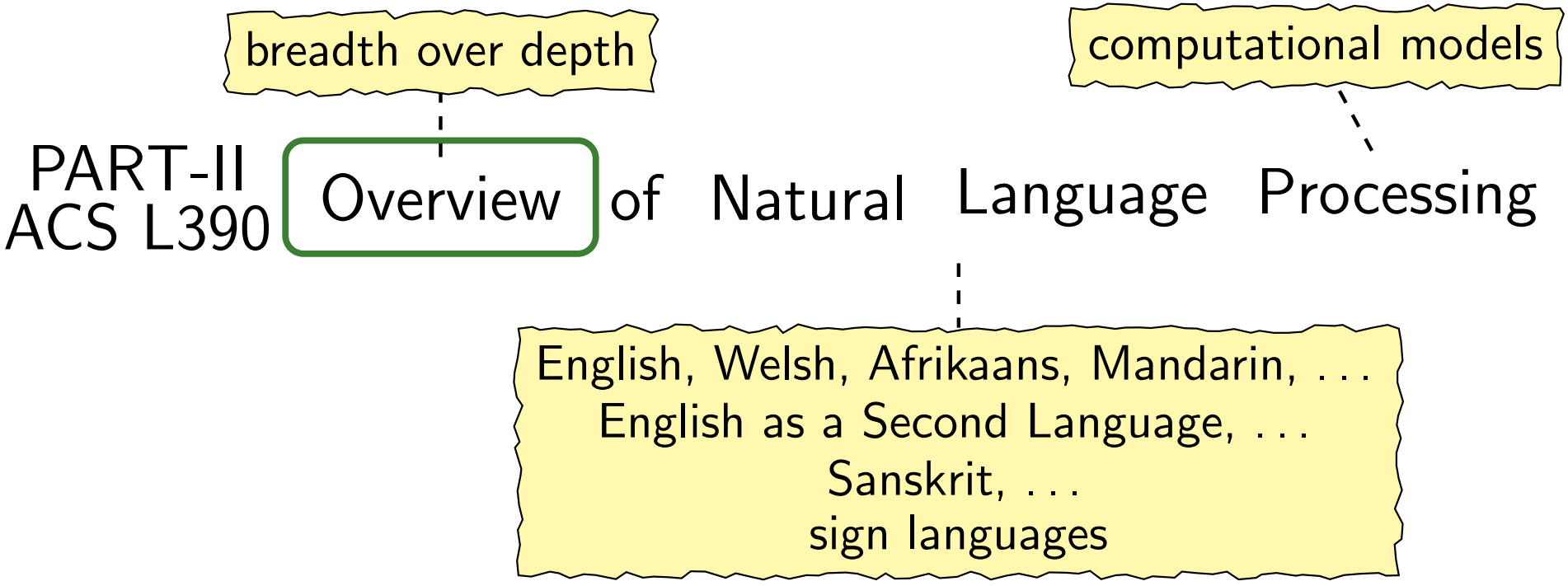
Overview of Natural Language Processing

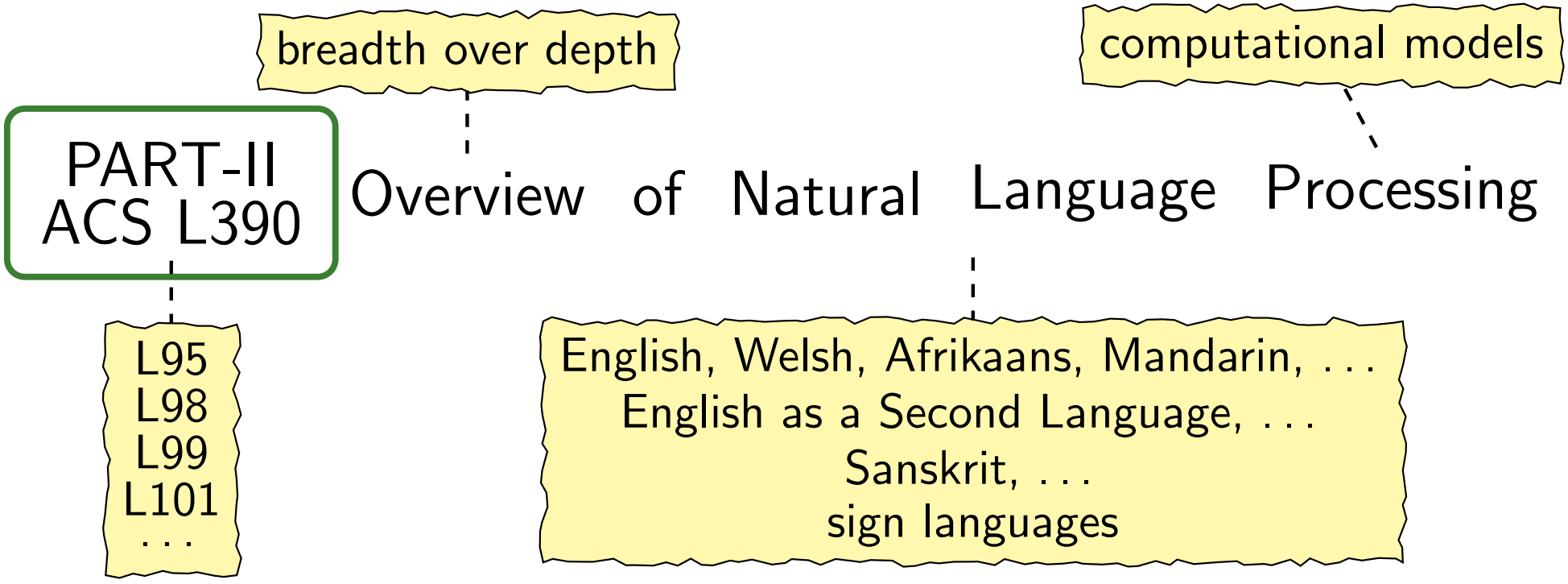


PART-II
ACS L390

Overview of Natural Language Processing




breadth over depth
computational models
PART-II
ACS L390
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 L390
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

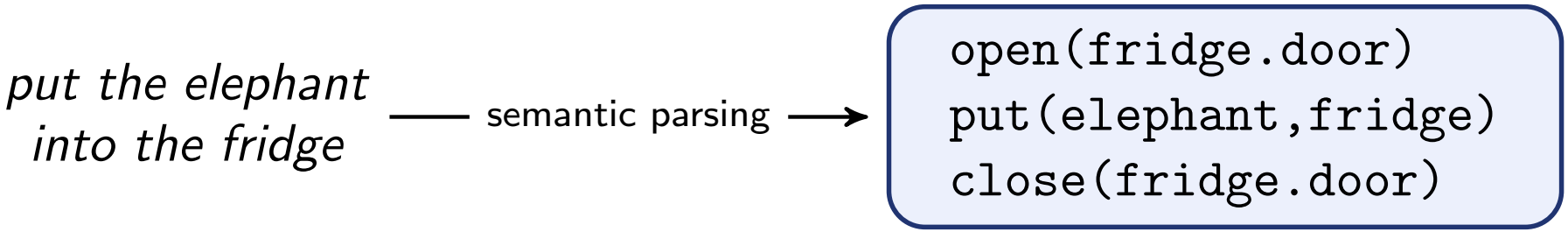
rude or offensive words

Conversational User Interface

How can siri put the elephant into the fridge?

Conversational User Interface

How can siri put the elephant into the fridge?



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

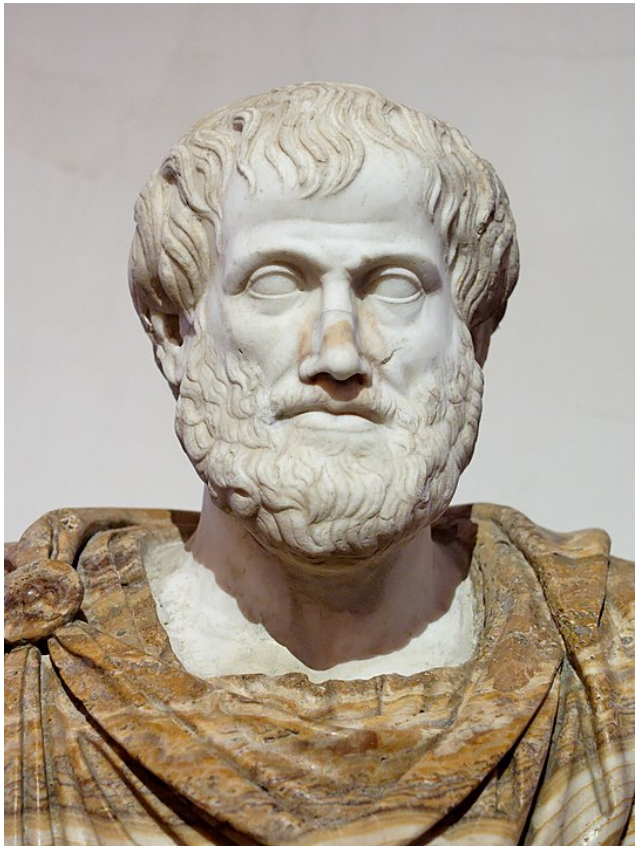
Dialogue System

Example

A Could you please close the door from the outside?

B [...]

