UPPSALA
UNIVERSITET

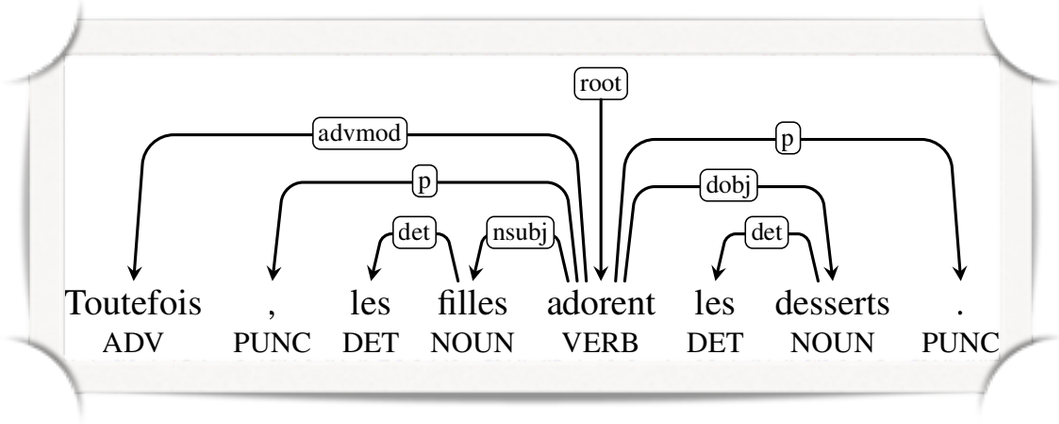
Dependency Parsing with Multiword Expressions

Joakim Nivre
Uppsala University

UPPSALA UNIVERSITET

Introduction

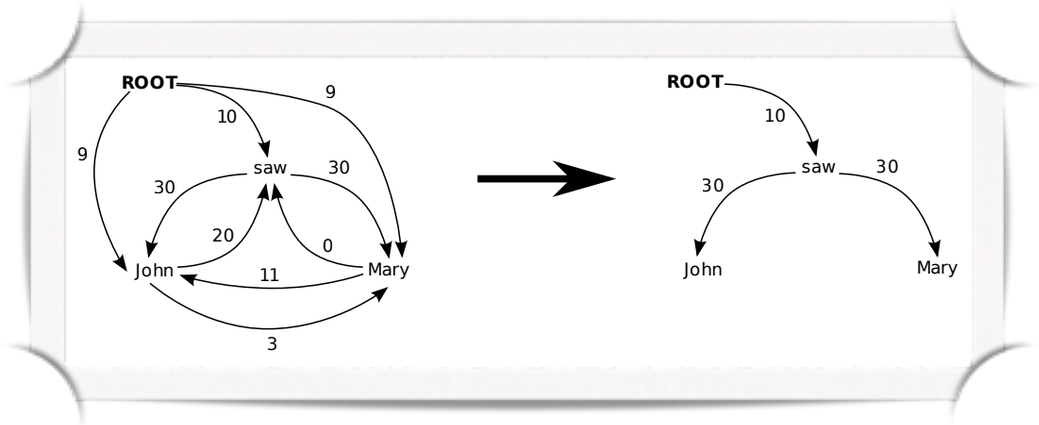
- Statistical dependency parsing
  - Map sentences to dependency trees
  - Learn mapping from (labeled) corpora



- Approaches
  - Graph-based: scoring function for trees, factored into subgraphs
  - Transition-based: scoring function for derivations, factored into transitions
- Recent convergence
  - Globally normalized models for structured prediction (perceptron style)
  - Rich features over the dependency tree (higher-order models)
  - Approximate search (beam search, dual decomposition, cube pruning)

Representations

- The spanning tree assumption:
  - Input is a sequence of tokens
  - Output is a directed spanning tree



- Problematic in two ways
  - A tree may not be sufficient to represent syntactic structure
  - There may not be a 1:1 mapping from tokens to nodes
    - French *du* = *de le* (1:*m*)
    - French *à cause de* = *à-cause-de* (*m*:1)
    - French *à cause du* = *à-cause-de le* (*m*:*n*)

UPPSALA
UNIVERSITET

This talk

- Giving up the spanning tree assumption
- Annotated resources
  - The Uni-Dep-TB project
- Parsing techniques
  - Transition-based model for parsing with MWEs

UPPSALA
UNIVERSITET

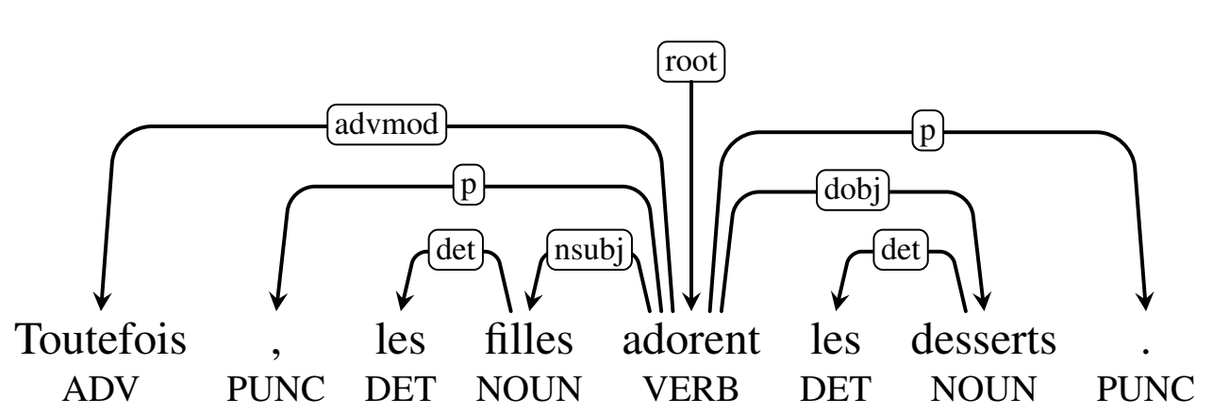
Uni-Dep-TB

- Treebank annotation schemes vary across languages
  - Hard to compare parsing results across languages (Nivre et al., 2007)
  - Hard to evaluate cross-lingual learning (McDonald et al., 2013)
- Recent initiatives:
  - HamleDT: Conversion of 29 existing treebanks to a PDT-like annotation scheme (Zeman et al., 2012)
  - Universal Dependency Treebank Project: New annotation, conversion and harmonization towards a cross-linguistically consistent annotation scheme (McDonald et al., 2013)

UPPSALA
UNIVERSITET

Uni-Dep-TB

https://code.google.com/p/uni-dep-tb/

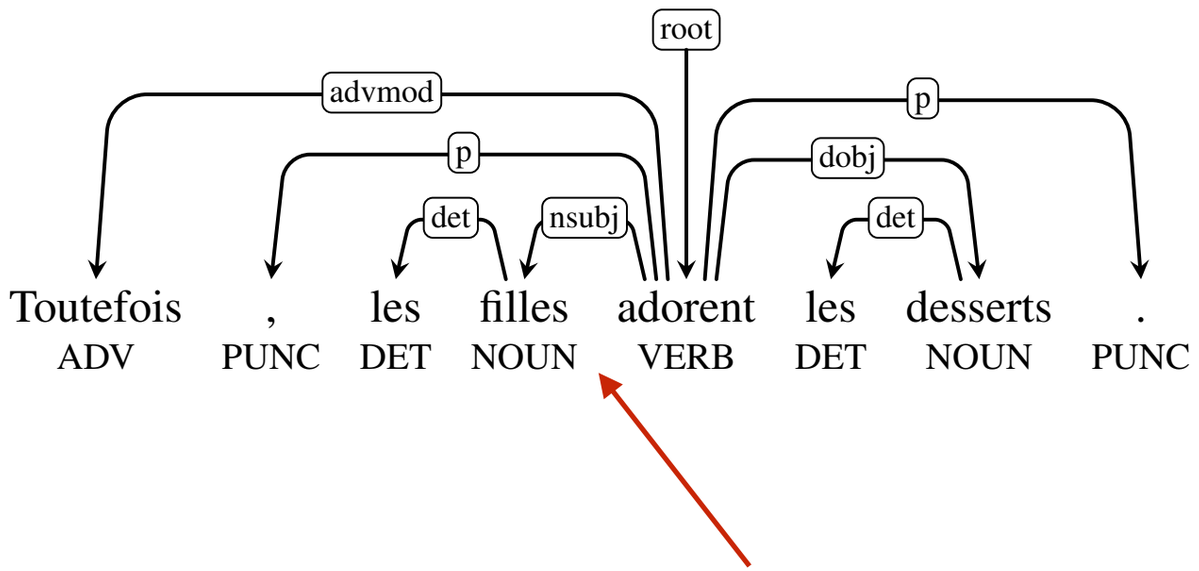


UPPSALA UNIVERSITET

Uni-Dep-TB

https://code.google.com/p/uni-dep-tb/

ADJ
ADP
ADV
CONJ
DET
NOUN
NUM
PRON
PRT
VERB
X
.



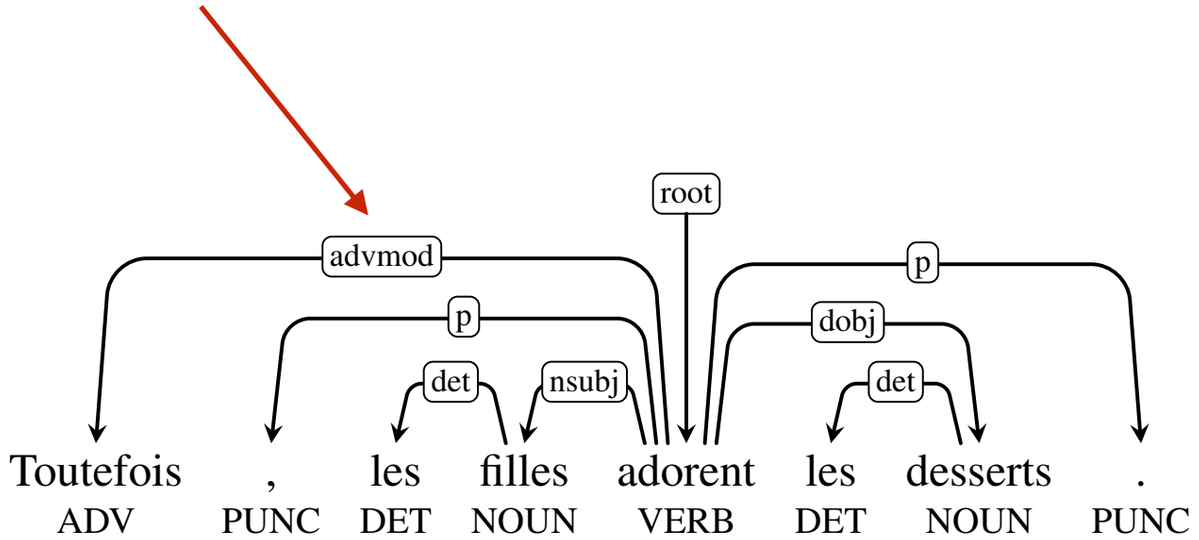
Google part-of-speech tags (Petrov et al, 2012), fine-grained language specific tags if available

