UPPSALA
UNIVERSITET

Enhanced Universal Dependencies

Joakim Nivre

Uppsala University
Department of Linguistics and Philology

Based on collaborative work with Filip Ginter, Jenna Kanerva, Paola Marongiu, Simonetta Montemagni, Sebastian Schuster and Maria Simi

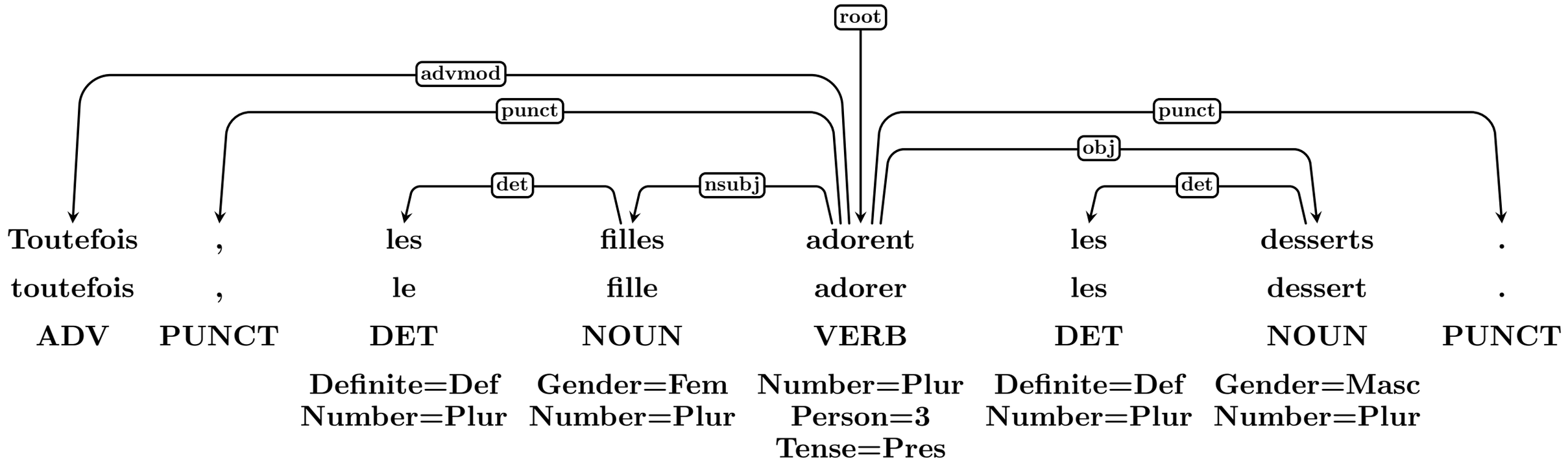
Goals of UD

Cross-linguistically consistent morphosyntactic annotation

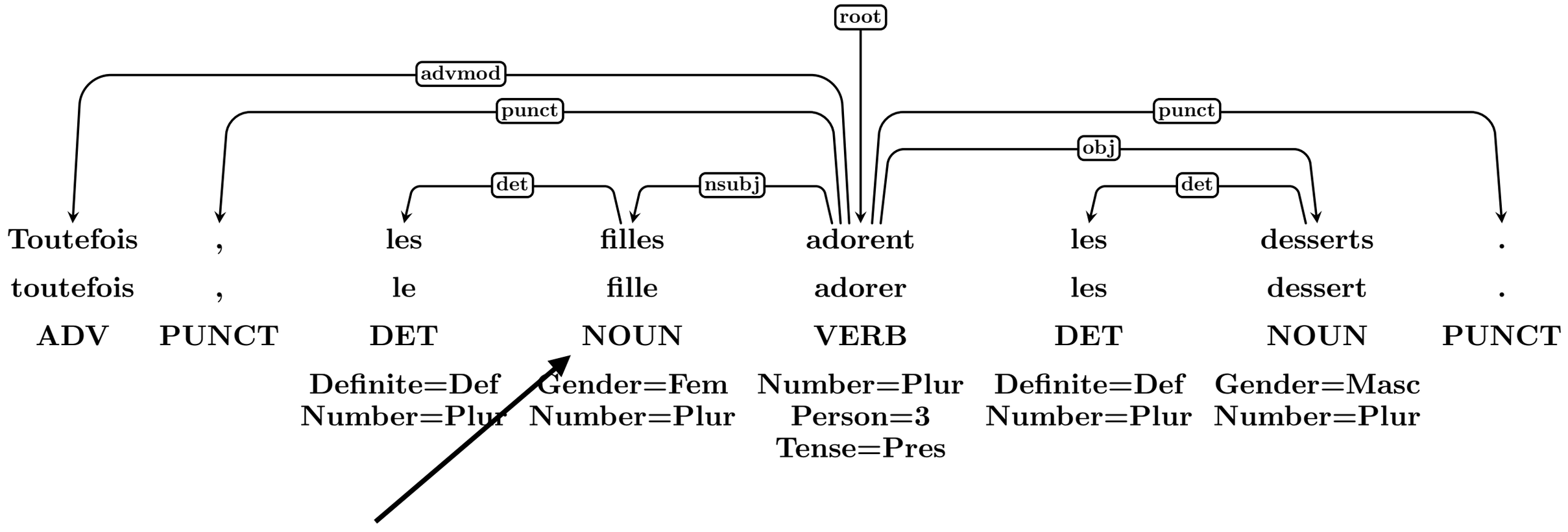
Facilitate multilingual research in NLP and linguistics

- Meaningful linguistic analysis across languages
- Syntactic parsing in cross-lingual settings
- NLP systems for multiple languages
- Facilitate resource-building for new languages

