

Universal Dependencies: Words and Morphology

Daniel Zeman

📅 July 21, 2026



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



unless otherwise stated

<https://ufal.mff.cuni.cz/courses/dz-2026-07-bh>



- 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)

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.

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



We are now in Valencia.

現在我們在瓦倫西亞。

Xiànzài 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 .

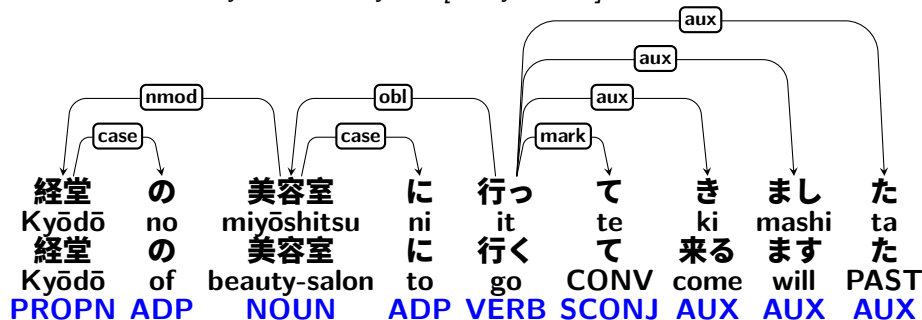
Now we in Valencia .

ADV PRON ADP PROP N PUNCT



Words in Japanese

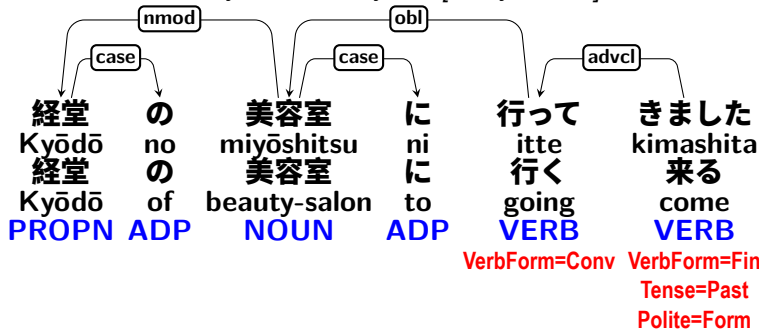
I went to the beauty salon of Kyōdō [, Beyond-R.]





Words in Japanese

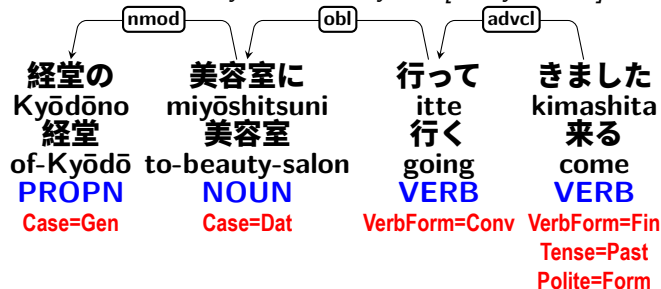
I went to the beauty salon of Kyōdō [, Beyond-R.]





Words in Japanese

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	...
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.



Numbers with Spaces

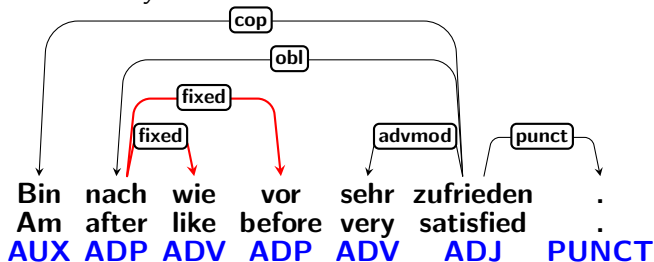
#	text = Il touche environ 100 000 sesterces par an.									
1	Il	il	PRON	...	2	nsubj	—	—		
2	touche	toucher	VERB	...	0	root	—	—		
3	environ	environ	ADV	...	4	advmod	—	—		
4	100 000	100 000	NUM	...	5	nummod	—	—		
5	sesterces	sesterce	NOUN	...	2	obj	—	—		
6	par	par	ADP	...	7	case	—	—		
7	an	an	NOUN	...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



