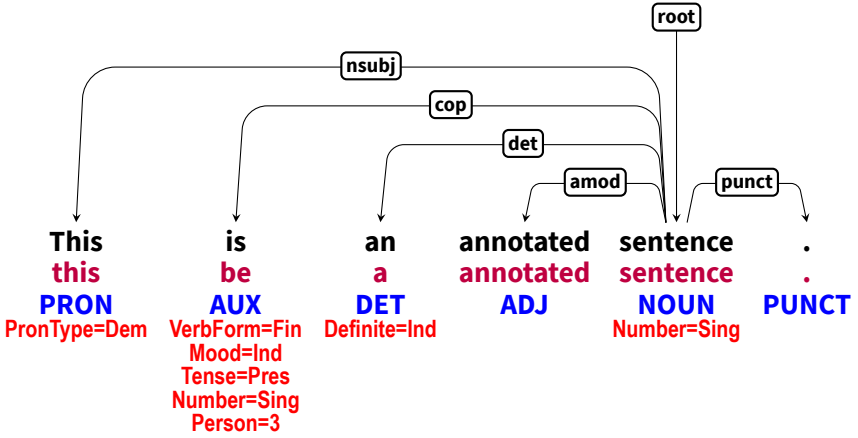
Enhanced and Deep Universal Dependencies for Many Languages

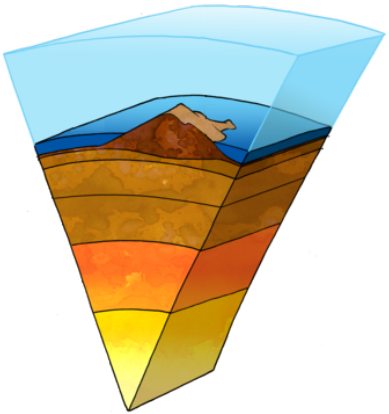
Kira Droganova, **Daniel Zeman**

{droganova,zeman}@ufal.mff.cuni.cz
http://universaldependencies.org/

Universal Dependencies



Multiple Layers of Dependencies

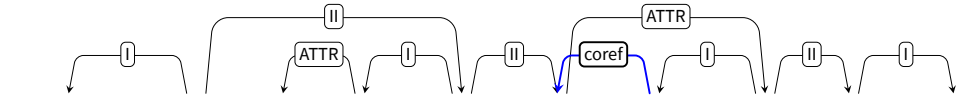


Form

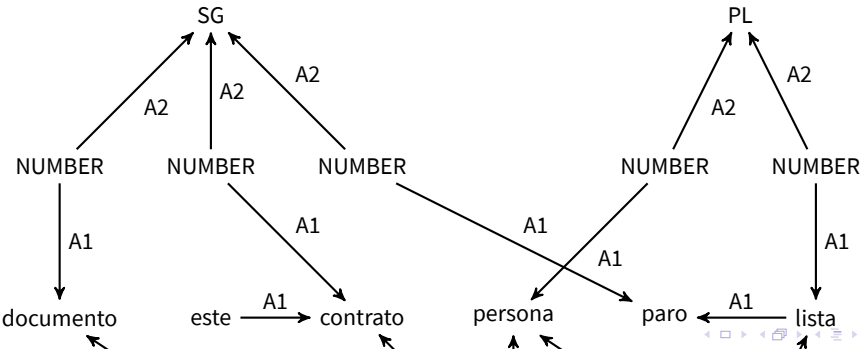
- Surface syntax
- Deep syntax
- Semantics

