Universal Dependencies

Daniel Zeman

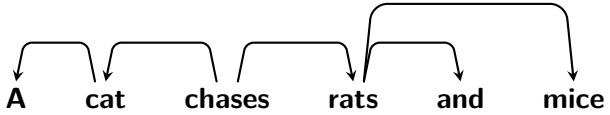
March 21, 2024

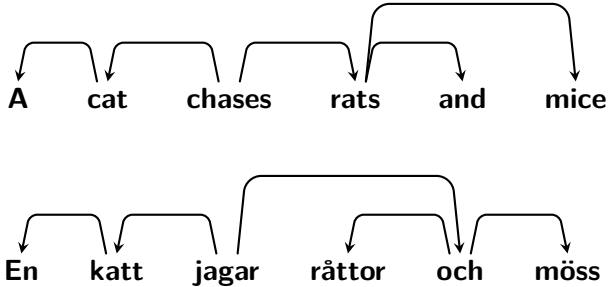
Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

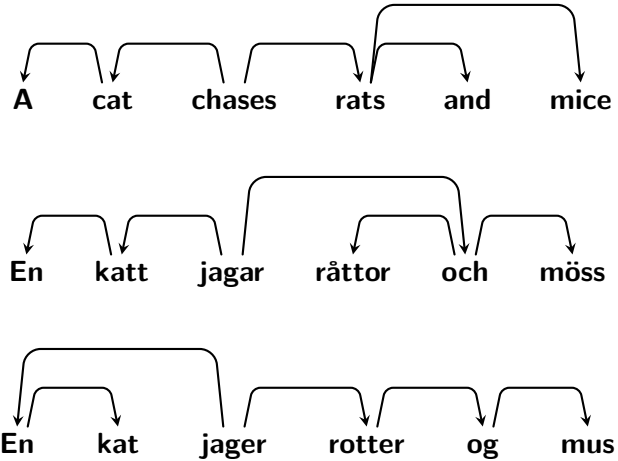
Introduction

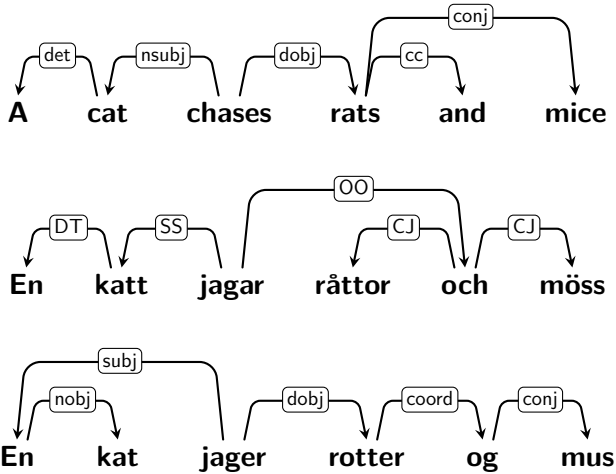
- Around 2010:

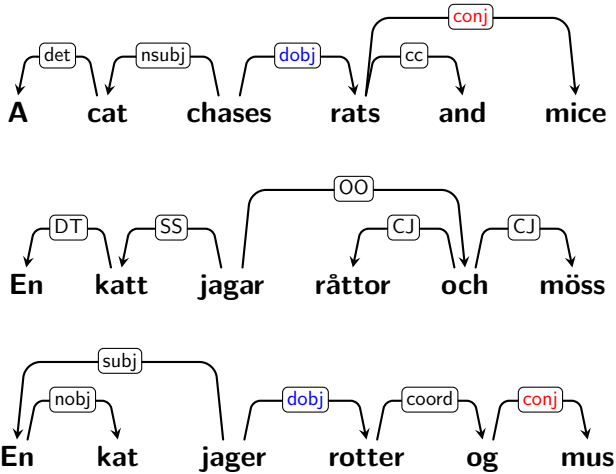
- Increasing interest in multilingual NLP
  - Multilingual evaluation campaigns to test generality
  - Cross-lingual learning to support low-resource languages
- Increasing awareness of methodological problems
  - Current NLP relies heavily on annotation
  - Annotation schemes vary across languages


A
cat
chases
rats
and
mice


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


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


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

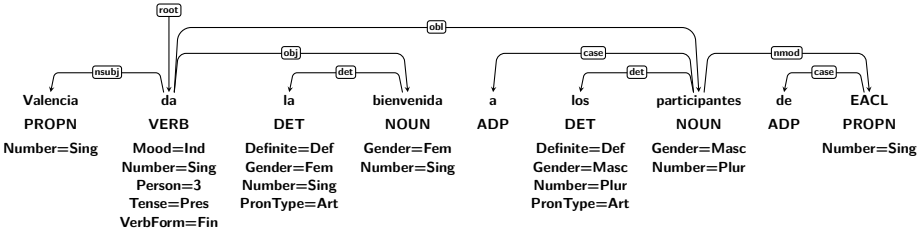

det
nsubj
dobj
cc
conj
A
cat
chases
rats
and
mice
DT
SS
OO
CJ
CJ
En
katt
jagar
råttor
och
möss
subj
nobj
dobj
coord
conj
En
kat
jager
rotter
og
mus

Why was this a problem?

