Core Arguments in Universal Dependencies




FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

Daniel Zeman

zeman@ufal.mff.cuni.cz
http://universaldependencies.org/workgroups/core.html

Universal Dependencies

- Unified annotation of…
  - Segmentation to “syntactic” words
  - Morphology (lemma, POS, features)
  - Surface syntax
- Same things annotated same way across languages…
- … while highlighting different coding strategies

Manning's Law

The secret to understanding UD is to realize that the design is a very subtle compromise between approximately 6 things:

1. UD must be satisfactory on linguistic analysis grounds for individual languages.

It's easy to come up with a proposal that improves UD on one of these dimensions. The interesting and difficult part is to improve UD while remaining sensitive to all these dimensions.

Manning's Law

The secret to understanding UD is to realize that the design is a very subtle compromise between approximately 6 things:

1. UD must be satisfactory on linguistic analysis grounds for individual languages.
2. UD must be good for linguistic typology, i.e., providing a suitable basis for bringing out cross-linguistic parallelism across languages and language families.

It's easy to come up with a proposal that improves UD on one of these dimensions. The interesting and difficult part is to improve UD while remaining sensitive to all these dimensions.

Manning's Law

The secret to understanding UD is to realize that the design is a very subtle compromise between approximately 6 things:

1. UD must be satisfactory on linguistic analysis grounds for individual languages.
2. UD must be good for linguistic typology, i.e., providing a suitable basis for bringing out cross-linguistic parallelism across languages and language families.
3. UD must be suitable for rapid, consistent annotation by a human annotator.

It's easy to come up with a proposal that improves UD on one of these dimensions. The interesting and difficult part is to improve UD while remaining sensitive to all these dimensions.

Manning's Law

The secret to understanding UD is to realize that the design is a very subtle compromise between approximately 6 things:

1. UD must be satisfactory on linguistic analysis grounds for individual languages.
2. UD must be good for linguistic typology, i.e., providing a suitable basis for bringing out cross-linguistic parallelism across languages and language families.
3. UD must be suitable for rapid, consistent annotation by a human annotator.
4. UD must be easily comprehended and used by a non-linguist, whether a language learner or an engineer with prosaic needs for language processing. … it leads us to favor traditional grammar notions and terminology.

It's easy to come up with a proposal that improves UD on one of these dimensions. The interesting and difficult part is to improve UD while remaining sensitive to all these dimensions.

Manning's Law

The secret to understanding UD is to realize that the design is a very subtle compromise between approximately 6 things:

1. UD must be satisfactory on linguistic analysis grounds for individual languages.
2. UD must be good for linguistic typology, i.e., providing a suitable basis for bringing out cross-linguistic parallelism across languages and language families.
3. UD must be suitable for rapid, consistent annotation by a human annotator.
4. UD must be easily comprehended and used by a non-linguist, whether a language learner or an engineer with prosaic needs for language processing. … it leads us to favor traditional grammar notions and terminology.
5. UD must be suitable for computer parsing with high accuracy.

It's easy to come up with a proposal that improves UD on one of these dimensions. The interesting and difficult part is to improve UD while remaining sensitive to all these dimensions.

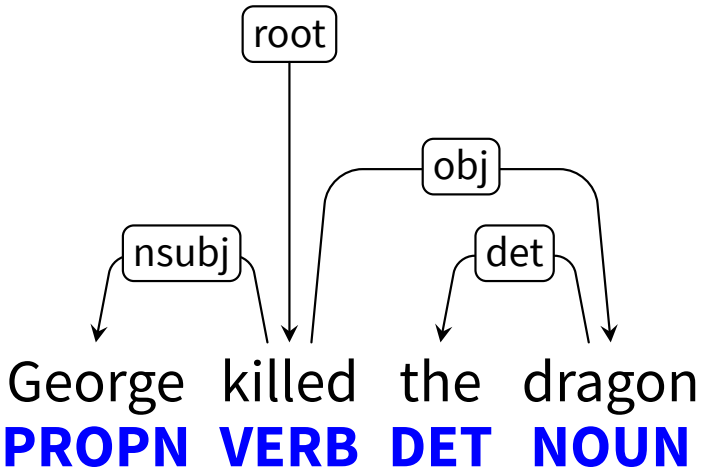
Manning's Law

The secret to understanding UD is to realize that the design is a very subtle compromise between approximately 6 things:

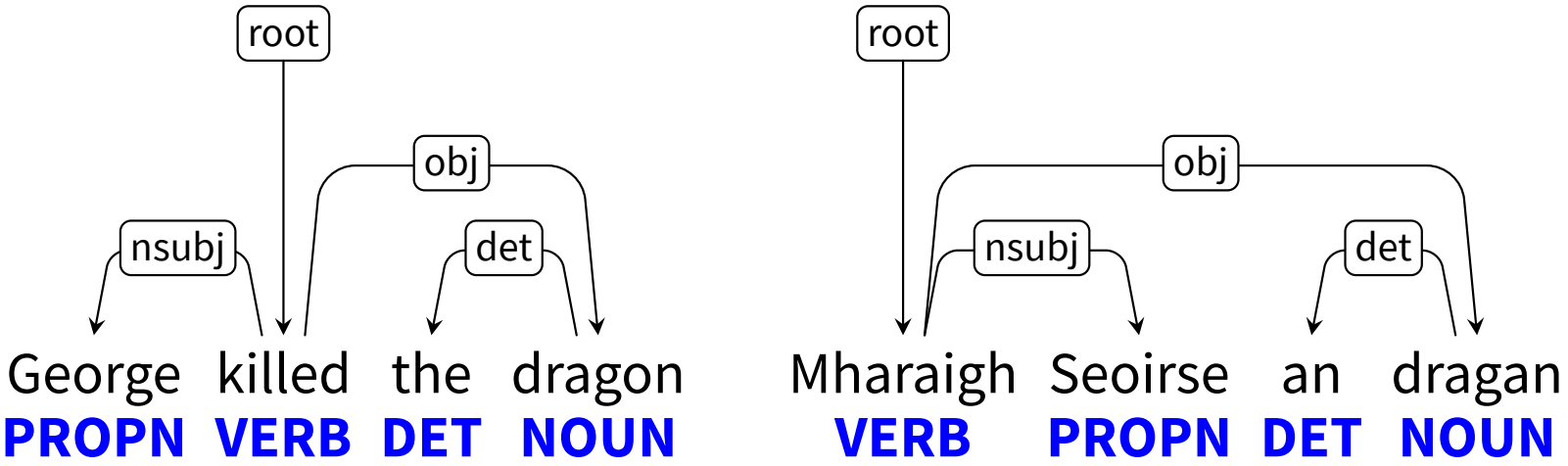
1. UD must be satisfactory on linguistic analysis grounds for individual languages.
2. UD must be good for linguistic typology, i.e., providing a suitable basis for bringing out cross-linguistic parallelism across languages and language families.
3. UD must be suitable for rapid, consistent annotation by a human annotator.
4. UD must be easily comprehended and used by a non-linguist, whether a language learner or an engineer with prosaic needs for language processing. … it leads us to favor traditional grammar notions and terminology.
5. UD must be suitable for computer parsing with high accuracy.
6. UD must support well downstream language understanding tasks (relation extraction, reading comprehension, machine translation, …)

It's easy to come up with a proposal that improves UD on one of these dimensions. The interesting and difficult part is to improve UD while remaining sensitive to all these dimensions.

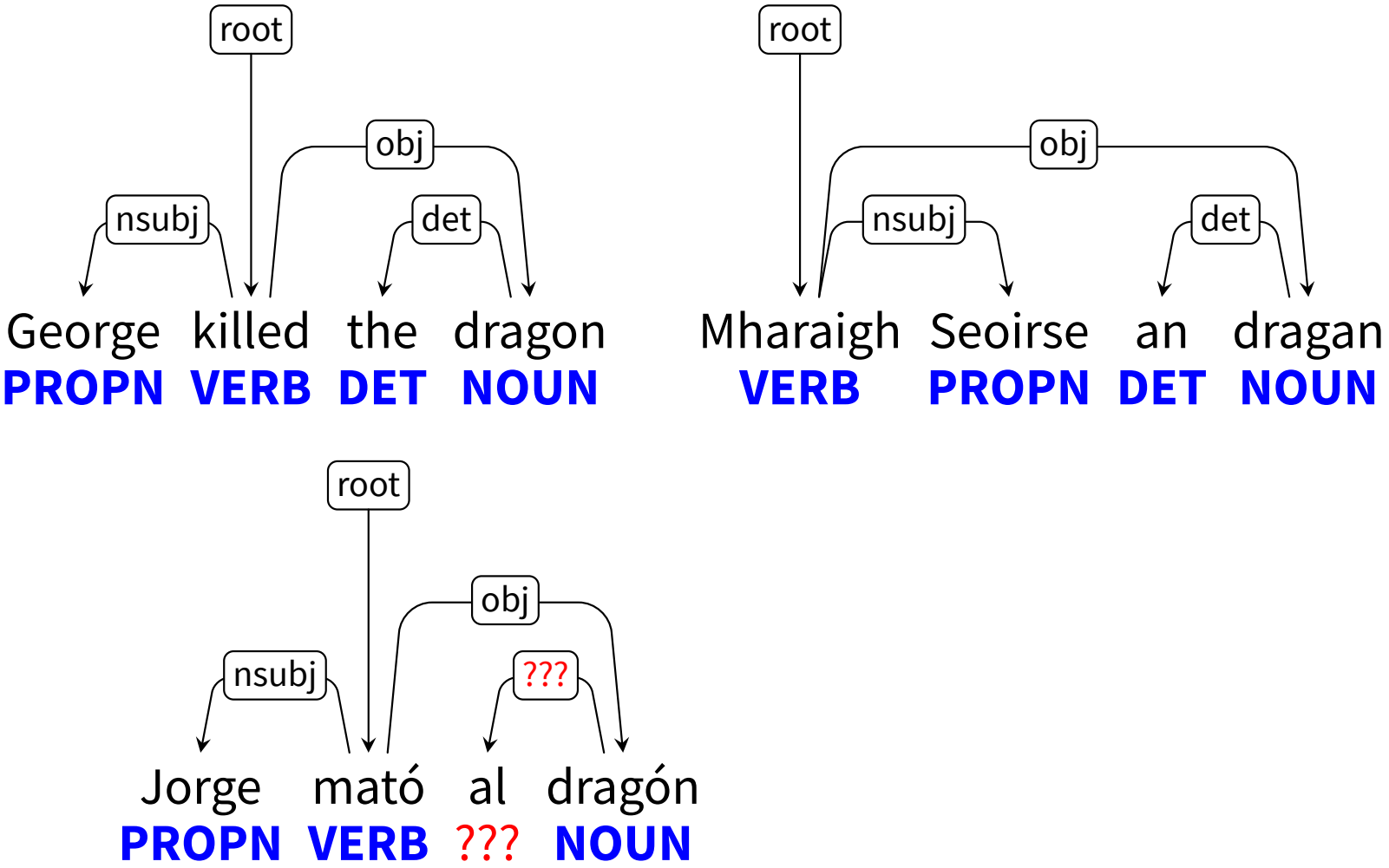
Same Thing Same Way



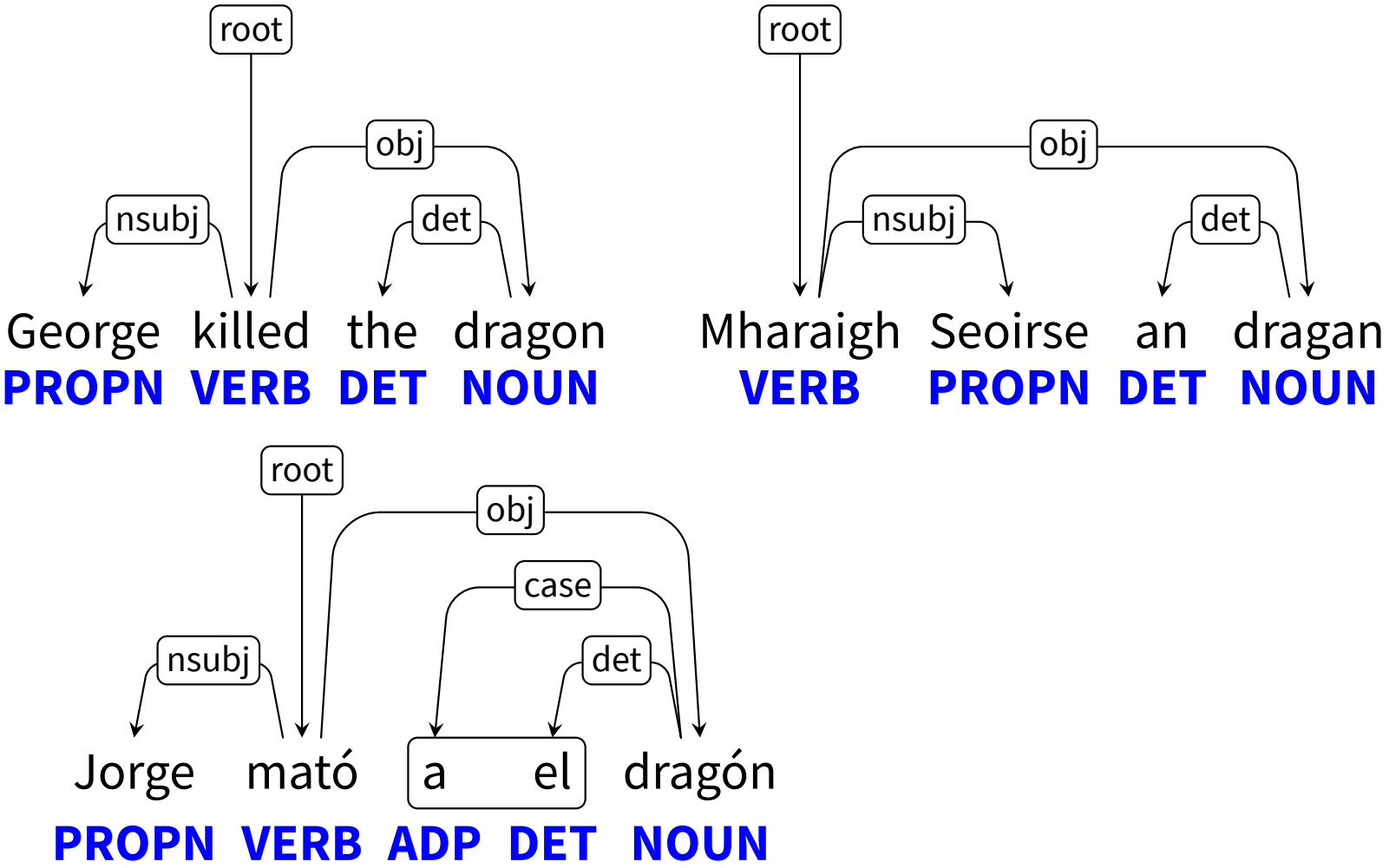
Same Thing Same Way



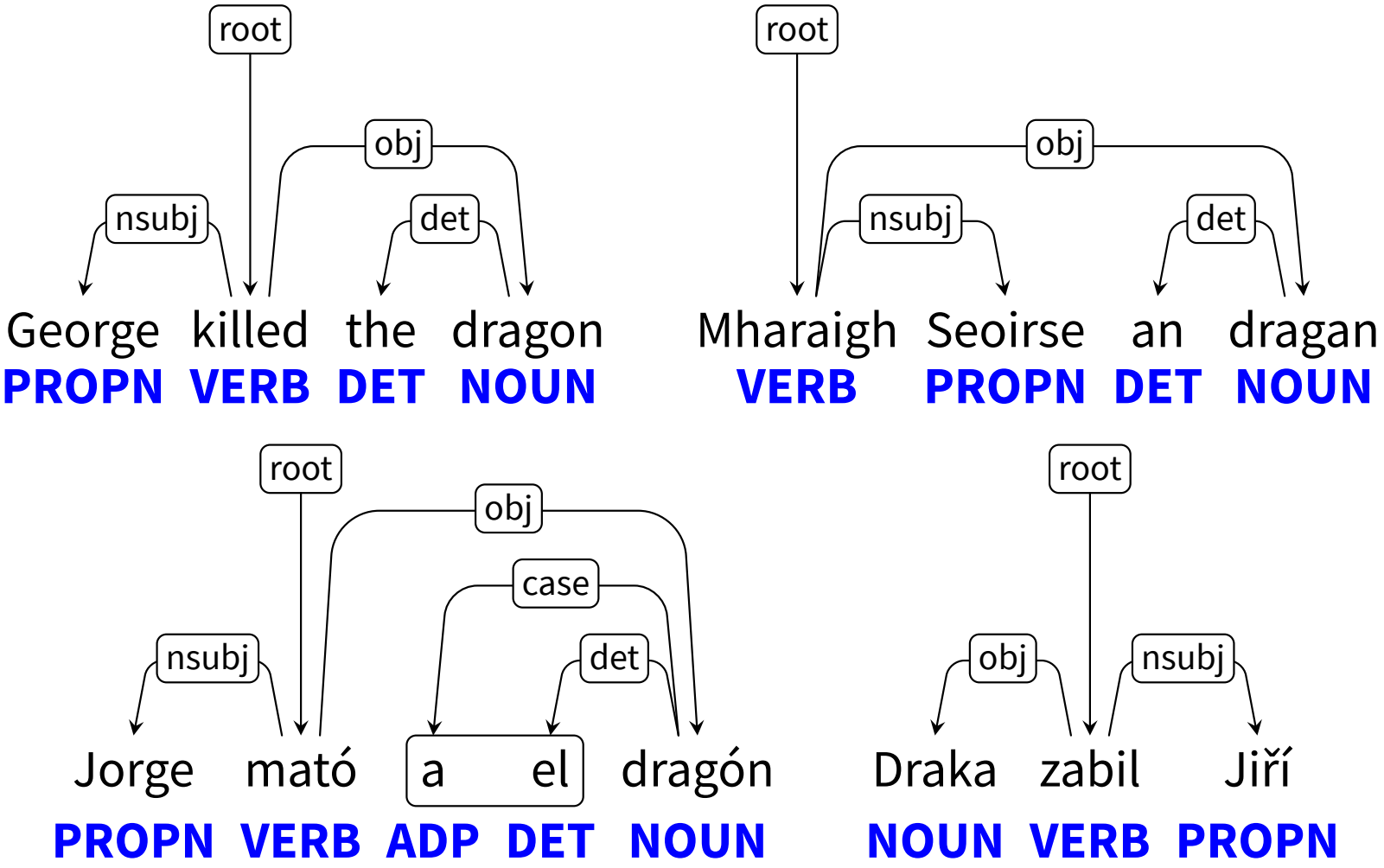
Same Thing Same Way



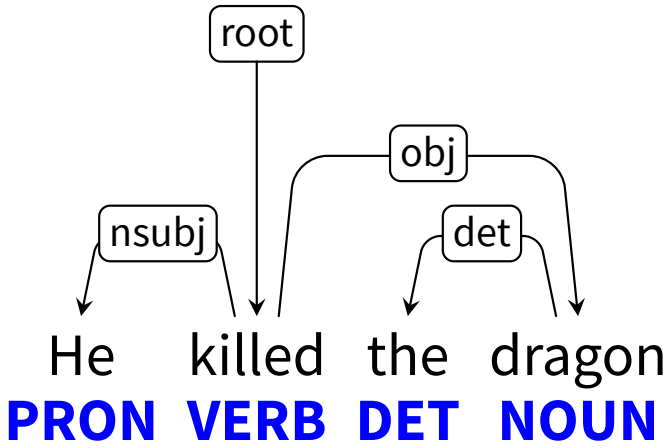
Same Thing Same Way



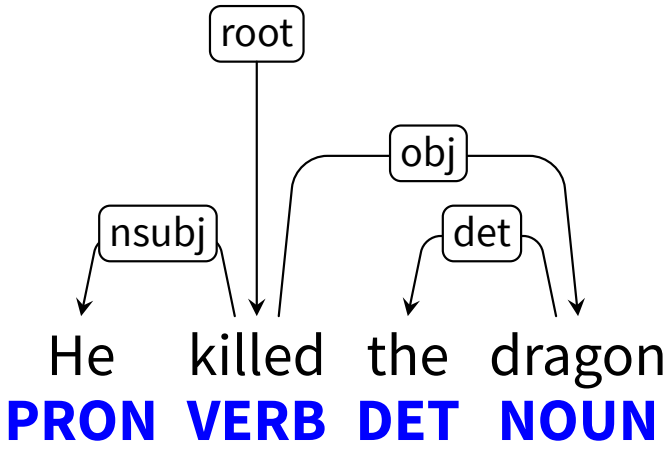
Same Thing Same Way

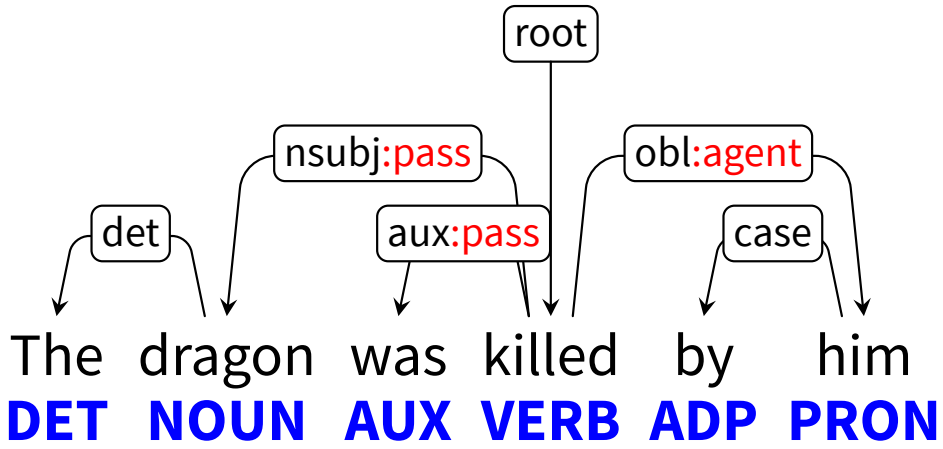


Same Meaning ≠ Same Construction!

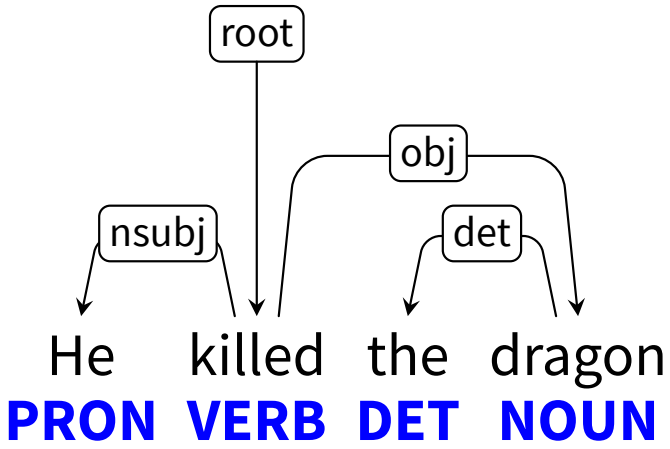


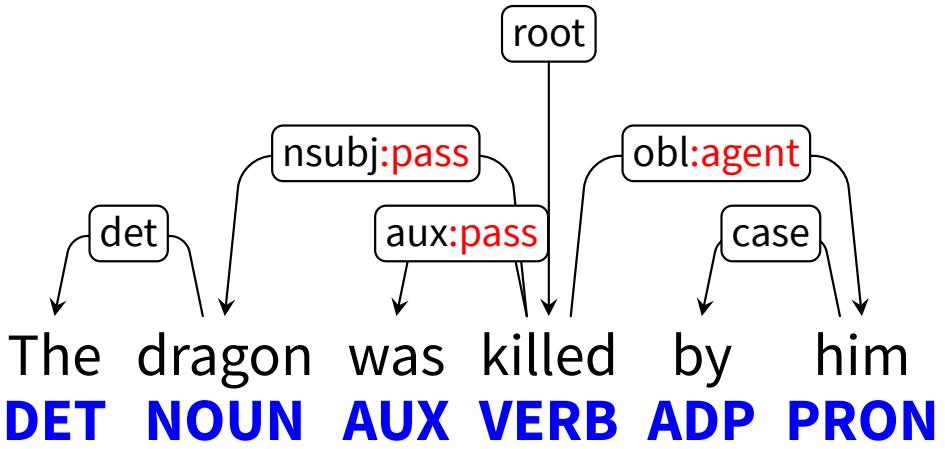
Same Meaning ≠ Same Construction!

