Lecture 1: Overview of Natural Language Processing

Machine Learning and Natural Language Processing

Weiwei Sun

Department of Computer Science and Technology
University of Cambridge

Summer 2023

What is 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)

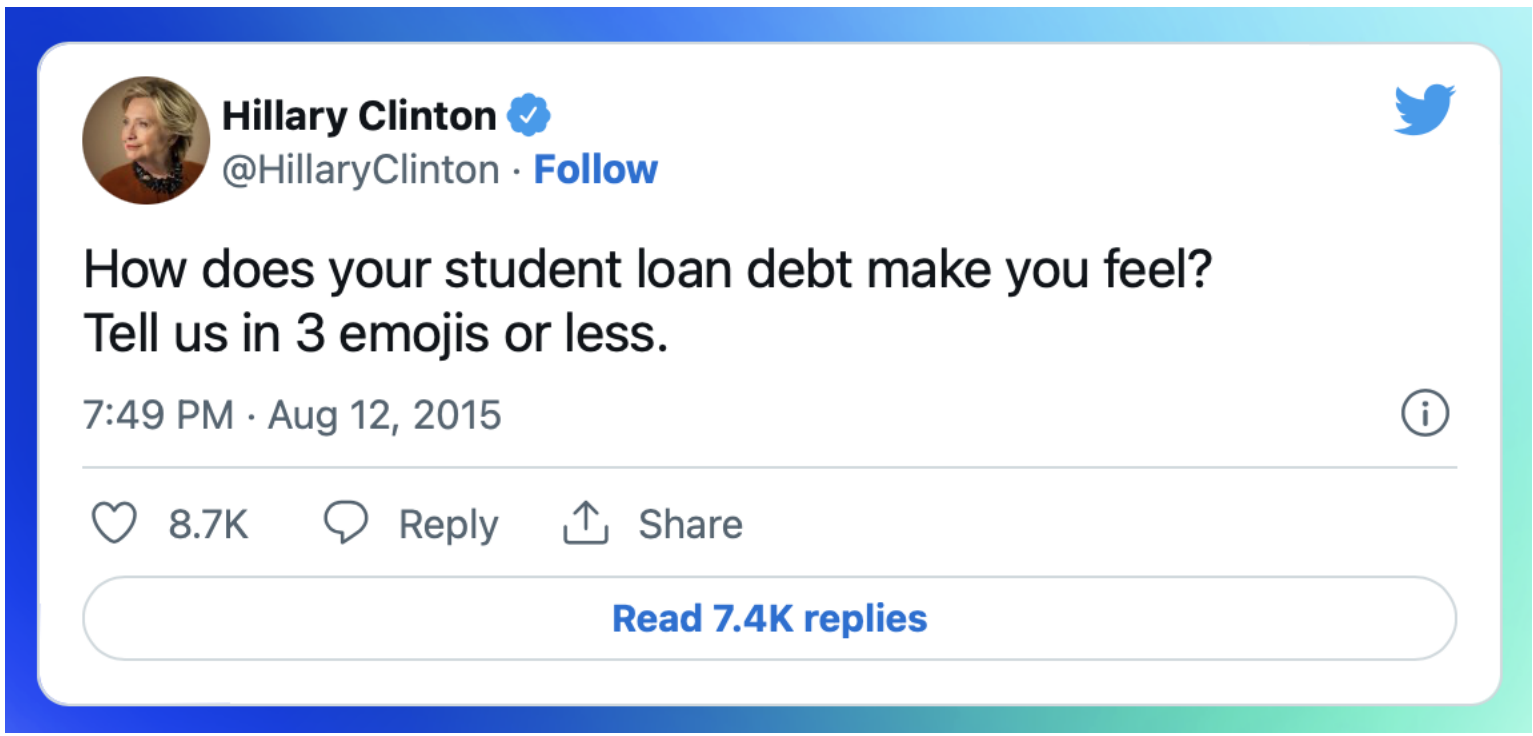
Are emojis part of your language?

What is language?

Obvious?

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

Are emojis part of your language?



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.

Example

$$L = \{ab, aabb, aaabbb, \ldots\}$$

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.

Example

$$L = \{ab, aabb, aaabbb, \ldots\}$$

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.

Example

$$L = \{ab, aabb, aaabbb, \ldots\}$$

for formal languages, we have a precise definition

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

Goal and Scope

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

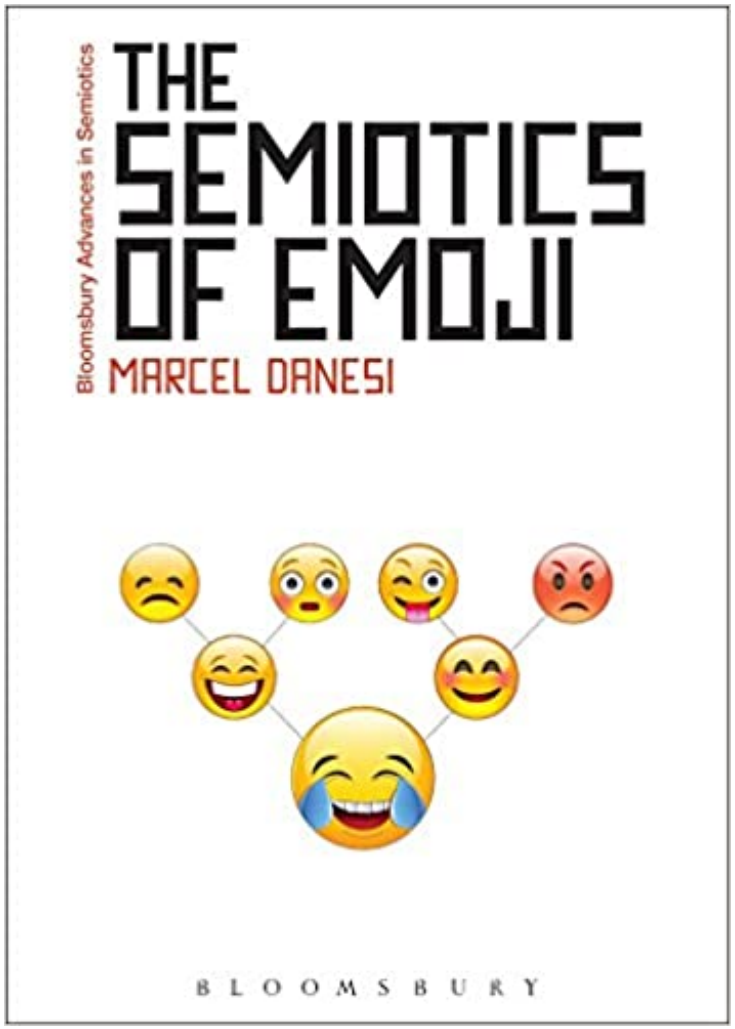
Dialogue System

Example

A Could you please close the door from outside?

B [...]

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 various 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.

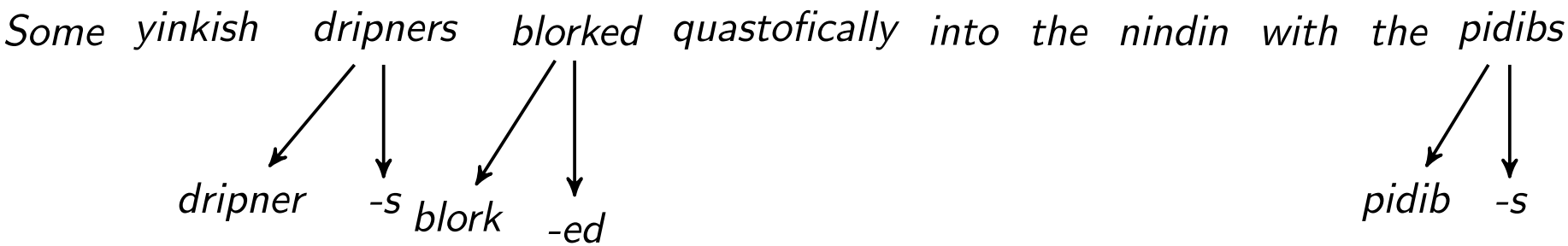
the example is partly from A Carnie's *Syntax: A Generative Introduction*

- there was a 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;

