Class 07: Semantic roles and PropBank

Introduction

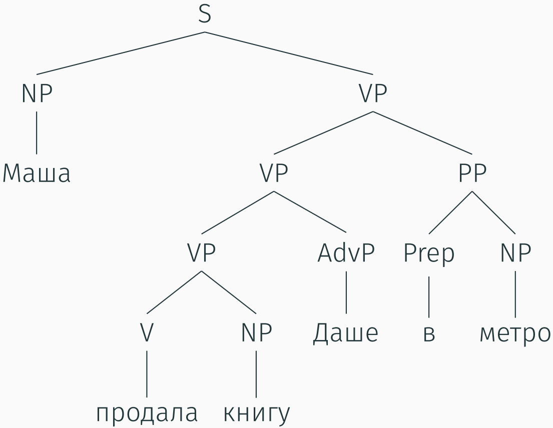
The grand quest of NLP:

Кто ?	сделал что ?	кому ?	где ?
Маша	продала книгу	Даше	в метро
Maša	*sold the book*	*to Daša*	*on the metro*

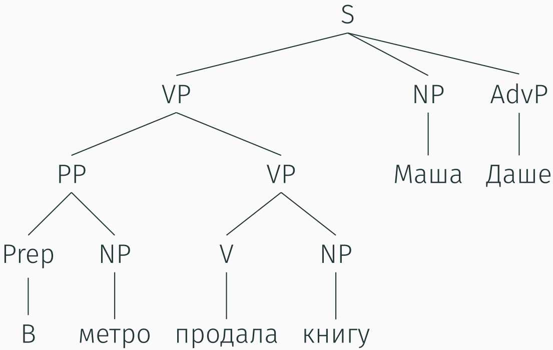
- Что было сделано ?
- Кто продал книгу?
- Кому продала Маша книгу ? / Кому Маша продала книгу ?
- Где Маша продала книгу ?

- Question answering
    - Determining if an event corresponds to a question
    - Event extraction and ontology filling
- Machine translation
    - Evaluation: Text coherence
    - Features for argument structure coherence
        - Making sure the “who did what to whom” is preserved in the output

Syntax/1

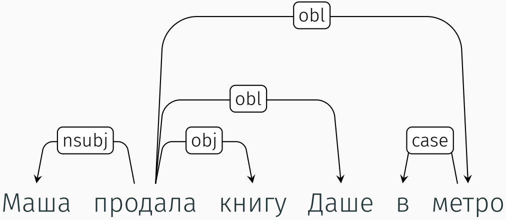

S
NP
VP
Маша
VP
PP
VP
AdvP
Prep
NP
V
NP
Даше
в
метро
продала
книгу

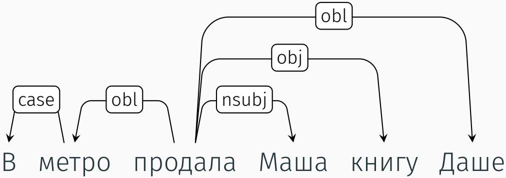
Syntax/1

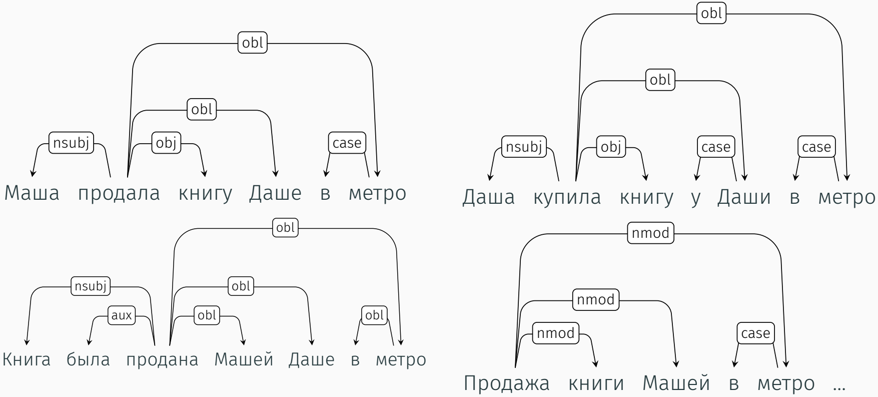


Syntax/2

Doesn't dependency parsing solve this ?