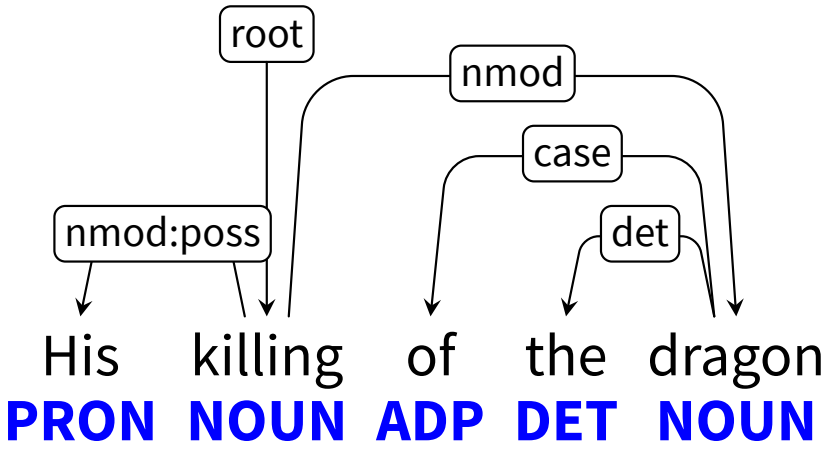




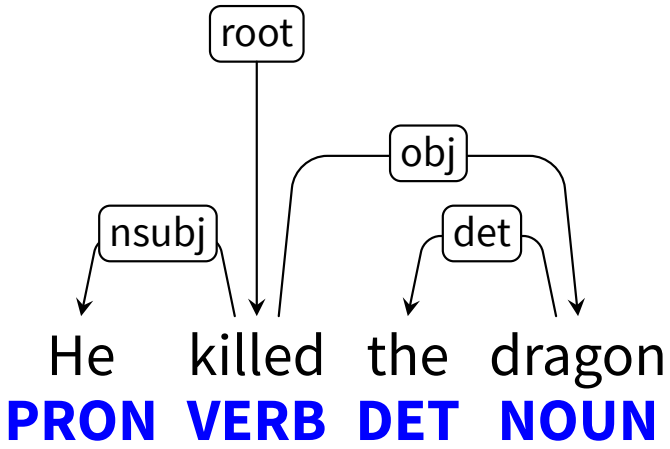
Same Meaning ≠ Same Construction!

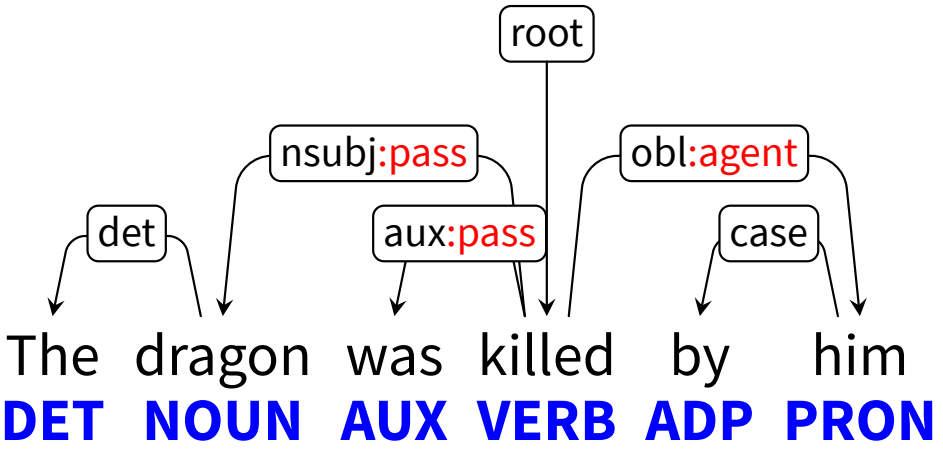


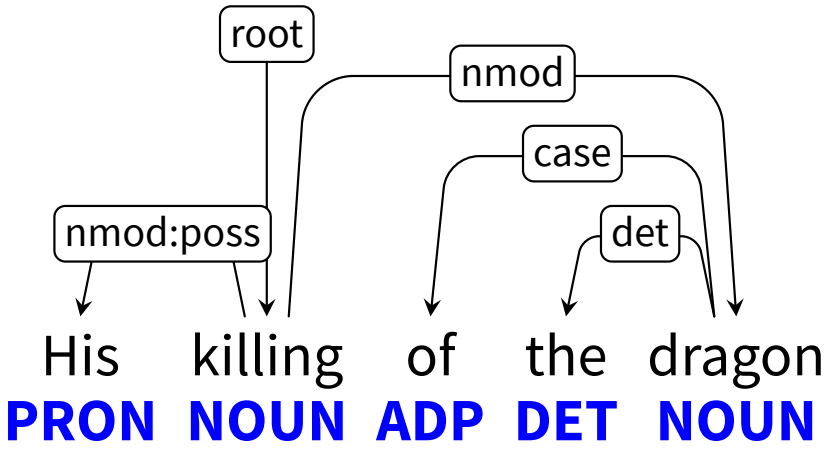


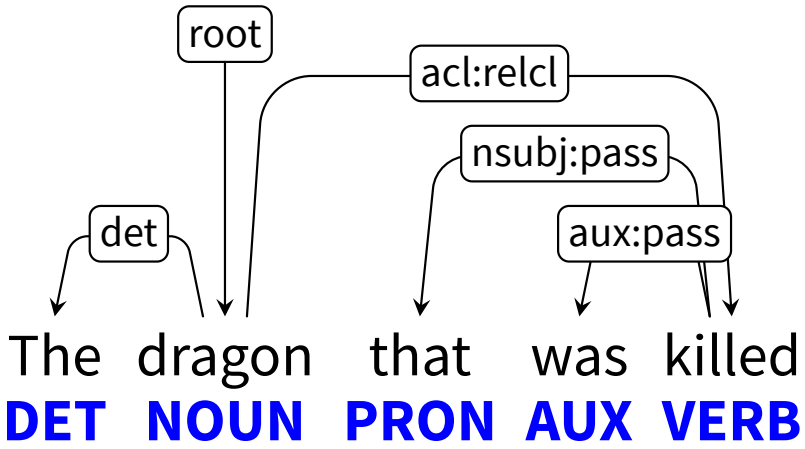


Same Meaning ≠ Same Construction!

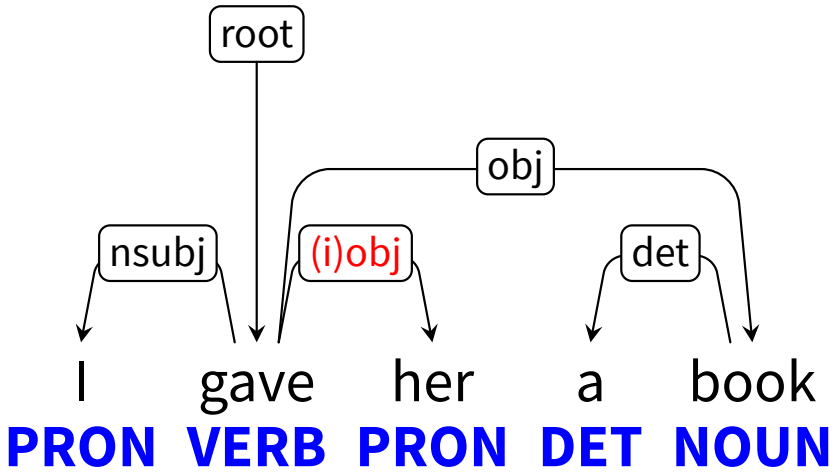




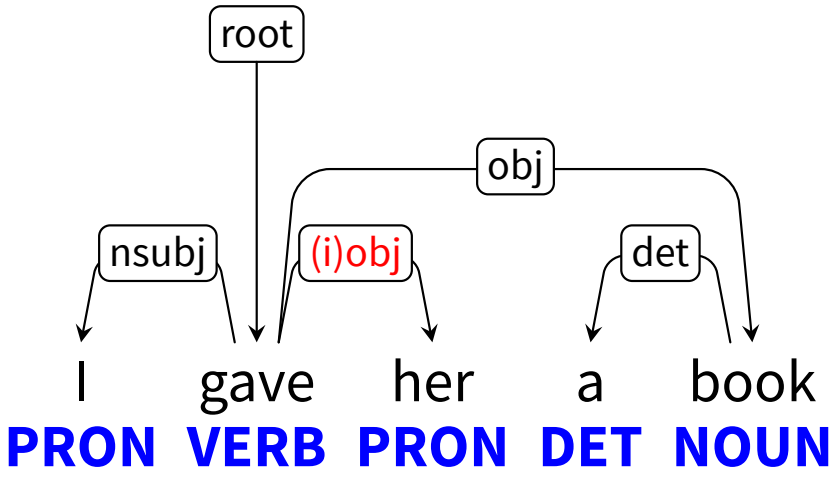


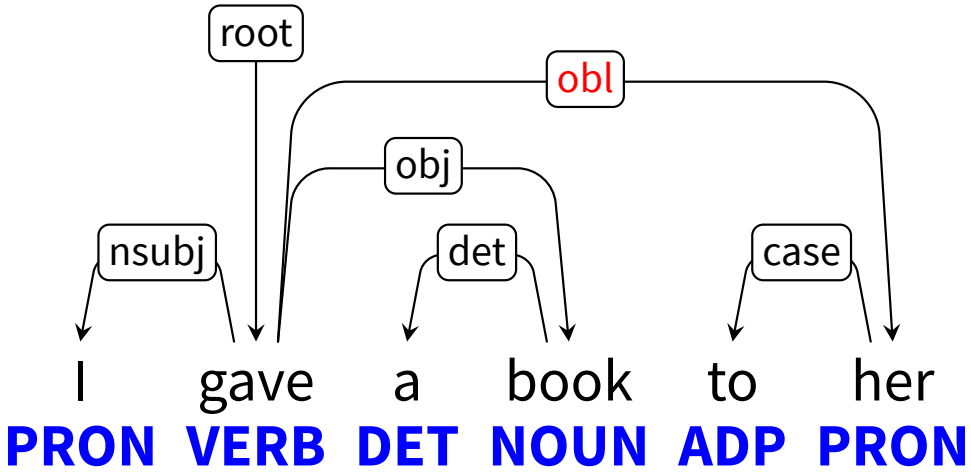


Information Packaging



Information Packaging





Dependents of Clauses (Verbal or Not)

	Nominal	Clausal	Modifier	Function
Core	**nsubj**	csubj		
Non-Core	**obl**	advcl	advmod	aux
	vocative		discourse	cop
	dislocated			mark
	expl			

Dependents of Clauses (Verbal or Not)

	Nominal	Clausal	Modifier	Function
Core	**nsubj**	csubj		
Non-Core	**obl**	advcl	advmod	aux
	vocative		discourse	cop
	dislocated			mark
	expl			

Dependents of Verbs, Adjectives and Adverbs

	Nominal	Clausal	Modifier
Core	**obj**	ccomp	
	iobj	xcomp	
Non-Core	**obl**	advcl	advmod
	expl		

Dependents of Clauses (Verbal or Not)

	Nominal	Clausal	Modifier	Function
Core	**nsubj**	csubj		
Non-Core	**obl**	advcl	advmod	aux
	vocative		discourse	cop
	dislocated			mark
	expl			

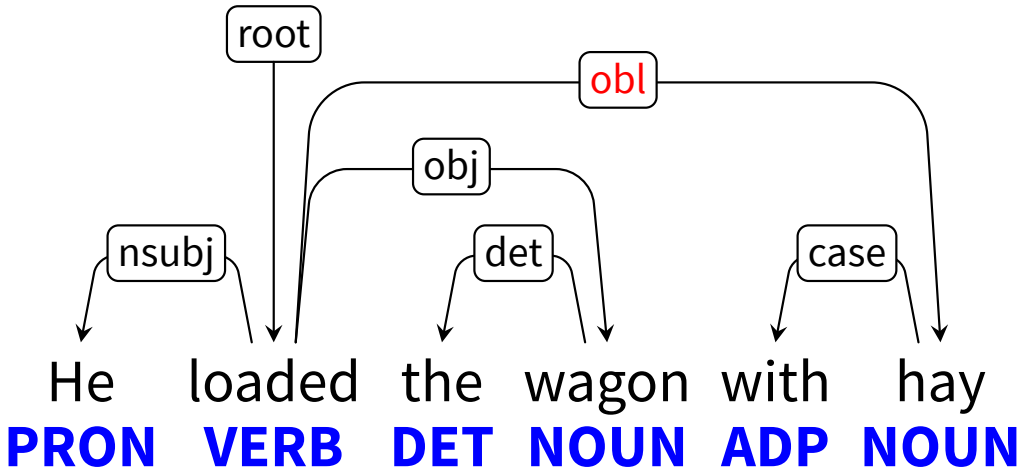
Dependents of Verbs, Adjectives and Adverbs

	Nominal	Clausal	Modifier
Core	**obj**	ccomp	
	iobj	xcomp	
Non-Core	**obl**	advcl	advmod
	expl		

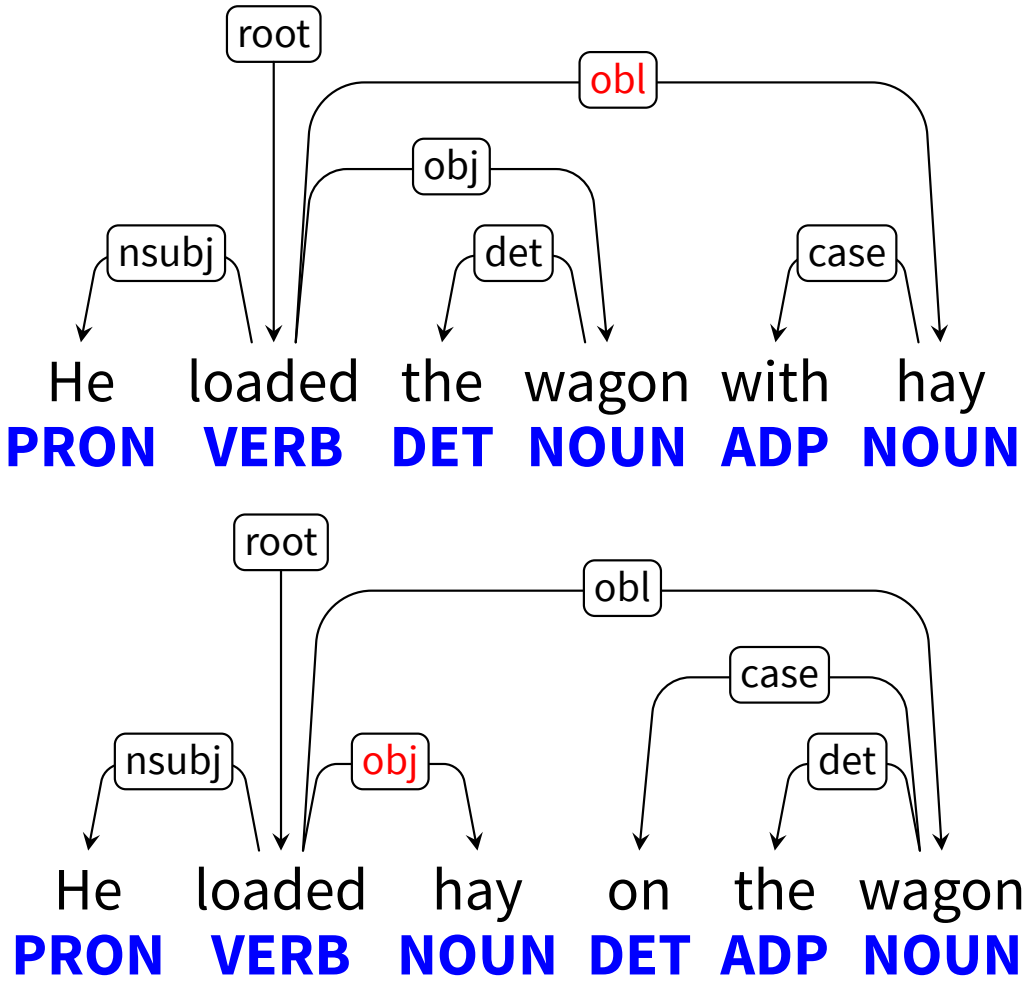
Dependents of Nominals

Nominal	Clausal	Modifier	Function
nmod	acl	amod	det
appos		nummod	case

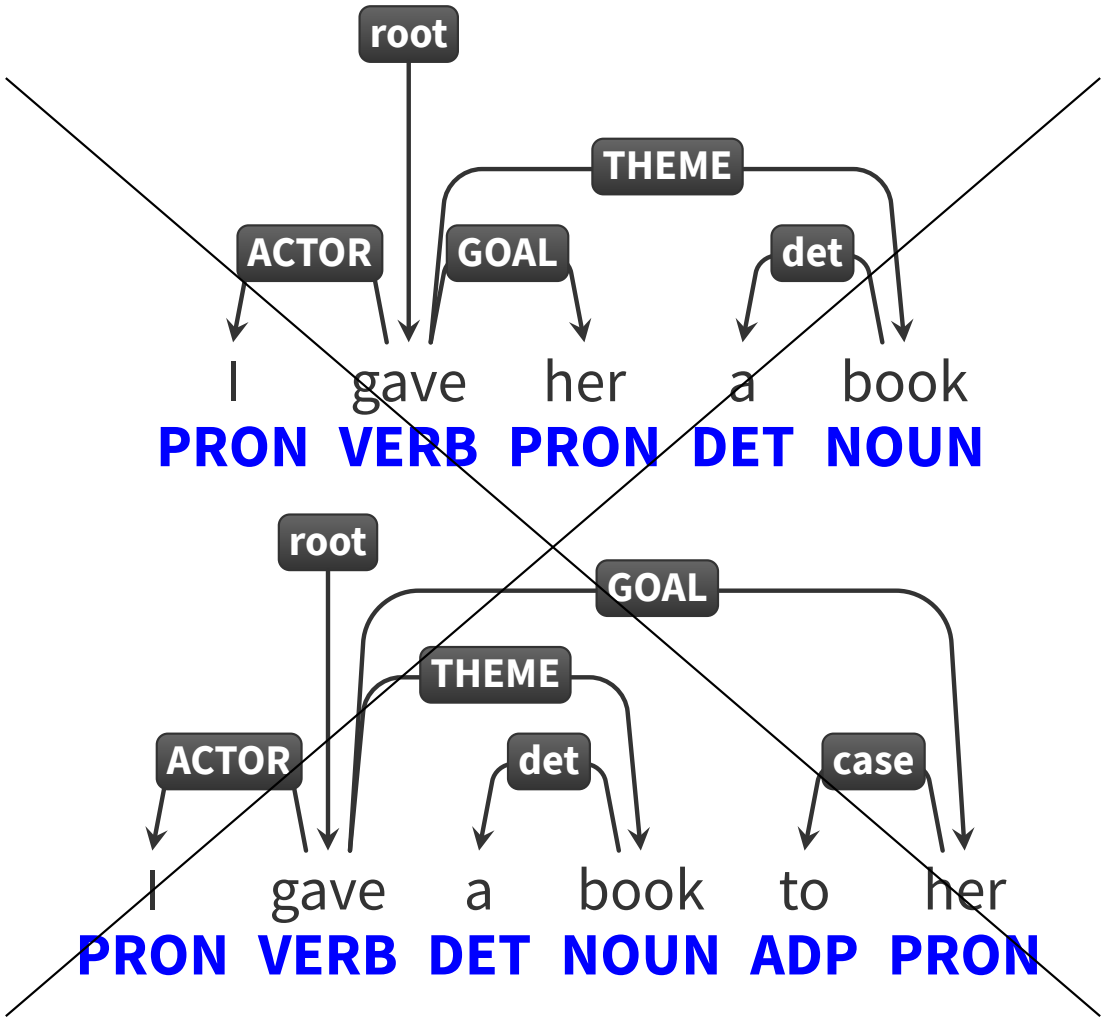
Information Packaging



Information Packaging



UD is NOT about Semantic Roles!



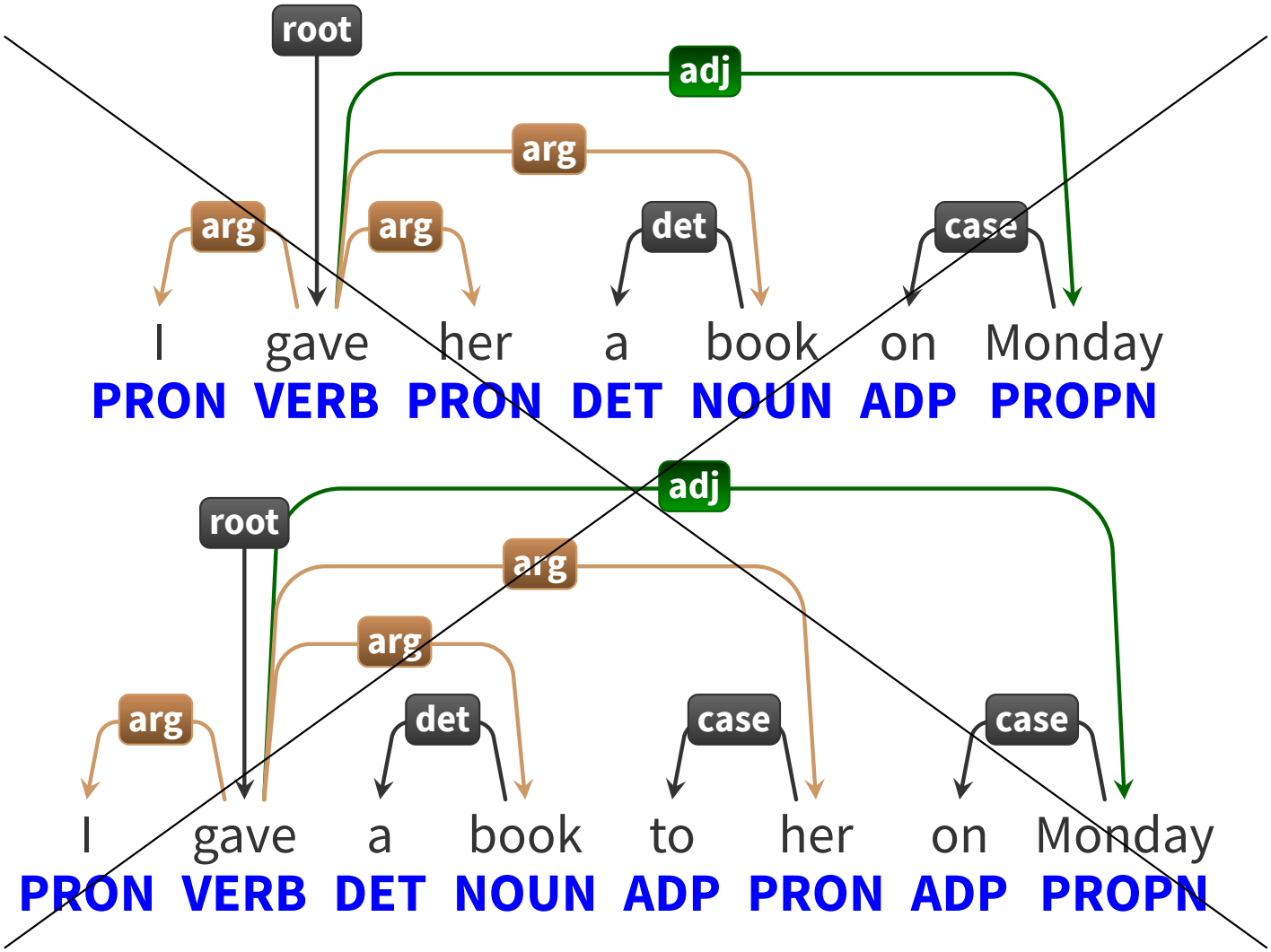
Manning's Law – What If We Do Semantic Roles?

The secret to understanding the design and current success of UD is to realize that the design is a very subtle compromise between approximately 6 things:

1. UD must be satisfactory on linguistic analysis grounds for individual languages.
2. UD must be good for linguistic typology, i.e., providing a suitable basis for bringing out cross-linguistic parallelism across languages and language families.
3. UD must be suitable for rapid, consistent annotation by a human annotator.
4. UD must be easily comprehended and used by a non-linguist, whether a language learner or an engineer with prosaic needs for language processing. … it leads us to favor traditional grammar notions and terminology.
5. UD must be suitable for computer parsing with high accuracy.
6. UD must support well downstream language understanding tasks (relation extraction, reading comprehension, machine translation, …)

It's easy to come up with a proposal that improves UD on one of these dimensions. The interesting and difficult part is to improve UD while remaining sensitive to all these dimensions.

UD Avoids Argument-Adjunct Distinction!



