Universal Dependencies

Daniel Zeman

March 23, 2023
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
A cat chases rats and mice

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

En katt jagar råttor och möss
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
http://universaldependencies.org

- Part-of-speech tags
- Morphological features
- Syntactic dependencies
• 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.
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.
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.
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.
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.
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.
Design Principles

- **Dependency**
  - Widely used in practical NLP systems
  - Available in treebanks for many languages

- **Lexicalism**
  - Basic annotation units are words – syntactic words
  - Words have morphological properties
  - Words enter into syntactic relations

- **Recoverability**
  - Transparent mapping from input text to word segmentation
• 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
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
• 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

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

- Revised and extended version of CoNLL-X format
- Two-level segmentation and enhanced dependencies
Where are we today?

• Brief history of UD:
  • First guidelines launched in October 2014
  • Treebank releases (roughly) every six months
  • Version 2 in December 2016 (guidelines) and March 2017 (treebanks)
  • New system of guidelines amendments in May 2022

• 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
  • Five UD workshops so far; next at Syntaxfest 2023, Washington
  • Next release in November 2022 (v2.11)
Basic Universal Dependencies: 138 (136) 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: Wolof, Yoruba • Other: Basque, Sw. Sign, Naija
Morphological Annotation in UD
Morphological Annotation in UD

- Tokenization / word segmentation
- Lemmatization (LEMMA)
- Universal part-of-speech tags (UPOS)
- Universal features (FEATS)
- Language-specific features
“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)
• 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.”
He abdicated in favour of his son Baudouin.

yatanāzalu ʿan al-ʿarši li+ibni+hi būdūān
surrendered on the throne to son his Baudouin
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。
Xiàn zài wǒ men zài wǎ lún xī yǎ.
Now we in Valencia.

ADV PRON ADP PROPN PUNCT
I went to the beauty salon of Kyōdō [Beyond-R.]
I went to the beauty salon of Kyōdō [Beyond-R.]

経堂の美容室にきました
Kyōdō no miyōshitsu ni itte kimashita

Kyōdō of beauty-salon to going come
VerbForm=Conv VerbForm=Fin Tense=Past Polite=Form
I went to the beauty salon of Kyōdō [Beyond-R.]

経堂の

Kyōdōno

美容室に

miyōshitsuni

行く

itte

きますた

kimashita

来る

come

PROPN

Case=Gen

NOUN

Case=Dat

VERB

VerbForm=Conv

VerbForm=Fin

Tense=Past

Polite=Form
All the concrete country roads are the result of...

- Spaces delimit monosyllabic morphemes, not words.
- Multiple syllables without space occur in loanwords (bêtông).
- Spaces are allowed to occur word-internally in Vietnamese UD.
Il touche environ 100 000 sesterces par an.
One syntactic word spans several orthographic words?

*I am still very satisfied.*
• When to split?
  • Only part of the token involved in a relation to something outside the token? Split!
• When to split?
  • Only part of the token involved in a relation to something outside the token? Split!
  • Hard time finding POS tag? Split!
• 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.
• 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

• 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.

<table>
<thead>
<tr>
<th>ID</th>
<th>FORM</th>
<th>LEMMA</th>
<th>UPOS</th>
<th>HEAD</th>
<th>_</th>
<th>MISC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-2</td>
<td>Vámonos</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>1</td>
<td>Vamos</td>
<td>ir</td>
<td>VERB</td>
<td>0</td>
<td>root</td>
<td>_</td>
</tr>
<tr>
<td>2</td>
<td>nos</td>
<td>nosotros</td>
<td>PRON</td>
<td>1</td>
<td>obj</td>
<td>_</td>
</tr>
<tr>
<td>3-4</td>
<td>al</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>3</td>
<td>a</td>
<td>a</td>
<td>ADP</td>
<td>5</td>
<td>case</td>
<td>_</td>
</tr>
<tr>
<td>4</td>
<td>el</td>
<td>el</td>
<td>DET</td>
<td>5</td>
<td>det</td>
<td>_</td>
</tr>
<tr>
<td>5</td>
<td>mar</td>
<td>mar</td>
<td>NOUN</td>
<td>1</td>
<td>obl</td>
<td>_</td>
</tr>
<tr>
<td>6</td>
<td>.</td>
<td>.</td>
<td>PUNCT</td>
<td>1</td>
<td>punct</td>
<td>_</td>
</tr>
</tbody>
</table>
### text = Vámonos al mar.