UPPSALA UNIVERSITET

Uni-Dep-TB

https://code.google.com/p/uni-dep-tb/

Stanford dependencies (de Marneffe et al, 2006), adapted and harmonised for cross-lingual consistency



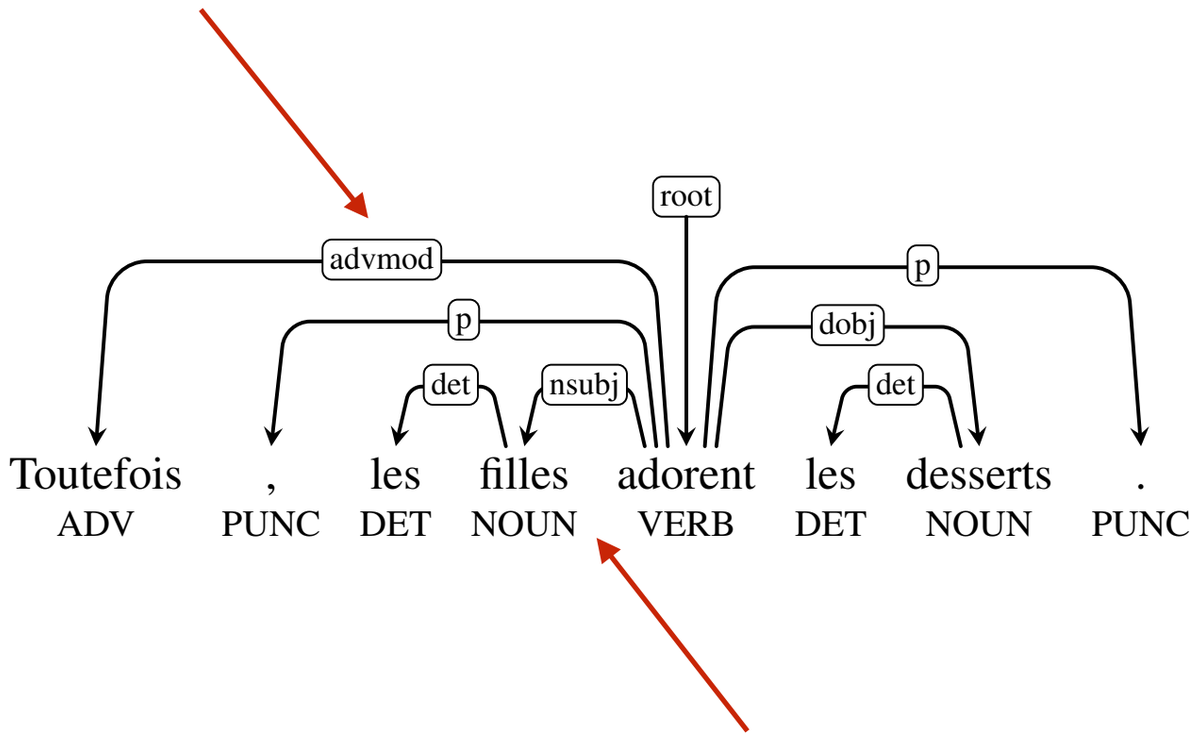
acomp	det
adp	dobj
adpcomp	expl
adpmod	infmod
adpobj	iobj
advcl	mark
advmod	mwe
amod	neg
appos	nmod
attr	nsubj
aux	nsubjpass
auxpass	num
cc	p
ccomp	parataxis
compmod	partmod
conj	poss
cop	prt
csubj	rcmod
csubjpass	rel
dep	xcomp

UPPSALA UNIVERSITET

Uni-Dep-TB

https://code.google.com/p/uni-dep-tb/

Stanford dependencies (de Marneffe et al, 2006), adapted and harmonised for cross-lingual consistency



Google part-of-speech tags (Petrov et al, 2012), fine-grained language specific tags if available

Version 1.0:
English
French
German
Korean
Spanish
Swedish
July 2013

UPPSALA UNIVERSITET

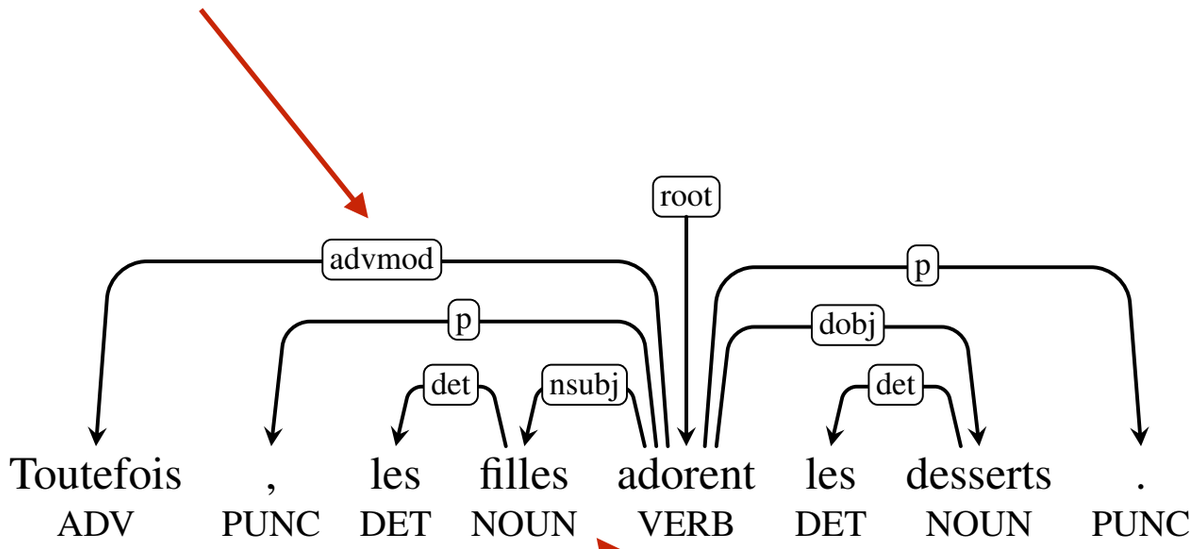
Uni-Dep-TB

https://code.google.com/p/uni-dep-tb/

Stanford dependencies (de Marneffe et al, 2006), adapted and harmonised for cross-lingual consistency

Version 2.0:
English
Finnish
French
German
Italian
Indonesian
Japanese
Korean
Portuguese
Spanish
Swedish
February 2014

Version 1.0:
English
French
German
Korean
Spanish
Swedish
July 2013



Google part-of-speech tags (Petrov et al, 2012), fine-grained language specific tags if available

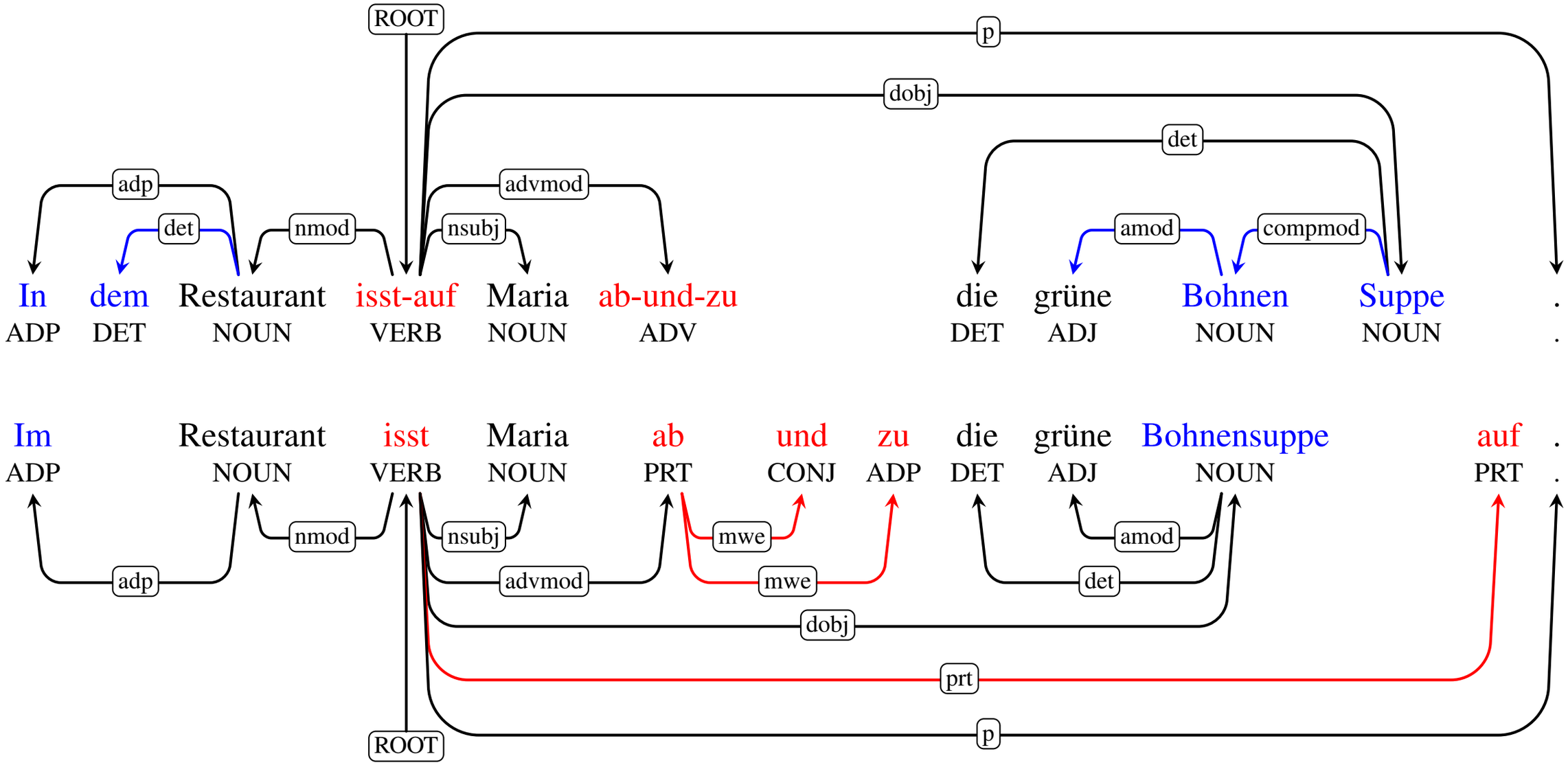
UPPSALA UNIVERSITET

Problems and Prospects

- Treebanks still satisfy the spanning tree assumption
  - For MWEs, the mwe relation can be used to annotate pseudo-dependencies — but no consistent guidelines
  - No consistent principles for segmenting contractions, clitics, etc.
- Recent initiative
  - Dagsthuhl seminar on MT for morphologically rich languages
  - Working group on extending the Uni-Dep-TB scheme