Fixed Expressions

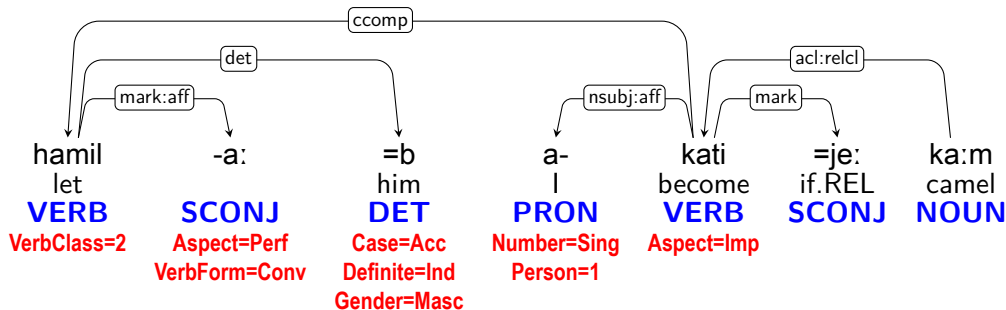
One syntactic word spans several orthographic words?

I am still very satisfied.





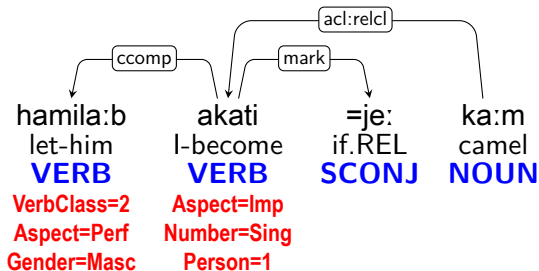
Fieldwork Glosses in Beja



“a camel that I had let loose”



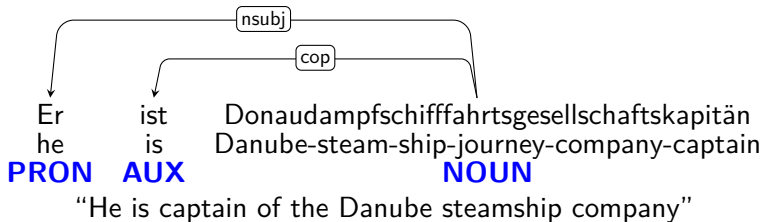
Fieldwork Glosses in Beja



"a camel that I had let loose"



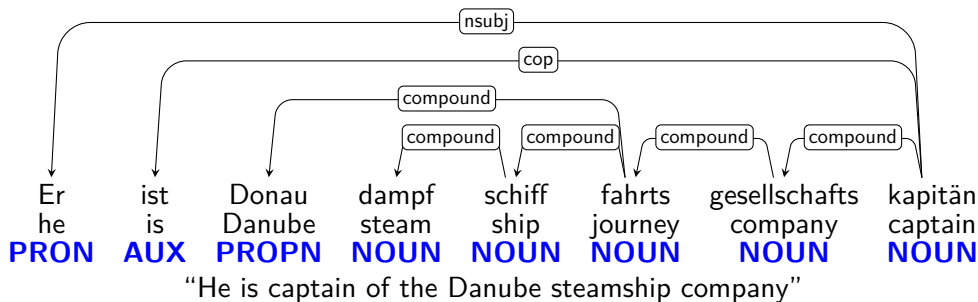
Compounds in German





Compounds in German

This is not done in UD!



Compounds in Parallel UD

PUD sentence id w05003050

- [en] Under his administration, the female vote was approved.
- [de] Unter seiner Regierung wurde das Frauenwahlrecht eingeführt.
- [sv] Under hans administration godkändes den kvinnliga rösträtten.
- [fi] Hänen hallintonsa aikana naisten äänioikeus hyväksyttiin.
- [it] Durante la sua amministrazione venne approvato il diritto di voto alle donne.
- [es] En su gestión se decretó el sufragio femenino.
- [tr] Yönetimi sırasında kadınların oy kullanması kabul edilmiş.
 - = Administration during women-of vote to-use acceptance been-taken
- [cs] Pod jeho správou směly volit ženy.
 - (Pod jeho správou bylo zavedeno volební právo pro ženy.)



