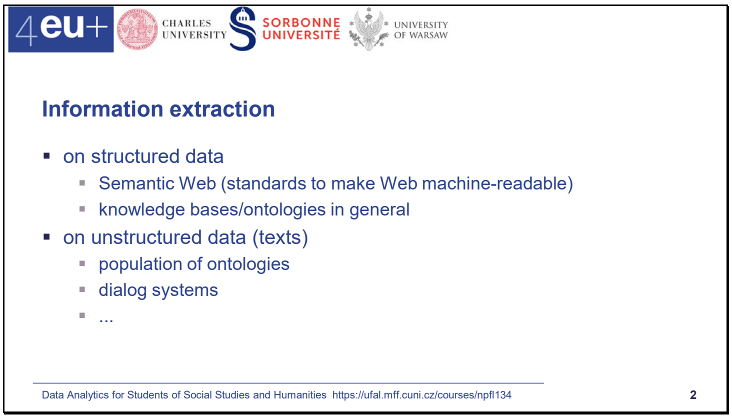
Slide 1

Introduction
Class #6, March 21 2023

Silvie Cinková cinkova@ufal.mff.cuni.cz

Slide 2



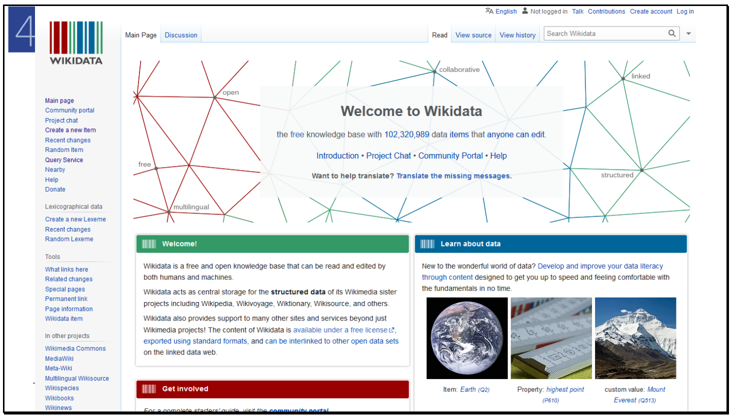
I am going to argue for using the linguistic markup for information extraction from unstructured data (text, usually). Present the difference of information extraction from structured data first, then get to the unstructured data.

Slide 3

Information extraction on structured data

- Resource Description Framework (RDF), Web Ontology Language (OWL)
  - concepts: *city, tree, event, ...*
  - entities *Sophia Loren, Bible, Volkswagen Beetle, Coca-Cola*
  - relations between entities: part of, place of birth, occupation, date of beginning
  - categories: humans, animals

Slide 4



A completely free knowledge base of Wikipedia, with links to other structured knowledge bases (national bibliographies etc.) The Wikidata repository consists of items, each one having a label and a description.

Slide 5

Item	Property	Value
Q42	P69	Q691283
Douglas Adams	educated at	St John's College

Wikidata property related to religions and beliefs [edit]

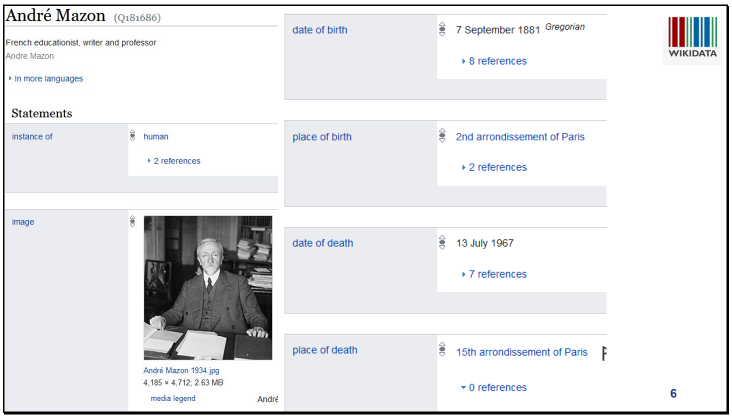
Title	ID	Data type	Description	Examples
place of burial	P119	Item	location of grave, resting place, place of ash-scattering, etc. (e.g., town/city or cemetery) for a person or animal. There may be several places: e.g., re-burials, parts of body buried separately.	Christian Doppler <*place of burial*> Cemetery of San Michele
religion or worldview	P140	Item	religion of a person, organization or religious building, or associated with this subject	Narendra Modi <*religion or worldview*> Hinduism
canonization status	P411	Item	stage in the process of attaining sainthood per the subject's religious organization	John Paul II <*canonization status*> saint
patron saint	P417	Item	patron saint adopted by the subject	Paris <*patron saint*> Genevieve

WIKIDATA

5

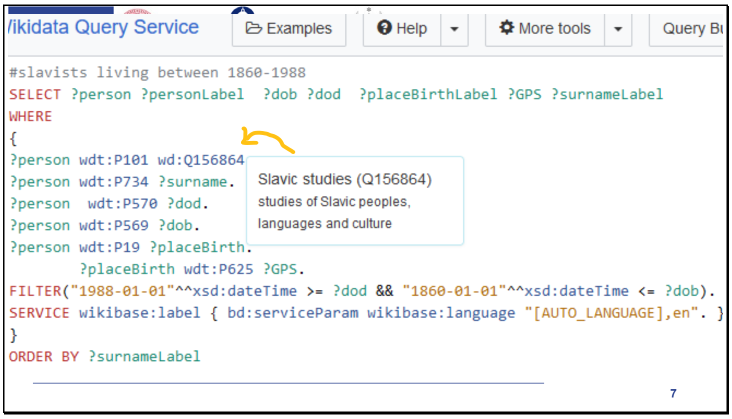
Item label starts with Q. When you describe an item, you make statements, which consist of the item, its properties and their values. The value of a property is very often another item.