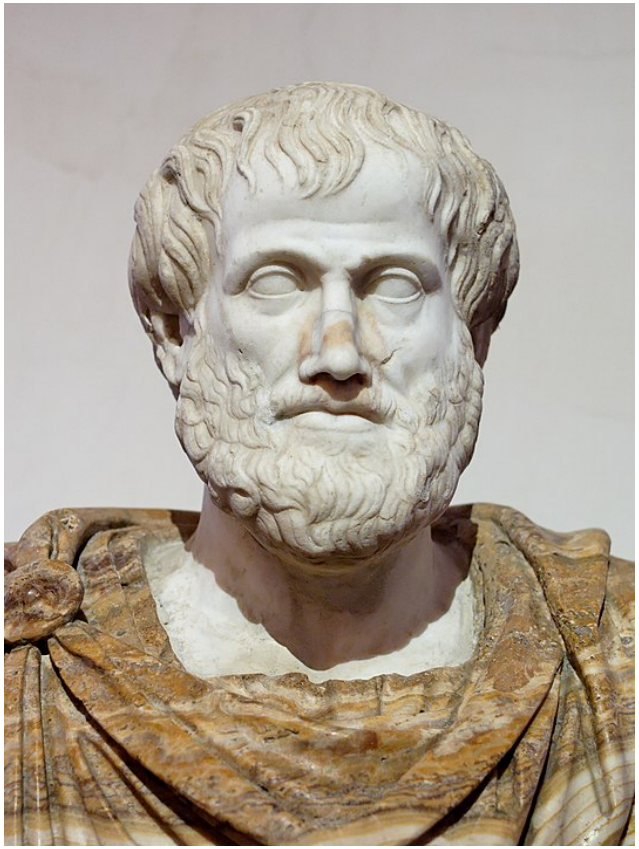
Aristotle's Syllogism



Syllogism=Syn- + logos

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

Aristotle's Syllogism

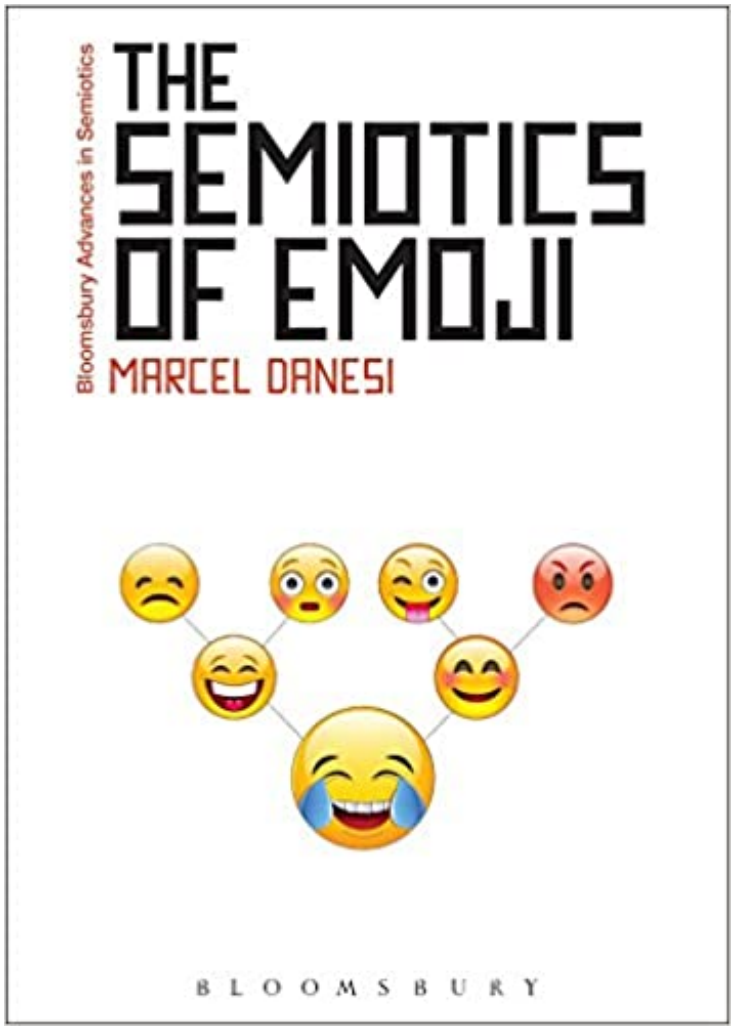


Syllogism=Syn- + logos

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

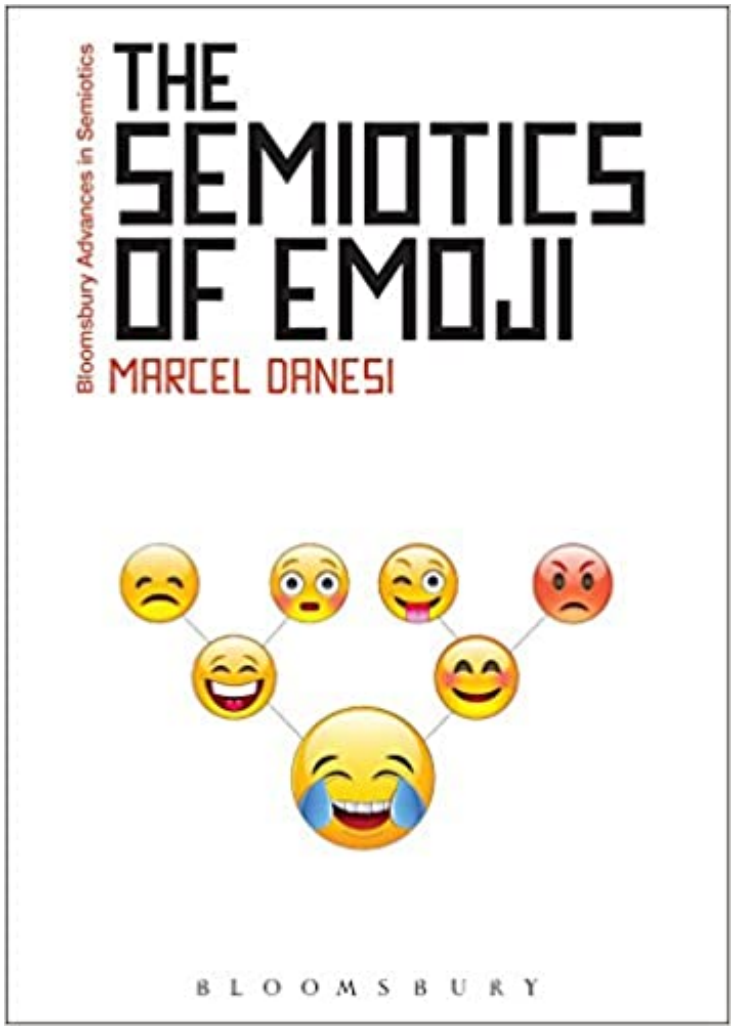
language goes beyond communication

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 2024 aims to have a broad technical program. Relevant topics for the conference include, but are not limited to, the following areas (in alphabetical order):

- Computational Social Science and Cultural Analytics
- Dialogue and Interactive Systems
- Discourse and Pragmatics
- Efficient/Low-Resource Methods for NLP
- Ethics, Bias, and Fairness
- Generation
- Information Extraction
- Information Retrieval and Text Mining
- Interpretability and Analysis of Models for NLP
- Linguistic theories, Cognitive Modeling and Psycholinguistics
- Machine Learning for NLP
- Machine Translation

A call-for-paper (2)

ACL 2024 aims to have a broad technical program. Relevant topics for the conference include, but are not limited to, the following areas (in alphabetical order):

- Multilinguality and Language Diversity
- Multimodality and Language Grounding to Vision, Robotics and Beyond
- NLP Applications
- Phonology, Morphology and Word Segmentation
- Question Answering
- Resources and Evaluation
- Semantics: Lexical
- Semantics: Sentence-level Semantics, Textual Inference and Other areas
- Sentiment Analysis, Stylistic Analysis, and Argument Mining
- Speech recognition, text-to-speech and spoken language understanding
- Summarization
- Syntax: Tagging, Chunking and Parsing

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.