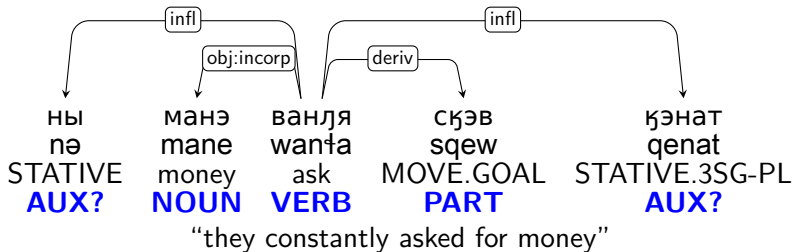
НЫМАНЭВАНЛЯСҚЭВҚЭНАТ
nəmanewanʃasqewqenat
constantly-they-money-asked
VERB

“they constantly asked for money”



Incorporation in Chukchi

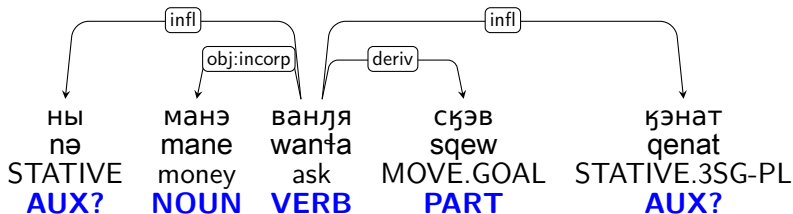
This is not done in UD!





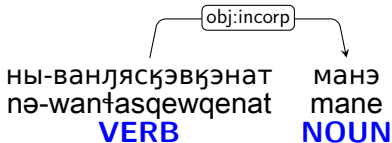
Incorporation in Chukchi

This is not done in UD!



"they constantly asked for money"

But you will find this in **enhanced** UD:





çöplüklerimizdekilerdenmiydi

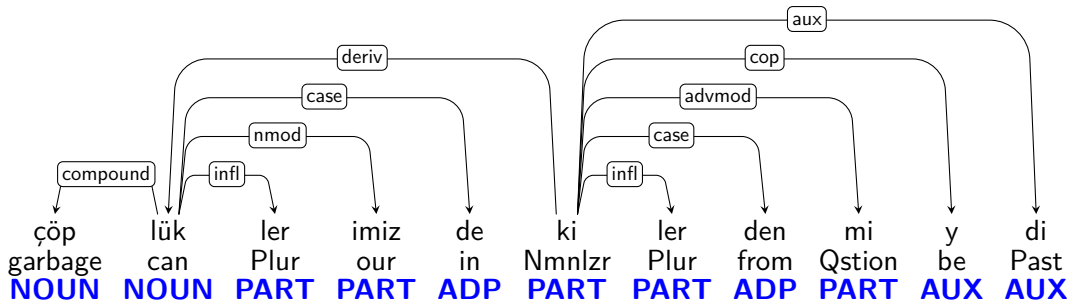
VERB

“was it from those that were in our garbage cans?”



Inflection vs. Derivation in Turkish

This is not done in UD!

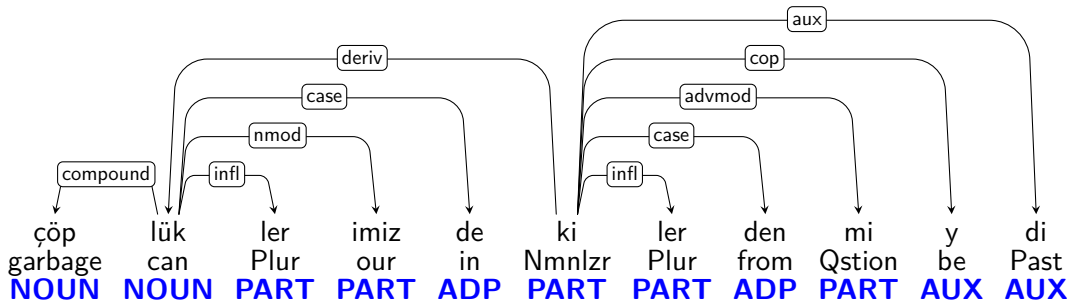


"was it from those that were in our garbage cans?"