Slide 6



Part of Wikidata entry of André Mazon with a few properties.

Slide 7



Who were Mazon's professional contemporaries and where were they from? Slavists who were one generation older to two generations younger.
SPARQL Semantic query language for databases able to retrieve and manipulate data stored in RDF;
Display the names, birth and death dates and birthplaces of people whose field of work (P101) was Slavic studies and limit the query to people who lived between 1860-1988 and were thus Mazon's contemporaries (one generation older or two generations younger).
Also provide the GPS coordinates of the birth places.

Slide 8

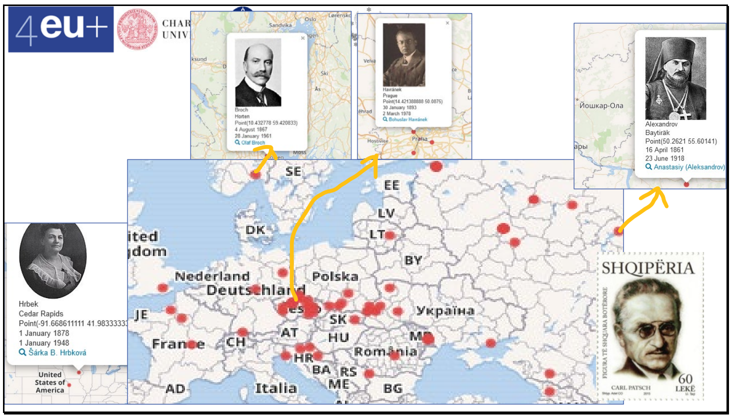
4eu+ CHARLES UNIVERSITY SORBONNE UNIVERSITÉ UNIVERSITY OF WARSAW

Table

person	personLabel	dob	dod	placeBirthLabel
wd:Q4469268	Zinaida Udalcova	5 March 1918	29 September 1987	Kislovodsk
wd:Q2361662	Dmitry Abramovich	7 August 1873	4 March 1955	Hulivka
wd:Q4064891	Anastasiy (Aleksandrov)	16 April 1861	23 June 1918	Baytiräk
wd:Q4069303	Fedor Aristov	26 October 1888	5 November 1932	Varnavino
wd:Q112548238	James Daniel Armstrong	1 January 1942	1 January 1979	Kansas City
wd:Q2637042	Artemy Artsikhovsky	26 December 1902	17 February 1978	Saint Petersburg
wd:Q7476739	Ioan Bogdan	25 July 1864	1 June 1919	Şcheii Braşovului
wd:Q2656990	Olaf Broch	4 August 1867	28 January 1961	Horten
wd:Q12084870	Ivan Bryk	8 July 1879	17 September 1947	Ustrzyki Dolne
wd:Q12084870	Ivan Bryk	8 July 1879	17 September 1947	Ustrzyki Dolne
wd:Q4097652	Nicolai von Bubnov	7 January 1880	4 August 1962	Saint Petersburg
wd:Q4097652	Nicolai von Bubnov	7 January 1880	4 August 1962	Saint Petersburg

The results in the alphabetical order of surnames. Error in the entry of Zinaida Udalcova. She appears first because her surname is missing in the entry. After her all are alphabetically sorted. Duplicates are annoying, due to small differences in GPS coordinates in different language versions of WikiData. You apparently cannot simply say *unique person ID* in SPARQL.

Slide 9



WikiData Query Service offers some plotting options beside table. Map with details and optional images (when available)

Slide 10

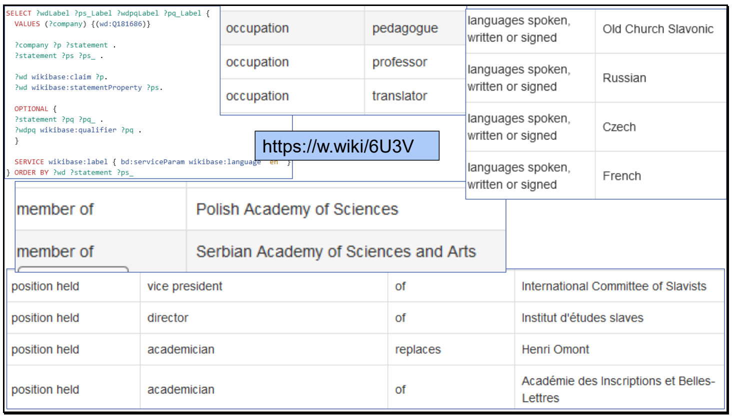
Kucharski	Eugeniusz Kucharski	Drohobych	12 December 1880	12 August 1952
Mach	Otto Mach	Brněnec	20 September 1917	25 December 1965
Malý	Jaroslav Malý	Daruvar	1 January 1907	1 January 1945
Manning	Clarence Manning	New York City	1 April 1893	4 October 1972

Mazon is missing here!!!!

Meillet	Antoine Meillet	Moulins	11 November 1866	21 September 1936
Mladenov	Stefan Mladenov	Vidin	27 December 1880	1 May 1963
Niederle	Lubor Niederle	Klatovy	20 September 1865	14 June 1944
Oblak	Vatroslav Oblak	Celje	15 May 1864	15 April 1896

The author(s) of the Mazon WikiData entries did not use the property field of work (P101). Nor did they use any other label that would have explicitly something to do with Slavic studies

Slide 11



This query retrieves all properties and their values associated with the given item. The relevant ones are displayed here – no explicit mention of Slavic studies or philology. Now imagine that you could go through the real Wikipedia and automatically complete the missing properties.