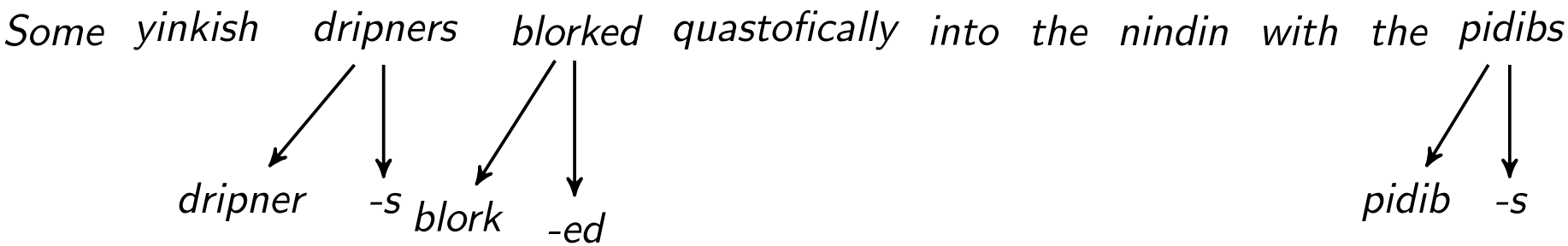
the example is partly from A Carnie's *Syntax: A Generative Introduction*

