The CoNLL-U Format

Daniel Zeman

Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University
Prague, Czechia

History

- CoNLL 2006 (CoNLL-X) shared task in dependency parsing … “CoNLL-X” format
  - Extension of a much older “vertical” corpus format
  - 1 token per line, tab-separated attributes, empty line between sentences
  - Extremely simple and successful
  - Other tasks: CoNLL 2007, ICON 2009 & 2010, SPMRL 2013 & 2014
  - Treebanks: CoNLL often a secondary (or primary) format
  - Dependency parsers: recognized input/output format
- Extensions
  - CoNLL 2008 & 2009, SDP 2014 & 2015 tasks
    - Problem: variable number of columns
    - Less popular, less readable
  - CoNLL-U (“Universal”)

CoNLL-U

- Universal Dependencies (`http://universaldependencies.org/`)
- Took CoNLL-X as a de-facto standard
  - Avoided XML-based formats (TEI, PML, ...) – more official standard, more complex, but less widespread (in treebanks)
  - We want people to use UD => give them something they use anyway

CoNLL-U

- Universal Dependencies (`http://universaldependencies.org/`)
- Took CoNLL-X as a de-facto standard
  - Avoided XML-based formats (TEI, PML, ...) – more official standard, more complex, but less widespread (in treebanks)
  - We want people to use UD => give them something they use anyway
- However: Certain extensions needed
  - Sentence-level comments
  - Redefined some columns (fields)
  - Two-level tokenization/word segmentation
  - Enhanced UD representation: empty nodes, graphs

Overview

- Plain text, UTF-8 (no signature), LF-only line breaks
- Optional comment lines before a sentence (start with #)
- Mandatory empty line after each sentence
  - Empty sentences not allowed
- One or more token/word lines:
  - 10 columns (fields), separated by TAB characters
  - _ (underscore) for empty field
  - no pre-defined escaping (when token = "_")

Example

sent_id = s1
text = A short phrase.
hmm, maybe I should add another comment

1	A	a	DET	DT	Definite=Ind	3	det	_	_
2	short	short	ADJ	JJ	_	3	amod	_	_
3	phrase	phrase	NOUN	NN	Number=Sing	0	root	_	SpaceAfter=No
4	.	.	PUNCT	.	_	3	punct	_	_

sent_id = s2
text = Another one.
...

Example

sent_id = s1
text = A short phrase.
hmm, maybe I should add another comment

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1	A	a	DET	DT	Definite=Ind	3	det	_	_
2	short	short	ADJ	JJ	_	3	amod	_	_
3	phrase	phrase	NOUN	NN	Number=Sing	0	root	_	SpaceAfter=No
4	.	.	PUNCT	.	_	3	punct	_	_

sent_id = s2
text = Another one.
...

Validation

CoNLL-U data can be (in)valid at various levels:

- Readable, can be processed by tools, not necessarily complete (e.g. UPOS = “_”)

Validation

CoNLL-U data can be (in)valid at various levels:

- Readable, can be processed by tools, not necessarily complete (e.g. UPOS = "_")
- UD-released treebanks formally valid (single root, UPOS, deprels, text vs. SpaceAfter=No, unique sent_id)
    - Possible branch: non-UD annotation, e.g. Prague-style treebanks

Validation

CoNLL-U data can be (in)valid at various levels:

- Readable, can be processed by tools, not necessarily complete (e.g. UPOS = "_")
- UD-released treebanks formally valid (single root, UPOS, deprels, text vs. SpaceAfter=No, unique sent_id)
  - Possible branch: non-UD annotation, e.g. Prague-style treebanks
- UD language-specific formally valid (lists of extra features, deprel subtypes, exceptional spaces in words)

Validation

CoNLL-U data can be (in)valid at various levels:

- Readable, can be processed by tools, not necessarily complete (e.g. UPOS = "_")
- UD-released treebanks formally valid (single root, UPOS, deprels, text vs. SpaceAfter=No, unique sent_id)
  - Possible branch: non-UD annotation, e.g. Prague-style treebanks
- UD language-specific formally valid (lists of extra features, deprel subtypes, exceptional spaces in words)
- Content passes automatic tests (e.g. the `conj` relation always goes left-to-right)
  - Fully complies with UD guidelines for the language – verifiable only manually

Lemmas

do roka se Alžírsko stane islámským státem
“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
  - UD treebanks: UPOS never empty, use X
  - Not all tags have to be used by all languages
  - Need extensions? Use features!