Meaning

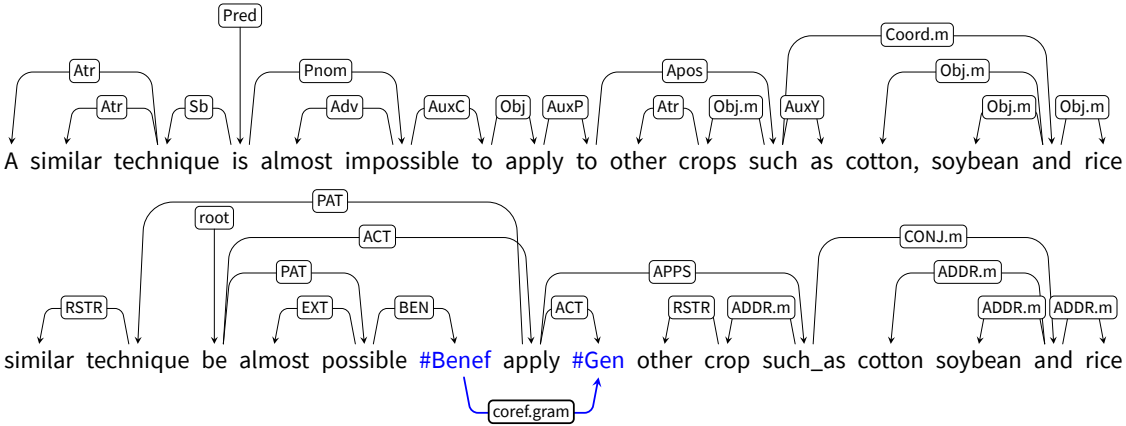
Meaning-Text Theory



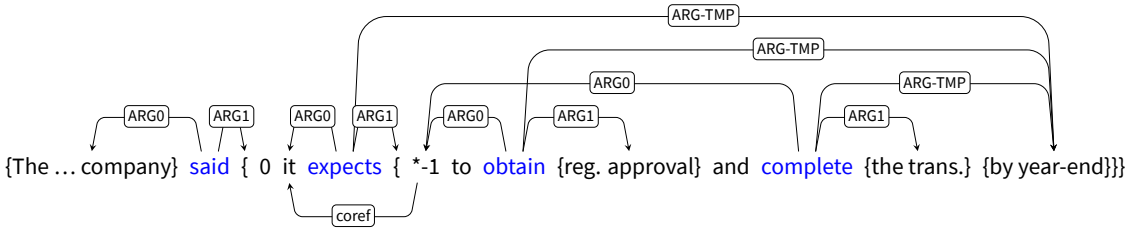
“The document suggests that this contract affect the persons who make the unemployment lists swell.”



Functional Generative Description (Prague Tectogrammatics)

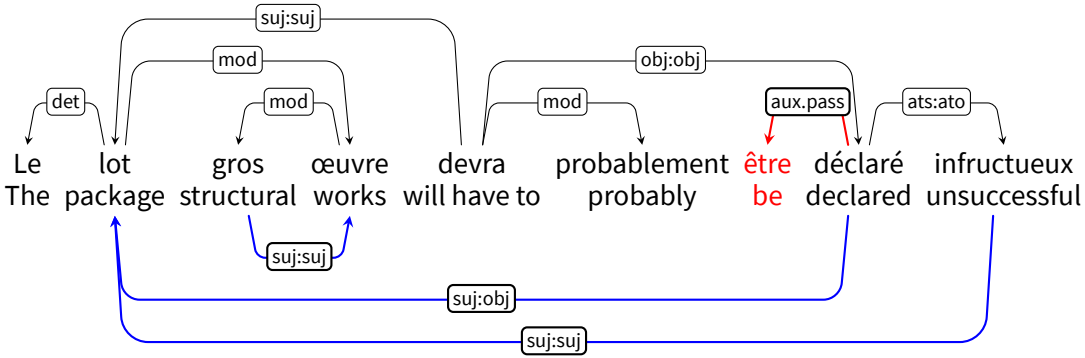


Proposition Banks



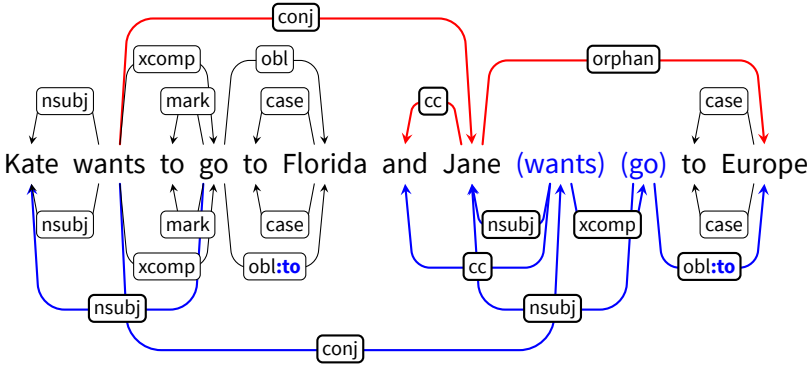
"The thrift holding company said it expects to obtain regulatory approval and complete the transaction by year-end."

Sequoia



“The structural system package should probably be declared unsuccessful.”

Enhanced Universal Dependencies



Multilingual Annotation

- **MTT:** Russian, English, Spanish, French
- **FGD:** Czech, English
- **PropBank:** English, Arabic, Chinese, Finnish, Hindi, Urdu, Persian, Portuguese, Turkish, German, French
- **AMR:** English, Chinese, Portuguese, Korean, Vietnamese, Spanish, French, German
- **Sequoia:** French

Multilingual Annotation

- **MTT:** Russian, English, Spanish, French
- **FGD:** Czech, English
- **PropBank:** English, Arabic, Chinese, Finnish, Hindi, Urdu, Persian, Portuguese, Turkish, German, French
- **AMR:** English, Chinese, Portuguese, Korean, Vietnamese, Spanish, French, German
- **Sequoia:** French

- **Enhanced UD:** Arabic, Bulgarian, Czech, Dutch, English, Estonian, Finnish, French, Italian, Latvian, Lithuanian, Polish, Russian, Slovak, Swedish, Tamil, Ukrainian

