Universal Dependencies: Principles and Tools

Daniel Zeman

July 1, 2022

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Introduction

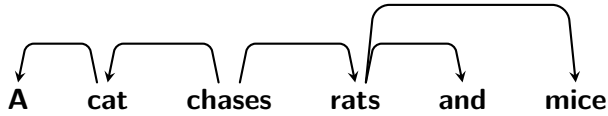
Outline

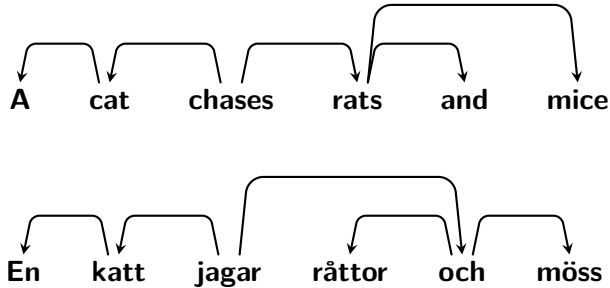
1. Introduction
2. Morphological Annotation in UD
3. Syntactic Annotation in UD
4. Core vs. Oblique
5. Enhanced Universal Dependencies
6. UD Tools

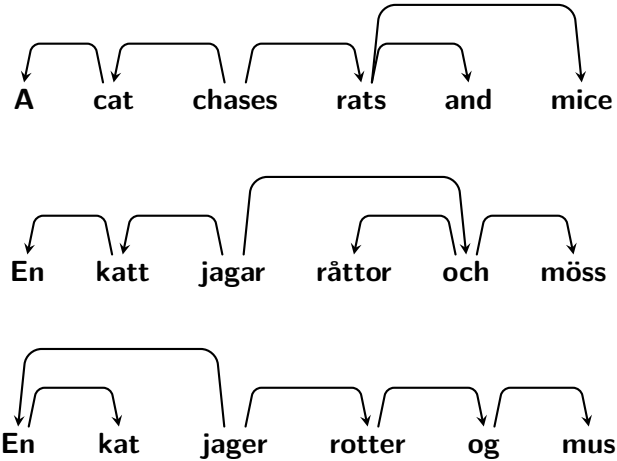
Introduction

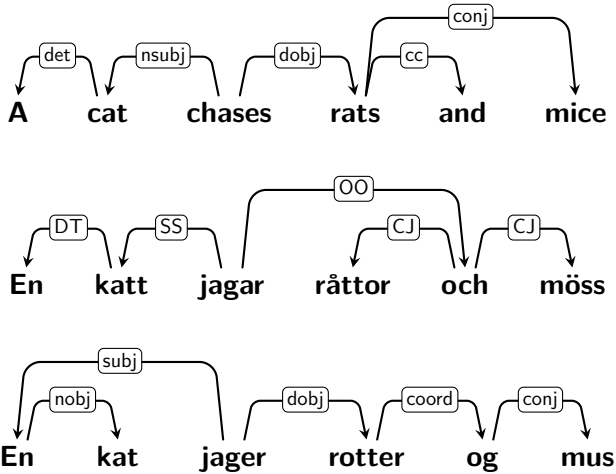
- Around 2010:

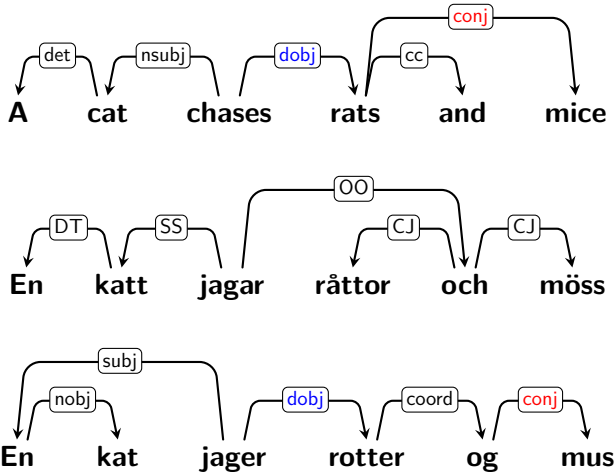
- Increasing interest in multilingual NLP
  - Multilingual evaluation campaigns to test generality
  - Cross-lingual learning to support low-resource languages
- Increasing awareness of methodological problems
  - Current NLP relies heavily on annotation
  - Annotation schemes vary across languages


A
cat
chases
rats
and
mice


A cat chases rats and mice
En katt jagar råttor och möss


A cat chases rats and mice
En katt jagar råttor och möss
En kat jager rotter og mus


det
nsubj
dobj
cc
conj
A
cat
chases
rats
and
mice
DT
SS
OO
CJ
CJ
En
katt
jagar
råttor
och
möss
subj
nobj
dobj
coord
conj
En
kat
jager
rotter
og
mus


det
nsubj
dobj
cc
conj
A
cat
chases
rats
and
mice
DT
SS
OO
CJ
CJ
En
katt
jagar
råttor
och
möss
subj
nobj
dobj
coord
conj
En
kat
jager
rotter
og
mus

Why was this a problem?

- Hard to compare empirical results across languages
- Hard to usefully do cross-lingual structure transfer
- Hard to evaluate cross-lingual learning
- Hard to build and maintain multilingual systems
- Hard to make comparative linguistic studies
- Hard to validate linguistic typology
- Hard to make progress towards a universal parser

Universal Dependencies

- https://universaldependencies.org/

- 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.

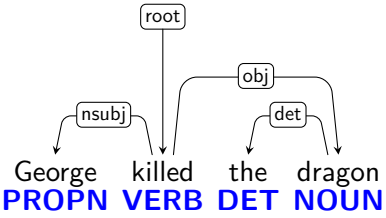
Manning's Law

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

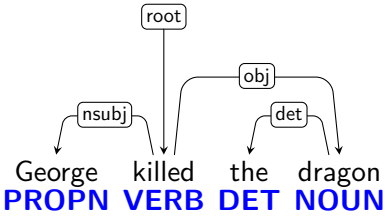
1. UD must be satisfactory on linguistic analysis grounds for individual languages.
2. UD must be good for linguistic typology, i.e., providing a suitable basis for bringing out cross-linguistic parallelism across languages and language families.
3. UD must be suitable for rapid, consistent annotation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

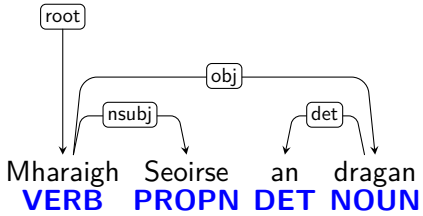
It's easy to come up with a proposal that improves UD on one of these dimensions. The interesting and difficult part is to improve UD while remaining sensitive to all these dimensions.

Same Thing Same Way

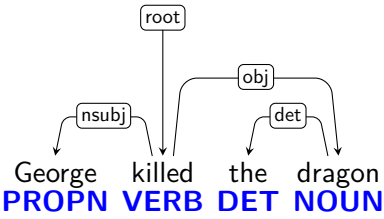


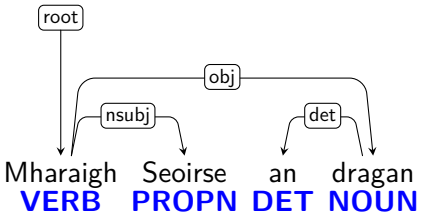
Same Thing Same Way

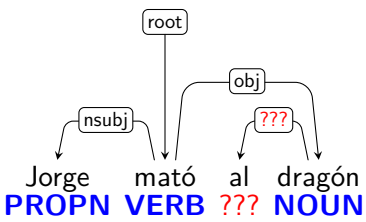




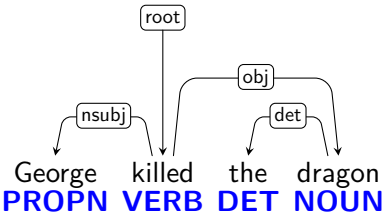
Same Thing Same Way

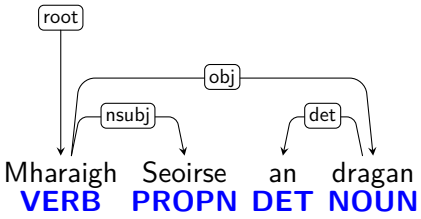


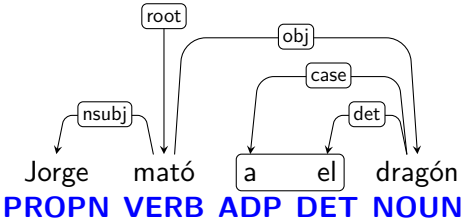




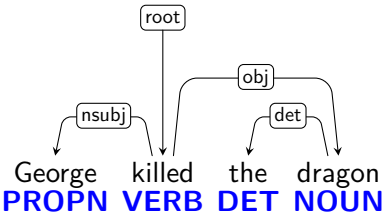
Same Thing Same Way

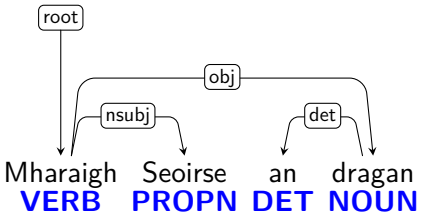


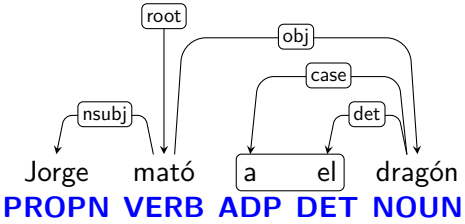


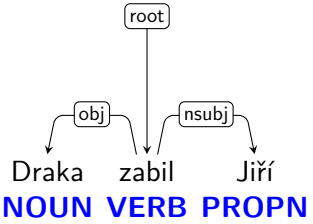


Same Thing Same Way

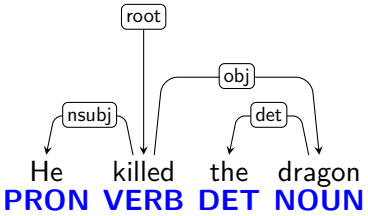




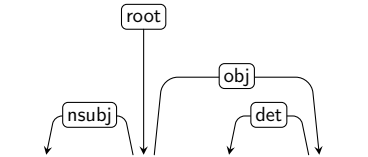


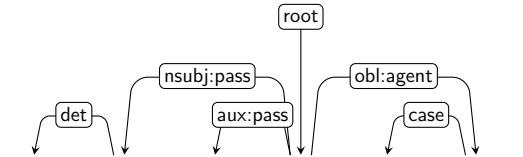


Same Meaning $\neq$ Same Construction!

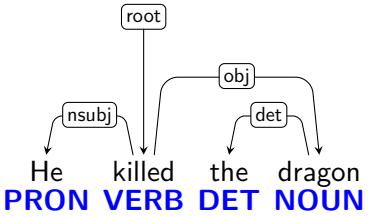


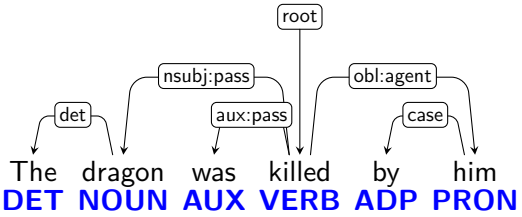
Same Meaning ≠ Same Construction!

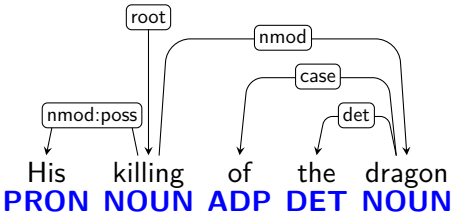




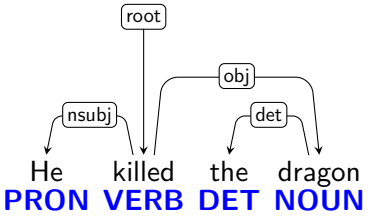
Same Meaning ≠ Same Construction!

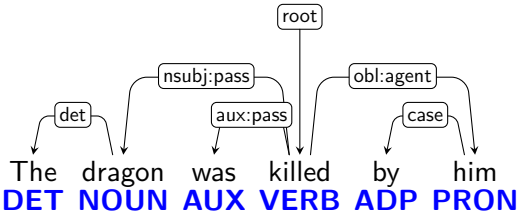


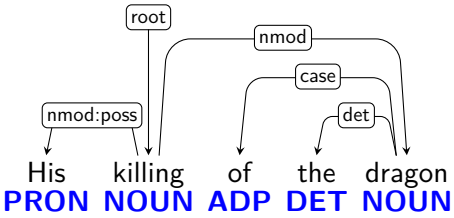


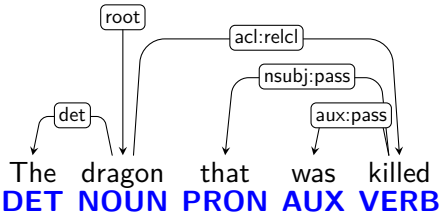


Same Meaning ≠ Same Construction!

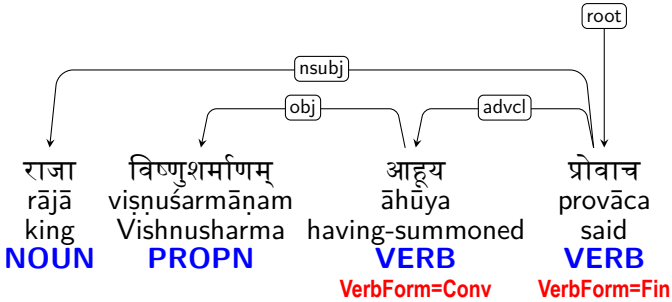




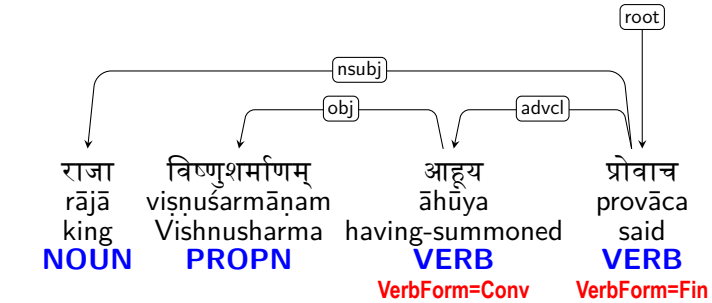




Language-specific Preferences



Language-specific Preferences



Morphological Annotation

Le	chat	chasse	les	chiens	.
le	chat	chasser	le	chien	.
DET	NOUN	VERB	DET	NOUN	PUNCT
Definite=Def Gender=Masc Number=Sing	Gender=Masc Number=Sing	Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm=Fin	Definite=Def Gender=Masc Number=Plur	Gender=Masc Number=Plur	

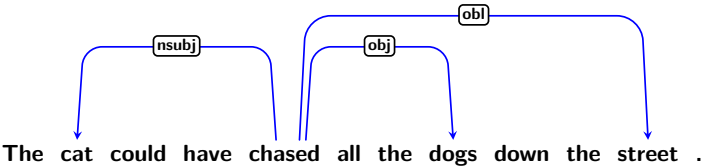
- Lemma representing the semantic content of a word
- Part-of-speech tag representing its grammatical class
- Features representing lexical and grammatical properties of the lemma or the particular word form

Syntactic Annotation

The cat could have chased all the dogs down the street .