Avoiding an Argument-Adjunct Distinction

- From the guidelines:
  - Subtle, unclear, and frequently argued over
  - Questionable as a categorical distinction
  - Best practical solution is to eliminate it

Avoiding an Argument-Adjunct Distinction

- From the guidelines:
  - Subtle, unclear, and frequently argued over
  - Questionable as a categorical distinction
  - Best practical solution is to eliminate it

- BUT:
  - Cannot be eliminated completely
  - Some people/data have it and want to keep it
    - ★ It aligns well with traditional grammars
  - ⇒ there is now a relation subtype `obl:arg`

Avoiding an Argument-Adjunct Distinction

- From the guidelines:
  - Subtle, unclear, and frequently argued over
  - Questionable as a categorical distinction
  - Best practical solution is to eliminate it
- BUT:
  - Cannot be eliminated completely
  - Some people/data have it and want to keep it
    - ★ It aligns well with traditional grammars
  - ⇒ there is now a relation subtype `obl:arg`
- AND I will argue that
  - Core-oblique distinction is unclear and argued over too

Avoiding an Argument-Adjunct Distinction

- From the guidelines:
  - Subtle, unclear, and frequently argued over
  - Questionable as a categorical distinction
  - Best practical solution is to eliminate it
- BUT:
  - Cannot be eliminated completely
  - Some people/data have it and want to keep it
    - ★ It aligns well with traditional grammars
  - ⇒ there is now a relation subtype `obl:arg`
- AND I will argue that
  - Core-oblique distinction is unclear and argued over too
  - (Though I will not propose to discard it.)

So What Is Core and Why?

Community Confusion

- UD v1 guidelines took core-oblique for granted
- English (simplified):
  - Bare noun phrase ⇒ core argument (`nsubj`, `obj`, `iobj`)
  - Prepositional phrase ⇒ oblique argument or adjunct (`obl`)

Community Confusion

- UD v1 guidelines took core-oblique for granted
- English (simplified):
  - Bare noun phrase ⇒ core argument (`nsubj`, `obj`, `iobj`)
  - Prepositional phrase ⇒ oblique argument or adjunct (`obl`)
- Other languages: not necessarily! (Spanish, Japanese)
  - But some people simply took the English rule…
  - Manning's law: non-linguists!

Community Confusion

- UD v1 guidelines took core-oblique for granted
- English (simplified):
  - Bare noun phrase ⇒ core argument (`nsubj`, `obj`, `iobj`)
  - Prepositional phrase ⇒ oblique argument or adjunct (`obl`)
- Other languages: not necessarily! (Spanish, Japanese)
  - But some people simply took the English rule…
  - Manning's law: non-linguists!
- Clash with traditional terminology
  - Grammars of German, Czech etc. define prepositional objects
  - But these are not necessarily core…
  - Yet some people took their national definition of object…

Language-specific Coding Strategy

- Idea:
  - Oblique arguments are marked similarly to adjuncts (prepositions, certain morphological cases…)
  - Core arguments are marked differently
    - ⇒ easy for annotators and non-linguists!
- Why are core arguments special?
  - They tend to be targeted by grammatical rules
    - Passivization
    - Control verbs
    - Reflexives
    - …

Language-specific Coding Strategy

- Core vs. oblique is not defined in traditional grammar
- How shall we define it?

Language-specific Coding Strategy

- Core vs. oblique is not defined in traditional grammar
- How shall we define it?

- Andrews, 2007 (In Shopen: Language Typology)
  - Identify primary transitive predicates
    - ★ We need semantic roles for this! (One-time only.)

Language-specific Coding Strategy

- Core vs. oblique is not defined in traditional grammar
- How shall we define it?

- Andrews, 2007 (In Shopen: Language Typology)
  - Identify primary transitive predicates
    - ★ We need semantic roles for this! (One-time only.)
  - Actor/agent = function A
  - Undergoer/patient = function P

Language-specific Coding Strategy

- Core vs. oblique is not defined in traditional grammar
- How shall we define it?

- Andrews, 2007 (In Shopen: Language Typology)
  - Identify primary transitive predicates
    - ★ We need semantic roles for this! (One-time only.)
  - Actor/agent = function A
  - Undergoer/patient = function P
  - Note the way they are coded
  - Note other grammatical rules that target them
  - Generalize to other predicates with same coding and rules

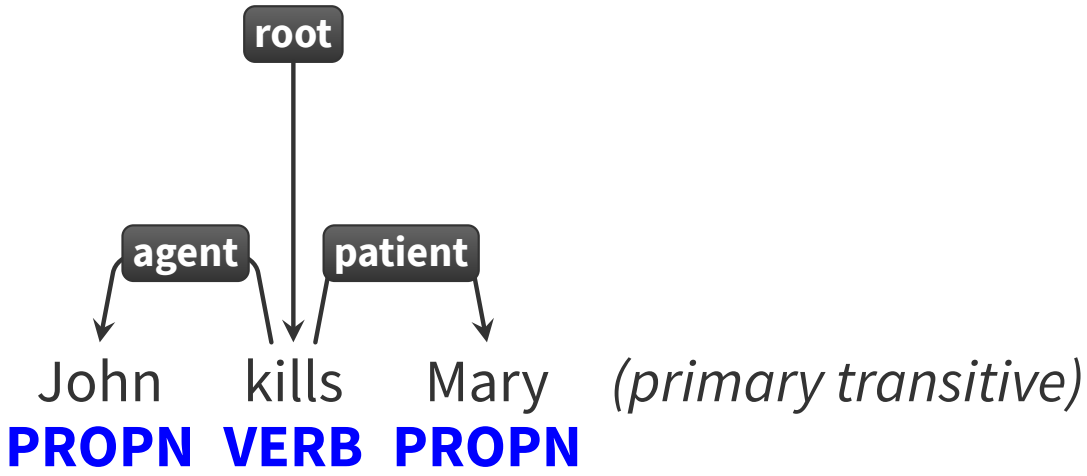
Language-specific Coding Strategy

- Core vs. oblique is not defined in traditional grammar
- How shall we define it?

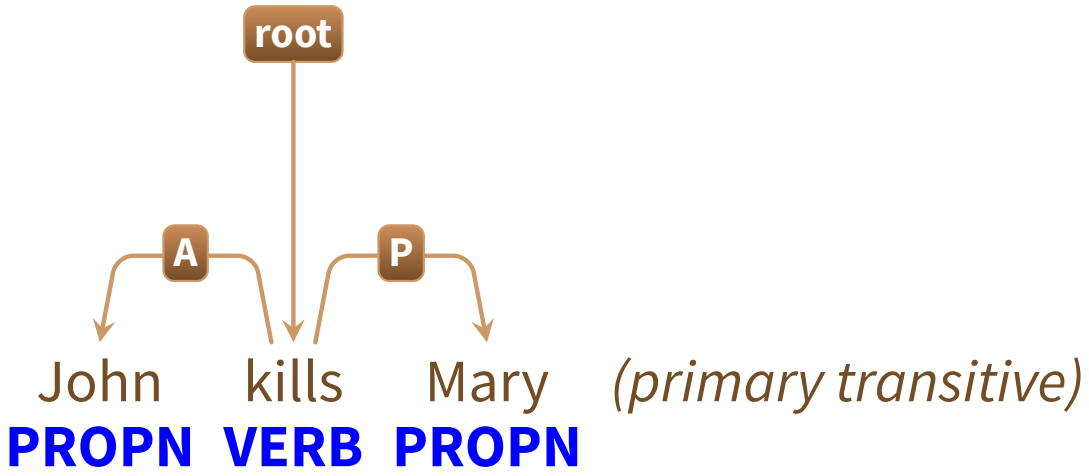
- Andrews, 2007 (In Shopen: Language Typology)
  - Identify primary transitive predicates
    - ★ We need semantic roles for this! (One-time only.)
  - Actor/agent = function A
  - Undergoer/patient = function P
  - Note the way they are coded
  - Note other grammatical rules that target them
  - Generalize to other predicates with same coding and rules

  - Then define:
    - ★ function A ⇒ **`nsubj`**
    - ★ function P ⇒ **`obj`**

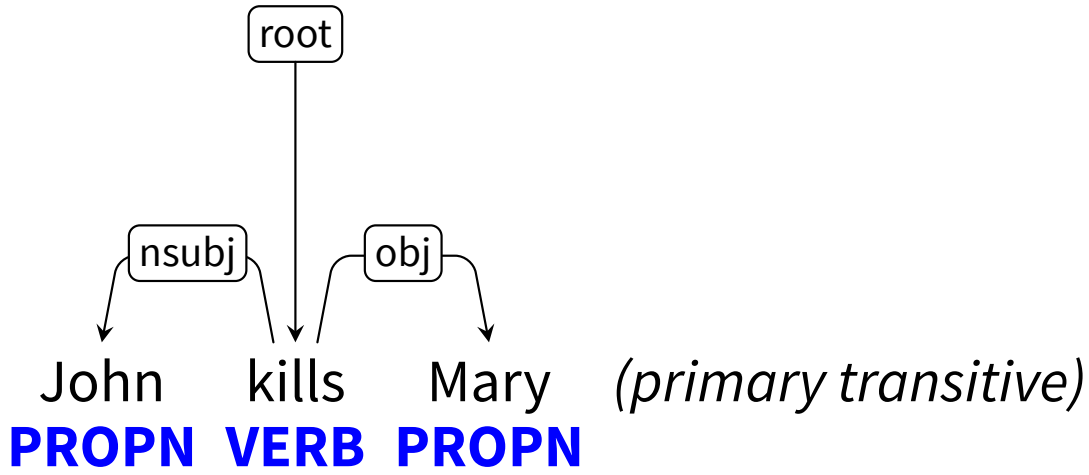
Transitive Predicates in English



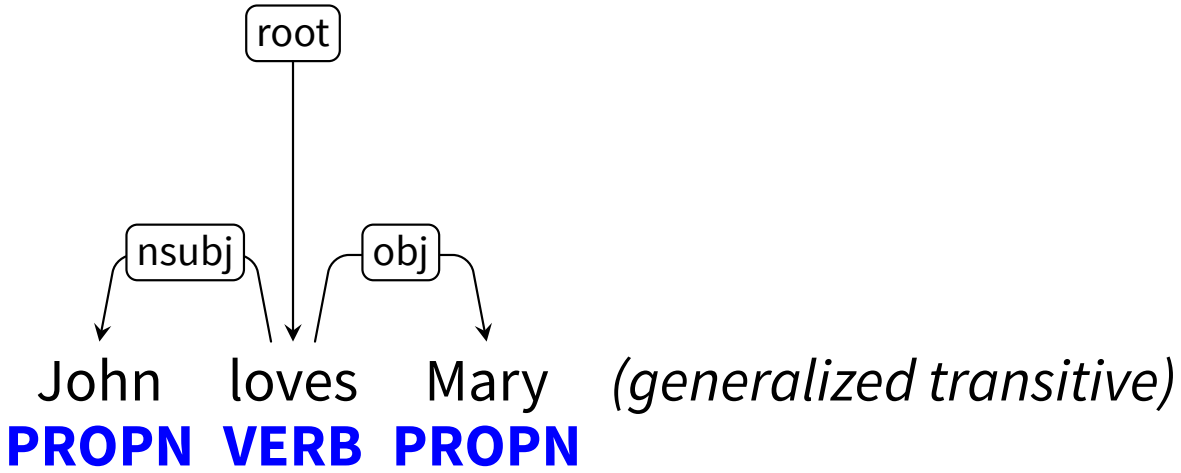
Transitive Predicates in English



Transitive Predicates in English

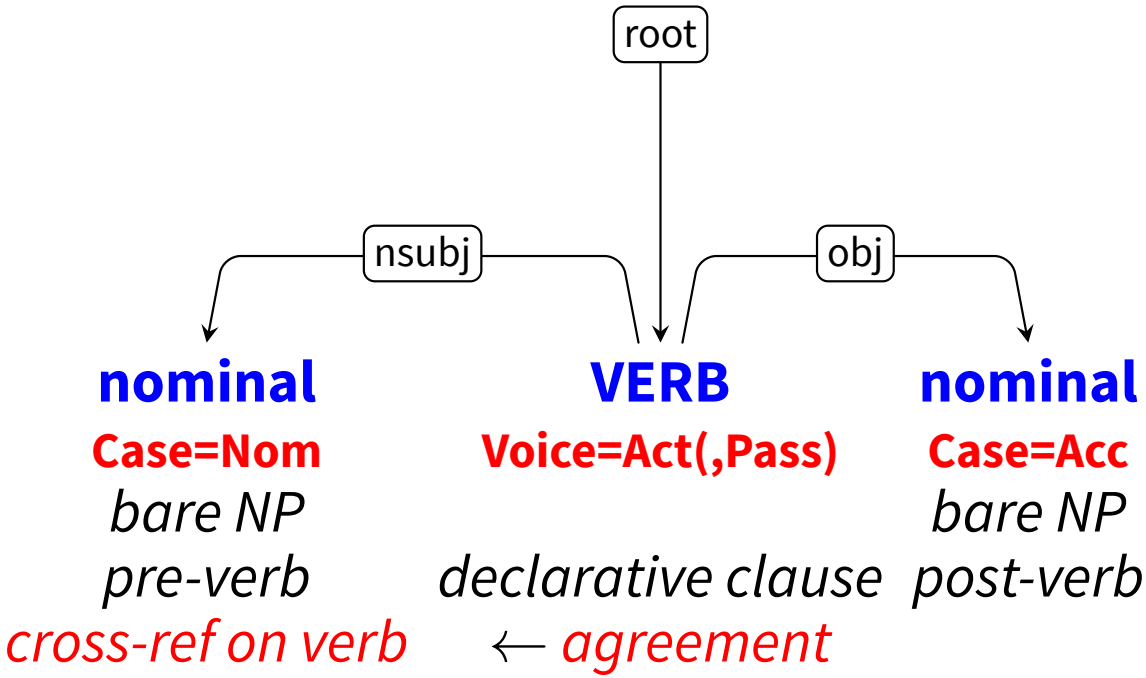


(primary transitive)

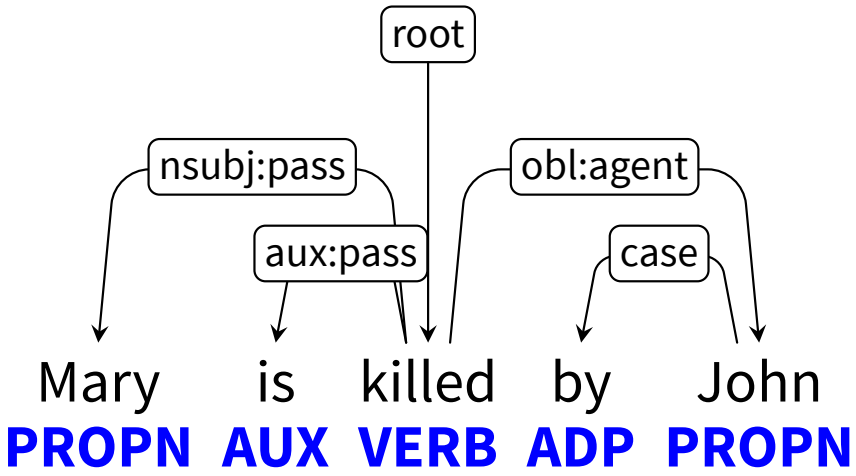


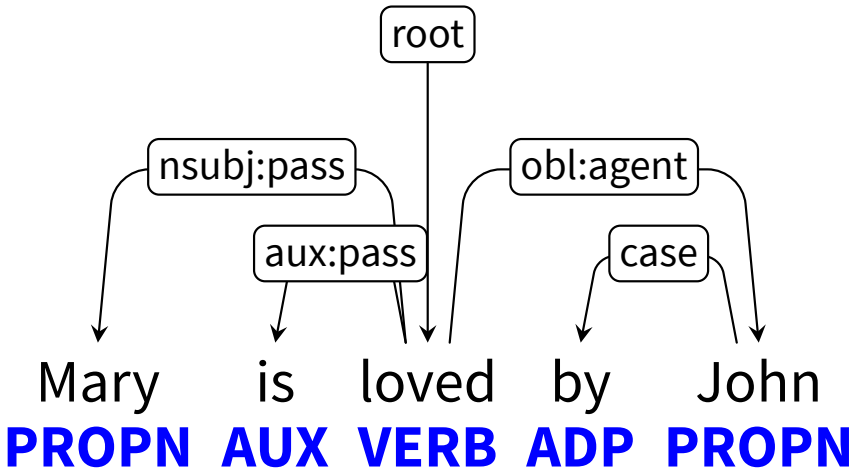
(generalized transitive)

Transitive Predicates in English

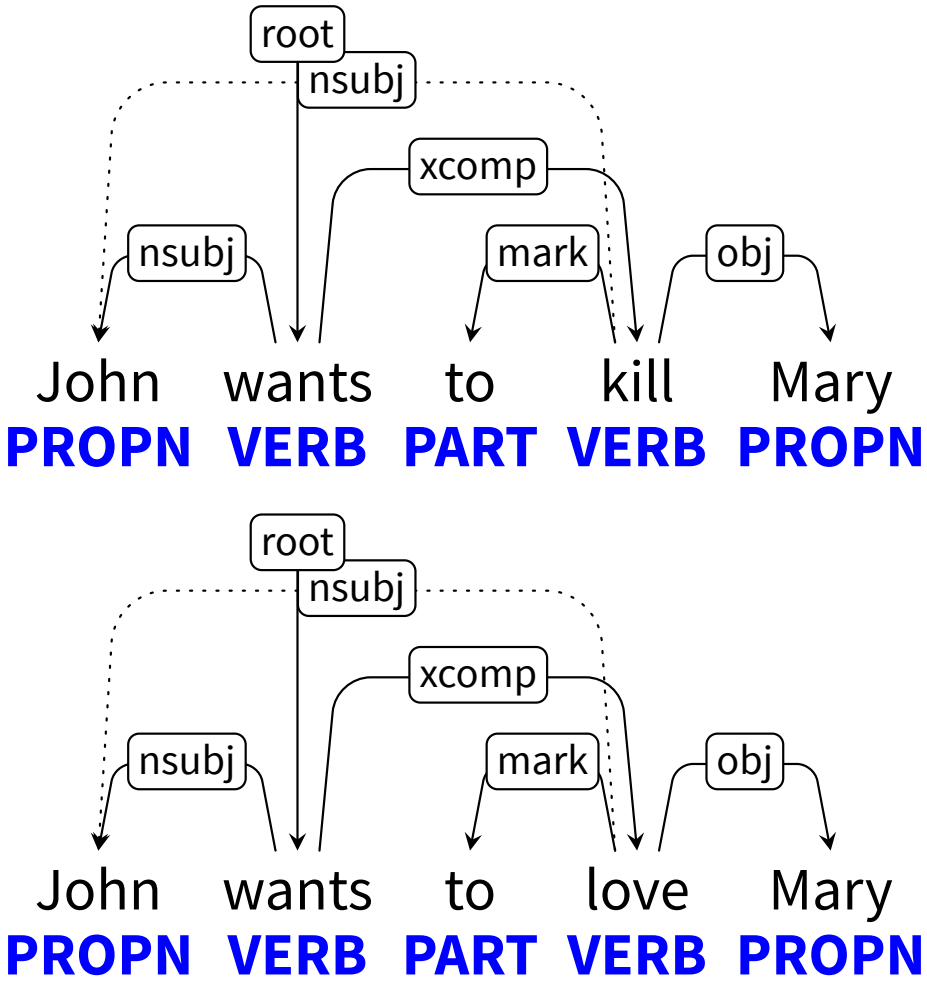


Passivization in English

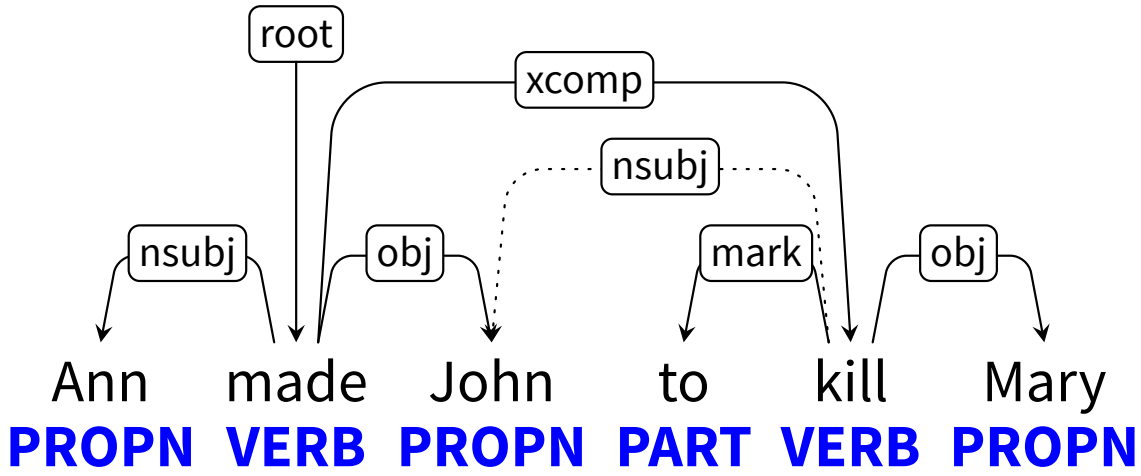


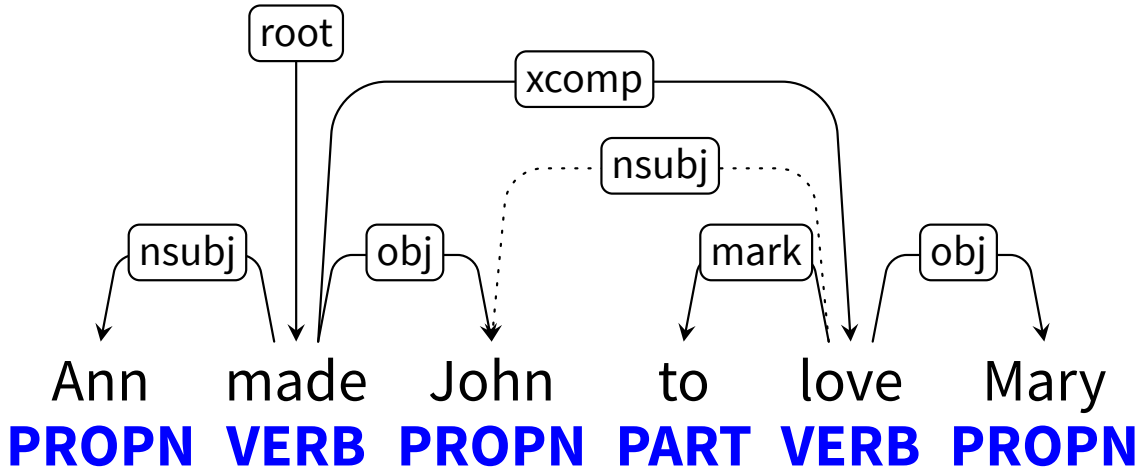


Subject Control in English



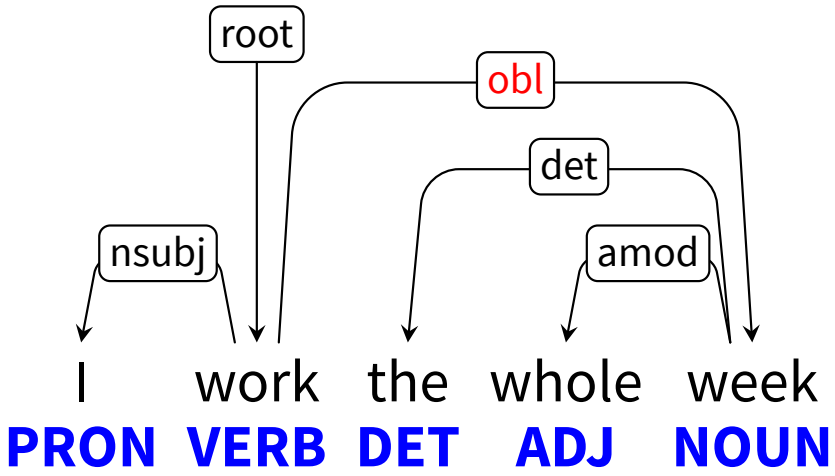
Object Control in English





Some Problems

- Some temporal adjuncts are bare noun phrases
  - *I work the whole week.*
  - *I work every Friday.*



- At least it cannot passivize:
  - **The whole week is worked by me.*
  - **Every Friday is worked by me.*
- But…

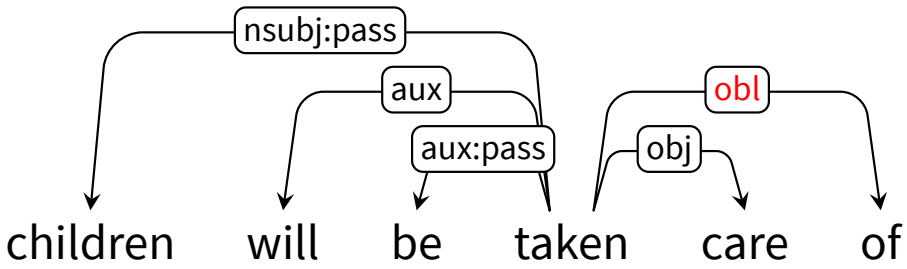
Some Problems

- Some transitive verbs cannot passivize
  - *John has a new car.*
    - ★ **A new car is had by John.*
  - *Friday does not suit me.*
    - ★ **I am not suited by Friday.*

Some Problems

- Some transitive verbs cannot passivize
  - *John has a new car.*
    - ★ **A new car is had by John.*
  - *Friday does not suit me.*
    - ★ **I am not suited by Friday.*

- Some prepositional verbs can passivize
  - *You can rely on Ben.*
    - ★ *Ben can be relied on.*
  - *They will take care of your children.*
    - ★ *Your children will be taken care of.*



Tentative Summary?