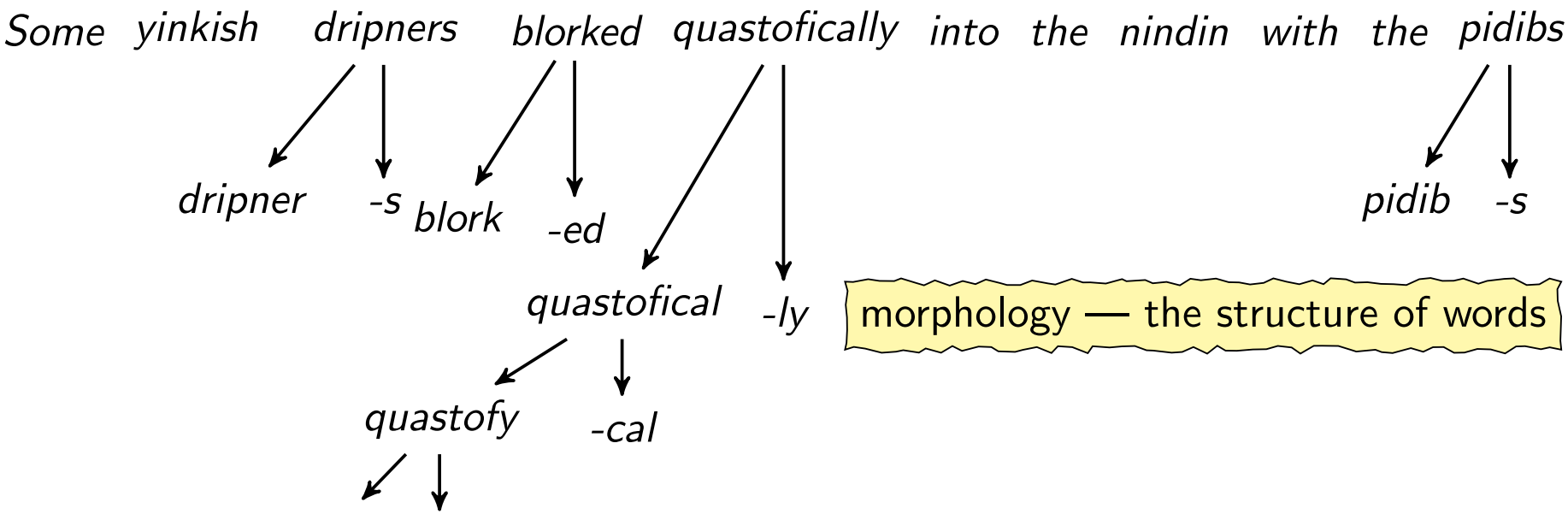
Structuring a sentence

Some yinkish dripners blorked quastofically into the nindin with the pidibs

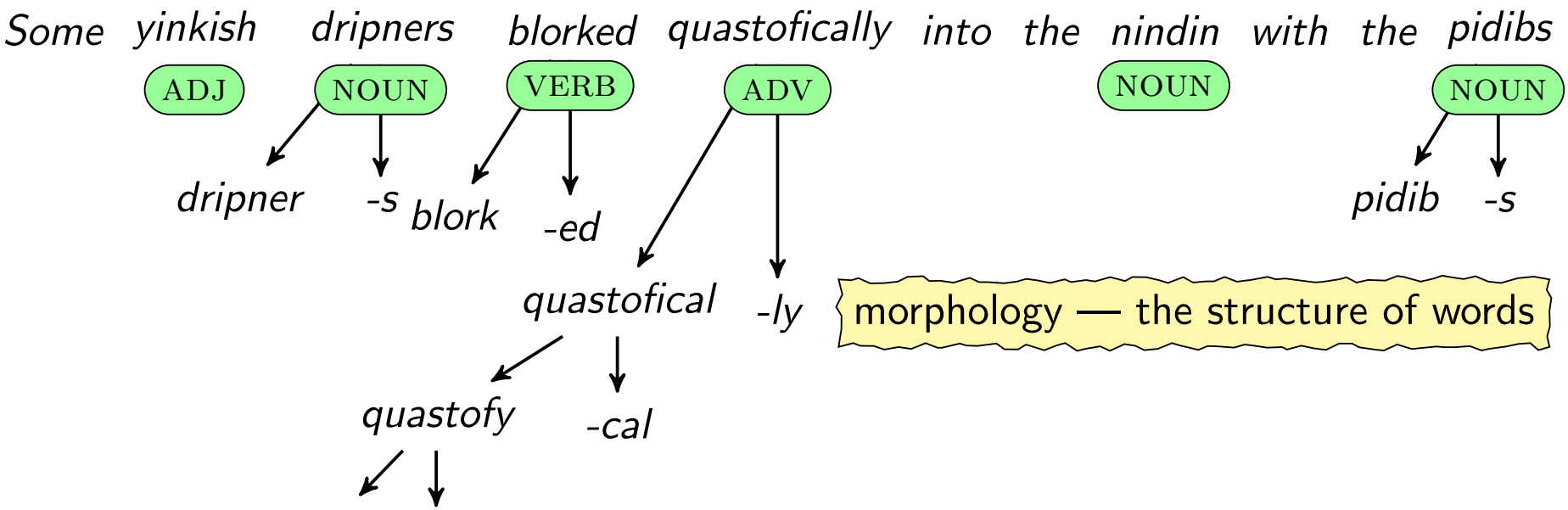
Structuring a sentence



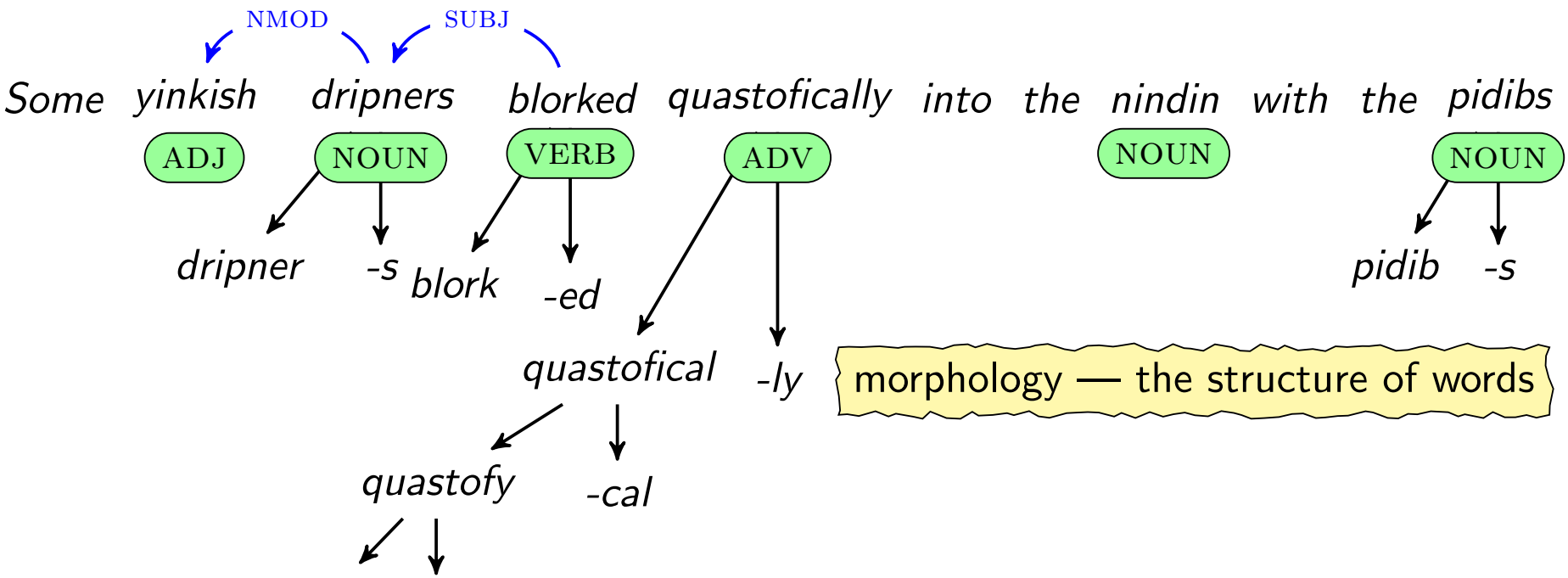
Structuring a sentence



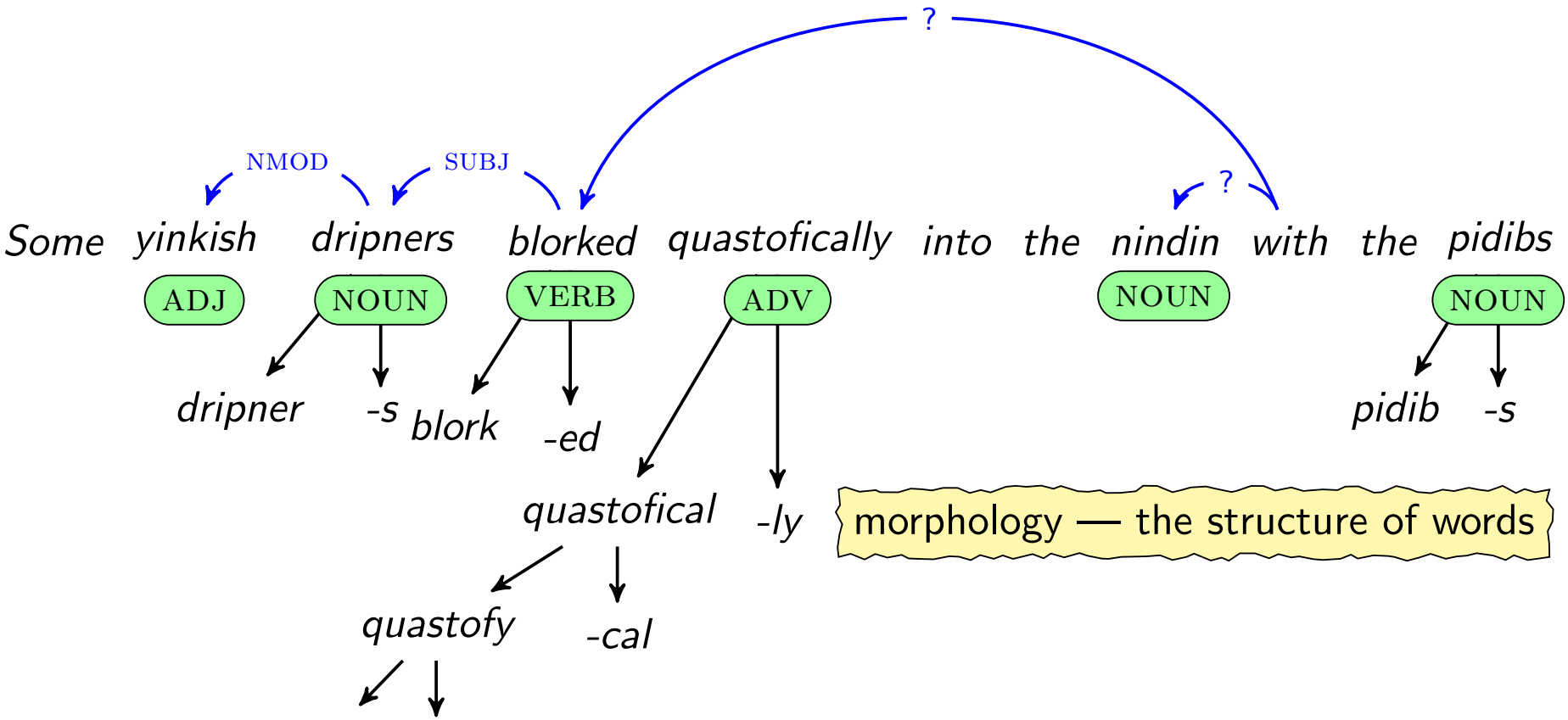
Structuring a sentence