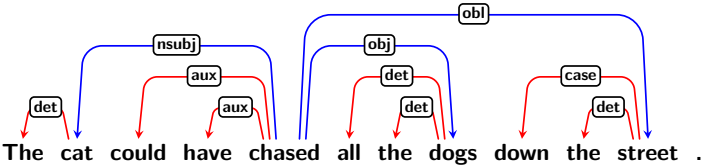
- Content words are related by dependency relations
- Function words attach to the content word they modify
- Punctuation attach to head of phrase or clause

Syntactic Annotation



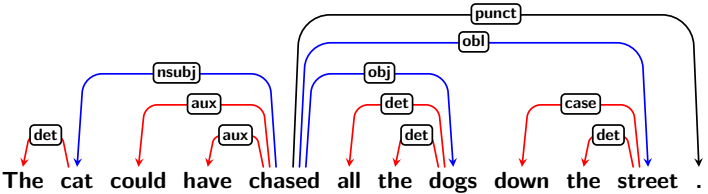
- Content words are related by dependency relations
- Function words attach to the content word they modify
- Punctuation attach to head of phrase or clause

Syntactic Annotation



- Content words are related by dependency relations
- Function words attach to the content word they modify
- Punctuation attach to head of phrase or clause

Syntactic Annotation



- Content words are related by dependency relations
- Function words attach to the content word they modify
- Punctuation attach to head of phrase or clause

CoNLL-U Format

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1	Le	le	DET	_	_	2	det	_	_
2	chat	chat	NOUN	_	_	3	nsubj	_	_
3	boit	boire	VERB	_	_	0	root	_	_
4-5	du	_	_	_	_	_	_	_	_
4	de	de	ADP	_	_	6	case	_	_
5	le	le	DET	_	_	6	det	_	_
6	lait	lait	NOUN	_	_	3	obj	_	SpaceAfter=No
7	.	.	PUNCT	_	_	3	punct	_	_

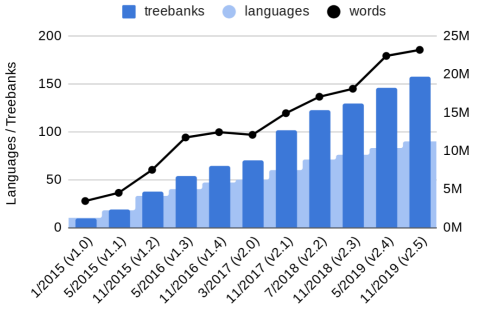
- Revised and extended version of CoNLL-X format
- Two-level segmentation and enhanced dependencies

Basic Universal Dependencies: 130 (128) Languages and Growing

▪ **I.-E.:** Armenian (+West), Greek (+Ancient), Albanian, Hittite, Breton, Irish, Manx, Scottish, Welsh, Afrikaans, Danish, Dutch, English, Faroese, Frisian, German, Gothic, Icelandic, Low Saxon, Norwegian, Swedish, Swiss German, Catalan, French, Galician, Italian, Latin, Ligurian, Neapolitan, Old French, Portuguese, Romanian, Spanish, Umbrian, Belarusian, Bulgarian, Church Slavonic, Croatian, Czech, Old Russian, Polish, Pomak, Russian, Serbian, Slovak, Slovenian, Ukrainian, Upper Sorbian, Latvian, Lithuanian, Kurmanji, Persian, Khunsari, Nayini, Soi, Urdu, Hindi, Kangri, Bhojpuri, Bengali, Marathi, Sanskrit ▪ **Dravidian:** Tamil, Telugu ▪ **Uralic:** Erzya, Estonian, Finnish, Hungarian, Karelian, Livvi, Komi Permyak+Zyrian, Moksha, Sámi North+Skolt ▪ **Turkic:** Kazakh, Old Turkish, Tatar, Turkish, Uyghur, Yakut ▪ Buryat ▪ Xibe ▪ Korean ▪ Japanese ▪ **Sino-T.:** Cantonese, Classical Chinese, Chinese ▪ **Tai-Kadai:** Thai ▪ **Aus.-As.:** Vietnamese ▪ **Austron.:** Indonesian, Javanese, Tagalog, Cebuano ▪ **Pama-Nyu.:** Warlpiri ▪ **Chu.-Kam.:** Chukchi ▪ **Esk.-Al.:** Yupik ▪ **Mayan:** Kiche ▪ **Arawakan:** Apurinã ▪ **Arawan:** Madi ▪ **Tupian:** Akuntsu, Guajajara, Kaapor, Karo, Makurap, Mundurukú, Tupinambá, Mbyá, Guaraní, Teko ▪ **Af.-As.:** Akkadian, Amharic, Arabic Standard+Levantine, Assyrian, Beja, Coptic, Hebrew (+Ancient), Maltese ▪ **Niger-Congo:** Yoruba ▪ **Other:** Basque, Sw. Sign, Naija

Where are we today?

- Brief history of UD:
  - First guidelines launched in October 2014
  - Treebank releases (roughly) every six months
  - Version 2 guidelines/treebanks in 2016–2017
  - New: guideline amendments since May 2022
  - Extensions: MWEs, PropBanks, Coreference
- UD in numbers:
  - 130 languages
  - 228 treebanks
  - 502 contributors
  - 150,000+ downloads
- Past and current UD events:
  - 4 CoNLL and IWPT shared tasks on UD parsing
  - UD workshops: next in Washington 2023
  - COST action: UniDive (since 2022)
  - Next release in November 2022 (v2.11)



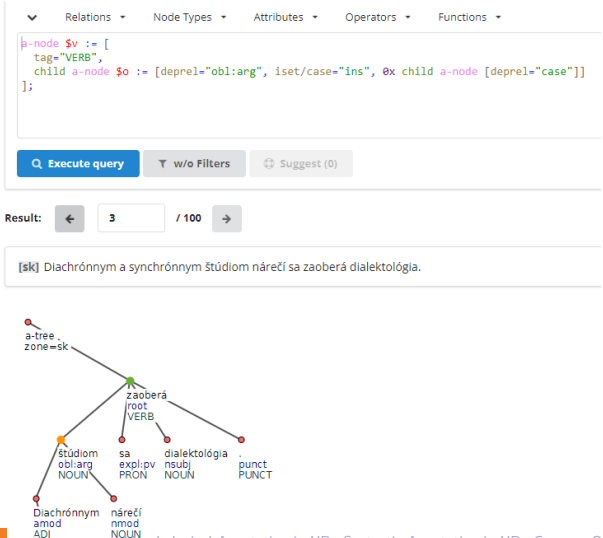
502 Contributors

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Gabrielė Aleksandravičiūtė, Ika Alfina, Avner Algom, Erik Andersen, Lene Antonsen, Katya Aplonova, Angelina Aquino, Carolina Aragon, Glyd Aranes, Maria Jesus Aranzabe, Bilge Nas Arıcan, Þórunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Deniz Baran Aslan, Cengiz Asmazoğlu, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Starkaður Barkarson, Rodolfo Basile, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Kepa Bengoetxea, Yifat Ben Moshe, Gözde Berk, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Kristín Bjarnadóttir, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Anouck Braggaar, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Lauren Cassidy, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Neslihan Cesur, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Shweta Chauhan, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Jayeol Chun, Juyeon Chung, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Daniela Corbetta, Marine Courtin, Mihaela Cristescu, Philemon Daniel, Elizabeth Davidson, Mathieu Dehouck, Martina de Laurentiis, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Sandra Eiche, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Federica Favero, Jannatul Ferdaousi, Marília Fernanda, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Federica Gamba, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Fabrício Ferraz Gerardi, Kim Gerdes, Filip Ginter, Gustavo Godoy, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Hinrik Hafsteinsson, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Muhammad Yudistira Hanifmuti, Takahiro Harada, Sam Hardwick, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Takumi Ikeda, Anton Karl Ingason, Radu Ion, Elena Irimia, Ọlájídé Ishola, Kaoru Ito, Siratun Jannat, Tomáš Jelínek, Apoorva Jha, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Sarveswaran K, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, Ritván Karahóǧa, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Elena Klyachko, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Mehmet Köse, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Parameswari Krishnamurthy, Sandra Kübler, Oğuzhan Kuyrukçu, Aslı Kuzgun, Sookyoung Kwak, Veronika Laippala, Lucia Lam, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phương Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Cheuk Ying Li, Josie Li, Keying Li, Yuan Li, KyungTae Lim, Bruna Lima Padovani, Krister Lindén, Nikola Ljubešić, Olga Loginova, Stefano Lusito, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Menel Mahamdi, Jean Maillard, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Büşra Marşan, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Stella Markantonatou, Héctor Martínez Alonso, Lorena Martín Rodríguez, André Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Alessandro Mazzei, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Tatiana Merzhevich, Niko Miekka, Karina Mischenkova, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Judit Molnár, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Giovanni Moretti, Keiko Sophie Mori, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhlé, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Manuela Nevaci, Lương Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Alireza Nourian, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayọ̀ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Noam Ordan, Petya Osenova, Robert Östling, Lilja Øvrelid, Şaziye Betül Özateş, Merve Özçelik, Arzucan Özgür, Balkız Öztürk Başaran, Teresa Paccosi, Alessio Palmero Aprosio, Hyunji Hayley Park, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Giulia Pedonese, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Andrea Peverelli, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Mizanur Rahoman, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Mathilde Regnault, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Putri Rizqiyah, Luisa Rocha, Eiríkur Rögnvaldsson, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Davide Rovati, Ben Rozonoyer, Olga Rudina, Jack Rueter, Kristján Rúnarsson, Shoval Sadde, Pegah Safari, Benoît Sagot, Aleksi Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Ezgi Sanıyar, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Shefali Saxena, Kevin Scannell, Salvatore Scarlata, Nathan Schneider, Sebastian Schuster, Lane Schwartz, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Syeda Shahzadi, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Yana Shishkina, Muh Shohibussirri, Dmitry Sichinava, Janine Siewert, Einar Freyr Sigurðsson, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Shafi Sourov, Carolyn Spadine, Rachele Sprugnoli, Vivian Stamou, Steinþór Steingrímsson, Antonio Stella, Milan Straka, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Shingo Suzuki, Daniel Swanson, Zsolt Szántó, Chihiro Taguchi, Dima Taji, Yuta Takahashi, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Dipta Tanaya, Mirko Tavoni, Samson Tella, Isabelle Tellier, Marinella Testori, Guillaume Thomas, Sara Tonelli, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Elena Vagnoni, Sowmya Vajjala, Rob van der Goot, Martine Vanhove, Daniel Van Niekerk, Gertjan van Noord, Viktor Varga, Uliana Vedenina, Eric Villemonte de la Clergerie, Veronika Vincze, Natalia Vlasova, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Shira Wigderson, Sri Hartati Wijono, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Arife Betül Yenice, Olcay Taner Yıldız, Zhuoran Yu, Arlisa Yuliawati, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, He Zhou, Hanzhi Zhu, Yilun Zhu, Anna Zhuravleva, Rayan Ziane

Use in Digital Humanities

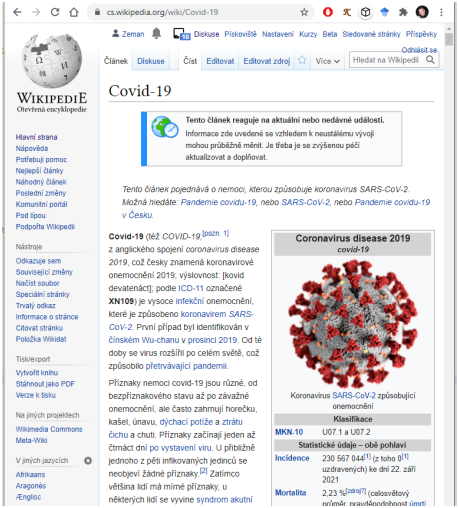
Linguists Can Search Treebanks

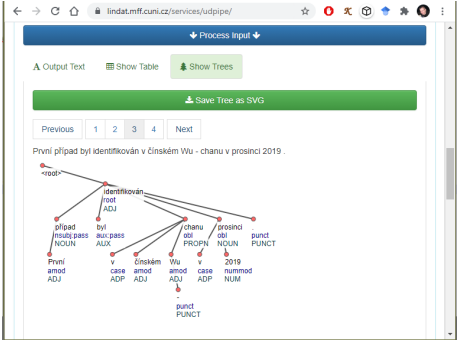
https://lindat.mff.cuni.cz/services/pmltq/



Linguists Can Parse and Search New Data

https://lindat.mff.cuni.cz/services/udpipe/



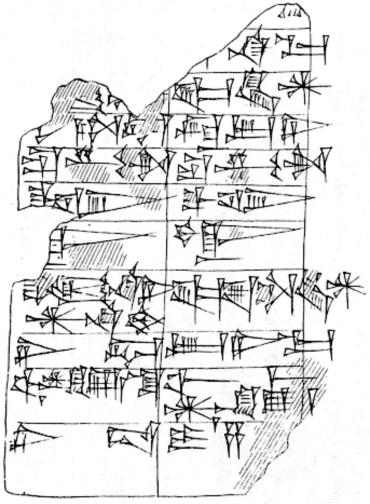


Language Learning

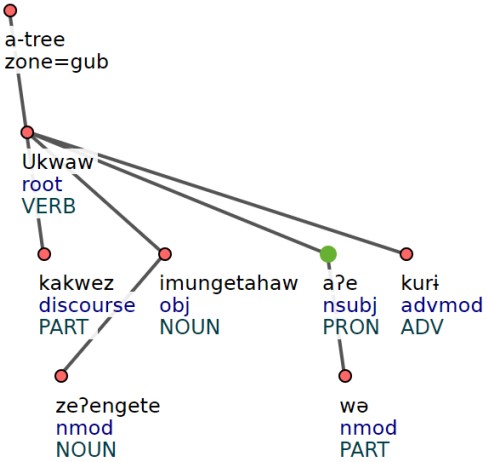
- Check grammar usage in the corpus
- Learner corpora

Historical Linguistics, Classical Languages

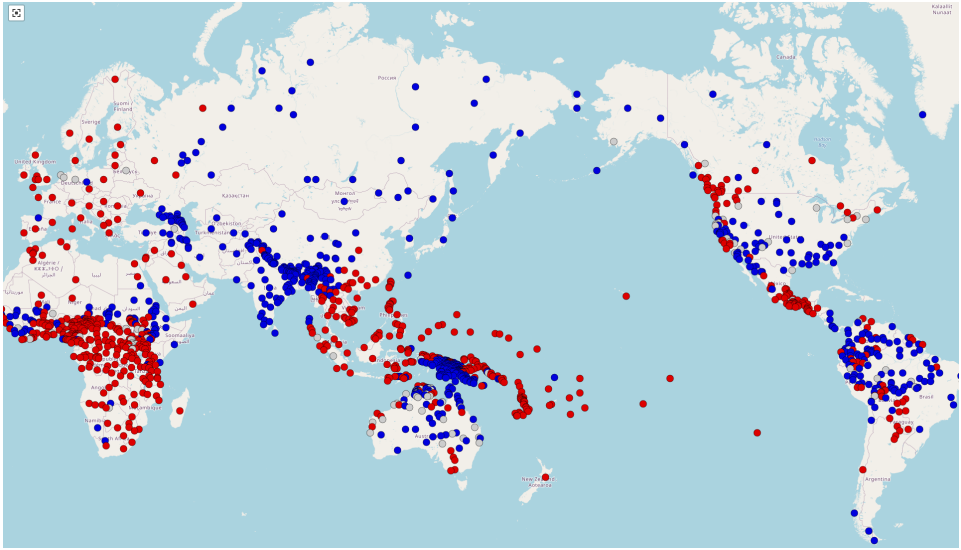
- Old Turkish
- Classical Chinese
- Sanskrit
- Hittite
- Akkadian
- Coptic
- Ancient Hebrew
- Ancient Greek
- Latin
- Old French
- Gothic
- Old Church Slavonic
- Old East Slavic



Documentation of Endangered Languages



Linguistic Typology



Linguistic Typology

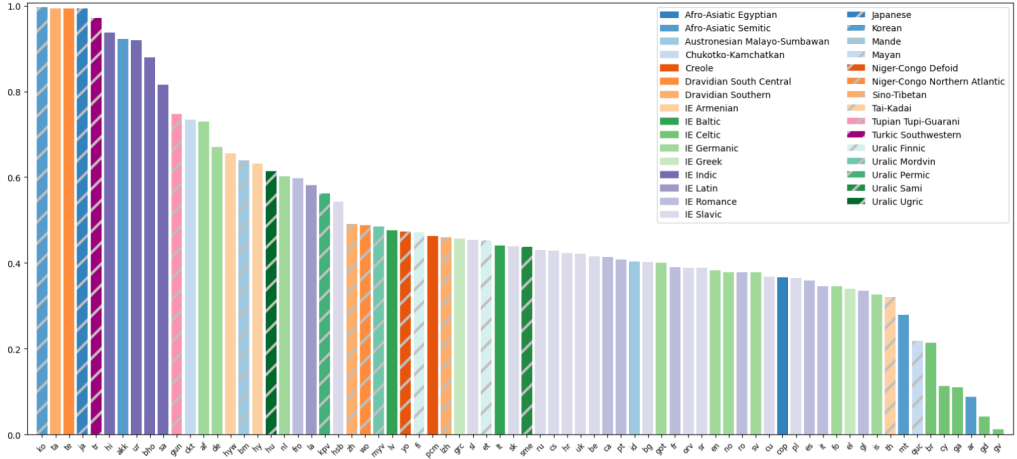


Figure 7 Percentage of head-final dependencies. Each bar is one language.