Slide 12

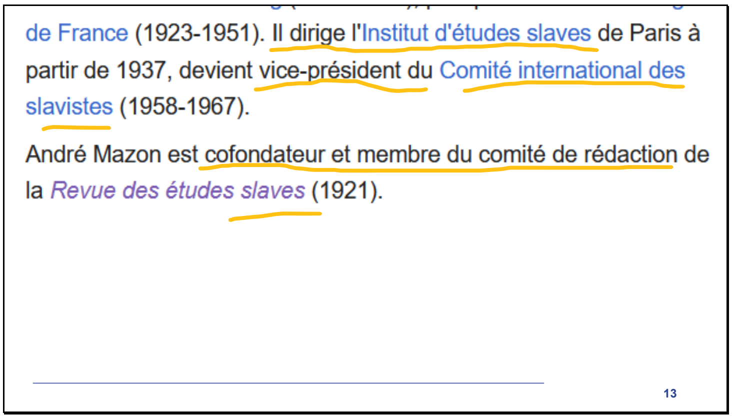
André Mazon

Article Discussion Lire

André Mazon (André Auguste Mazon), né le 7 septembre 1881 à Paris 2e et mort le 13 juillet 1967 dans le 15e arrondissement de Paris[1], est un slaviste français, professeur au Collège de France (1923) et membre de l'Académie des inscriptions et belles-lettres (1941). Ses travaux portent sur la littérature en slavon et en russe classique, sur la langue russe et la langue tchèque, ainsi que sur le folklore slave.

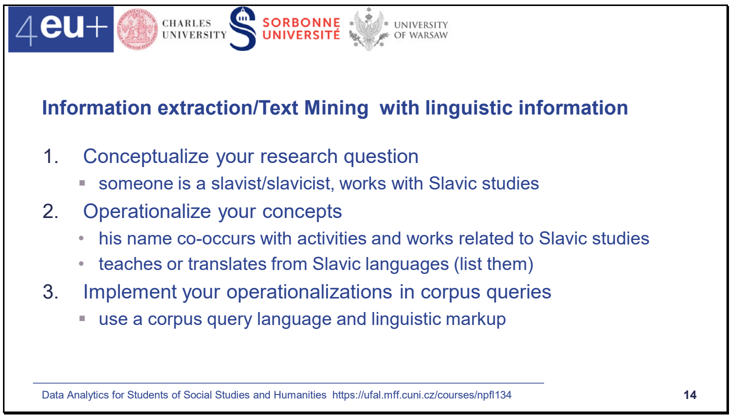
But when you read the more verbose entry on Wikipedia, you immediately understand that Mazon was a slavist. Explicitly said and also some implicit hints.

Slide 13



With Jirka, you will learn how to formulate such templates with a corpus query language, next lesson. I am going to tell you more about the information extraction strategies and the currently most common markup.

Slide 14



To extract conceptual information from unstructured text, you will have to rely on linguistic structures: how do people usually/typically refer to a concept? Guesswork with evidence.

Slide 15

Information extraction with subsequent Machine Learning

- Baroni, M., Murphy, B., Barbu, E., & Poesio, M. (2010). Strudel: A Corpus-Based Semantic Model Based on Properties and Types. *Cognitive Science*, *34*(2), 222–254. https://doi.org/10.1111/j.1551-6709.2009.01068.x

"Petr Novák, Wikipedia".

Table 1
Examples of input and output of the Strudel pattern template component

Input	Output	Notes
Layer from an onion	P_from_a_C	*an* normalized to *a*
Layers in a red onion	P_in_a_ADJ_C	*red* mapped to *ADJ*
Onion with different layers	C_with_different_P	Frequent adj *different* preserved
Onions and with their layers	Ø	Conjunction blocks pattern extraction

Two projects a decade ago: they populate a knowledge base with templates made on a very large corpus. Strudel: structured dimension extraction and labeling.
Property P, Concept C. They started with a number of nouns and wrote templates to capture context that could help characterize each noun.

Slide 16

CHARLES UNIVERSITY SORBONNE UNIVERSITÉ UNIVERSITY OF WARSAW

Table 2
Examples of Strudel output with type sketches

Concept	Property	Log-likelihood	Type Sketch
child	parent-n	11,726.7	P_of_C (40%), P_with_C (11%)
child	parent-v	120.8	P_C (79%)
lion	mane-n	259.1	C_'s_P (50%), C_with_P (15%), C_have_P (12%), P_of_C (10%
wolf	forest-n	78.3	C_in_P (32%), P_of_C (31%), C_through_P (14%)
wolf	pack-n	251.2	P_of_C (70%), C_in_P (15%)
egg	female-n	1,603.4	P_produce_C (13%), C_by_P (12%)
breakfast	croissant-n	257.2	P_for_C (46%), C_of_P (34%), C_with_P (12%)
beach	walk-v	687.6	P_C (29%), P_from_C (24%), P_along_C (23%), P_on_C (13%)
grass	green-a	277.6	P_C (58%), C_is_P (25%), C_is_ADV_P (16%)

Data Analytics for Students of Social Studies and Humanities
https://ufal.mff.cuni.cz/courses/npfl134 16

For each unique *property collocate* they computed how typical it was for the given noun (compared to all other nouns, using the *log-likelihood ratio*).