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

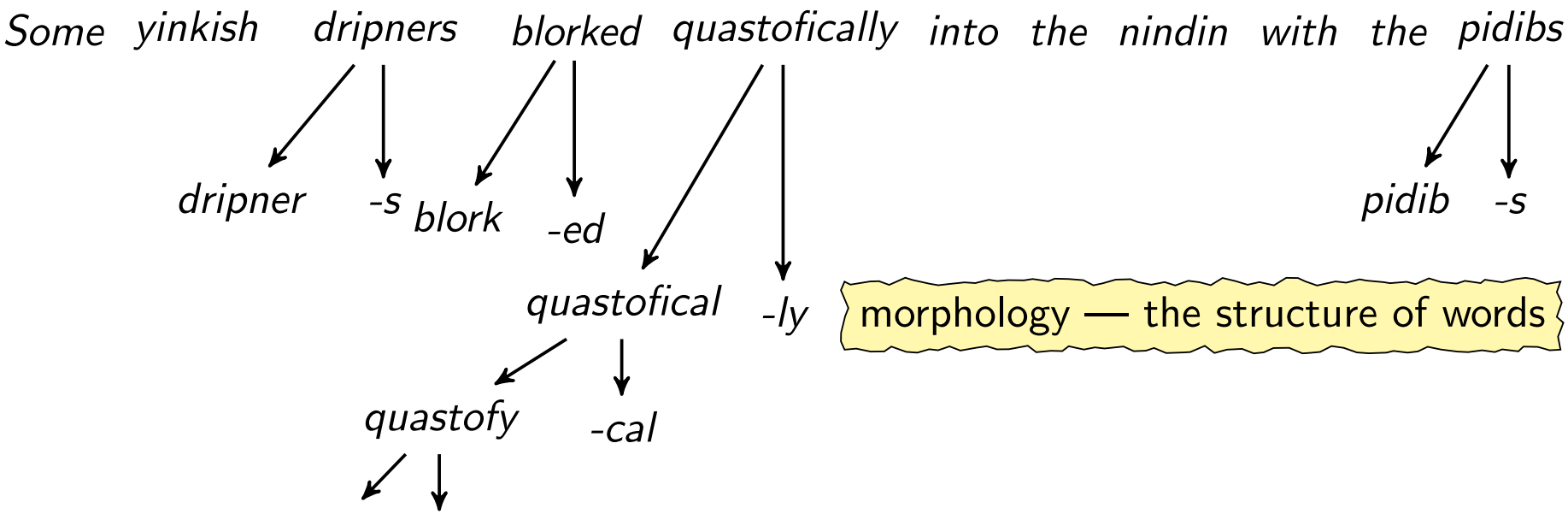
Structuring a sentence

Some yinkish dripners blorked quastofically into the nindin with the pidibs

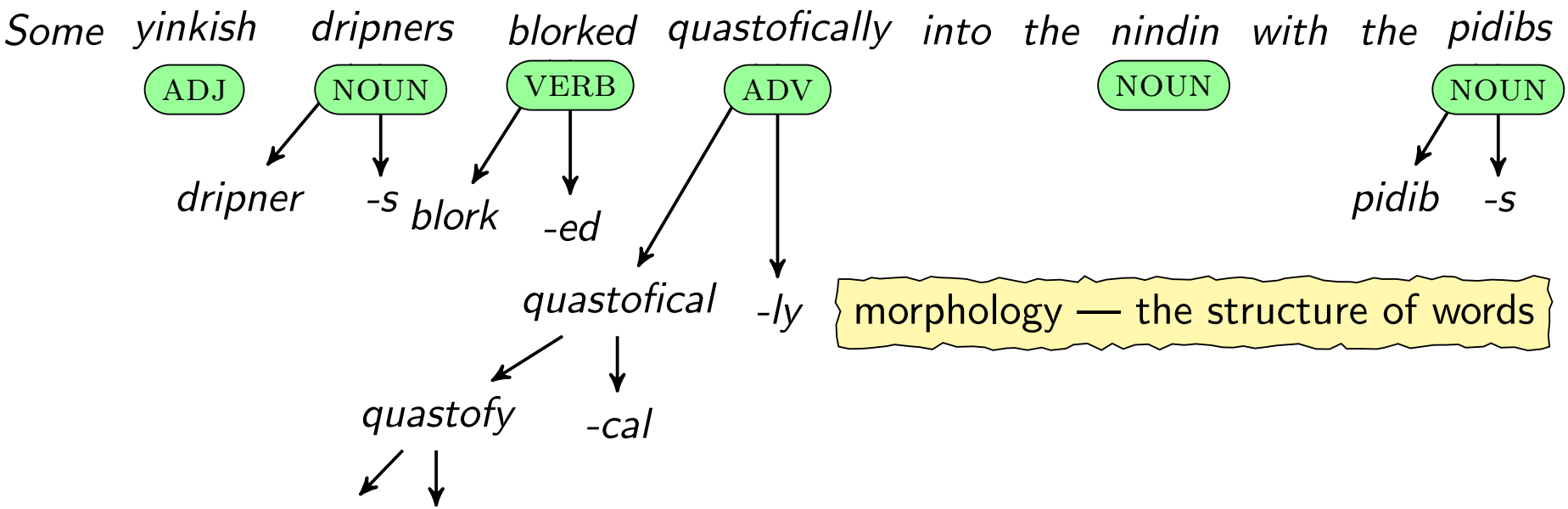
Structuring a sentence



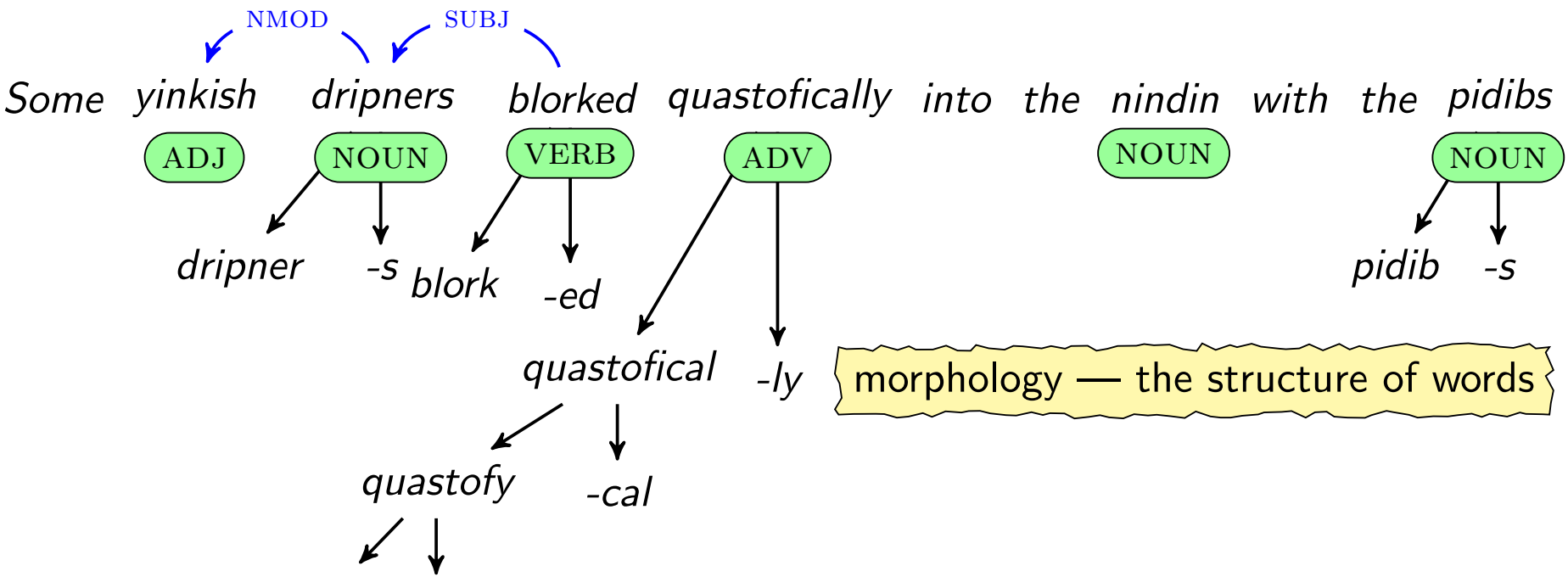
Structuring a sentence



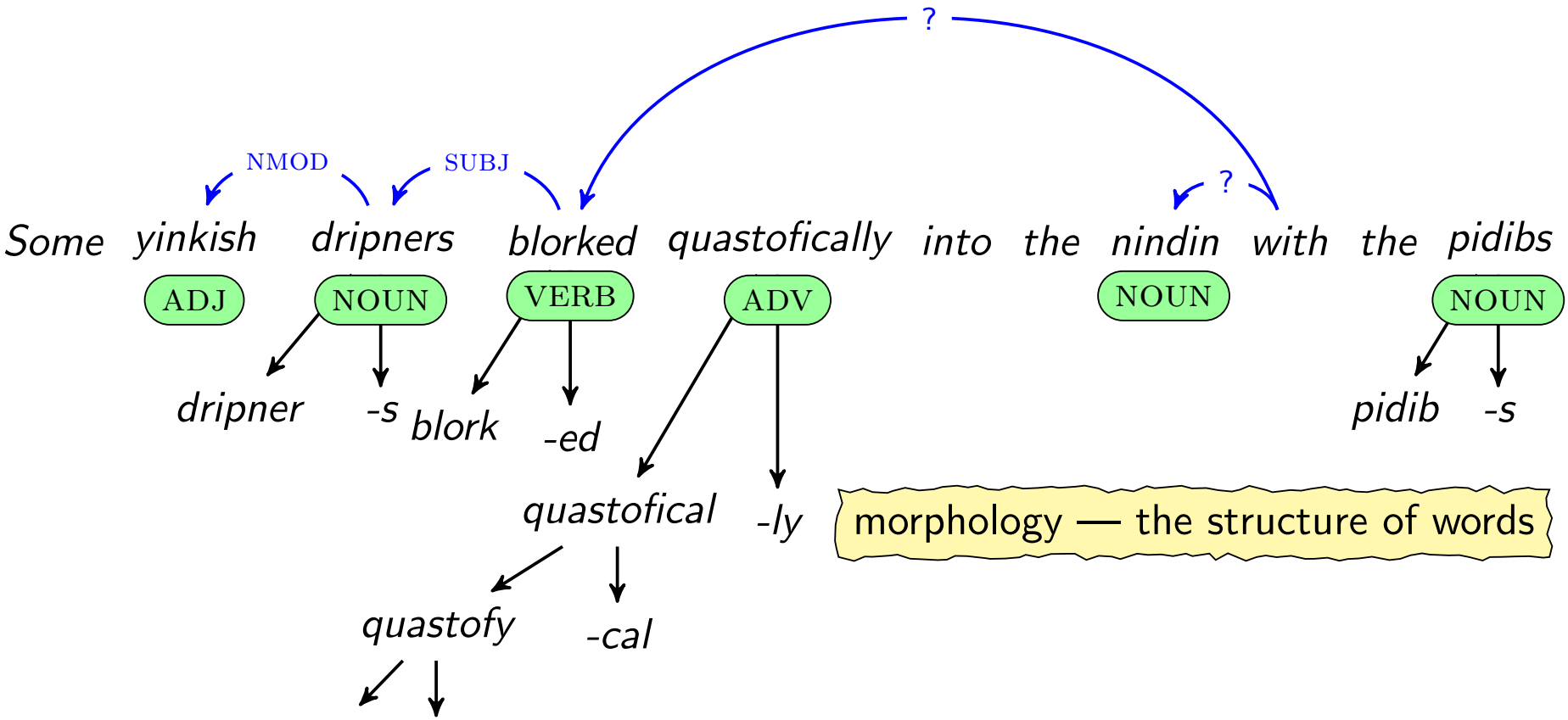
Structuring a sentence



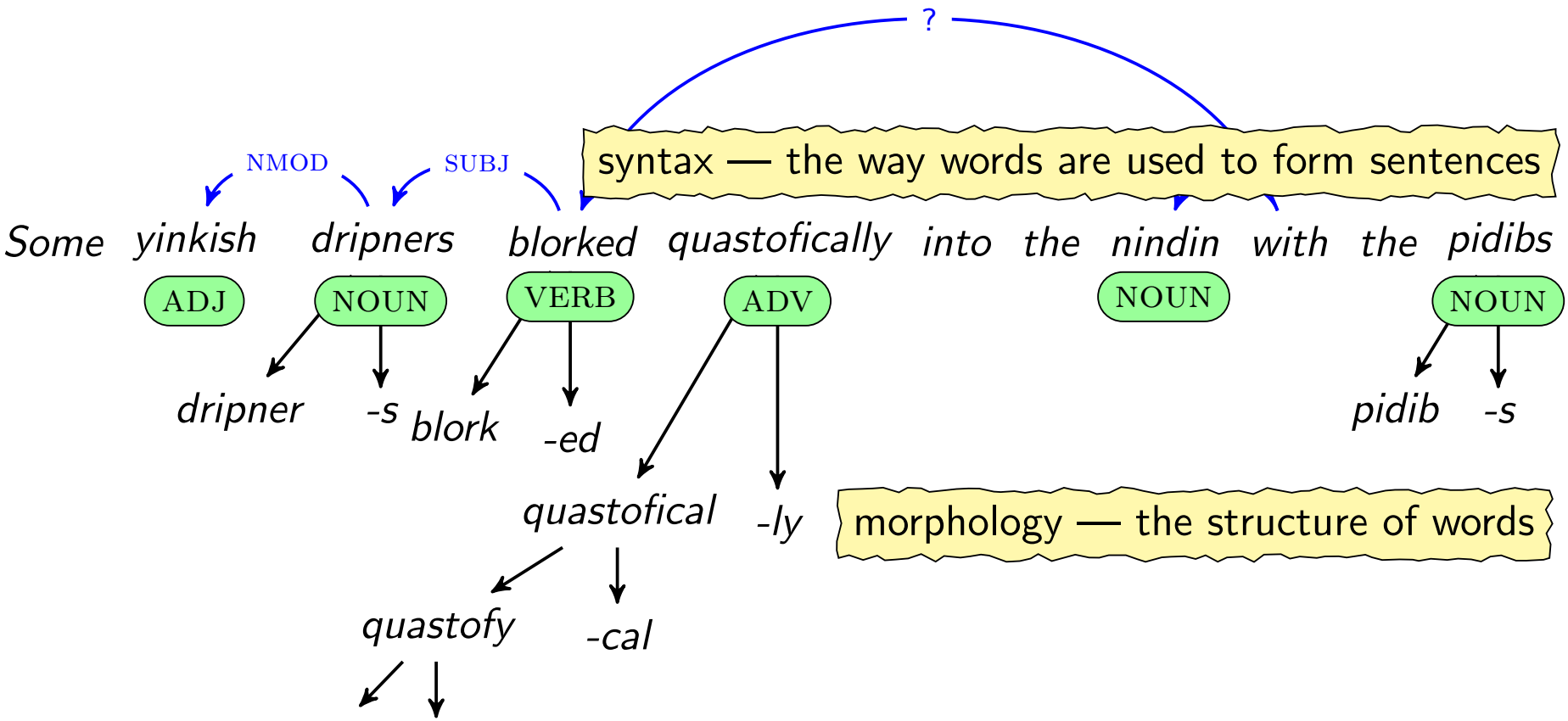
Structuring a sentence



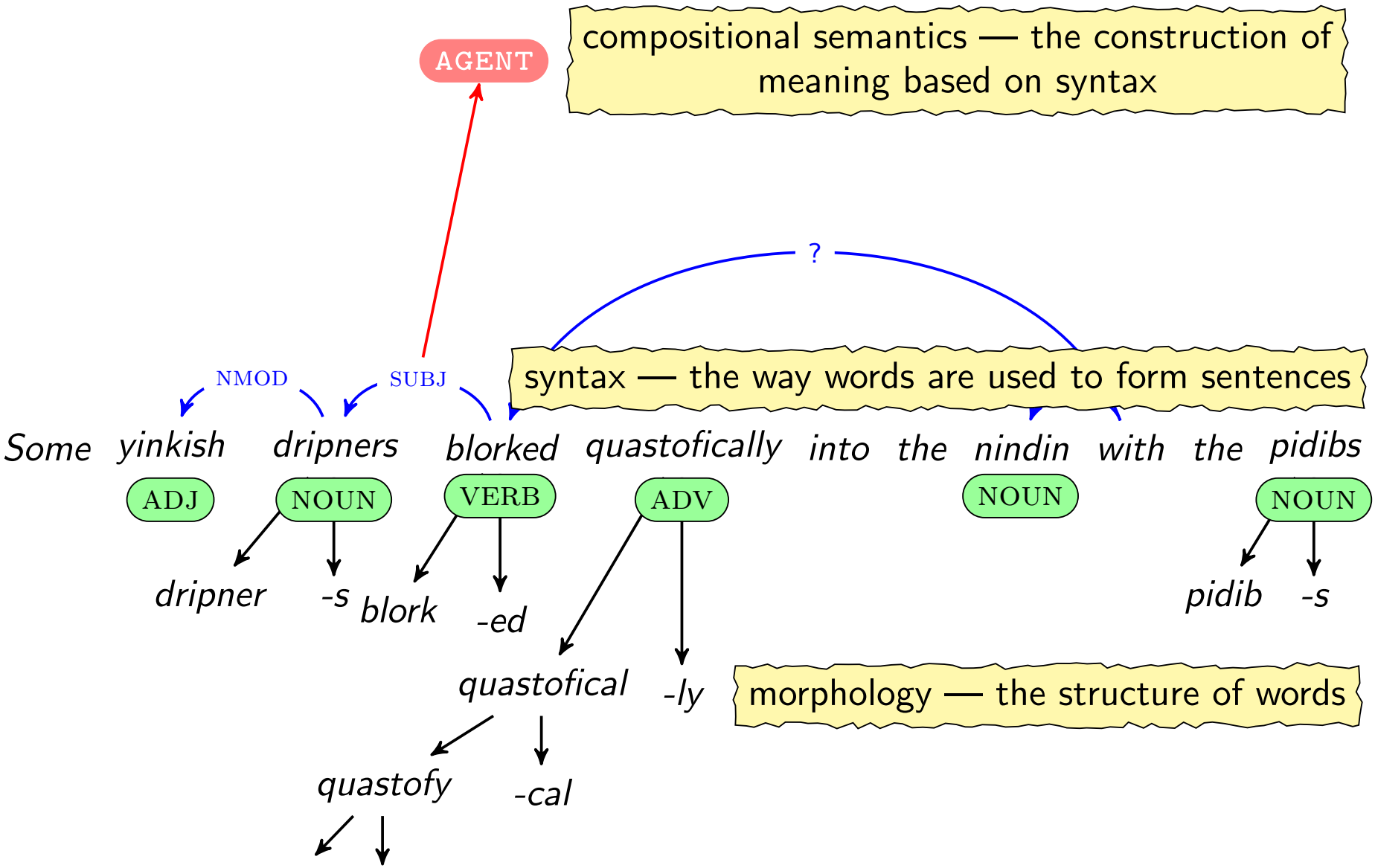
Structuring a sentence



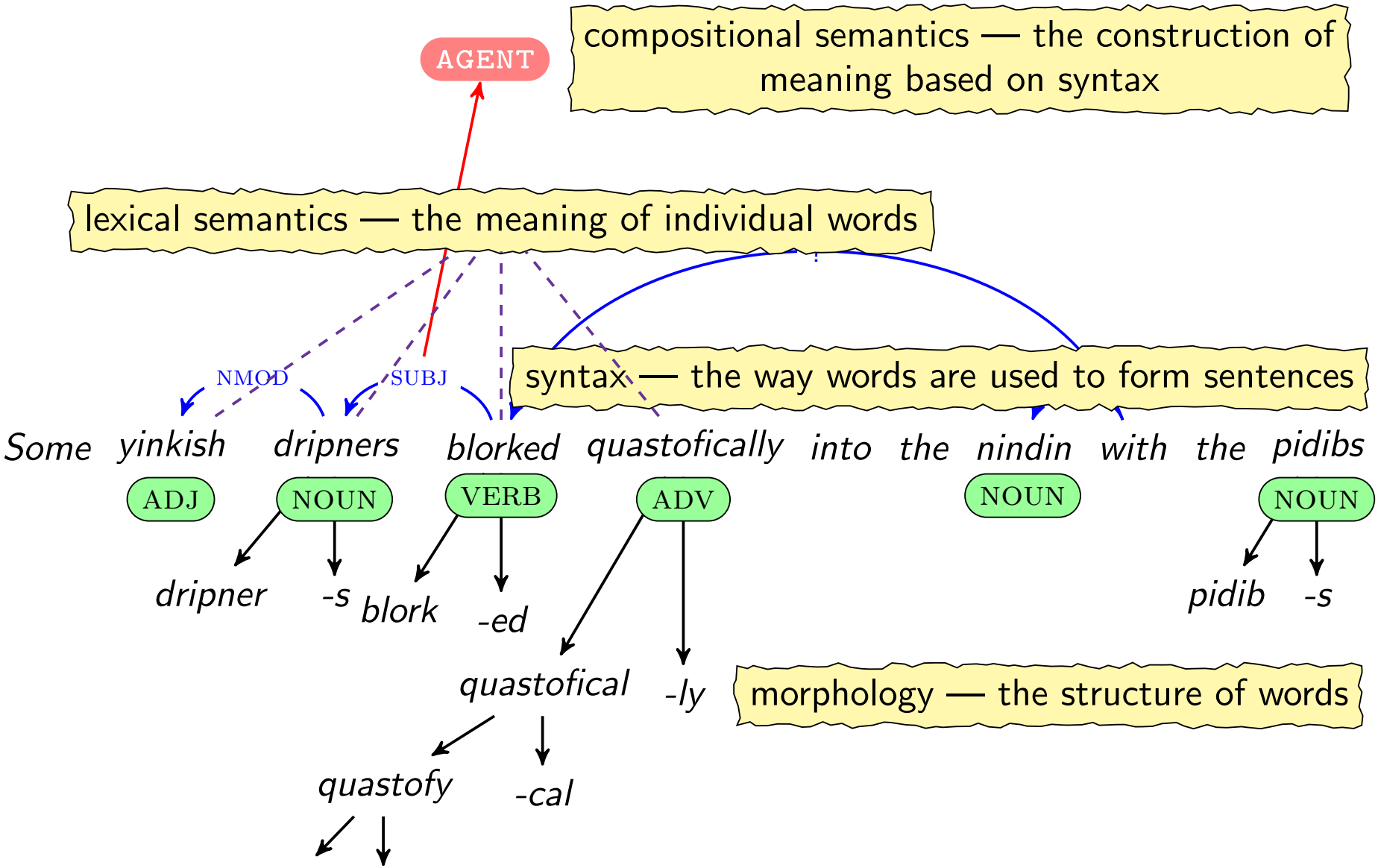
Structuring a sentence