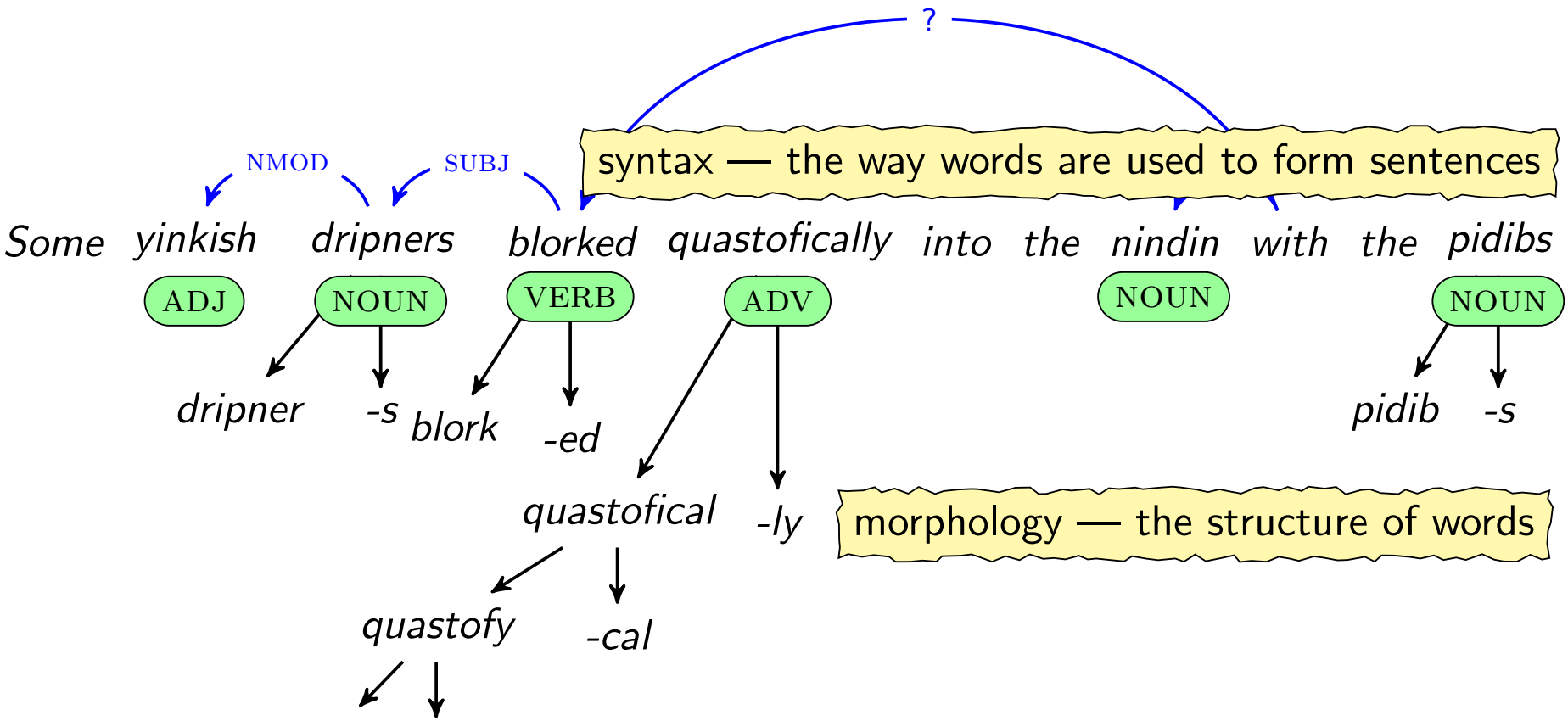
Structuring a sentence



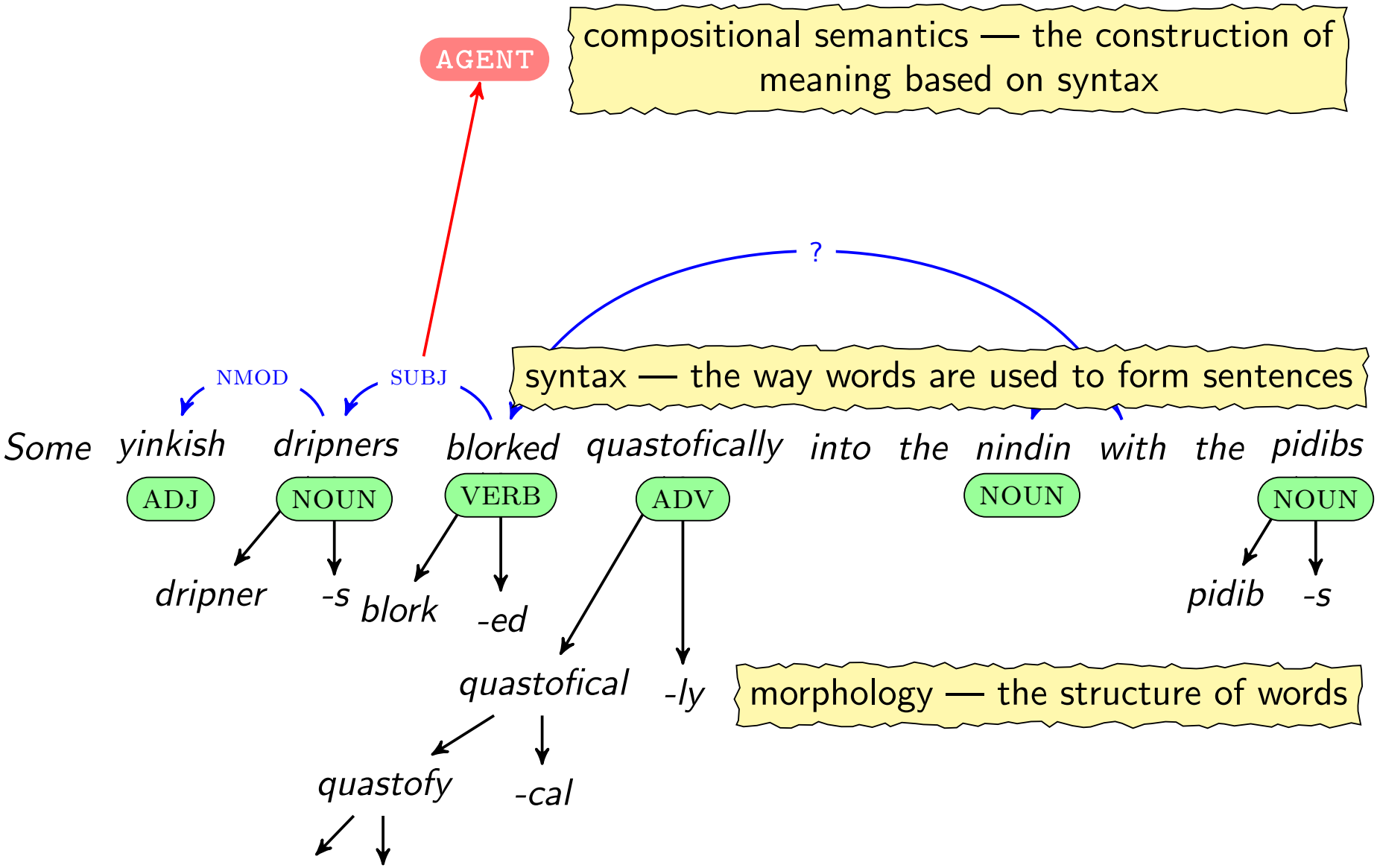
Structuring a sentence



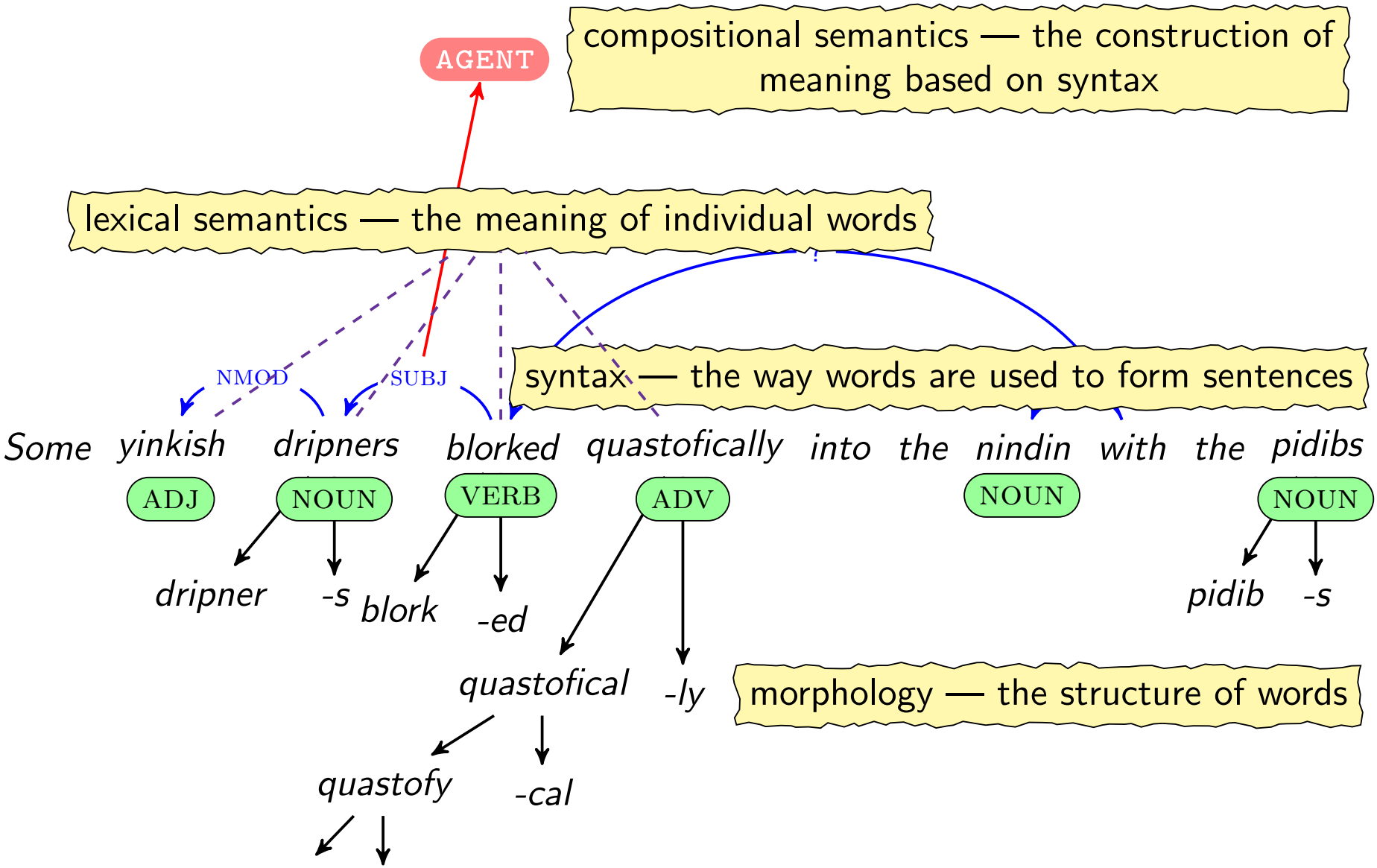
Structuring a sentence



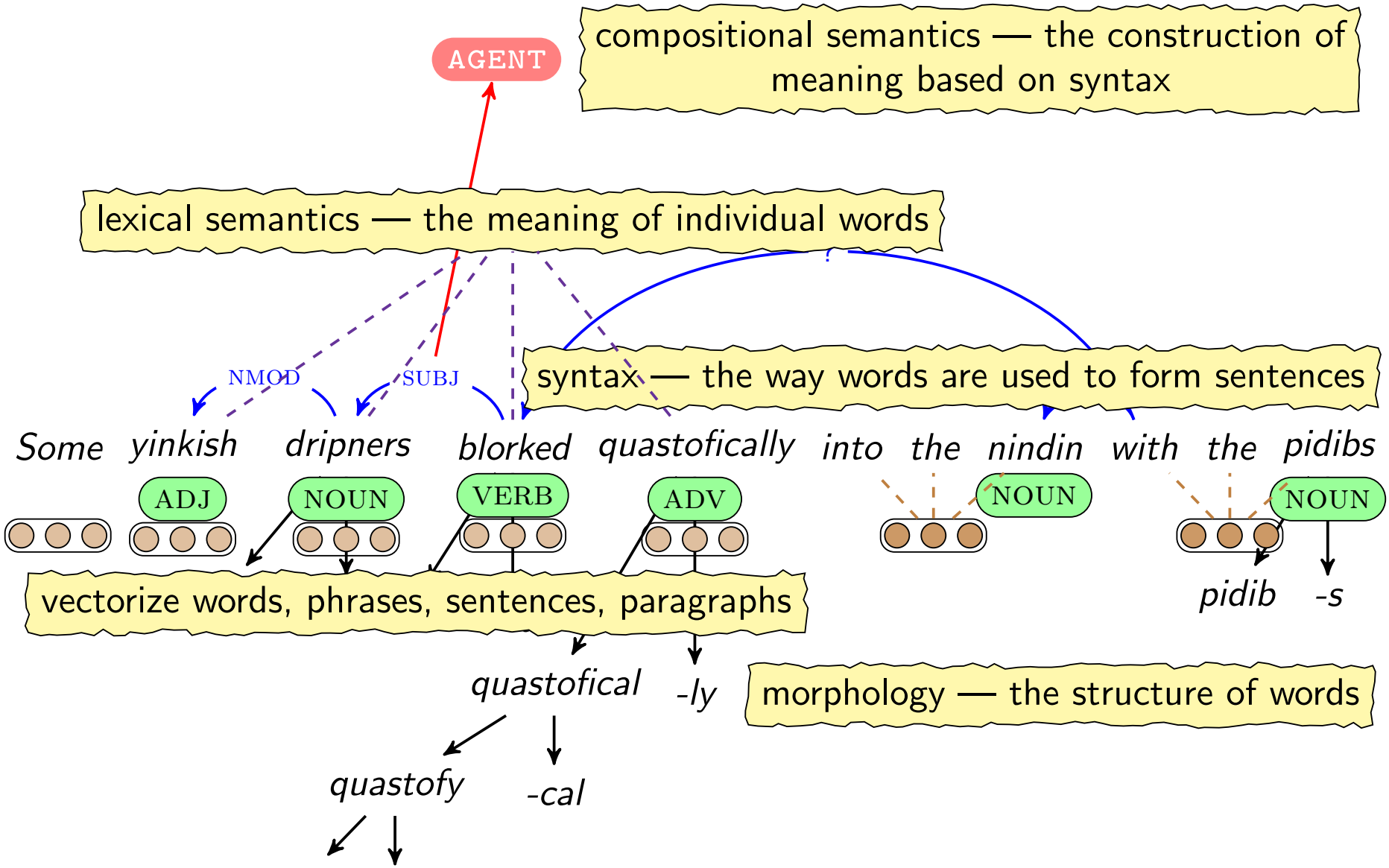
Structuring a sentence



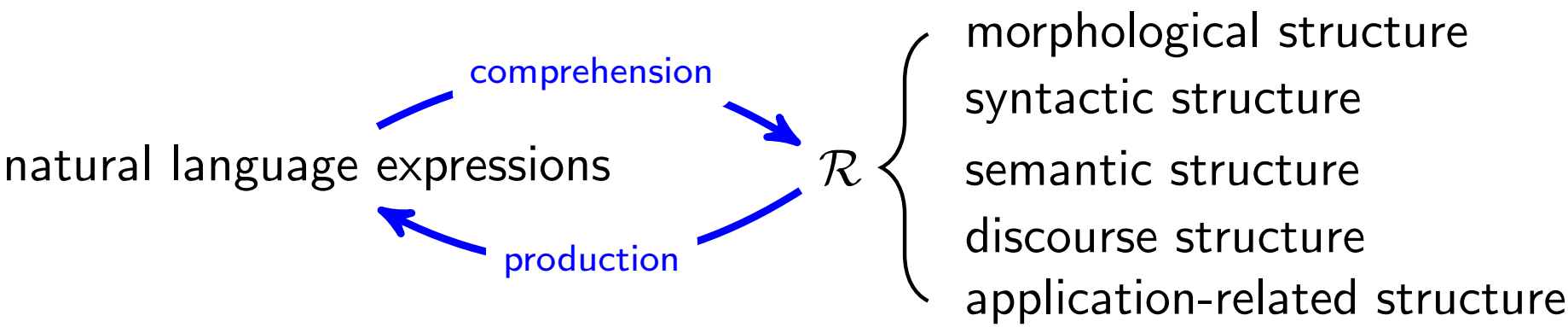
Structuring a sentence



Structuring a sentence