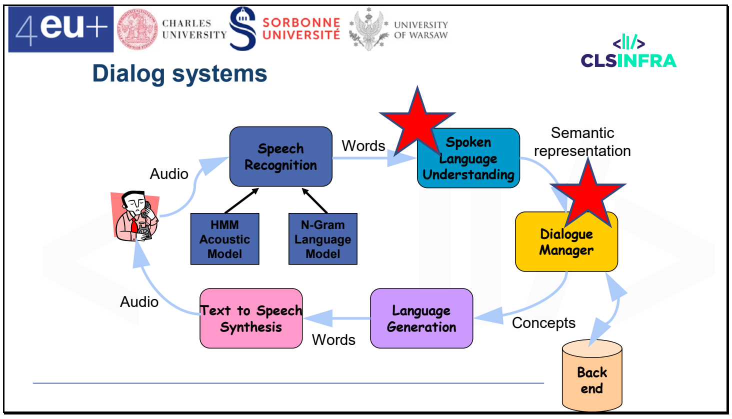
Slide 17


Read the Web
Research Project at Carnegie Mellon University
https://rtw.ml.cmu.edu/rtw/
Home
Project Overview
Resources & Data
Publications
People
NELL: Never-Ending Language Learning
Can computers learn to read? We think so. "Read the Web" is a research project that attempts to create a computer system that learns over time to read the web. Since January 2010, our computer system called NELL (Never-Ending Language Learner) has been running continuously, attempting to perform two tasks each day:
• First, it attempts to "read," or extract facts from text found in hundreds of millions of web pages (e.g., playsInstrument(George_Harrison, guitar)).
• Second, it attempts to improve its reading competence, so that tomorrow it can extract more facts from the web, more accurately.
Browse the Knowledge Base!
So far, NELL has accumulated over 50 million candidate beliefs by reading the web, and it is considering these at different levels of confidence. NELL has high confidence in 2,810,379 of these beliefs — these are displayed on this website. It is not perfect, but NELL is learning. You can track NELL's progress below or @cmunell on Twitter, browse and download its knowledge base, read more about our technical approach, or join the discussion group.
17

Slide 18

instance	iteration	date learned	confidence
blyth_s_hornbill is a bird	1111	06-jul-2018	100.0
test_plants is a plant	1111	06-jul-2018	99.7
fion_lim is a chef	1111	06-jul-2018	96.3
restaurant_breakfast is a visualizable thing	1111	06-jul-2018	96.8
disney_s_fairies_magazine is a magazine	1111	06-jul-2018	99.9
michael is a person who moved to the state pennsylvania	1113	15-aug-2018	93.8
standard_chartered is a bank in china	1114	25-aug-2018	96.9
salmon is a fish that can be served with the food introduction in a meal (or dish)	1116	12-sep-2018	99.9
majestic_sierra_nevada is a mountain in the state or province california	1116	12-sep-2018	93.8
rafael_nadal is an athlete who wins roland_garros	1116	12-sep-2018	99.9

Slide 19



information extraction is a crucial element in dialog systems: developers write templates that capture what the computer is supposed to watch out for hearing.

Slide 20

Semantic grammar PHOENIX

- Grammar #1:

ORIGIN_CITY → [from | beginning in] [Atlanta | Pittsburgh | Boston | …]

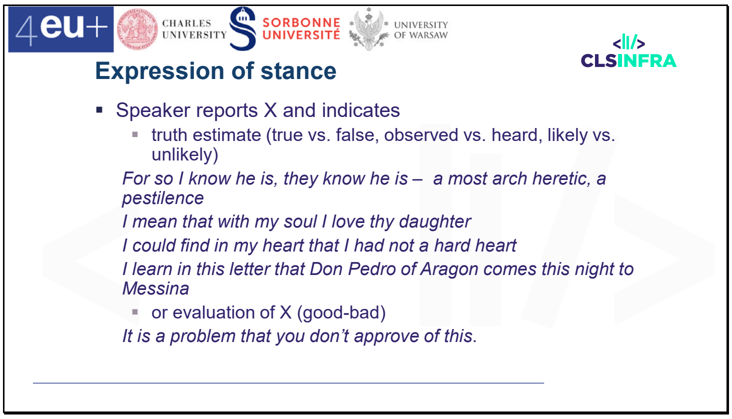
- Grammar #2:

DEPARTURE_TIME → [leaving at | on] TIME_EXPRESSION
TIME_EXPRESSION → [DAY_OF_WEEK]
TIME_EXPRESSION → [DAY_OF_WEEK] [TIME_OF_DAY]

Slide 21

Information extraction from non-fiction: usually content. In other contexts, style can be more interesting.
Interesting: style + pragmatics (content + form, context)
Biber: investigating linguistic variation in texts. Extracted 67 English linguistic patterns (e. g. past tense, perfect tense, definite noun) from 481 texts across genres, also spoken.
Features for co-occurrence clusters: passive and nominalizations vs. 2nd person + contracted verb forms
Each text got a score for each feature according to feature frequency per 100 words – multidimensional space, features clustered – statistical reduction of the dimensions.
When you have that, you can say about an unknown text to which text genres or registers it is similar (e. g. this is probably an academic text by style).

