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

Universal Dependencies
Morphological Annotation in UD

3/47
A cat chases rats and mice

En katt jagar råttor och möss

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
Valencia
PROPN
Number=Sing
Mood=Ind
Number=Sing
Person=3
Tense=Pres
VerbForm=Fin

da
VERB
Definite=Def
Gender=Fem
Number=Sing

la
DET

bienvenida
NOUN
Gender=Fem
Number=Sing

a
ADP

los
DET

participantes
NOUN
Definite=Def
Gender=Masc
Number=Plur

de
ADP

EACL
PROPN
Number=Sing

• 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.

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.
• 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
### Morphological Annotation

<table>
<thead>
<tr>
<th></th>
<th>le</th>
<th>chat</th>
<th>chasse</th>
<th>les</th>
<th>chiens</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DET</strong></td>
<td><strong>NOUN</strong></td>
<td><strong>VERB</strong></td>
<td><strong>DET</strong></td>
<td><strong>NOUN</strong></td>
<td><strong>PUNCT</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Definite=Def</strong></td>
<td><strong>Gender=Masc</strong></td>
<td><strong>Gender=Masc</strong></td>
<td><strong>Definite=Def</strong></td>
<td><strong>Gender=Masc</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Number=Sing</strong></td>
<td><strong>Mood=Ind</strong></td>
<td><strong>Number=Sing</strong></td>
<td><strong>Number=Plur</strong></td>
<td><strong>Number=Plur</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Person=3</strong></td>
<td><strong>Tense=Pres</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>VerbForm=Fin</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 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

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
• Content words are related by dependency relations
• Function words attach to the content word they modify
• Punctuation attach to head of phrase or clause
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:
  - 148 languages
  - 259 treebanks
  - 577 contributors
  - 150,000+ downloads

- Past and current UD events:
  - 4 CoNLL and IWPT shared tasks on UD parsing
  - Six UD workshops so far; next at LREC-COLING 2024, Torino
  - Next release in May 2024 (v2.14)
Basic Universal Dependencies: 148 (145) Languages and Growing

- I.-E.: Armenian (+West +Class.), Greek (+Ancient), Albanian, Gheg, Hittite, Breton, Irish (+Old), Manx, Scottish, Welsh, Afrikaans, Danish, Dutch, English, Faroese, Frisian, German, Gothic, Icelandic, Low Saxon, Norwegian, Swedish, Swiss German, Catalan, French (+Old +Mid.), Galician, Italian, Latin, Ligurian, Neapolitan, Portuguese, Romanian, Spanish, Umbrian, Belarusian, Bulgarian, Church Slavonic, Croatian, Czech, Macedonian, Polish, Pomak, Russian (+Old), Serbian, Slovak, Slovenian, Ukrainian, Upper Sorbian, Latvian, Lithuanian, Kurmanji, Persian, Khunsari, Nayini, Soi, Urdu, Hindi, Kangri, Bhojpuri, Bengali, Marathi, Sinhala, Sanskrit
- Dravidian: Malayalam, Tamil, Telugu
- Uralic: Erzya, Estonian, Finnish, Hungarian, Karelian, Livvi, Komi Permyak+Zyrian, Moksha, Sámi North+Skolt, Veps
- Turkic: Kazakh, Kyrgyz, Old Turkish, Tatar, Turkish, Uyghur, Yakut
- Buryat
- Xibe
- Korean
- Japanese
- Sino-T.: Cantonese, Chinese (+Class.)
- Tai-K.: Thai
- Aus.-As.: Vietnamese
- Austron.: Indonesian, Javanese, Tagalog, Cebuano
- Pama-Nyu.: Warlpiri
- Chu.-Kam.: Chukchi
- Esk.-Al.: Yupik
- U.-Az.: Nahuatl
- West+High
- Mayan: Kiche
- Arawakan: Apurinã
- Arawan: Madi
- Tupian: Akuntsu, Guajajara, Kaapor, Karo, Makurap, Mundurukú, Nheengatu, Tupinambá, Mbyá, Guarani, Teko
- M.-Je: Xavante, Bororo
- Af.-As.: Akkadian, Amharic, Arabic Levantine, Assyrian, Beja, Coptic, Hebrew (+Ancient), Maltese
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
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 \textit{nos} vamos mañana?
    • \textit{Nos} despertamos a las cinco.
      “We wake up at five.”
    • \textit{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。

We are now in Valencia.

現在我們在瓦倫西亞。

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.]

经堂の美容室に行った
Kyōdō no miyōshitsu ni itte kimashita

经堂 of beauty-salon to going come
VerbForm=Conv VerbForm=Fin Tense=Past Polite=Form
I went to the beauty salon of Kyōdō [Beyond-R.]
All the concrete country roads are the result of...

- Tất cả đường bê tông nội đồng là thành quả ...