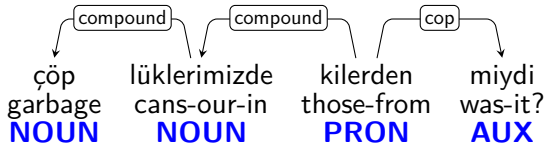


Inflection vs. Derivation in Turkish

This is not done in UD!



“was it from those that were in our garbage cans?”

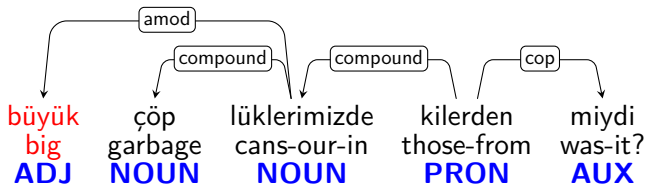




Inflection vs. Derivation in Turkish



"was it from those that were in our **big** garbage cans?"



Word Segmentation Summary

- 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!

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.

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).

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.

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

- 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 ($\Rightarrow$ 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)



within a year Algeria will become an islamic state

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

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

Part-of-Speech Tags

<https://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

- 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**)

Features

<https://universaldependencies.org/u/feat/index.html>

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

- 27 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”

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



Language-Specific Features

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				



Language-Specific Features

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	<u>InfForm=1</u>		<u>InfForm=3</u>	
	Tense=Pres			Case=Ade	



Layered Features

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

otcův
father's
ADJ

Gender=Masc
Gender[psor]=Masc

bratr
brother
NOUN

Gender=Masc

matčin
mother's
ADJ

Gender=Masc
Gender[psor]=Fem

bratr
brother
NOUN

Gender=Masc

otcova
father's
ADJ

Gender=Fem
Gender[psor]=Masc

sestra
sister
NOUN

Gender=Fem

matčina
mother's
ADJ

Gender=Fem
Gender[psor]=Fem

sestra
sister
NOUN

Gender=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 **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 => 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)

- 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
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

schlafen
sleep
VERB

VerbForm=Inf



Participle Types

некурящий человек
 nekurjaščij čelovek
 non-smoking person
ADJ NOUN

VerbForm=Part

Tense=Pres

Gender=Masc

Number=Sing

Case=Nom

Gender=Masc

Number=Sing

Case=Nom

начавшийся разговор
 načavšijsja razgovor
 that-has-started conversation
ADJ NOUN

VerbForm=Part

Tense=Past

Gender=Masc

Number=Sing

Case=Nom

Gender=Masc

Number=Sing

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

- 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**)
 - Romance: closer to participles? (**VerbForm=Part | Tense=Pres**) or converbs? (**VerbForm=Conv**)
 - 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**)
 - Romance: closer to participles? (**VerbForm=Part | Tense=Pres**) or converbs? (**VerbForm=Conv**)
 - 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

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				<u>Tense=Past?</u>	



Conflicting Traditional Terminologies

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				<u>Tense=Past?</u>	

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			<u>Tense=Past??</u>	



Conflicting Traditional Terminologies

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				<u>Tense=Past?</u>	

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			<u>Tense=Past??</u>	

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	<u>Tense=Past???</u>		

Conflicting Traditional Terminologies

Slavic *I-participle*:

- 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

Slavic *I-participle*:

- 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

Slavic *I-participle*:

- 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)
- “I-participle”
 - But that would be a language-specific verb form (**VerbForm=LPart**).

- UPOS: <https://universaldependencies.org/u/pos/index.html>
- Universal FEATS: <https://universaldependencies.org/u/feat/index.html>
- Features used in data (including language-specific):
<https://universaldependencies.org/survey-feats.html>

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.mff.cuni.cz/courses/dz-2026-07-bh>