### text_en = Let’s go to the sea.

<table>
<thead>
<tr>
<th>ID</th>
<th>FORM</th>
<th>LEMMA</th>
<th>UPOS</th>
<th>HEAD</th>
<th>MISC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-2</td>
<td>Vámonos</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>1</td>
<td>Vamos</td>
<td>_</td>
<td>_</td>
<td>0</td>
<td>root</td>
</tr>
<tr>
<td>2</td>
<td>nos</td>
<td>_</td>
<td>_</td>
<td>1</td>
<td>obj</td>
</tr>
<tr>
<td>3-4</td>
<td>al</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>3</td>
<td>a</td>
<td>a</td>
<td>ADP</td>
<td>5</td>
<td>case</td>
</tr>
<tr>
<td>4</td>
<td>el</td>
<td>el</td>
<td>DET</td>
<td>5</td>
<td>det</td>
</tr>
<tr>
<td>5-6</td>
<td>mar.</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>5</td>
<td>mar</td>
<td>mar</td>
<td>NOUN</td>
<td>1</td>
<td>obl</td>
</tr>
<tr>
<td>6</td>
<td>.</td>
<td>.</td>
<td>PUNCT</td>
<td>1</td>
<td>punct</td>
</tr>
</tbody>
</table>
Tokenization vs. Multi-word Tokens

- 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

- Punctuation involved? Low level!
  - Exceptions: Spanish-Catalan parallelism.
Tokenization vs. Multi-word Tokens Summary

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

- Boundary between two letters? Typically high level.
  - Exceptions: Chinese, Japanese.
Tokenization vs. Multi-word Tokens Summary

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

• Boundary between two letters? Typically high level.
  • Exceptions: Chinese, Japanese.

• Non-concatenative? High level!
• Basic or citation form (⇒ 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)
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 stáne 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


<table>
<thead>
<tr>
<th>Open</th>
<th>Closed</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOUN</td>
<td>common noun</td>
<td>PRON</td>
</tr>
<tr>
<td>PROPN</td>
<td>proper noun</td>
<td>DET</td>
</tr>
<tr>
<td>VERB</td>
<td>verb</td>
<td>AUX</td>
</tr>
<tr>
<td>ADJ</td>
<td>adjective</td>
<td>NUM</td>
</tr>
<tr>
<td>ADV</td>
<td>adverb</td>
<td>ADP</td>
</tr>
<tr>
<td>INTJ</td>
<td>interjection</td>
<td>SCONJ</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CCONJ</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PART</td>
</tr>
</tbody>
</table>

- 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

• 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** (*druh zájmena*)
- **NumType** (*druh číslovky*)
- **Poss** (*přivlastňovací*)
- **Reflex** (*zvratné*)
- **Foreign** (*cizí slovo*)
- **Abbr** (*zkratka*)
- **Typo** (*překlep*)
- **Gender** (*rod*)
- **Animacy** (*životnost*)
- **NounClass** (*jmenná třída*)
- **Number** (*číslo*)
- **Case** (*pád*)
- **Definite(ness)** (*určitost*)
- **Degree** (*stupeň*)
- **VerbForm** (*slovesný tvar*)
- **Mood** (*způsob*)
- **Tense** (*čas*)
- **Aspect** (*vid*)
- **Voice** (*slovesný rod*)
- **Evident(iality)** (*zjevnost*)
- **Polarity** (*zápor*)
- **Person** (*osoba*)
- **Polite(ness)** (*zdvořilost*)
- **Clusivity** (*kluzivita*)
<table>
<thead>
<tr>
<th>Features</th>
<th>Inflectional (&quot;Nominal&quot;)</th>
<th>Inflectional (&quot;Verbal, Pronominal&quot;)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PronType</td>
<td>Gender</td>
<td>VerbForm</td>
</tr>
<tr>
<td>NumType</td>
<td>Animacy</td>
<td>Mood</td>
</tr>
<tr>
<td>Poss</td>
<td>NounClass</td>
<td>Tense</td>
</tr>
<tr>
<td>Reflect</td>
<td>Number</td>
<td>Aspect</td>
</tr>
<tr>
<td>Foreign</td>
<td>Case</td>
<td>Voice</td>
</tr>
<tr>
<td></td>
<td>Definite</td>
<td>Evident</td>
</tr>
<tr>
<td></td>
<td>Degree</td>
<td>Polarity</td>
</tr>
<tr>
<td>Abbr</td>
<td></td>
<td>Person</td>
</tr>
<tr>
<td>Typo</td>
<td></td>
<td>Polite</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Clusivity</td>
</tr>
</tbody>
</table>

- 24 features, each with a number of possible values
- Languages select relevant features
- May add language-specific features or values
Three types of infinitives in Finnish:

Example: *olla* “to be”

<table>
<thead>
<tr>
<th></th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
</tr>
</thead>
<tbody>
<tr>
<td>olla</td>
<td>ollena</td>
<td>olemassa</td>
<td>olematta</td>
</tr>
<tr>
<td>olle</td>
<td>olleena</td>
<td>olemaan</td>
<td>olemaan</td>
</tr>
<tr>
<td>ollan</td>
<td>olleena</td>
<td>olemasta</td>
<td>olemaan</td>
</tr>
<tr>
<td>olla</td>
<td>ollena</td>
<td>olematta</td>
<td>olemaan</td>
</tr>
</tbody>
</table>

Universal Dependencies
Joku yrittää piristää itsään värjämällä hiuksensa
Someone tries to uplift oneself by-staining their-hair
Someone tries to uplift oneself by staining their hair.