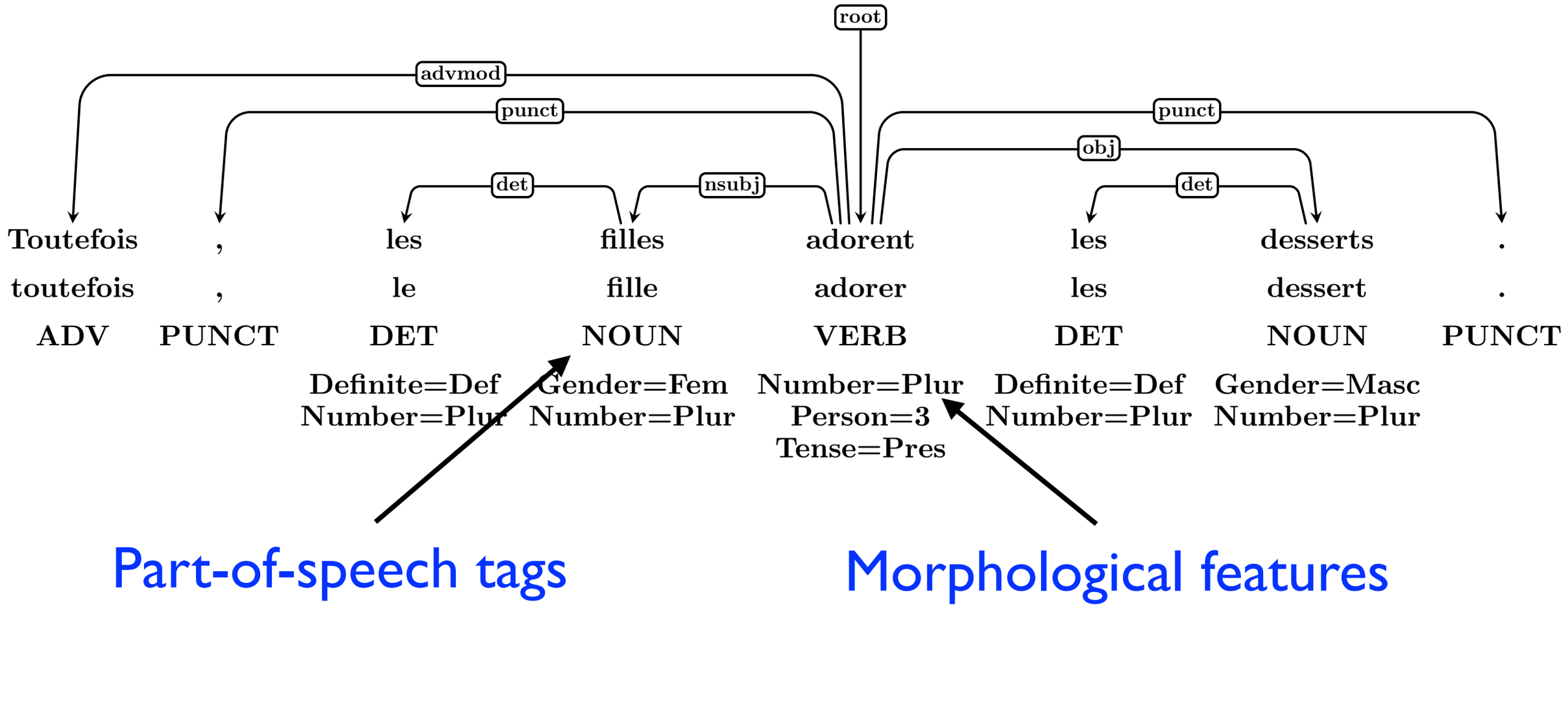
Annotation in UD



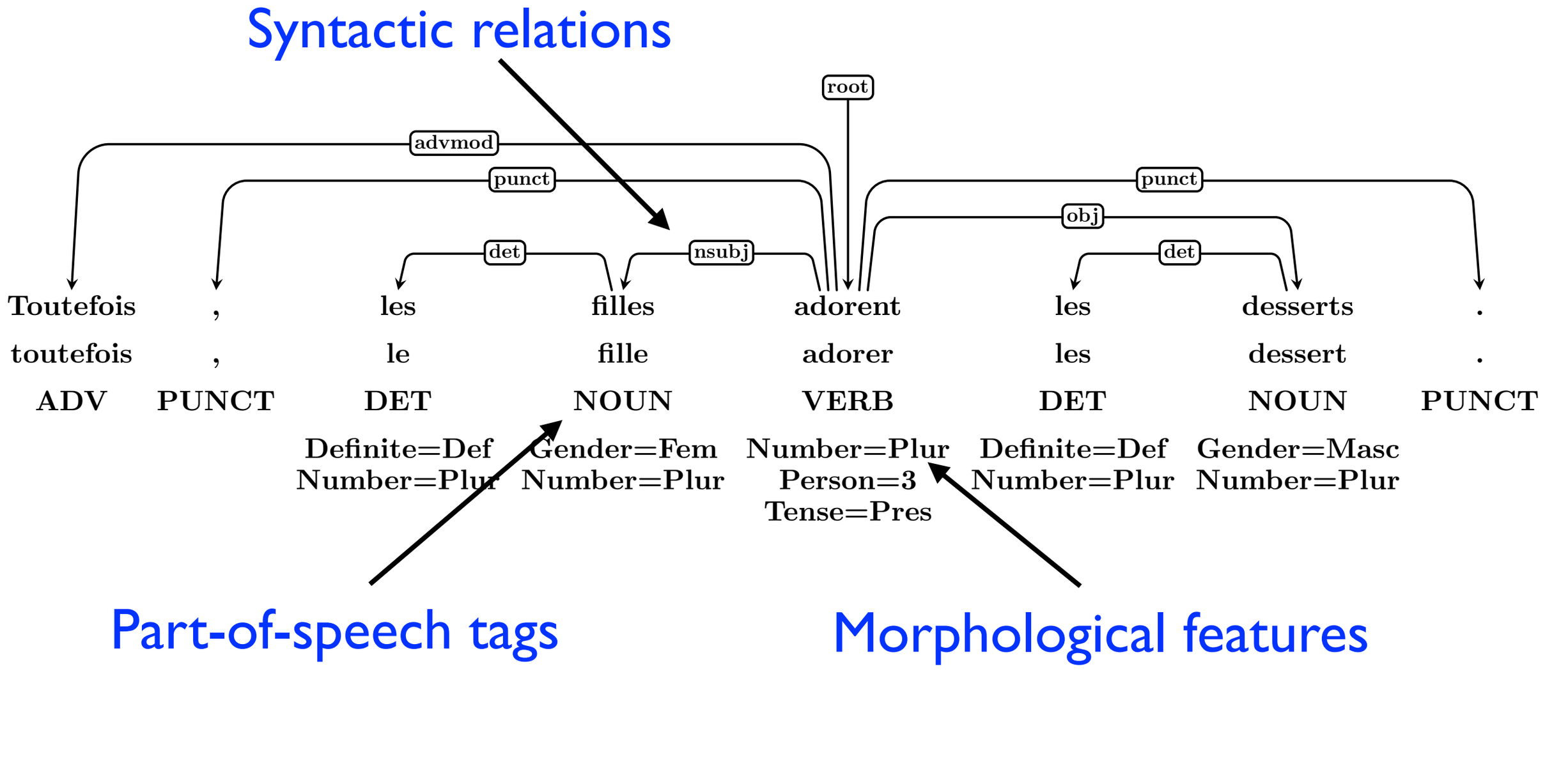
Annotation in UD



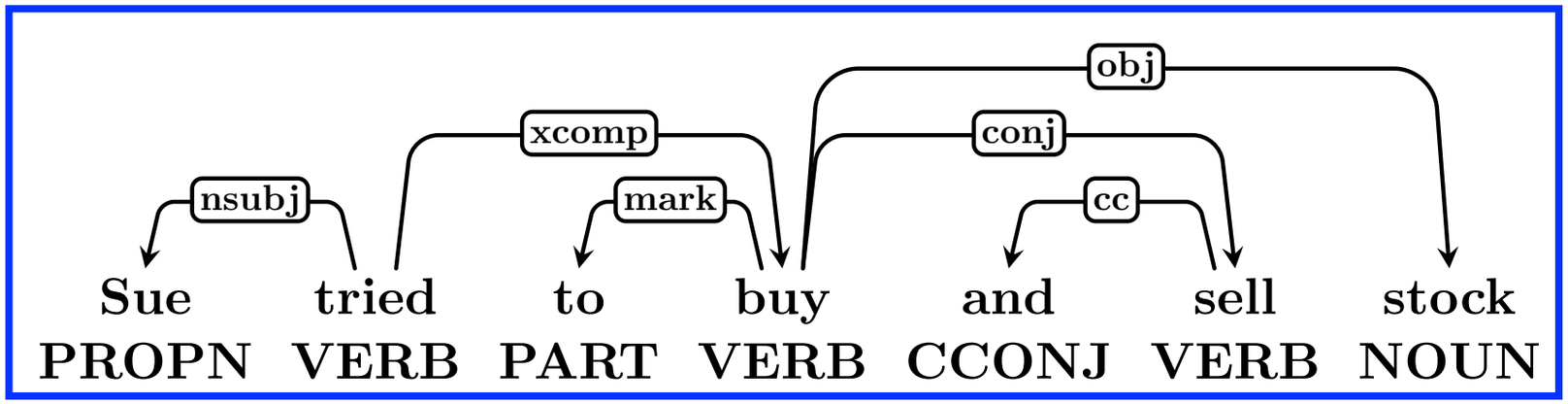
Annotation in UD



Annotation in UD



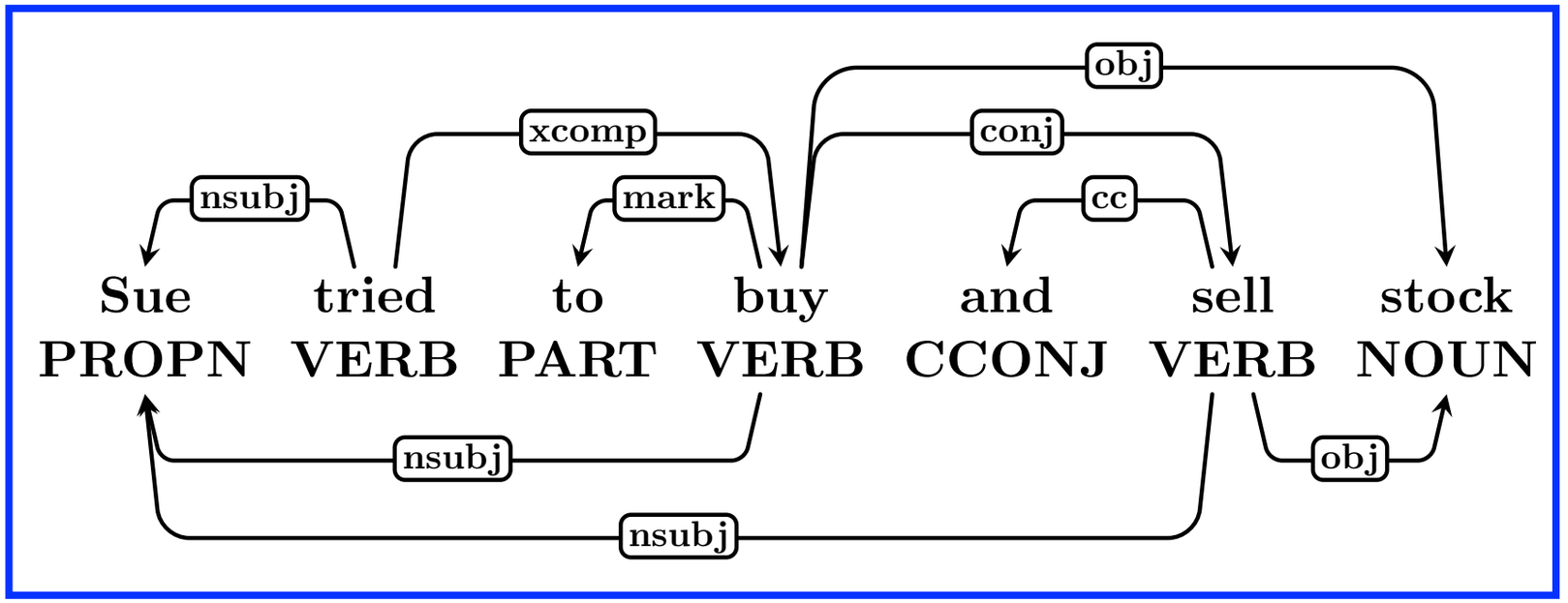
Basic Dependencies



Properties of the basic syntactic representation:

- Spanning tree over the words of the sentence
- One-to-one mapping from words to nodes – no empty nodes
- Every word related to (at most) one other word
- Underspecified representation of predicate-argument structure
- Suitable for parsing but not for (all) downstream applications

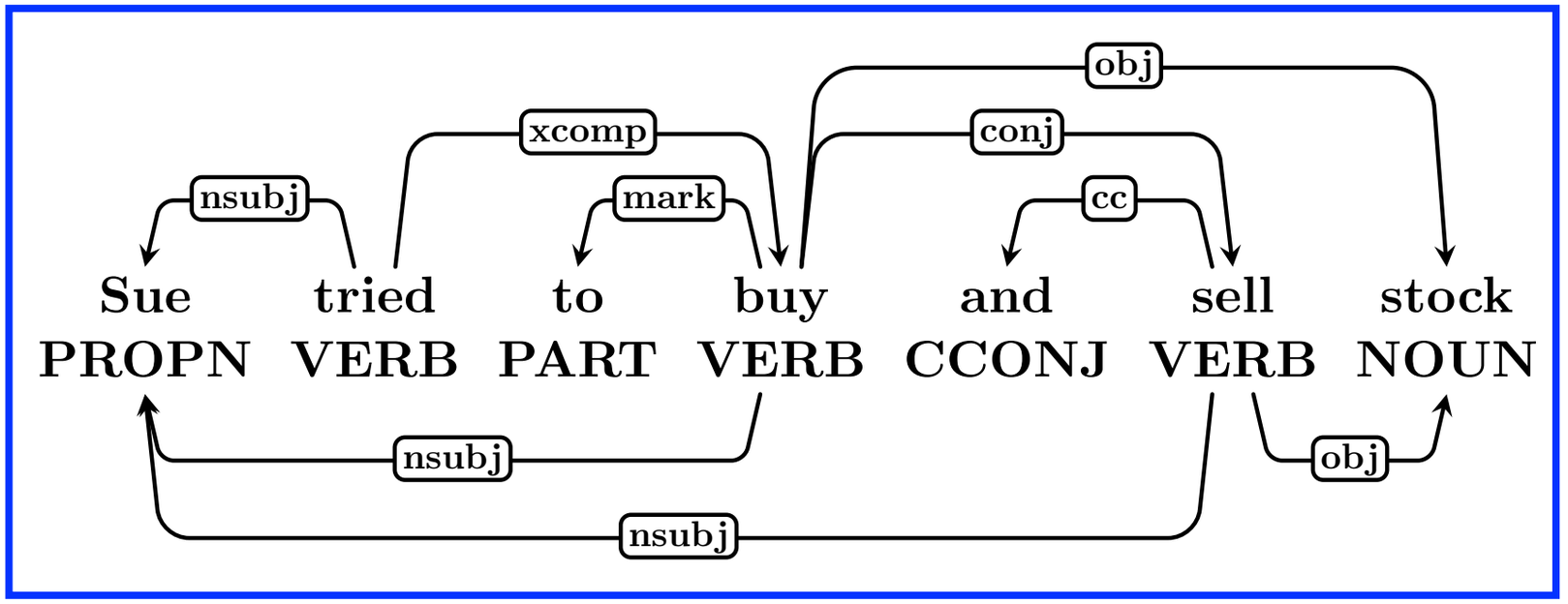
Enhanced Dependencies



Properties of the enhanced syntactic representation:

- General graph structure – not a tree (and not spanning)
- Partial mapping from words to nodes – and vice versa
- Not a monotonic extension of basic dependencies
- Disambiguates aspects of predicate-argument structure
- Collapses paths into single arcs – for practical convenience

Enhanced Dependencies



Properties of the enhanced syntactic representation:

- General graph structure – not a tree (and not spanning)
- Partial mapping from words to nodes – and vice vers
- Not a monotonic extension of basic depende
- Disambiguates aspects of predicate-arg tructure
- Collapses paths into single arcs – ctical convenience

Required by UDepLambda

Enhancements in UD v2

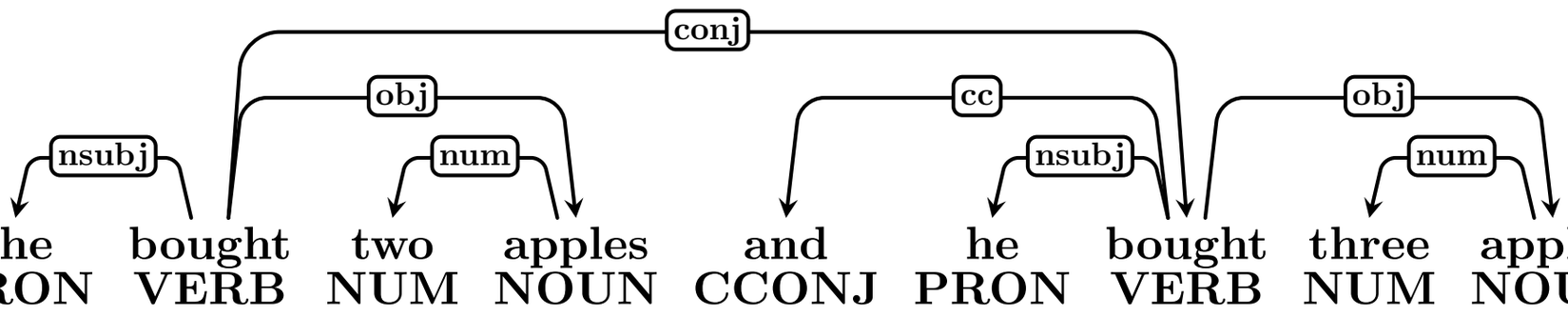
1. Null nodes for ellided predicates
2. Shared heads and dependents in coordination
3. Added subject relations in control and raising
4. Coreference in relative clause constructions
5. Augmented modifier relations