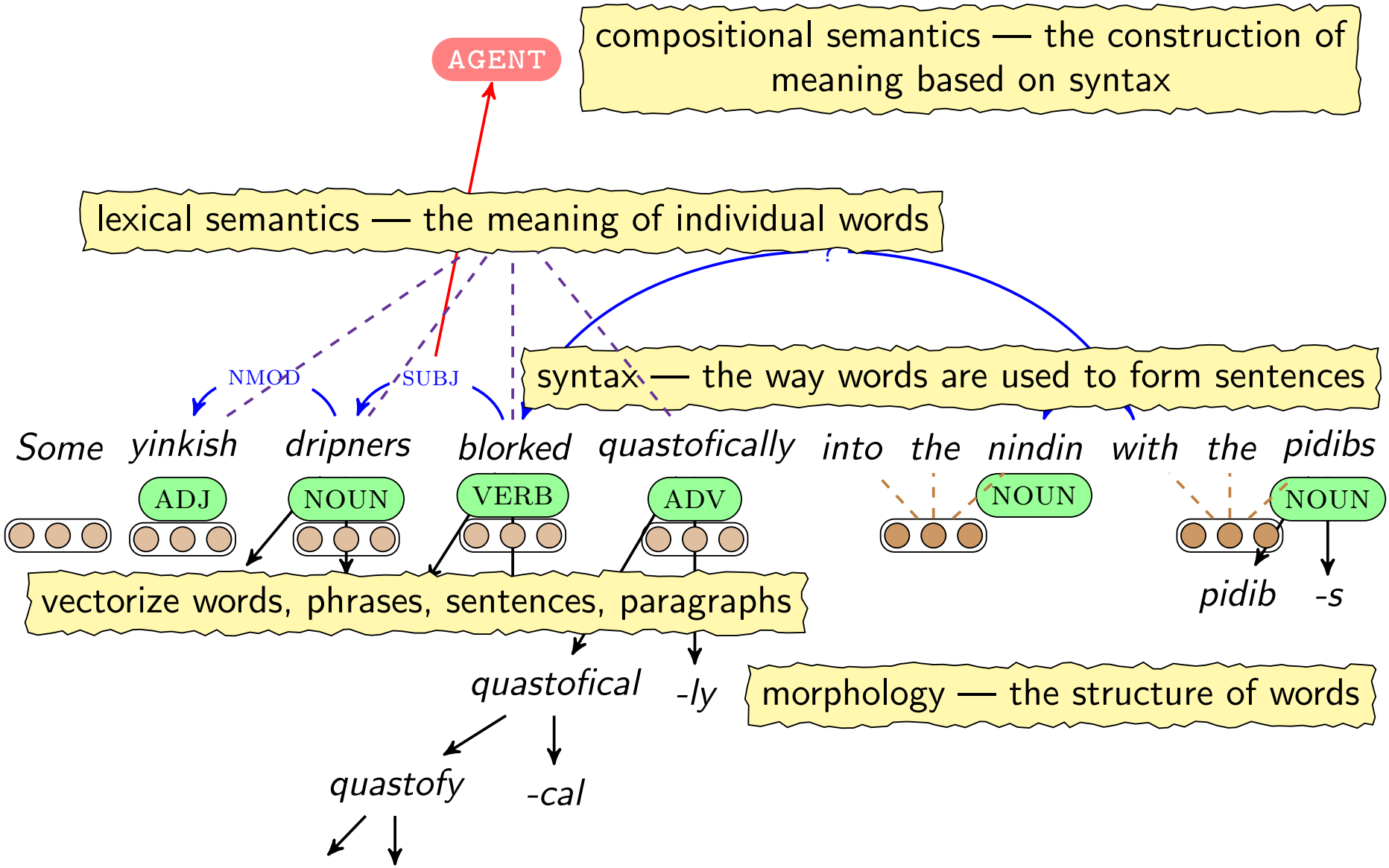
Structuring a sentence



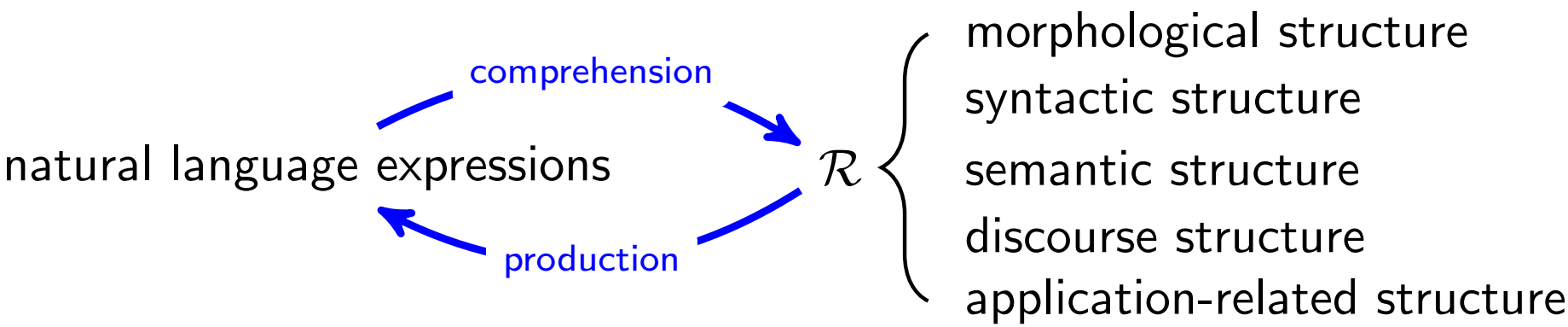
Structuring a sentence



Structuring a sentence



Form transformation

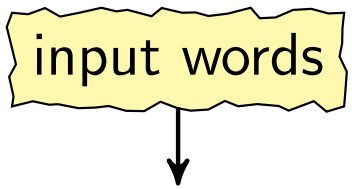


CoNLL shared tasks

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

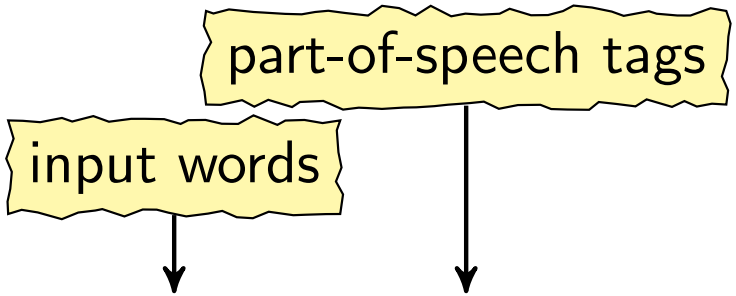
2019	Cross-Framework Meaning Representation Parsing
2018/2017	Multilingual Parsing from Raw Text to Universal Dependencies
2018/2017	Universal Morphological Reinflection
2016/2016	(Multilingual) Shallow Discourse Parsing
2014/2013	Grammatical Error Correction