Form transformation

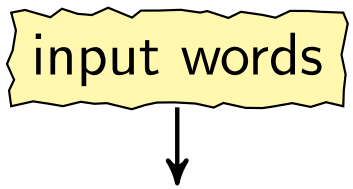


CoNLL shared tasks

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

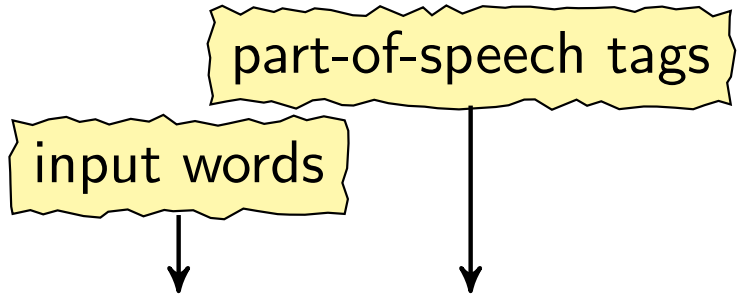
2019	Cross-Framework Meaning Representation Parsing
2018/2017	Multilingual Parsing from Raw Text to Universal Dependencies
2018/2017	Universal Morphological Reinflection
2016/2016	(Multilingual) Shallow Discourse Parsing
2014/2013	Grammatical Error Correction
2012/2011	Modelling (Multilingual) Unrestricted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



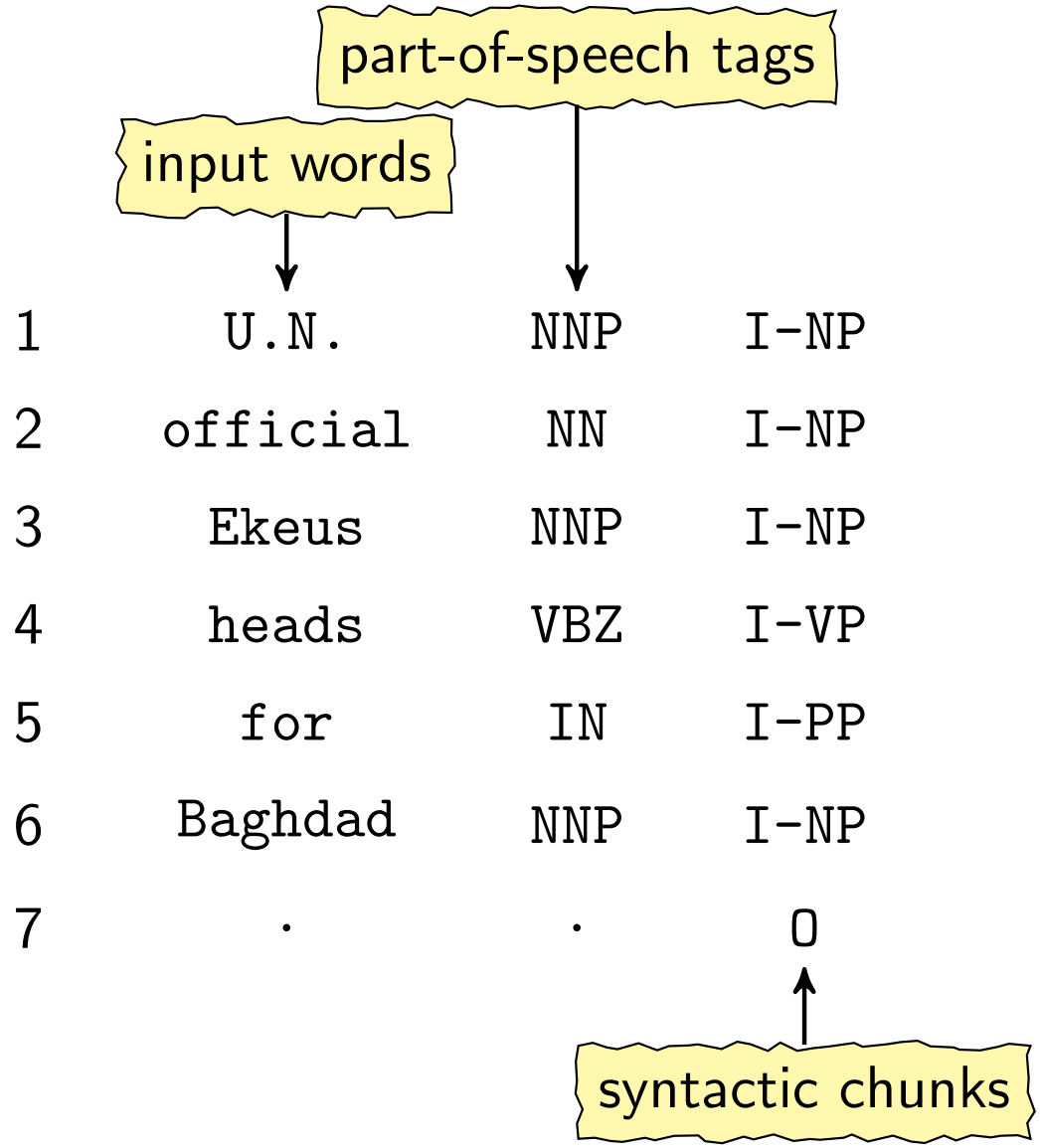

```
1       U.N.
2     official
3      Ekeus
4      heads
5       for
6     Baghdad
7        .
```