Universal Dependencies

Morphological Annotation in UD
Czech adjectives agree with nouns in gender.

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

velká  big  sestra  sister
ADJ   NOUN
Gender=Fem  Gender=Fem
### Possessive adjectives: agreement gender vs. lexical gender

<table>
<thead>
<tr>
<th></th>
<th>Gender</th>
<th>Gender[psor]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>otcův</strong></td>
<td>Masc</td>
<td>Masc</td>
</tr>
<tr>
<td><strong>bratr</strong></td>
<td>Masc</td>
<td>Masc</td>
</tr>
<tr>
<td><strong>otcova</strong></td>
<td>Fem</td>
<td>Masc</td>
</tr>
<tr>
<td><strong>sestra</strong></td>
<td>Fem</td>
<td>Fem</td>
</tr>
<tr>
<td><strong>matčin</strong></td>
<td>Masc</td>
<td>Fem</td>
</tr>
<tr>
<td><strong>matčina</strong></td>
<td>Fem</td>
<td>Fem</td>
</tr>
<tr>
<td><strong>bratr</strong></td>
<td>Masc</td>
<td>Fem</td>
</tr>
<tr>
<td><strong>sestra</strong></td>
<td>Fem</td>
<td>Fem</td>
</tr>
</tbody>
</table>
Multi-valued Features (Disjunction / Parallel Application)

- 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)

• Currently used in Turkish (language-specific values)

• Two or more morphemes in chain, affecting the same feature

• Example: **Voice=CauPass** (causative + passive $\Rightarrow$ someone is caused to do something)
  - *yanıll* “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)

- Currently used in Turkish (language-specific values)

- Two or more morphemes in chain, affecting the same feature

- Example: **Voice=CauPass** (causative + passive $\Rightarrow$ 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"
Future tense in Spanish and German: no \textit{Tense=Fut} in German!

\begin{tabular}{llll}
\textbf{Dormirá} & \textbf{Er} & \textbf{wird} & \textbf{schlafen} \\
\textbf{He-will-sleep} & \textbf{He} & \textbf{will} & \textbf{sleep} \\
\textbf{VERB} & \textbf{PRON} & \textbf{AUX} & \textbf{VERB} \\
\text{VerbForm=Fin} & \text{PronType=Prs} & \text{VerbForm=Fin} & \text{VerbForm=Inf} \\
\text{Mood=Ind} & \text{Number=Sing} & \text{Mood=Ind} & \\
\text{Tense=Fut} & \text{Person=3} & \text{Tense=Pres} & \\
\text{Number=Sing} & \text{Gender=Masc} & \text{Number=Sing} & \\
\text{Person=3} & \text{Case=Nom} & \text{Person=3} & \\
\end{tabular}
### Participle Types

| некурящий | человек | начавшийся | разговор |
| nekurjaščij | čelovek | načavšijsja | razgovor |
| non-smoking | person | that-has-started | conversation |

- **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 **ParticT ype=Past**
  - Reasonably close to the grammatical meaning
Conflicting Traditional Terminologies

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

- 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**)

Universal Dependencies
Conflicting Traditional Terminologies

Ako so leta 1942 vračali ...

And as they-were returning ...

VerbForm=Fin
VerbForm=Part
Tense=Past?
And as they were returning in-year 1942, they would not come in Athens.

Future they will drive just wood.

Past? Past?
Conflicting Traditional Terminologies

And as they were returning in-year 1942, they would not come in Athens this future year, they will not drive just wood...
Conflicting Traditional Terminologies

- 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

- 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: наруши́вший / нару́шивши́й
  - bg “participle + past (aorist) / imperfect” (two subtypes)
  - cu “participle + resultative aspect” (lang-spec)
Conflicting Traditional Terminologies

• West/South Slavic: \texttt{VerbForm=Part}
• Russian: \texttt{VerbForm=Fin} (past tense)
  • \texttt{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: \texttt{нарушивший} / \texttt{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.
Universal Dependencies

**Summary**

- 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
  - Layered features
  - Multi-valued features
- Lemmas, tags, and features apply to words (tree nodes), not to multi-word expressions and not to sub-word units (morphemes)
- Categories are comparable (but not identical) across languages

[https://ufal.cz/courses/nplfl075](https://ufal.cz/courses/nplfl075)