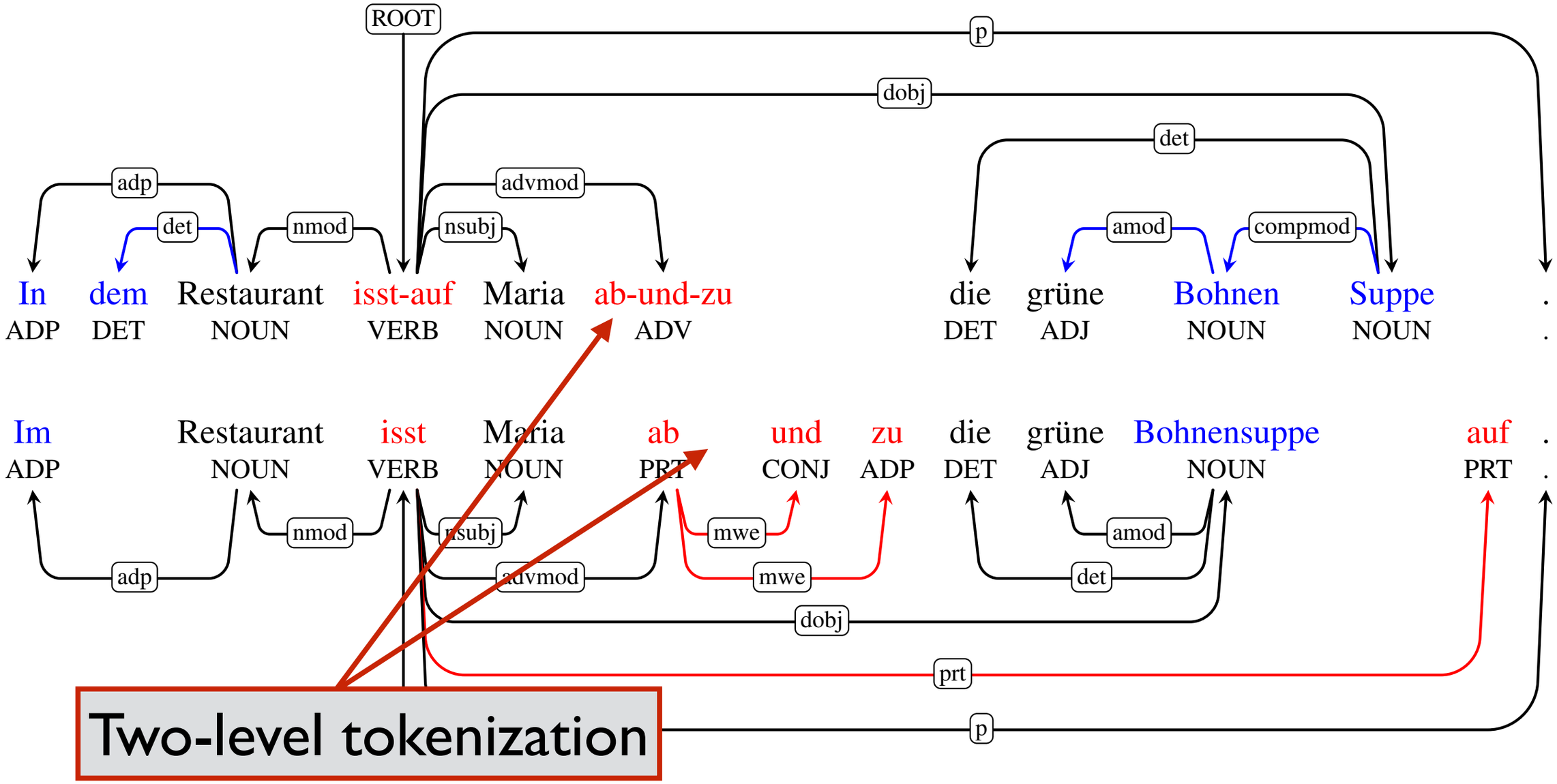
UPPSALA UNIVERSITET

Dagstuhl Proposal



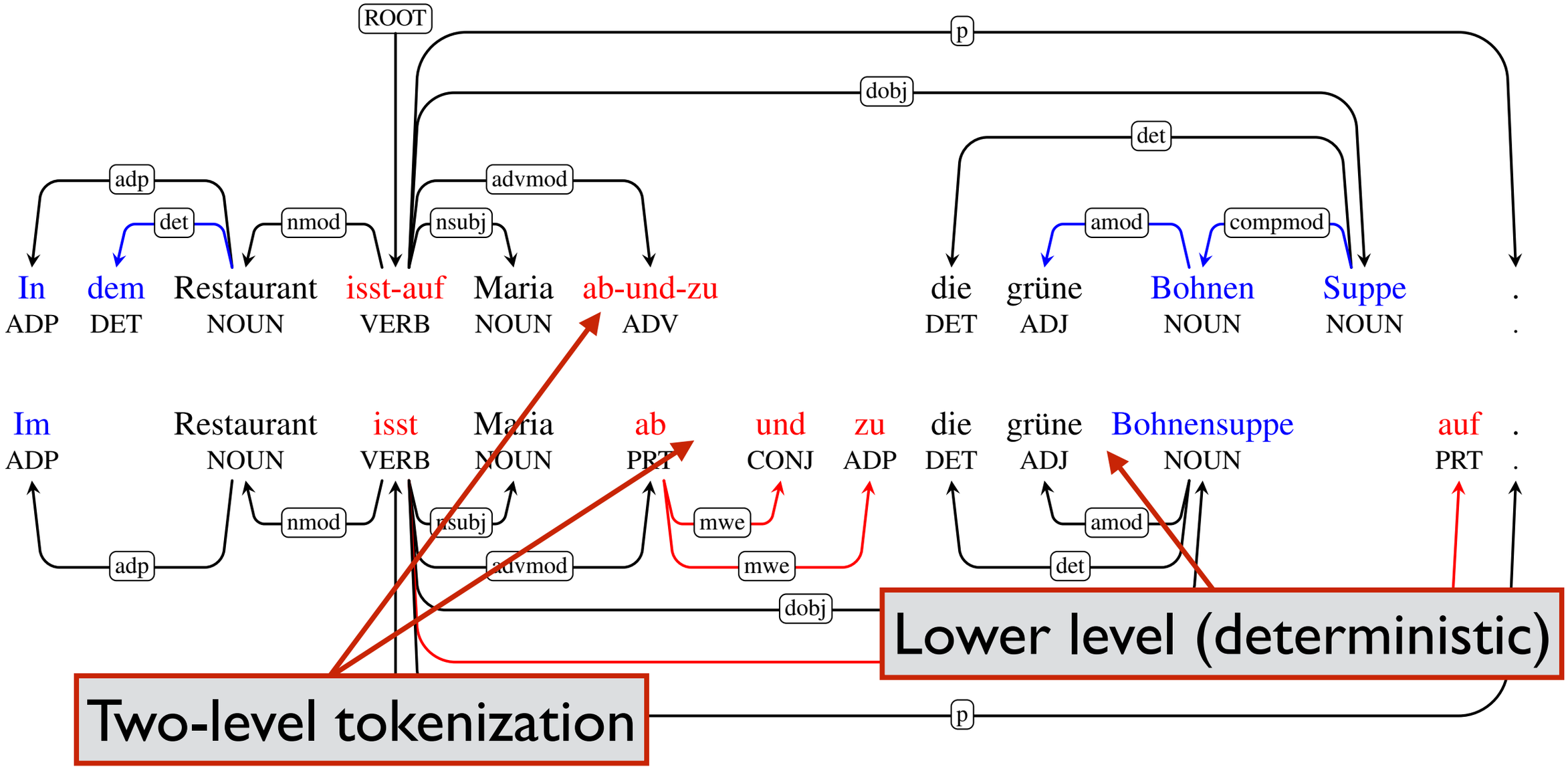
UPPSALA UNIVERSITET

Dagstuhl Proposal



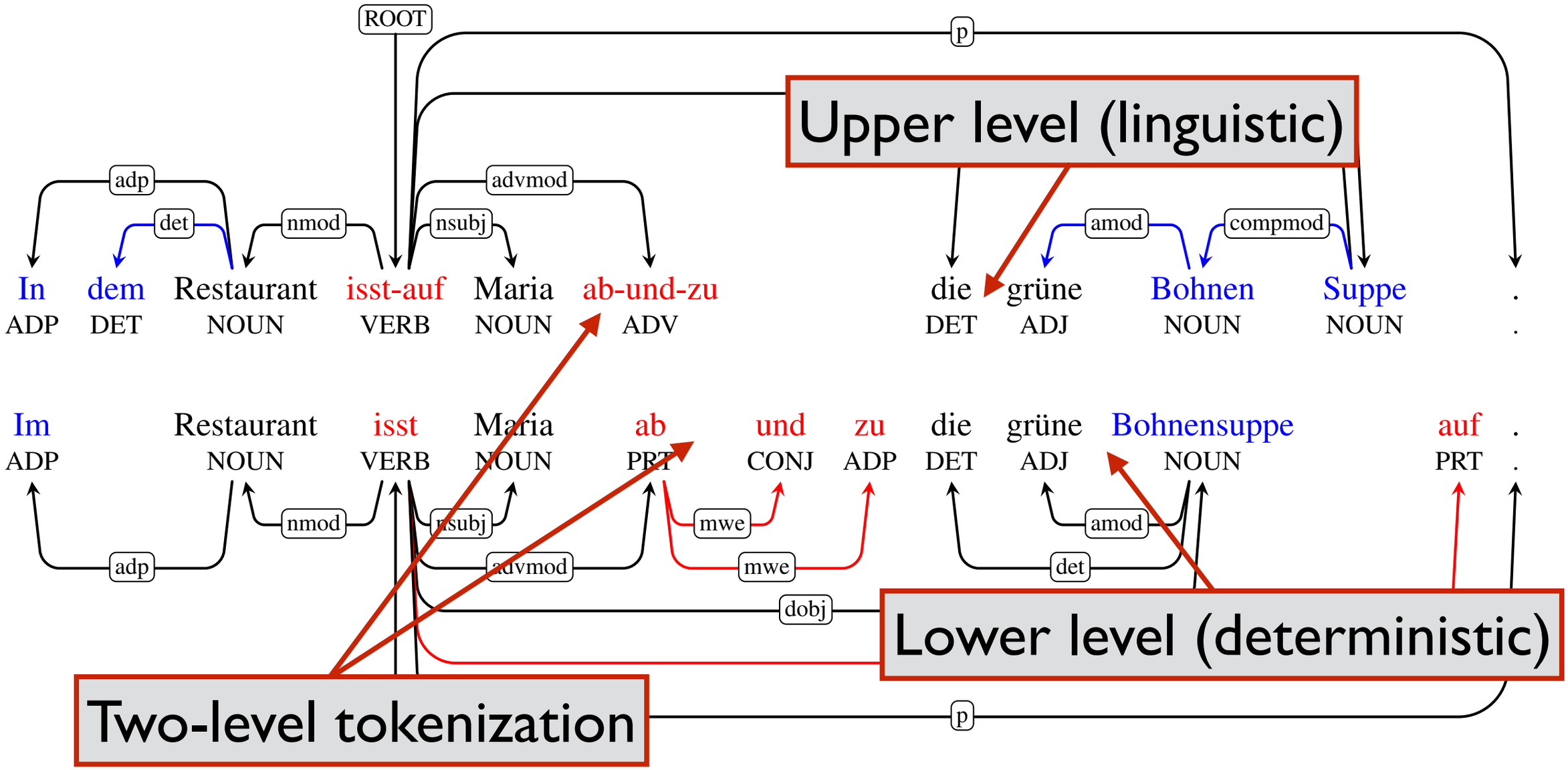
UPPSALA UNIVERSITET

Dagstuhl Proposal



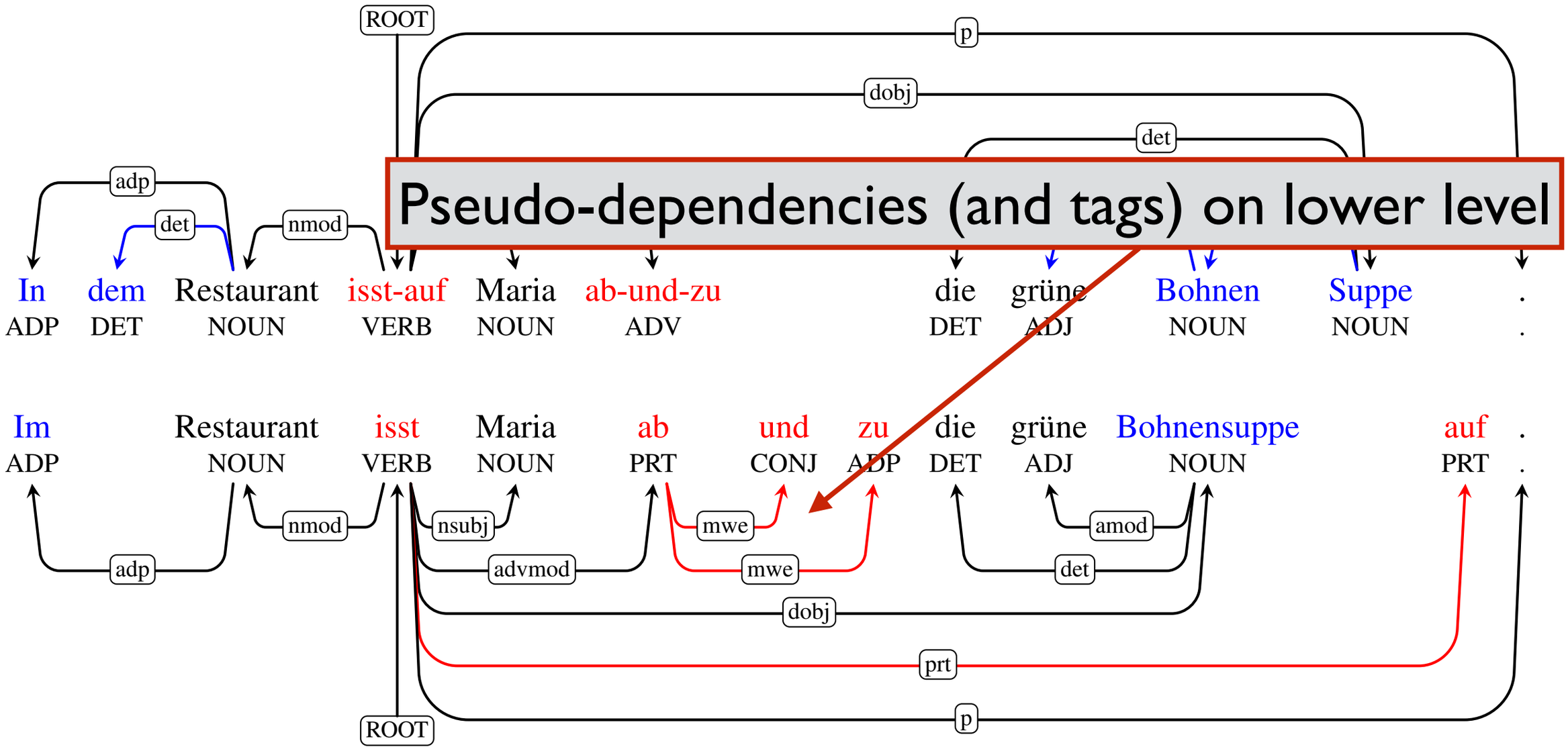
UPPSALA UNIVERSITET

Dagstuhl Proposal



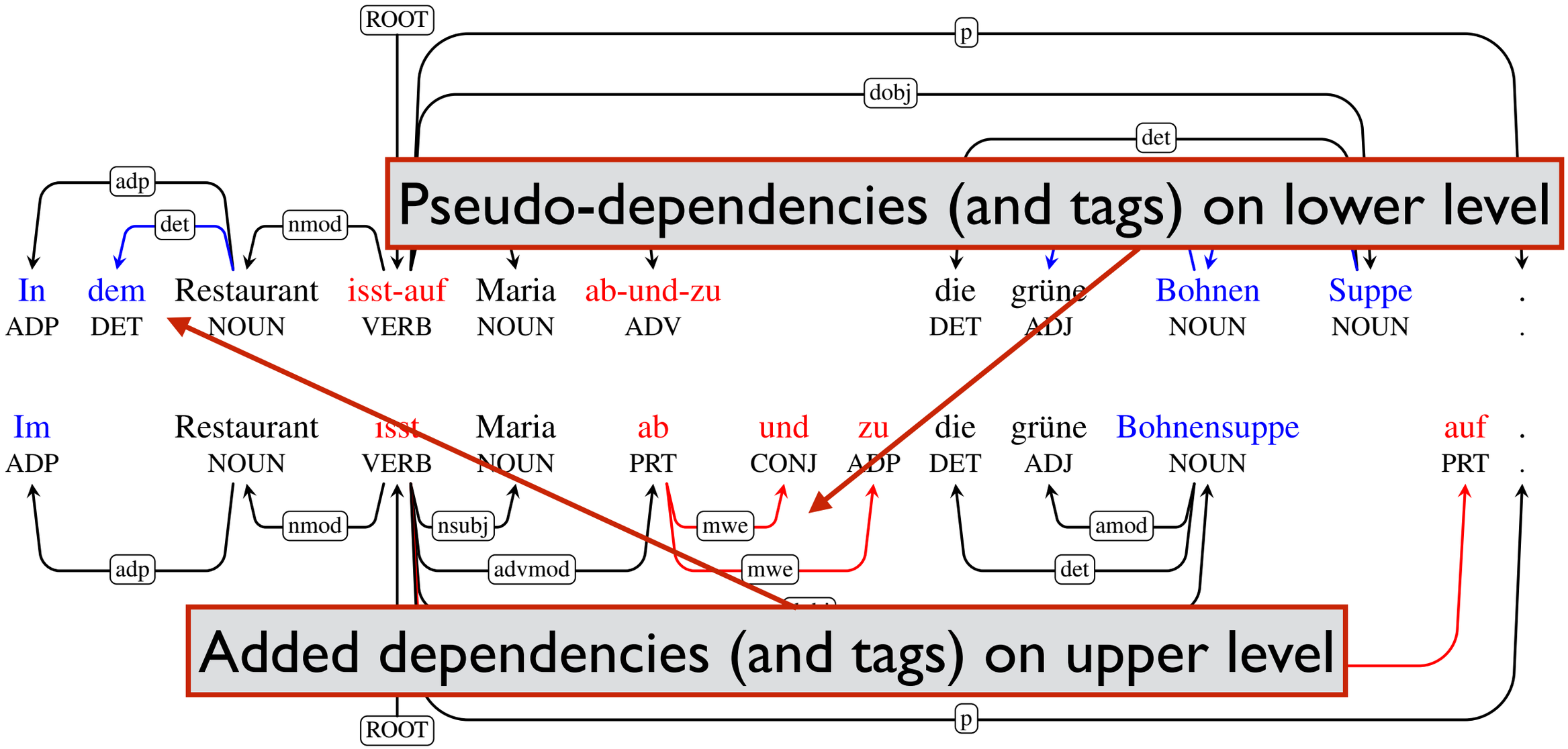
UPPSALA UNIVERSITET

Dagstuhl Proposal



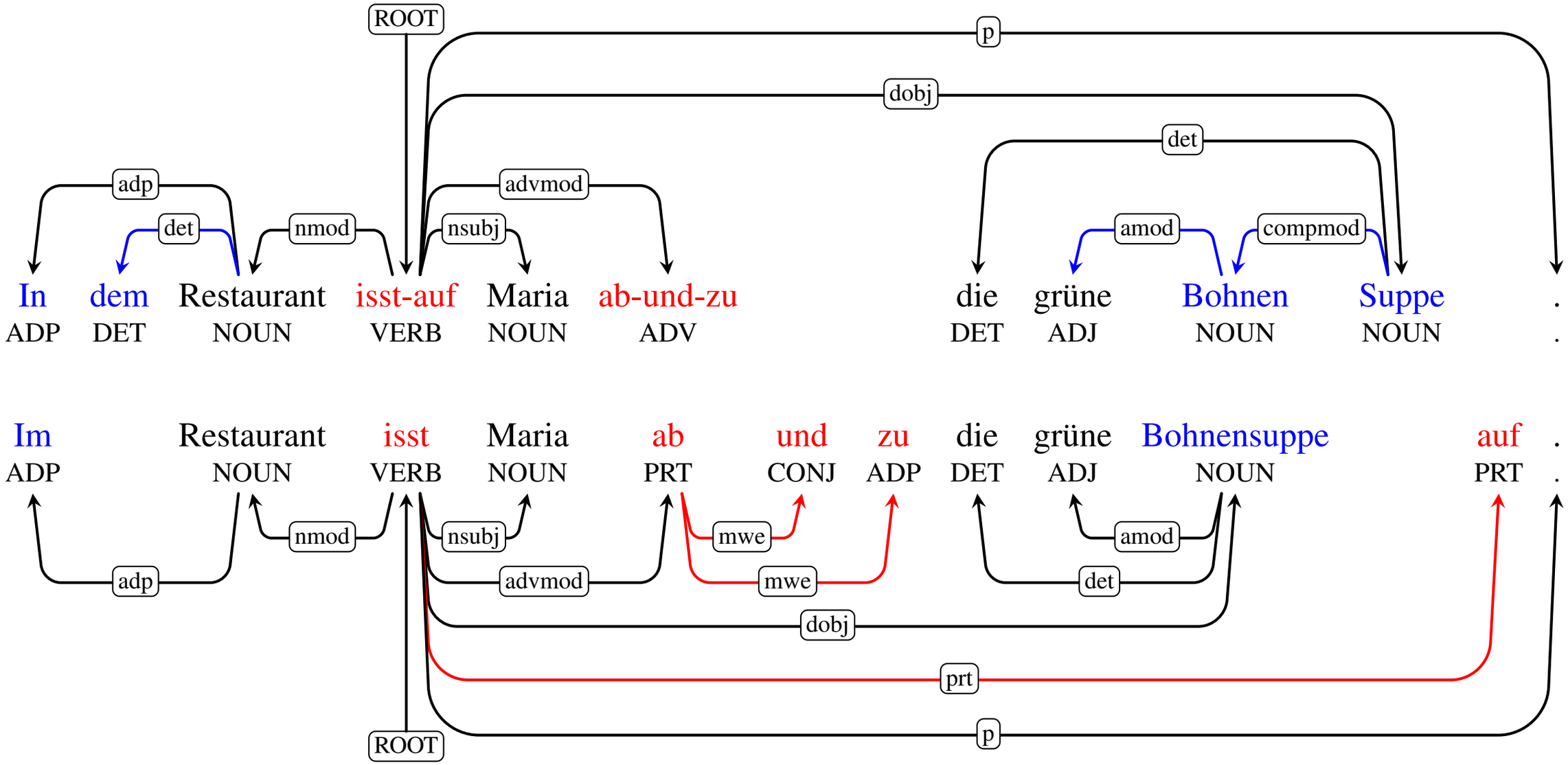
UPPSALA UNIVERSITET

Dagstuhl Proposal



UPPSALA UNIVERSITET

Dagstuhl Proposal



UPPSALA UNIVERSITET

Outlook

- Two-level tokenization adds flexibility (and usefulness)
  - If you want to continue playing the treebank game, ignore lower level
  - If you want to build robust applications, ignore upper level
  - If you want to handle all linguistic phenomena, use both