Basic Universal Dependencies: 90 (89) Languages and Growing

- I.-E.: Armenian, Ancient Greek, Greek, Breton, Irish, Scottish, Welsh
  - Germanic: Afrikaans, Danish, Dutch, English, Faroese, German, Gothic, Norwegian, Swedish, Swiss German
  - Romance: Catalan, French, Galician, Italian, Latin, Old French, Portuguese, Romanian, Spanish
  - Balto-Slavic: Belarusian, Bulgarian, Croatian, Czech, Church Slavonic, Old Russian, Polish, Russian, Serbian, Slovak, Slovenian, Ukrainian, Upper Sorbian, Latvian, Lithuanian
  - Indo-Iranian: Kurmanji, Persian, Hindi, Bhojpuri, Marathi, Sanskrit, Urdu
- Uralic: Erzya, Estonian, Finnish, Hungarian, Karelian, Livvi, Komi Permyak+Zyrian, Moksha, Sámi North+Skolt
- Dravidian: Tamil, Telugu; Turkic: Kazakh, Turkish, Uyghur
- Af.-As.: Akkadian, Amharic, Arabic, Assyrian, Coptic, Hebrew, Maltese
- Sino-Tibetan: Cantonese, Classical Chinese, Chinese; Aus.-As.: Vietnamese
- Tai-Kadai: Thai; Austronesian: Indonesian, Tagalog
- Other: Buryat, Japanese, Korean, Basque, Sw. Sign, Naija, Bambara, Wolof, Yoruba, Warlpiri, Mbyá Guaraní

Two-Speed Approach

- Automatic part: derived from basic UD
- Optional manual extras

Two-Speed Approach

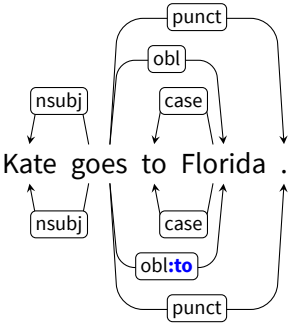
- Automatic part: derived from basic UD
  - Significant part of Enhanced UD can be derived automatically!
  - We can go beyond current guidelines (“Enhanced++”)
    - ★ Normalize syntactic alternations (cf. Candito et al. 2017)
- Optional manual extras

Two-Speed Approach

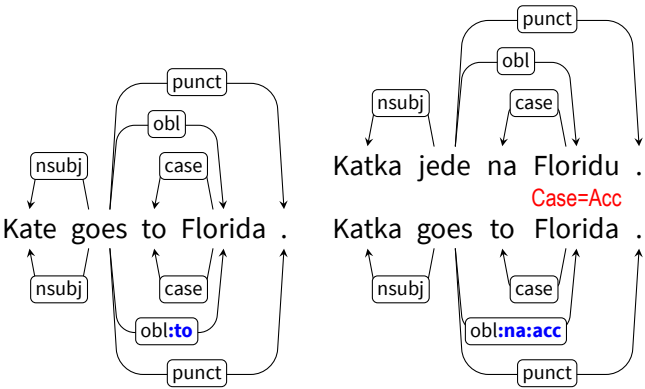
- Automatic part: derived from basic UD

- Optional manual extras
  - Existing language-specific resources converted to common format
    - ★ Frames, semantic roles
    - ★ Textual coreference
    - ★ Everything else…
    - ★ … and improve the automatic part above

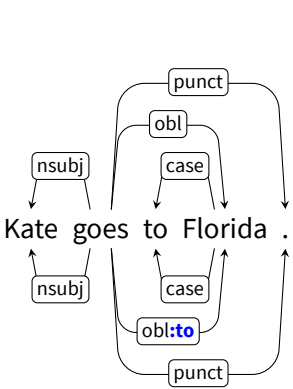
Enhanced UD: Case Information in Dependency Label

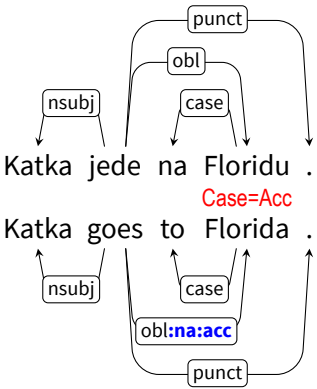


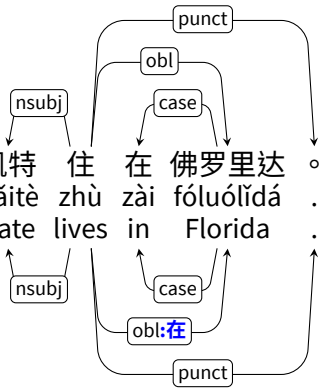
Enhanced UD: Case Information in Dependency Label



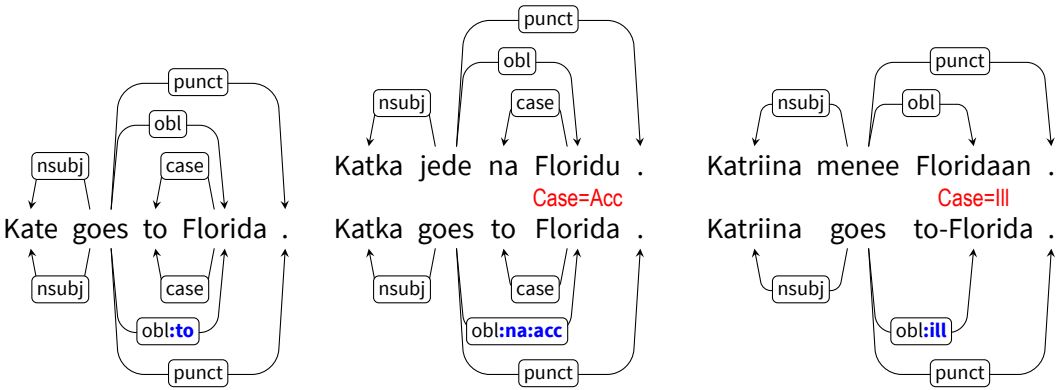
Enhanced UD: Case Information in Dependency Label