Morphological Annotation in UD

Outline

1. Introduction
2. Morphological Annotation in UD
3. Syntactic Annotation in UD
4. Core vs. Oblique
5. Enhanced Universal Dependencies
6. UD Tools

Morphological Annotation in UD

- Tokenization / word segmentation

- Lemmatization **(LEMMA)**
- Universal part-of-speech tags **(UPOS)**
- Universal features **(FEATS)**
- Language-specific features

Tokenization

"María, I love you!" Juan exclaimed.

«¡María, te amo!», exclamó Juan.
X PRON X VERB X

« ¡ María , te amo ! » ,
PUNCT PUNCT PROPN PUNCT PRON VERB PUNCT PUNCT PUNCT

- Classic tokenization:
  - Separate punctuation from words
  - Recognize certain clusters of symbols like "..."
  - Perhaps keep together things like `user@mail.x.edu`

Word Segmentation

Let's go to the sea.

Vámonos	al	mar	.	Vamos	nos	a	el	mar	.
VERB?	X	NOUN	PUNCT	VERB	PRON	ADP	DET	NOUN	PUNCT

- **Syntactic word** vs. orthographic word
- **Multi-word tokens**
- Two-level scheme:
  - Tokenization (low level, punctuation, concatenative)
  - Word segmentation (higher level, not necessarily concatenative)

Word Segmentation

- Lexicalist hypothesis:
  - Words (not morphemes) are the basic units in syntax
  - Words enter in dependency relations
  - Words are forms of lemmas and have morphological features

- Orthographic vs. syntactic word
  - Syntactically autonomous part of orthographic word
  - Contractions *(al = a + el)*
  - Clitics *(vámonos = vamos + nos)*
    - *¿A qué hora nos vamos mañana?*
    - *Nos despertamos a las cinco.*
      "We wake up at five."
    - *Nuestro guía nos despierta a las cinco.*
      "Our guide wakes us up at five."

Contractions in Arabic (EXTRA)

He abdicated in favour of his son Baudouin.

يتنازل	عن	العرش	لابنه	بودوان
yatanāzalu	**ᶜan**	**al-ᶜarši**	**li+ibni+hi**	**būdūān**
surrendered	**on**	**the throne**	**to son his**	**Baudouin**
VERB	**ADP**	**NOUN**	**ADP+NOUN+PRON**	**PROPN**

Chinese Word Segmentation

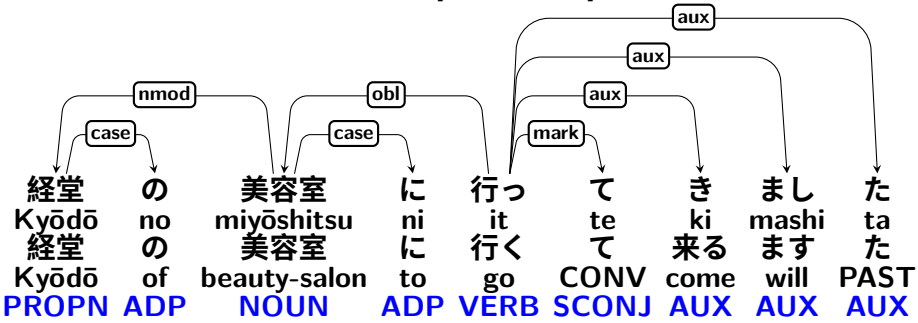
We are now in Valencia.

現在我們在瓦倫西亞。
Xiàn zài wǒ men zài wǎ lún xī yǎ.
We are now in Valencia.

現在	我們	在	瓦倫西亞	。
Xiànzài	wǒmen	zài	Wǎlúnxīyǎ	.
Now	we	in	Valencia	.
ADV	PRON	ADP	PROPN	PUNCT

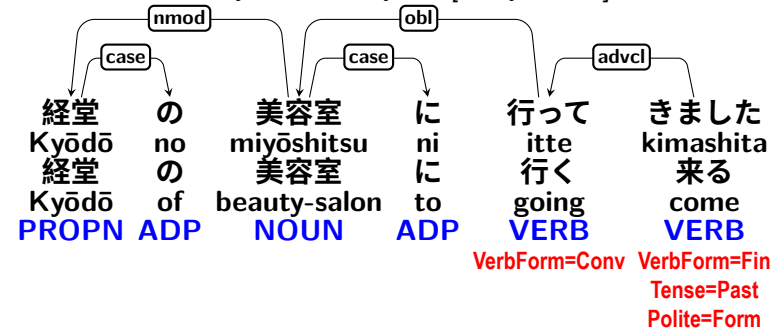
Words in Japanese (EXTRA)

I went to the beauty salon of Kyōdō [, Beyond-R.]



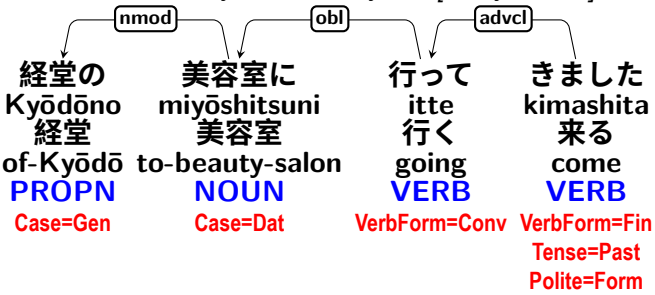
Words in Japanese (EXTRA)

I went to the beauty salon of Kyōdō [, Beyond-R.]



Words in Japanese (EXTRA)

I went to the beauty salon of Kyōdō [, Beyond-R.]



Vietnamese: Words with Spaces

All the concrete country roads are the result of...

Tất cả	đường	bêtông	nội đồng	là	thành quả	...
All	road	concrete	country	is	achievement	...
PRON	NOUN	NOUN	NOUN	AUX	NOUN	PUNCT

- Spaces delimit monosyllabic morphemes, not words.
- Multiple syllables without space occur in loanwords *(bêtông)*.
- Spaces are allowed to occur word-internally in Vietnamese UD.

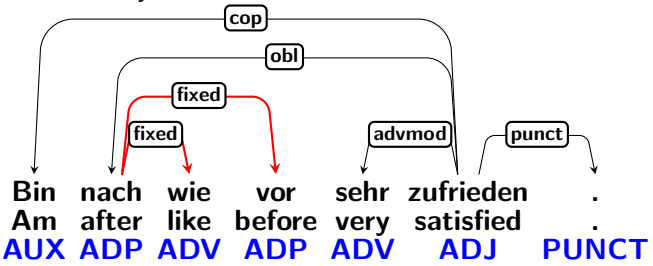
Numbers with Spaces

```
# text = Il touche environ 100 000 sesterces par an.
1  Il          il          PRON  ...  2  nsubj  _ _
2  touche      toucher     VERB  ...  0  root   _ _
3  environ     environ     ADV   ...  4  advmod _ _
4  100 000     100 000     NUM   ...  5  nummod _ _
5  sesterces   sesterce     NOUN  ...  2  obj    _ _
6  par         par         ADP   ...  7  case   _ _
7  an          an          NOUN  ...  2  obl    _ SpaceAfter=No
8  .           .           PUNCT ...  2  punct  _ _
```

Fixed Expressions

One syntactic word spans several orthographic words?

I am still very satisfied.



Word Segmentation Summary (EXTRA)

- When to split?
  - Only part of the token involved in a relation to something outside the token? Split!

Word Segmentation Summary (EXTRA)

- When to split?
  - Only part of the token involved in a relation to something outside the token? Split!
  - Hard time finding POS tag? Split!

Word Segmentation Summary (EXTRA)

- When to split?
  - Only part of the token involved in a relation to something outside the token? Split!
  - Hard time finding POS tag? Split!
  - Hard time finding dependency relation? Don't split!
    - Or not hard time but the relation would be compound, flat, fixed or goeswith.

Word Segmentation Summary (EXTRA)

- When to split?
  - Only part of the token involved in a relation to something outside the token? Split!
  - Hard time finding POS tag? Split!
  - Hard time finding dependency relation? Don't split!
    - Or not hard time but the relation would be compound, flat, fixed or goeswith.
  - Border case? Keep orthographic words (if they exist).

Word Segmentation Summary (EXTRA)

- When to split?
  - Only part of the token involved in a relation to something outside the token? Split!
  - Hard time finding POS tag? Split!
  - Hard time finding dependency relation? Don't split!
    - Or not hard time but the relation would be compound, flat, fixed or goeswith.
  - Border case? Keep orthographic words (if they exist).

- Words with spaces
  - Vietnamese writing system
  - Very restricted set of exceptions (numbers)
  - Special relations elsewhere (`fixed`, `compound`)

Recoverability: CoNLL-U Format

text = Vámonos al mar.
text_en = Let's go to the sea.

ID	FORM	LEMMA	UPOS	...	HEAD		_ MISC
1-2	Vámonos	_	_	...	_	_	_ _
1	Vamos	ir	VERB	...	0	root	_ _
2	nos	nosotros	PRON	...	1	obj	_ _
3-4	al	_	_	...	_	_	_ _
3	a	a	ADP	...	5	case	_ _
4	el	el	DET	...	5	det	_ _
5	mar	mar	NOUN	...	1	obl	_ SpaceAfter=No
6	.	.	PUNCT	...	1	punct	_ _

Recoverability: CoNLL-U Format

text = Vámonos al mar.
text_en = Let's go to the sea.

ID	FORM	LEMMA	UPOS	...	HEAD		_ MISC
1-2	Vámonos	_	_	...	_	_	_ _
1	Vamos	ir	VERB	...	0	root	_ _
2	nos	nosotros	PRON	...	1	obj	_ _
3-4	al	_	_	...	_	_	_ _
3	a	a	ADP	...	5	case	_ _
4	el	el	DET	...	5	det	_ _
5-6	**mar.**	**_**	**_**	**...**	**_**	**_**	**_ _**
5	mar	mar	NOUN	...	1	obl	_ _
6	.	.	PUNCT	...	1	punct	_ _

Tokenization vs. Multi-word Tokens (EXTRA)

- Parallelism among closely related languages
  - ca: **informar-se sobre el patrimoni cultural**
  - es: **informarse sobre el patrimonio cultural**
  - en: *learn about cultural heritage*

- ca: L'únic que veig és => **L' únic que veig és**
- en: don't => **do n't**

- No strict guidelines for tokenization (yet)
  - UD English: **non-stop**, **post-war**: single-word tokens
  - UD Czech: **non-stop** would be split to three tokens

Tokenization vs. Multi-word Tokens Summary (EXTRA)

- Punctuation involved? Low level!
  - Exceptions: Spanish-Catalan parallelism.

Tokenization vs. Multi-word Tokens Summary (EXTRA)

- Punctuation involved? Low level!
  - Exceptions: Spanish-Catalan parallelism.
- Boundary between two letters? Typically high level.
  - Exceptions: Chinese, Japanese.

Tokenization vs. Multi-word Tokens Summary (EXTRA)

- Punctuation involved? Low level!
  - Exceptions: Spanish-Catalan parallelism.
- Boundary between two letters? Typically high level.
  - Exceptions: Chinese, Japanese.
- Non-concatenative? High level!

Lemmas

- Basic or citation form ($\Rightarrow$ it is an existing word in most cases)
- Disambiguating ids, if available, go to MISC
- Derivational vs. inflectional morphology (if participles are ADJ, their lemma should not be infinitive)

Lemmas (EXTRA)

within a year Algeria will become an islamic state

13	do	do	ADP	...	LId=do-1
14	roka	rok	NOUN	...	_
15	se	se	PRON	...	LGloss=(zvr._zájmeno/částice)
16	Alžírsko	Alžírsko	PROPN	...	_
17	stane	stát	VERB	...	LId=stát-2
18	islámským	islámský	ADJ	...	_
19	státem	stát	NOUN	...	LId=stát-1|LGloss=(státní_útvar)|SpaceAfter=No

- Basic or citation form
- Disambiguating ids, if available, go to MISC

Part-of-Speech Tags

http://universaldependencies.org/u/pos/index.html

Open		Closed		Other	
NOUN	common noun	**PRON**	pronoun	**PUNCT**	punctuation
PROPN	proper noun	**DET**	determiner	**SYM**	symbol
VERB	verb	**AUX**	auxiliary	**X**	unknown
ADJ	adjective	**NUM**	numeral		
ADV	adverb	**ADP**	adposition		
INTJ	interjection	**SCONJ**	subordinator		
		CCONJ	coordinator		
		PART	particle		

- Taxonomy of 17 universal POS tags
- All languages use the same inventory
  - Not all tags have to be used by all languages
  - Need extensions? Use features!