Features

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

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

- 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~~

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	čelovek	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

Tagset Conversion

- Other tagsets can be mapped on UPOS + features
  - Universal features are based on Interset (`http://ufal.mff.cuni.cz/interset/`)

TRUNC	=>	X	Hyph=Yes	*Staats-, Industrie-, Finanz-, Öl-, Lohn-*
VAFIN	=>	AUX	Mood=Ind|VerbForm=Fin	*ist, hat, wird, sind, sei*
VAIMP	=>	AUX	Mood=Imp|VerbForm=Fin	*Seid, werde, Sei*
VAINF	=>	AUX	VerbForm=Inf	*werden, sein, haben, worden, Dabeisein*
VAPP	=>	AUX	Aspect=Perf|VerbForm=Part	*worden, gewesen, geworden, gehabt, werden*
VMFIN	=>	VERB	Mood=Ind|VerbForm=Fin|VerbType=Mod	*kann, soll, will, muß, sollen*
VMINF	=>	VERB	VerbForm=Inf|VerbType=Mod	*können, müssen, wollen, dürfen, sollen*
VMPP	=>	VERB	Aspect=Perf|VerbForm=Part|VerbType=Mod	*gewollt*
VVFIN	=>	VERB	Mood=Ind|VerbForm=Fin	*sagte, gibt, geht, steht, kommt*
VVIMP	=>	VERB	Mood=Imp|VerbForm=Fin	*siehe, sprich, schauen, Sagen, gestehe*
VVINF	=>	VERB	VerbForm=Inf	*machen, lassen, bleiben, geben, bringen*
VVIZU	=>	VERB	VerbForm=Inf	*einzusetzen, durchzusetzen, aufzunehmen, abzubau*
VVPP	=>	VERB	Aspect=Perf|VerbForm=Part	*gemacht, getötet, gefordert, gegeben, gestellt*
XY	=>	X		*dpa, ap, afp, rtr, wb*

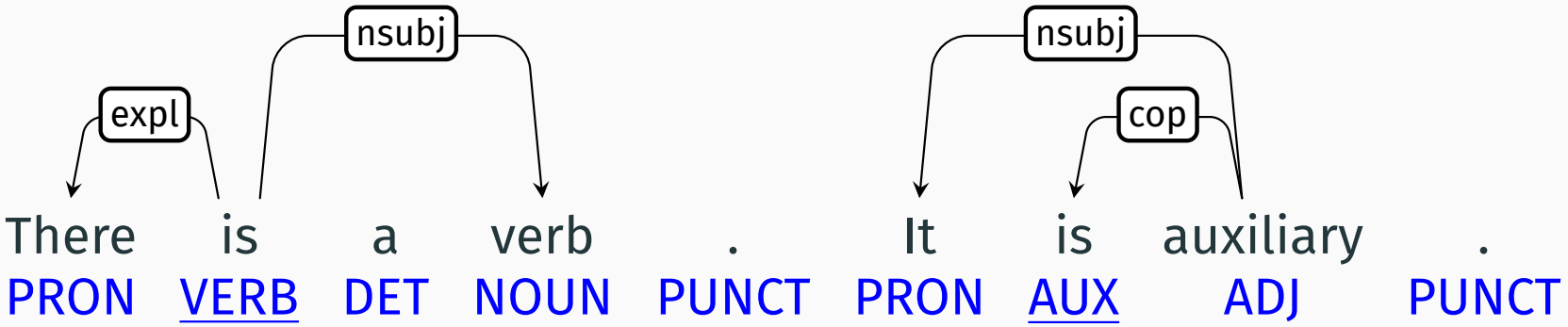
Tagset Conversion

- Sometimes tag mapping is not enough!
- Need to look at lemmas and/or dependency relations