- The borderline is inherently fuzzy
- No universally applicable and exact algorithm
- Better described in terms of probability

Tentative Summary?

- The borderline is inherently fuzzy
- No universally applicable and exact algorithm
- Better described in terms of probability

- Core coding not favored by adjuncts
- Oblique coding similar to most adjuncts
- Passivization etc. may help…
- … but does not work as strict criterion

Tentative Summary?

- The borderline is inherently fuzzy
- No universally applicable and exact algorithm
- Better described in terms of probability

- Core coding not favored by adjuncts
- Oblique coding similar to most adjuncts
- Passivization etc. may help…
- … but does not work as strict criterion

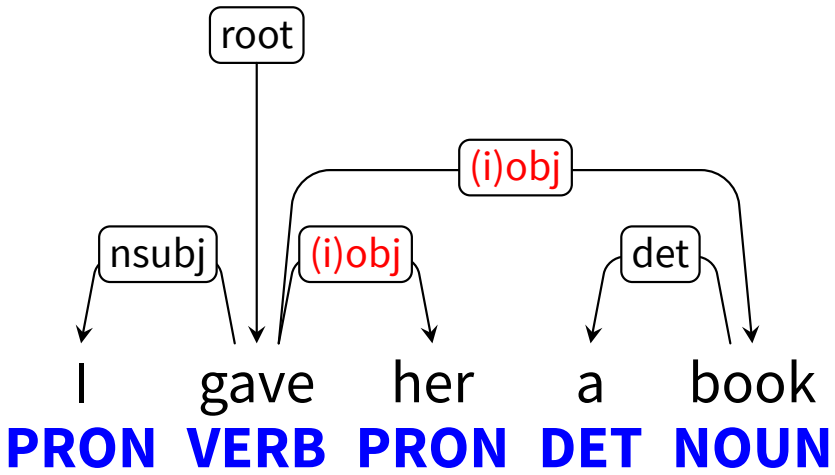
- Semantic roles needed when starting a new language
- Argument-adjunct needed to describe exceptions *(the whole week)*

Intransitive Predicates

- Just one core argument
  - We already “know” how to find out if there are two
- ⇒ function S
  - Regardless of semantic role:
    - *John runs.*
    - *John sleeps.*
    - *John falls.*
- Then define:
  - function S ⇒ **nsubj**

Ditransitive Predicates

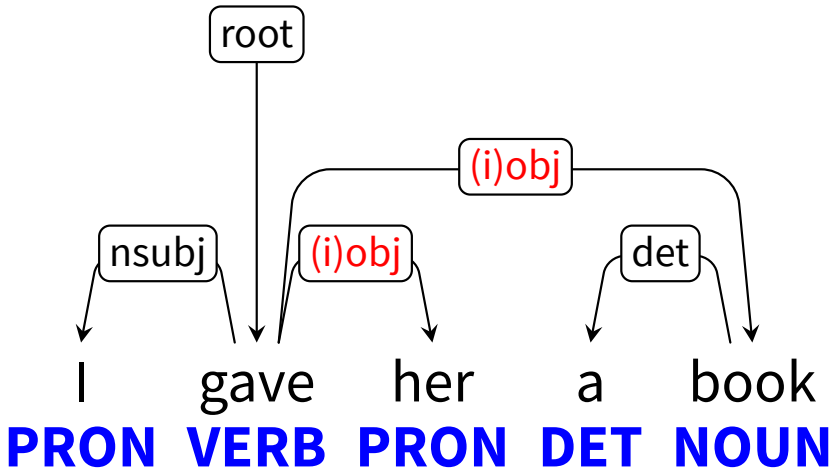
- Three core arguments
- Is one of them “least core”? ⇒ **`iobj`**
- (Alternatively, we could look at the semantic roles once again.)



- Passivization:
  - *She was given a book by me.*
  - *?A book was given her by me.*

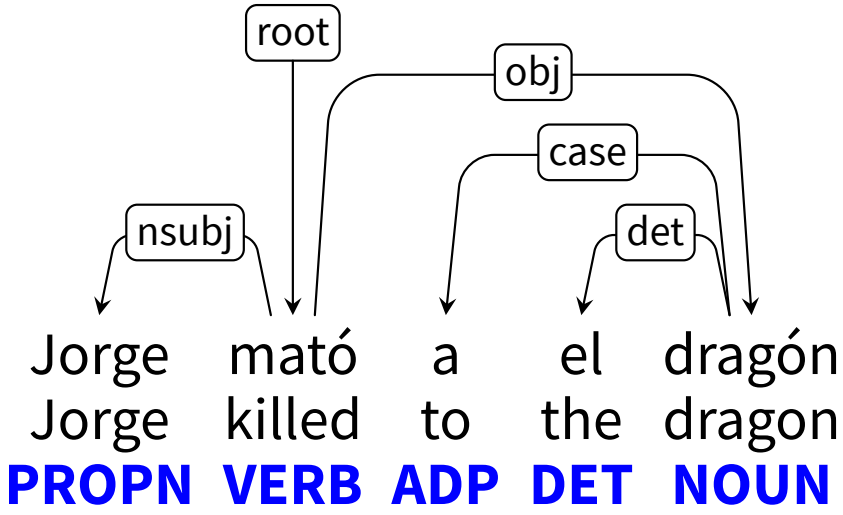
Ditransitive Predicates

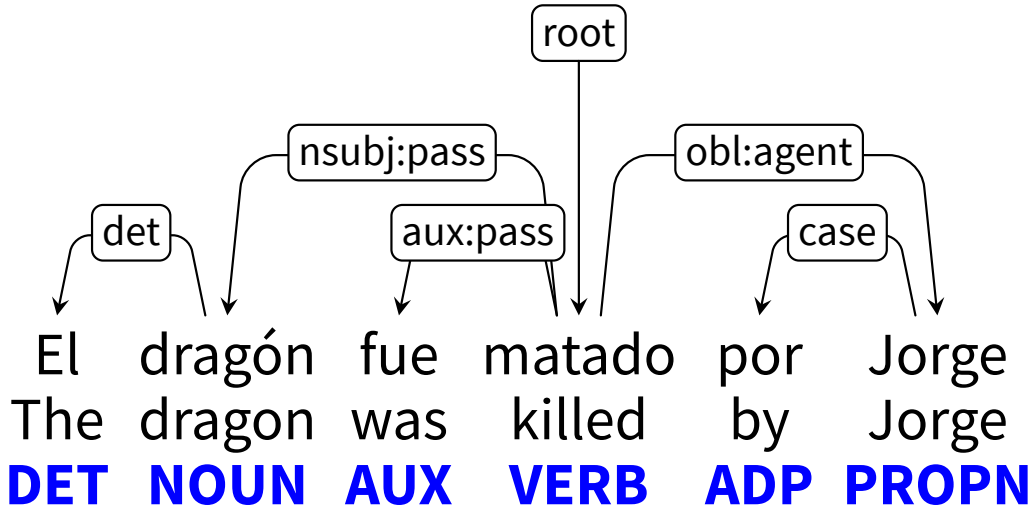
- Three core arguments
- Is one of them “least core”? ⇒ **`iobj`**
- (Alternatively, we could look at the semantic roles once again.)



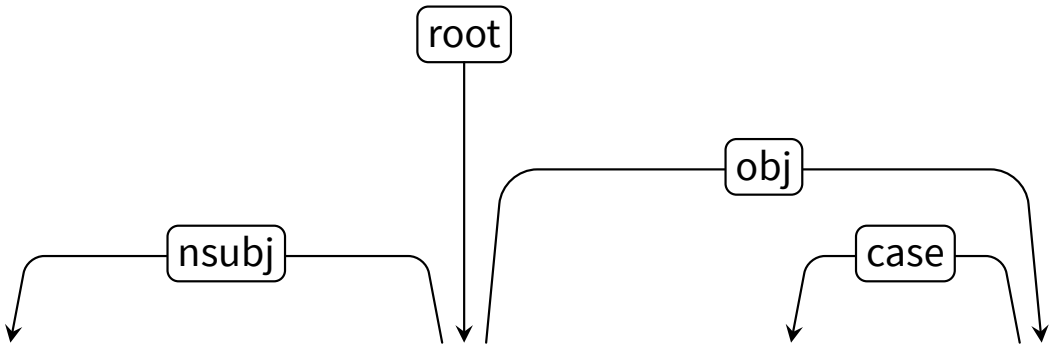
- Andrews (2007): *the status of the notion of ‘indirect object’ is problematic and difficult to sort out. The top priority is to work out what properties recipients and themes do and do not share with P arguments of primary transitive verbs.*

Spanish


root
obj
case
nsubj
det
Jorge mató a el dragón
Jorge killed to the dragon
PROPN VERB ADP DET NOUN


root
nsubj:pass
obl:agent
det
aux:pass
case
El dragón fue matado por Jorge
The dragon was killed by Jorge
DET NOUN AUX VERB ADP PROPN

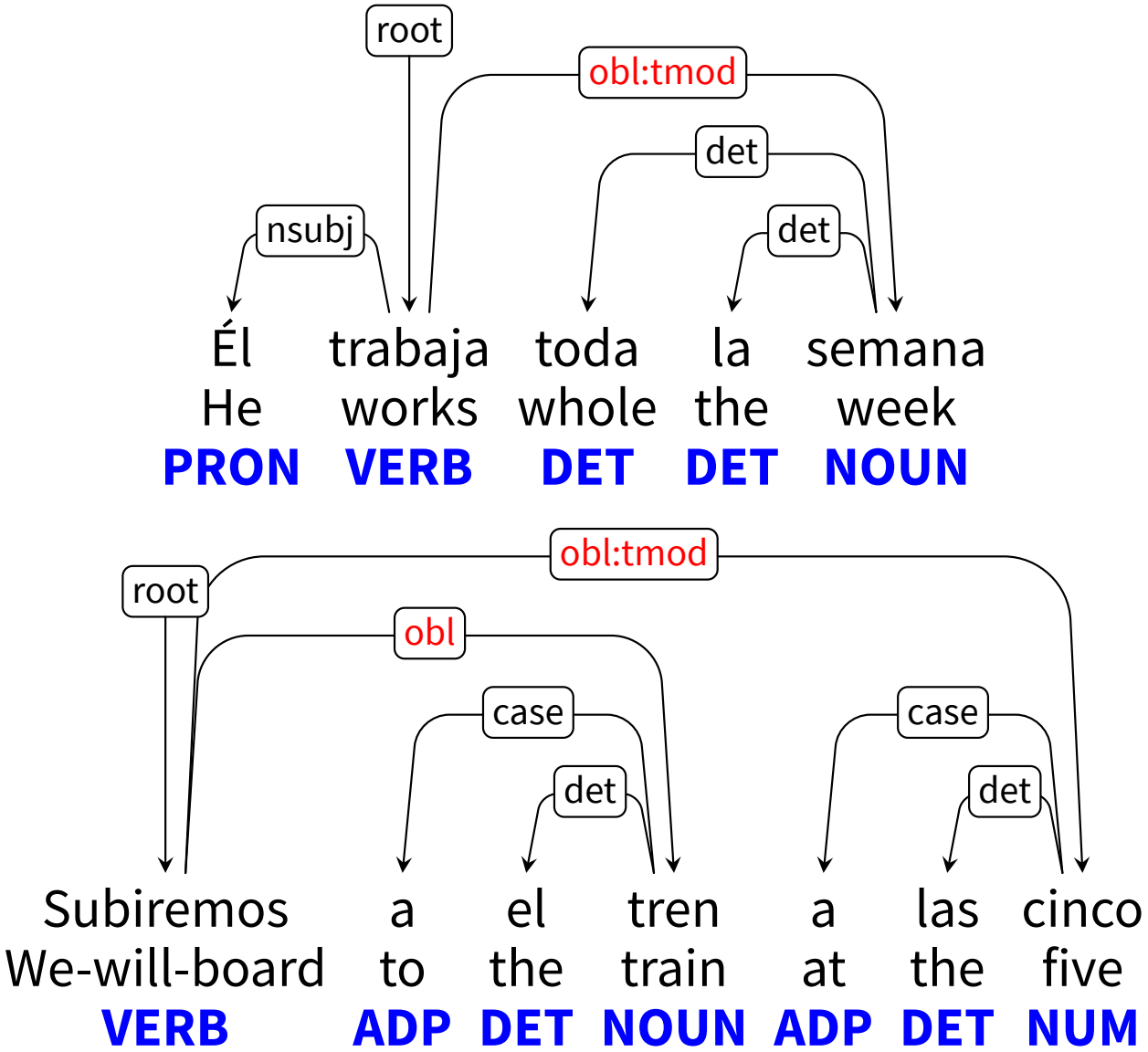
Spanish Transitive Clauses



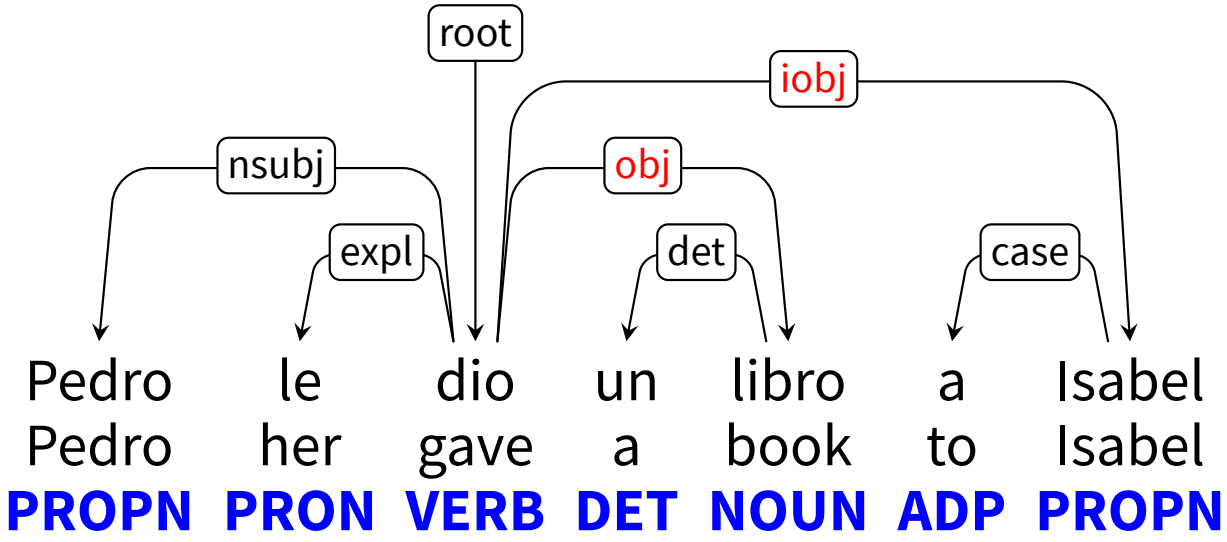
(a)

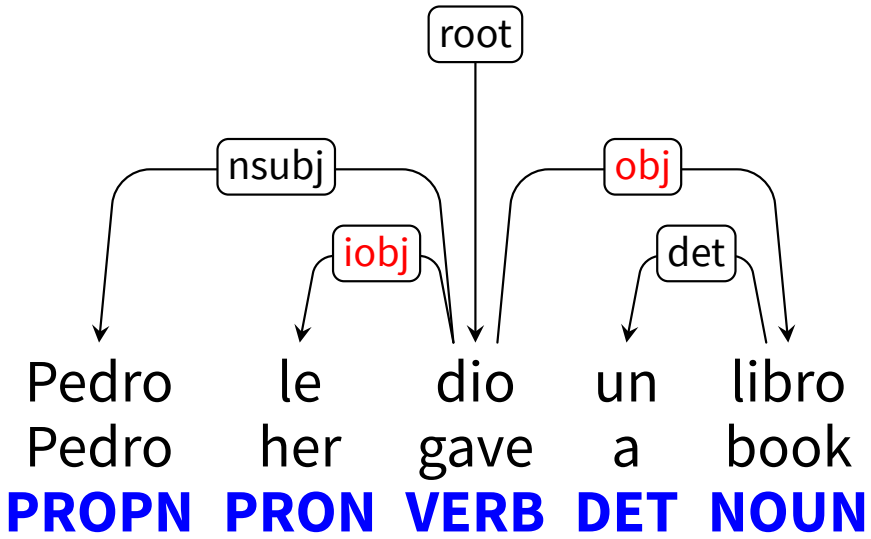
nominal	**VERB**	**(ADP)**	**nominal**
Case=Nom	**Voice=Act(,Pass)**		**Case=Acc**
bare NP			*(or bare NP)*
~~pre-verb~~	*declarative clause*		*~~post-verb~~*
cross-ref on verb	*← agreement*		