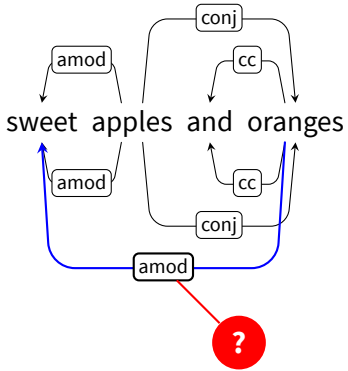




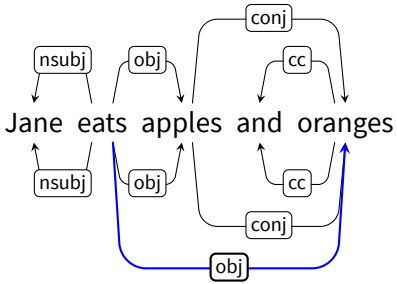
Enhanced UD: Case Information in Dependency Label



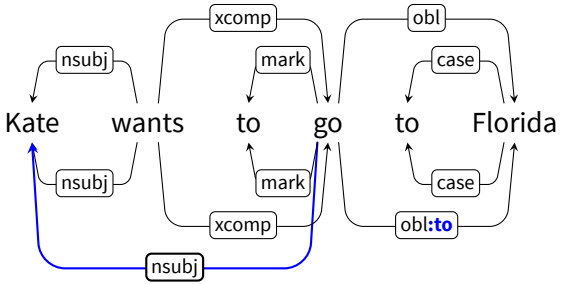
Enhanced UD: Shared Dependent of Coordination



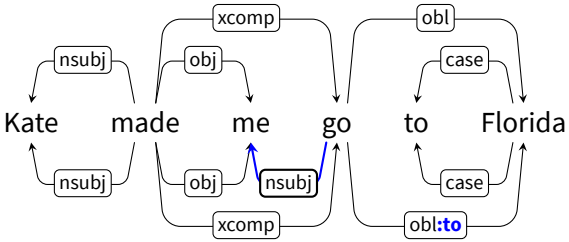
Enhanced UD: Parent of Coordination



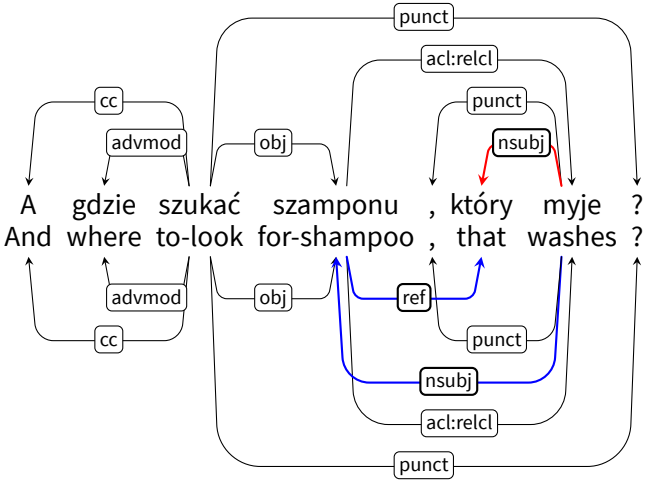
Enhanced UD: External Subject of Controlled Predicate



Enhanced UD: External Subject in Object-Control Construction

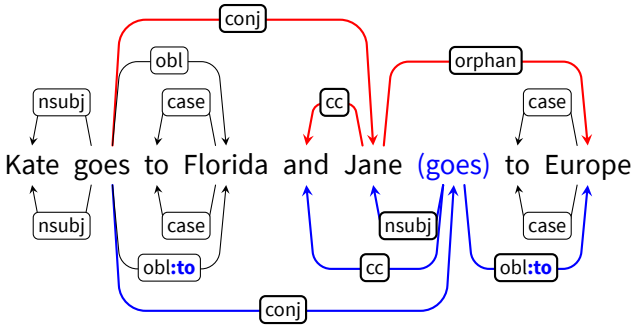


Enhanced UD: Relative Clauses

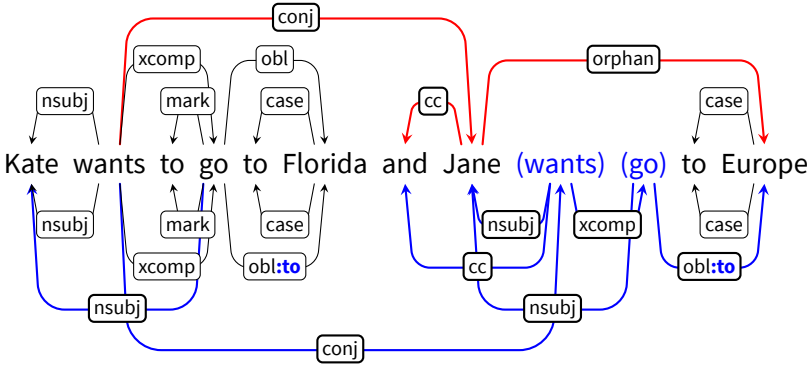


“And where to look for shampoo that works?”