Part-of-Speech Tags (EXTRA)

- Traditionally a mixture of morphological, syntactic/distributional and semantic/notional criteria
- Prefer grammatical > semantic criteria
  - Language-particular definition of a category
- But the **name** of the category is universal
  - Translated words: overlapping categories, but not perfect match
    - UPOS of English *dog* is **NOUN**; so is French *chien* or Russian *собака*
- Preferably POS is encoded in lexicon, not heavily usage-dependent
  - But not for incompatible syntactic functions (e.g. **PRON** vs. **SCONJ**)

Universal Features

http://universaldependencies.org/u/feat/index.html

- **PronType** *(pronoun type)*
- **NumType** *(numeral type)*
- **Poss** *(possessive)*
- **Reflex** *(reflexive)*
- **Foreign** *(foreign)*
- **Abbr** *(abbreviation)*
- **Typo** *(typo)*
- **Gender** *(gender)*
- **Animacy** *(animacy)*
- **NounClass** *(noun class)*
- **Number** *(number)*
- **Case** *(case)*
- **Definite** *(definiteness)*
- **Degree** *(degree of comparison)*
- **VerbForm** *((de)verbal form)*
- **Mood** *(mood)*
- **Tense** *(tense)*
- **Aspect** *(aspect)*
- **Voice** *(voice)*
- **Evident** *(evidentiality)*
- **Polarity** *(polarity)*
- **Person** *(person)*
- **Polite** *(politeness)*
- **Clusivity** *(clusivity)*

Features (EXTRA)

Lexical	Inflectional ("Nominal")	Inflectional ("Verbal, Pronominal")
PronType	Gender	VerbForm
NumType	Animacy	Mood
Poss	NounClass	Tense
Reflect	Number	Aspect
Foreign	Case	Voice
	Definite	Evident
	Degree	Polarity
Abbr		Person
Typo		Polite
		Clusivity

- 24 features, each with a number of possible *values*
- Languages select relevant features
- May add language-specific features or values

Language-Specific Features (EXTRA)

Three types of infinitives in Finnish:
Example: *olla* "to be"

1st	2nd	3rd
olla	ollessa	olemassa
	ollen	olemaan
		olemasta
		olemalla
		olematta

Language-Specific Features (EXTRA)

Joku	yrittää	piristää	itseään	värjäämällä	hiuksensa
Someone	tries	to-uplift	oneself	by-staining	their-hair
PRON	VERB	VERB	PRON	VERB	NOUN
	VerbForm=Fin Mood=Ind Tense=Pres	VerbForm=Inf		VerbForm=Inf3 Case=Ade	

Language-Specific Features (EXTRA)

Joku	yrittää	piristää	itseään	värjäämällä	hiuksensa
Someone	tries	to-uplift	oneself	by-staining	their-hair
PRON	VERB	VERB	PRON	VERB	NOUN
	VerbForm=Fin	VerbForm=Inf		VerbForm=~~Inf3~~	
	Mood=Ind			Case=Ade	
	Tense=Pres				

Joku	yrittää	piristää	itseään	värjäämällä	hiuksensa
Someone	tries	to-uplift	oneself	by-staining	their-hair
PRON	VERB	VERB	PRON	VERB	NOUN
	VerbForm=Fin	VerbForm=Inf		VerbForm=Inf	
	Mood=Ind	InfForm=1		InfForm=3	
	Tense=Pres			Case=Ade	

Layered Features (EXTRA)

Czech adjectives agree with nouns in gender.

velký	bratr
big	brother
ADJ	NOUN
Gender=Masc	Gender=Masc

velká	sestra
big	sister
ADJ	NOUN
Gender=Fem	Gender=Fem

Layered Features (EXTRA)

Possessive adjectives: agreement gender vs. lexical gender

otcův	bratr	matčin	bratr
father's	brother	mother's	brother
ADJ	NOUN	ADJ	NOUN
Gender=Masc	Gender=Masc	Gender=Masc	Gender=Masc
Gender[psor]=Masc		Gender[psor]=Fem	

otcova	sestra	matčina	sestra
father's	sister	mother's	sister
ADJ	NOUN	ADJ	NOUN
Gender=Fem	Gender=Fem	Gender=Fem	Gender=Fem
Gender[psor]=Masc		Gender[psor]=Fem	

Multi-valued Features (Disjunction / Parallel Application) (EXTRA)

- Feature can have two or more values
- Interpreted as disjunction
- Example: in some languages, many pronouns function both as interrogative and relative, but some pronouns are only relative. The former will have **PronType=Int,Rel**
- In other cases, it is desirable to disambiguate by context. Polish *którym* (form of *który* "which") can be **Case=Ins**, **Loc** in singular or **Dat** in plural but we do not want to annotate **Case=Dat,Ins,Loc**!
- All values of the feature/language? Omit the feature completely! Polish: ~~**Gender=Fem,Masc,Neut**~~. Spanish: ~~**Gender=Fem,Masc**~~

Multi-valued Features (Serial Application) (EXTRA)

- Currently used in Turkish (language-specific values)
- Two or more morphemes in chain, affecting the same feature
- Example: **Voice=CauPass** (causative + passive => someone is caused to do something)
  - *yanıl* "be wrong"
  - *yanılmışım* **Voice=Act** "I was wrong"
  - *okuru yanılttığını* **Voice=Cau** "mislead the reader"
  - *okurlar yanıltılmıştır* **Voice=CauPass** "readers were misled"

Multi-valued Features (Serial Application) (EXTRA)

- Currently used in Turkish (language-specific values)
- Two or more morphemes in chain, affecting the same feature
- Example: **Voice=CauPass** (causative + passive => someone is caused to do something)
  - *yanıl* "be wrong"
  - *yanılmışım* **Voice=Act** "I was wrong"
  - *okuru yanılttığını* **Voice=Cau** "mislead the reader"
  - *okurlar yanıltılmıştır* **Voice=CauPass** "readers were misled"
  - Hypothetical: **Voice=PassCau** (not used in Turkish) could mean "to cause something to be done by someone"

Features Apply to Individual Words

Future tense in Spanish and German: no Tense=Fut in German!

Dormirá	Er	wird	schlafen
He-will-sleep	He	will	sleep
VERB	PRON	AUX	VERB
VerbForm=Fin	PronType=Prs	VerbForm=Fin	VerbForm=Inf
Mood=Ind	Number=Sing	Mood=Ind	
Tense=Fut	Person=3	Tense=Pres	
Number=Sing	Gender=Masc	Number=Sing	
Person=3	Case=Nom	Person=3	

Participle Types (EXTRA)

некурящий	человек	начавшийся	разговор
nekurjaščij	čelovek	načavšijsja	razgovor
non-smoking	person	that-has-started	conversation
ADJ	NOUN	ADJ	NOUN
VerbForm=Part		VerbForm=Part	
Tense=Pres		Tense=Past	
Gender=Masc	Gender=Masc	Gender=Masc	Gender=Masc
Number=Sing	Number=Sing	Number=Sing	Number=Sing
Case=Nom	Case=Nom	Case=Nom	Case=Nom

- Sometimes features like Tense help distinguish participle types
- Not the same tense as with finite verbs (reference point)
- But useful because:
  - We use known UD primitives rather than language-specific labels such as ~~VerbForm=PastPart~~, or even ~~ParticType=Past~~
  - Reasonably close to the grammatical meaning

Conflicting Traditional Terminologies (EXTRA)

- If possible, stay compatible with traditional grammar
- Often it is not possible: terminology conflicts
- **VerbForm=Conv** – *converb*, *transgressive*, *adverbial participle*, *gerund*

Conflicting Traditional Terminologies (EXTRA)

- If possible, stay compatible with traditional grammar
- Often it is not possible: terminology conflicts
- **VerbForm=Conv** – *converb*, *transgressive*, *adverbial participle*, *gerund*
- *Gerund* (**VerbForm=Ger**)
  - English: close to verbal nouns (**VerbForm=Vnoun**)
  - Spanish: more like present participle (**VerbForm=Part | Tense=Pres**)
  - Slavic: converb (**VerbForm=Conv**)

Conflicting Traditional Terminologies (EXTRA)

- If possible, stay compatible with traditional grammar
- Often it is not possible: terminology conflicts
- **VerbForm=Conv** – *converb*, *transgressive*, *adverbial participle*, *gerund*
- *Gerund* (**VerbForm=Ger**)
  - English: close to verbal nouns (**VerbForm=Vnoun**)
  - Spanish: more like present participle (**VerbForm=Part | Tense=Pres**)
  - Slavic: converb (**VerbForm=Conv**)
- *Aorist*
  - Ancient Greek, Turkish: neutral non-past tense (they use a language-specific value **Tense=Aor**)
  - Slavic languages: simple past tense (**Tense=Past**)

Conflicting Traditional Terminologies (EXTRA)

A	ko	so	se	leta	1942	vračali	...
And	as	they-were	REFL	in-year	1942	returning	...
CCONJ	SCONJ	AUX	PRON	NOUN	NUM	VERB	
		VerbForm=Fin				VerbForm=Part	
		Tense=Pres				Tense=Past?	

Conflicting Traditional Terminologies (EXTRA)

A	ko	so	se	leta	1942	vračali	...
And	as	they-were	REFL	in-year	1942	returning	...
CCONJ	SCONJ	AUX	PRON	NOUN	NUM	VERB	
		VerbForm=Fin Tense=Pres				VerbForm=Part Tense=Past?	

da	ne	bi	v	Atene	prišli	...
that	not	would	in	Athens	they-come	...
SCONJ	PART	AUX	ADP	PROPN	VERB	
		VerbForm=Fin Mood=Cnd			VerbForm=Part Tense=Past??	

Conflicting Traditional Terminologies (EXTRA)

A	ko	so	se	leta	1942	vračali	...
And	as	they-were	REFL	in-year	1942	returning	...
CCONJ	SCONJ	AUX	PRON	NOUN	NUM	VERB	
		VerbForm=Fin				VerbForm=Part	
		Tense=Pres				Tense=Past?	

da	ne	bi	v	Atene	prišli	...
that	not	would	in	Athens	they-come	...
SCONJ	PART	AUX	ADP	PROPN	VERB	
		VerbForm=Fin			VerbForm=Part	
		Mood=Cnd			Tense=Past??	

v	prihodnje	ne	bodo	vozili	zgolj	les
in	future	not	they-will	drive	just	wood
ADP	NOUN	PART	AUX	VERB	PART	NOUN
			VerbForm=Fin	VerbForm=Part		
			Tense=Fut	Tense=Past???		

Conflicting Traditional Terminologies (EXTRA)

- West/South Slavic: **VerbForm=Part**
- Russian: **VerbForm=Fin** (past tense)
  - **Tense=Past** useful to distinguish from other participles (especially in Bulgarian)
  - But it is also used for the conditional (any tense)
  - In Slovenian even for the future tense!

Conflicting Traditional Terminologies (EXTRA)

- West/South Slavic: **VerbForm=Part**
- Russian: **VerbForm=Fin** (past tense)
  - **Tense=Past** useful to distinguish from other participles (especially in Bulgarian)
  - But it is also used for the conditional (any tense)
  - In Slovenian even for the future tense!
- Terminology – options:

  cs "active participle" / "past tense"
  ru "past tense" / "finite!"
    - Active participle is something else: *нарушивший* / *narušivšij*

  bg "participle + past (aorist) / imperfect" (two subtypes)
  cu "participle + resultative aspect" (lang-spec)

Conflicting Traditional Terminologies (EXTRA)

- West/South Slavic: **VerbForm=Part**
- Russian: **VerbForm=Fin** (past tense)
  - **Tense=Past** useful to distinguish from other participles (especially in Bulgarian)
  - But it is also used for the conditional (any tense)
  - In Slovenian even for the future tense!
- Terminology – options:

  cs "active participle" / "past tense"
  ru "past tense" / "finite!"
    - Active participle is something else: *нарушивший* / *narušivšij*

  bg "participle + past (aorist) / imperfect" (two subtypes)
  cu "participle + resultative aspect" (lang-spec)

- "l-participle"
  - But that would be a language-specific verb form.

Summary of Morphology

- Multi-word tokens: 1 orthographic token = N syntactic words
- Lemma = citation form of the word
- UPOS = universal part-of-speech tag (17 coarse-grained tags)
- Morphological features (feature-value pairs)
  - Universal feature-value pairs
  - Language-specific values or even features
- Lemmas, tags, and features apply to words (tree nodes), not to multi-word expressions and not to sub-word units (morphemes)
- (EXTRA)
  - Categories are comparable (but not identical) across languages
  - Layered features
  - Multi-valued features

https://universaldependencies.org/

Syntactic Annotation in UD

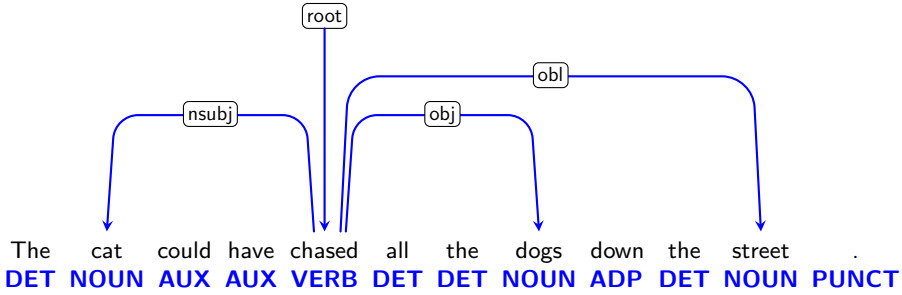
Outline

1. Introduction
2. Morphological Annotation in UD
3. Syntactic Annotation in UD
4. Core vs. Oblique
5. Enhanced Universal Dependencies
6. UD Tools

Syntax

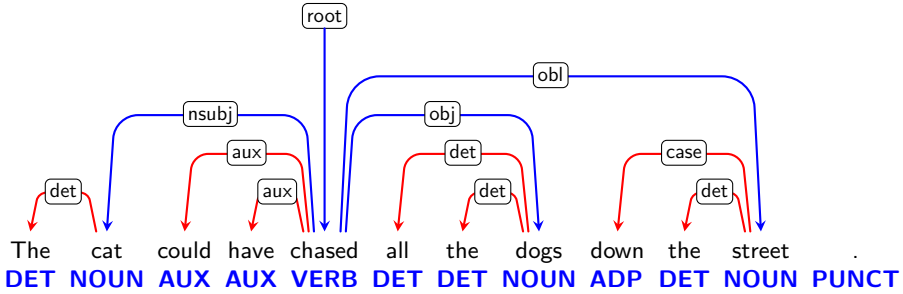
The	cat	could	have	chased	all	the	dogs	down	the	street	.
DET	**NOUN**	**AUX**	**AUX**	**VERB**	**DET**	**DET**	**NOUN**	**ADP**	**DET**	**NOUN**	**PUNCT**

Syntax



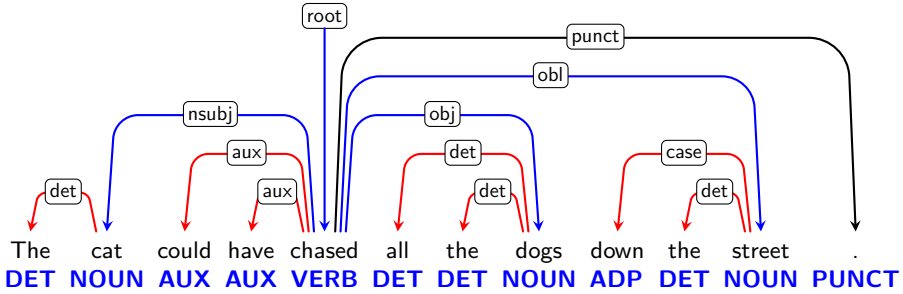
- Content words are related by dependency relations

Syntax



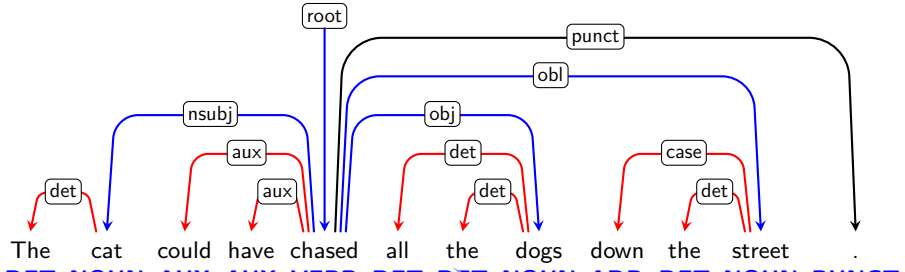
- Content words are related by dependency relations
- Function words attach to closest content words