- The way forward
  - No annotated resources exist in this format yet
  - Uni-Dep-TB is meant to be a community effort
  - Any volunteers?

UPPSALA UNIVERSITET

Transition-Based Parsing

- Transition system
  - Abstract state machine for mapping sentences to dependency trees
  - Configurations = parser states
  - Transitions = parser actions
- Scoring model
  - Statistical model for scoring possible transitions out of a configuration
  - Usually a linear model learned from treebank derivations
- Search algorithm
  - Algorithm for finding the highest scoring sequence of transitions
  - Usually approximate search (greedy search, beam search)

UPPSALA UNIVERSITET

Arc-Standard System

- Configurations
  - A buffer B of input tokens
  - A stack S of tree nodes
  - A set of A dependency arcs

node = token

Shift:	$(S, w\|B, A)$	$\Rightarrow$	$(S\|w, B, A)$
Right-Arc:	$(S\|w\|w', B, A)$	$\Rightarrow$	$(S\|w, B, A[w \rightarrow w'])$
Left-Arc:	$(S\|w\|w', B, A)$	$\Rightarrow$	$(S\|w', B, A[w' \rightarrow w])$

UPPSALA
UNIVERSITET

Parsing Example

she	found	the	word

UPPSALA
UNIVERSITET

Parsing Example

she

found | the | word

UPPSALA
UNIVERSITET

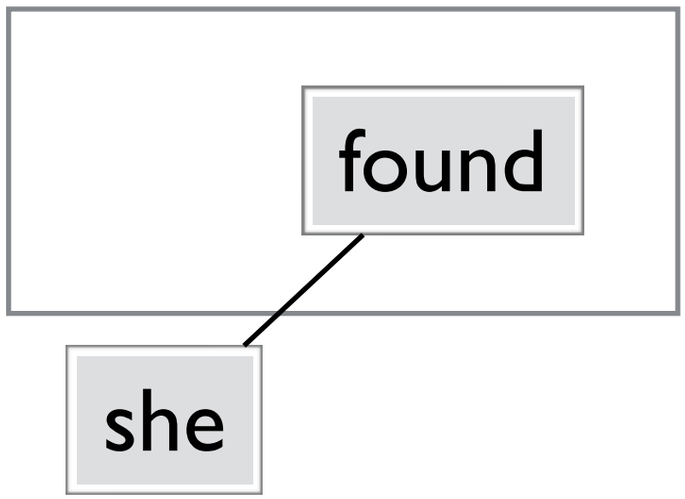
Parsing Example

she found

the word

UPPSALA UNIVERSITET

Parsing Example

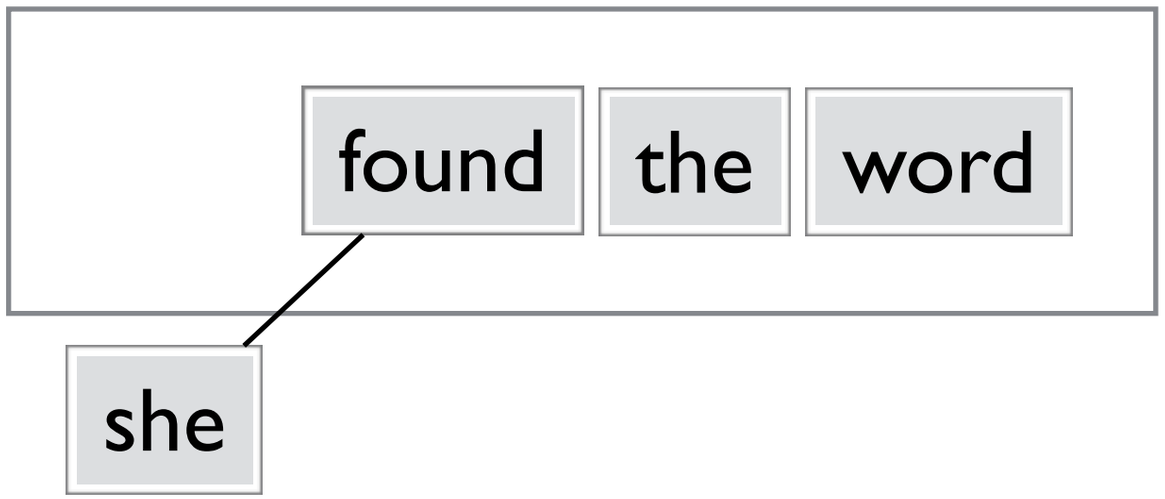


the word

Parsing Example

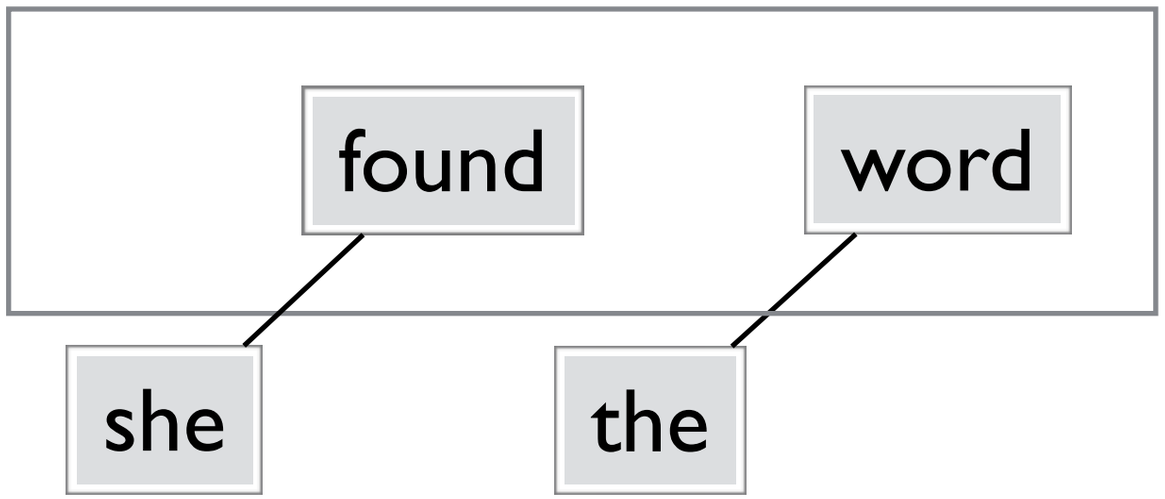
UPPSALA
UNIVERSITET

Parsing Example



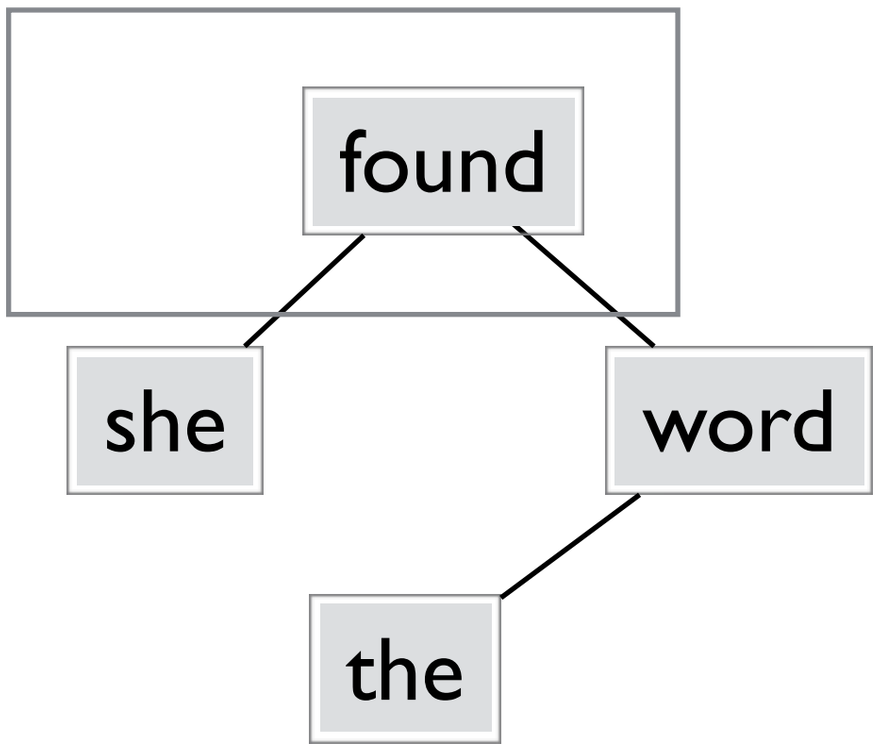
UPPSALA
UNIVERSITET

Parsing Example



UPPSALA
UNIVERSITET

Parsing Example



UPPSALA
UNIVERSITET

What about MWEs?