Enhanced UD: Gapping and Stripping



Enhanced UD: Gapping + Control Verb Construction



Enhanced UD: Five (Six) Enhancements

- Null nodes for gapping (12 treebanks in UD 2.5)
- Dependency propagation in coordination (22 treebanks)
  - Common parent of coordination
  - Shared dependents of coordination
- External subjects of controlled predicates (12 treebanks)
- Cyclic dependencies to/from relative clauses (9 treebanks)
- Case-enhanced dependency labels (10 treebanks)

- All 5 types: 6 treebanks, 3 languages
- At least 1 type: 24 treebanks, 16 languages
- Only basic UD: 133 treebanks

Stanford Enhancer

- Part of Stanford CoreNLP (Java)
- Rules from basic to enhanced UD

Stanford Enhancer

- Part of Stanford CoreNLP (Java)
- Rules from basic to enhanced UD

- Gapping: embeddings for similarity of arguments

Stanford Enhancer

- Part of Stanford CoreNLP (Java)
- Rules from basic to enhanced UD

- Gapping: embeddings for similarity of arguments
- Coordination:
  - Parent propagation: deterministic
  - Shared dependents: heuristics (human desirable!)

Stanford Enhancer

- Part of Stanford CoreNLP (Java)
- Rules from basic to enhanced UD

- Gapping: embeddings for similarity of arguments
- Coordination:
  - Parent propagation: deterministic
  - Shared dependents: heuristics (human desirable!)
- External subjects: heuristics (subject vs. object control)

Stanford Enhancer

- Part of Stanford CoreNLP (Java)
- Rules from basic to enhanced UD

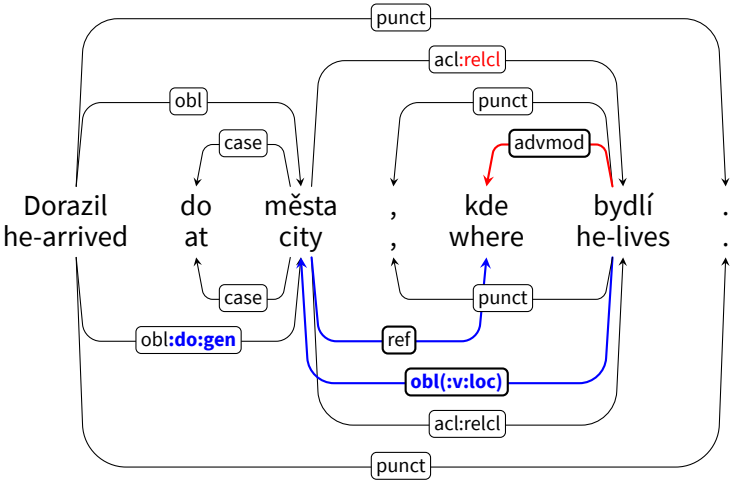
- Gapping: embeddings for similarity of arguments
- Coordination:
  - Parent propagation: deterministic
  - Shared dependents: heuristics (human desirable!)
- External subjects: heuristics (subject vs. object control)
- Relative clauses: need **`acl:relcl`** and list of relative pronouns

Stanford Enhancer

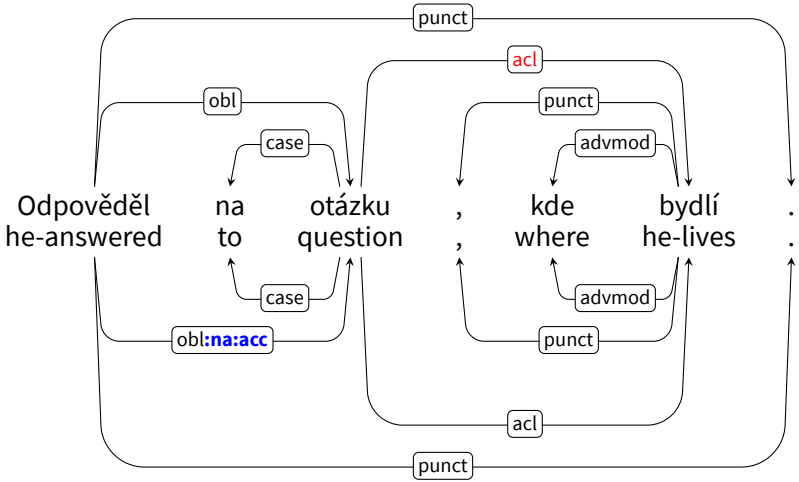
- Part of Stanford CoreNLP (Java)
- Rules from basic to enhanced UD

- Gapping: embeddings for similarity of arguments
- Coordination:
  - Parent propagation: deterministic
  - Shared dependents: heuristics (human desirable!)
- External subjects: heuristics (subject vs. object control)
- Relative clauses: need **`acl:relcl`** and list of relative pronouns
- Case-enhanced labels: deterministic