Ellipsis

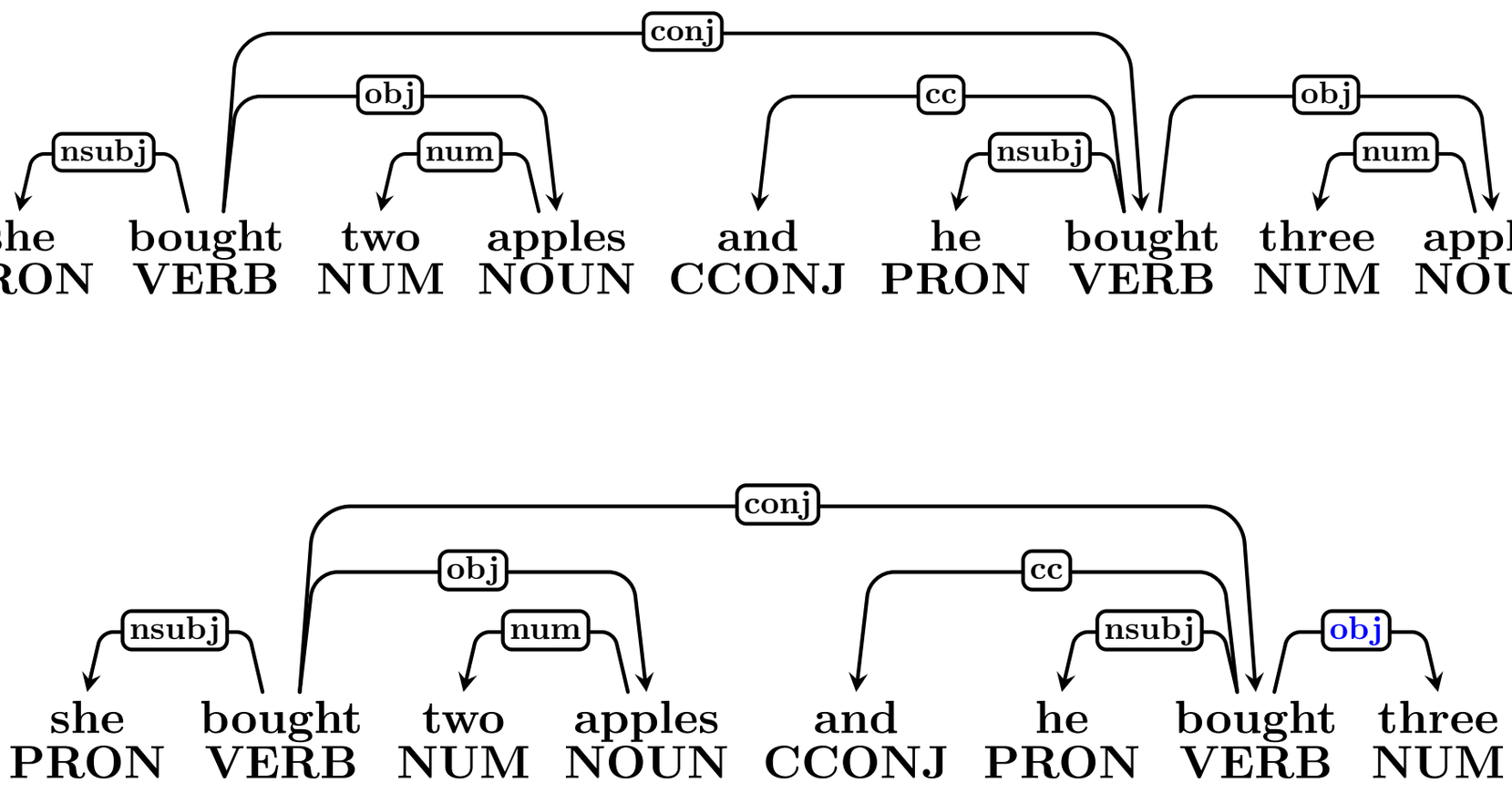
Ellipsis in basic dependencies:

1. If the elided element has no overt dependents, we do nothing.
2. If the elided element has overt dependents, we promote one of these to take the role of the head.
3. If the elided element is a predicate and the promoted element a core argument or modifier, we use the orphan relation to attach other non-functional dependents to the promoted head.

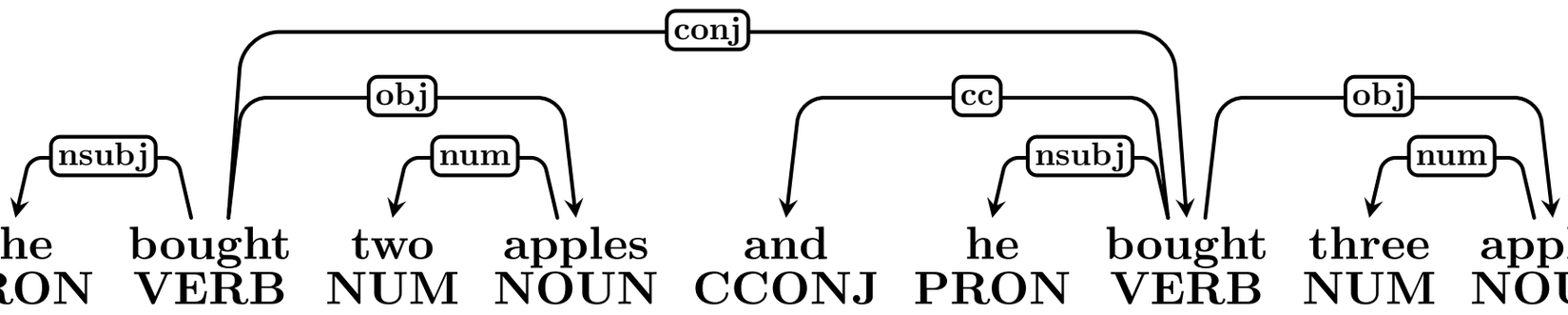
Ellipsis

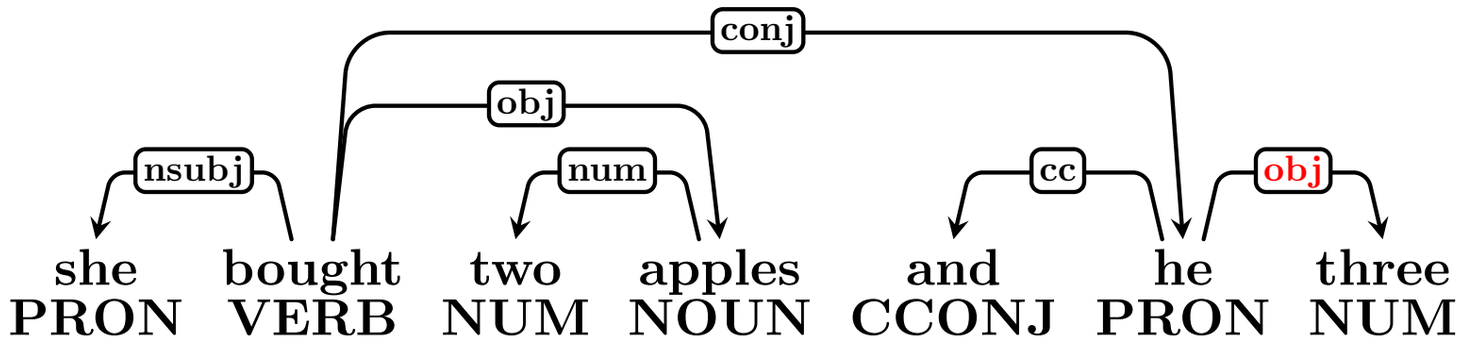

conj
obj
cc
obj
nsubj
num
nsubj
num
she bought two apples and he bought three apples
PRON VERB NUM NOUN CCONJ PRON VERB NUM NOUN

Ellipsis

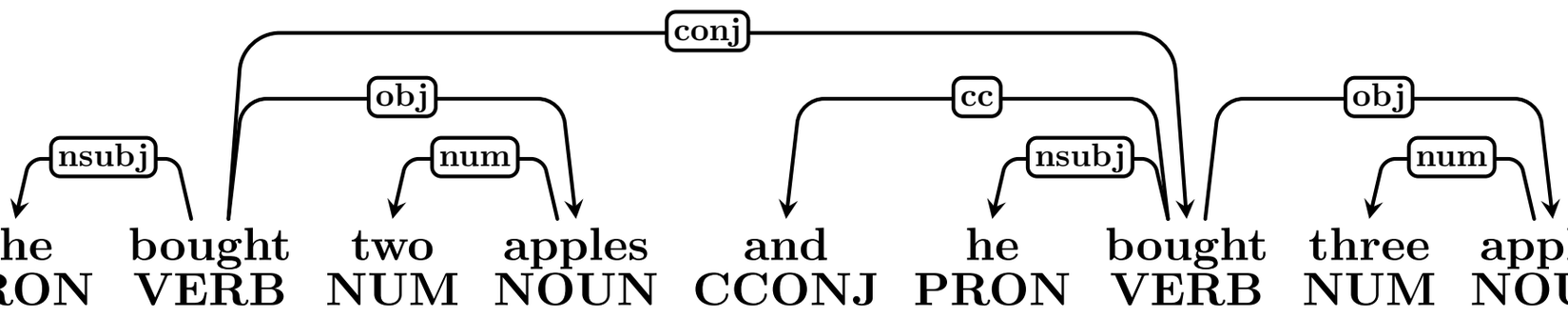

conj
obj
cc
obj
nsubj
num
nsubj
num
she bought two apples and he bought three apples
PRON VERB NUM NOUN CCONJ PRON VERB NUM NOUN
conj
obj
cc
nsubj
num
nsubj
obj
she bought two apples and he bought three
PRON VERB NUM NOUN CCONJ PRON VERB NUM

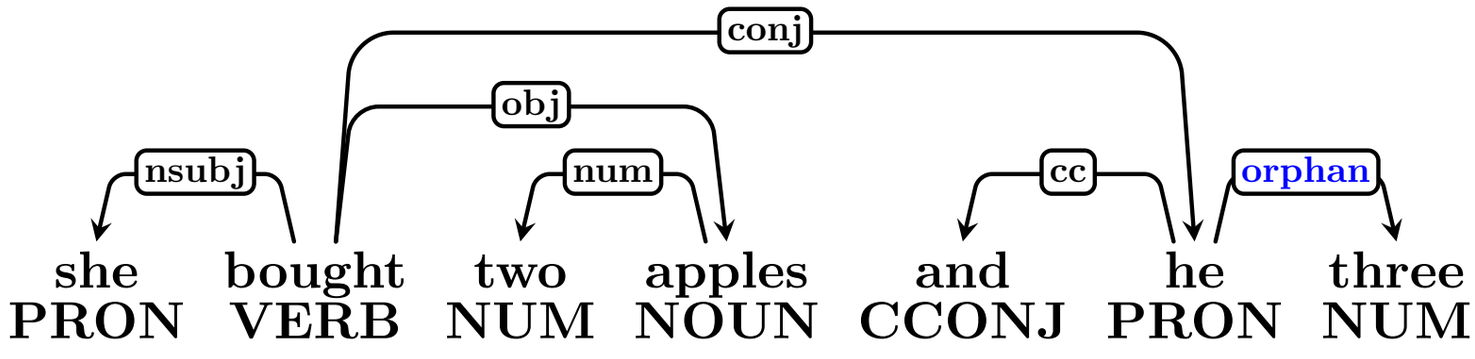
Ellipsis


conj
obj
cc
obj
nsubj
num
nsubj
num
she bought two apples and he bought three apples
PRON VERB NUM NOUN CCONJ PRON VERB NUM NOUN


conj
obj
nsubj
num
cc
obj
she bought two apples and he three
PRON VERB NUM NOUN CCONJ PRON NUM

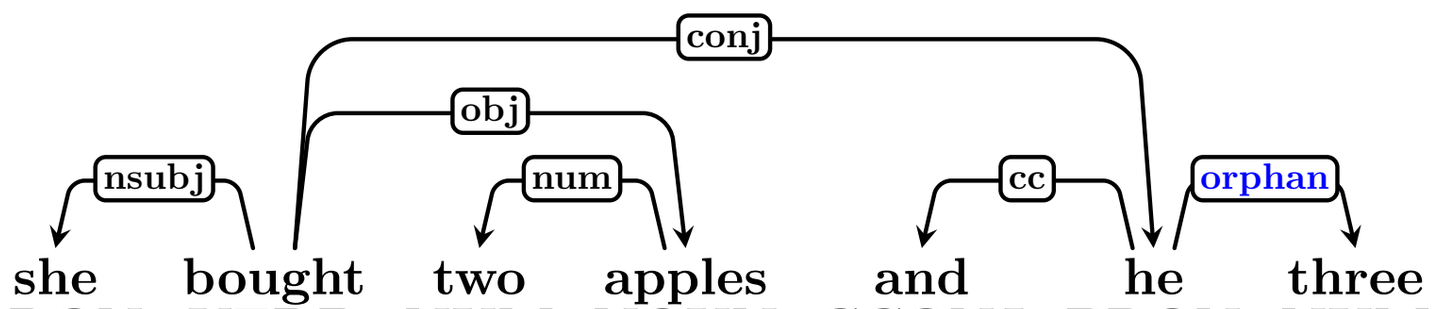
Ellipsis


conj
obj
cc
obj
nsubj
num
nsubj
num
she bought two apples and he bought three apples
PRON VERB NUM NOUN CCONJ PRON VERB NUM NOUN

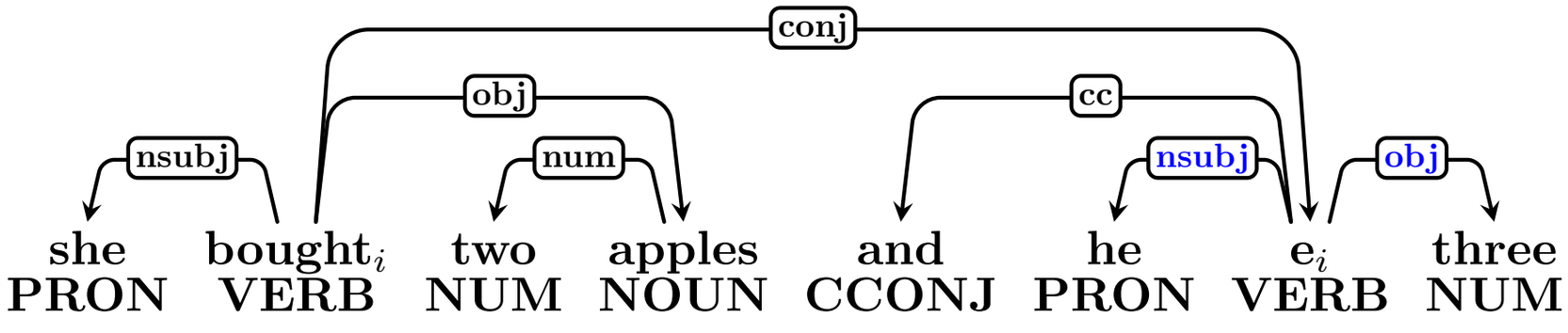

conj
obj
nsubj
num
cc
orphan
she bought two apples and he three
PRON VERB NUM NOUN CCONJ PRON NUM

