Tutorial on Universal Dependencies

Word segmentation and morphological annotation

Joakim Nivre[1]   **Daniel Zeman**[2]   Filip Ginter[3]   Francis M. Tyers[45]

[1]Department of Linguistics and Philology, Uppsala University, Sweden

[2]Institute of Formal and Applied Linguistics, Charles University, Prague, Czechia

[3]Department of Information Technology, University of Turku, Finland

[4]Giela ja kultuvrra instituhtta, UiT Norgga árktalaš universitehta, Tromsø, Norway

[5]Arvutiteaduse instituut, Tartu Ülikool, Estonia

- Tokenization / word segmentation

- Lemmatization
- Universal part-of-speech tags
- Universal features
- Language-specific features

- Errors in text

"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	.
VERB?	X	NOUN	PUNCT

Vamos	nos	a	el	mar	.
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。
Xiànzàiwǒmenzàiwǎlúnxīyǎ.
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
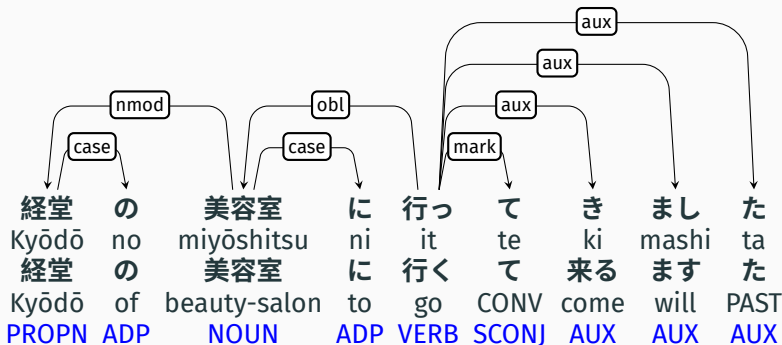
I went to the beauty salon of Kyōdō [, Beyond-R.]

経堂	の	美容室	に	行っ	て	き	まし	た
Kyōdō	no	miyōshitsu	ni	it	te	ki	mashi	ta
経堂	の	美容室	に	行く	て	来る	ます	た
Kyōdō	of	beauty-salon	to	go	CONV	come	will	PAST
PROPN	ADP	NOUN	ADP	VERB	SCONJ	AUX	AUX	AUX

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	miyōshitsu	ni	itte	kimashita
経堂	の	美容室	に	行く	来る
Kyōdō	of	beauty-salon	to	going	come
PROPN	ADP	NOUN	ADP	VERB	VERB

VerbForm=Conv VerbForm=Fin
Tense=Past
Polite=Form

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

nmod	obl	advcl

経堂の	美容室に	行って	きました
Kyōdōno	miyōshitsuni	itte	kimashita
経堂	**美容室**	**行く**	**来る**
of-Kyōdō	to-beauty-salon	going	come
PROPN	NOUN	VERB	VERB
Case=Gen	Case=Dat	VerbForm=Conv	VerbForm=Fin
			Tense=Past
			Polite=Form
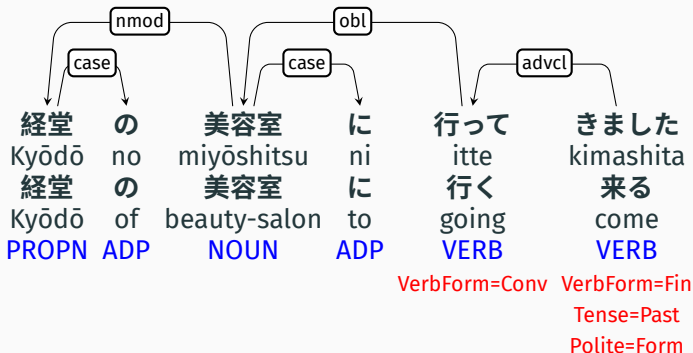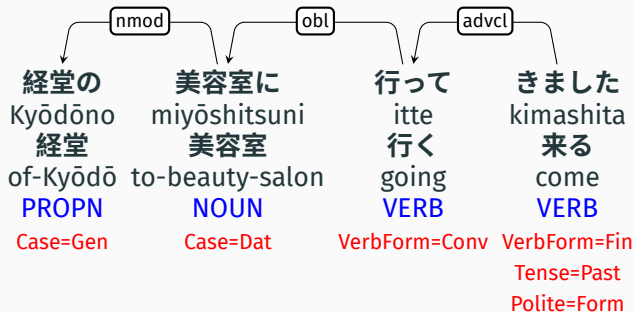
All the concrete country roads are the result of...