Spanish Adjunct Exceptions

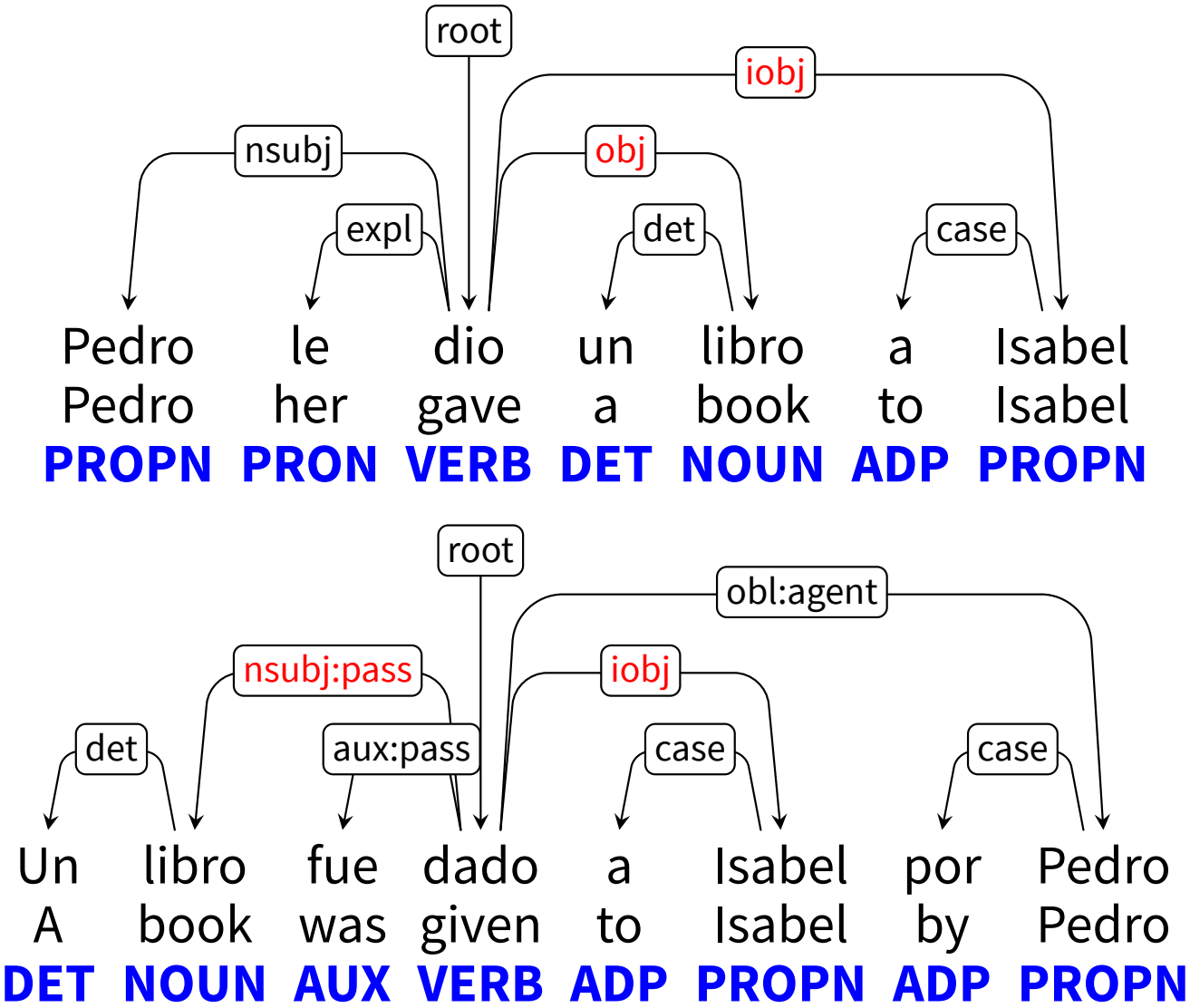


Spanish Ditransitive Clauses

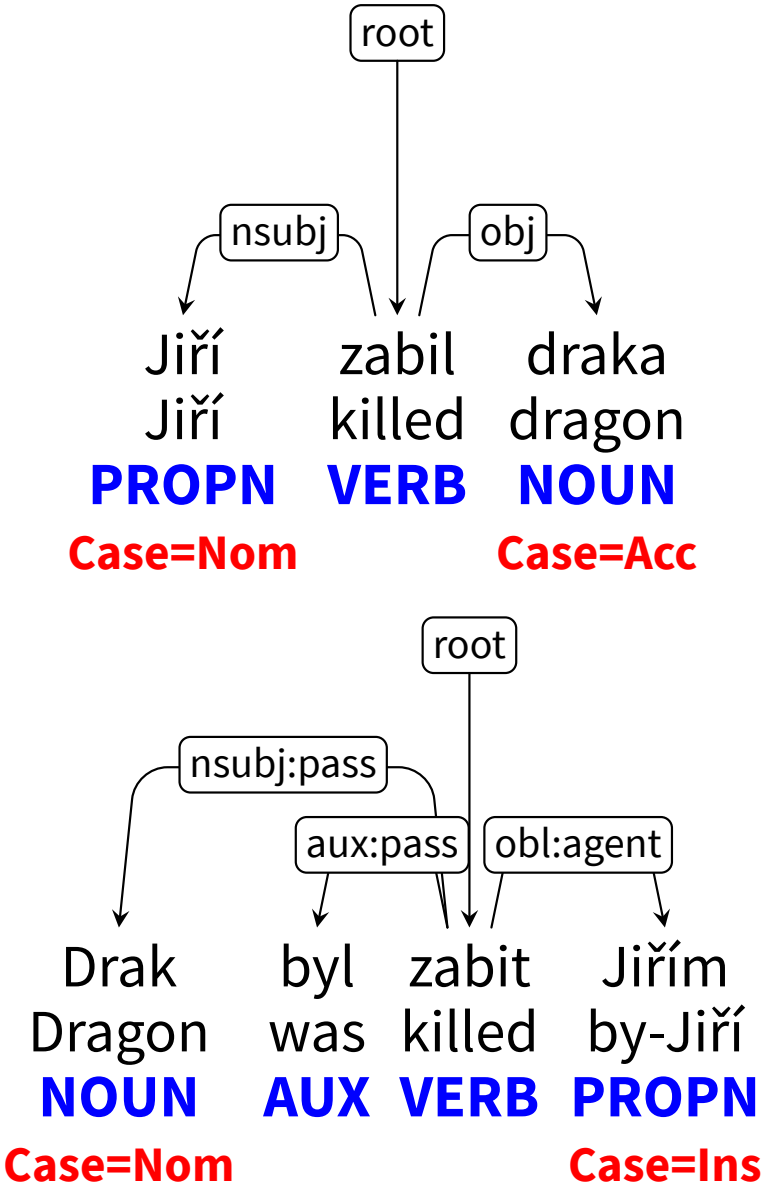




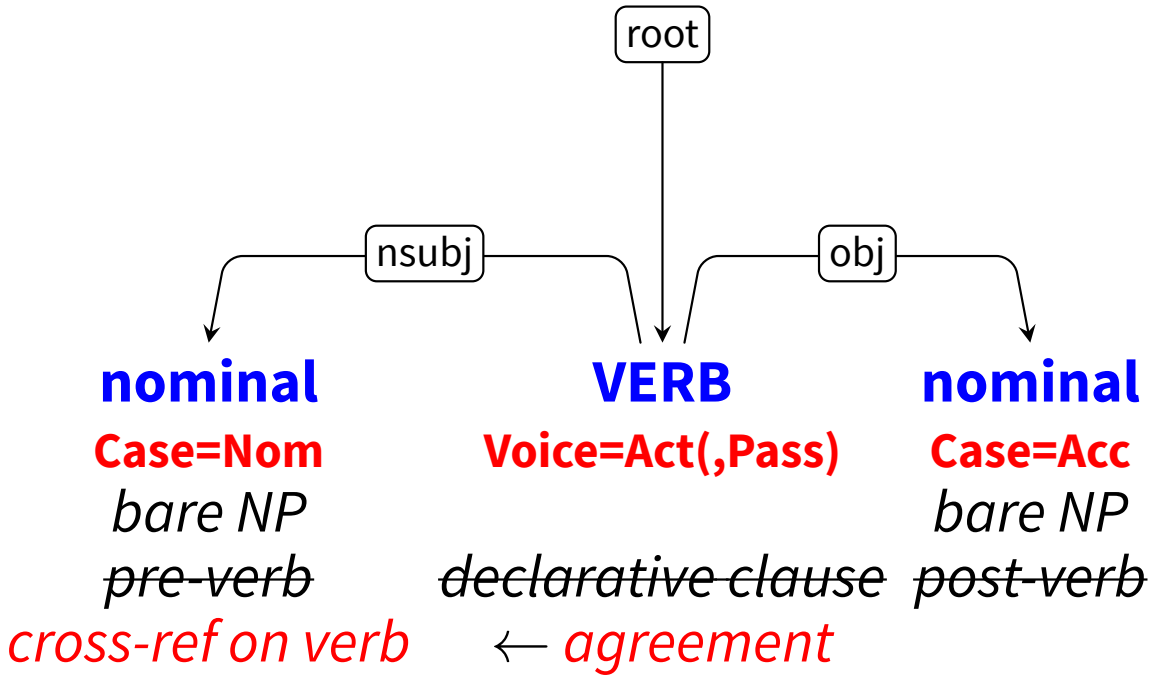
Spanish Ditransitive Clauses



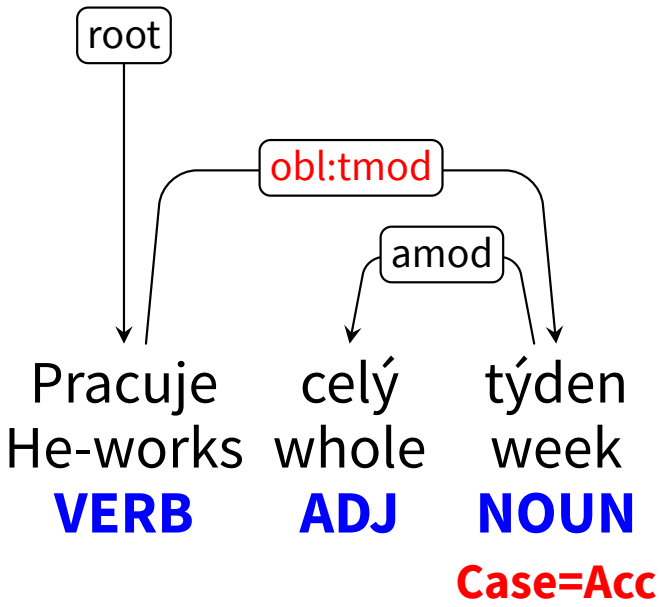
Czech


root
nsubj
obj
Jiří
Jiří
PROPN
Case=Nom
zabil
killed
VERB
draka
dragon
NOUN
Case=Acc
root
nsubj:pass
aux:pass
obl:agent
Drak
Dragon
NOUN
Case=Nom
byl
was
AUX
zabit
killed
VERB
Jiřím
by-Jiří
PROPN
Case=Ins

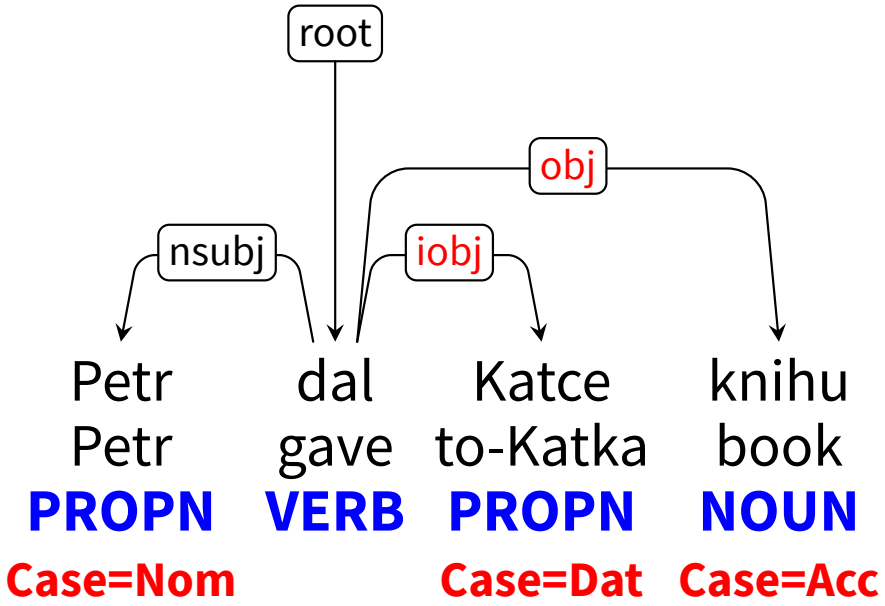
Czech Transitive Clauses

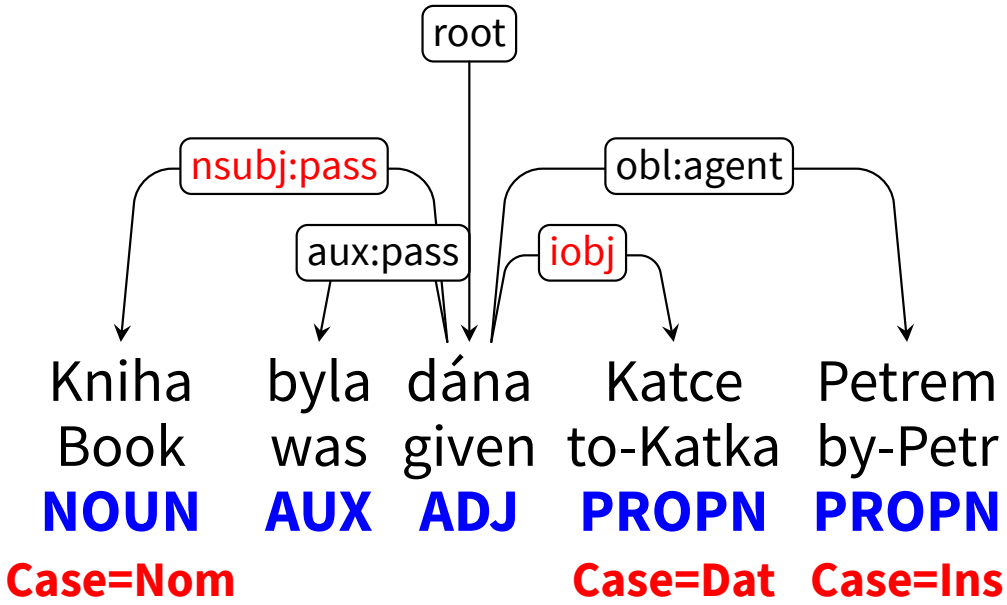


Czech Adjunct Exceptions

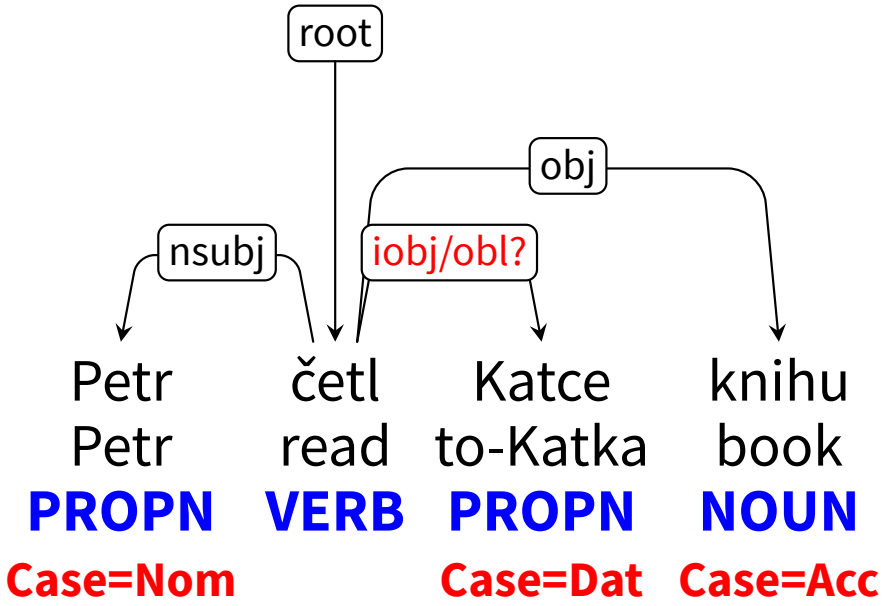


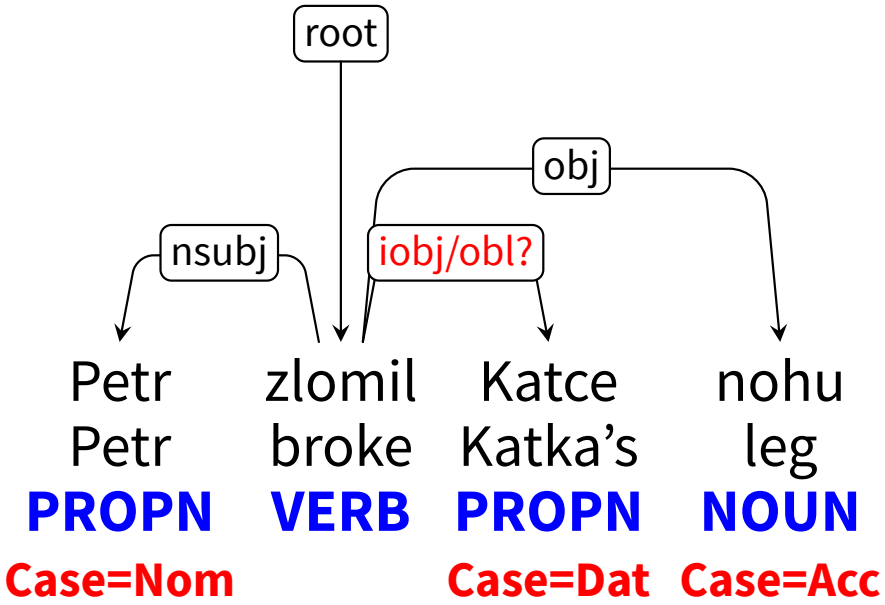
Czech Ditransitive Clauses



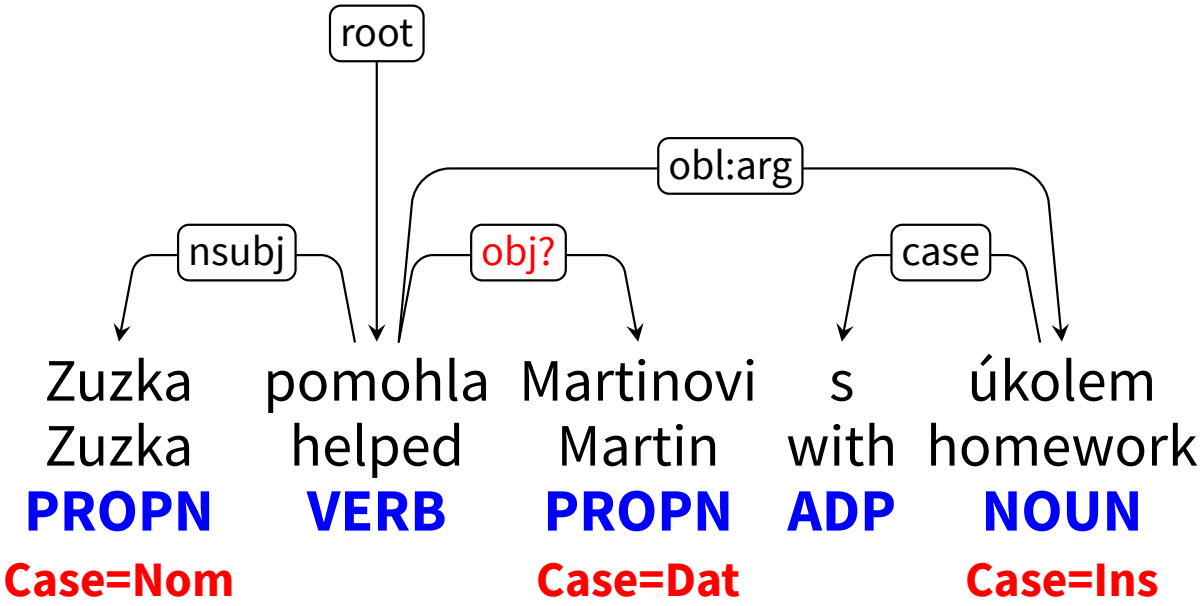


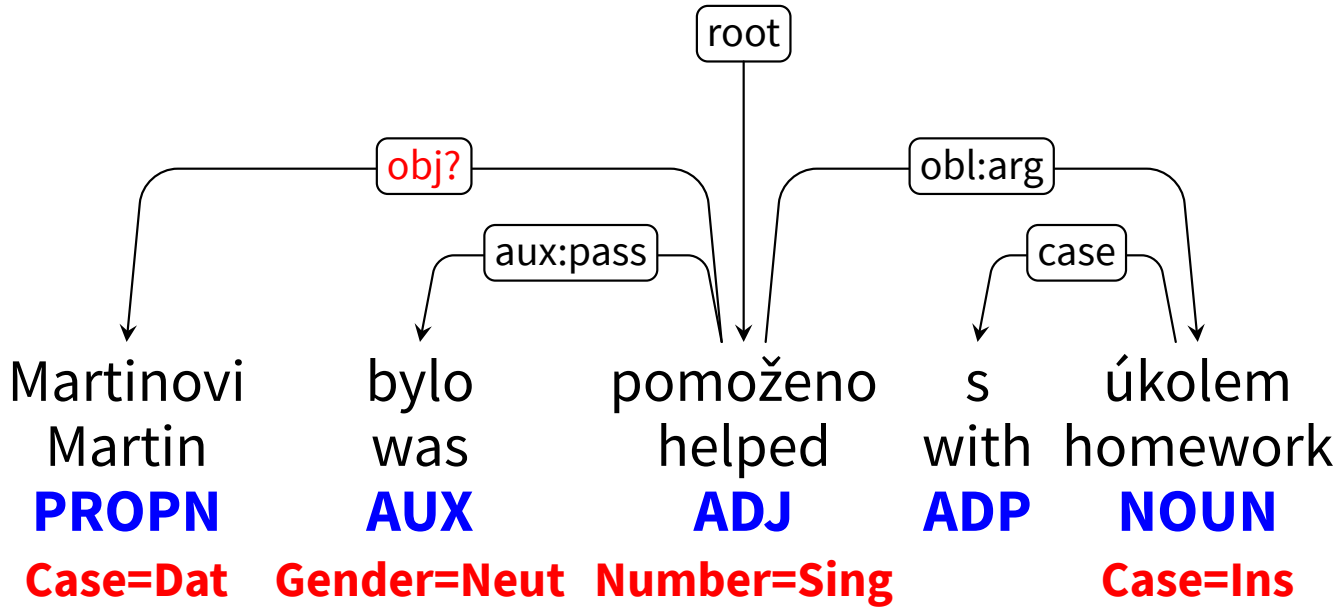
Dative: Recipient vs. Beneficiary



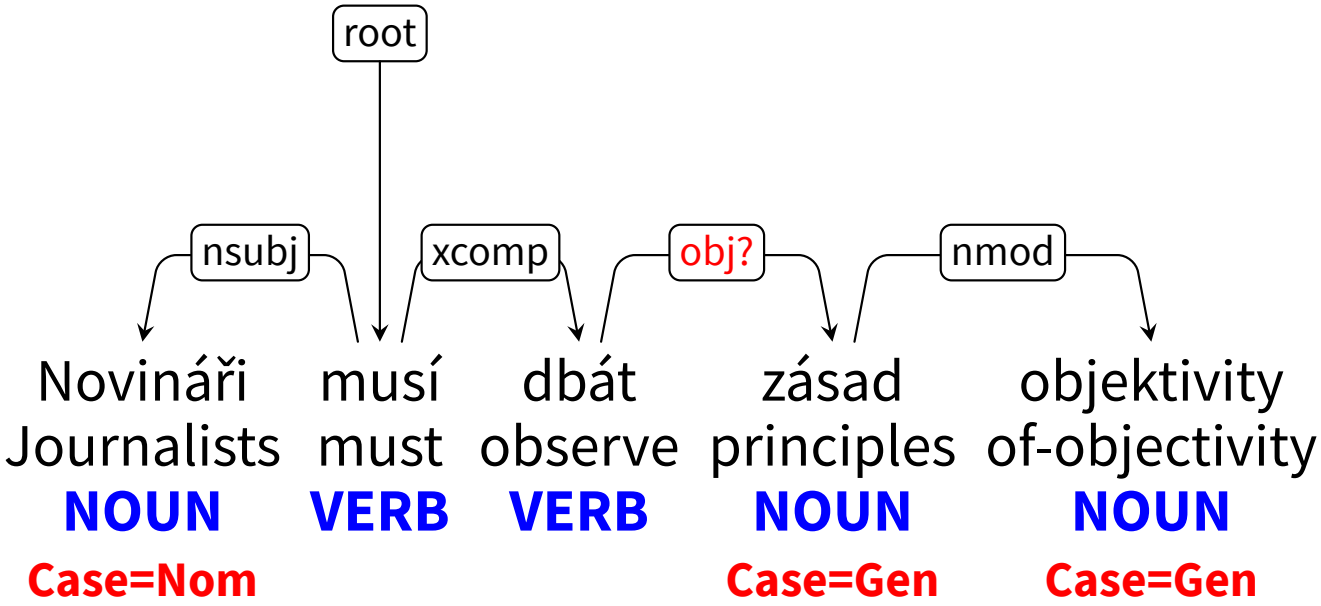


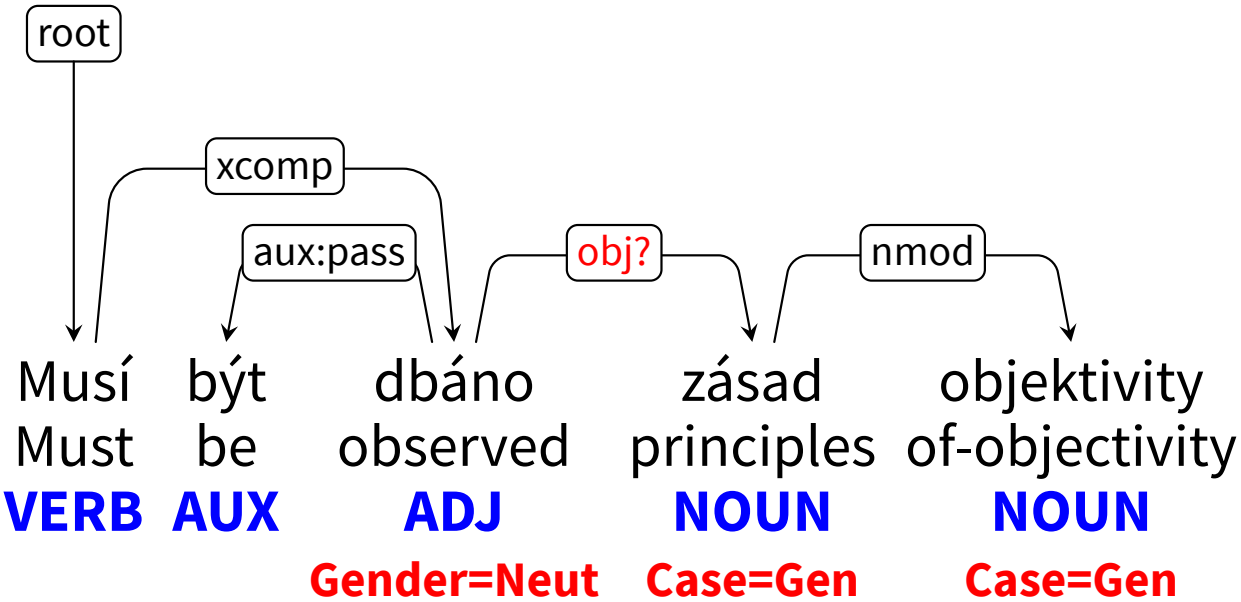
Monotransitive with Dative?



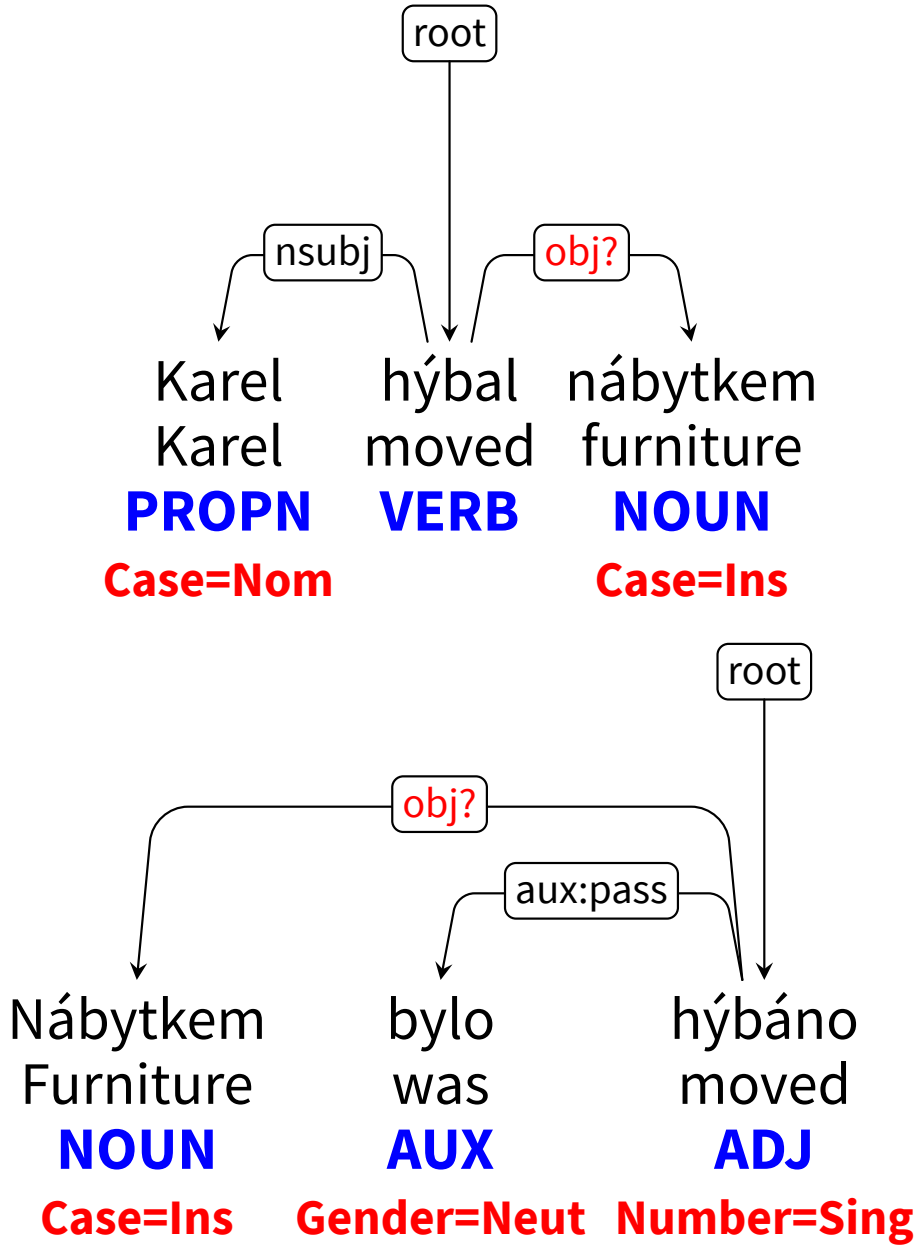


Monotransitive with Genitive?