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

Universal Dependencies

http://universaldependencies.org



- Part-of-speech tags
- Morphological features
- Syntactic dependencies

Universal 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.

Manning's Law

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

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

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

Manning's Law

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

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

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

Manning's Law

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

1. UD must be satisfactory on linguistic analysis grounds for individual languages.
2. UD must be good for linguistic typology, i.e., providing a suitable basis for bringing out cross-linguistic parallelism across languages and language families.
3. UD must be suitable for rapid, consistent annotation by a human annotator.
4. UD must be easily comprehended and used by a non-linguist, whether a language learner or an engineer with prosaic needs for language processing. … it leads us to favor traditional grammar notions and terminology.

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

Manning's Law

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

1. UD must be satisfactory on linguistic analysis grounds for individual languages.
2. UD must be good for linguistic typology, i.e., providing a suitable basis for bringing out cross-linguistic parallelism across languages and language families.
3. UD must be suitable for rapid, consistent annotation by a human annotator.
4. UD must be easily comprehended and used by a non-linguist, whether a language learner or an engineer with prosaic needs for language processing. … it leads us to favor traditional grammar notions and terminology.
5. UD must be suitable for computer parsing with high accuracy.

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

Manning's Law

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

1. UD must be satisfactory on linguistic analysis grounds for individual languages.
2. UD must be good for linguistic typology, i.e., providing a suitable basis for bringing out cross-linguistic parallelism across languages and language families.
3. UD must be suitable for rapid, consistent annotation by a human annotator.
4. UD must be easily comprehended and used by a non-linguist, whether a language learner or an engineer with prosaic needs for language processing. … it leads us to favor traditional grammar notions and terminology.
5. UD must be suitable for computer parsing with high accuracy.
6. UD must support well downstream language understanding tasks (relation extraction, reading comprehension, machine translation, …)

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

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

Morphological Annotation

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

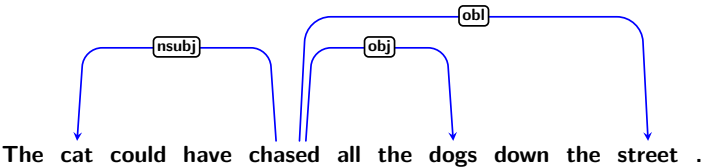
- Lemma representing the semantic content of a word
- Part-of-speech tag representing its grammatical class
- Features representing lexical and grammatical properties of the lemma or the particular word form

Syntactic Annotation

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

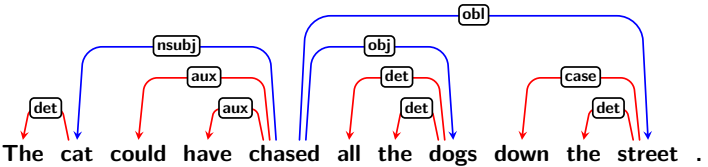
- Content words are related by dependency relations
- Function words attach to the content word they modify
- Punctuation attach to head of phrase or clause

Syntactic Annotation



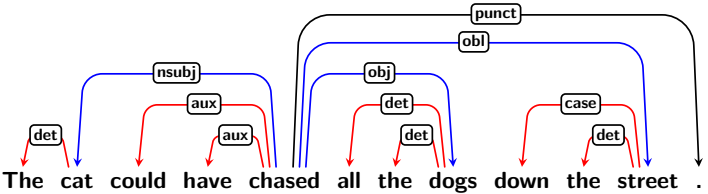
- Content words are related by dependency relations
- Function words attach to the content word they modify
- Punctuation attach to head of phrase or clause

Syntactic Annotation



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

Syntactic Annotation



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

CoNLL-U Format

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1	Es	es	PRON	_	_	2	nsubj	_	_
2	unterscheidet	unterscheiden	VERB	_	_	0	root	_	_
3	sich	sich	PRON	_	_	2	expl:pv	_	_
4-5	vom	_	_	_	_	_	_	_	_
4	von	von	ADP	_	_	7	case	_	_
5	dem	der	DET	_	_	7	det	_	_
6	westlichen	westlich	ADJ	_	_	7	amod	_	SpaceAfter=No
7	Teil	Teil	NOUN	_	_	2	obl	_	_
8	des	der	DET	_	_	9	det	_	_
9	Landes	Land	NOUN	_	_	7	nmod	_	SpaceAfter=No
10	.	.	PUNCT	_	_	2	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:
  - 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, Guaraní, 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

Tokenization

"María, I love you!" Juan exclaimed.