Ellipsis

Basic

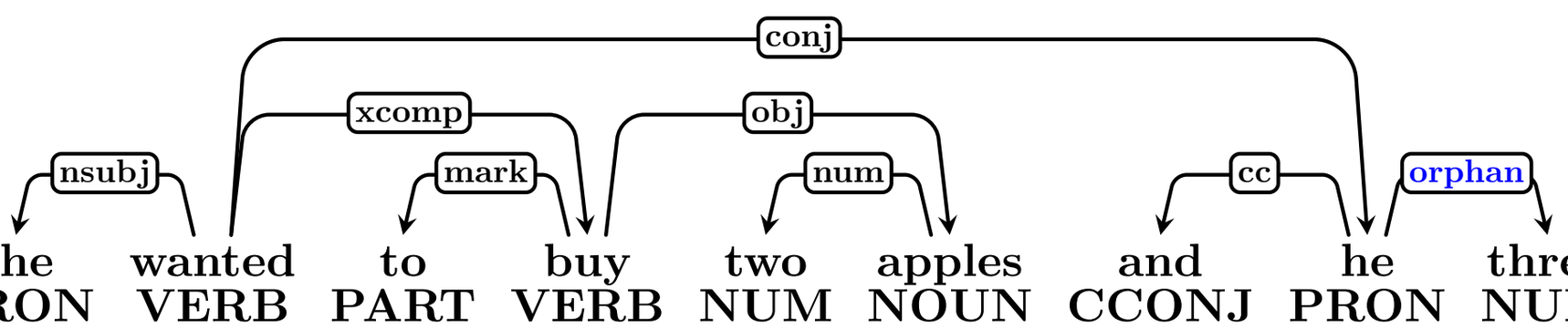


Enhanced

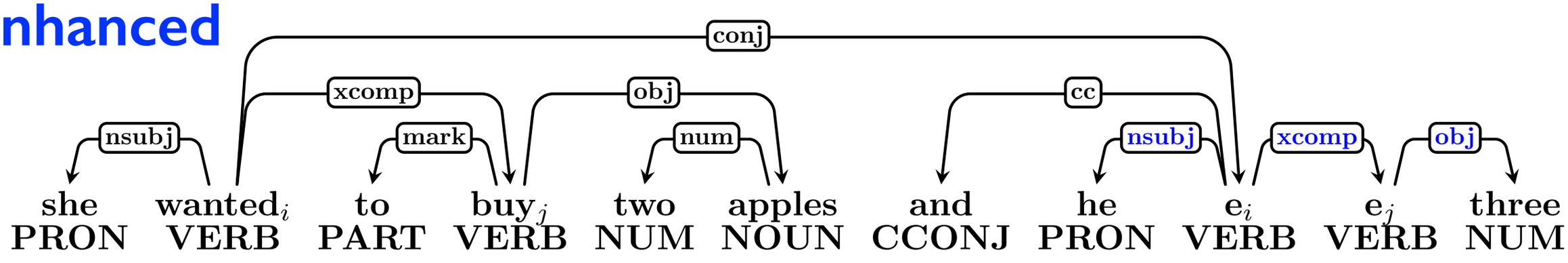


Ellipsis

Basic



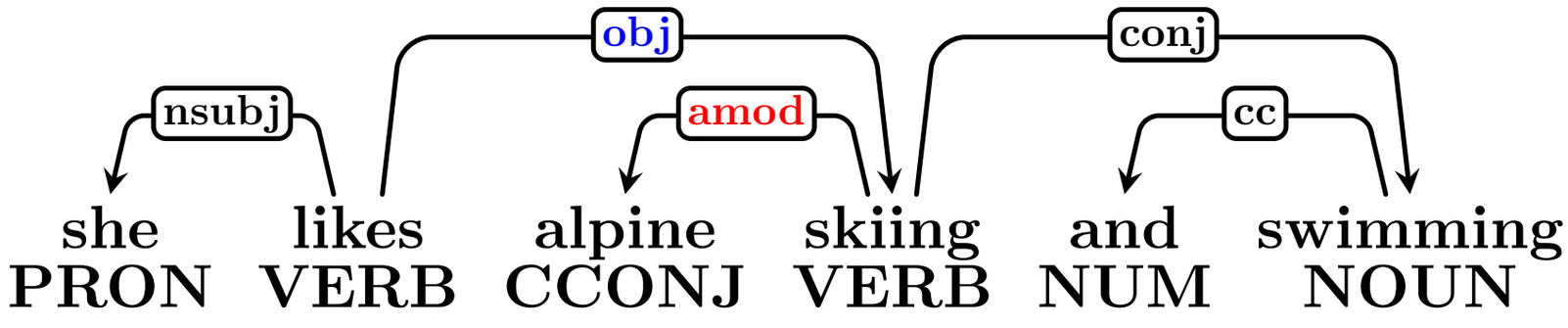
Enhanced



Coordination

Basic dependencies underspecify dependency relations into and out of coordinated phrases:

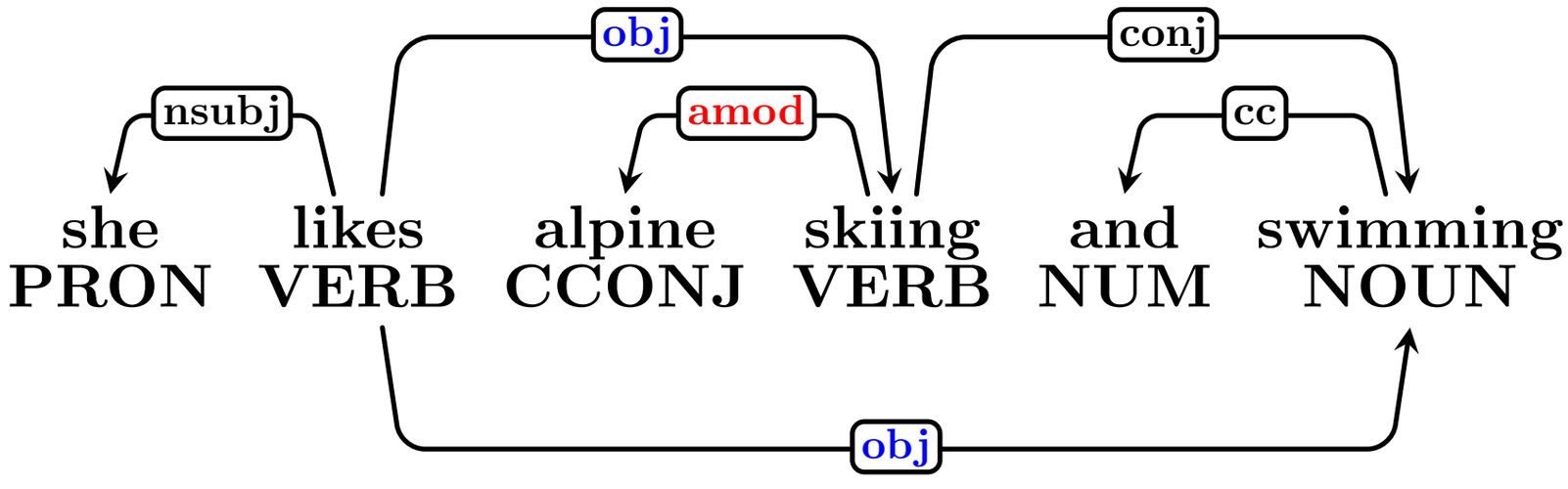
1. The shared head is attached only to the first conjunct (unambiguous).
2. Shared dependents are attached only to the first conjunct (ambiguous).



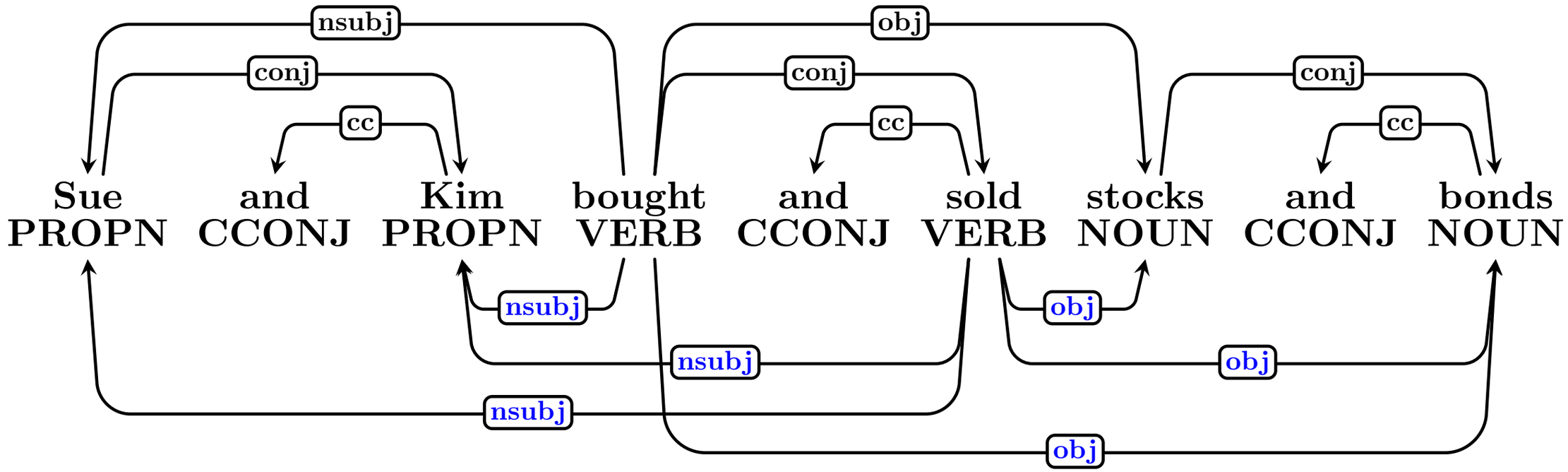
Coordination

Basic dependencies underspecify dependency relations into and out of coordinated phrases:

1. The shared head is attached only to the first conjunct (unambiguous).
2. Shared dependents are attached only to the first conjunct (ambiguous).



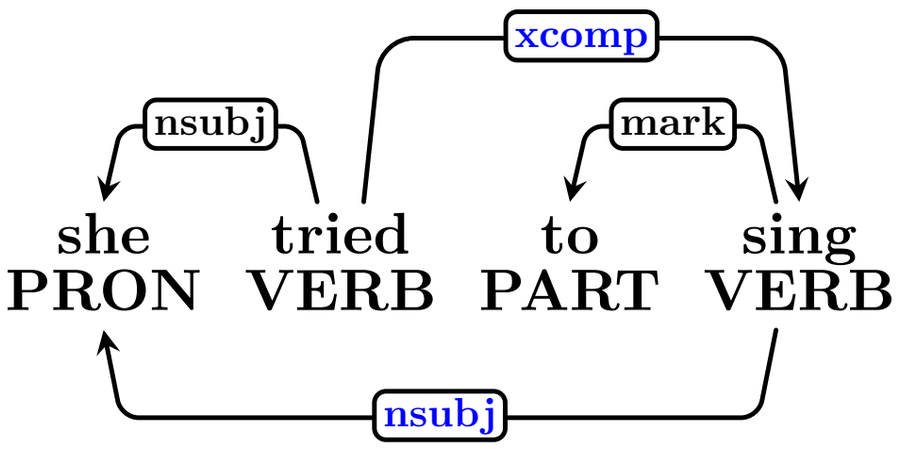
Coordination

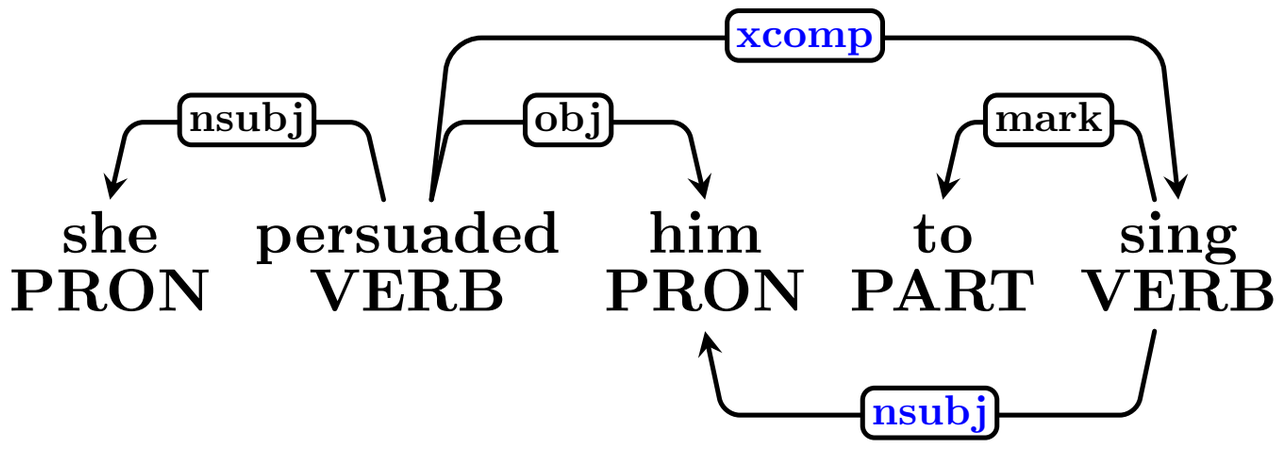


Raising and Control

Enhanced dependencies add an explicit subject relation out of open clausal complements (xcomp)