Syntax

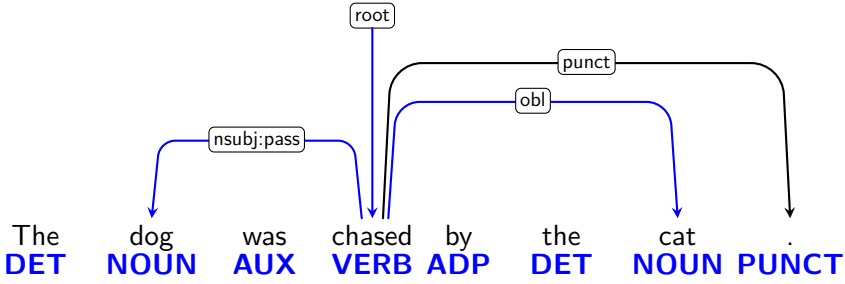


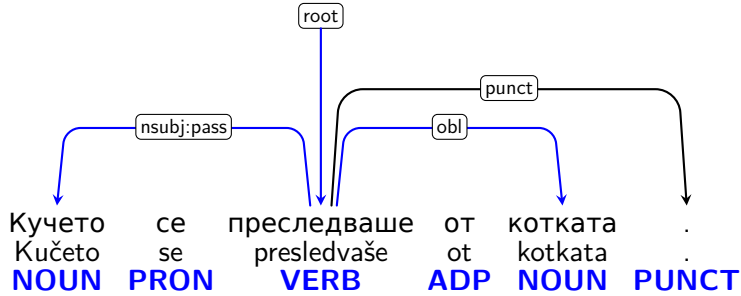
- Content words are related by dependency relations
- Function words attach to closest content words
- Punctuation attach to head of phrase or clause

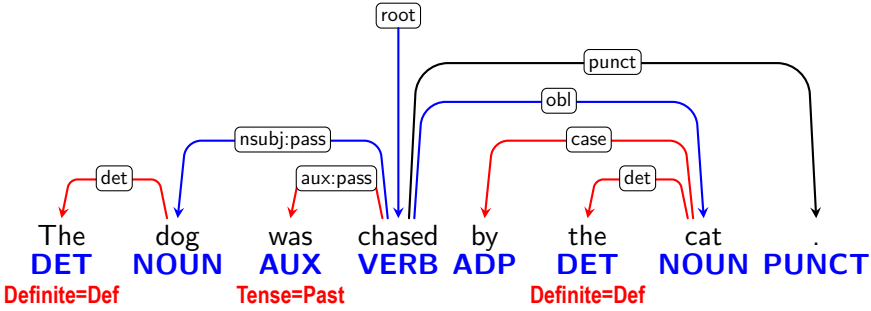
Syntax

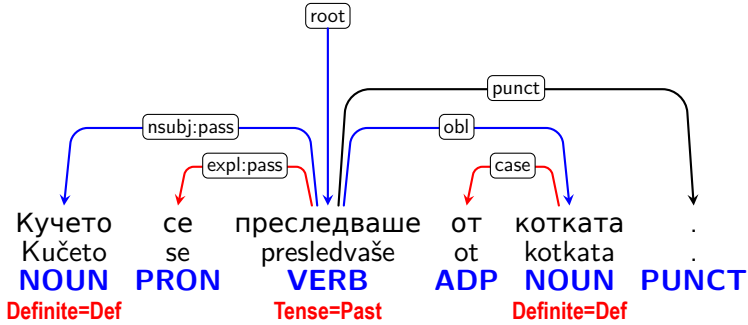


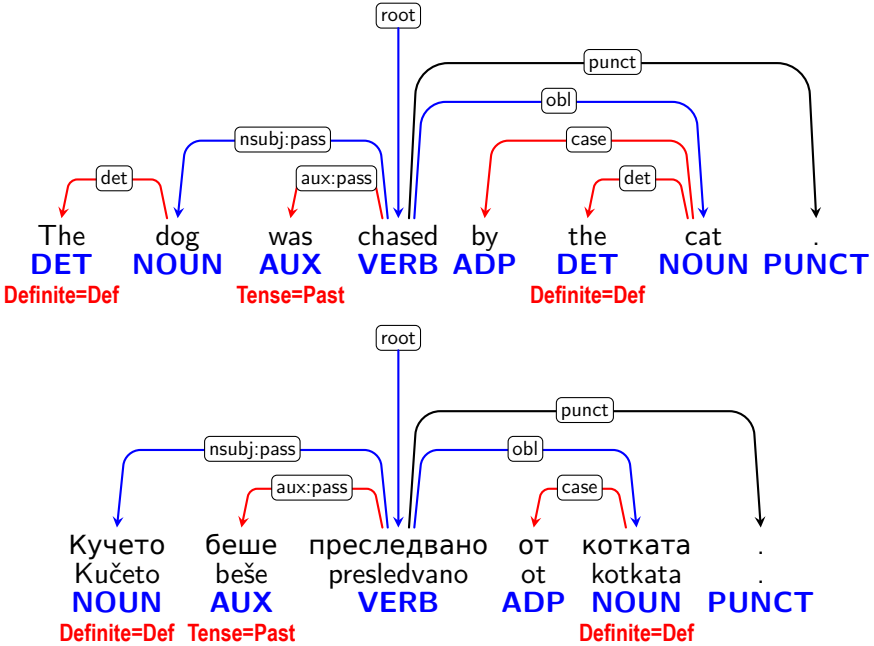
Not "dependency" in the strictly syntactic sense!

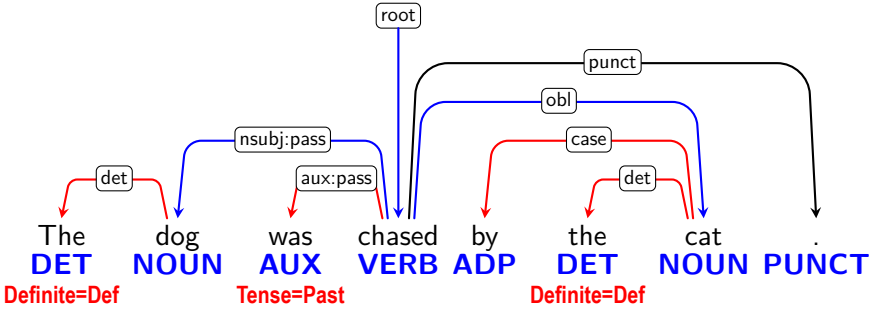

root
punct
obl
nsubj:pass
The dog was chased by the cat .
DET NOUN AUX VERB ADP DET NOUN PUNCT

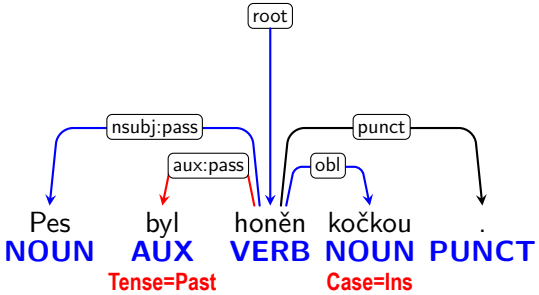

root
punct
obl
nsubj:pass
Кучето се преследваше от котката .
Kučeto se presledvaše ot kotkata .
NOUN PRON VERB ADP NOUN PUNCT


root
punct
obl
nsubj:pass
case
det
aux:pass
det
The dog was chased by the cat .
DET NOUN AUX VERB ADP DET NOUN PUNCT
Definite=Def
Tense=Past
Definite=Def


root
punct
nsubj:pass
obl
expl:pass
case
Кучето се преследваше от котката .
Kučeto se presledvaše ot kotkata .
NOUN PRON VERB ADP NOUN PUNCT
Definite=Def
Tense=Past
Definite=Def


root
punct
obl
nsubj:pass
case
det
aux:pass
det
The dog was chased by the cat .
DET NOUN AUX VERB ADP DET NOUN PUNCT
Definite=Def
Tense=Past
Definite=Def
root
punct
nsubj:pass
obl
aux:pass
case
Кучето беше преследвано от котката .
Kučeto beše presledvano ot kotkata .
NOUN AUX VERB ADP NOUN PUNCT
Definite=Def
Tense=Past
Definite=Def


root
punct
obl
nsubj:pass
case
det
aux:pass
det
The dog was chased by the cat .
DET NOUN AUX VERB ADP DET NOUN PUNCT
Definite=Def
Tense=Past
Definite=Def


root
nsubj:pass
punct
aux:pass
obl
Pes byl honěn kočkou .
NOUN AUX VERB NOUN PUNCT
Tense=Past
Case=Ins

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		

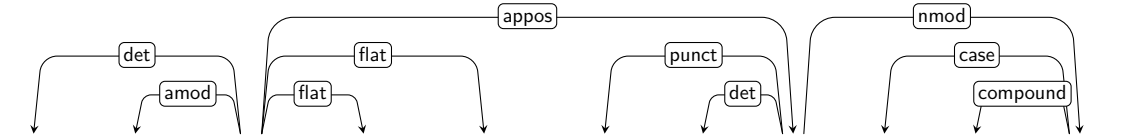
Dependents of Nominals

Nominal	Clausal	Modifier	Function
nmod	acl	amod	det
		nummod	case

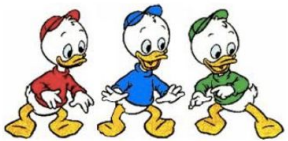
Noun Phrase

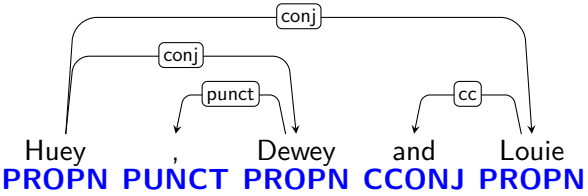
Dependents of Nominals

Nominal	**Clausal**	**Modifier**	**Function**
nmod	acl	amod	det
appos		nummod	case
compound			clf
flat			



Coordination





- Coordinate structures are headed by the first conjunct
  - Subsequent conjuncts depend on it via the conj relation
  - Conjunctions depend on the next conjunct via the cc relation
  - Punctuation marks depend on the next conjunct via the punct relation

Multiword Expressions

Relation	Examples
fixed	*as well, by and large, according to, more than*
flat	*president Havel, New York, four thousand*
compound	*phone book, dress up*
goeswith	*notwith standing, with out*

- UD annotation almost does not permit "words with spaces"
  - Multiword expressions are analyzed using special relations
  - The fixed, flat and goeswith relations are always head-initial
  - The compound relation reflects the internal structure
- Words with spaces allowed in exceptional cases:
  - Vietnamese (spaces delimit syllables, not words)
  - Numbers ("1 000 000")
  - Possibly other approved cases, e.g. multi-word abbreviations

Other Relations

Other Relations

Relation	Explanation
parataxis	Loosely linked clauses of same rank
list	Lists without syntactic structure
orphan	Orphans in ellipsis linked together
reparandum	Disfluency linked to (speech) repair
dep	Unspecified dependency
root	The single syntactically independent element of the sentence

Language-specific Relation Subtypes

- Language-specific relations are subtypes of universal relations added to capture important phenomena
- Subtyping permits us to "back off" to universal relations

Language-specific Relation Subtypes

Relation	Explanation
acl:relcl	Relative clause (the boy who lived)
compound:prt	Verb particle (dress up)
nmod:poss	Possessive nominal (Mary 's book)
obl:agent	Agent in passive (saved by the bell)

Core vs. Oblique

Outline

1. Introduction
2. Morphological Annotation in UD
3. Syntactic Annotation in UD
4. Core vs. Oblique
5. Enhanced Universal Dependencies
6. UD Tools

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			

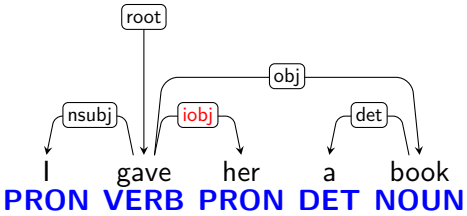
Dependents of Verbs, Adjectives and Adverbs

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

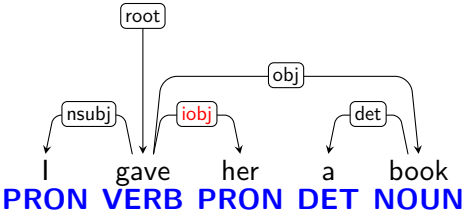
Dependents of Nominals

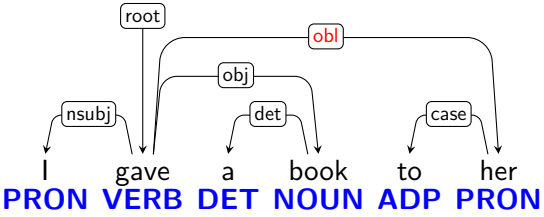
Nominal	Clausal	Modifier	Function
nmod	acl	amod	det
		nummod	case

Information Packaging

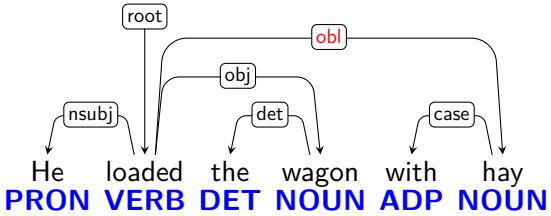


Information Packaging

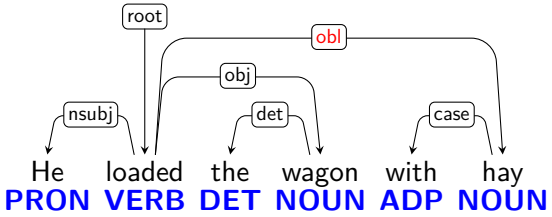




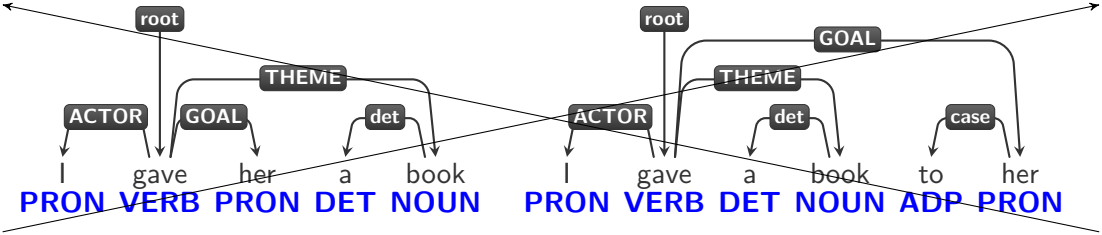
Information Packaging



Information Packaging



UD is NOT about Semantic Roles!



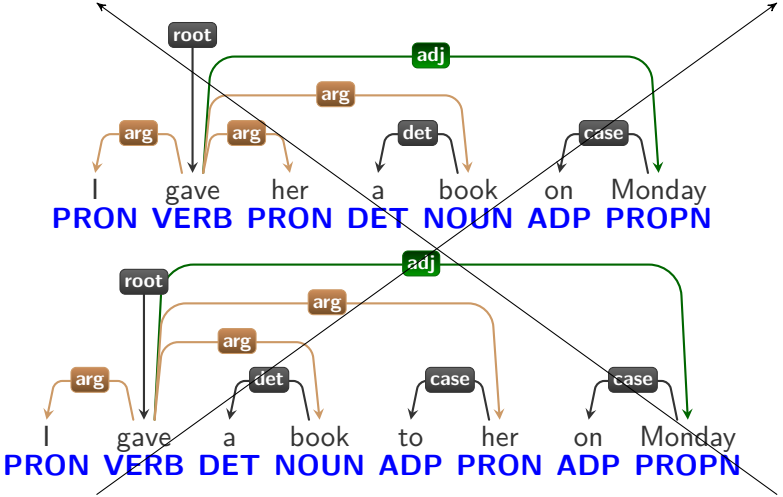
Manning's Law – What If We Do Semantic Roles?