UPPSALA
UNIVERSITET

What about MWEs?

- MWEs can be handled in preprocessing
  - Only works (well) for fixed, unambiguous expressions

by-and-large

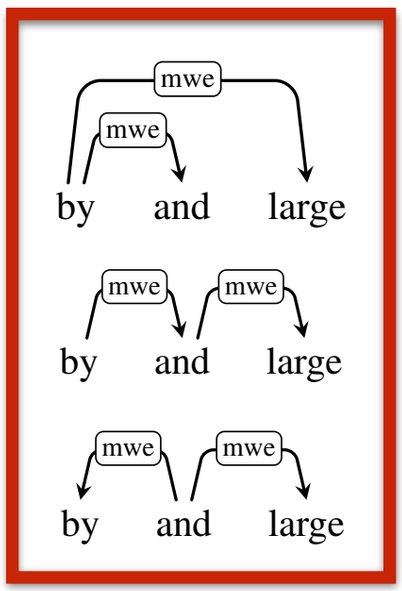
UPPSALA UNIVERSITET

What about MWEs?

- MWEs can be handled in preprocessing
  - Only works (well) for fixed, unambiguous expressions

by-and-large

- MWEs can be encoded with pseudo-dependencies
  - Structure is (sometimes) arbitrary
  - Lexical and grammatical features are distorted



UPPSALA UNIVERSITET

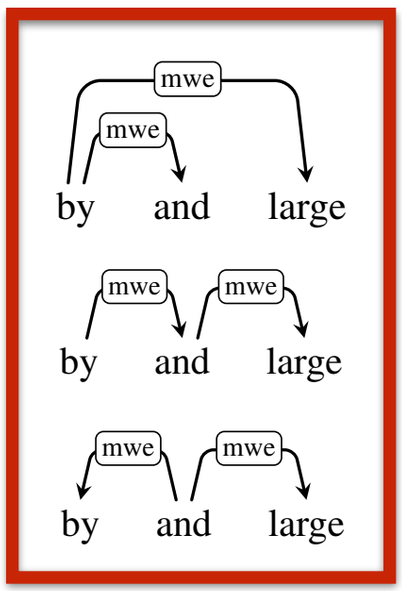
What about MWEs?

- MWEs can be handled in preprocessing
  - Only works (well) for fixed, unambiguous expressions

by-and-large

- MWEs can be encoded with pseudo-dependencies
  - Structure is (sometimes) arbitrary
  - Lexical and grammatical features are distorted

Nivre and Nilsson (2004)	MWE	Other
Oracle preprocessing	100.0	81.6
Pseudo-dependencies	71.1	80.7



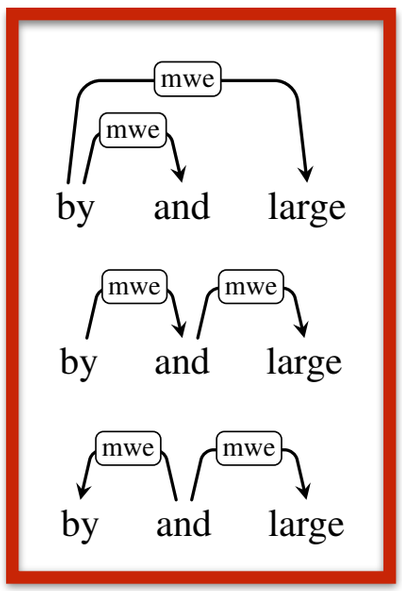
UPPSALA UNIVERSITET

What about MWEs?

- MWEs can be handled in preprocessing
  - Only works (well) for fixed, unambiguous expressions

by-and-large

- MWEs can be encoded with pseudo-dependencies
  - Structure is (sometimes) arbitrary
  - Lexical and grammatical features are distorted



- Can we do better?
  - Make MWEs first class citizens in parsing land
  - Add transition for merging tokens into MWEs

UPPSALA UNIVERSITET

New Transition System

- Configurations
  - A buffer B of input tokens
  - A stack S of tree nodes
  - A set of A dependency arcs

node = string of tokens

Shift:	$(S, w\|B, A)$	$\Rightarrow$	$(S\|w, B, A)$
Chunk:	$(S\|w\|w', B, A)$	$\Rightarrow$	$(S\|w\text{-}w', B, A)$
Swap:	$(S\|w\|w', B, A)$	$\Rightarrow$	$(S\|w', w\|B, A)$
Right-Arc:	$(S\|w\|w', B, A)$	$\Rightarrow$	$(S\|w, B, A[w \rightarrow w'])$
Left-Arc:	$(S\|w\|w', B, A)$	$\Rightarrow$	$(S\|w', B, A[w' \rightarrow w])$