Slide 22


4eu+
CHARLES UNIVERSITY
SORBONNE UNIVERSITÉ
UNIVERSITY OF WARSAW
CLSINFRA
Expression of stance
Speaker reports X and indicates
truth estimate (true vs. false, observed vs. heard, likely vs. unlikely)
For so I know he is, they know he is – a most arch heretic, a pestilence
I mean that with my soul I love thy daughter
I could find in my heart that I had not a hard heart
I learn in this letter that Don Pedro of Aragon comes this night to Messina
or evaluation of X (good-bad)
It is a problem that you don't approve of this.

Slide 23

Narrativity

- + simple past tense
- - 2nd person
- + past/present progressive tense
- - simple present tense
- - passive voice

Slide 24

Descriptivity

- + adjectives in attributive positions
- + relative clauses
- + copula predicates
- + present tense
- - progressive tense
- - modal verbs

Slide 25

Interactivity

- 2nd person
- questions
- vocatives
- imperatives

Slide 26

Uncertainty or distance

- + hedge expressions (*maybe, basically, a bit*)
- + indefinite pronouns (*some, any*)
- + some modal verbs (*can, may*)
- + conditional markers (*would, if, when, whether*)

Slide 27

4eu+ CHARLES UNIVERSITY SORBONNE UNIVERSITÉ UNIVERSITY OF WARSAW

CLSINFRA

Emotionality

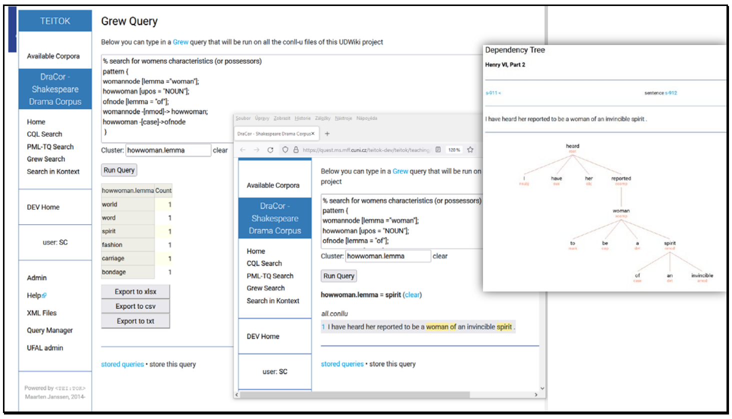
- + interjections
- + exclamation marks

- Shakespeare: short lines by one speaker – one verse in his iambic pentameter is comprised of several speakers’ lines

Slide 28

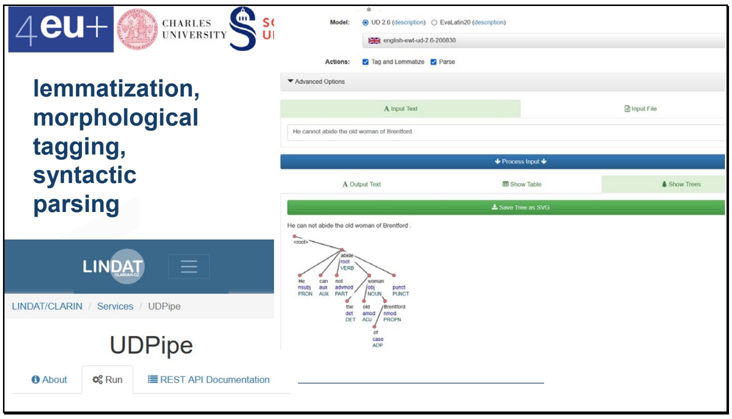
You collect the text of interest into a corpus and query the corpus things you want to know. Either you read the matches individually, or you extract them in a big amount and further process to make some automatic decisions. Like here, we are trying to find out how Shakespeare characterized women and we had known, that the attribute could be expressed by of and something coming after it. (besides adjective before woman of course).

Slide 29



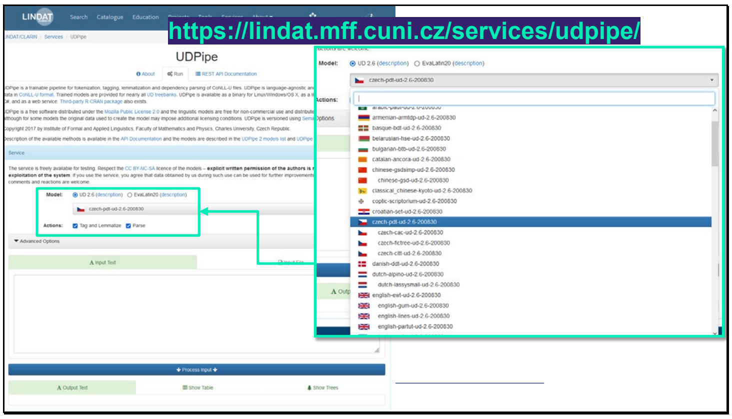
With full linguistic markup, you can abstract from the word order and grab the noun governed by the preposition *of*. You can also say that you do not want to match proper nouns after *of*.
Tree query language: you write a query and can see all matches. You can display them all at once or ask for an aggregation and then select which to view. You can even view (and edit) the syntactic trees.