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
  - $\Rightarrow$ 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
  - $\Rightarrow$ 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
    - $\Rightarrow$ 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

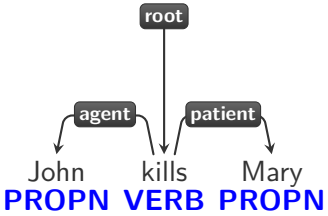
Language-specific Coding Strategy

- 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

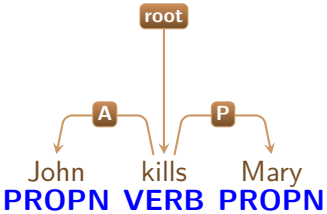
  - Then define:
    - function A ⇒ `nsubj`
    - function P ⇒ `obj`

Transitive Predicates in English

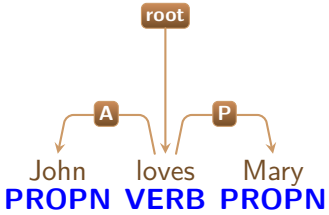


(primary transitive)

Transitive Predicates in English

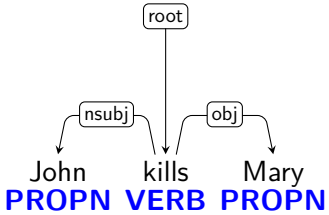


(primary transitive)

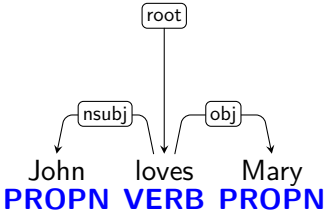


(generalized transitive)

Transitive Predicates in English

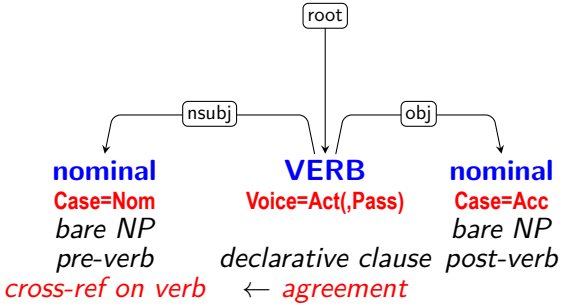


(primary transitive)

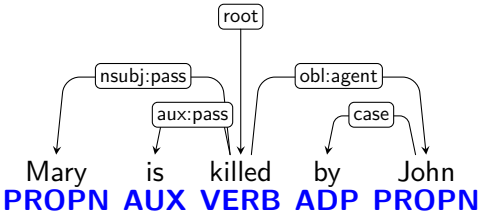


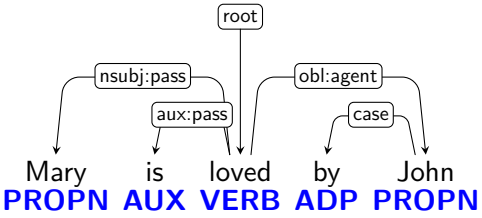
(generalized transitive)

Transitive Predicates in English

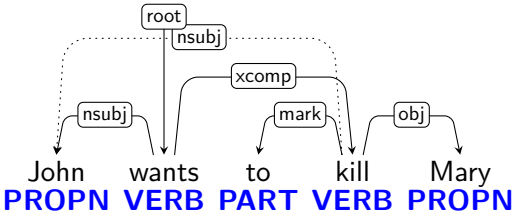


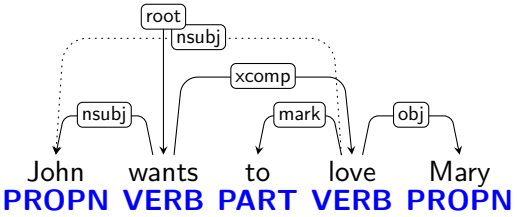
Passivization in English



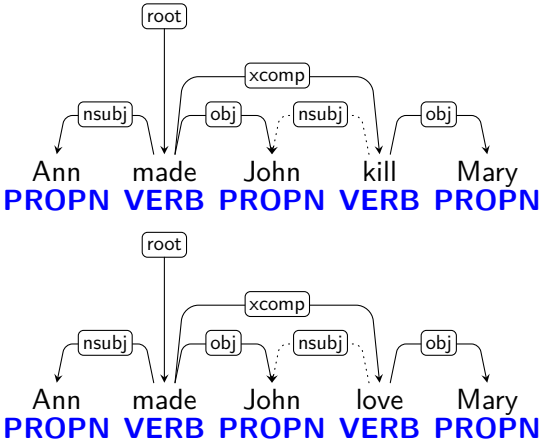


Subject Control in English



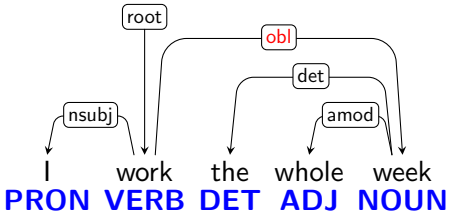


Object Control in English



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...

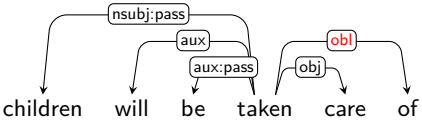
Some Problems

- 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.*



Bare Temporal Adjuncts: Any Other Criteria?

- *I work the whole week.*
- *I work every Friday.*

- English has a fixed word order; adjuncts are less fixed than objects:
  - *I work every Friday in Paris.*
  - *I work in Paris every Friday.*
  - *I spend every Friday in Paris.*
  - **I spend in Paris every Friday.*

- Unlike objects, adjuncts cannot be replaced by pronouns:
  - *Where do you spend this Friday? I spend it in Paris.*
  - *Where do you work this Friday? *I work it in Paris.*

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

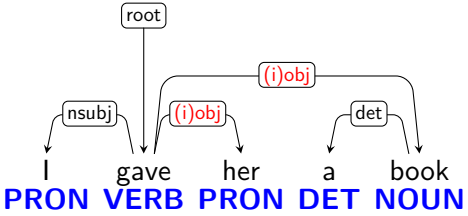
- Semantic roles needed when starting a new language
- Argument-adjunct might help with exceptions
  - Although we managed to explain *the whole week* without it

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

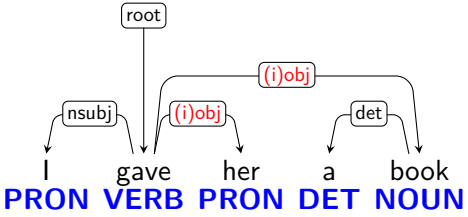
- Three core arguments
- Is one of them "least core"? ⇒ `iobj`
- (Alternatively, we could look at the semantic roles once again.)



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

Ditransitive Predicates

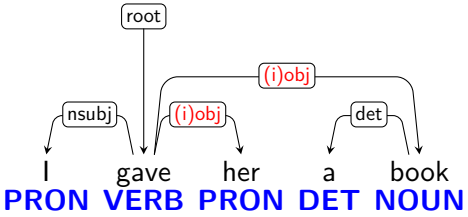
- Three core arguments
- Is one of them "least core"? ⇒ `iobj`
- (Alternatively, we could look at the semantic roles once again.)



- Fronting in questions:
  - *What did I give her?*
  - **Who did I give a book?*

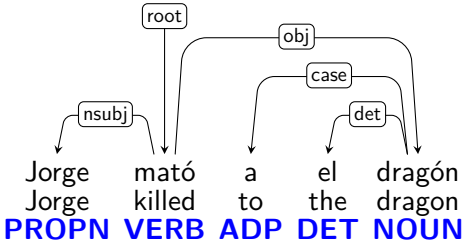
Ditransitive Predicates

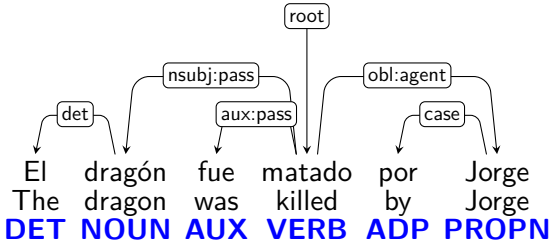
- Three core arguments
- Is one of them "least core"? ⇒ **iobj**
- (Alternatively, we could look at the semantic roles once again.)



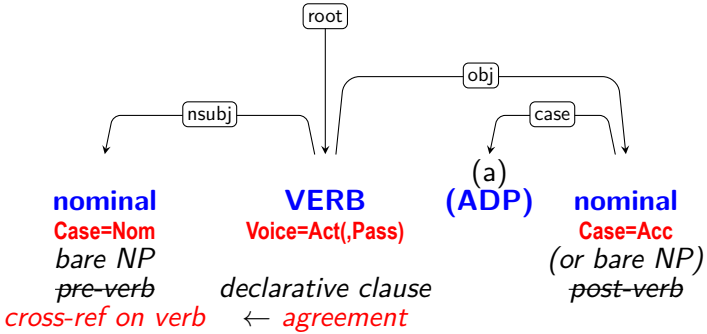
- Andrews (2007): *the status of the notion of 'indirect object' is problematic and difficult to sort out. The top priority is to work out what properties recipients and themes do and do not share with P arguments of primary transitive verbs.*

Spanish (EXTRA)

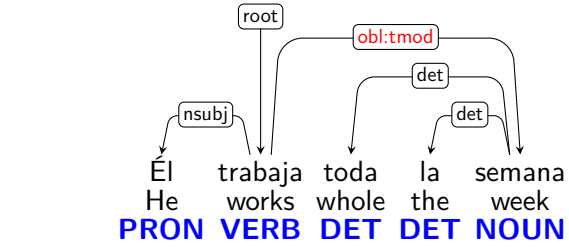


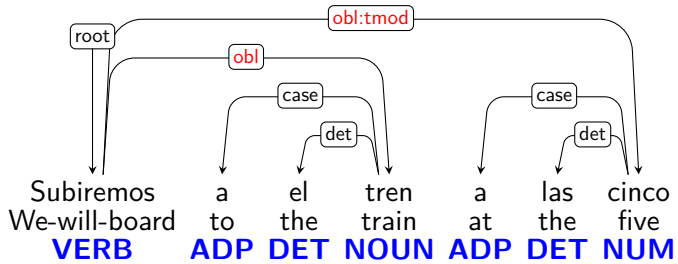


Spanish Transitive Clauses (EXTRA)

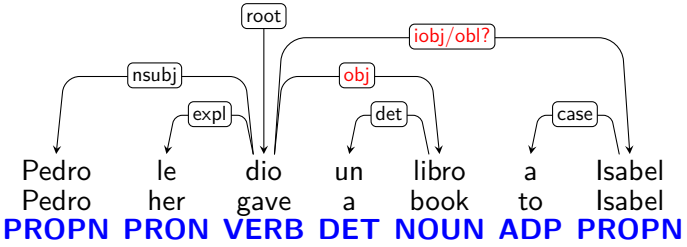


Spanish Adjunct Exceptions (EXTRA)

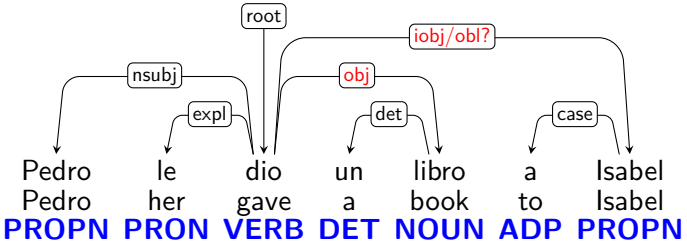




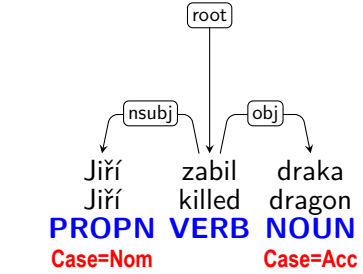
Spanish Ditransitive Clauses (EXTRA)

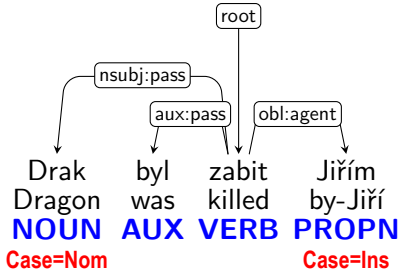


Spanish Ditransitive Clauses (EXTRA)

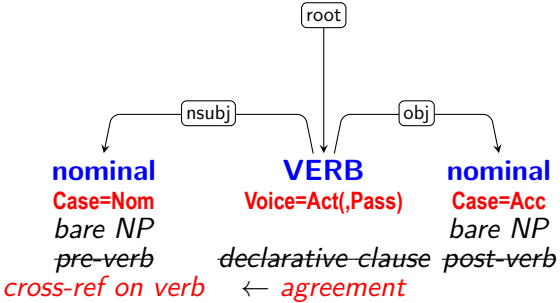


Czech (EXTRA)

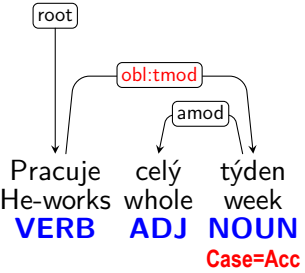

root
nsubj
obj
Jiří
Jiří
PROPN
Case=Nom
zabil
killed
VERB
draka
dragon
NOUN
Case=Acc


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

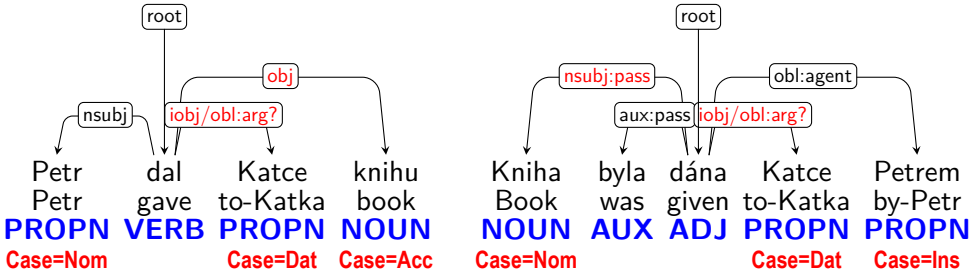
Czech Transitive Clauses (EXTRA)



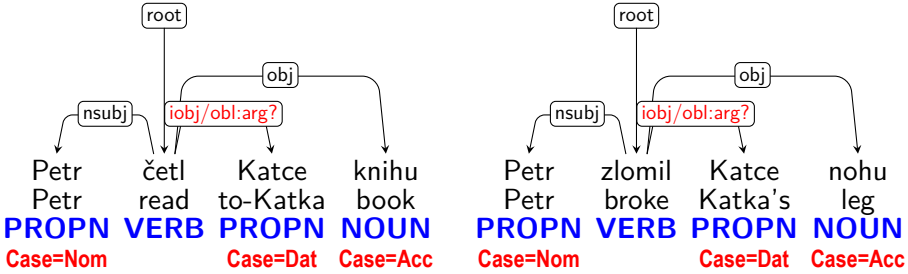
Czech Adjunct Exceptions (EXTRA)



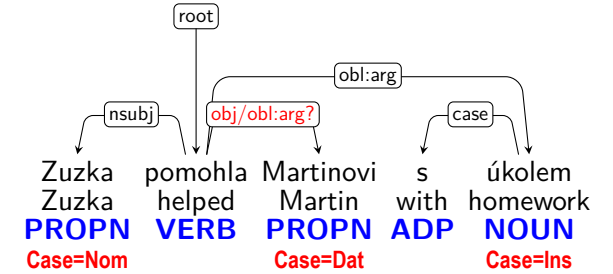
Czech Ditransitive Clauses (EXTRA)