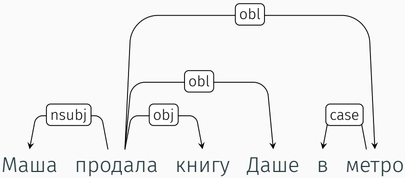
Маша продала книгу Даше в метро

Даша купила книгу у Даши в метро

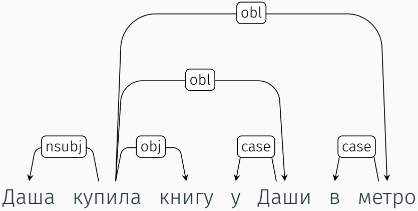
Книга была продана Машей Даше в метро

Продажа книги Машей в метро …

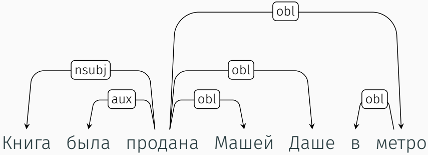
Could these refer to the same event ?



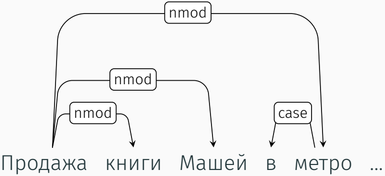
Маша продала книгу Даше в метро



Даша купила книгу у Даши в метро



Книга была продана Машей Даше в метро



Продажа книги Машей в метро …

Could these refer to the same event ?

Can you think of more ways of saying the same thing?

Semantic roles

Shallow representation

Predicates and arguments/roles.
Predicates (like продать, купить) represent an **event**.
Semantic roles (like Agent, Theme) express the abstract role of the arguments of the predicate.

Buyer Agent Proto-Agent

← →

More specific **More general**

Deep roles

Specific for a predicate,

- Maša broke the window
- Saša opened the door

Subjects of *break* and *open*: Breaker and Opener

The objects are: BrokenThing and OpenedThing

Hard to reason with for applications

Thematic roles/1

But both Breaker and Opener have something in common:

- Volitional actors
- Often animate
- Direct causal responsibility for their events

Thematic roles capture this similarity,

- Breaker and Opener are both AGENTS
    - Volitional actors with causal responsibility for an event
- BrokenThing and OpenedThing are both THEMES
    - Inanimate objects affected in some way by an action

Thematic roles/2

One of the first linguistic models:

- Introduced by the grammarian Pāṇini between the 7th and 4th centuries BCE
- Called *kāraka* in Sanskrit/Indo-Aryan linguistics

Modern formulation by Fillmore (1966):

- Influenced by Tesnière (1959)'s dependency syntax
- Called first *actants* (following Tesnière) and then later *case*

The terminology is confusing.

Thematic roles/3

Role	Definition
AGENT	The volitional causer of an event ...
EXPERIENCER	The experiencer of an event ...
FORCE	Non-volitional causer of an event ...
THEME	Participant most directly affected by an event ...
INSTRUMENT	An instrument used in an event ...
BENEFICIARY	The beneficiary of an event ...
SOURCE	Origin of a transfer event ...
GOAL	The destination of a transfer event ...

Thematic roles/3

Role	Definition
AGENT	The volitional causer of an event
	Маша разбила окно
EXPERIENCER	The experiencer of an event
	У Саши болит голова
FORCE	Non-volitional causer of an event
	Ветер сдувал снег
THEME	Participant most directly affected by an event
	Маша продала **книгу**
INSTRUMENT	An instrument used in an event
	Она написала письмо **ручкой**
BENEFICIARY	The beneficiary of an event
	Я купил **тебе** кофе
SOURCE	Origin of a transfer event
	Ты не приехала **из Кызыла?**
GOAL	The destination of a transfer event
	Я хочу **в Якутск**

Thematic «grid»

разбить:

- AGENT
- THEME
- INSTRUMENT

Realisations:

- AGENT/Subject THEME/Object
- AGENT/Subject THEME/Object INSTRUMENT/NP$_{ins}$
- THEME/Subject

Маша	*разбила*	*окно*	
AGENT		THEME	
Маша	*разбила*	*окно*	*молотком*
AGENT		THEME	INSTRUMENT
$^?$ *Молоток*	*разбил*	*окно*	
INSTRUMENT		THEME	
Окно	*разбилось*		
THEME			
Окно	*было*	*разбито*	*Машей*
THEME			AGENT
Окно	*было*	*разбито*	*молотком*
THEME			INSTRUMENT

Problems

Very hard to create a standard set of roles or formally define them.