Slide 30



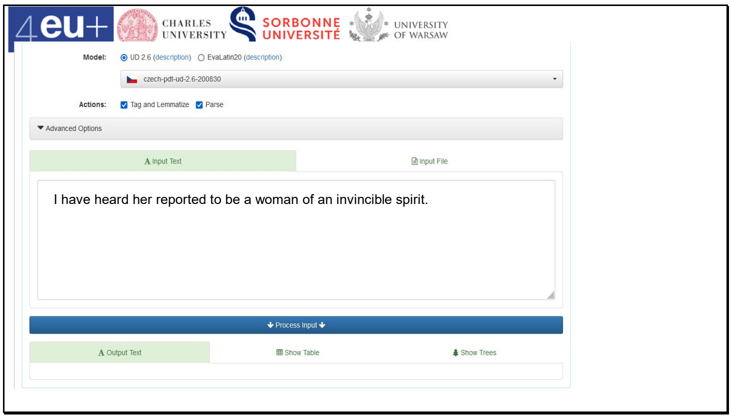
With the morphological markup, each sentence is a tree diagram in which you can see that each word syntactically depends as “child” on another word – its “parent”. The parent is modified by the child; e.g., *warm* modifies *weather;*
in *chair of wood*, *wood* modifies *chair;* in *read a book* the *book* modifies *read*.
Warm is an adjectival attribute of *weather*, *book* is a direct object of *read*. When *Peter reads a book*, *Peter* is the nominal subject of *read*. These relations are encoded as labels on the children and denote their syntactic dependency on the parent. You can imagine them as the edges (the lines) in the graph. The words are obviously its nodes (the points). Each node can only have one parent. The main predicate is the top of the tree (it has the *root* label); and it hangs on a technical node. Each node stores some additional information: the actual word form, the lemma (dictionary form), the part of speech (noun, verb), and morphological details such as case and number in nouns and tense in verbs. These are called morphological features.

Slide 31



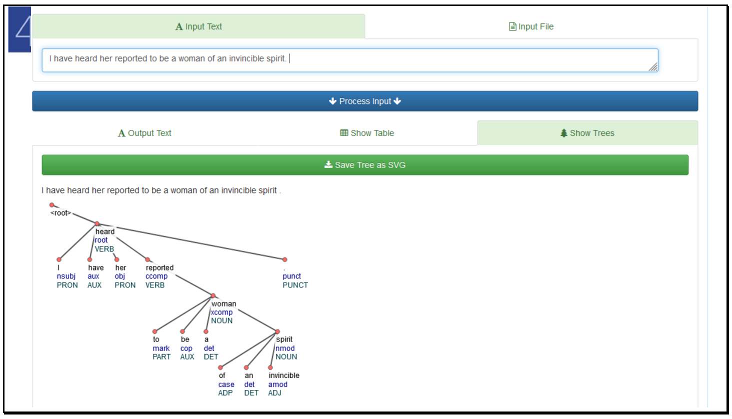
This is a modern formalism of language description that is universal across languages. It uses the same labels for parts of speech for all languages, and a common pool of features and their values. But these are selected and interpreted in a language-specific way. Some languages do not have cases in nouns, or just some , or they do not have gender in verbs, some languages have special polite forms in verbs or pronouns, etc. The authors struggle to keep the syntax universal, but some languages insist of their language-specific ways, sometimes with very solid arguments to do so. Anyway, it is much easier to handle than learning totally different tagsets and description principles for each language separately. We will use English examples, but the Russian, Italian, and French description would be very similar.

Slide 32



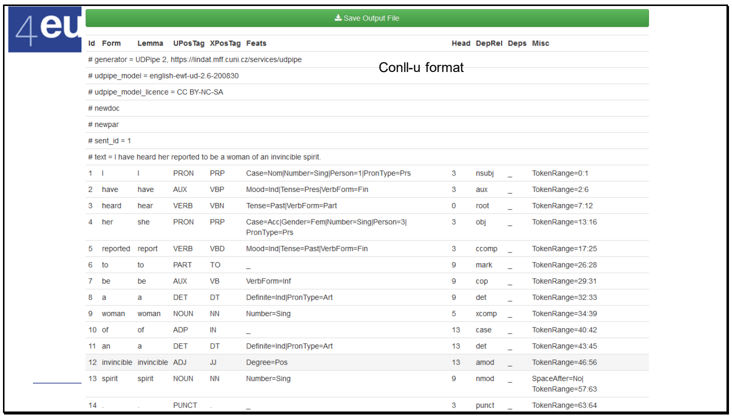
4eu+
CHARLES UNIVERSITY
SORBONNE UNIVERSITÉ
UNIVERSITY OF WARSAW
Model: UD 2.6 (description) EvaLatin20 (description)
czech-pdt-ud-2.6-200830
Actions: Tag and Lemmatize Parse
Advanced Options
Input Text
Input File
I have heard her reported to be a woman of an invincible spirit.
Process Input
Output Text
Show Table
Show Trees

Slide 33



Again a tree. Main predicate, subject, auxiliary verb always child of the full verb; in this formalism. This is just a visualization. In the reality, the format is plain text.