2012/2011	Modelling (Multilingual) Unrestricted 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

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



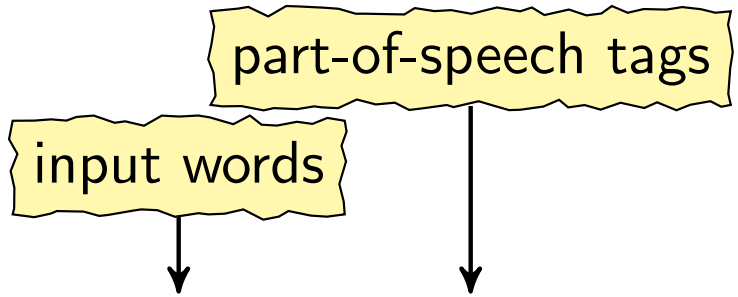

```
1       U.N.
2     official
3      Ekeus
4      heads
5       for
6     Baghdad
7        .
```

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

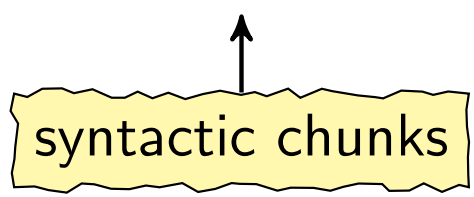


1	U.N.	NNP
2	official	NN
3	Ekeus	NNP
4	heads	VBZ
5	for	IN
6	Baghdad	NNP
7	.	.

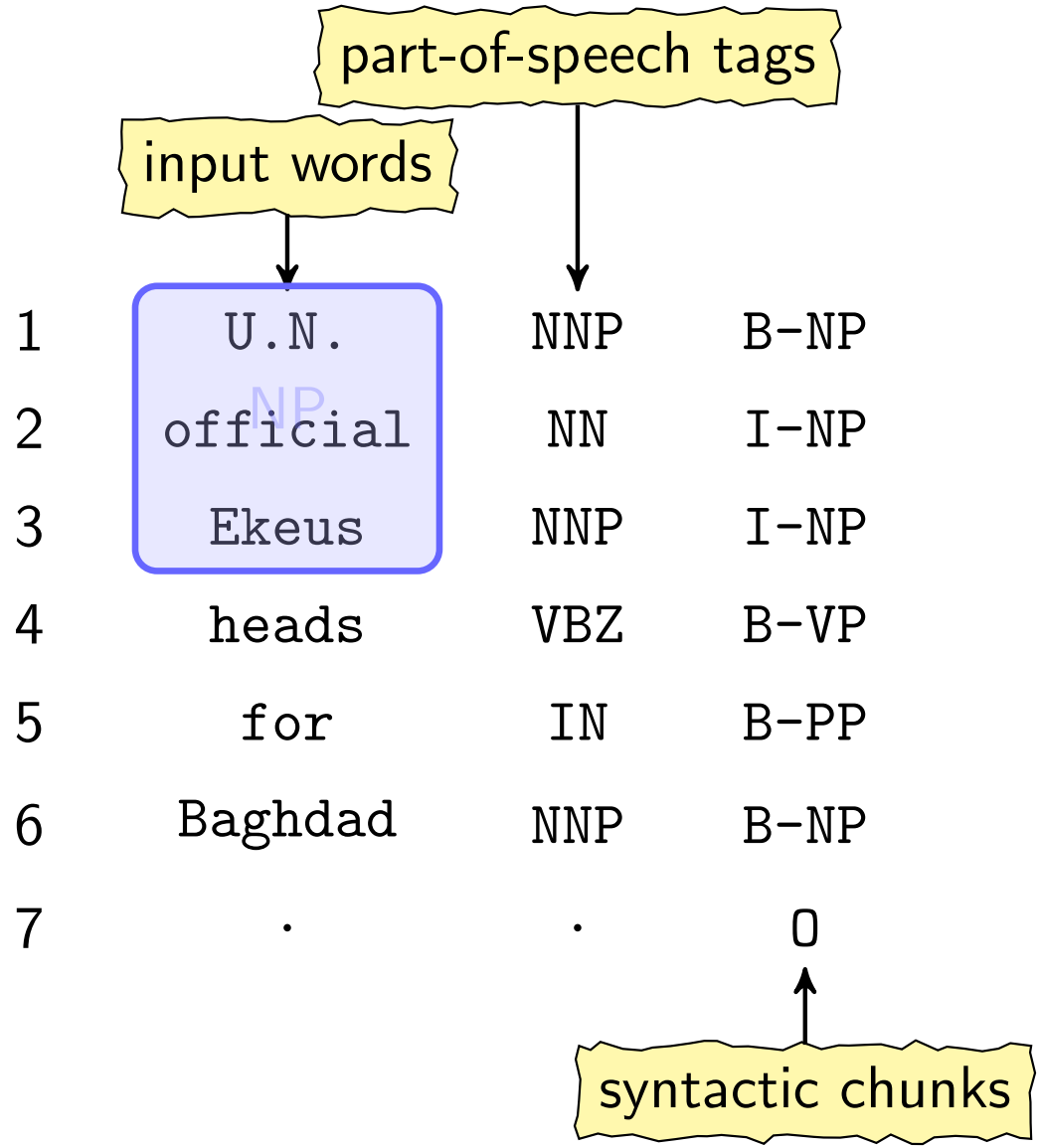
CoNLL ST 1999/2000/2002/2003/2006/2007/2017/2018