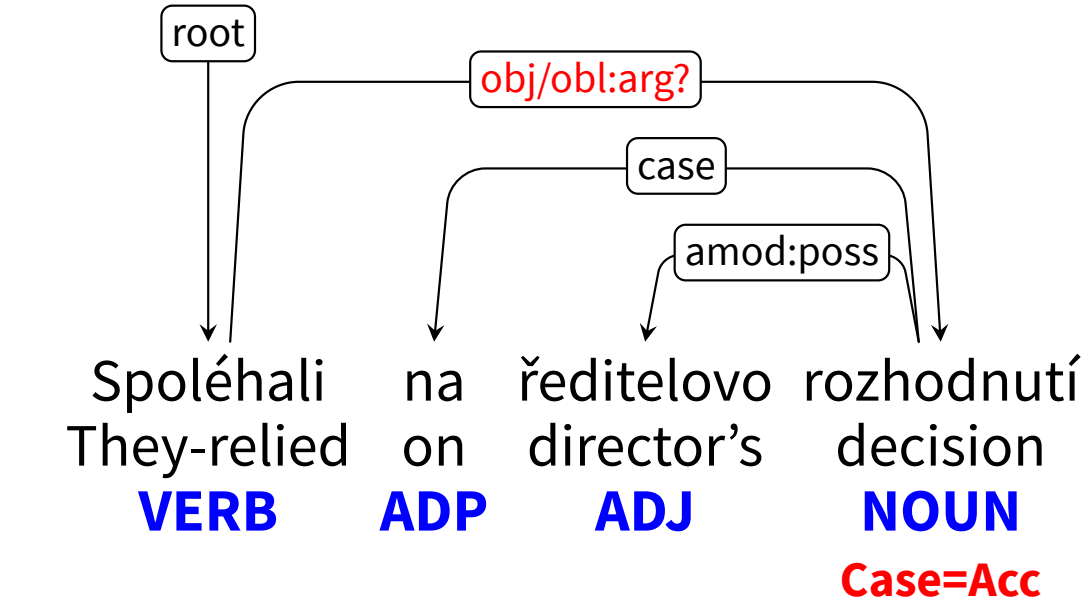


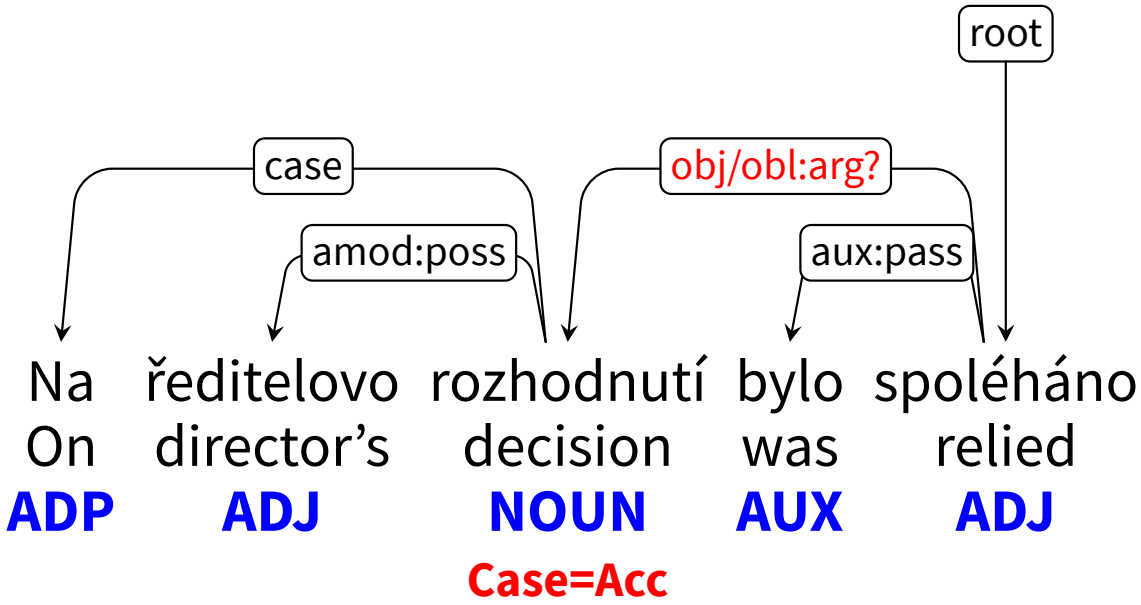


Monotransitive with Instrumental?



Monotransitive with Preposition?





Tentative Summary 2

- There is a core-oblique scale:
- **Nom** > **Acc** > **Gen,Dat** > **Ins** > **preposition**
- Where is the borderline?

Tentative Summary 2

- There is a core-oblique scale:
- **Nom** > **Acc** > **Gen,Dat** > **Ins** > **preposition**
- Where is the borderline?
- UD Czech 1.0: object = argument
  - Nom, Acc, Gen, Dat, Ins, ADP > “adverbial”

Tentative Summary 2

- There is a core-oblique scale:
- **Nom** > **Acc** > **Gen,Dat** > **Ins** > **preposition**
- Where is the borderline?
- UD Czech 1.0: object = argument
  - Nom, Acc, Gen, Dat, Ins, ADP > “adverbial”
- UD Czech 2.1: bare NP > PP
  - Nom, Acc, Gen, Dat, Ins > ADP + adjuncts

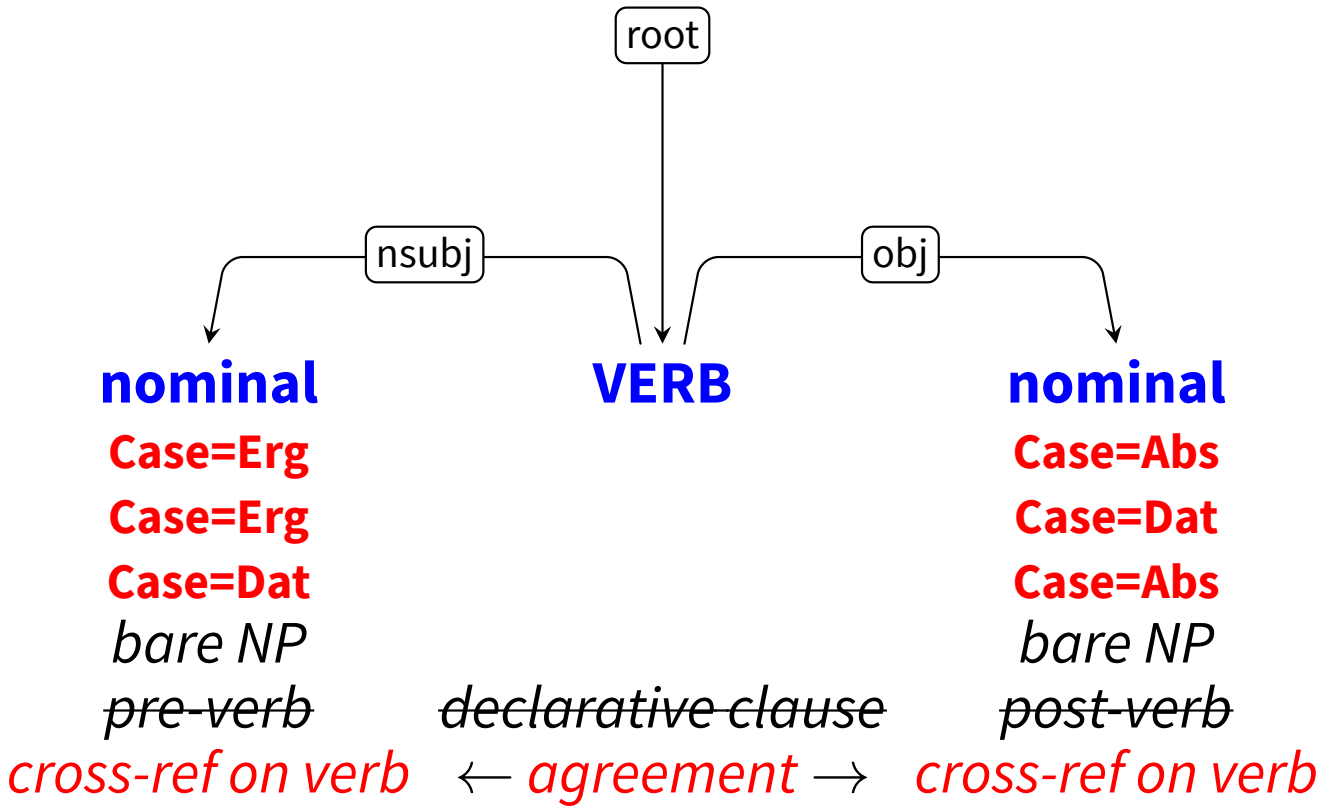
Tentative Summary 2

- There is a core-oblique scale:
- **Nom** > **Acc** > **Gen,Dat** > **Ins** > **preposition**
- Where is the borderline?
- UD Czech 1.0: object = argument
  - Nom, Acc, Gen, Dat, Ins, ADP > “adverbial”
- UD Czech 2.1: bare NP > PP
  - Nom, Acc, Gen, Dat, Ins > ADP + adjuncts
- But perhaps we should go even further?
  - Nom, Acc > Gen, Dat, Ins, ADP + adjuncts

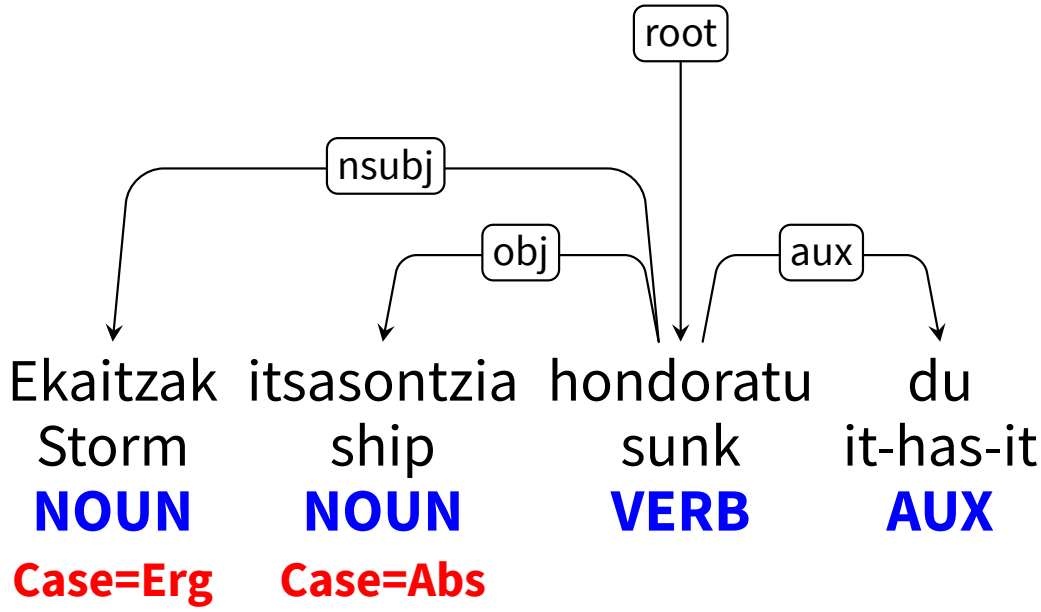
Tentative Summary 2

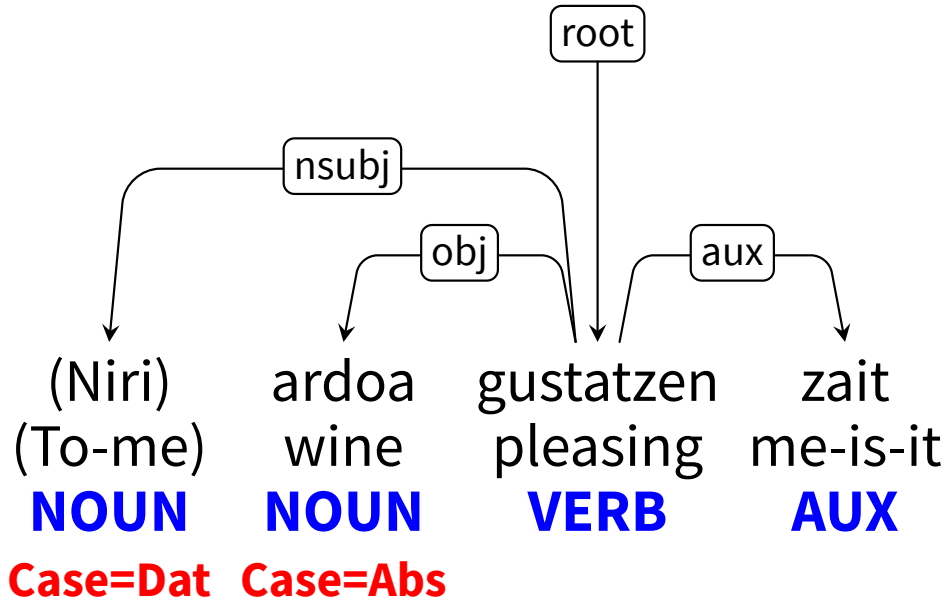
- There is a core-oblique scale:
- **Nom** > **Acc** > **Gen,Dat** > **Ins** > **preposition**
- Where is the borderline?
- UD Czech 1.0: object = argument
  - Nom, Acc, Gen, Dat, Ins, ADP > "adverbial"
- UD Czech 2.1: bare NP > PP
  - Nom, Acc, Gen, Dat, Ins > ADP + adjuncts
- But perhaps we should go even further?
  - Nom, Acc > Gen, Dat, Ins, ADP + adjuncts
  - $\Rightarrow$ No ditransitives in Czech!