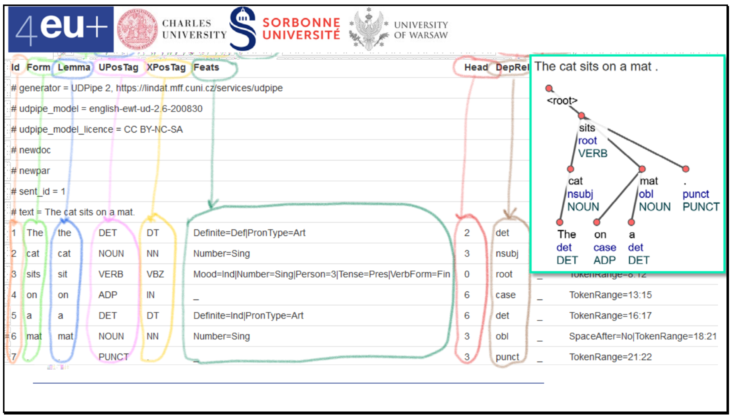
Slide 34



Id	Form	Lemma	UPosTag	XPosTag	Feats	Head	DepRel	Deps	Misc
# generator = UDPipe 2, https://lindat.mff.cuni.cz/services/udpipe									
# udpipe_model = english-ewt-ud-2.6-200830									
# udpipe_model_licence = CC BY-NC-SA									
# newdoc									
# newpar									
# sent_id = 1									
# text = I have heard her reported to be a woman of an invincible spirit.									
1	I	I	PRON	PRP	Case=Nom|Number=Sing|Person=1|PronType=Prs	3	nsubj	_	TokenRange=0:1
2	have	have	AUX	VBP	Mood=Ind|Tense=Pres|VerbForm=Fin	3	aux	_	TokenRange=2:6
3	heard	hear	VERB	VBN	Tense=Past|VerbForm=Part	0	root	_	TokenRange=7:12
4	her	she	PRON	PRP	Case=Acc|Gender=Fem|Number=Sing|Person=3|PronType=Prs	3	obj	_	TokenRange=13:16
5	reported	report	VERB	VBD	Mood=Ind|Tense=Past|VerbForm=Fin	3	ccomp	_	TokenRange=17:25
6	to	to	PART	TO	_	9	mark	_	TokenRange=26:28
7	be	be	AUX	VB	VerbForm=Inf	9	cop	_	TokenRange=29:31
8	a	a	DET	DT	Definite=Ind|PronType=Art	9	det	_	TokenRange=32:33
9	woman	woman	NOUN	NN	Number=Sing	5	xcomp	_	TokenRange=34:39
10	of	of	ADP	IN	_	13	case	_	TokenRange=40:42
11	an	a	DET	DT	Definite=Ind|PronType=Art	13	det	_	TokenRange=43:45
12	invincible	invincible	ADJ	JJ	Degree=Pos	13	amod	_	TokenRange=46:56
13	spirit	spirit	NOUN	NN	Number=Sing	9	nmod	_	SpaceAfter=No|TokenRange=57:63
14	.	.	PUNCT	.	_	3	punct	_	TokenRange=63:64

This is the actual format. Like a table with a few commented lines. Each row represents one word (token). ID (word order in the sentence) Form, lemma, Universal POS, traditional POS tag (just ignore), universal features; ID of the parent node and the syntactic relation to the parent. *I* is a child of *heard*. So is *have*, *her*, and *reported*.

Slide 35

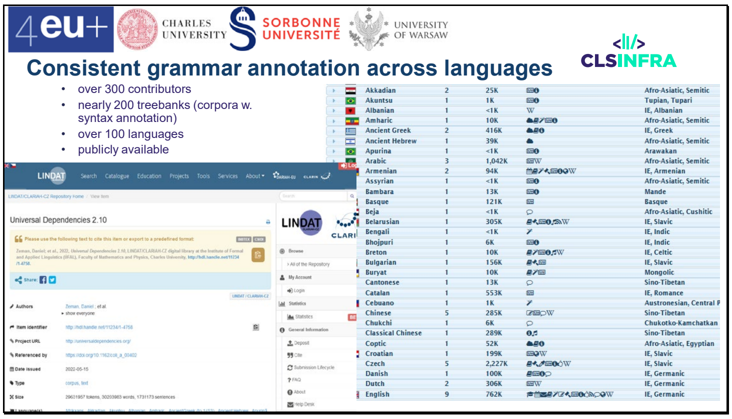


Once again, let’s compare the rows to the tree.

Slide 36

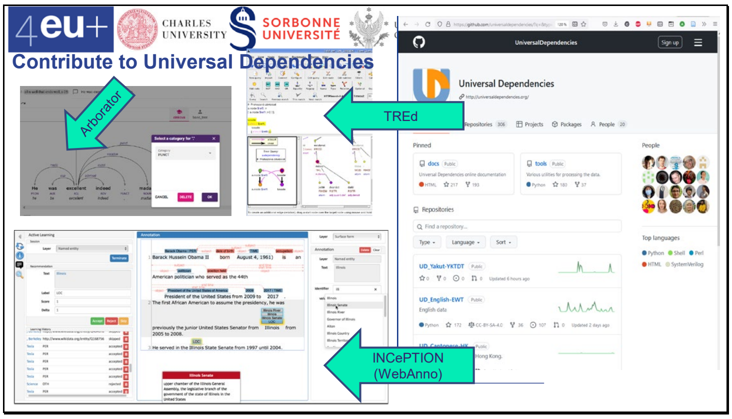

4eu+
CHARLES UNIVERSITY
SORBONNE UNIVERSITÉ
UNIVERSITY OF WARSAW
CLSINFRA
Universal Dependencies
universaldependencies.org