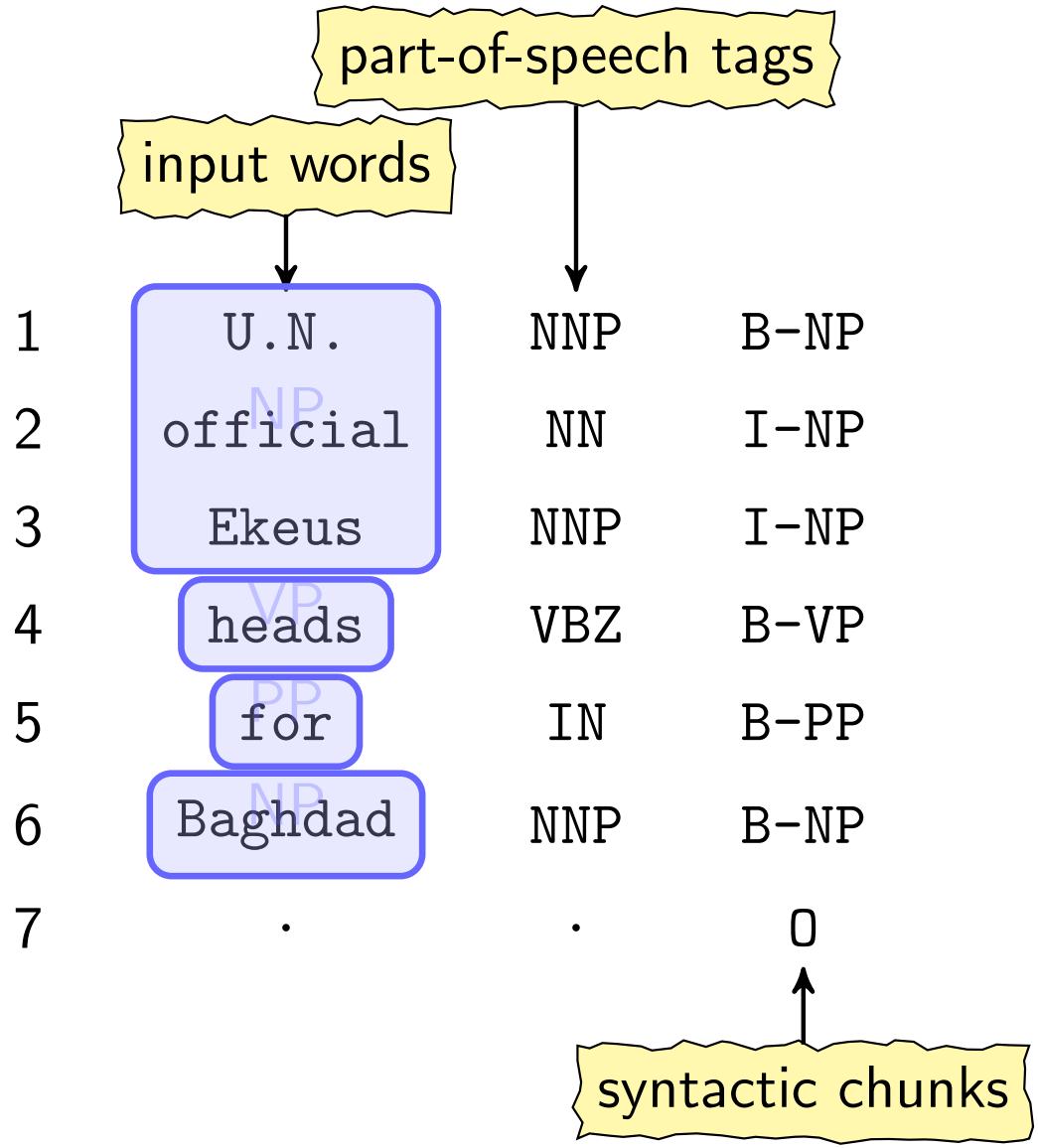
1	U.N.	NNP	B-NP
2	official	NN	I-NP
3	Ekeus	NNP	I-NP
4	heads	VBZ	B-VP
5	for	IN	B-PP
6	Baghdad	NNP	B-NP
7	.	.	O



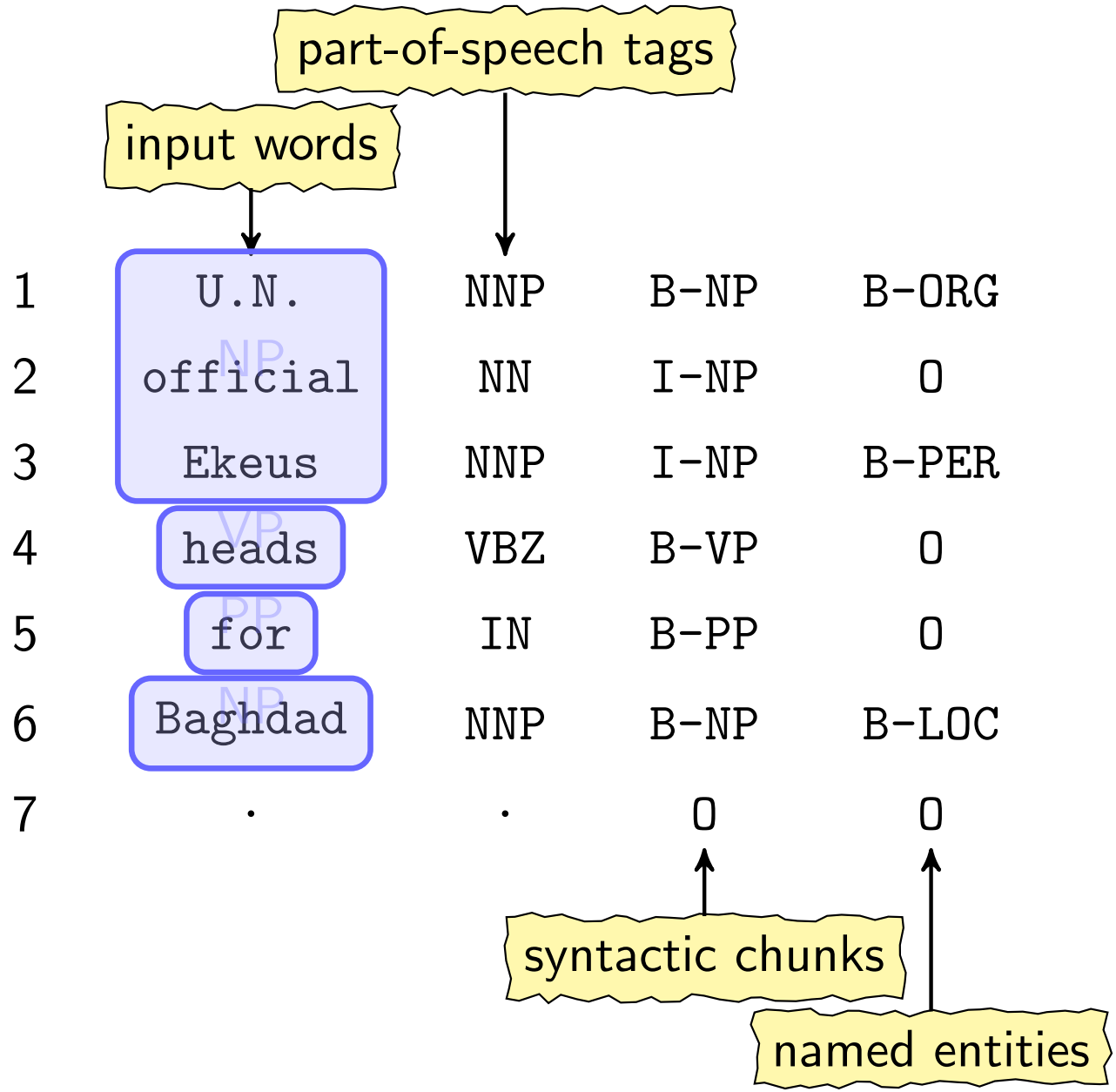
CoNLL ST 1999/2000/2002/2003/2006/2007/2017/2018



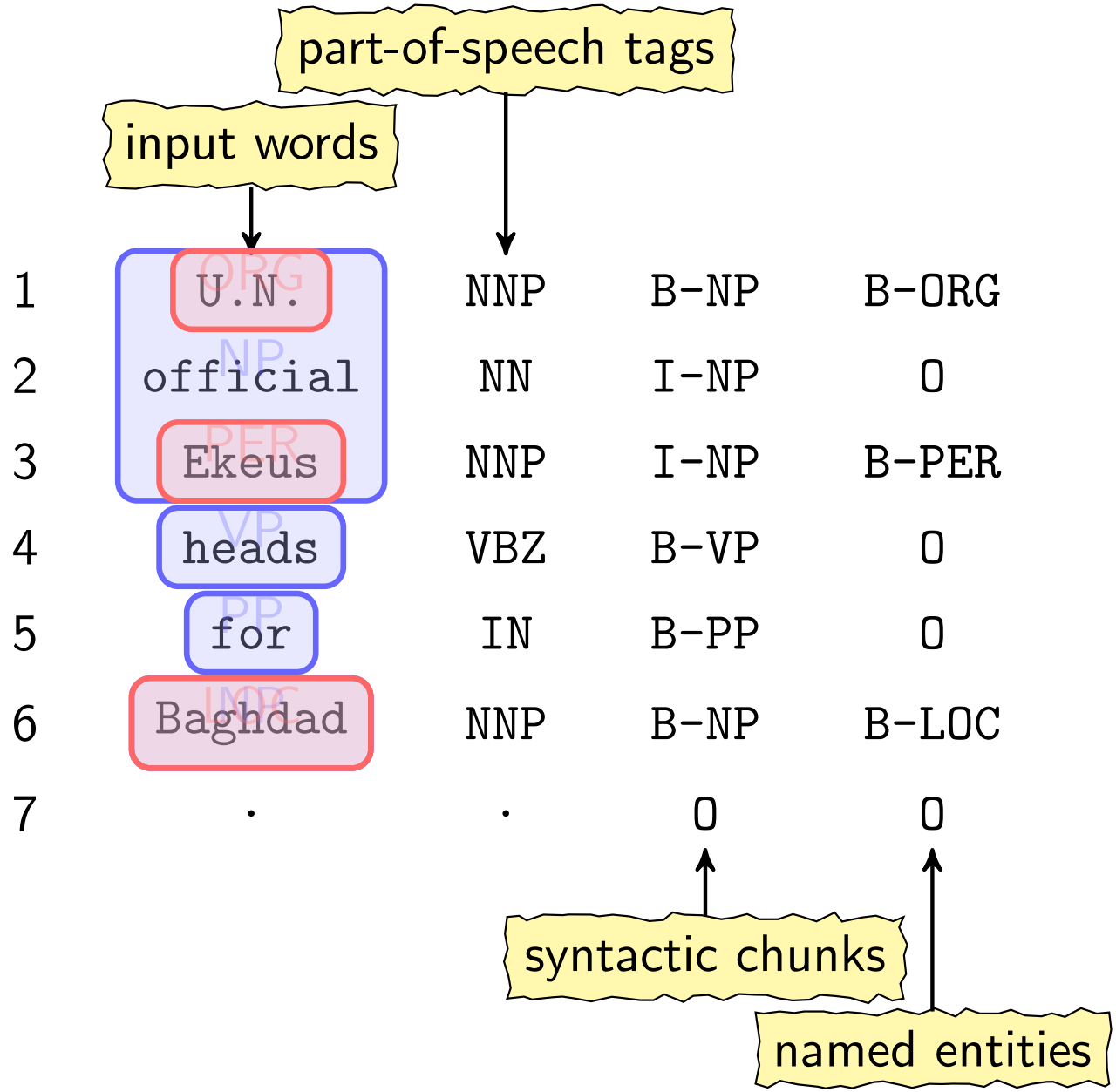
CoNLL ST 1999/2000/2002/2003/2006/2007/2017/2018