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

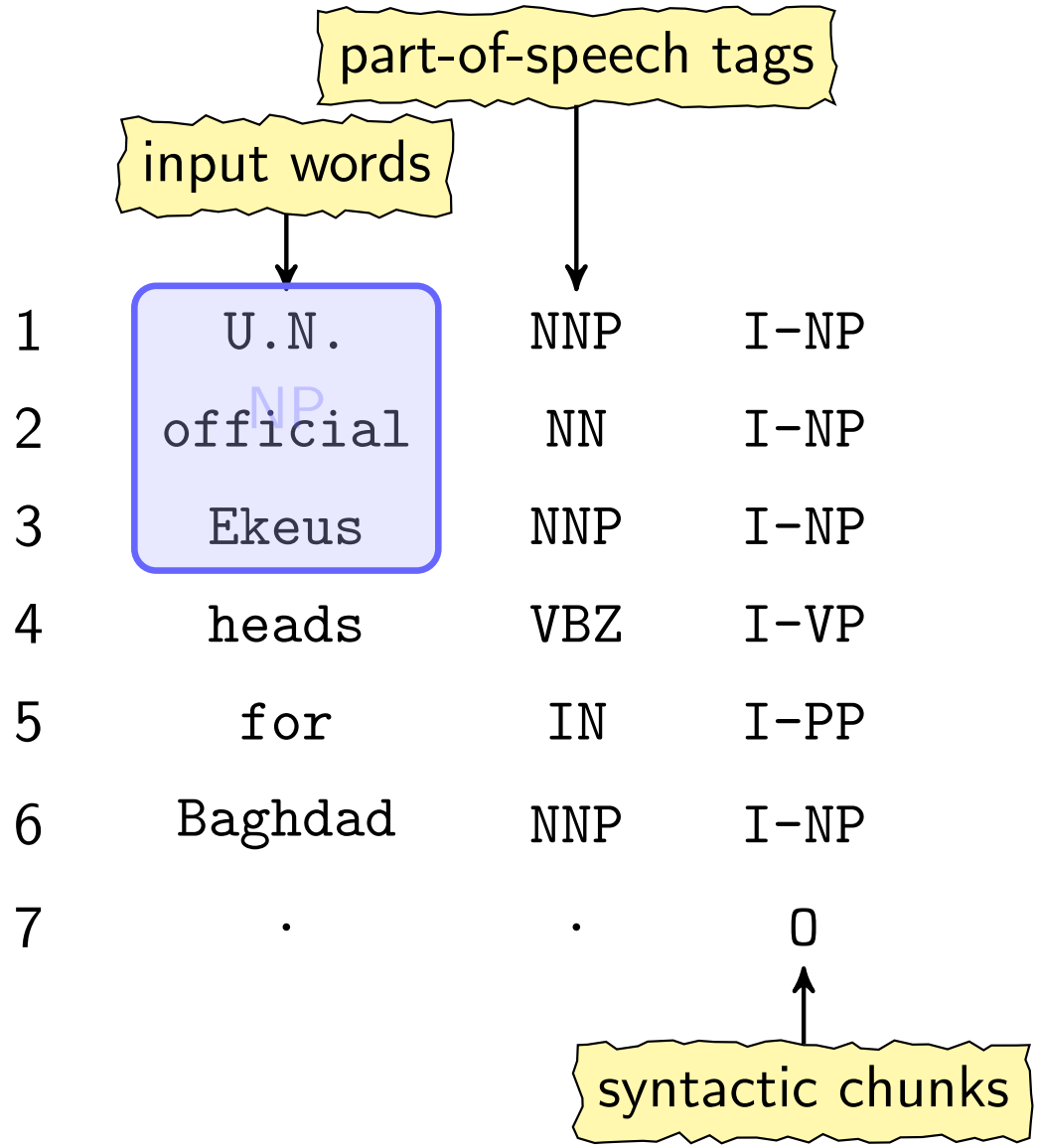


1	U.N.	NNP
2	official	NN
3	Ekeus	NNP
4	heads	VBZ
5	for	IN
6	Baghdad	NNP
7	.	.

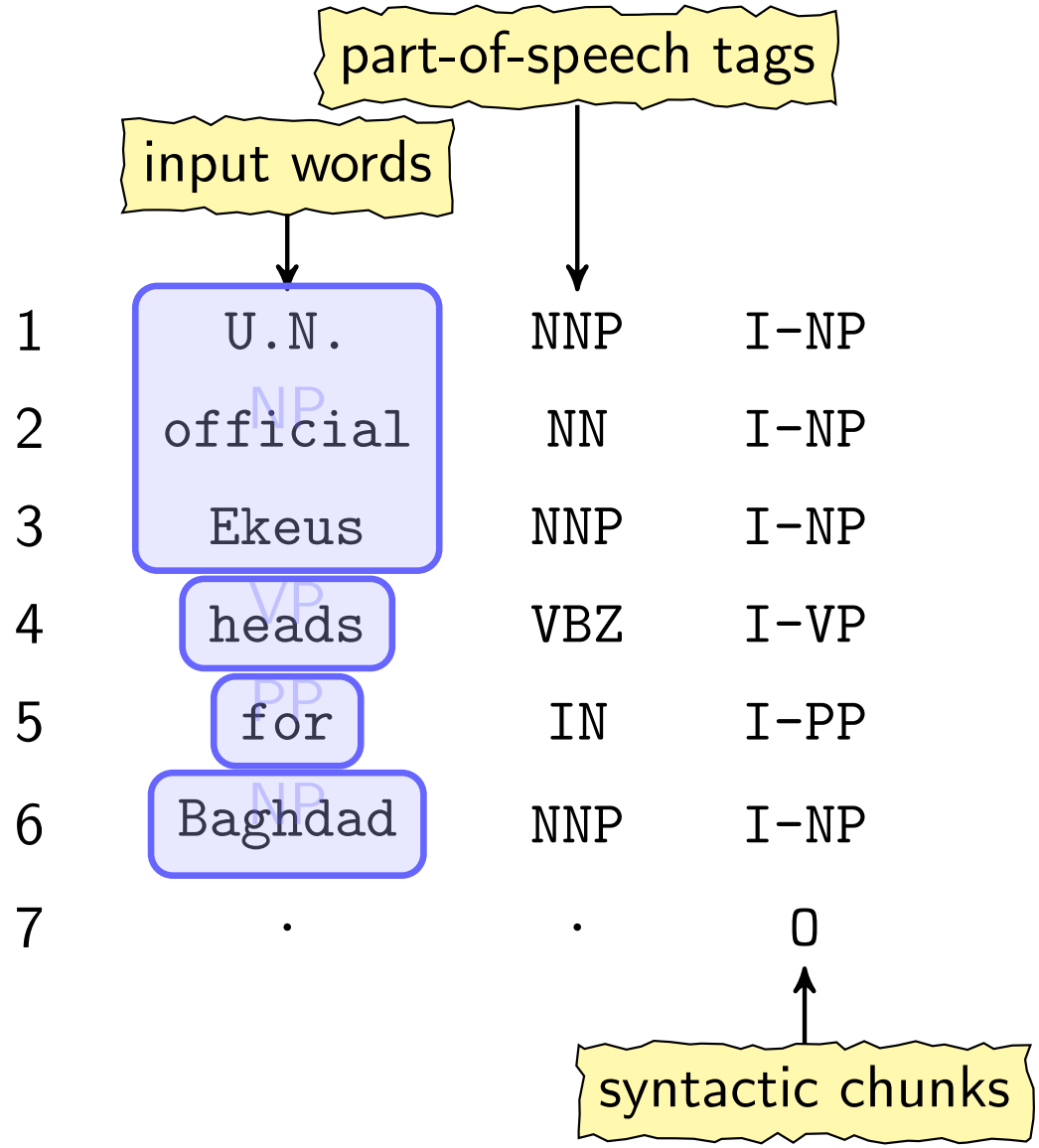
CoNLL ST 1999/2000/2002/2003/2006/2007/2017/2018