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

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

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

Word Segmentation

Let's go to the sea.

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

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

Word Segmentation

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

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

Contractions in Arabic

He abdicated in favour of his son Baudouin.

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

Chinese Word Segmentation

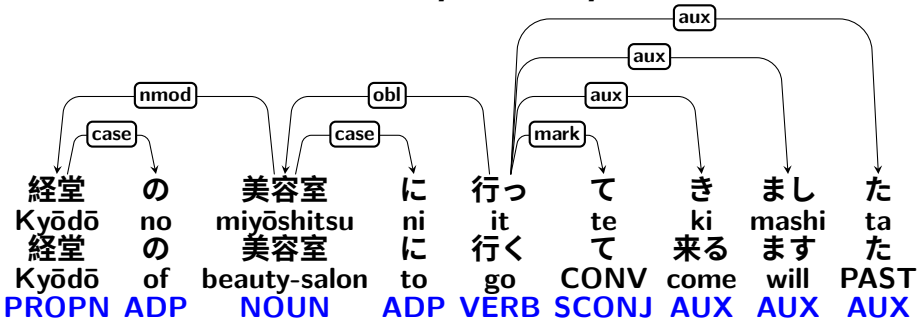
We are now in Valencia.

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

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

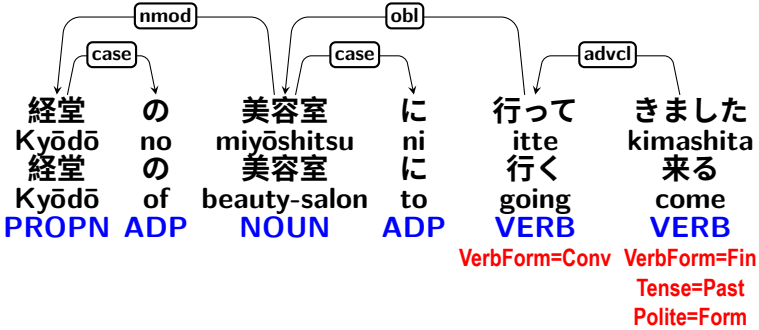
Words in Japanese

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



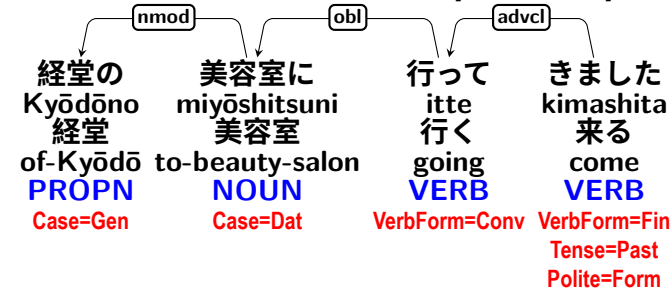
Words in Japanese

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



Words in Japanese

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



Vietnamese: Words with Spaces

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

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

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

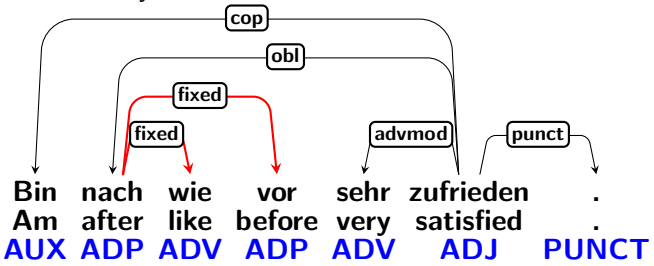
Numbers with Spaces

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

Fixed Expressions

One syntactic word spans several orthographic words?

I am still very satisfied.



Word Segmentation Summary

- 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!

Lemmas

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

Lemmas

within a year Algeria will become an islamic state

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

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

Part-of-Speech Tags

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

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

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

Part-of-Speech Tags

- 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)*
- **ess)** *(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

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

- 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”

1st	2nd	3rd
olla	ollessa	olemassa
	ollen	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	InfForm=1		InfForm=3	
	Tense=Pres			Case=Ade	

Layered Features

Czech adjectives agree with nouns in gender.

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

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

Layered Features

Possessive adjectives: agreement gender vs. lexical gender

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

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

Multi-valued Features (Disjunction / Parallel Application)

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

Participle Types

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

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

Conflicting Traditional Terminologies

- 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

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

Conflicting Traditional Terminologies

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

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

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				Tense=Past?	

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

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

Conflicting Traditional Terminologies

- 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