Basque Transitive Clauses

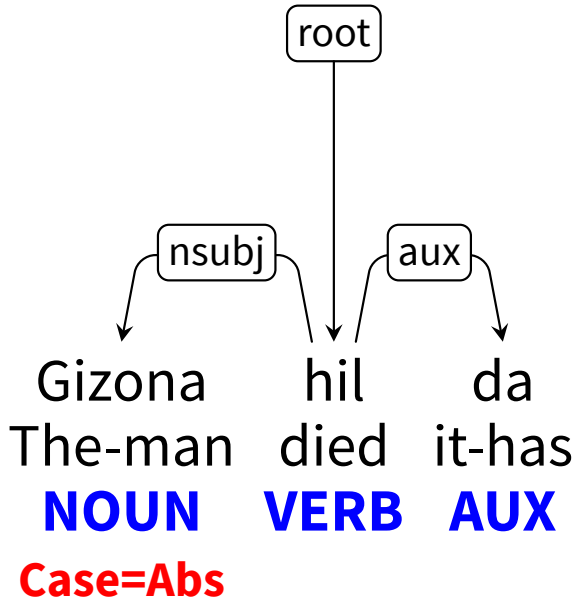


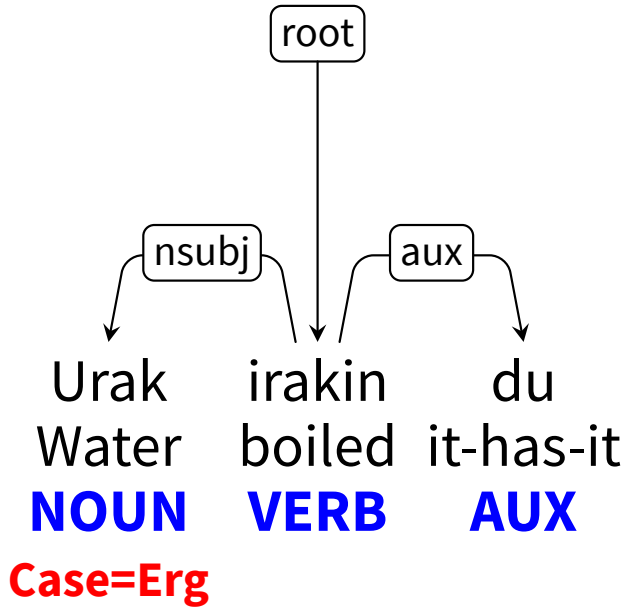
Basque Transitive Clauses



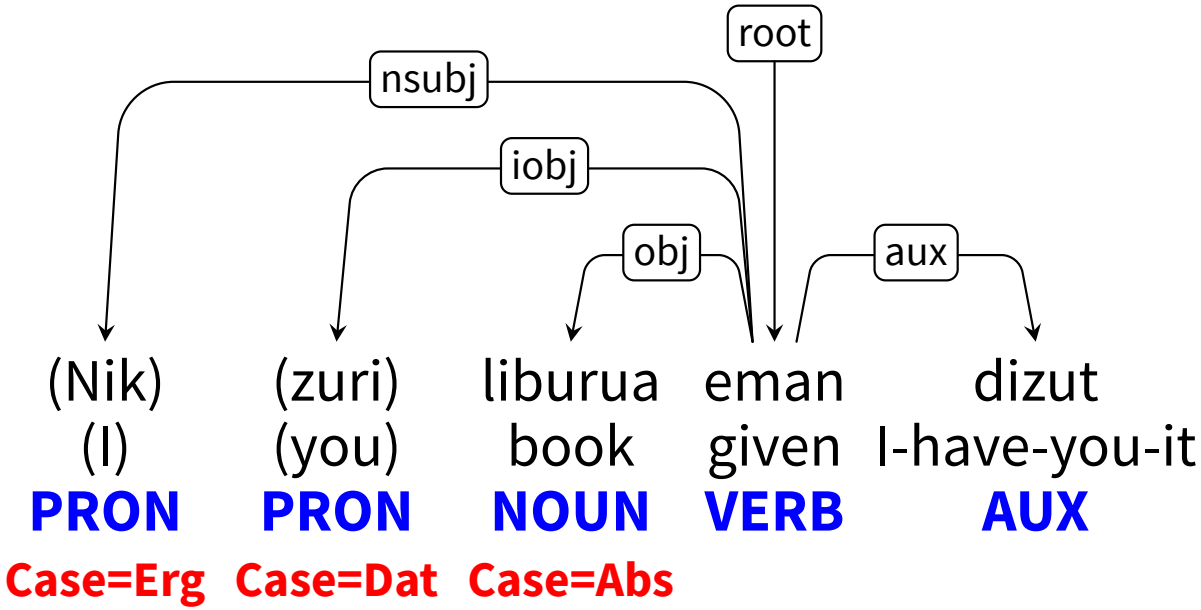


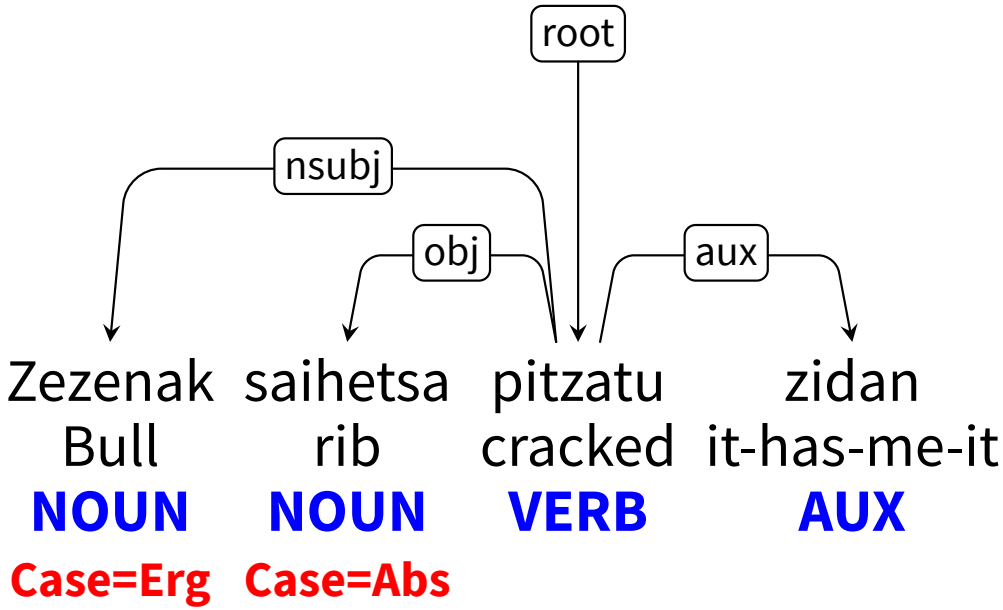
Basque Intransitive Clauses



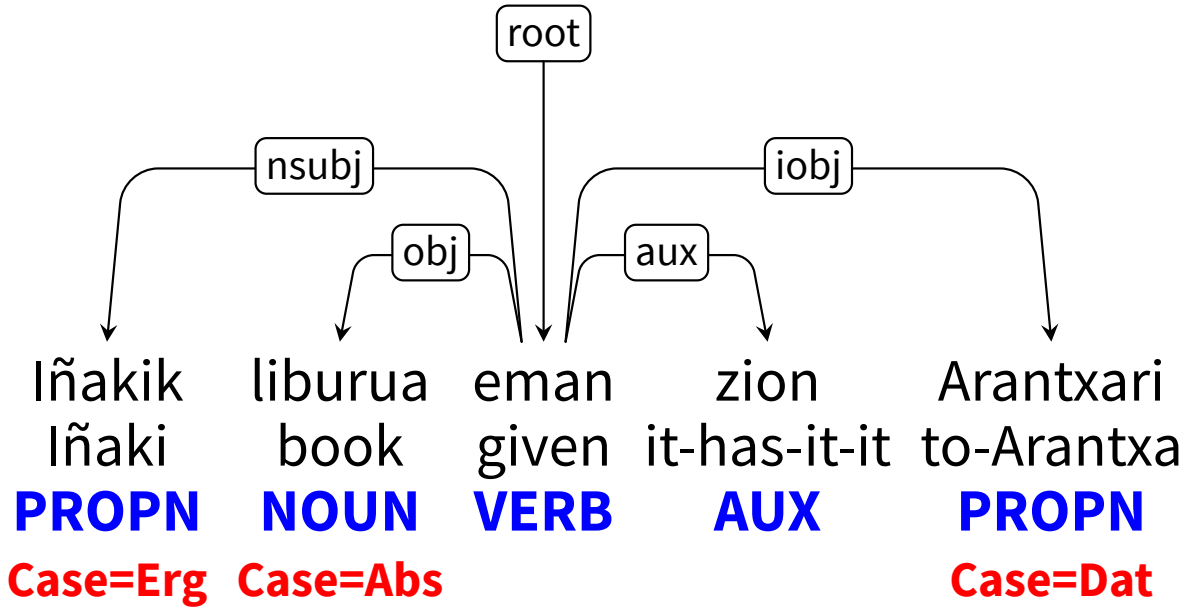


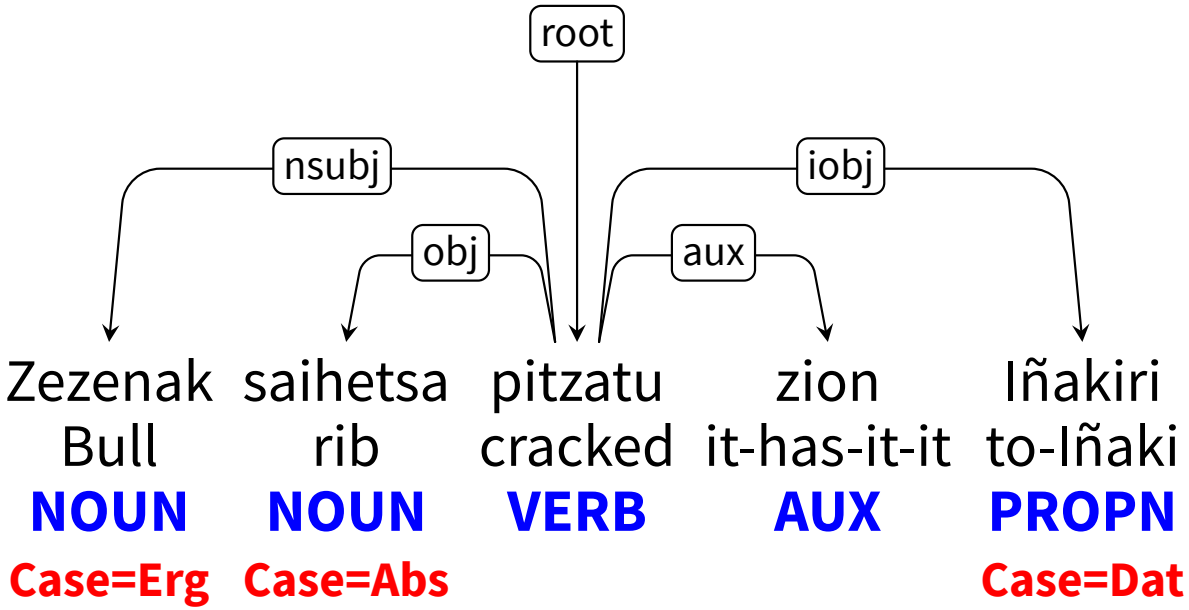
Basque Ditransitive Clauses



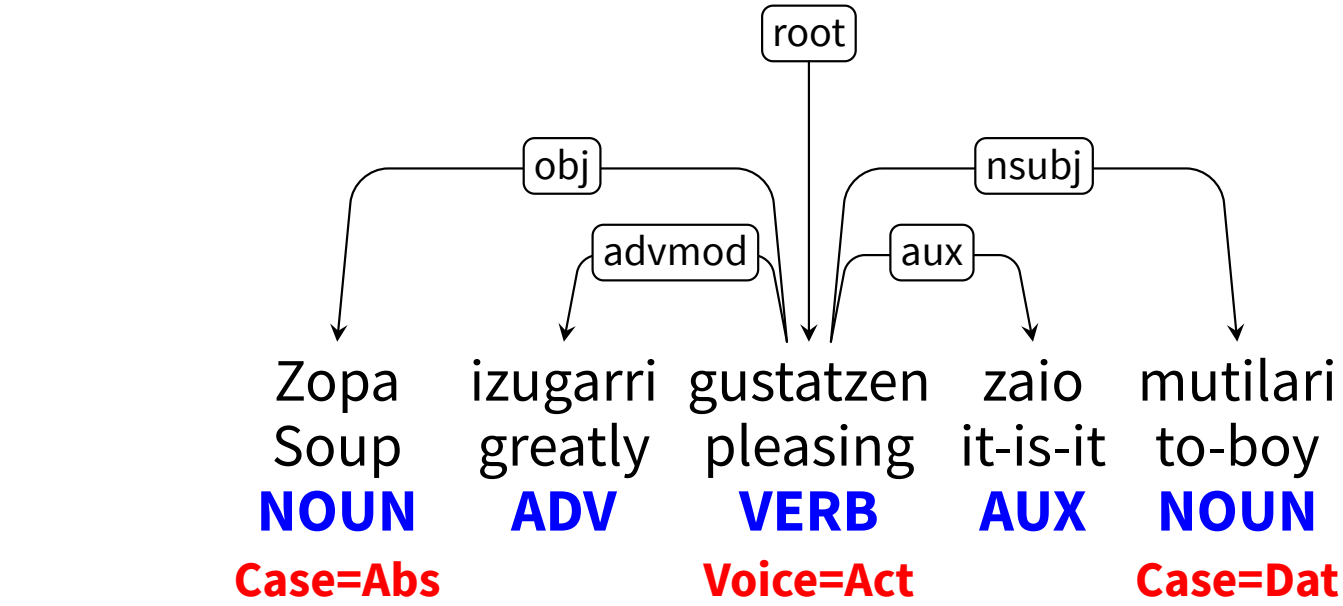


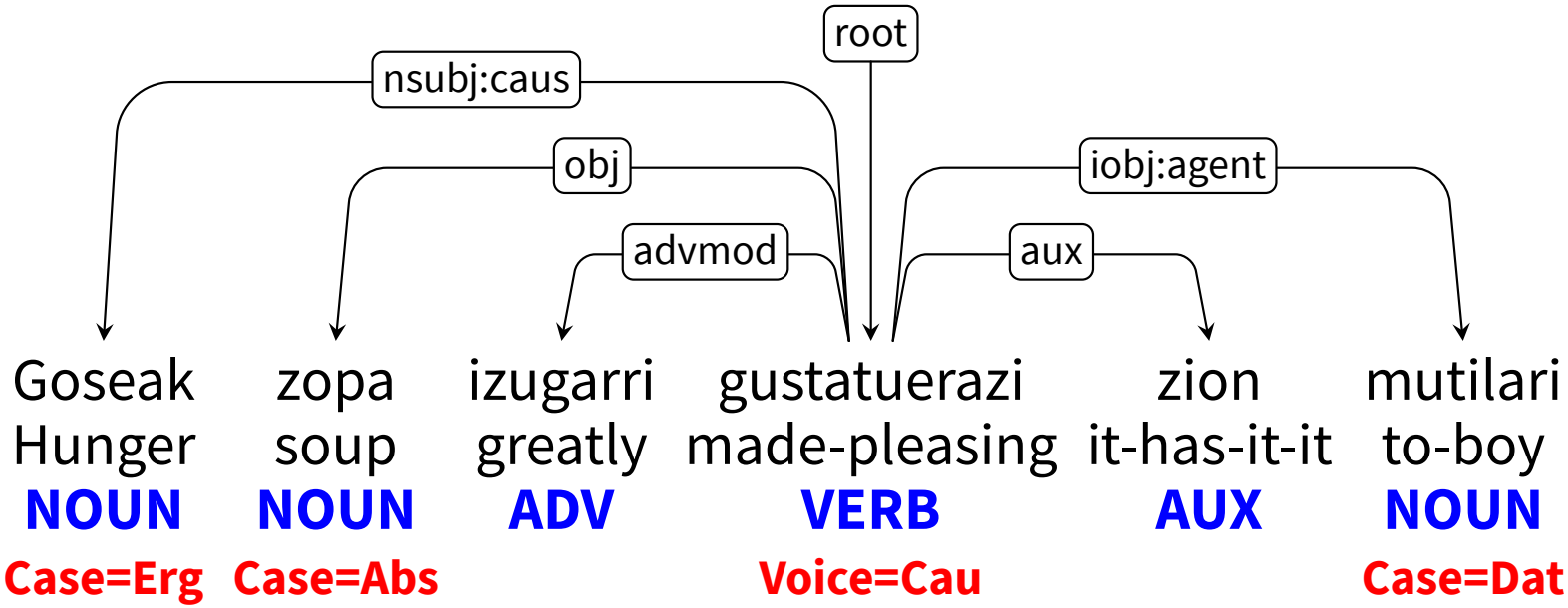
Basque Ditransitive Clauses



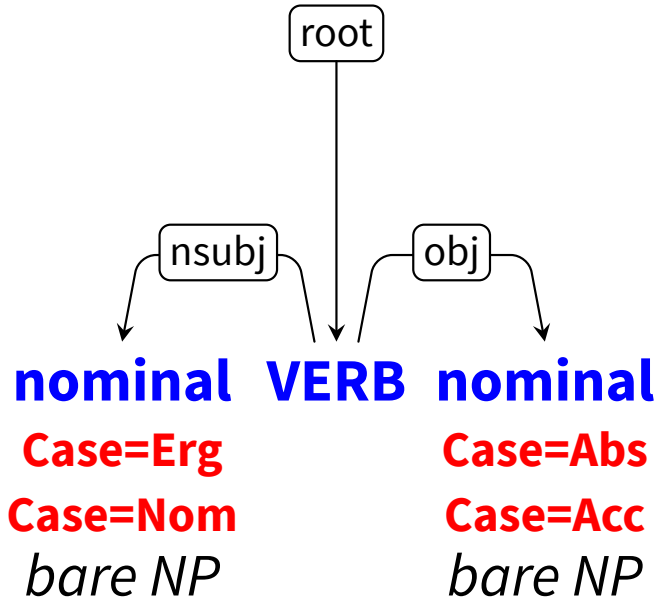


Basque Causative Applied to Dative Subject



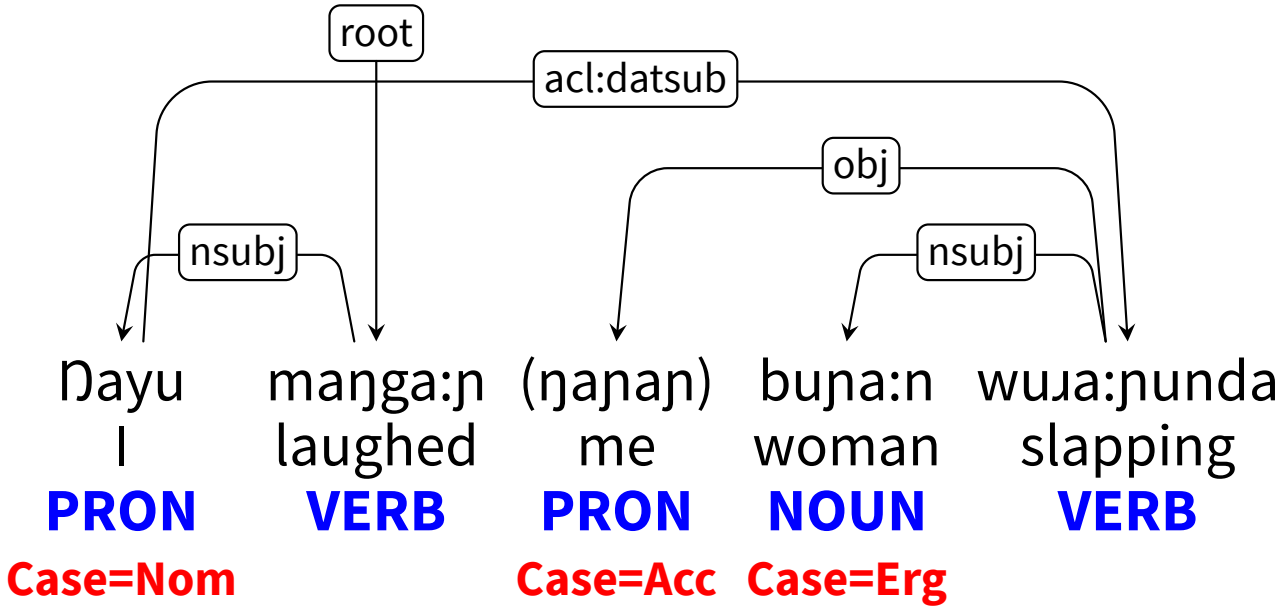


Yidiɲ Transitive Clauses



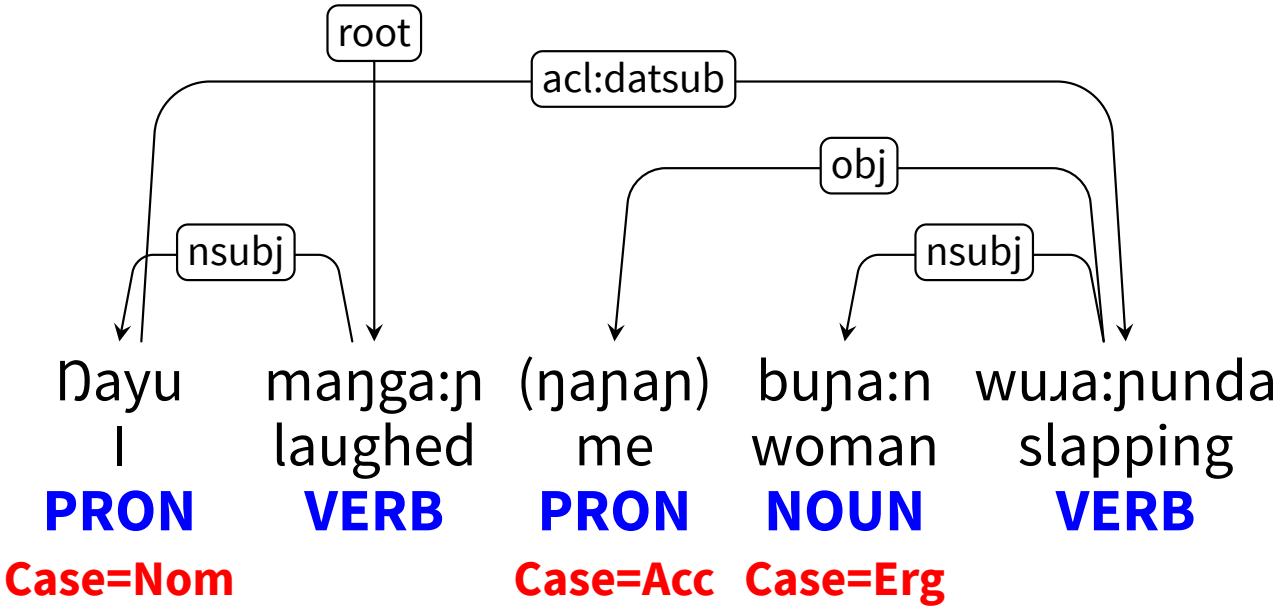
Yidiɲ “Dative” Adnominal Clauses

“I, (who) was slapped by the woman, laughed”



Yidiɲ “Dative” Adnominal Clauses

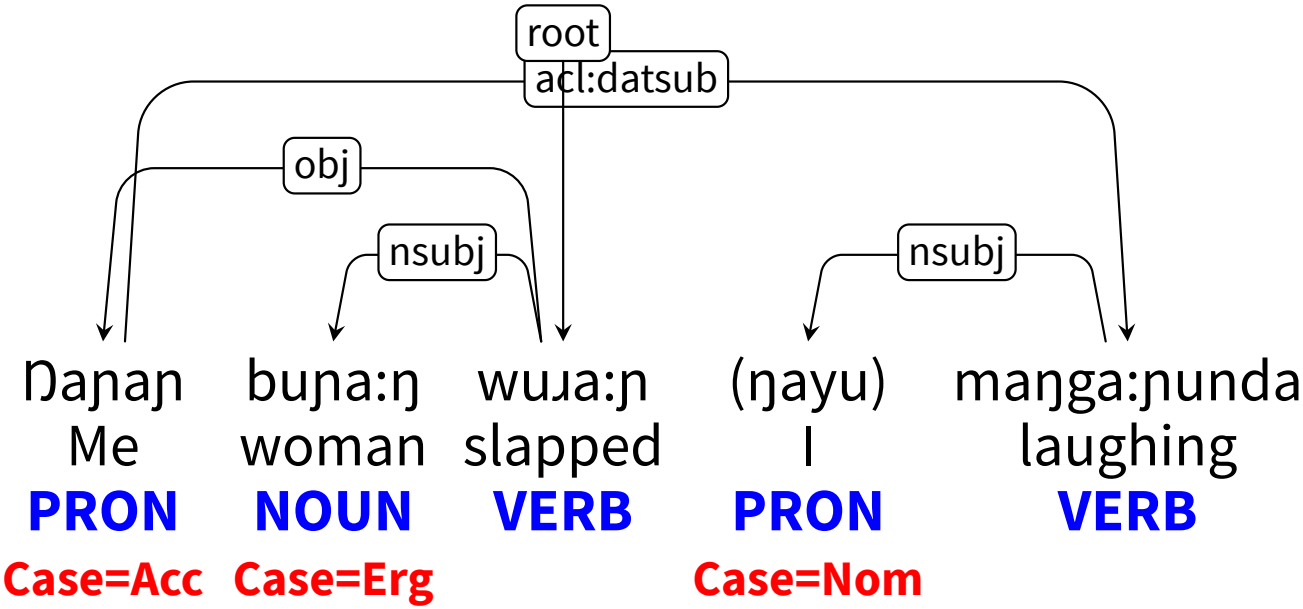
“I, (who) was slapped by the woman, laughed”



The coreferential (and elidable) NP must have S or P function.

Yidiɲ “Dative” Adnominal Clauses

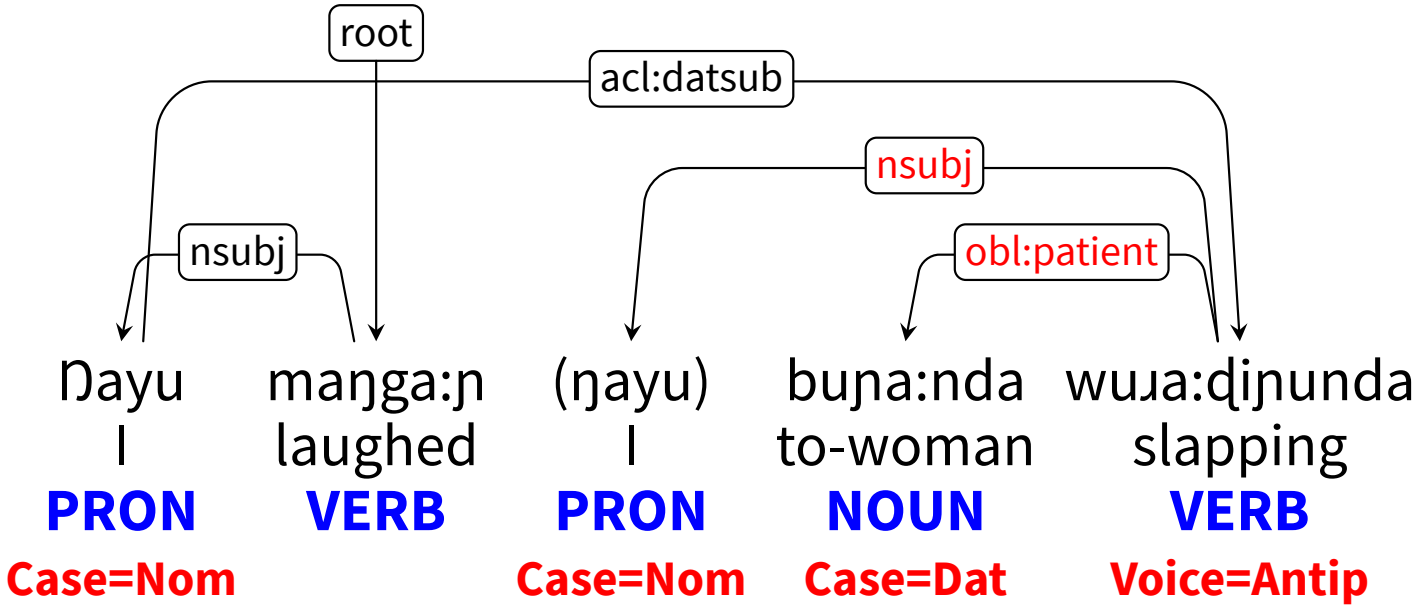
“I, (who) was lauging, was slapped by the woman”



The coreferential (and elidable) NP must have S or P function.

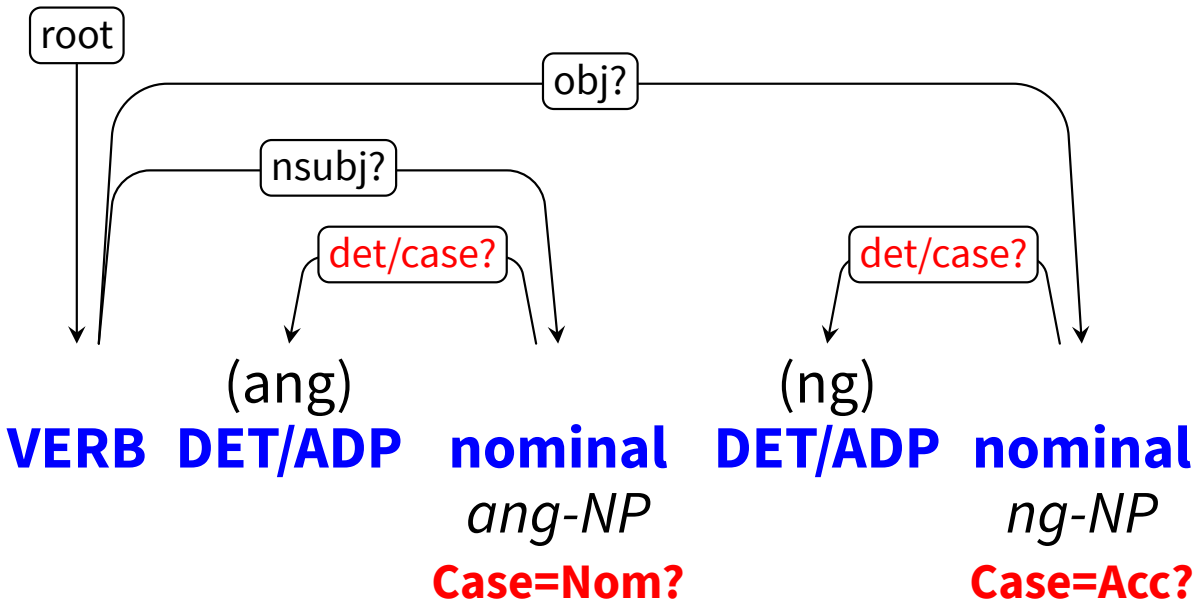
Yidiɲ Antipassive

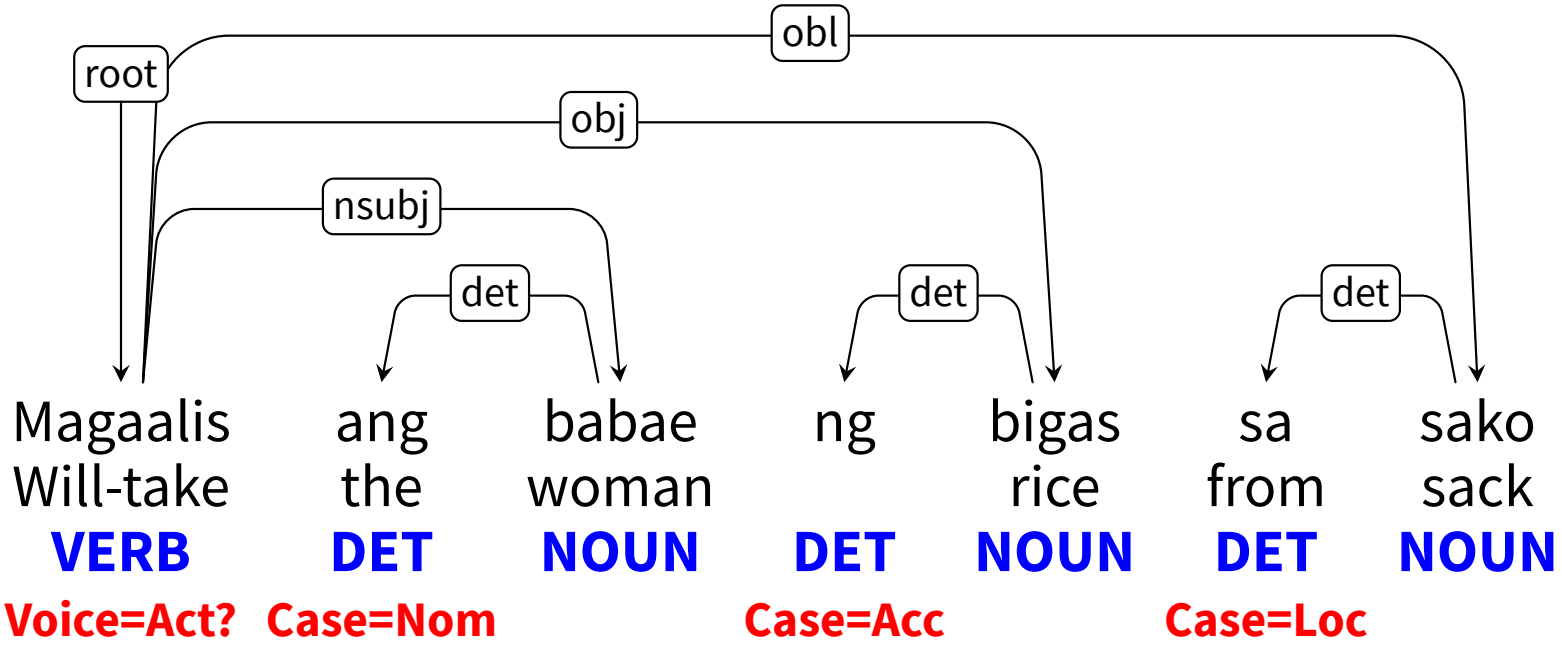
“I, (who) was slapping the woman, laughed”

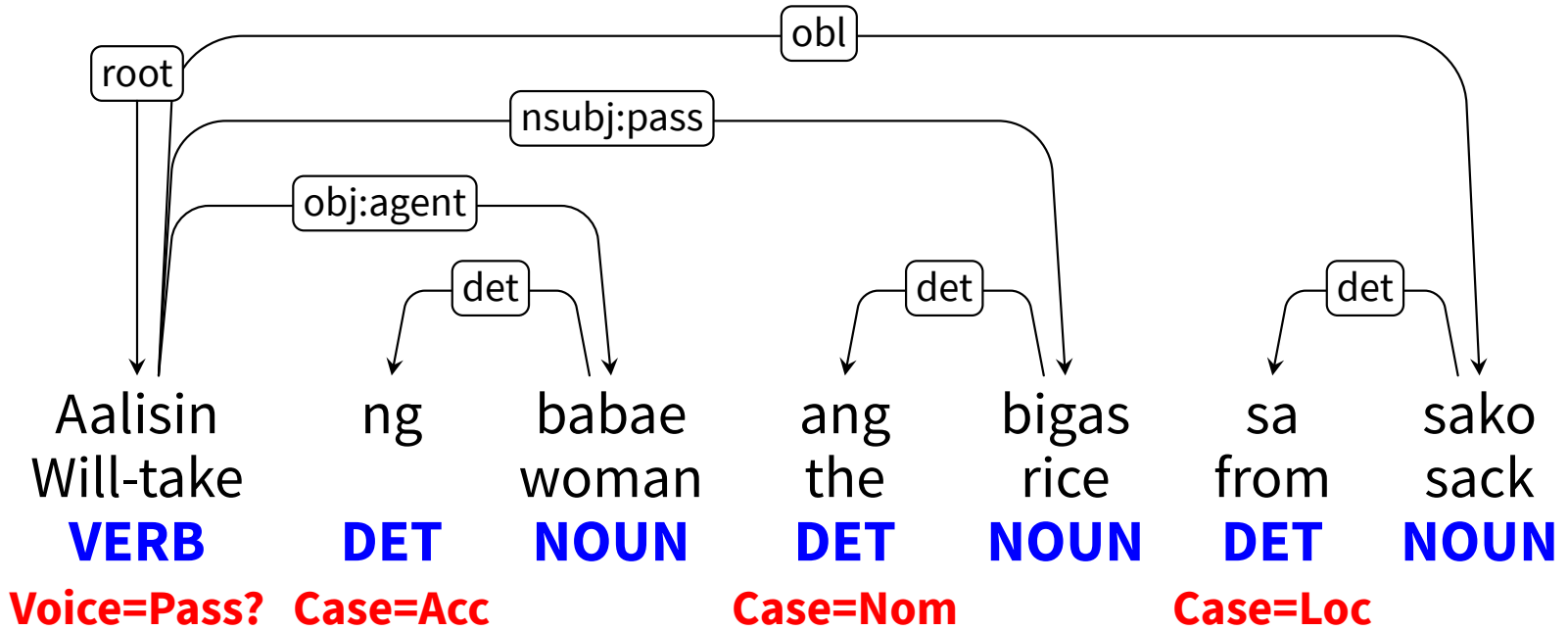


Original P is now oblique and original A is now S.