UPPSALA
UNIVERSITET

Parsing Example 1

she	looked	up	the	word

UPPSALA
UNIVERSITET

Parsing Example 1

she

looked | up | the | word

UPPSALA
UNIVERSITET

Parsing Example I

she looked

up the word

UPPSALA
UNIVERSITET

Parsing Example 1

she | looked | up

the | word

UPPSALA
UNIVERSITET

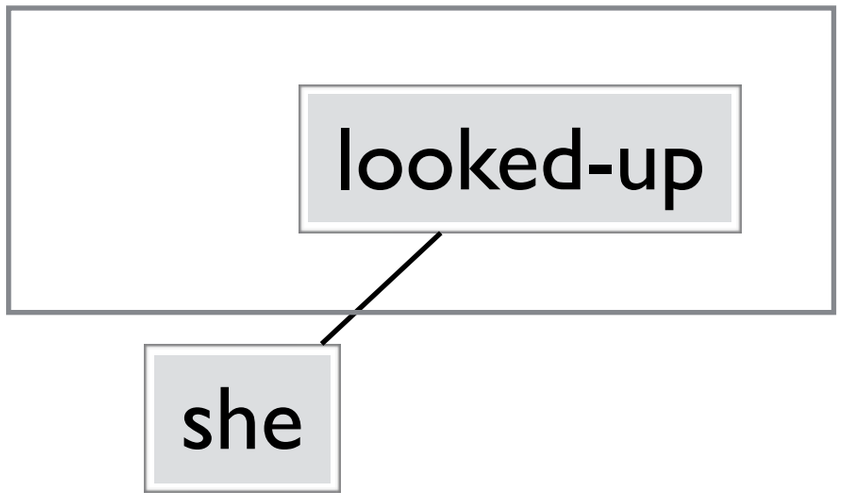
Parsing Example 1

she | looked-up

the | word

UPPSALA
UNIVERSITET

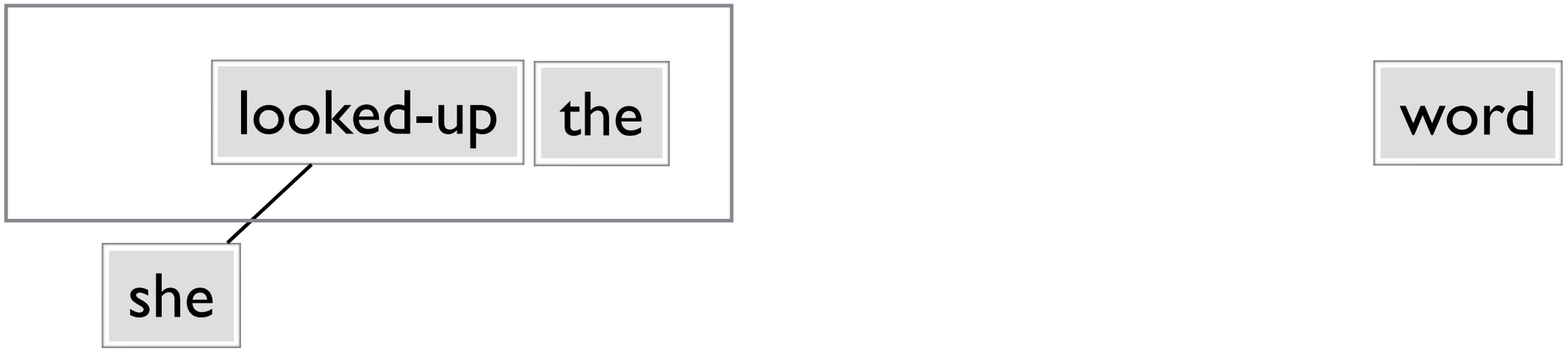
Parsing Example 1



the word

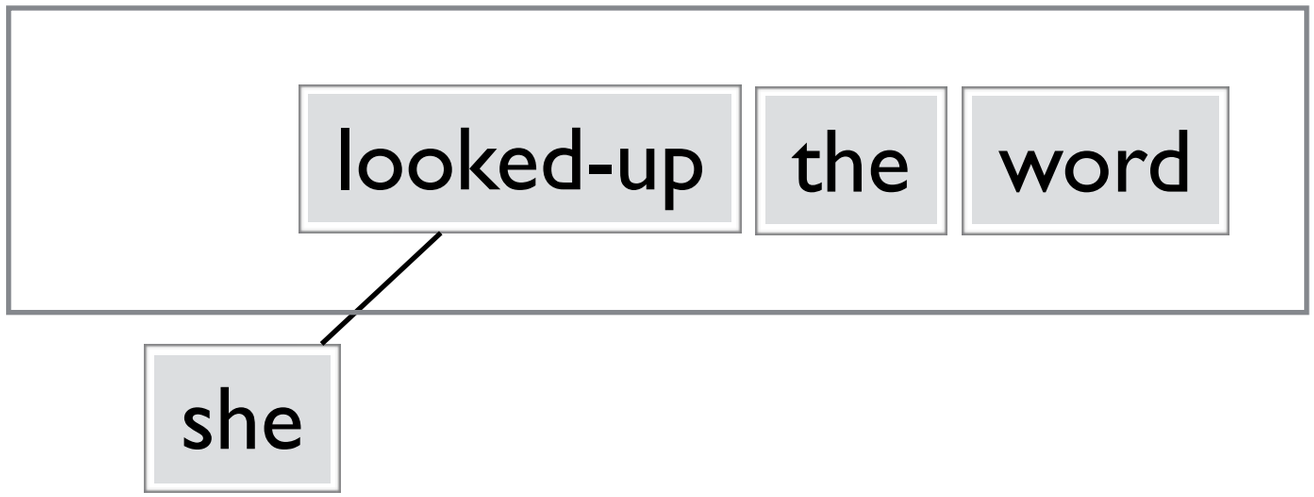
UPPSALA
UNIVERSITET

Parsing Example 1



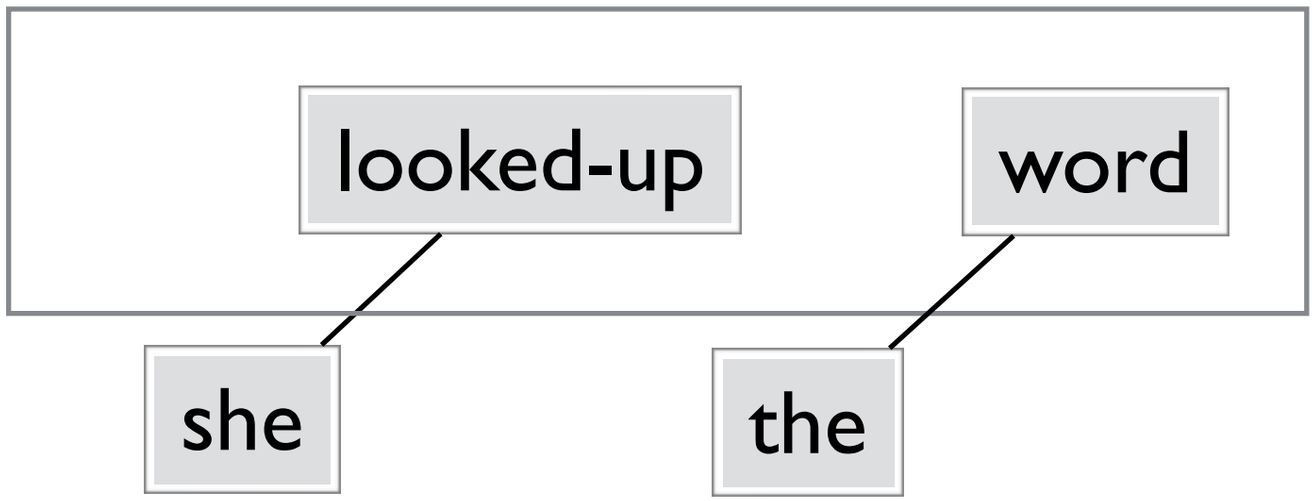
UPPSALA
UNIVERSITET

Parsing Example 1



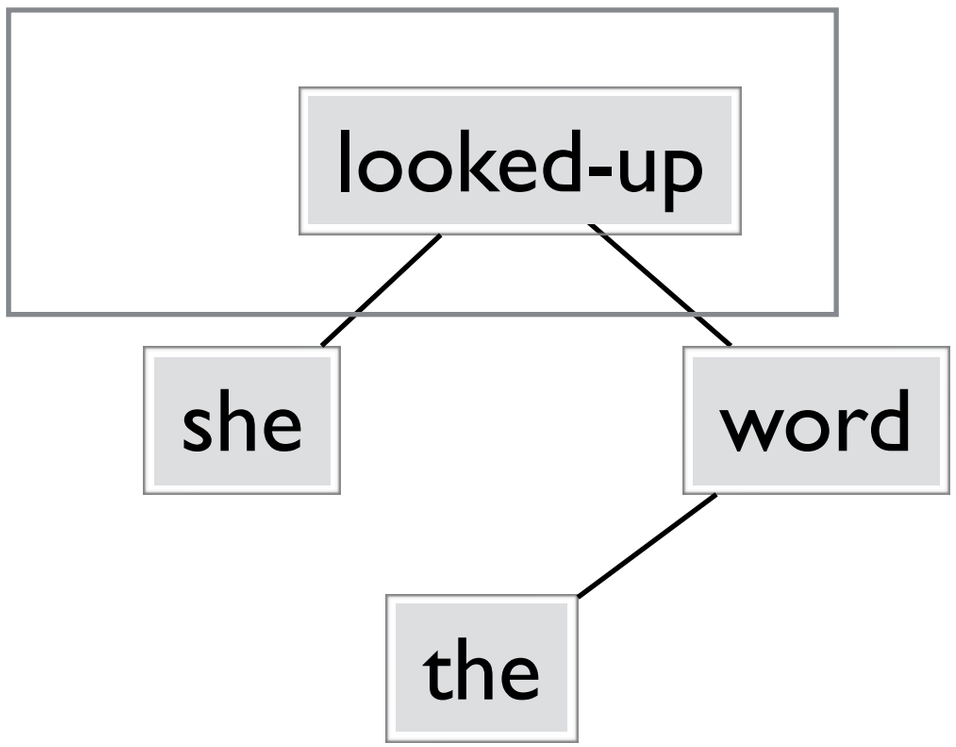
UPPSALA
UNIVERSITET

Parsing Example 1



UPPSALA
UNIVERSITET

Parsing Example 1



UPPSALA
UNIVERSITET

Parsing Example 2

she	looked	the	word	up

UPPSALA
UNIVERSITET

Parsing Example 2

she

looked | the | word | up

UPPSALA
UNIVERSITET

Parsing Example 2

she looked

the word up

UPPSALA
UNIVERSITET

Parsing Example 2

UPPSALA UNIVERSITET

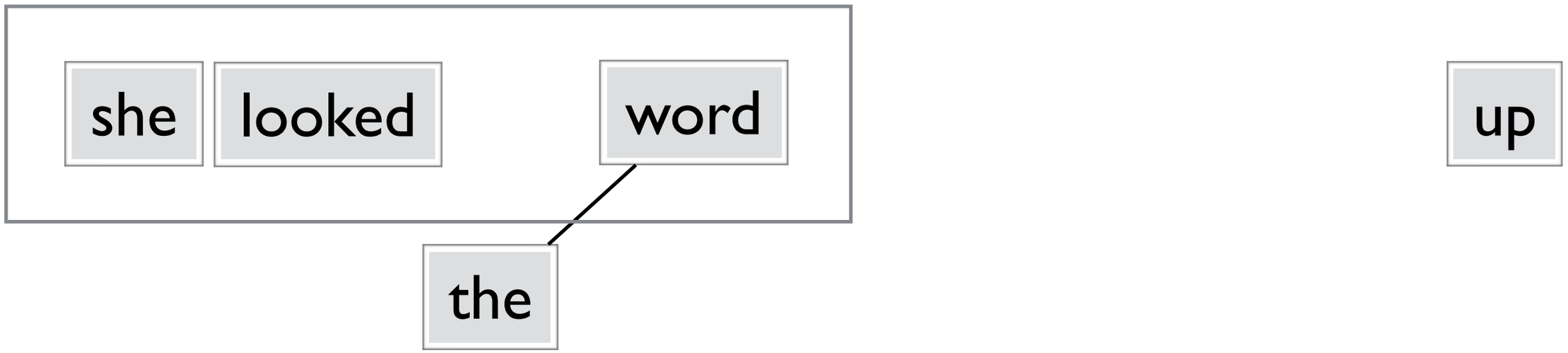
Parsing Example 2

she | looked | the | word

up

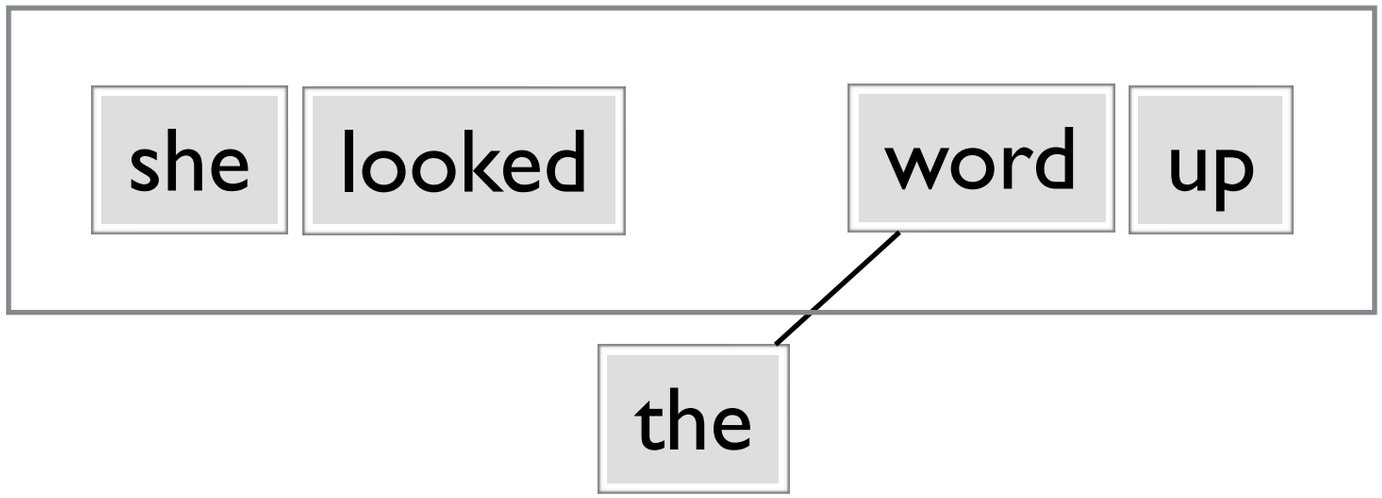
UPPSALA
UNIVERSITET

Parsing Example 2



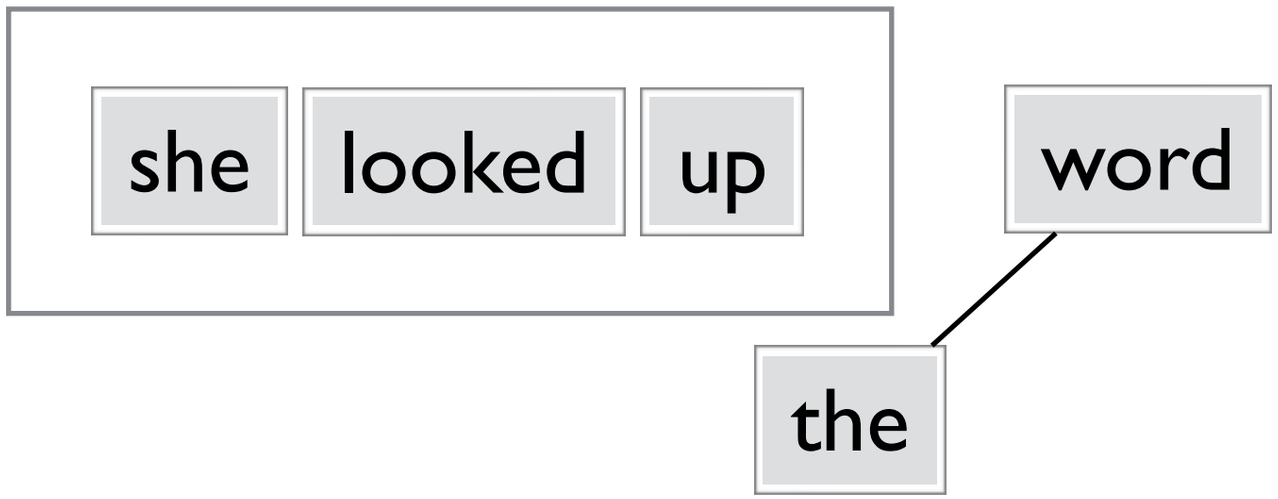
