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

En kat jager rotter og mus
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
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. … 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.
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.
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.
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.
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
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
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
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
### 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
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), Ἕλληνικά (+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:
    - Erzja, 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: APURINĀ, 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:
          - BASK, SWF. SIGN, NAJAVA
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.

PRON
VERB

«¡María, te amo!», ,
PUNCT PUNCT PROPN PUNCT PRON VERB 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)
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.”
He abdicated in favour of his son 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。

Xiànzài wǒmen zài Wǎlúnxīyǎ.

Now we in Valencia.

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

経堂の美容室に行った
Kyōdō no miyōshitsu ni ite kimashita

経堂の美容室に行くて来るた
Kyōdō no miyōshitsu ni go-te kieru ta
I went to the beauty salon of Kyōdō [Beyond-R.]
I went to the beauty salon of Kyōdō [Beyond-R.]

経堂の美容室に行ってきました

Kyōdō no miyōshitsuni itte kimashita

経堂美容室に行く来る

Kyōdō no miyōshitsuni going come

PROPN NOUN VERB VERB

Case=Gen Case=Dat VerbForm=Conv VerbForm=Fin

Tense=Past Polite=Form
Vietnamese: Words with Spaces

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

Tất cả đường bê tông 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 (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!
• When to split?
  • Only part of the token involved in a relation to something outside the token? Split!
  • Hard time finding POS tag? Split!

Vietnamese writing system
• Very restricted set of exceptions (numbers)
• Special relations elsewhere (fixed, compound)

Universal Dependencies
Morphological Annotation in UD
• 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

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

# text = Vámonos al mar.
# text_en = Let’s go to the sea.
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 |
• 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.
• 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

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

- 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       | PRON       | PUNCT        |
| PROPN      | DET        | SYM          |
| VERB       | AUX        | X            |
| ADJ        | NUM        |              |
| ADV        | ADP        |              |
| INTJ       | SCONJ      |              |
|            | CCONJ      |              |
|            | PART       |              |

- common noun
- proper noun
- verb
- adjective
- adverb
- interjection
- pronoun
- determiner
- auxiliary
- numeral
- adposition
- subordinator
- coordinator
- particle
- punctuation
- symbol
- unknown

- 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)*
# Features

| 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                            |
|                 |                          | Person                              |
| Abbr            |                          | Polite                              |
| Typo            |                          | Clusivity                           |

- 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”

|   | 1st  | 2nd  | 3rd         |
|---|------|------|-------------|
| olla | ollena | olemassa |
|    |      | olemaan |
|    |      | olemasta |
|    |      | olemalla |
|    |      | olematta |
Joku yrittää piristää itseään värjäämällä hiuksensa.

Someone tries to uplift oneself by-staining their-hair.

- **Joku**: PRON, **yrittää**: VERB, **piristää**: VERB, **itseään**: PRON, **värjäämällä**: VERB, **hiuksensa**: NOUN
- **Mood=Ind**, **VerbForm=Fin**, **VerbForm=Inf**, **VerbForm=Inf3**, **Tense=Pres**, **Case=Ade**
Joku yrittää piristää itseään värjämällä hiuksensa.

Someone tries to uplift oneself by-staining their-hair.

Universal Dependencies
Czech adjectives agree with nouns in gender.

velký  
big  
ADJ  
Gender=Masc

bratr  
big  
brother  
NOUN  
Gender=Masc

velká  
big  
ADJ  
Gender=Fem

sestra  
sister  
NOUN  
Gender=Fem
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]=Masc | Gender[psor]=Fem | 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]=Masc | Gender[psor]=Fem | Gender[psor]=Fem |
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 \textbf{PronType=Int,Rel}
- In other cases, it is desirable to disambiguate by context. Polish \textit{którym} (form of \textit{który} “which”) can be \texttt{Case=Ins}, \texttt{Loc} in singular or \texttt{Dat} in plural but we do not want to annotate \texttt{Case=Dat,Ins,Loc}!
- All values of the feature/language? Omit the feature completely! Polish: \texttt{Gender=Fem,Masc,Neut}. Spanish: \texttt{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)
  - *yanılmıș* "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)
  - *yanı̇l* “be wrong”
  - *yanı̇lmişim* **Voice=Act** “I was wrong”
  - *okuru yanı̇lttiğini* **Voice=Cau** “mislead the reader”
  - *okurlar yanı̇ltilmiş tir* **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 Tense=Fut in German!

Dormirá
He-will-sleep
VERB
VerbForm=Fin
Mood=Ind
Tense=Fut
Number=Sing
Person=3

Er
He
PRON
PronType=Prs
Number=Sing
Person=3
Gender=Masc
Case=Nom

wird
will
AUX
VerbForm=Fin
Mood=Ind
Tense=Pres
Number=Sing
Person=3

schläfen
sleep
VERB
VerbForm=Inf
Mood=Ind
Number=Sing
Person=3
### Participle Types

| нежуряшкичий | человек | начавшийся | разговор |
|---------------|---------|------------|---------|
| некурящий | человек | начался | разговор |
| 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 **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*, *adverbia ial participle*, *ger und*
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
• \texttt{VerbForm=Conv} – 
  \textit{converb}, transgressive, adverbial participle, gerund
• \textit{Gerund} (\texttt{VerbForm=Ger})
  • English: close to verbal nouns (\texttt{VerbForm=Vnoun})
  • Spanish: more like present participle (\texttt{VerbForm=Part} | \texttt{Tense=Pres})
  • Slavic: converb (\texttt{VerbForm=Conv})
• \textit{Aorist}
  • Ancient Greek, Turkish: neutral \underline{non-past} tense (they use a language-specific value \texttt{Tense=Aor})
  • Slavic languages: simple \underline{past} tense (\texttt{Tense=Past})
And as they were returning...
Conflicting Traditional Terminologies
Conflicting Traditional Terminologies

And as they were returning in year 1942, they would come to Athens and then they will drive just wood.

da ne bi v prihodnje future ne bodo vozili just drive zgolj les

Universal Dependencies

45/47
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/npfl075