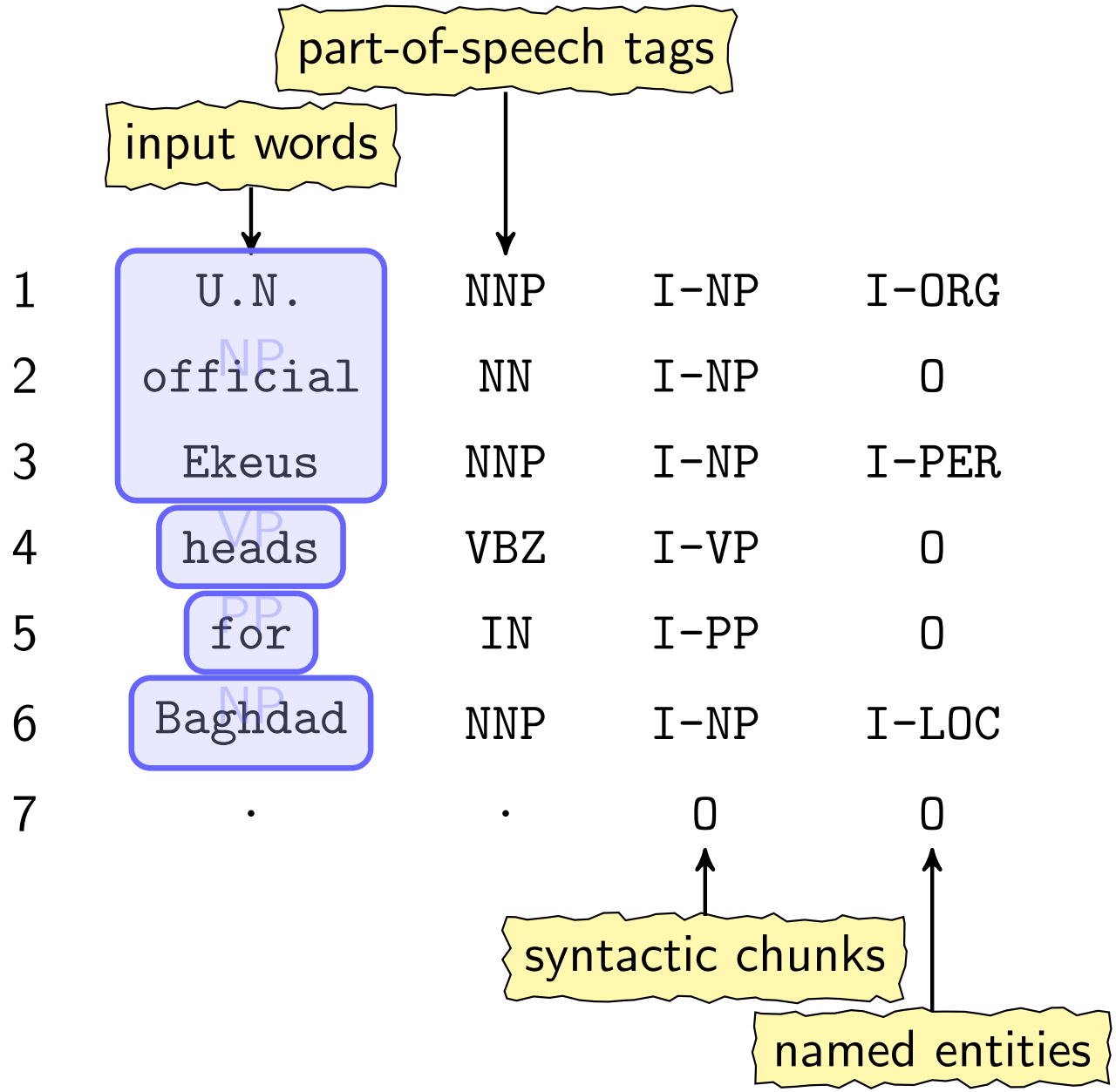
CoNLL ST 1999/2000/2002/2003/2006/2007/2017/2018



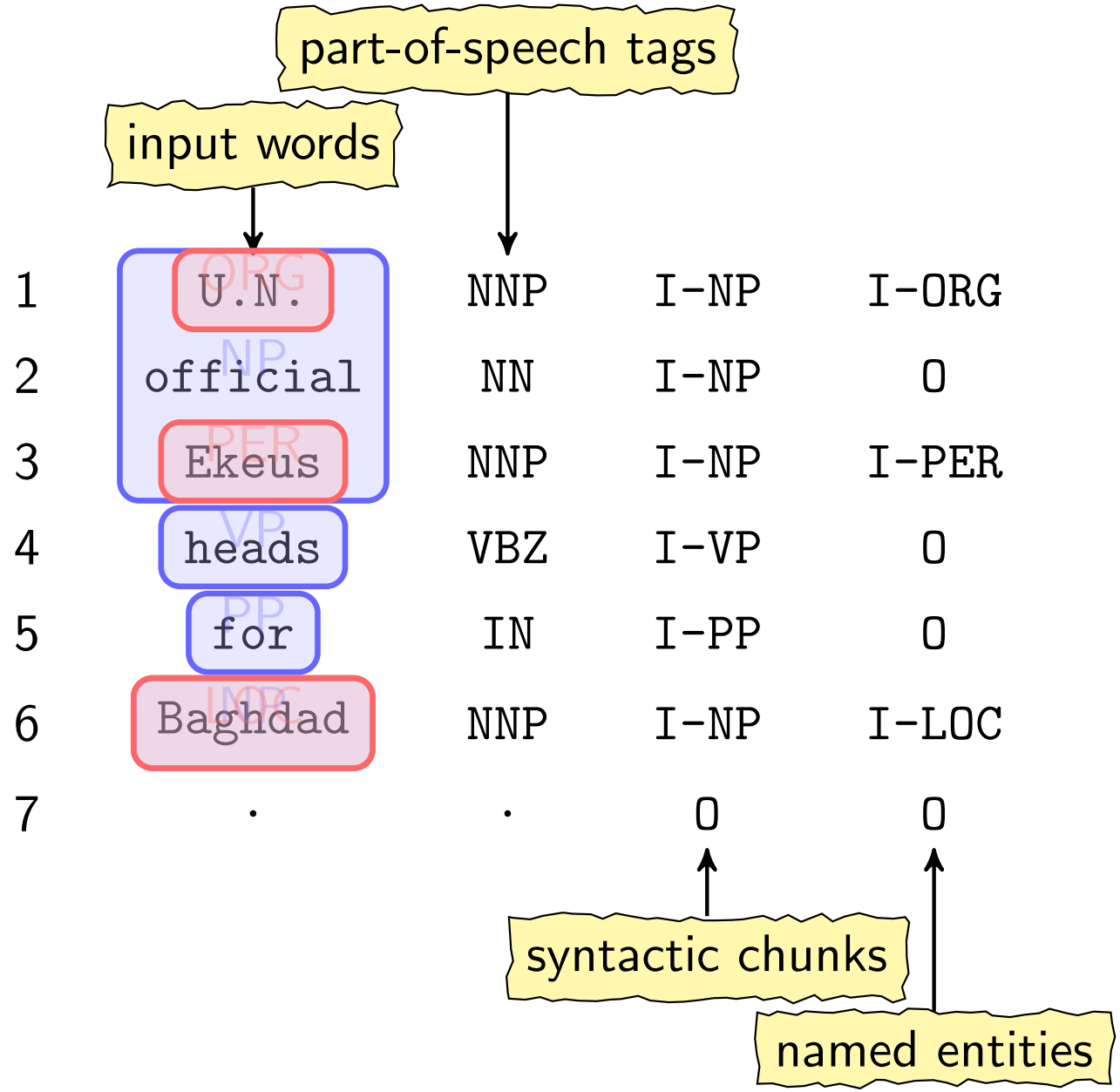
CoNLL ST 1999/2000/2002/2003/2006/2007/2017/2018