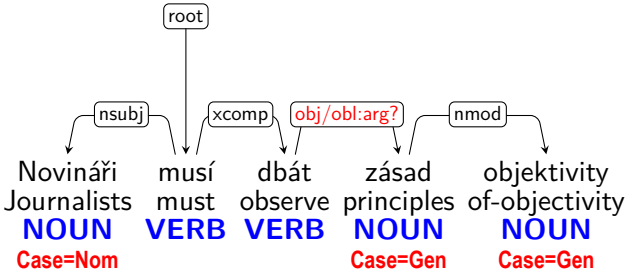
Dative: Recipient vs. Beneficiary (EXTRA)



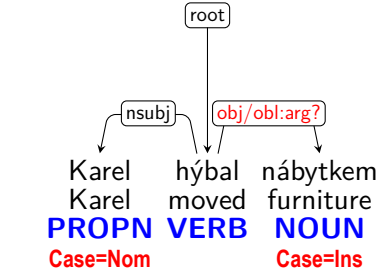
Monotransitive with Dative? (EXTRA)



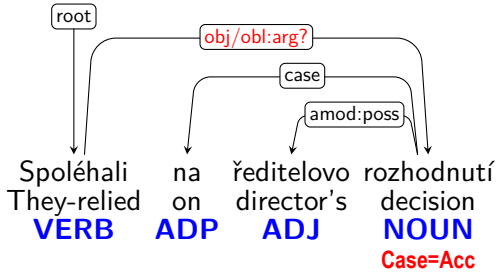
Monotransitive with Genitive? (EXTRA)

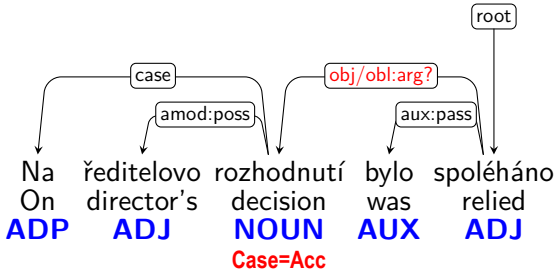


Monotransitive with Instrumental? (EXTRA)



Monotransitive with Preposition? (EXTRA)





Tentative Summary 2 (EXTRA)

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

Tentative Summary 2 (EXTRA)

- 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 (EXTRA)

- 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 2.1–2.5: bare NP > PP
  - Nom, Acc, Gen, Dat, Ins > ADP + adjuncts

Tentative Summary 2 (EXTRA)

- 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 2.1–2.5: bare NP > PP
  - Nom, Acc, Gen, Dat, Ins > ADP + adjuncts
- UD Czech 2.6 (May 2020):
  - Nom, Acc > Gen, Dat, Ins, ADP + adjuncts

Tentative Summary 2 (EXTRA)

- 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 2.1–2.5: bare NP > PP
  - Nom, Acc, Gen, Dat, Ins > ADP + adjuncts
- UD Czech 2.6 (May 2020):
  - Nom, Acc > Gen, Dat, Ins, ADP + adjuncts
  - ⇒ No ditransitives in Czech!
  - (Exception: *učit* "to teach" takes two Acc.)

Summary of Basic Syntax

- Universal Dependencies
  - Content words higher than function words ... better parallelism
  - Clauses – nominals – modifier words
    - Distinguished both as heads and as dependents
  - Core arguments vs. oblique dependents
  - Language-specific subtypes of relations

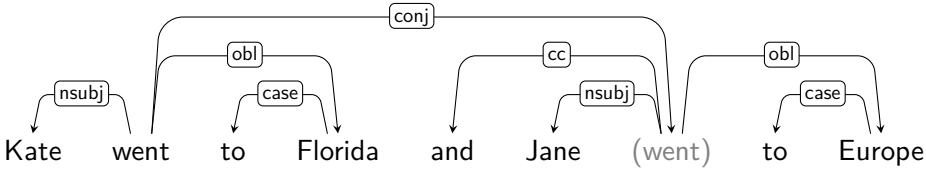
https://universaldependencies.org/

Enhanced Universal Dependencies

Outline

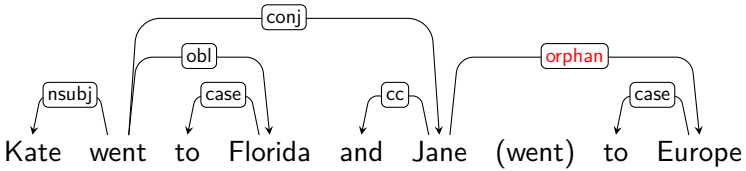
1. Introduction
2. Morphological Annotation in UD
3. Syntactic Annotation in UD
4. Core vs. Oblique
5. Enhanced Universal Dependencies
6. UD Tools

Deleted Predicates in Coordination

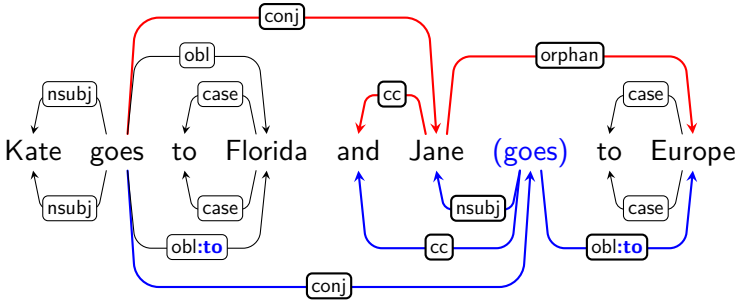


- Some treebanks would use an empty node to represent the second *went*.
- UD enhanced representation allows empty nodes!
- But the basic representation sticks with the overt words.

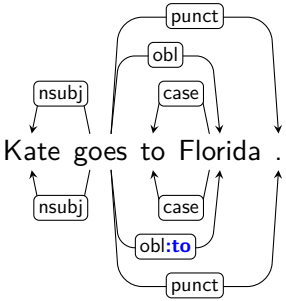
UD V2 Basic Dependencies: The `orphan` Relation



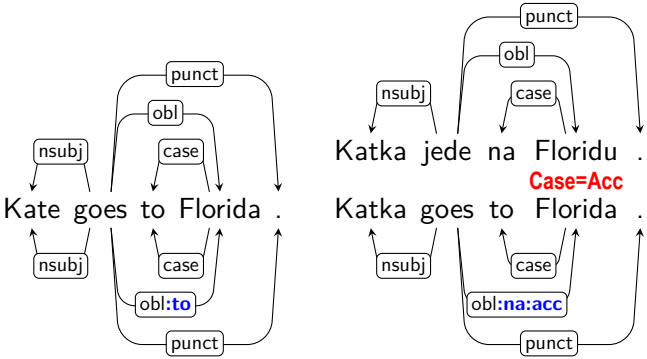
UD Enhanced Dependencies: Gapping and Stripping



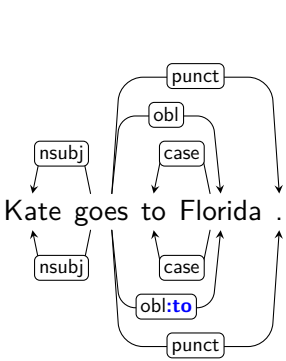
Enhanced UD: Case Information in Dependency Label

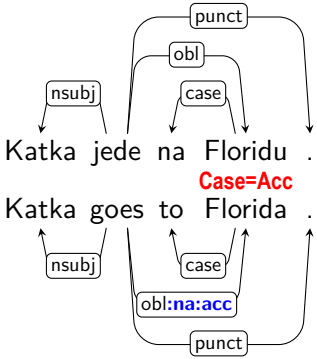


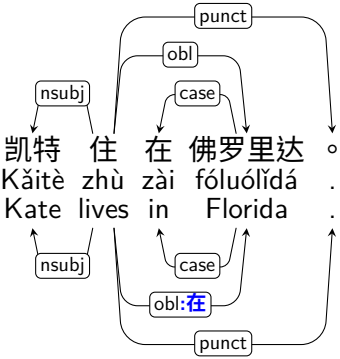
Enhanced UD: Case Information in Dependency Label



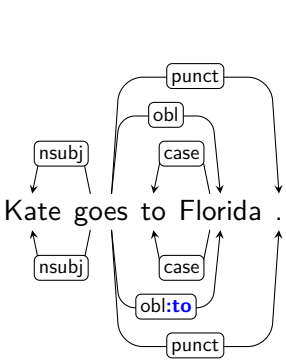
Enhanced UD: Case Information in Dependency Label

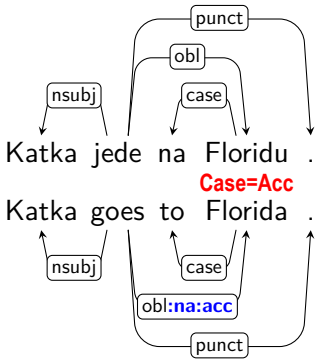


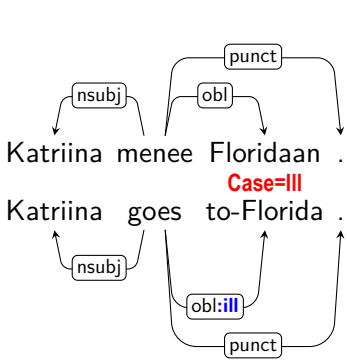




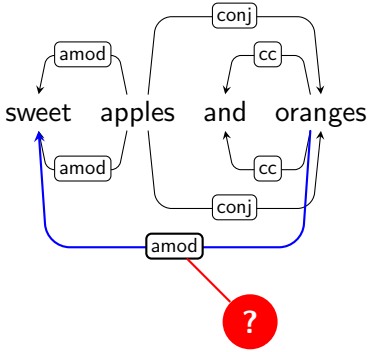
Enhanced UD: Case Information in Dependency Label



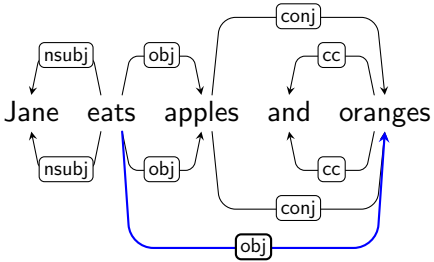




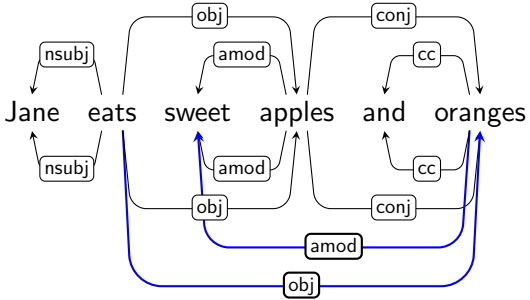
Enhanced UD: Shared Dependent of Coordination



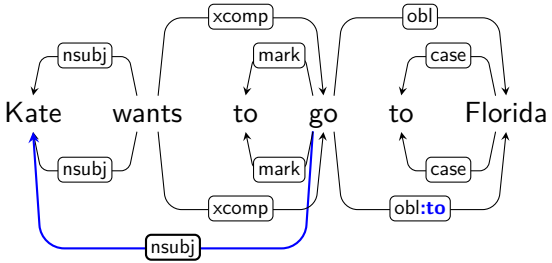
Enhanced UD: Parent of Coordination



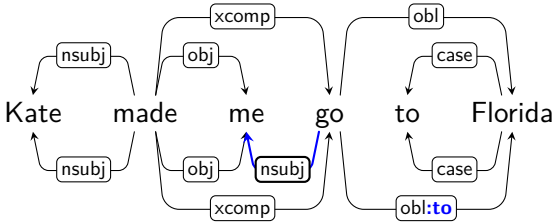
Enhanced UD: Coordination



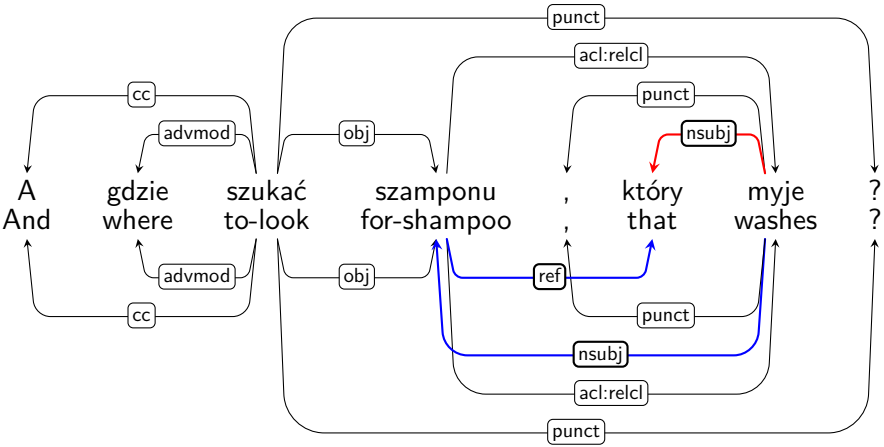
Enhanced UD: External Subject of Controlled Predicate



Enhanced UD: External Subject in Object-Control Construction

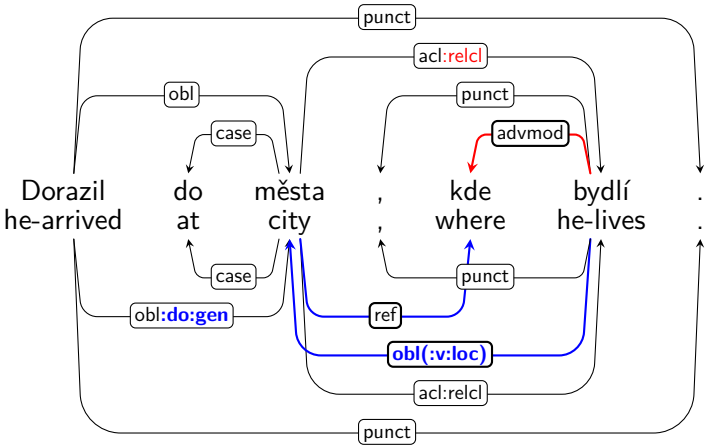


Enhanced UD: Relative Clauses

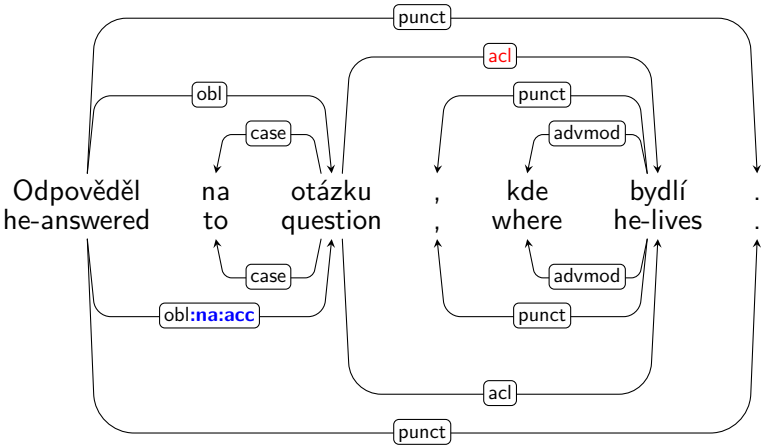


"And where to look for shampoo that works?"