For example for INSTRUMENT,

- **intermediary instruments** can appear as subjects:
    - The cook opened the jar with the new gadget
    - The new gadget opened the jar
- **enabling instruments** cannot:
    - They ate rice with chopsticks
    - *The chopsticks ate rice

Alternatives

Buyer Agent Proto-Agent

← →

FrameNet **More roles** **Fewer roles** PropBank

PropBank:

- Generalised roles defined as prototypes

FrameNet:

- Define roles specific to a group of predicates

Pause for thought:

- If we want to use this in a practical NLP system, does the label matter or does the distribution matter?
- If we can generalise over different things that look different but refer to the same event (buy, sell; kick, is kicked) does the precise formalism matter?

PropBank and FrameNet

PropBank/1

A **PropBank**[1] is a corpus annotated with predicates and arguments

The English PropBank:

- Annotated on top of the Penn Treebank
- Not freely available

Uses numbered arguments:

- Arg0: PROTO-AGENT
- Arg1: PROTO-PATIENT
- Arg2: BENEFACTIVE, INSTRUMENT, ATTRIBUTE END STATE
- ...

PropBanks exist for: English*, Chinese*, Arabic*, Finnish, Russian?

[1] Martha Palmer, Daniel Gildea and Paul Kingsbury (2005) “The Proposition Bank: An Annotated Corpus of Semantic Roles”. *Computational Linguistics* 31(1):71–106

PropBank/2

Proto-Agent:[2]

- Volitional involvement in event or state
- Sentience (and/or perception)
- Causes an event or change of state in another participant
- Movement (relative to position of another participant)

Proto-Patient:

- Undergoes change of state
- Causally affected by another participant
- Stationary relative to movement of another participant

[2] David Dowty (1991) “Thematic Proto-Roles and Argument Selection”. *Language*, 67(3) pp. 547–619.

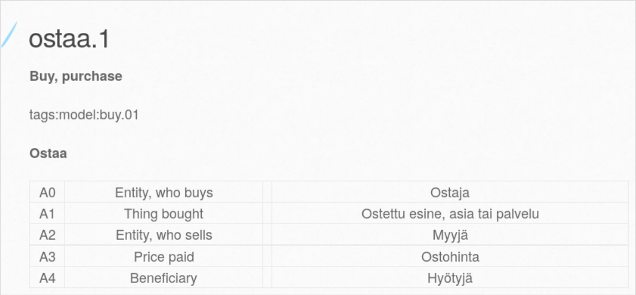
PropBank/3

There is a special prefix, ArgM-, for modifiers of the predicate:

ArgM-TMP	Когда ?	yesterday evening, now
-LOC	Где ?	in the metro, in Moscow
-DIR	Куда ?	down, to Kyzyl
-MNR	Как ?	clearly, enthusiastically
-PRP	Почему ?	because, in response to the ruling
-ADV	Miscellaneous	–
-PRD	II-predication	painted the room naked

PropBank/4

PropBank comes with **frame files** which contain predicates and their argument structure.



ostaa.1

Buy, purchase

tags:model:buy.01

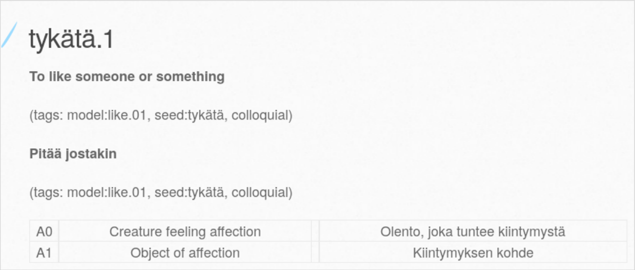
Ostaa

A0	Entity, who buys	Ostaja
A1	Thing bought	Ostettu esine, asia tai palvelu
A2	Entity, who sells	Myyjä
A3	Price paid	Ostohinta
A4	Beneficiary	Hyötyjä