- Croatian pronouns/pronominal words:
- *ja, ti, on, mi, vi, oni, se, tko, što, svatko, sve, nitko, ništa* …PRON
- *moj, tvoj, njegov, njezin, njen, naš, vaš, njihov, svoj, kakav, koji, koliki, čiji, nekakav, svaki, nikakav* … DET
- *gdje, odakle, kuda, kada, kad, otkada, kako, zašto, tu, tamo, ovdje, ondje, sada, tada, onda, tako, stoga, negdje, odnekud, ponekad, nekada, nekako, svuda, uvijek, svakako, nigdje, ikad, nikako* … ADV

Tagset Conversion

- Sometimes tag mapping is not enough!
- Need to look at lemmas and/or dependency relations



Tokenization

«¡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	.
VERB?	X	NOUN	PUNCT

Vamos	nos	a	el	mar	.
VERB	PRON	ADP	DET	NOUN	PUNCT

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

Recoverability

```
#     text = Vámonos al mar.
#     text_en = Let's go to the sea.
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

#	text = Vámonos al mar.						
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

#	text = Vámonos al mar.						
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	_ _

Contractions in Arabic

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

Chinese Word Segmentation

現在我們在布拉格。
Xiànzàiwǒmenzàibùlāgé.
We are now in Prague.

現在	**我們**	**在**	**布拉格**	**。**
Xiànzài	wǒmen	zài	Bùlāgé	.
Now	we	in	Prague	.
ADV	PRON	ADP	PROPN	PUNCT

Solution 1: Low Level

```
#   text = 現在我們在布拉格。
1   現在      現在      ADV     ...   4   obl     _ SpaceAfter=No
2   我們      我        PRON    ...   4   nsubj   _ SpaceAfter=No
3   在        在        ADP     ...   4   case    _ SpaceAfter=No
4   布拉格    布拉格    PROPN   ...   0   root    _ SpaceAfter=No
5   。        。        PUNCT   ...   4   punct   _ SpaceAfter=No
```

Solution 2: High Level

#	text = 現在我們在布拉格。						
1-4	現在我們在布拉格	_	_	...	_	_	_ SpaceAfte
1	現在	現在	ADV	...	4	obl	_ _
2	我們	我	PRON	...	4	nsubj	_ _
3	在	在	ADP	...	4	case	_ _
4	布拉格	布拉格	PROPN	...	0	root	_ _
5	。	。	PUNCT	...	4	punct	_ _

Low Level in Chinese: Sometimes There Is a Space!

```
#   text = 現在我們在MFF UK。
1   現在   現在   ADV     ...   5   obl        _ SpaceAfter=No
2   我們   我     PRON    ...   5   nsubj      _ SpaceAfter=No
3   在     在     ADP     ...   5   case       _ SpaceAfter=No
4   MFF    MFF    X       ...   4   compound   _ _
5   UK     UK     X       ...   0   root       _ SpaceAfter=No
6   。     。     PUNCT   ...   5   punct      _ SpaceAfter=No
```

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    _ _
```