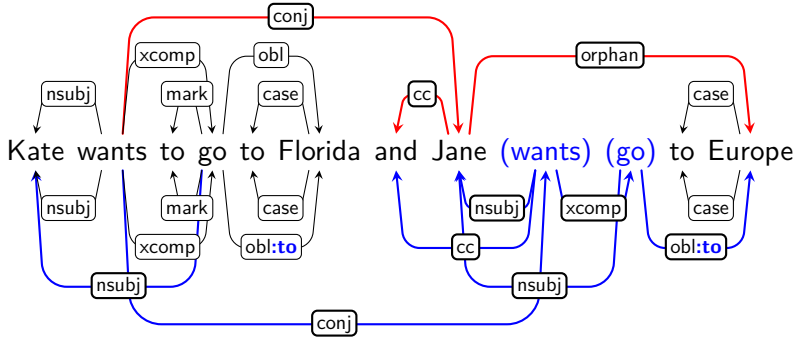
How to Recognize Relative Clauses (EXTRA)



How to Recognize Relative Clauses (EXTRA)



Enhanced UD: Gapping + Control Verb Construction



Enhanced UD: Six Enhancements (EXTRA)

- Null nodes for gapping (24 treebanks in UD 2.8)
- Dependency propagation in coordination
  - Common parent of coordination (28 treebanks)
  - Shared dependents of coordination (25 treebanks)
- External subjects of controlled predicates (21 treebanks)
- Cyclic dependencies to/from relative clauses (22 treebanks)
- Case-enhanced dependency labels (21 treebanks)

- All 6 types: 15 treebanks, 8 languages
- At least 1 type: 30 treebanks, 18 languages
- Only basic UD: 172 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

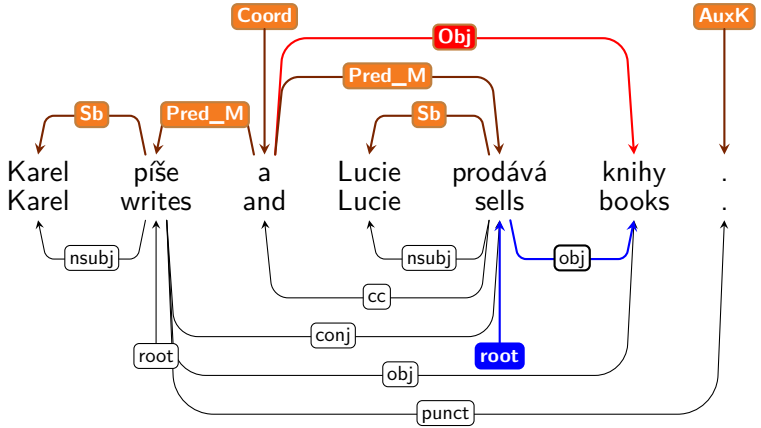
Stanford Enhancer

- 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

Conversion from non-UD Data: Extra Information? (EXTRA)

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



Summary of Enhanced Syntax

- Some deep-syntactic relations
- Directed graph (not always tree)
- Six enhancement types
  - Empty nodes for gapping
  - Shared parents in coordination
  - Shared dependents in coordination
  - External subjects in control/raising
  - Cyclic dependency in relative clauses
  - Case-enhanced modifier relations
- Some (not all) can be guessed from the basic tree

https://universaldependencies.org/

UD Tools

Outline

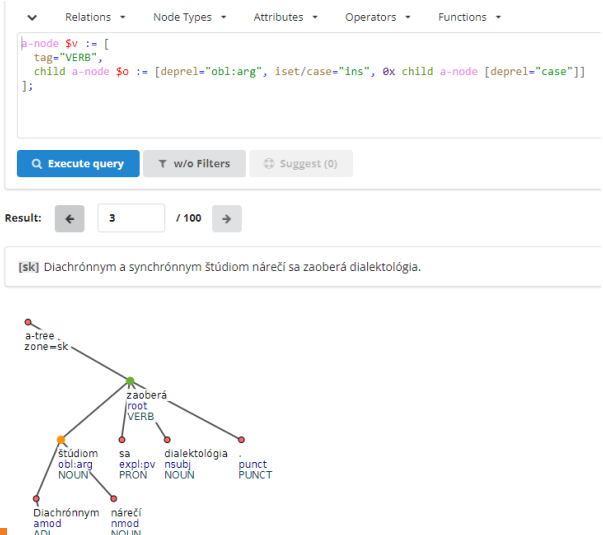
1. Introduction
2. Morphological Annotation in UD
3. Syntactic Annotation in UD
4. Core vs. Oblique
5. Enhanced Universal Dependencies
6. UD Tools

Off-the-Shelf Tools

- https://universaldependencies.org/tools.html

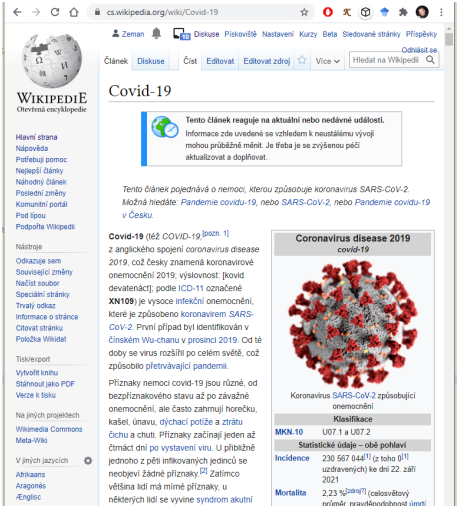
Linguists Can Search Treebanks

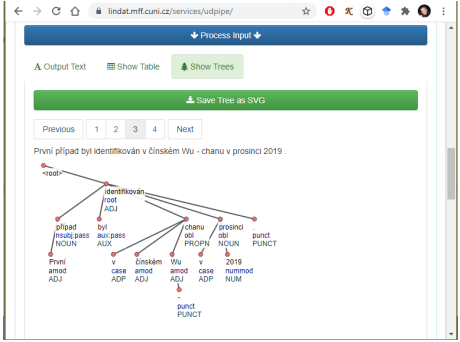
https://lindat.mff.cuni.cz/services/pmltq/



Linguists Can Parse and Search New Data

https://lindat.mff.cuni.cz/services/udpipe/



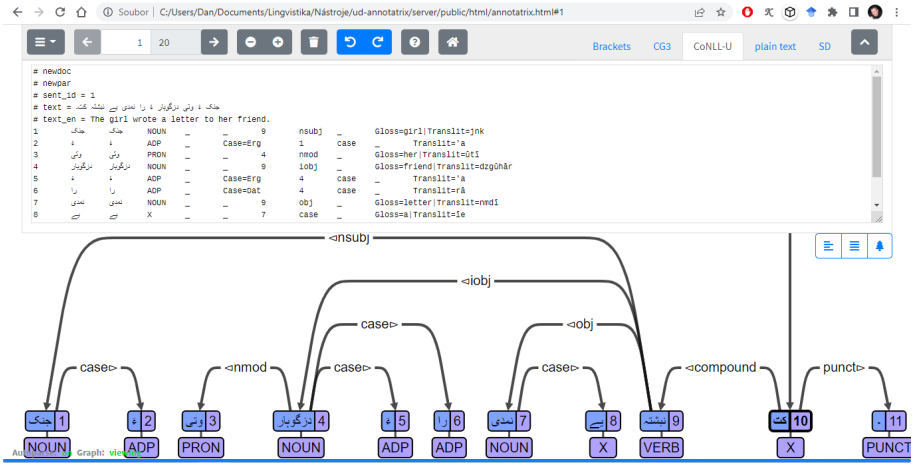


Parsers and Processing Pipelines

- UDPipe (`https://lindat.mff.cuni.cz/services/udpipe/`)
  - `https://ufal.mff.cuni.cz/udpipe`
- Stanza (`https://stanfordnlp.github.io/stanza/`)

Annotation

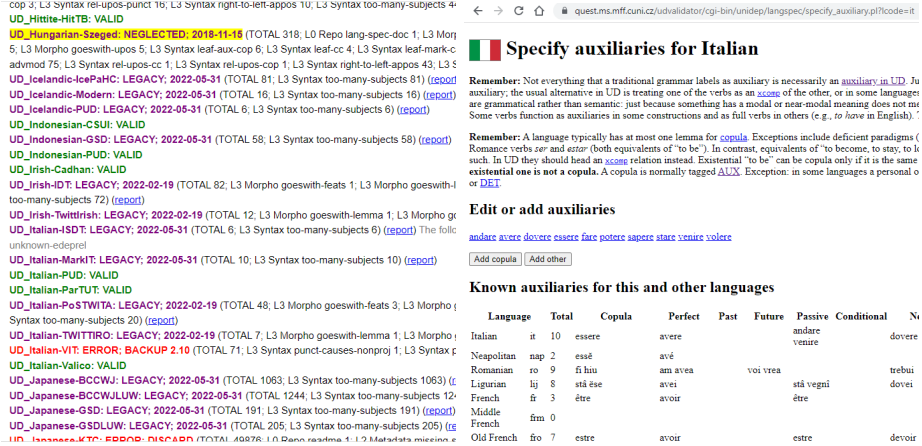
- (Tred) `https://ufal.mff.cuni.cz/tred/`
- UD Annotatrix `https://github.com/jonorthwash/ud-annotatrix`



- Morphology: Spreadsheet (Excel / LibreOffice Calc)

Validation and Releasing

- (Tred) `https://ufal.mff.cuni.cz/tred/`
- UD Annotatrix `https://github.com/jonorthwash/ud-annotatrix`



Documentation and Statistics

universaldependencies.org/cs/index.html

UD for Czech

Tokenization and Word Segmentation

- In general, words are delimited by whitespace characters. Description of exceptions follows.
- According to typographical rules, many punctuation marks are attached to a neighboring word. We always tokenize them as separate tokens (words); that holds even for hyphenated compounds such as *česko-slovenský* "Czech-Slovak" (three tokens) and for abbreviations such as *atd.* "etc." (two tokens).
- A whitespace separating digits in a large number is not treated as a word separator. For example, *1 000 000* ("1,000,000" by English rules) is one token.
- There are several closed classes of contractions that are treated as multi-word tokens and segmented to individual syntactic words. The most prominent type is a subordinating conjunction fused with a conditional auxiliary: *kdybych = když + bych* "if I". For more details, see tokenization.

Morphology

Tags

This is an overview only. For more detailed discussion and examples, see the list of Czech POS tags and Czech features.

- Czech uses all 17 universal POS categories, including particles (PART). At present, more than 70 word types are tagged PART. This is a legacy of an existing Czech morphological analyzer and many of these words should probably belong to another category in UD; however, the exact list has yet to be worked out.
- The pronoun (PRON) vs. determiner (DET) distinction is based on word lists because the traditional grammar does not define determiners. In general, words that inflect for gender, to be able to agree with a modified noun, are tagged DET, even if they act independently in a given sentence; that includes possessives. Pronominal quantifiers (which the traditional grammar includes in numerals) are DET as well.
- Czech has just one auxiliary verb (AUX), *být* ("to be"), but lemmas *bývat* and *bývávat* are also possible. They are in fact just variants of *být*, but they are separate lemmas because the morphological process that relates them to *být* is considered derivational. The auxiliary verb is used in

universaldependencies.org/treebanks/la-comparison.html

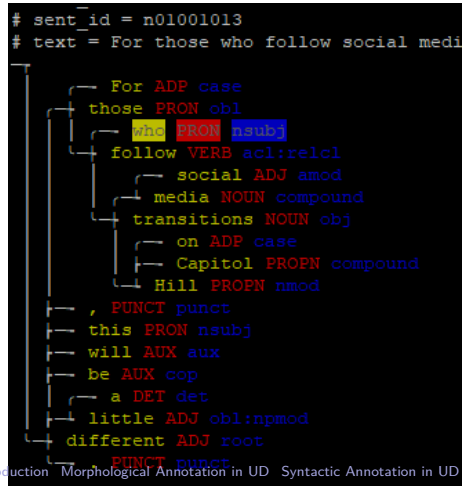
Pronouns, Determiners, Quantifiers	Pronouns, Determiners, Quantifiers	Pronouns, Determiners, Quantifiers	Pronouns, Determiners, Quantifiers	Pronouns, Determiners, Quantifiers
• PronType	• PronType	• PronType		• PronType
• Art ◦ DET: ly				
• Con ◦ ADV: solum, alias ◦ DET: alia, aliud, aliis, alio, alterius, aliam, alii, sola, aliorum, utrumque				• Con ◦ ADV: solum, alite ◦ DET: aliud, alia, al alii, aliorum, sola, alterum, solus, sc ◦ SCONJ: utrum
• Dem ◦ ADV: ecce, ita, tam, idem ◦ DET: hoc, ipsum, haec, his, illud, ipsa, ipse, ipsius, ipso, idem ◦ PRON: quod	• Dem ◦ ADV: taliter, tantum, hic, ita, tam, ecce, tantu, hecce, hinc, ic ◦ DET: ipsa, suprascripta, ipsius, hanc, ipse, illa, ipso, suprascripte, suprascripto, hec			• Dem ◦ ADV: sic, tamen, t tantum, tunc, ita, hic, tanto, pretere ◦ DET: hoc, illa, illuc ipsum, huius, hiis ipse, ille ◦ PART: itaque, Ecc

universaldependencies.org/treebanks/cu_proiel/cu_proiel-feat-Case.html

Paradigm *и*	Nom	Acc	Dat	Gen	Loc	Ins
Gender=Masc,Neut|Number=Sing|Person=2			*немоу*			
Gender=Masc,Neut|Number=Sing|Person=3			*емоу*		*немь*	*нимь*
Gender=Masc|Number=Sing|Person=3		*и, і, нь, и҆, нъ, й, ı*	*емоу, немоу, моу, е҆моу, немѫ*	*его, него*	*немь, емь, немъ, емъ*	*нимь, нимъ, имъ, имь*
Gender=Masc|Number=Dual|Person=3		*ѣ*	*има, нима*	*ею, ею҆, нею, неѭ, неѭ҆*		*нима, има*
Gender=Masc|Number=Plur|Person=3		*ѩ, ѩ҆, нѧ*	*имъ, нимъ, нимь, имь, и҆мъ, ımъ*	*ихъ, нихъ, ихь, и҆хъ, ıхъ, і҆хъ*	*нихъ, ихъ*	*ними, ими, ни*
Gender=Fem|Number=Sing|Person=3		*ѭ, ѭ҆, нѭ, ѭ҆, нѫѭ*	*еи, неи, еи҆, е҆и҆*	*еѩ, еѩ҆, неѩ, ѩ*	*неи*	*неѭ, еѭ, неѭ҆, еѭ҆*
Gender=Fem|Number=Dual|Person=3		*и*	*има*	*ею, нею*		
Gender=Fem|Number=Plur|Person=3		*ѩ, нѧ*	*имъ, нимъ*	*ихъ, нихъ*	*нихъ*	*ними*
Gender=Neut|Number=Sing|Person=3		*е, не, е҆*		*его, него*	*немь*	*нимь*
Gender=Neut|Number=Dual|Person=3				*ею*		
Gender=Neut|Number=Plur|Person=3	ѣ	ѣ		*ихъ*		

Udapi

- http://udapi.github.io/
- https://ufal.mff.cuni.cz/~zeman/vyuka/deptreebanks/NPFL075-working-with-UD.pdf



Summary of Tools

- PML-TQ (tree query)
- UDPipe, Stanza (parsers)
- Annotatrix, Tred (annotation + tree view)
- UD web infrastructure (documentation, validation, statistics)
- Udapi (searching, batch processing, visualization)
- Many other tools listed on the UD website

https://universaldependencies.org/

Thanks!
Grazie!

https://universaldependencies.org/