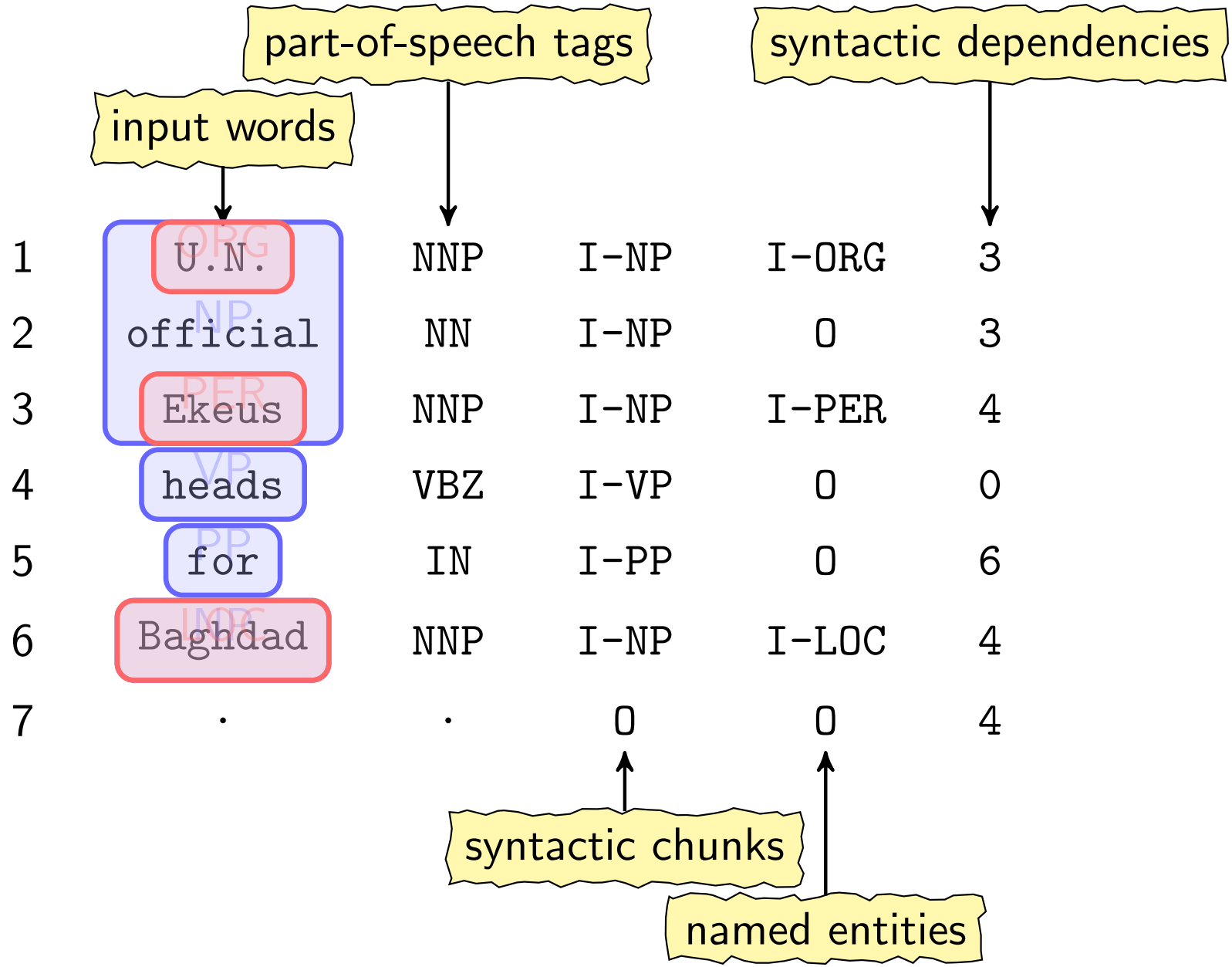
CoNLL ST 1999/2000/2002/2003/2006/2007/2017/2018



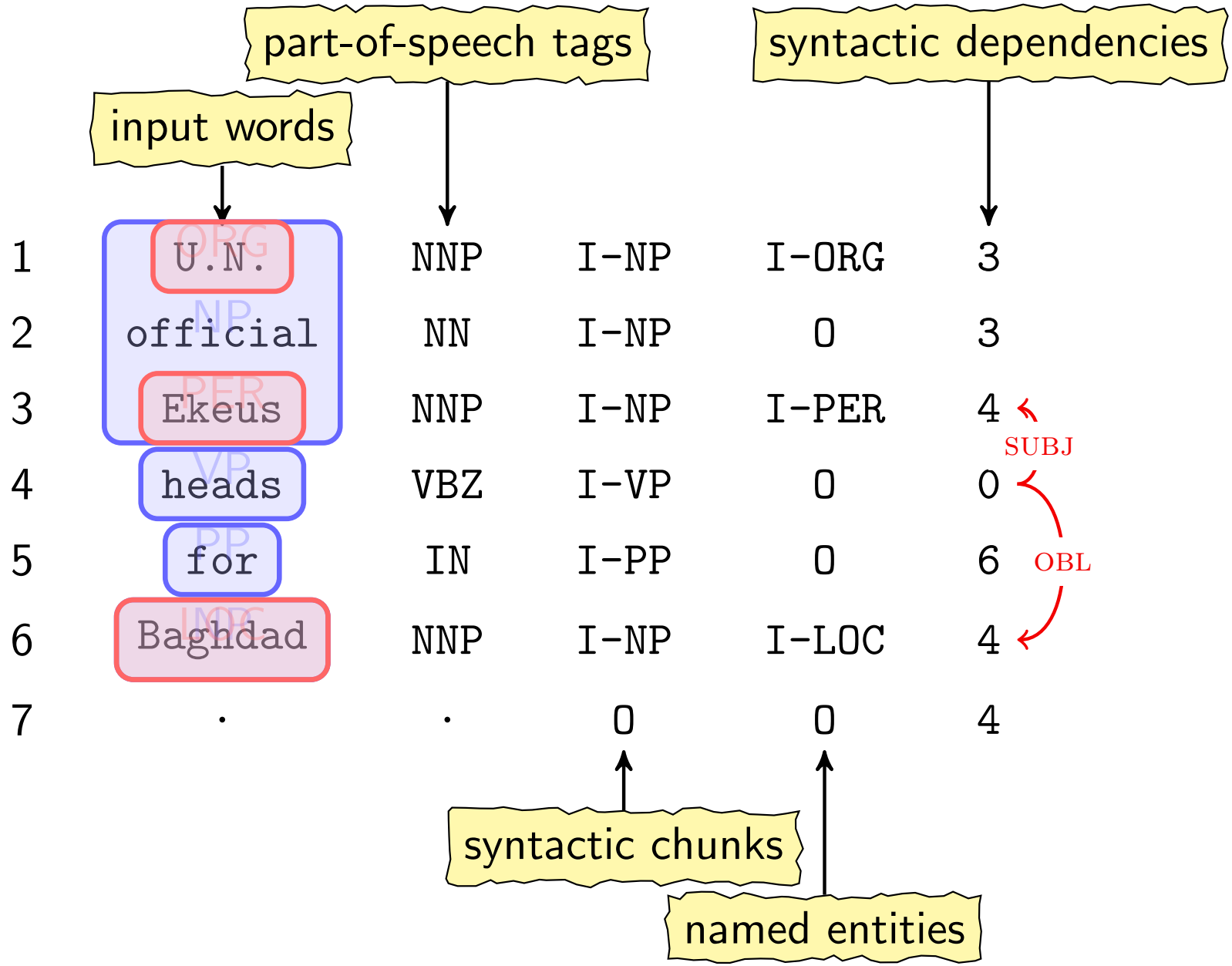
CoNLL ST 1999/2000/2002/2003/2006/2007/2017/2018



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



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



Reading

Sang and Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition.
`https://aclanthology.org/W03-0419.pdf`

Syllabus

1. Overview of Natural Language Processing (today)
2. Word: Morphology and Part-of-Speech (today)
3. Supervised machine learning and Perceptron (24 June)
4. Phrase structure and dependency (24 June)
5. Discussion (1 July)
6. Parsing models (1 July)
7. Neural parsing (8 July)

Project: Parsing for code-switching

- 17 June – 1 July: Data preparation
- 1 July – 15 July: Data annotation
- 15 July – 22 July: Running syntactic parsers
- after 8 July: Analysing parsing models

Code-switching

Code-switching: a speaker alternates between two or more languages in the context of a single conversation or situation.

Code-switching in Hong Kong

The English word "sure" / "cute" is mixed into an otherwise Cantonese sentence.

- 我唔sure
- cu唔cute啊