- **PRON NOUN NOUN NOUN AUX NOUN PUNCT**

- Spaces delimit monosyllabic morphemes, not words.
- Multiple syllables without space occur in loanwords (*bê tông*).
- 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!
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!
• 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

• 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)

Universal Dependencies
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>

SpaceAfter=No
<table>
<thead>
<tr>
<th>ID</th>
<th>FORM</th>
<th>LEMMA</th>
<th>UPOS</th>
<th>...</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>
<td>_</td>
</tr>
<tr>
<td>1</td>
<td>Vamos</td>
<td>ir</td>
<td>VERB</td>
<td>...</td>
<td>0</td>
<td>root</td>
</tr>
<tr>
<td>2</td>
<td>nos</td>
<td>nosotros</td>
<td>PRON</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>
<td>_</td>
</tr>
<tr>
<td>3</td>
<td>a</td>
<td>a</td>
<td>ADP</td>
<td>...</td>
<td>5</td>
<td>case</td>
</tr>
<tr>
<td>4</td>
<td>el</td>
<td>el</td>
<td>DET</td>
<td>...</td>
<td>5</td>
<td>det</td>
</tr>
<tr>
<td>5-6</td>
<td>mar.</td>
<td>_</td>
<td>_</td>
<td>...</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>5</td>
<td>mar</td>
<td>mar</td>
<td>NOUN</td>
<td>...</td>
<td>1</td>
<td>obl</td>
</tr>
<tr>
<td>6</td>
<td>.</td>
<td>.</td>
<td>PUNCT</td>
<td>...</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
• 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!
Lemmas

• 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></th>
<th>Open</th>
<th>Closed</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOUN</td>
<td>common noun</td>
<td>PRON pronoun</td>
<td>PUNCT punctuation</td>
</tr>
<tr>
<td>PROPN</td>
<td>proper noun</td>
<td>DET determiner</td>
<td>SYM symbol</td>
</tr>
<tr>
<td>VERB</td>
<td>verb</td>
<td>AUX auxiliary</td>
<td>X unknown</td>
</tr>
<tr>
<td>ADJ</td>
<td>adjective</td>
<td>NUM numeral</td>
<td></td>
</tr>
<tr>
<td>ADV</td>
<td>adverb</td>
<td>ADP adposition</td>
<td></td>
</tr>
<tr>
<td>INTJ</td>
<td>interjection</td>
<td>SCONJ subordinator</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CCONJ coordinator</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>PART particle</td>
<td></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 Dependencies
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*)
- **Deixis** (*vzdálenost*)
- **DeixisRef** (*referenční bod*)
### Features

<table>
<thead>
<tr>
<th>Lexical</th>
<th>Inflectional (“Nominal”)</th>
<th>Inflectional (“Verbal, Pronominal”)</th>
</tr>
</thead>
<tbody>
<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>Abbr</td>
<td>DeixisRef</td>
<td>Polarity</td>
</tr>
<tr>
<td>Typo</td>
<td>Degree</td>
<td>Person</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Polite</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Clusivity</td>
</tr>
</tbody>
</table>

- 26 features, each with a number of possible values
- Languages select relevant features
- May add language-specific features or values
Language-Specific Features

Three types of infinitives in Finnish:

Example: *olla* “to be”

<table>
<thead>
<tr>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
</tr>
</thead>
<tbody>
<tr>
<td>olla</td>
<td>ollessa</td>
<td>olemassa</td>
</tr>
<tr>
<td>ollen</td>
<td>olemaan</td>
<td>olemasta</td>
</tr>
<tr>
<td></td>
<td>olemalla</td>
<td></td>
</tr>
<tr>
<td></td>
<td>olematta</td>
<td></td>
</tr>
<tr>
<td>Joku</td>
<td>yrittää</td>
<td>piristää</td>
</tr>
<tr>
<td>--------------</td>
<td>--------------</td>
<td>------------</td>
</tr>
<tr>
<td>Someone</td>
<td>tries</td>
<td>to-uplift</td>
</tr>
<tr>
<td>PRON</td>
<td>VERB</td>
<td>VERB</td>
</tr>
<tr>
<td>VerbForm=Fin</td>
<td>VerbForm=Inf</td>
<td>Inf</td>
</tr>
<tr>
<td>Case=Ade</td>
<td>Tense=Pres</td>
<td>Case=Ade</td>
</tr>
</tbody>
</table>
Joku yrittää piristää itseään värjäämällä hiuksensa

Someone tries to uplift oneself by-staining their-hair

<table>
<thead>
<tr>
<th>Joku</th>
<th>yrittää tries</th>
<th>piristää to-uplift</th>
<th>itseään oneself</th>
<th>värjäämällä by-staining</th>
<th>hiuksensa their-hair</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRON</td>
<td>VERB</td>
<td>VERB</td>
<td>PRON</td>
<td>VERB</td>
<td>NOUN</td>
</tr>
<tr>
<td>Mood=Ind</td>
<td>VerbForm=Fin</td>
<td>VerbForm=Inf</td>
<td>InfForm=1</td>
<td>InfForm=3</td>
<td>Case=Ade</td>
</tr>
<tr>
<td>Tense=Pres</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Universal Dependencies
Czech adjectives agree with nouns in gender.