- Finnish PropBank is freely available
- `https://github.com/TurkuNLP/Finnish_PropBank` (data branch)

PropBank/4

PropBank comes with **frame files** which contain predicates and their argument structure.



tykätä.1

To like someone or something

(tags: model:like.01, seed:tykätä, colloquial)

Pitää jostakin

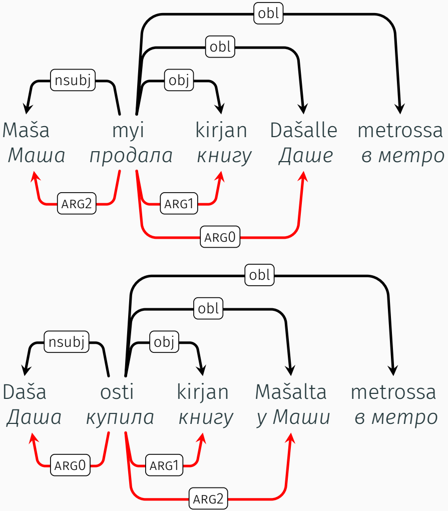
(tags: model:like.01, seed:tykätä, colloquial)

A0	Creature feeling affection	Olento, joka tuntee kiintymystä
A1	Object of affection	Kiintymyksen kohde

- Finnish PropBank is freely available
- `https://github.com/TurkuNLP/Finnish_PropBank` (data branch)

PropBank/5

PropBank-style annotation allows us to see commonalities:


obl
obl
nsubj
obj
Maša myi kirjan Dašalle metrossa
Маша продала книгу Даше в метро
ARG2
ARG1
ARG0
obl
obl
nsubj
obj
Daša osti kirjan Mašalta metrossa
Даша купила книгу у Маши в метро
ARG0
ARG1
ARG2

PropBank/6

Summary:

- A propbank is a corpus annotated with predicate–argument structure
- Predicate–argument structure generalises over syntax
- There is a free PropBank for Finnish

But how about Russian?

- There is a semantically-annotated corpus based on FrameNet
- It could be converted into a PropBank
- For more info ask Olya Lyashevskaya

FrameNet/1

FrameNet is very popular:

- Semantically-annotated database/electronic resource

It contains (for English):

- 1,200 frames
- 13,000 lexical units (word–meaning correspondence)
- 202,000 example sentences

FrameNet/2

Frames:

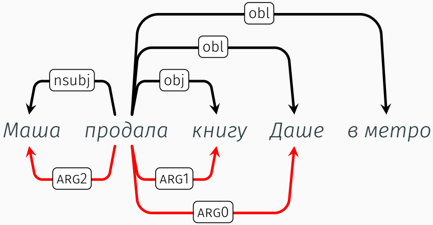
- Conceptual structure involving participants, events and background knowledge
- Extremely specific, e.g.
  - `Commerce_goods-transfer`
  - `Being_born`
  - `Criminal_process`

Frame elements:

- **Core:** essential to the meaning of the Frame
  - `Seller`, `Buyer`, `Goods`
- **Non-core:** descriptive, e.g. time, place, manner
  - `Place`, `Purpose`

vs. PropBank

PropBank:



FrameNet:

Маша	*продала*	*книгу*	*Даше*	*в метро*
`Seller`		`Goods`	`Buyer`	

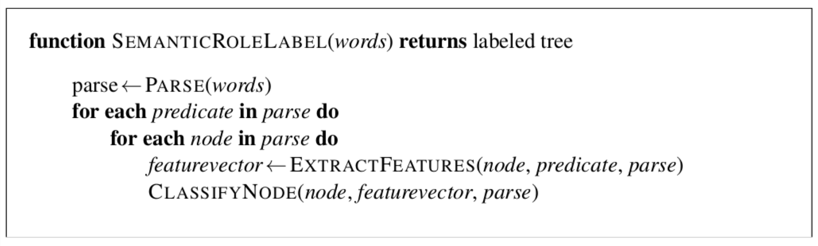
продать.1

`Commerce_goods-transfer`

Semantic role labelling

Semantic role labelling

A generic algorithm:



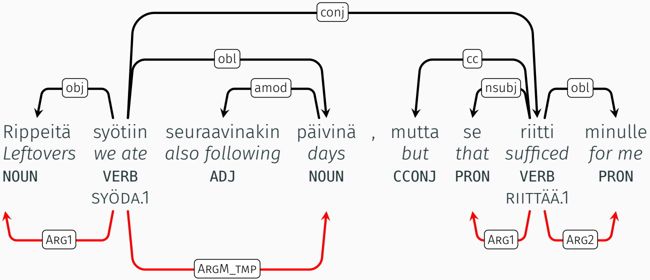
```
function SEMANTICROLELABEL(words) returns labeled tree

    parse ← PARSE(words)
    for each predicate in parse do
        for each node in parse do
            featurevector ← EXTRACTFEATURES(node, predicate, parse)
            CLASSIFYNODE(node, featurevector, parse)
```

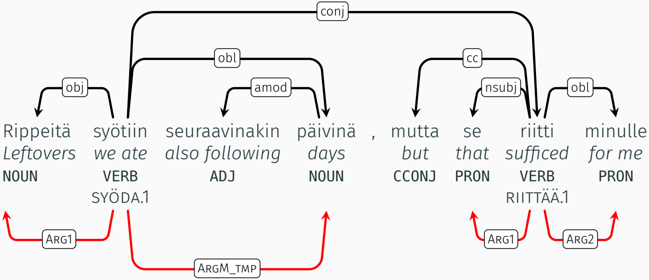
How do we decide what is a predicate ?

- PropBank: Use the verbs
- FrameNet: Use what was labelled as such in the training data

Features


conj
obl
cc
obj
amod
nsubj
obl
Rippeitä
Leftovers
NOUN
syötiin
we ate
VERB
SYÖDA.1
seuraavinakin
also following
ADJ
päivinä
days
NOUN
,
mutta
but
CCONJ
se
that
PRON
riitti
sufficed
VERB
RIITTÄÄ.1
minulle
for me
PRON
ARG1
ARGM_TMP
ARG1
ARG2

Features



Headword of constituent	Rippeitä
Headword POS	NOUN
Headword Morph. features	Case=Par
Voice of clause	Active
Linear position (wrt. predicate)	before
Path features	...
First and last words in constituent	...

One step or three step

One step:

- Classify argument type

Three step:

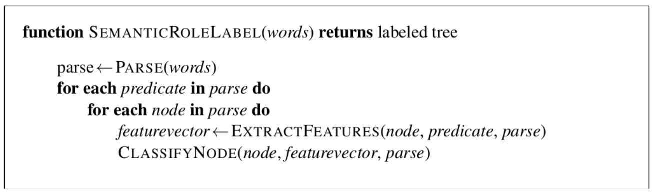
- Prune unlikely nodes
- Identify if a node is an argument or not
- Classify argument type

Why add pruning and identification steps?

Why add pruning and identification steps?

- Algorithm is looking at one predicate at a time
- Very few of the nodes in the tree could possible be arguments of that one predicate
- Imbalance between:
  - (+) positive samples (constituents/nodes that are arguments of predicate)
  - (-) negative samples (constituents/nodes that are not arguments of predicate)
- Imbalanced data can be hard for many classifiers
- So we prune the very unlikely constituents first, and then use a classifier to get rid of the rest.