How to Recognize Relative Clauses

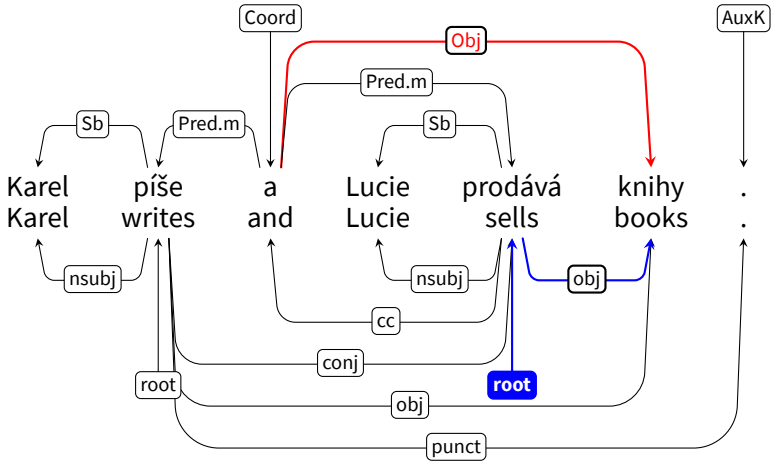


How to Recognize Relative Clauses



Conversion from non-UD Data: Extra Information?

- Analytical layer of Prague-style treebanks: shared dependents of coordination are known!



Stanford Enhancer Applied to UD 2.5

- Null nodes for gapping (12 treebanks → 62 treebanks)
- Dependency propagation in coordination (22 treebanks → 119 treebanks)
  - Common parent of coordination
  - Shared dependents of coordination
- External subjects of controlled predicates (12 treebanks → 117 treebanks)
- Cyclic dependencies to/from relative clauses (9 treebanks → 51 treebanks)
- Case-enhanced dependency labels (10 treebanks → 122 treebanks)

- All 5 types: 35 treebanks, 17 languages
- At least 1 type: 125 treebanks, 79 languages

IWPT 2020 Shared Task in Parsing EUD

- https://universaldependencies.org/iwpt20/
- Test phase (postponed): April 1–22
- End-to-end from raw text to EUD
- Data in 17 languages (UD 2.5 + French)
- Possible baseline: UDPipe + Stanford Enhancer

Our “Enhanced Plus”

- Enhanced UD help us identify more predicate-argument relations
- But some patterns are still not handled…

- Adverbial **infinitives**
  - *They will meet to discuss a contract.*
  - But not `ccomp` infinitives: *He recommended to replace the tyres.*

Our “Enhanced Plus”

- Enhanced UD help us identify more predicate-argument relations
- But some patterns are still not handled…

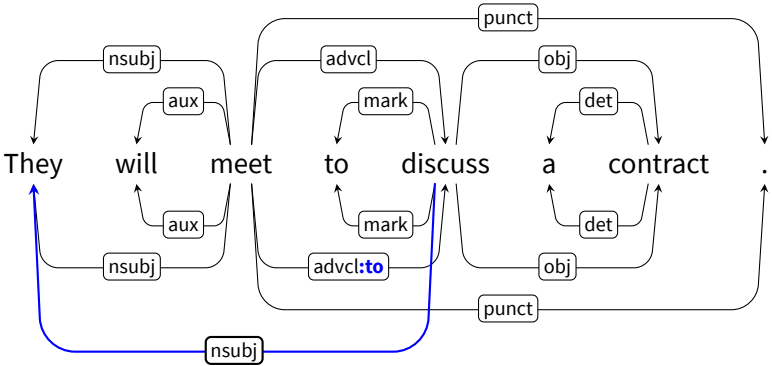
- Adverbial **infinitives**
  - *They will meet to discuss a contract.*
  - But not `ccomp` infinitives: *He recommended to replace the tyres.*
- Adverbial **converbs (gerunds)**
  - *Terrorists detonated a bomb killing five people.*

Our “Enhanced Plus”

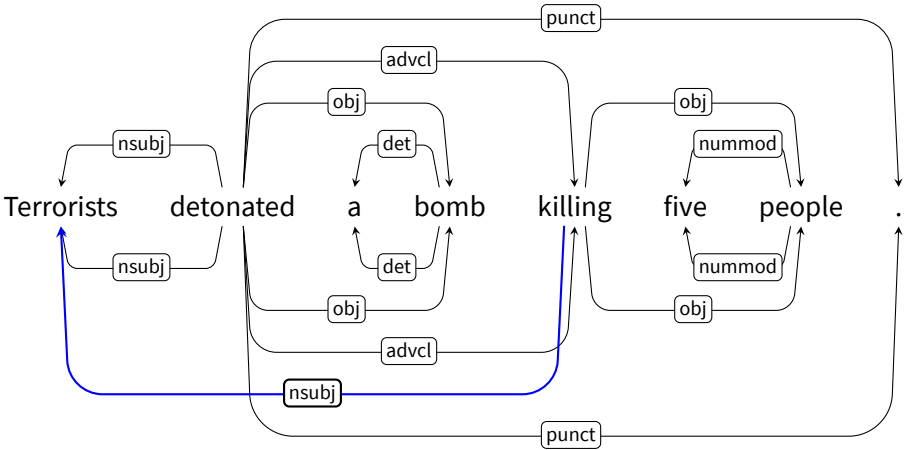
- Enhanced UD help us identify more predicate-argument relations
- But some patterns are still not handled…