velký  
big  
ADJ  
Gender=Masc

bratr  
brother  
NOUN  
Gender=Masc

velká  
big  
ADJ  
Gender=Fem

sestra  
sister  
NOUN  
Gender=Fem
**Layered Features**

Possessive adjectives: agreement gender vs. lexical gender

<table>
<thead>
<tr>
<th></th>
<th>otcův</th>
<th>bratr</th>
<th>matčin</th>
<th>bratr</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>father’s</td>
<td>brother</td>
<td>mother’s</td>
<td>brother</td>
</tr>
<tr>
<td>ADJ</td>
<td>Gender=Masc</td>
<td>Gender=Masc</td>
<td>Gender=Masc</td>
<td>Gender=Masc</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gender=Masc</td>
<td>Gender=Masc</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Otcova</td>
<td>Sestra</td>
<td>Matčina</td>
<td>Sestra</td>
</tr>
<tr>
<td>ADJ</td>
<td>Gender=Fem</td>
<td>Gender=Fem</td>
<td>Gender=Fem</td>
<td>Gender=Fem</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gender=Masc</td>
<td>Gender=Masc</td>
<td></td>
</tr>
</tbody>
</table>

Universal Dependencies

Morphological Annotation in UD 39/47
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 yanılttığıni* **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”
Features Apply to Individual Words

Future tense in Spanish and German: no $\text{Tense=Fut}$ in German!

<table>
<thead>
<tr>
<th>Dormirá</th>
<th>Er</th>
<th>wird</th>
<th>schlafen</th>
</tr>
</thead>
<tbody>
<tr>
<td>He-will-sleep</td>
<td>He</td>
<td>will sleep</td>
<td>sleep</td>
</tr>
<tr>
<td><strong>VERB</strong></td>
<td><strong>PRON</strong></td>
<td><strong>AUX</strong></td>
<td><strong>VERB</strong></td>
</tr>
<tr>
<td>VerbForm=Fin</td>
<td>PronType=Prs</td>
<td>VerbForm=Fin</td>
<td>VerbForm=Inf</td>
</tr>
<tr>
<td>Mood=Ind</td>
<td>Number=Sing</td>
<td>Mood=Ind</td>
<td>Number=Sing</td>
</tr>
<tr>
<td>$\text{Tense=Fut}$</td>
<td>Person=3</td>
<td>$\text{Tense=Pres}$</td>
<td>Person=3</td>
</tr>
<tr>
<td>Number=Sing</td>
<td>Gender=Masc</td>
<td>Number=Sing</td>
<td></td>
</tr>
<tr>
<td>Person=3</td>
<td>Case=Nom</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Participle Types

<table>
<thead>
<tr>
<th>Russian</th>
<th>English</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>некурящий человек</td>
<td>non-smoking person</td>
<td>VerbForm=Part Tense=Pres Gender=Masc Number=Sing Case=Nom</td>
</tr>
<tr>
<td>начавшийся разговор</td>
<td>that-has-started conversation</td>
<td>VerbForm=Part Tense=Past Gender=Masc Number=Sing Case=Nom</td>
</tr>
</tbody>
</table>

- 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

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

And as they were returning ...
Conflicting Traditional Terminologies

And as they-were returning in-year 1942 vračali...

da that not would in Athens prišli...

Universal Dependencies

45/47
Conflicting Traditional Terminologies

<table>
<thead>
<tr>
<th>Lexical</th>
<th>Morphological</th>
<th>Syntactical</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>ko</td>
<td>se</td>
</tr>
<tr>
<td>And</td>
<td>as</td>
<td>se</td>
</tr>
<tr>
<td>so</td>
<td>they-were</td>
<td>in-year</td>
</tr>
<tr>
<td>1942</td>
<td>vračali</td>
<td>returning</td>
</tr>
<tr>
<td>1942</td>
<td>da</td>
<td>ne</td>
</tr>
<tr>
<td>not</td>
<td>v</td>
<td>in</td>
</tr>
<tr>
<td>Atene</td>
<td>bi</td>
<td>vrišli</td>
</tr>
<tr>
<td>that</td>
<td>would</td>
<td>they-come</td>
</tr>
<tr>
<td>SCONJ</td>
<td>AUX</td>
<td>ADP</td>
</tr>
<tr>
<td>PART</td>
<td>PROPN</td>
<td>VERB</td>
</tr>
<tr>
<td>V</td>
<td>future</td>
<td>not</td>
</tr>
<tr>
<td>ne</td>
<td>they-will</td>
<td>drive</td>
</tr>
<tr>
<td>bodo</td>
<td>vozili</td>
<td>zgolj</td>
</tr>
<tr>
<td>vzolj</td>
<td>les</td>
<td>just</td>
</tr>
<tr>
<td>les</td>
<td>wood</td>
<td></td>
</tr>
</tbody>
</table>
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: нарушивший / narušivšij
  - bg “participle + past (aorist) / imperfect” (two subtypes)
  - cu “participle + resultative aspect” (lang-spec)
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: *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

- 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/npfl1075