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      _ _
```
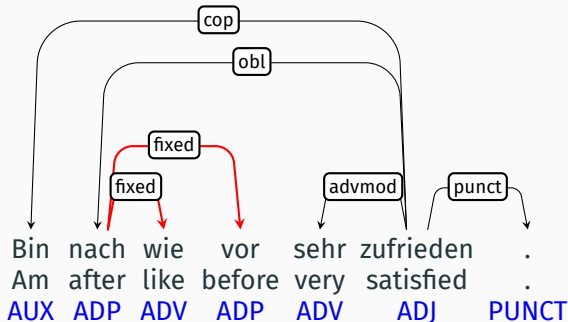
One syntactic word spans several orthographic words?

I am still very satisfied.

Bin	nach	wie	vor	sehr	zufrieden	.
Am	after	like	before	very	satisfied	.
AUX	ADP	ADV	ADP	ADV	ADJ	PUNCT

- 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!

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`)

```
#      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	_	_

```
#      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	_ _

- Punctuation involved? Low level!

- Punctuation involved? Low level!

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

Tokenization vs. Multi-word Tokens Summary

- Punctuation involved? Low level!

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

- Non-concatenative? High level!

Lemmas
UPOS tags
Features

- 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	...	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_útva

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

Part-of-Speech Tags

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

- 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!

- 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**)

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

- 21 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	ollessa	olemassa
	ollen	olemaan
		olemasta
		olemalla
		olematta

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=<u>Inf3</u>	
	Mood=Ind			Case=Ade	
	Tense=Pres				

22

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	

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

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	

- 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	

некурящий nekurjaščij non-smoking **ADJ**	человек čelovek person **NOUN**	начавшийся načavšijsja that-has-started **ADJ**	разговор razgovor conversation **NOUN**
VerbForm=Part		VerbForm=Part	
<u>Tense=Pres</u>		<u>Tense=Past</u>	
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

28

Conflicting Traditional Terminologies

- If possible, stay compatible with traditional grammar
- Often it is not possible: terminology conflicts
- `VerbForm=Conv` – <u>*converb*</u>*, 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** – <u>*converb*</u>*, 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 <u>non-past</u> tense (they use a language-specific value **Tense=Aor**)
  - Slavic languages: simple <u>past</u> tense (**Tense=Past**)

Errors in Underlying Text

- Currently not covered by the guidelines
- We do not want to hide errors (learning robust parsers!)

- Currently not covered by the guidelines
- We do not want to hide errors (learning robust parsers!)
- Possibilities:
- Typo not involving word boundary
  - FORM = *anotation*; LEMMA = *annotation*; FEATS: Typo=Yes; MISC: Correct=annotation
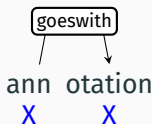
- Currently not covered by the guidelines
- We do not want to hide errors (learning robust parsers!)
- Possibilities:
- Typo not involving word boundary
    - FORM = *anotation*; LEMMA = *annotation*; FEATS: Typo=Yes; MISC: Correct=annotation

- Wrongly split word:

- Currently not covered by the guidelines
- We do not want to hide errors (learning robust parsers!)
- Possibilities:
- Typo not involving word boundary
  - FORM = *anotation*; LEMMA = *annotation*; FEATS: **Typo=Yes**; MISC: **Correct=annotation**

```
                    ┌──────────┐
                    │ goeswith │
                    └──────────┘
                     /        ↘
                  ann      otation
                   X          X
```

- Wrongly split word:
- Wrongly merged words: *thecar*
  - Fix tokenization (i.e. two lines); first line MISC: **SpaceAfter=No |** **CorrectSpaceAfter=Yes**
  - Sentence segmentation can be affected, too!

- Wrong morphology: *the cars is produced in Detroit*

- Wrong morphology: *the cars is produced in Detroit*
  - Not like normal typo *(the car iss produced...)*

- Wrong morphology: *the cars is produced in Detroit*
  - Not like normal typo *(the car iss produced…)*
  - Not obvious what is correct
    - *the car is*
    - *the cars are*

Errors in Underlying Text

- Wrong morphology: *the cars is produced in Detroit*
  - Not like normal typo *(the car iss produced...)*
  - Not obvious what is correct
    - *the car is*
    - *the cars are*

- Suggestion: select which word to fix, e.g. *cars* to *car*

- FORM = *cars*; FEATS: `Number=Plur`; MISC: `Correct=car` | `CorrectNumber=Sing`

Errors in Underlying Text

- Wrong morphology: *the cars is produced in Detroit*
  - Not like normal typo *(the car iss produced…)*
  - Not obvious what is correct
    - *the car is*
    - *the cars are*

- Suggestion: select which word to fix, e.g. *cars* to *car*

- FORM = *cars*; FEATS: `Number=Plur`; MISC: `Correct=car` | `CorrectNumber=Sing`

- cs: *viděl moři* "he saw the sea"
  - Should be *moře*
  - Would be `Case=Acc` (disambiguated from `Case=Acc,Gen,Nom,Voc`)
  - This form is `Case=Dat,Loc` (but which one?)

- *cestoval k moři* "he traveled to the sea" `Case=Dat`

- *plavil se po moři* "he sailed the sea" `Case=Loc`

Questions?