- Adverbial **infinitives**
  - *They will meet to discuss a contract.*
  - But not `ccomp` infinitives: *He recommended to replace the tyres.*
- Adverbial **converbs (gerunds)**
  - *Terrorists detonated a bomb killing five people.*
- Attributive **participles**
  - *The shares reflected on your statement.*

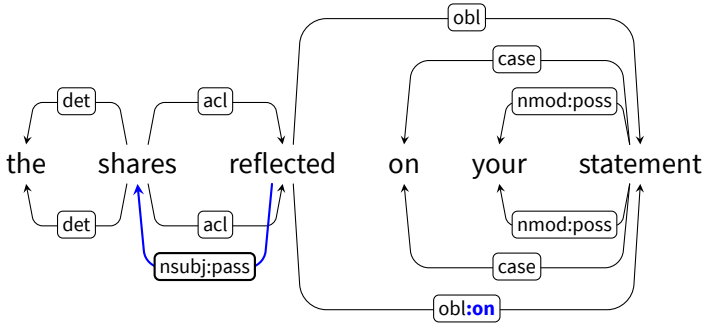
External Subject of Adverbial Infinitive



External Subject of Adverbial Gerund

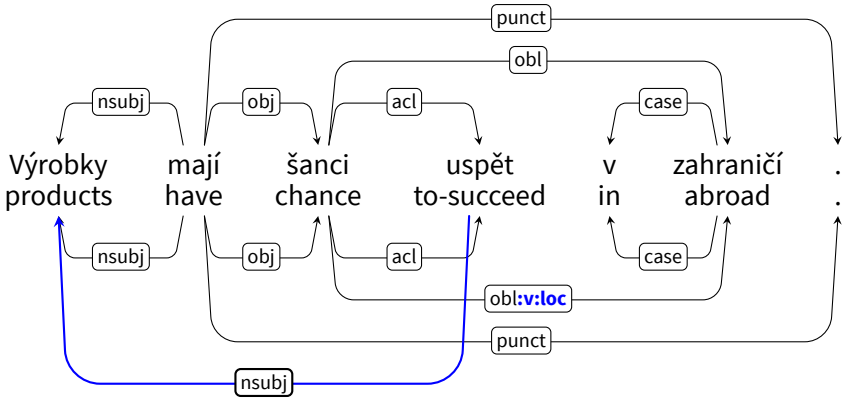


External Subject of Participle



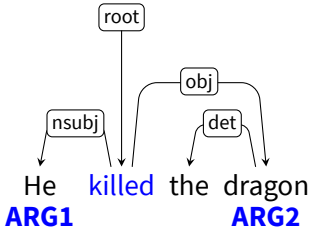
Possible Addition: Control by Light-Verb Constructions

We do not do this yet:

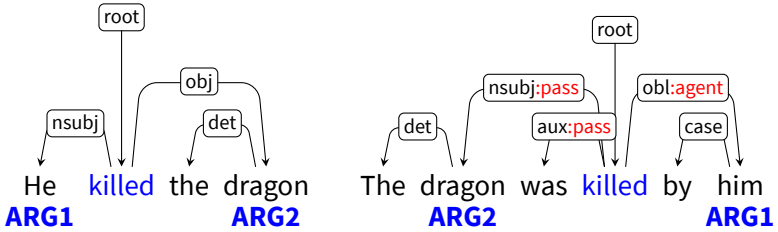


“The products have a chance to succeed abroad.”

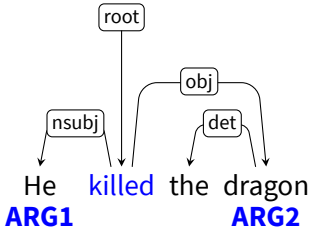
Deep UD: Normalization of Syntactic Alternations

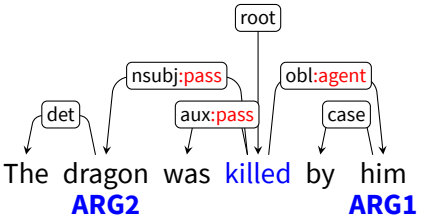


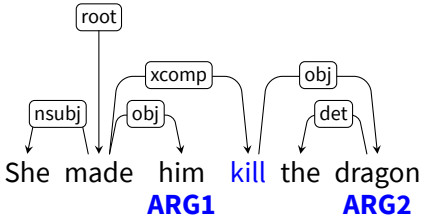
Deep UD: Normalization of Syntactic Alternations



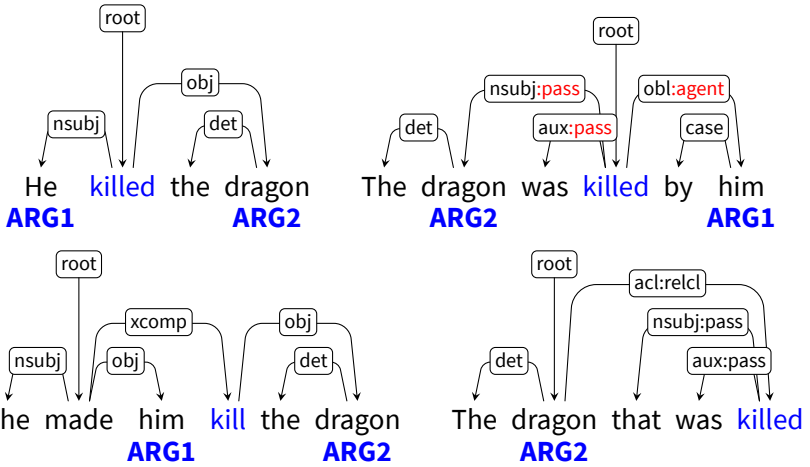
Deep UD: Normalization of Syntactic Alternations