1. Subject relation to raised subject under raising verbs
2. Subject relation to controller under control verbs

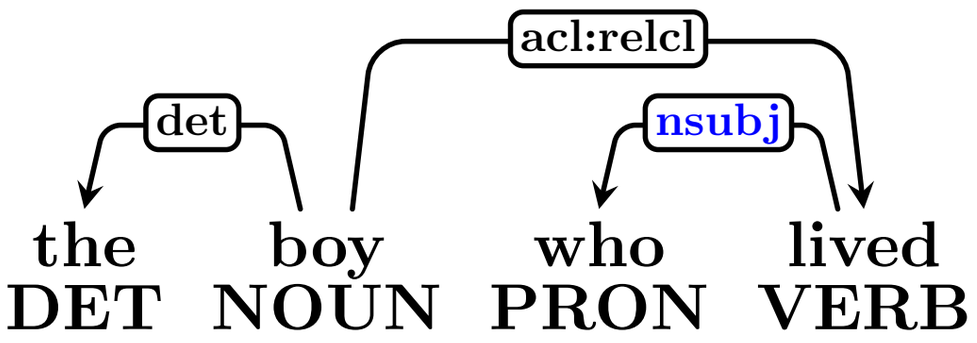


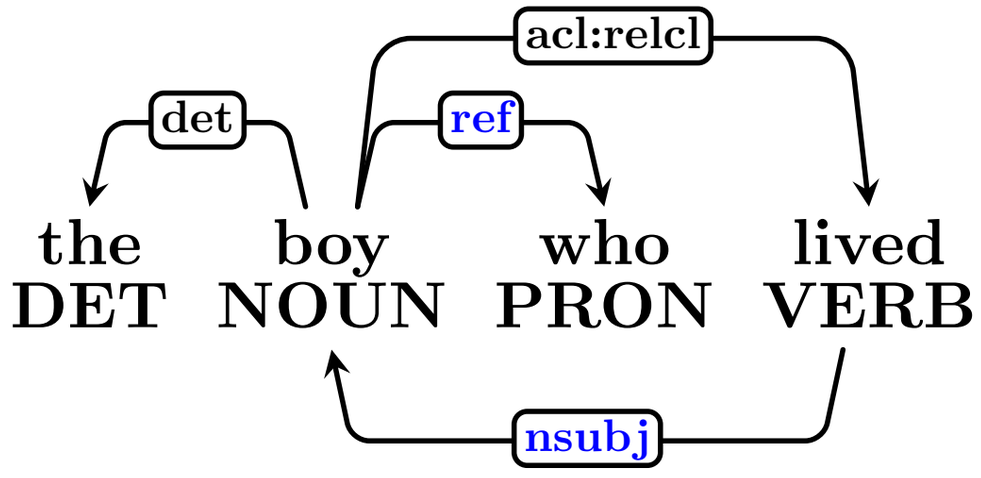


Relative Clauses

Enhanced dependencies adds two relations:

1. Co-reference relation from antecedent to relative pronoun
2. Core argument relation from relative clause predicate to antecedent

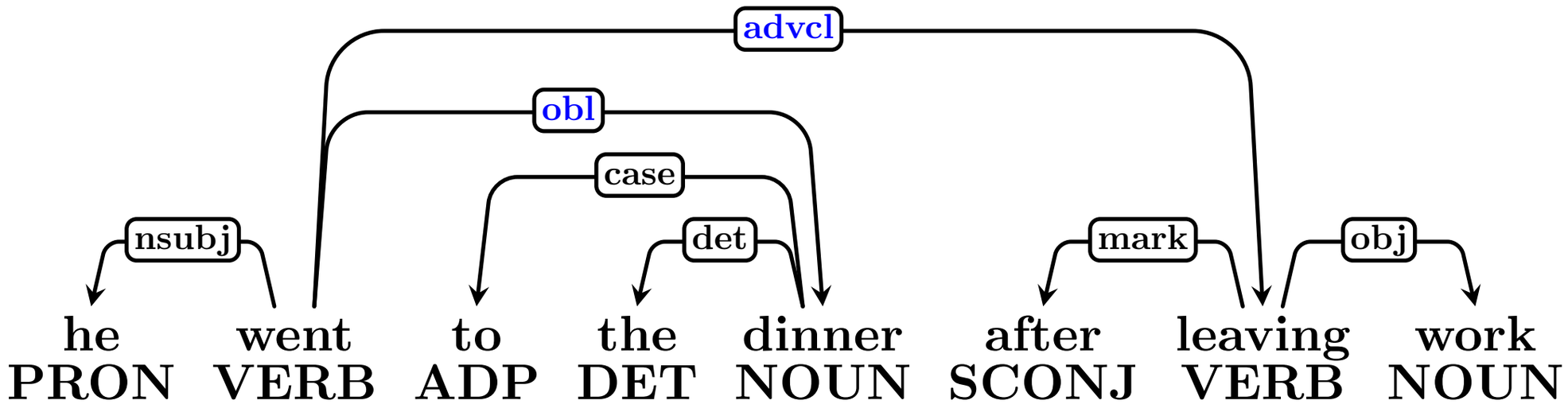




Augmented Labels

Enhanced dependencies augments relation labels:

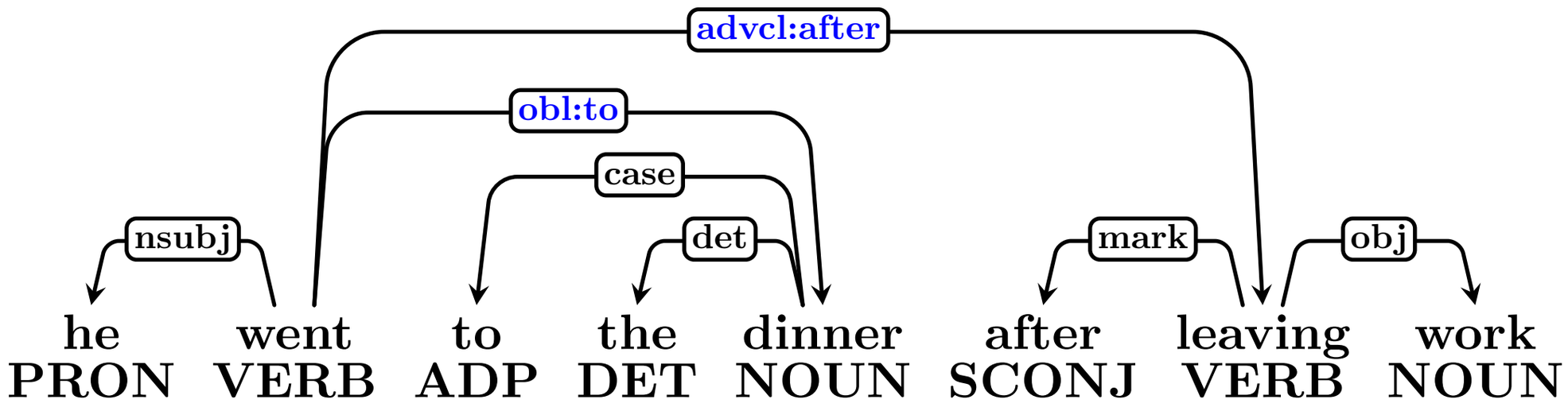
1. Adds case markers (adpositions) for obl and nmod
2. Adds markers (conjunctions) for advcl



Augmented Labels

Enhanced dependencies augments relation labels:

1. Adds case markers (adpositions) for obl and nmod
2. Adds markers (conjunctions) for advcl



Enhanced UD Treebanks

Enhanced UD Treebanks

UD v2.2: 5 out of 102 treebanks

- English (EWT)
- Finnish (TDT, PUD)
- Latvian (LVTB)
- Russian (SynTagRus)

Enhanced UD Treebanks

UD v2.2: 5 out of 102 treebanks

- English (EWT)
- Finnish (TDT, PUD)
- Latvian (LVTB)
- Russian (SynTagRus)

Case study on bootstrapping enhanced dependencies

- Joint work with Sebastian Schuster, Filip Ginter, Jenna Kanerva, Paola Marongiu, Simonetta Montemagni, Maria Simi

Case Study

Case Study

Two enhancers:

- Stanford – rule-based system developed for English
- Turku – data-driven system trained on Finnish

Case Study

Two enhancers:

- Stanford – rule-based system developed for English
- Turku – data-driven system trained on Finnish

Two target languages:

- Swedish
- Italian

Case Study

Two enhancers:

- Stanford – rule-based system developed for English
- Turku – data-driven system trained on Finnish

Two target languages:

- Swedish
- Italian

Three enhanced dependency types:

- Added subject relations in raising and control constructions
- Shared heads and dependents in coordination
- Null nodes for elided predicates

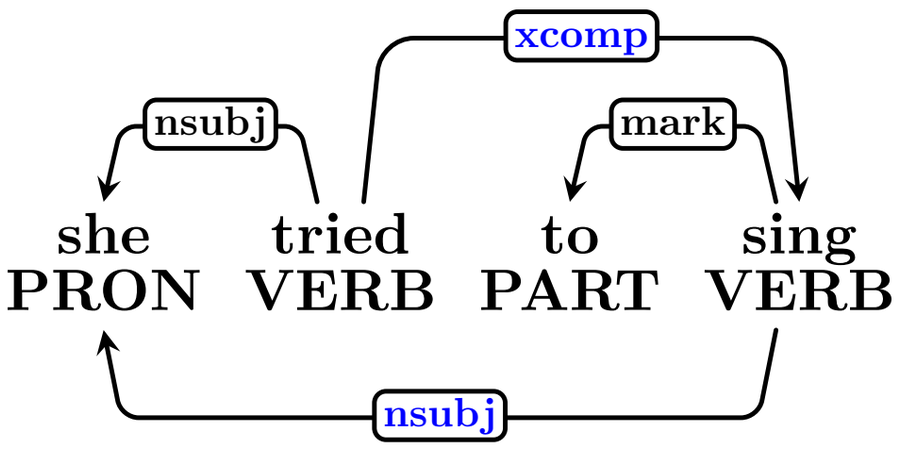
The Stanford System

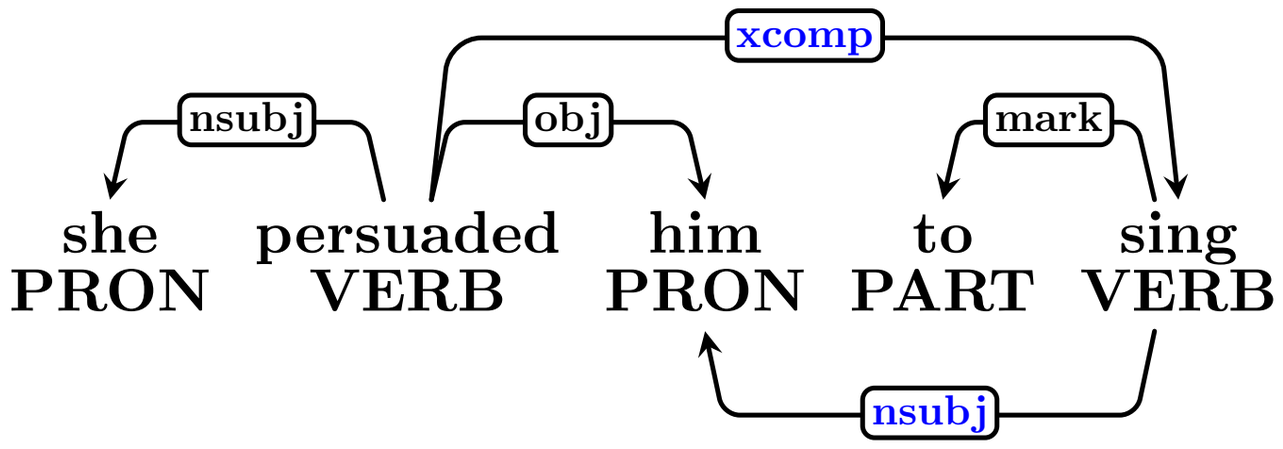
- Based on English system by Schuster and Manning (2016)
- Pattern matching to detect structures to enhance
- Heuristics to predict enhanced dependencies
- Novel method for ellipsis by Schuster et al. (2018)