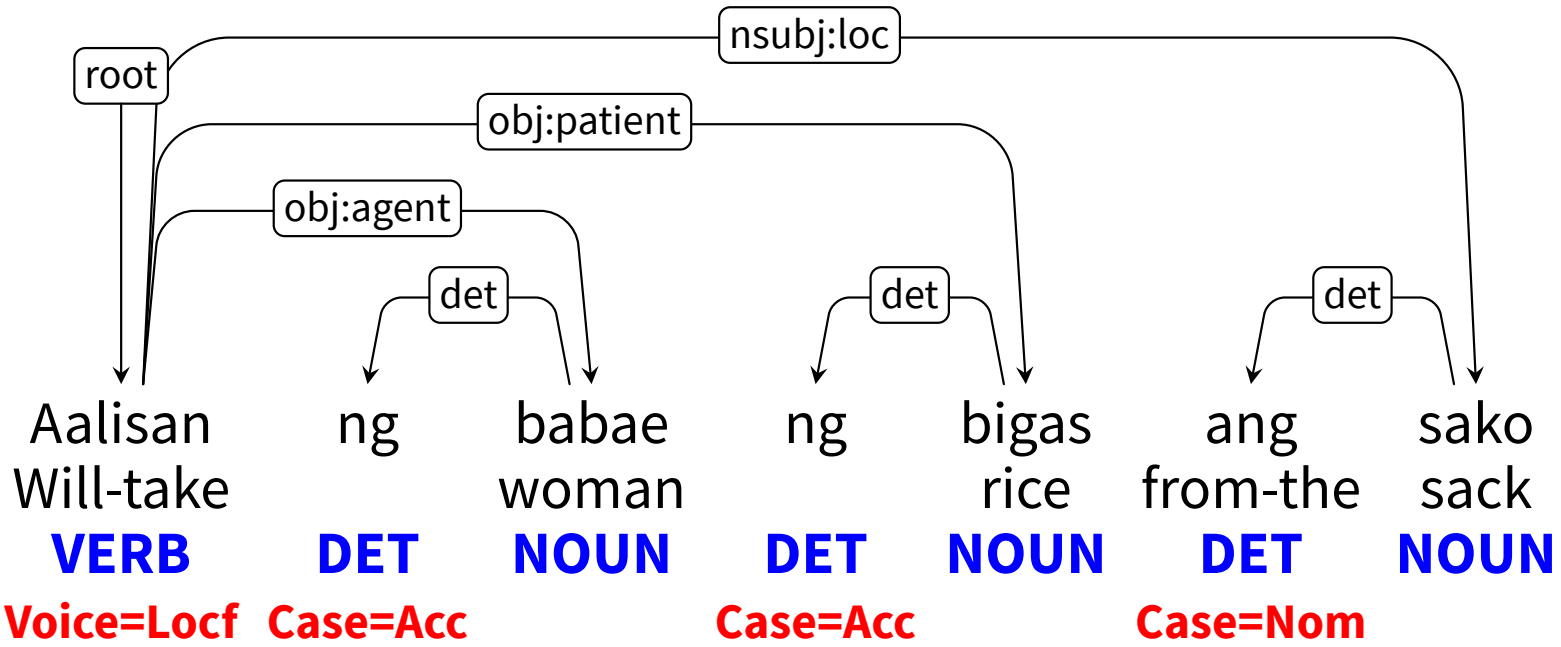
Tagalog Transitive Clauses



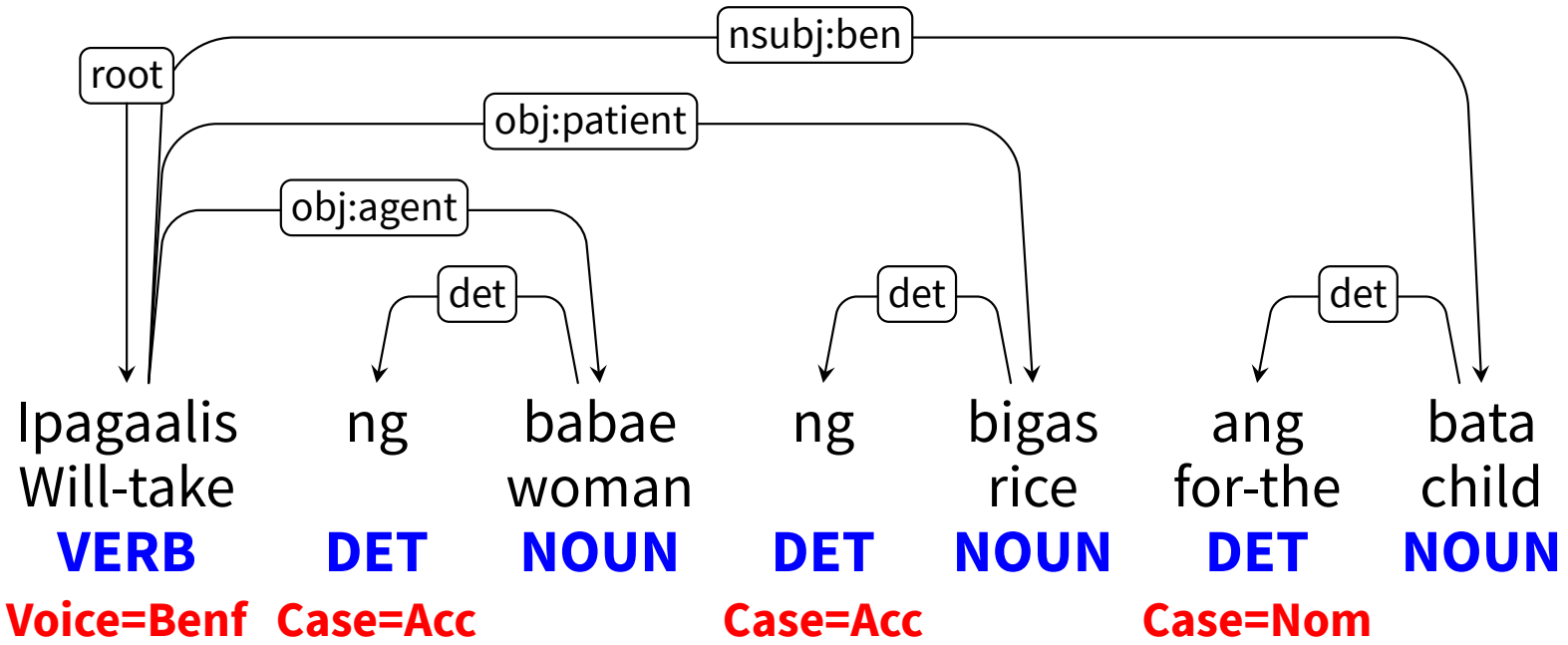

root
obl
obj
nsubj
det
det
det
Magaalis
Will-take
VERB
Voice=Act?
ang
the
DET
Case=Nom
babae
woman
NOUN
ng
DET
Case=Acc
bigas
rice
NOUN
sa
from
DET
Case=Loc
sako
sack
NOUN


root
obl
nsubj:pass
obj:agent
det
det
det
Aalisin
Will-take
VERB
Voice=Pass?
ng
DET
Case=Acc
babae
woman
NOUN
ang
the
DET
Case=Nom
bigas
rice
NOUN
sa
from
DET
Case=Loc
sako
sack
NOUN

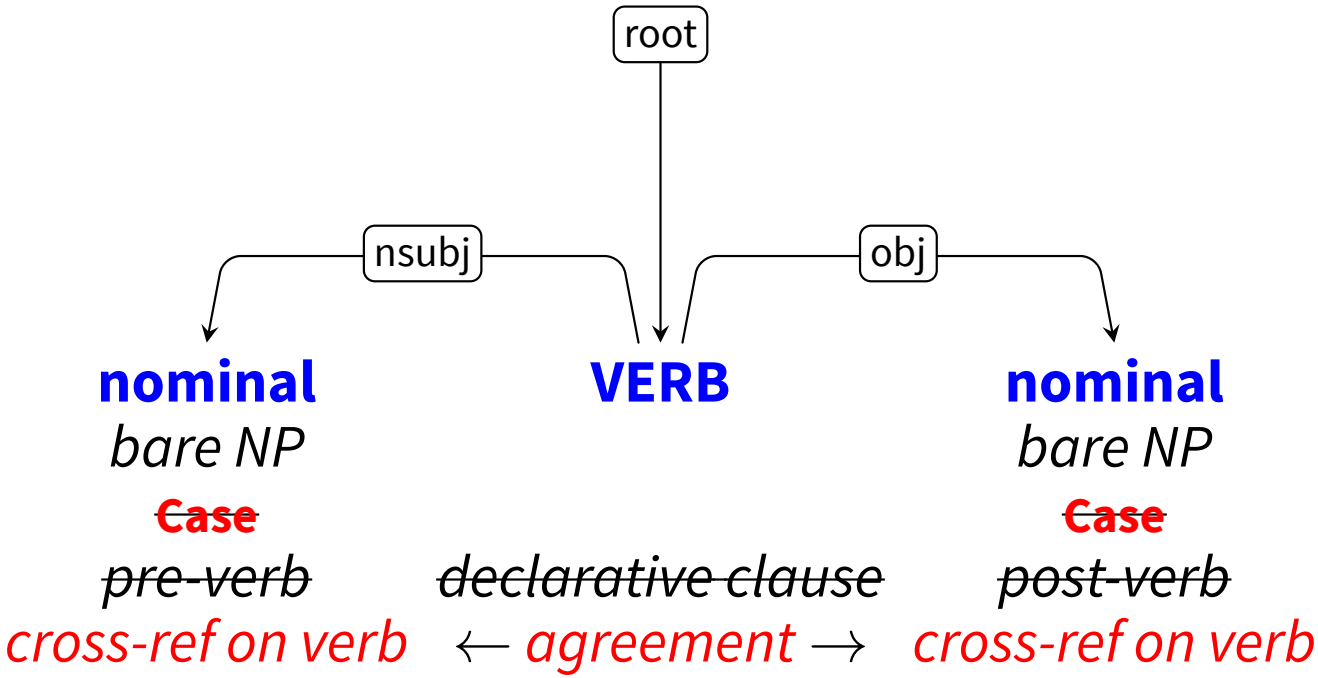
Tagalog Locative Voice ⇒ Ditransitive!



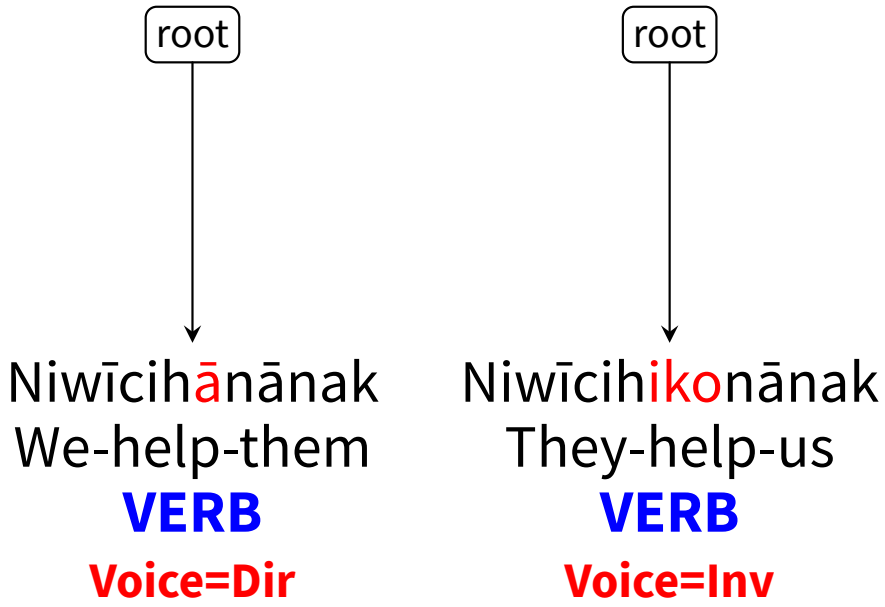
Tagalog Benefactive Voice ⇒ Ditransitive!



Plains Cree Transitive Clauses

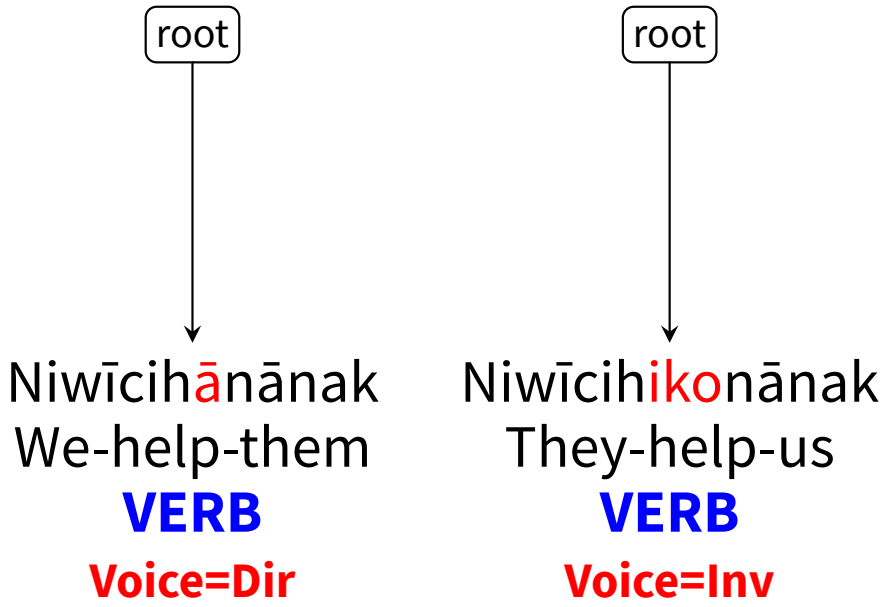


Direct-Inverse Voice in Plains Cree



Animacy hierarchy: 1st person > 3rd person

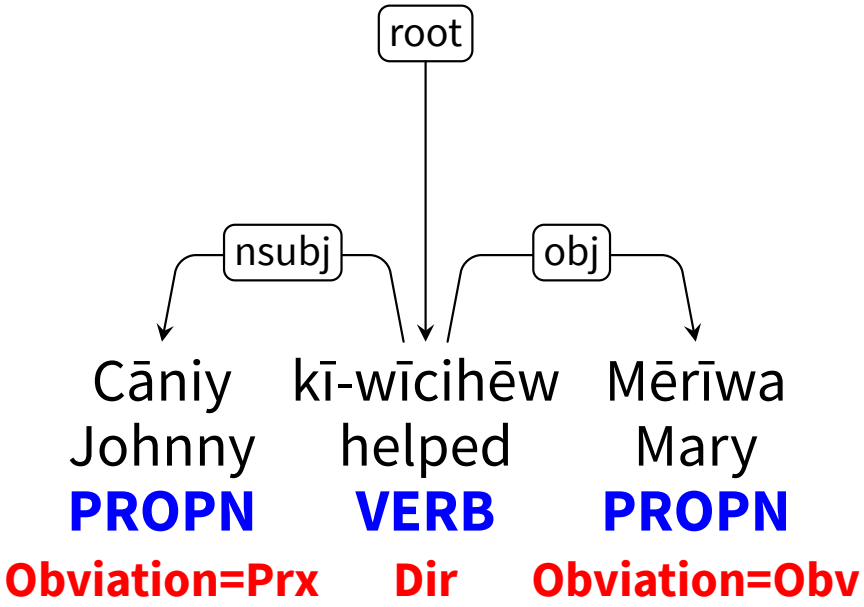
Direct-Inverse Voice in Plains Cree

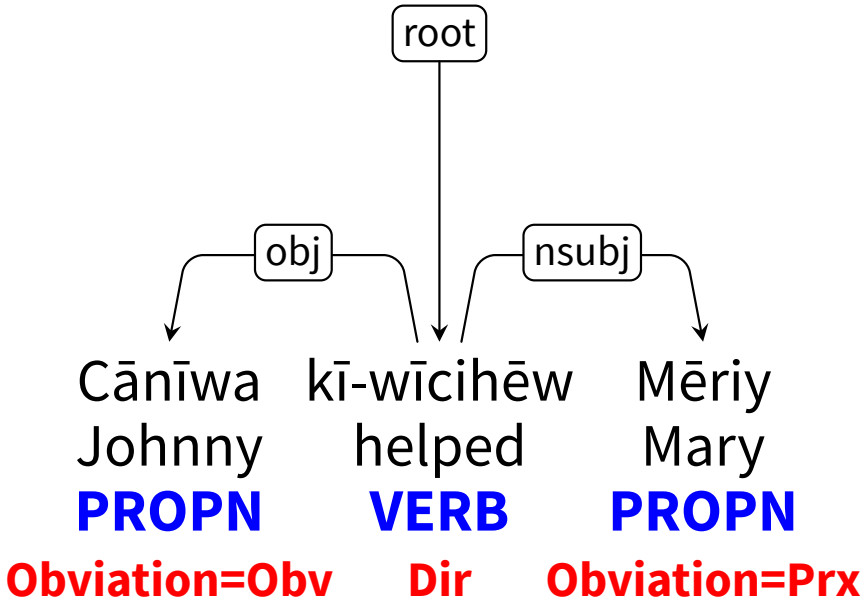


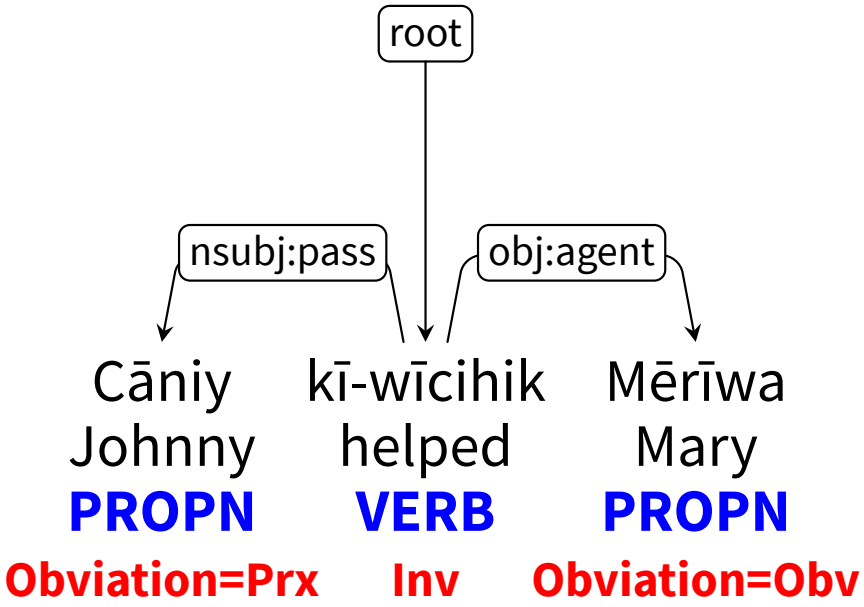
Animacy hierarchy: 1st person > 3rd person

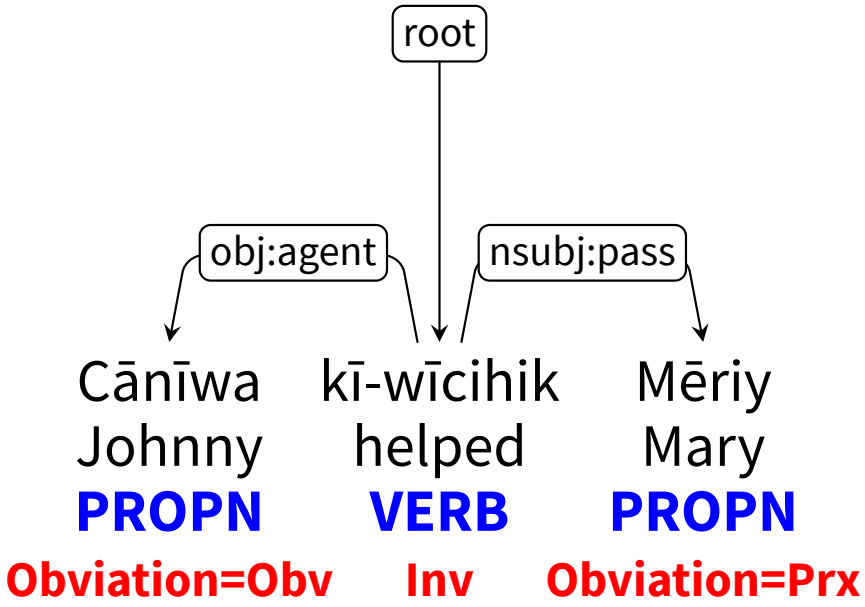
Should we set **nsubj** > **obj**?

Direct-Inverse Voice in Plains Cree

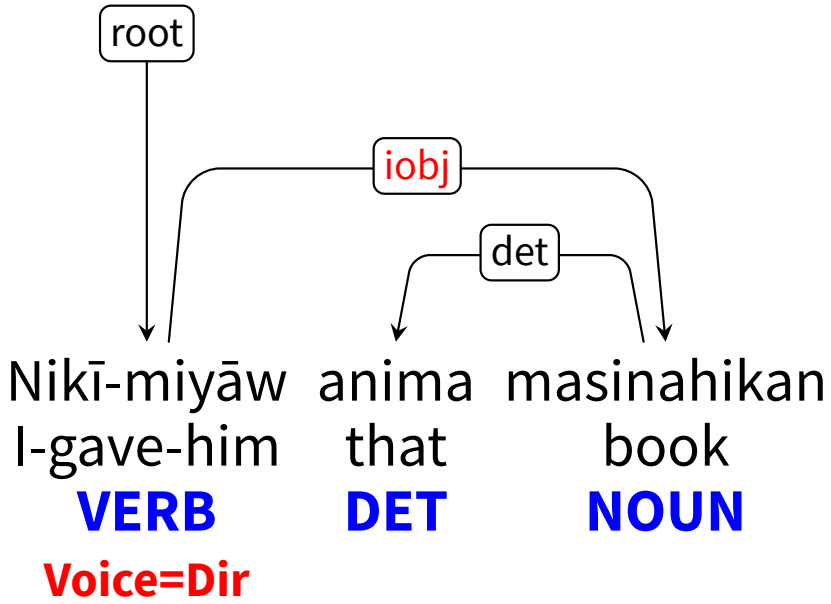








Plains Cree Ditransitive Clauses



The theme (not the recipient) is indirect object because it is not cross-referenced on the verb (it is inanimate, while the verb references an animate object).

Summary

UD guidelines: *"UD taxonomy is centered around the fairly clear distinction between core arguments versus other dependents."*

I have tried to find out how clear it is.
There is a fair number of open questions.

Summary

UD guidelines: *"UD taxonomy is centered around the fairly clear distinction between core arguments versus other dependents."*

I have tried to find out how clear it is.
There is a fair number of open questions.

ANSWERS WELCOME!
(Questions too.)