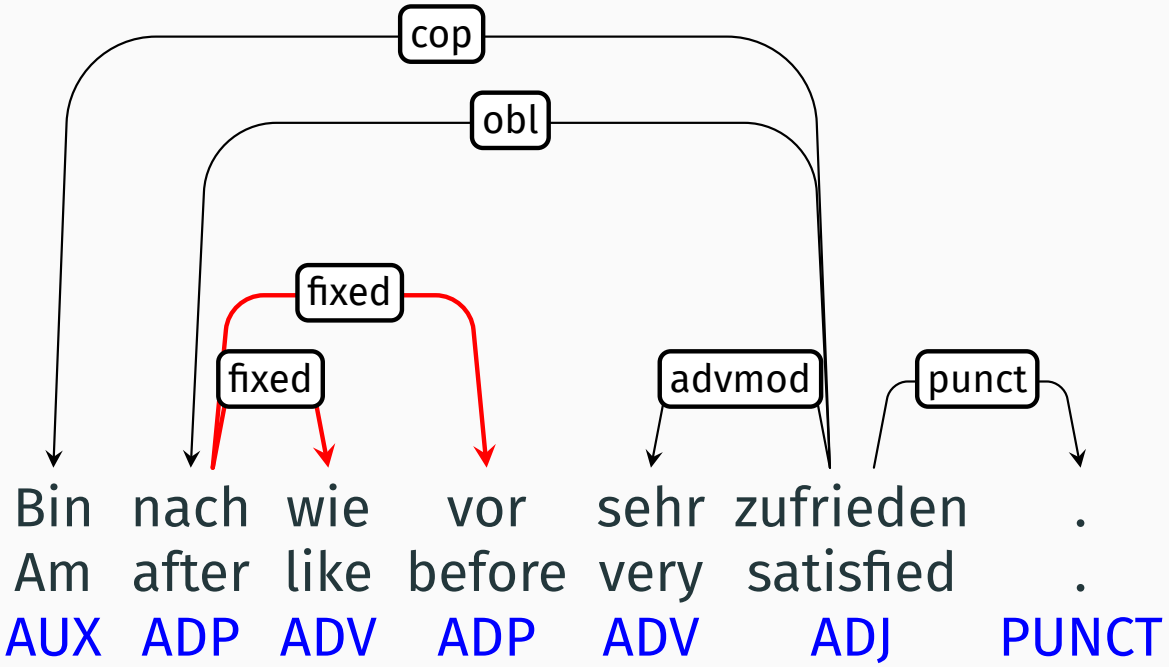
Fixed Expressions

One syntactic word spans several orthographic words?

```
#   text = Bin nach wie vor sehr zufrieden.
1   Bin          sein         AUX    ...  6   cop     _ _
2   nach         nach         ADP    ...  6   obl     _ _
3   wie          wie          ADV    ...  2   fixed   _ _
4   vor          vor          ADP    ...  2   fixed   _ _
5   sehr         sehr         ADV    ...  6   advmod  _ _
6   zufrieden    zufrieden    ADJ    ...  0   root    _ SpaceAfter=No
7   .            .            PUNCT  ...  6   obl     _ _
```

Fixed Expressions

One syntactic word spans several orthographic words?



Errors in Underlying Text

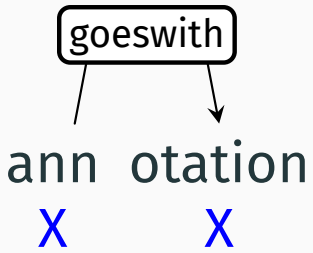
- Currently not covered by the guidelines
- We do not want to hide errors (learning robust parsers!)

Errors in Underlying Text

- 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

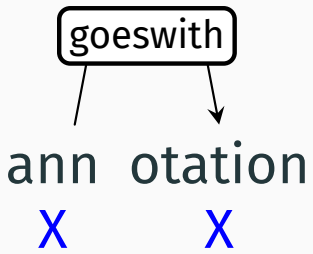
Errors in Underlying Text

- 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:



Errors in Underlying Text

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

Errors in Underlying Text

- Currently not covered by the guidelines
- We do not want to hide errors (learning robust parsers!)
- Possibilities:
- Wrong morphology: *the cars is produced in Detroit*

Errors in Underlying Text

- Currently not covered by the guidelines
- We do not want to hide errors (learning robust parsers!)
- Possibilities:
- Wrong morphology: *the cars is produced in Detroit*
    - Not like normal typo *(the car iss produced…)*

Errors in Underlying Text

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

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

Conversion of Syntax

- Depends on how far the source style is from UD
- Coordination (Prague vs. Stanford vs. Mel'čuk and variants)
- Prepositions head vs. leaf
- Auxiliaries, copulas
- Core vs. oblique arguments

- Look at source structure, POS tags, lemmas...