Stanford: Subjects

From any node attached as an xcomp to a higher predicate, add an nsubj dependency to:

1. The obj of the predicate if such a dependent exists
2. The nsubj of the predicate otherwise (raising, subject control)

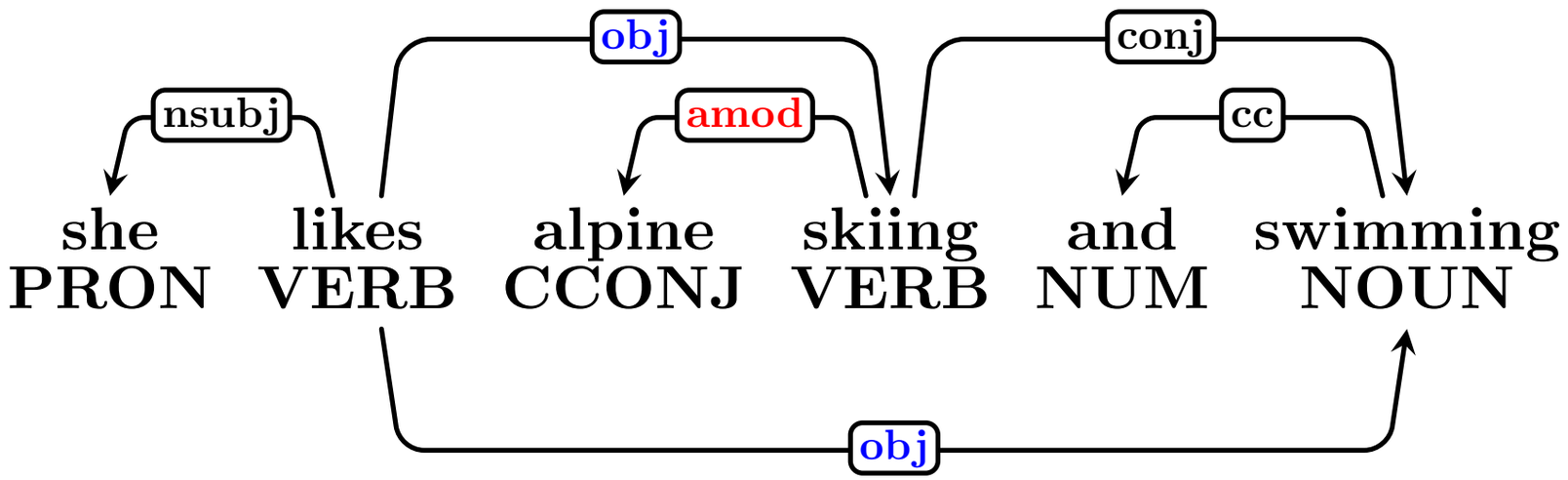




Stanford: Coordination

Two types of added dependencies:

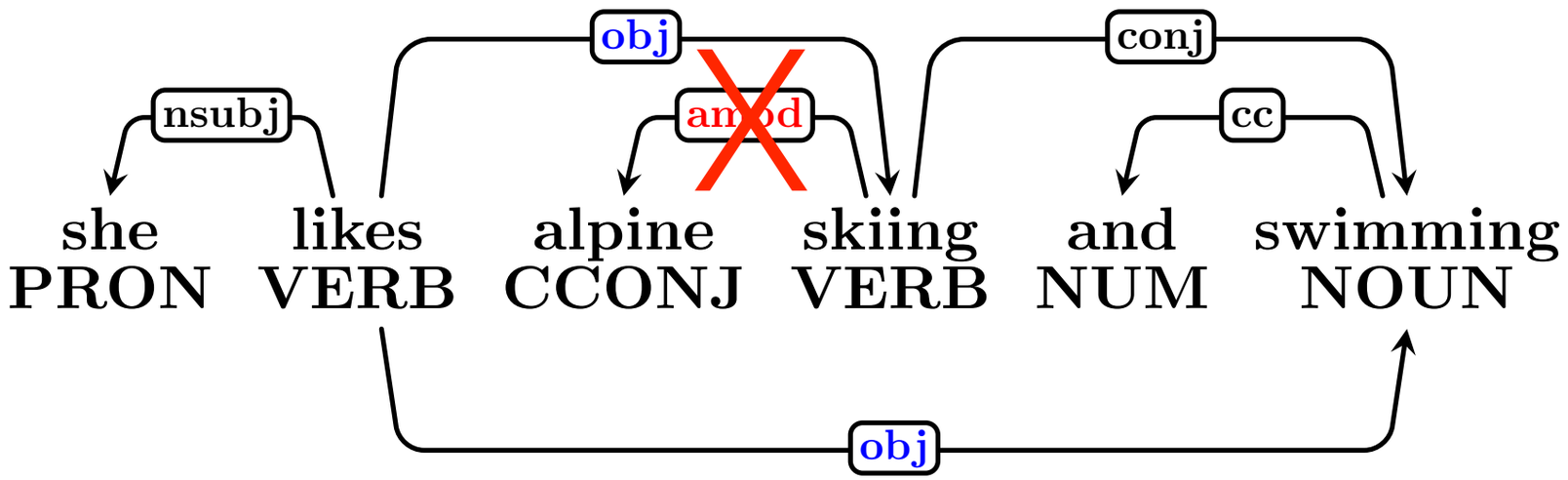
1. Shared heads for all coordinated elements
2. Shared dependents limited to core arguments of conjoined predicates: (i)obj, n/csubj, c/xcomp
3. Aims for high precision (rather than recall)



Stanford: Coordination

Two types of added dependencies:

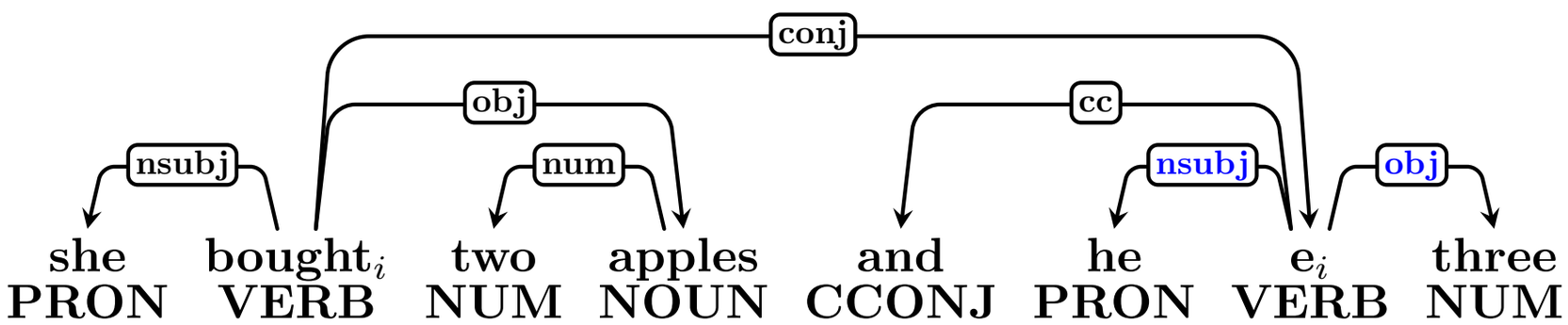
1. Shared heads for all coordinated elements
2. Shared dependents limited to core arguments of conjoined predicates: (i)obj, n/csubj, c/xcomp
3. Aims for high precision (rather than recall)



Stanford: Ellipsis

Enhanced dependencies augments relation labels:

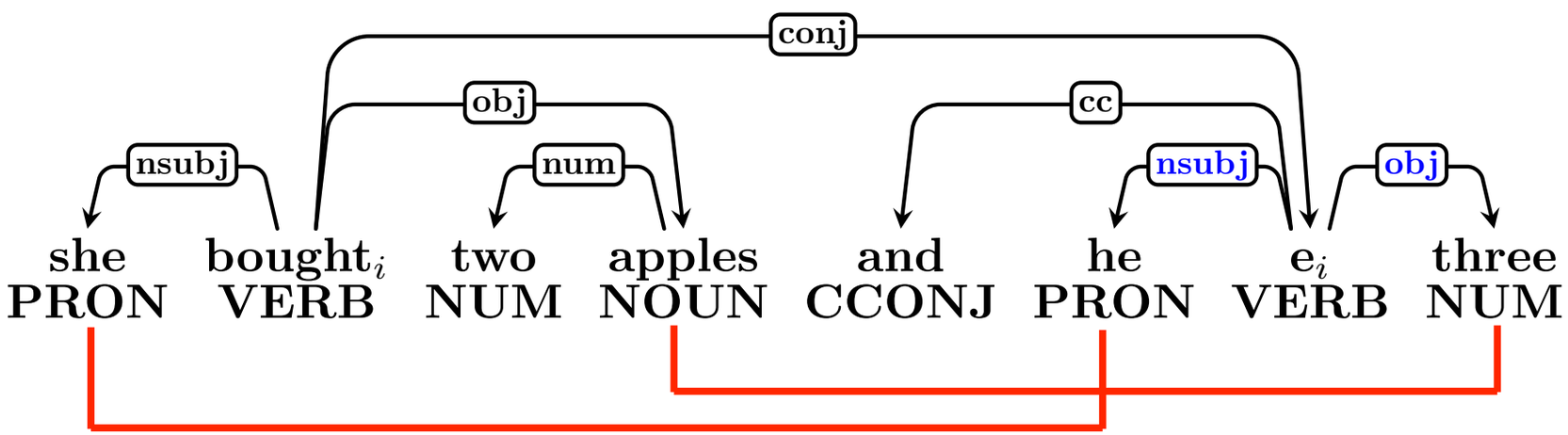
1. Align arguments and modifiers in the complete and gapped clause using similarity of word embeddings
2. Add null predicates corresponding to non-matched items and add dependencies based on matchings



Stanford: Ellipsis

Enhanced dependencies augments relation labels:

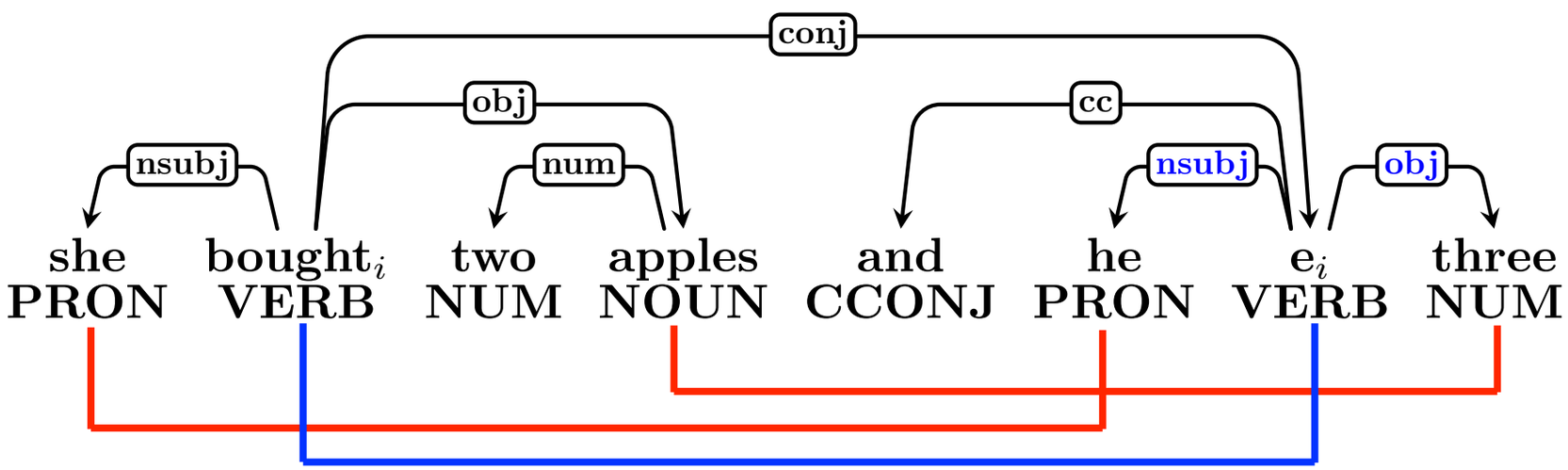
1. Align arguments and modifiers in the complete and gapped clause using similarity of word embeddings
2. Add null predicates corresponding to non-matched items and add dependencies based on matchings



Stanford: Ellipsis

Enhanced dependencies augments relation labels:

1. Align arguments and modifiers in the complete and gapped clause using similarity of word embeddings
2. Add null predicates corresponding to non-matched items and add dependencies based on matchings



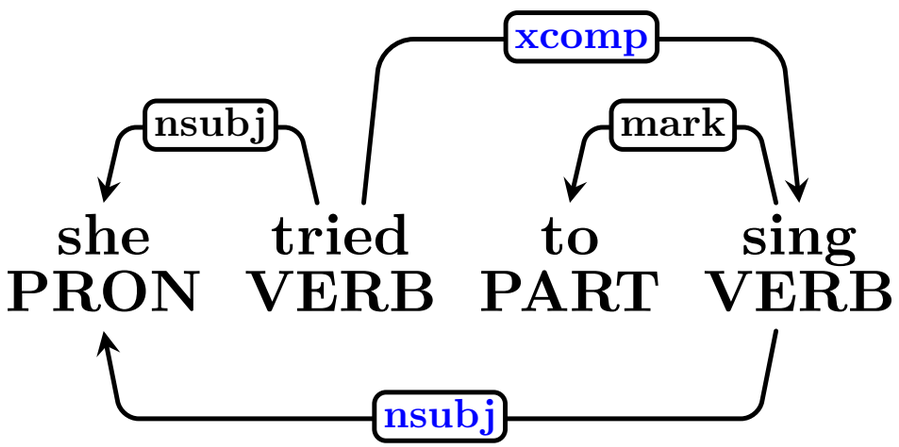
The Turku System

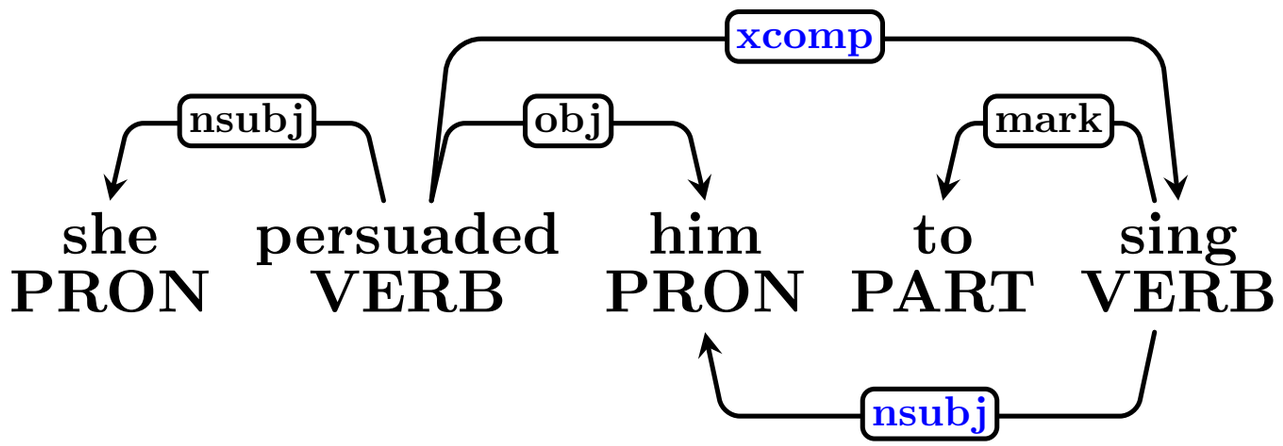
- Based on Finnish system by Nyblom et al. (2016)
- Pattern matching to detect structures to enhance
- SVM classifier selects candidate dependencies
- Language-specific features omitted for generality
- Does not handle null nodes for elided predicates

Turku: Subjects

From any infinitive verb attached as an xcomp to a higher predicate, consider adding an nsubj dependency to the nsubj of the predicate

1. Binary SVM classifier decides if dependency is added or not
2. Object control is not considered at all

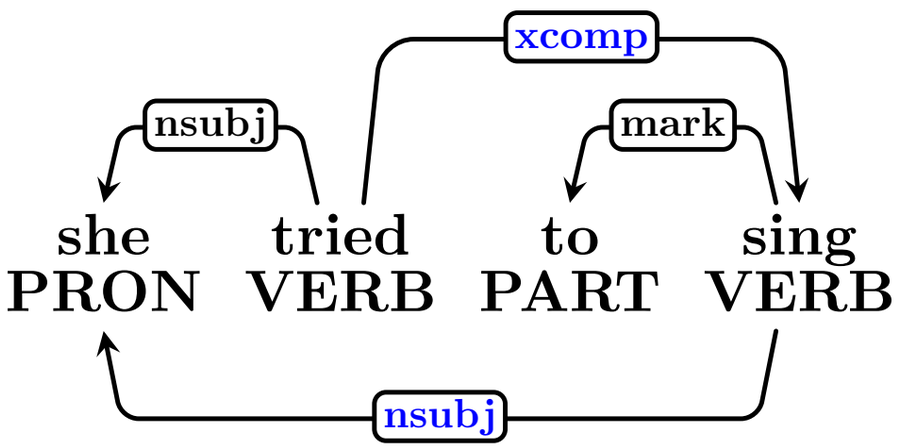


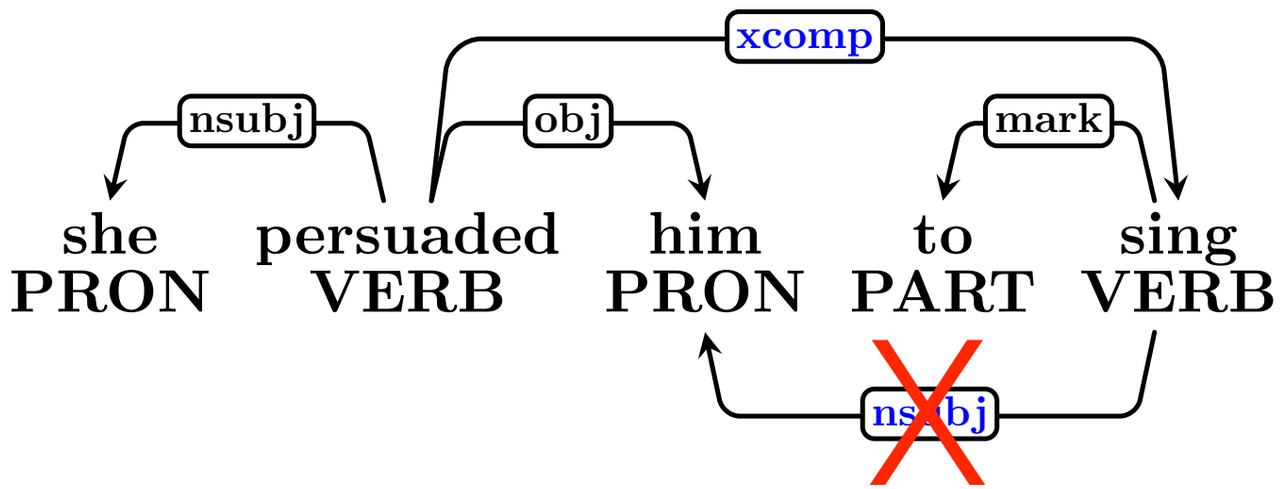


Turku: Subjects

From any infinitive verb attached as an xcomp to a higher predicate, consider adding an nsubj dependency to the nsubj of the predicate

1. Binary SVM classifier decides if dependency is added or not
2. Object control is not considered at all

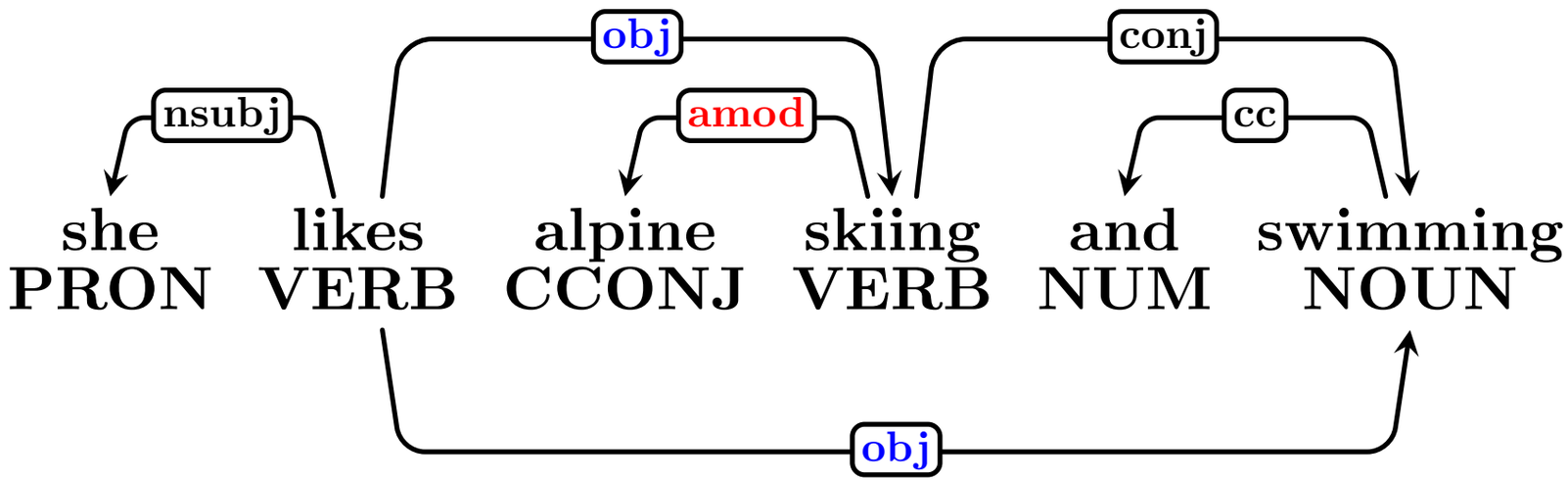




Turku: Coordination

The head and all dependents of the first conjunct are considered candidate head/dependents of all conjuncts

1. SVM classifier selects dependency label or null
2. Aims for high recall (rather than precision)



Evaluation

Evaluation

Manual evaluation:

- 1000 sentences from the training set for subjects and coordination
- The entire training sets for ellipsis (rare)

Evaluation

Manual evaluation:

- 1000 sentences from the training set for subjects and coordination
- The entire training sets for ellipsis (rare)

Error classification:

- Basic errors – errors caused by incorrect basic dependencies
- Enhanced errors – errors in spite of correct basic dependencies

Evaluation

Manual evaluation:

- 1000 sentences from the training set for subjects and coordination
- The entire training sets for ellipsis (rare)

Error classification:

- Basic errors – errors caused by incorrect basic dependencies
- Enhanced errors – errors in spite of correct basic dependencies

Evaluation metrics:

- Precision – percentage of predicted dependencies that are valid
- Recall – percentage of valid dependencies predicted (relative)

Results

Results

Subjects	Swe
Count	127
Precision	0.87
Recall (relative)	0.98
Basic errors	12
Enhanced errors	4

Results

Subjects	Swe	
	S	T
Count	127	36
Precision	0.87	0.83
Recall (relative)	0.98	0.27
Basic errors	12	1
Enhanced errors	4	5

Results

	Subjects		
	Swe		Ita
	S	T	S
Count	127	36	115
Precision	0.87	0.83	0.80
Recall (relative)	0.98	0.27	0.91
Basic errors	12	1	14
Enhanced errors	4	5	9

Results

	Subjects			
	Swe		Ita	
	S	T	S	T
Count	127	36	115	43
Precision	0.87	0.83	0.80	0.95
Recall (relative)	0.98	0.27	0.91	0.41
Basic errors	12	1	14	0
Enhanced errors	4	5	9	2

Results

	Subjects				Coordination
	Swe		Ita		Swe
	S	T	S	T	S
Count	127	36	115	43	559
Precision	0.87	0.83	0.80	0.95	0.94
Recall (relative)	0.98	0.27	0.91	0.41	0.55
Basic errors	12	1	14	0	25
Enhanced errors	4	5	9	2	9

Results

	Subjects				Coordination	
	Swe		Ita		Swe	
	S	T	S	T	S	T
Count	127	36	115	43	559	981
Precision	0.87	0.83	0.80	0.95	0.94	0.91
Recall (relative)	0.98	0.27	0.91	0.41	0.55	0.97
Basic errors	12	1	14	0	25	28
Enhanced errors	4	5	9	2	9	69

Results

	Subjects				Coordination		
	Swe		Ita		Swe		Ita
	S	T	S	T	S	T	S
Count	127	36	115	43	559	981	421
Precision	0.87	0.83	0.80	0.95	0.94	0.91	0.89
Recall (relative)	0.98	0.27	0.91	0.41	0.55	0.97	0.67
Basic errors	12	1	14	0	25	28	12
Enhanced errors	4	5	9	2	9	69	34