Joint inference/1



```
function SemanticRoleLabel(words) returns labeled tree

    parse ← Parse(words)
    for each predicate in parse do
        for each node in parse do
            featurevector ← ExtractFeatures(node, predicate, parse)
            ClassifyNode(node, featurevector, parse)
```

- The algorithm so far classifies everything locally – each decision about a constituent is made independently of all others
- But: Lots of global or joint interactions between arguments and constraints
  - e.g. PropBank does not allow multiple identical arguments, so
  - Labelling one constituent as Arg0 should increase the probability of another being Arg1

Joint inference/2

Reranking:

- The first stage SRL system produces multiple possible labels for each constituent
    - The second stage classifier the best global label for all constituents
    - Often a classifier that takes all the inputs along with other features (sequences of labels)

Summary

Semantic Role Labelling:

- A level of shallow semantics for representing **events** and their **participants**
- Intermediate between parses and full semantics
- Two common architectures, for various languages
  - FrameNet: frame-specific roles
  - PropBank: Proto-roles
- Current systems extract by
  - parsing sentence
  - Finding predicates in the sentence
  - For each one, classify each parse tree constituent

Practical

Practical

Option 1:

- Download Finnish PropBank
  - `https://github.com/TurkuNLP/Finnish_PropBank`
  - `https://github.com/TurkuNLP/Finnish_PropBank/tree/data`
  - `https://://github.com/TurkuNLP/Finnish_PropBank/tree/data/gen_lemmas`
- Write a semantic role labeller
- Train on `train`, find good feature combination on `dev` and test on `test`.

Option 2:

- Olya Lyaševskaya has given me a file with semantically annotated sentences for Russian
- Combination of TSV + XML
- Produce something approximating the PropBank style annotation.

Data format

Uusi elämä myös tuoksuu uudelta! :)

‘New life also smells fresh! :)’

.conllu file:

ID	TOKEN	LEM	POS		FEATS	HEAD	DEPREL	DEPRELS	MISC
1	Uusi	_	_	_	_	2	amod	_	_
2	elämä	_	_	_	_	4	nsubj	4:PBArg_1	_
3	myös	_	_	_	_	4	advmod	4:PBArgM_dis	_
4	tuoksuu	_	_	_	_	0	root	_	PBSENSE=tuoksua.1
5	uudelta	_	_	_	_	4	xcomp	4:PBArg_2	_
6	!	_	_	_	_	4	punct	_	_
7	:)	_	_	_	_	4	discourse	_	_

.tsv file:

```
base|number|argnum|definition|note|definition_fin|note_fin
tuoksua|1|1|Stinky thing|NULL|Tuoksuva asia|NULL
tuoksua|1|2|Attribute of arg1|NULL|Mille tuoksuu|NULL
```

Combination of TSV + XML

```
# new FrameAnno<br />
# FrameAnchor = беречься<br />
# ConstrID = 11655<br />
# ConstrName = 1.5_Берегись, чтобы не упасть.<br />
# ConstrPattern = Snom V чтобы + CL<br />
# ExampleID = 43401<br />
# SentType = sp<br />
# SentXml = <p class="verse"><se><w><ana lex="плакать" gr="V,2p,act,imper,ipf,norm,sg" sem="ca:noncaus t:physiol d:root" sem2="ca:noncaus
d:root"/>Плачь</w>, <w><ana lex="но" gr="CONJ,norm"/>но</w> <w><ana lex="беречься" gr="V,2p,act,imper,ipf,norm,sg" sem="ca:noncaus"
sem2="ca:noncaus"/>берегись</w>, <w><ana lex="чтобы" gr="CONJ,norm"/>чтобы</w> <w><ana lex="хоть" gr="PART,norm"/>хоть</w> <w><ana
lex="один" gr="APRO,f,nom,norm,sg" sem2="r:indet"/>одна</w> <w><ana lex="твой" gr="APRO,f,nom,norm,sg" sem="r:poss"/>твоя</w> ...
# SentText = <u>Плачь, но <b>берегись</b>, чтобы хоть одна твоя слеза скатилась по острию пера и примешалась к чернилам. </u>Ренар
<br />
# FEtable:FE_ID "Word" Role FE_Status SyntRank Morph LexClass "Group"<br />
21184 "" агенс Core Не выражен ""<br />
21185 "берегись" - Core Предикат беречься - "-"<br />
21186 "что / скатилась" потенциальная угроза Core Клауза чтобы + CL - "чтобы хоть одна твоя слеза скатилась по острию пера и примешалась
к чернилам"<br /><br /><br />
```

Ženya: “How was it born like that?”