Conversion of Syntax

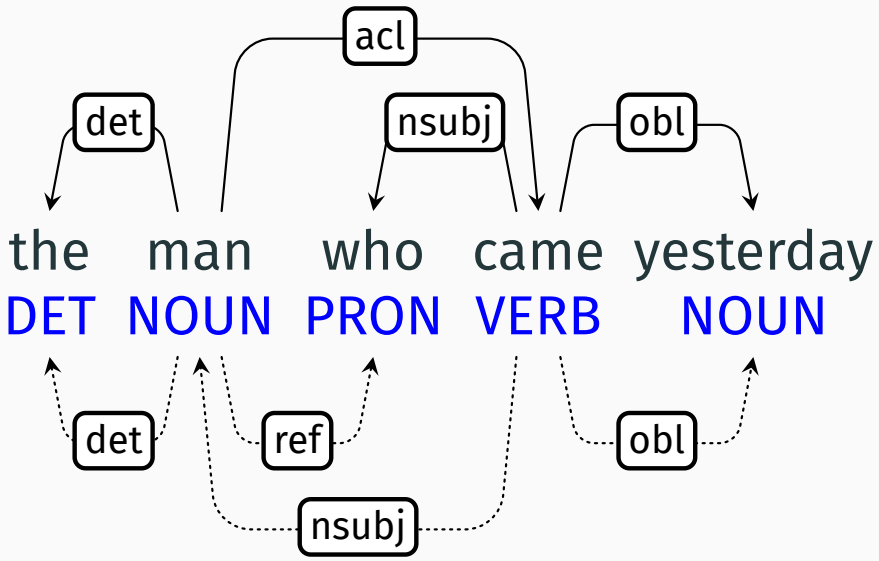
- Depends on how far the source style is from UD
- Coordination (Prague vs. Stanford vs. Mel'čuk and variants)
- Prepositions head vs. leaf
- Auxiliaries, copulas
- Core vs. oblique arguments

- Look at source structure, POS tags, lemmas...

- CoNLL-U format can hold non-UD dependencies as well!

Enhanced Graphs

- Additional relations
- Even remove basic relations
- Graph (not necessarily tree)
- DEPS column