UPPSALA
UNIVERSITET

Parsing Example 2



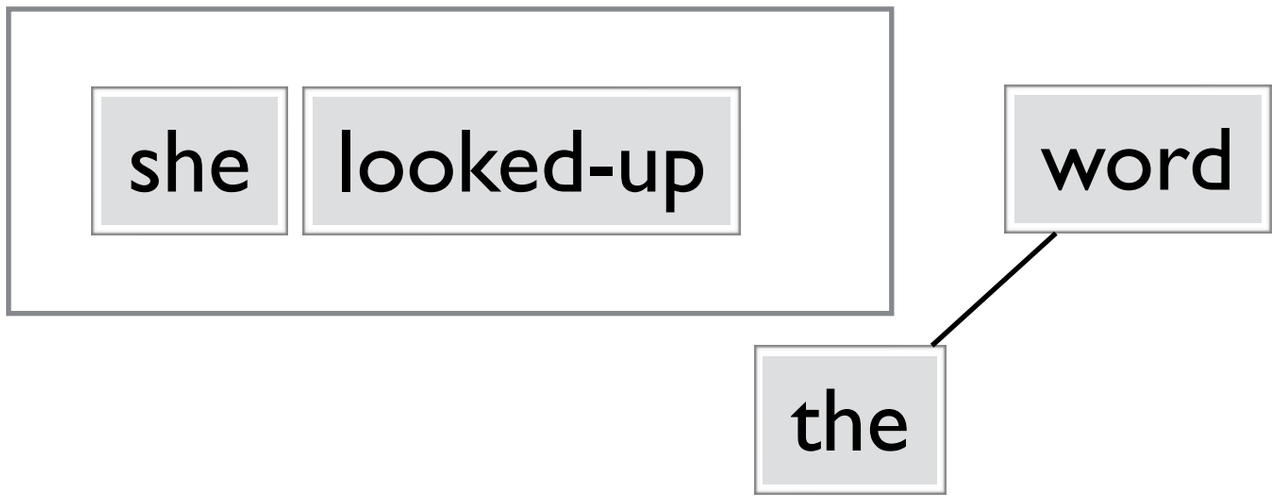
UPPSALA
UNIVERSITET

Parsing Example 2



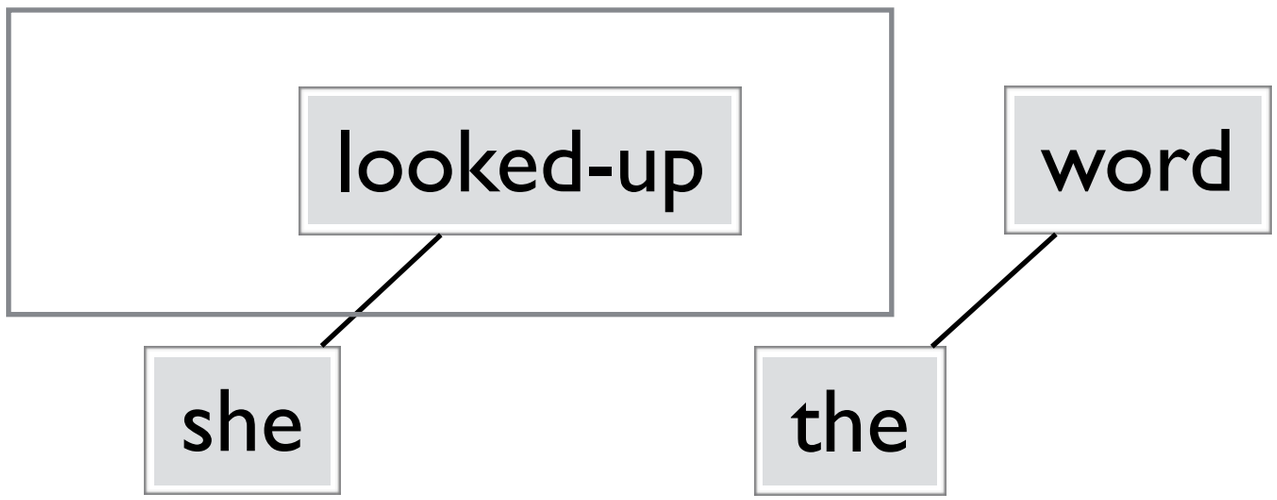
UPPSALA UNIVERSITET

Parsing Example 2



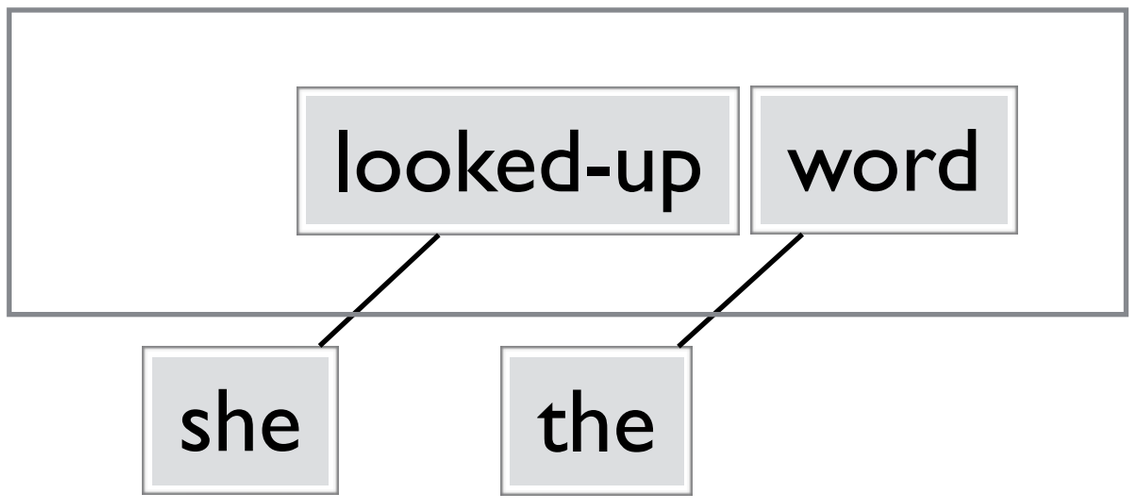
UPPSALA
UNIVERSITET

Parsing Example 2



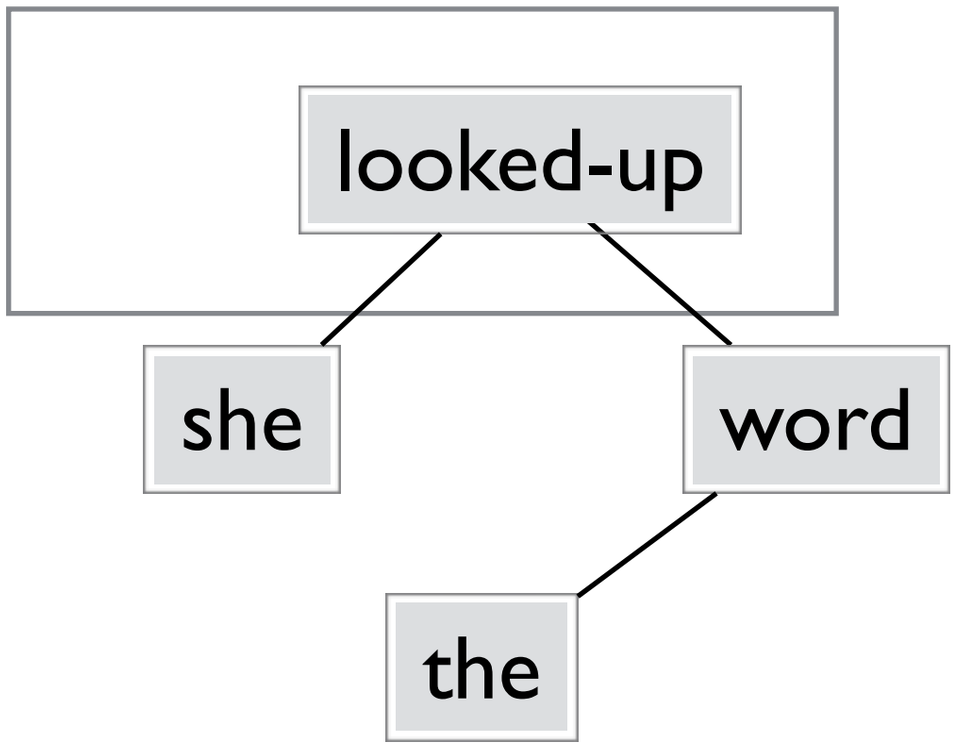
UPPSALA
UNIVERSITET

Parsing Example 2



UPPSALA UNIVERSITET

Parsing Example 2



UPPSALA
UNIVERSITET

Why stop here?

- Keep tokens and nodes ontologically distinct
  - Tokens live in the buffer and can be consumed by transitions
  - Nodes live on the stack and can be created by transitions
- Complete flexibility
  - Consume 1 token (*du*), create *m* nodes (*de le*)
  - Consume *m* tokens (*à cause de*), create 1 node (*à-cause-de*)
  - Consume 0 tokens, create 1 node (ellipsis?)
  - Consume 1 token, create 0 nodes (punctuation?)

UPPSALA UNIVERSITET

Discussion

- Open issues
  - How to integrate morphology?
  - How to do feature extraction?
  - How to make use of lexical resources?
- MWEs do not form a homogeneous class
  - Preprocessing for unambiguous fixed expressions?
  - Chunking for semi-flexible (possibly discontiguous) expressions?
  - Full dependency parsing for compositional expressions?