Deep UD: Normalization of Syntactic Alternations



Numbered Arguments

- Degree of salience of arguments derived from surface syntax:
  - Subject of active clause ⇒ **ARG1**
  - Direct object of active clause ⇒ **ARG2**
  - Indirect object of active clause ⇒ **ARG3**
  - Subject of passive clause ⇒ **ARG2**
  - etc…
- Numbers can be mapped to semantic roles if we have a valency dictionary
- Like in PropBank… SOMEWHAT!

Numbered Arguments

- Degree of salience of arguments derived from surface syntax:
  - Subject of active clause ⇒ **ARG1**
  - Direct object of active clause ⇒ **ARG2**
  - Indirect object of active clause ⇒ **ARG3**
  - Subject of passive clause ⇒ **ARG2**
  - etc…
- Numbers can be mapped to semantic roles if we have a valency dictionary
- Like in PropBank… SOMEWHAT!

	Deep UD	PropBank
John broke the window.	*John*.ARG1 *window*.ARG2	*John*.ARG0 *window*.ARG1
The window was broken by John.	*John*.ARG1 *window*.ARG2	*John*.ARG0 *window*.ARG1
The window broke.	*window*.ARG1	*window*.ARG1

Predicate Identfiers

- They could be sense/frame identifiers
- But now we just take lemmas

- Exception:
  - Germanic phrasal verbs: *come_up*
  - Inherently reflexive verbs: `[cs]` *smát_se* “laugh”
  - Other compound verbs (incl. light & serial verb constructions)

Summary and Next Steps

- Deep UD 2.4 (`http://hdl.handle.net/11234/1-3022`)
- 121 treebanks of 73 languages
- Enhanced graphs in all treebanks
- (Enhanced Plus: infinitives, gerunds, participles)
- Normalized active-passive

- Can be regenerated after each UD release
- (Deep UD 2.5 now ready to be released)

Summary and Next Steps

- Deep UD 2.4 (`http://hdl.handle.net/11234/1-3022`)
- 121 treebanks of 73 languages
- Enhanced graphs in all treebanks
- (Enhanced Plus: infinitives, gerunds, participles)
- Normalized active-passive

- Can be regenerated after each UD release
- (Deep UD 2.5 now ready to be released)

- Evaluate precision and recall (no gold standard yet)
- Test mapping to a valency dictionary
- Oblique arguments?
- Other alternations than passives?
- Non-verbal predicates
- …

Future: Link Arguments to Frame Dictionaries

kill

kill ACT() PAT() ?MEANS()

- *"Kill it," he says.*
- *"Ford probably would try *trace* to kill the proposal by enlisting support from U.S. takeover-stock speculators and holding out the carrot of a larger bid later, said Stephen Reitman, European auto analyst at London brokers UBS Phillips & Drew.*
- *John killed Mary with a lead pipe, in the conservatory.*

Corpus example(s): Close [X]

pedt **She** ACT was untrained and, in one botched job **killed a** PAT **client** PAT.

pedt Dallas District Judge Jack Hampton had sparked calls for a judicial inquiry with his remarks to the press last December, two weeks after *-1 sentencing an 18-year-old defendant to 30 years in state prison **for** *-2 **killing** two homosexual **men** PAT in a city park.

pedt The judge was quoted *-1 as *-4 referring to the **victims**←PAT as ``queers'' and *-4 saying **they** PAT **would**n't **have been killed** *-2 ``if they hadn't been cruising the streets *-3 picking up teenage boys.''

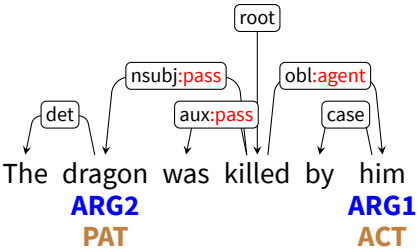
pedt Strong dividend growth, he says *T*-1, is ``the black widow of valuation'' -- a reference to the female **spiders**←ACT **that** ACT *T*-231 attract **males**←PAT and then **kill them** PAT after mating.

pedt Instead, Mr. Nixon reminded his host, Chinese President Yang Shangkun, that Americans haven't forgiven China's leaders for the military **assault**←ACT of June 3-4 **that** ACT *T*-241 **killed hundreds** PAT, PAT and perhaps **thousands** PAT, PAT of demonstrators.

EngVallex

(`http://lindat.mff.cuni.cz/services/EngVallex/EngVallex.html`)

- Select the correct frame of the verb
- Map observed arguments to frame slots
  - Use their syntactic function
  - Use their morphological form



Thanks!
Děkuji!