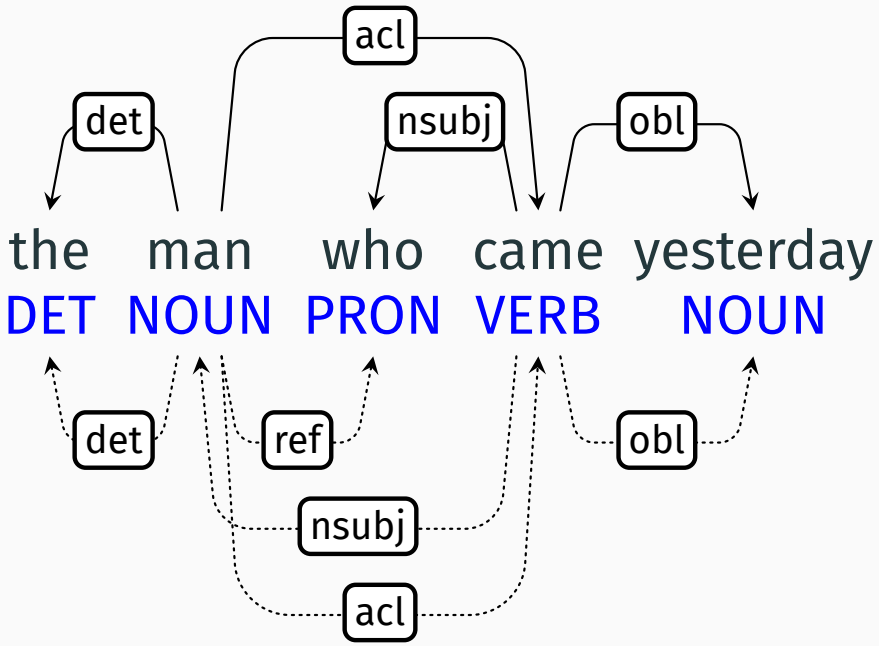
Enhanced Graphs

- Additional relations
- Even remove basic relations
- Graph (not necessarily tree)
- DEPS column

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1	the	the	DET	_	_	2	det	2:det	_
2	man	man	NOUN	_	_	0	root	4:nsubj	_
3	who	who	PRON	_	_	4	nsubj	2:ref	_
4	came	come	VERB	_	_	2	acl	0:root	_
5	yesterday	yesterday	NOUN	_	_	4	obl	4:obl	_

Enhanced Graphs

- Additional relations
- Even remove basic relations
- Graph (not necessarily tree)
- DEPS column



Enhanced Graphs

- Additional relations
- Even remove basic relations
- Graph (not necessarily tree)
- DEPS column

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1	the	the	DET	_	_	2	det	2:det	_
2	man	man	NOUN	_	_	0	root	0:root|4:nsubj	_
3	who	who	PRON	_	_	4	nsubj	2:ref	_
4	came	come	VERB	_	_	2	acl	2:acl	_
5	yesterday	yesterday	NOUN	_	_	4	obl	4:obl	_

Enhanced Graphs

- Additional relations
- Even remove basic relations
- Graph (not necessarily tree)
- DEPS column
- Empty nodes are also possible

Outside UD Extensions

- Parallel treebanks (CzEng, Martin Popel)
  - Sentence id identifies language “zone”
  - DEPS (or something similar in MISC) describes alignments
  - Target: either word ID (number), or sentence+word ID
  - (could be also used for coreference etc.)

Outside UD Extensions

- Parallel treebanks (CzEng, Martin Popel)
  - Sentence id identifies language “zone”
  - DEPS (or something similar in MISC) describes alignments
  - Target: either word ID (number), or sentence+word ID
  - (could be also used for coreference etc.)
- Stand-off annotation, addition to UD treebanks
  - Semantic roles (Alan Akbik)
  - Could be added as extra columns

Outside UD Extensions

- Parallel treebanks (CzEng, Martin Popel)
  - Sentence id identifies language “zone”
  - DEPS (or something similar in MISC) describes alignments
  - Target: either word ID (number), or sentence+word ID
  - (could be also used for coreference etc.)
- Stand-off annotation, addition to UD treebanks
  - Semantic roles (Alan Akbik)
  - Could be added as extra columns
- In general:
  - CoNLL-U is simple but you can make it complex
  - Sentence comments + MISC can carry almost anything

Outside UD Extensions

- Parallel treebanks (CzEng, Martin Popel)
  - Sentence id identifies language “zone”
  - DEPS (or something similar in MISC) describes alignments
  - Target: either word ID (number), or sentence+word ID
  - (could be also used for coreference etc.)
- Stand-off annotation, addition to UD treebanks
  - Semantic roles (Alan Akbik)
  - Could be added as extra columns
- In general:
  - CoNLL-U is simple but you can make it complex
  - Sentence comments + MISC can carry almost anything
  - At some point it may have been better to use XML
  - (although grep + regex still can filter out most of it)

Tools

- `http://universaldependencies.org/tools.html`
- `https://github.com/UniversalDependencies/tools`
- Validator
- UDAPI (`http://udapi.github.io/`)

- Annotation: some adaptations but nothing perfect
- Morphology: spreadsheet works reasonably well!

- Corpus search engine: problem with two-level word segmentation
  - (search for *al*, or *a+el*?)

Questions?