Results

	Subjects				Coordination			
	Swe		Ita		Swe		Ita	
	S	T	S	T	S	T	S	T
Count	127	36	115	43	559	981	421	653
Precision	0.87	0.83	0.80	0.95	0.94	0.91	0.89	0.82
Recall (relative)	0.98	0.27	0.91	0.41	0.55	0.97	0.67	0.96
Basic errors	12	1	14	0	25	28	12	32
Enhanced errors	4	5	9	2	9	69	34	86

Results

	Subjects				Coordination				Ellipsis
	Swe		Ita		Swe		Ita		Swe
	S	T	S	T	S	T	S	T	S
Count	127	36	115	43	559	981	421	653	112
Precision	0.87	0.83	0.80	0.95	0.94	0.91	0.89	0.82	0.85
Recall (relative)	0.98	0.27	0.91	0.41	0.55	0.97	0.67	0.96	
Basic errors	12	1	14	0	25	28	12	32	15
Enhanced errors	4	5	9	2	9	69	34	86	2

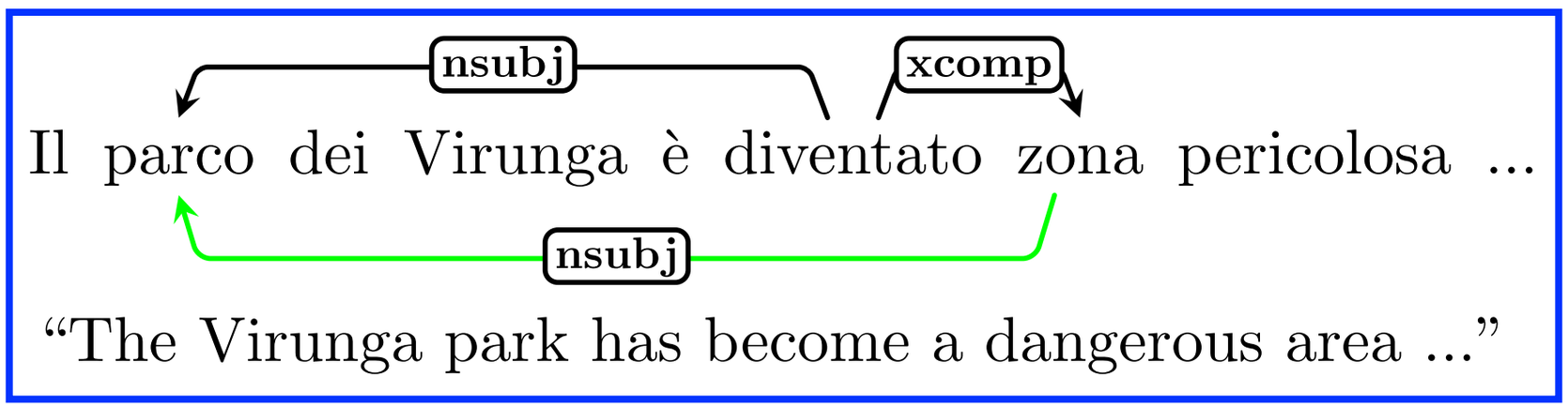
Results

	Subjects				Coordination				Ellipsis	
	Swe		Ita		Swe		Ita		Swe	Ita
	S	T	S	T	S	T	S	T	S	S
Count	127	36	115	43	559	981	421	653	112	162
Precision	0.87	0.83	0.80	0.95	0.94	0.91	0.89	0.82	0.85	0.76
Recall (relative)	0.98	0.27	0.91	0.41	0.55	0.97	0.67	0.96		
Basic errors	12	1	14	0	25	28	12	32	15	0
Enhanced errors	4	5	9	2	9	69	34	86	2	35

Error Analysis: Subjects

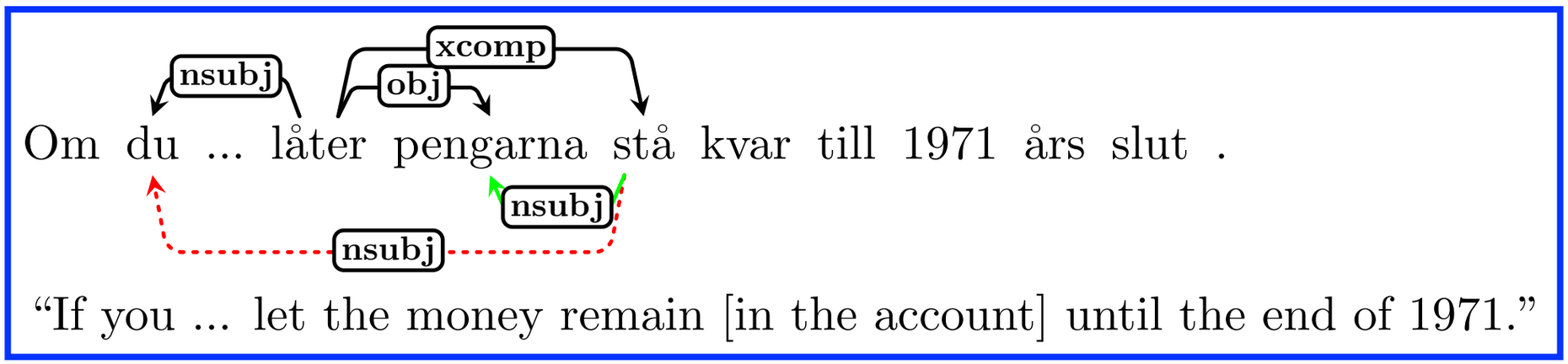
Error Analysis: Subjects

- Stanford has higher recall for both languages because it considers all xcomp nodes (not just infinitives)



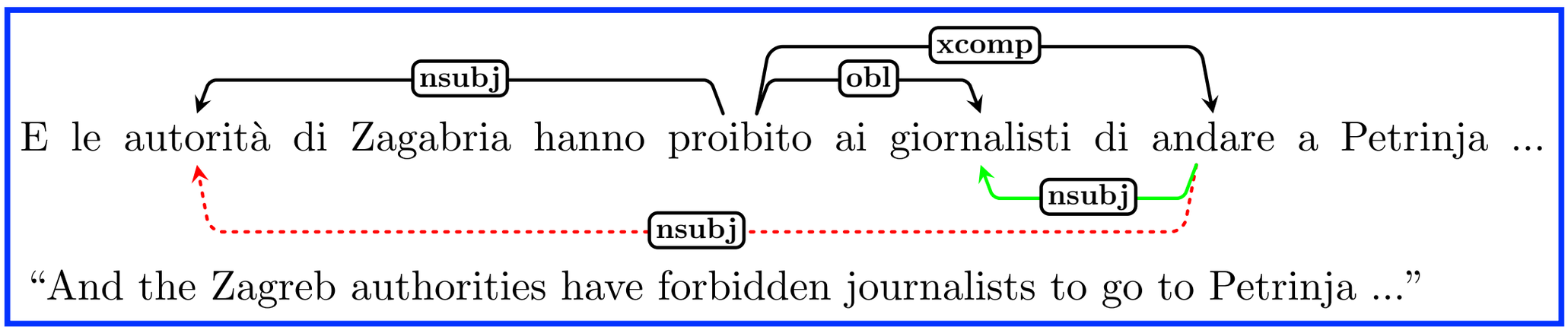
Error Analysis: Subjects

- Stanford has higher recall for both languages because it considers all xcomp nodes (not just infinitives)
- Stanford has higher precision for Swedish because it handles object control



Error Analysis: Subjects

- Stanford has higher recall for both languages because it considers all xcomp nodes (not just infinitives)
- Stanford has higher precision for Swedish because it handles object control
- Stanford has lower precision for Italian because it handles object control (sic)



Error Analysis: Coordination

Error Analysis: Coordination

- Stanford has higher precision for both languages because it limits projected dependencies to core arguments

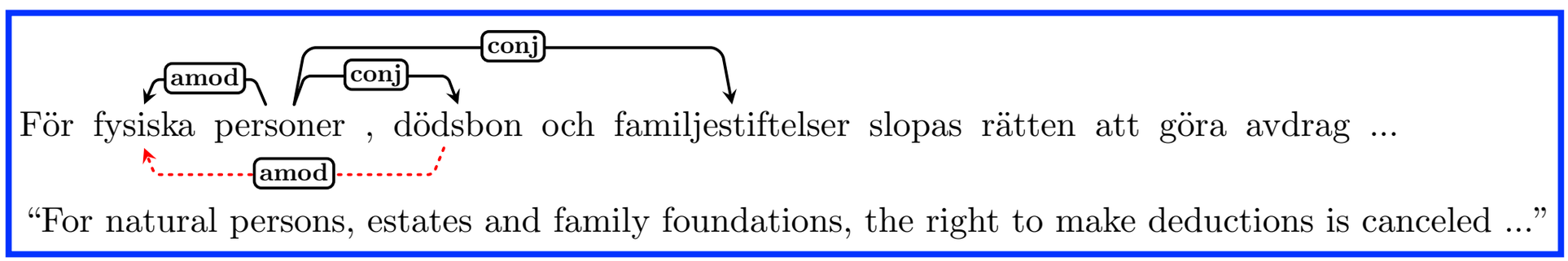
Error Analysis: Coordination

- Stanford has higher precision for both languages because it limits projected dependencies to core arguments
- Turku has higher recall for both languages because it projects all kinds of dependents

Error Analysis: Coordination

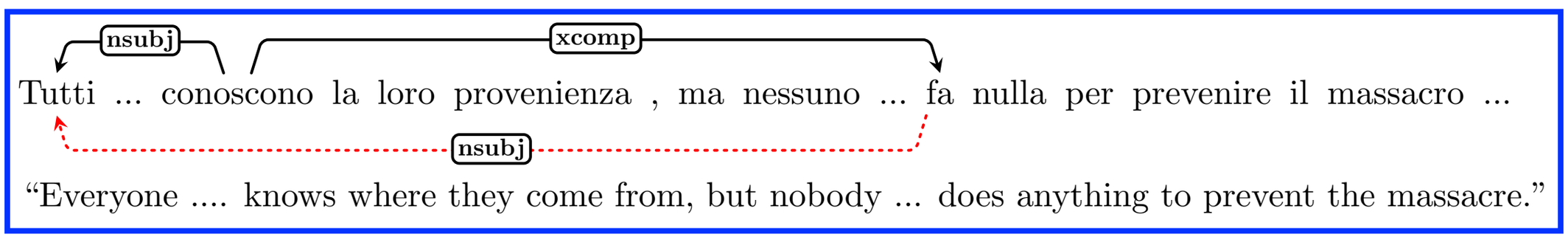
- Stanford has higher precision for both languages because it limits projected dependencies to core arguments
- Turku has higher recall for both languages because it projects all kinds of dependents

- Most common error type (especially for Turku): predicting shared left-dependents of the first conjunct



Error Analysis: Coordination

- Stanford has higher precision for both languages because it limits projected dependencies to core arguments
- Turku has higher recall for both languages because it projects all kinds of dependents
- Most common error type (especially for Turku): predicting shared left-dependents of the first conjunct



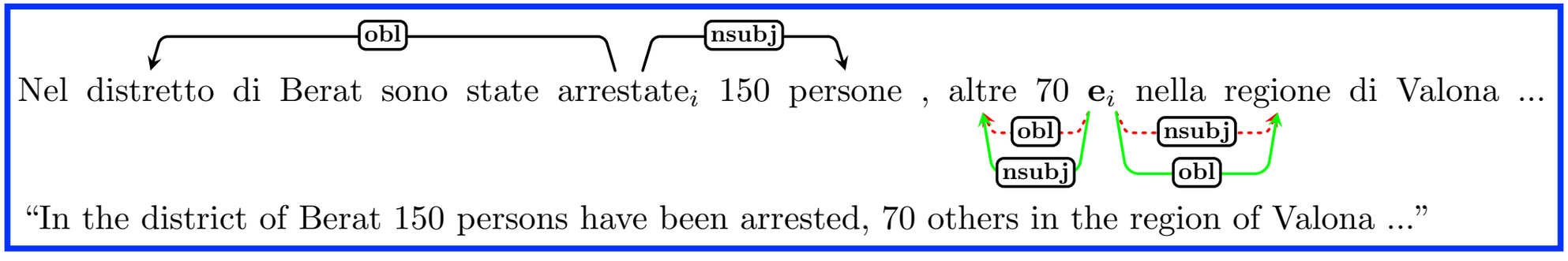
Error Analysis: Ellipsis

Error Analysis: Ellipsis

- The Stanford system developed for English works equally well (or better) for Swedish

Error Analysis: Ellipsis

- The Stanford system developed for English works equally well (or better) for Swedish
- The system has lower precision on Italian due (in part) to different word order constraints



Conclusion

Encouraging results for bootstrapping UD treebanks

- High cross-lingual accuracy – even for unrelated languages
- A few language-specific adaptations may help a lot
- Swedish v2.2 will have enhanced dependencies! 😀🇸🇪

Future work

- Use enhancers to post-process basic dependencies output by parsers