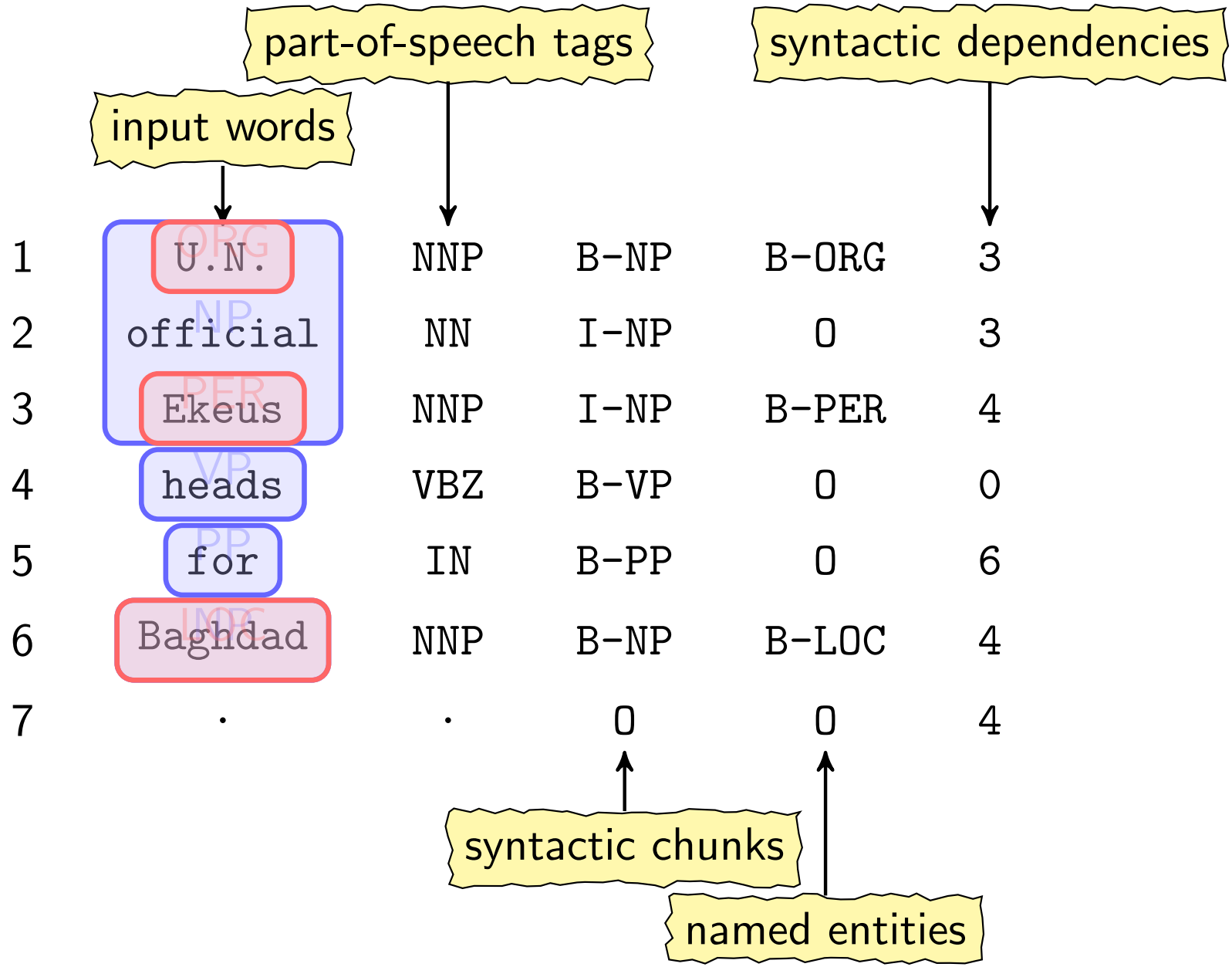
CoNLL ST 1999/2000/2002/2003/2006/2007/2017/2018



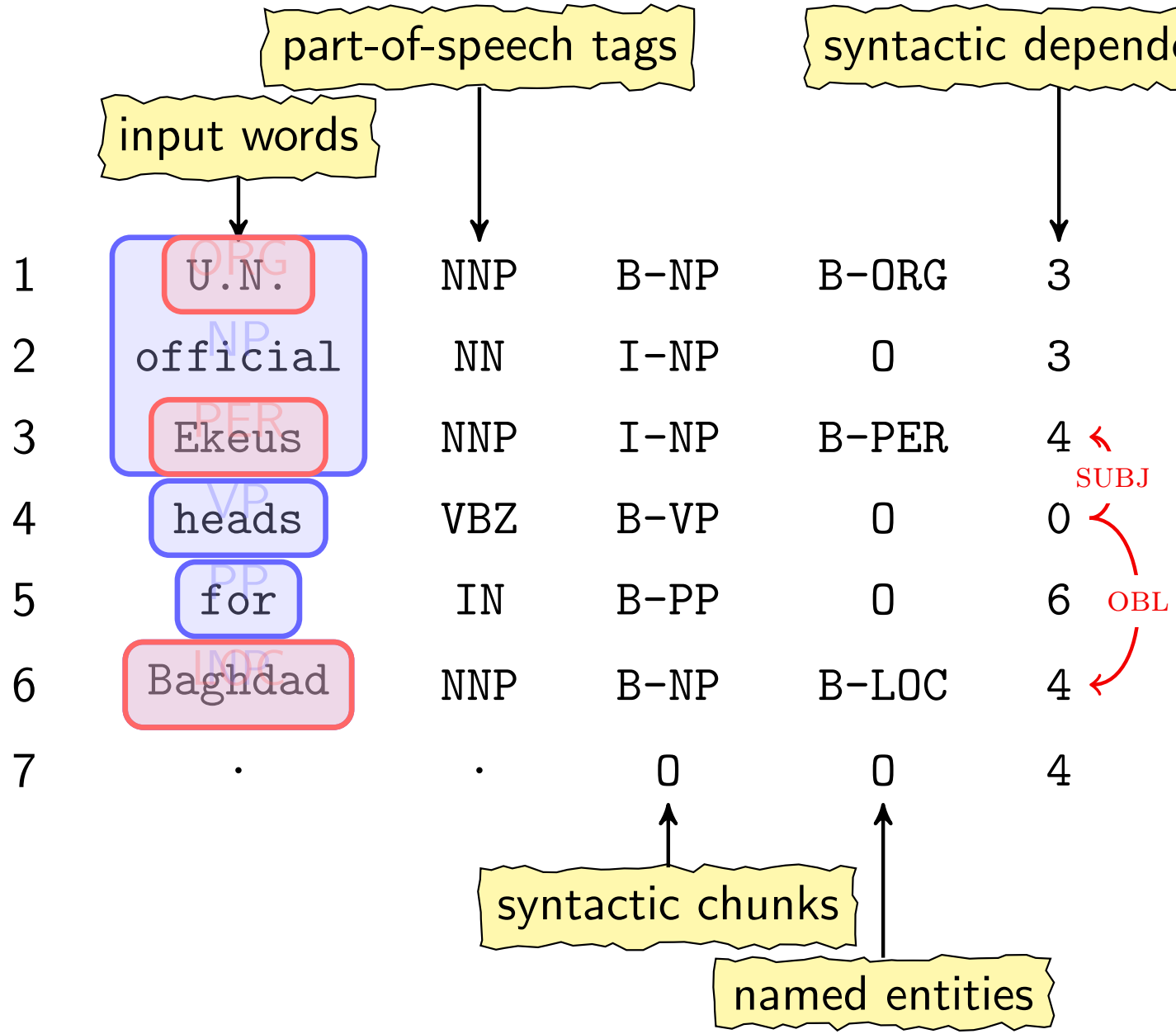
CoNLL ST 1999/2000/2002/2003/2006/2007/2017/2018



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 5 and 6.
- *Semantics* — the meaning of words and sentences: lecture 11.

- *Symbolic models* — finite-state machines, context-free grammars, logic: lectures 2, 5 and 11.
- *Statistical models* — (structured) prediction: lectures 3, 5 and 7.
- *Neural models* — representation learning: lectures 4, 9 and 10.

- *Language generation* — lecture 10 and 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/