Slide 37



4eu+
CHARLES UNIVERSITY
SORBONNE UNIVERSITÉ
UNIVERSITY OF WARSAW
CLSINFRA
Consistent grammar annotation across languages
over 300 contributors
nearly 200 treebanks (corpora w. syntax annotation)
over 100 languages
publicly available
LINDAT
Universal Dependencies 2.10
Akkadian 2 25K Afro-Asiatic, Semitic
Akuntsu 1 1K Tupian, Tupari
Albanian 1 <1K IE, Albanian
Amharic 1 10K Afro-Asiatic, Semitic
Ancient Greek 2 416K IE, Greek
Ancient Hebrew 1 39K Afro-Asiatic, Semitic
Apurina 1 <1K Arawakan
Arabic 3 1,042K Afro-Asiatic, Semitic
Armenian 2 94K IE, Armenian
Assyrian 1 <1K Afro-Asiatic, Semitic
Bambara 1 13K Mande
Basque 1 121K Basque
Beja 1 <1K Afro-Asiatic, Cushitic
Belarusian 1 305K IE, Slavic
Bengali 1 <1K IE, Indic
Bhojpuri 1 6K IE, Indic
Breton 1 10K IE, Celtic
Bulgarian 1 156K IE, Slavic
Buryat 1 10K Mongolic
Cantonese 1 13K Sino-Tibetan
Catalan 1 553K IE, Romance
Cebuano 1 1K Austronesian, Central P
Chinese 5 285K Sino-Tibetan
Chukchi 1 6K Chukotko-Kamchatkan
Classical Chinese 1 289K Sino-Tibetan
Coptic 1 52K Afro-Asiatic, Egyptian
Croatian 1 199K IE, Slavic
Czech 5 2,227K IE, Slavic
Danish 1 100K IE, Germanic
Dutch 2 306K IE, Germanic
English 9 762K IE, Germanic

Slide 38


4eu+
CHARLES UNIVERSITY
SORBONNE UNIVERSITÉ
Contribute to Universal Dependencies
Arborator
TREd
INCePTION
(WebAnno)
Universal Dependencies

Slide 39

Universal Parts of Speech (upos)

UD Morphology

Slide 40

4eu+ CHARLES UNIVERSITY SORBONNE UNIVERSITÉ UNIVERSITY OF WARSAW

Morphological categories

- Universal Parts of Speech (**upos**)
  - NOUN, PROPN
  - VERB, AUX
  - ADJ, ADV
  - PRON, DET, NUM
  - SCONJ, CCONJ, ADP
  - PART, INTJ
  - PUNCT, SYM, X
- Universal Features (**feats**)
  - morphological categories relevant to the given upos

Slide 41

NOUN vs. PROPN vs. neither

strawberries	

Slide 42

CHARLES UNIVERSITY SORBONNE UNIVERSITÉ UNIVERSITY OF WARSAW

NOUN vs. PROPN vs. neither

strawberries	NOUN
cat	

Slide 43

NOUN vs. PROPN vs. neither

strawberries	NOUN
cat	NOUN
small	

Slide 44

NOUN vs. PROPN vs. neither

strawberries	NOUN
cat	NOUN
small	neither
Peter	

Slide 45

4eu+ CHARLES UNIVERSITY SORBONNE UNIVERSITÉ UNIVERSITY OF WARSAW

NOUN vs. PROPN vs. neither

strawberries	NOUN
cat	NOUN
small	neither
Peter	PROPN
butter	

Slide 46

NOUN vs. PROPN vs. neither

strawberries	NOUN
cat	NOUN
small	neither
Peter	PROPN
butter	NOUN
beer	

Slide 47

NOUN vs. PROPN vs. neither

strawberries	NOUN
cat	NOUN
small	neither
Peter	PROPN
butter	NOUN
beer	NOUN
Dutchman	

Slide 48

NOUN vs. PROPN vs. neither

strawberries	NOUN
cat	NOUN
small	neither
Peter	PROPN
butter	NOUN
beer	NOUN
Dutchman	PROPN
until	

Slide 49

NOUN
vs.
PROPN
vs.
neither

strawberries	NOUN
cat	NOUN
small	neither
Peter	PROPN
butter	NOUN
beer	NOUN
Dutchman	PROPN
until	neither

Slide 50

4eu+ CHARLES UNIVERSITY SORBONNE UNIVERSITÉ UNIVERSITY OF WARSAW

VERB vs. AUX vs. neither

are	

Slide 51

4eu+ CHARLES UNIVERSITY SORBONNE UNIVERSITÉ UNIVERSITY OF WARSAW

VERB vs. AUX vs. neither

are	AUX
can	

Slide 52

VERB vs. AUX vs. neither

are	AUX
can	AUX
(He) did (it)	

Slide 53

VERB vs. AUX vs. neither

are	AUX
can	AUX
(He) did (it)	VERB
Do (you smoke?)	

Slide 54

VERB vs. AUX vs. neither

are	AUX
can	AUX
(He) did (it)	VERB
Do (you smoke?)	AUX
(be) flying	

Slide 55

4eu+ CHARLES UNIVERSITY SORBONNE UNIVERSITÉ UNIVERSITY OF WARSAW

VERB vs. AUX vs. neither

are	AUX
can	AUX
(He) did (it)	VERB
Do (you smoke?)	AUX
(be) flying	VERB
(He) used (to swim)	

Slide 56

VERB vs. AUX vs. neither

are	AUX
can	AUX
(He) did (it)	VERB
Do (you smoke?)	AUX
(be) flying	VERB
(He) used (to swim)	VERB
(She is) going (to win.)	VERB
(You) ought (to smile).	VERB

Slide 57

VERB vs. AUX vs. neither

(a) winning (strategy)	

Slide 58

VERB vs. AUX vs. neither

(a) winning (strategy)	VERB
(a) rotting (tooth)	

Slide 59

VERB vs. AUX vs. neither

(a) winning (strategy)	VERB
(a) rotting (tooth)	VERB
(a) lost (war)	

Slide 60

4eu+ CHARLES UNIVERSITY SORBONNE UNIVERSITÉ UNIVERSITY OF WARSAW

VERB vs. AUX vs. neither

(a) winning (strategy)	VERB
(a) rotting (tooth)	VERB
(a) lost (war)	VERB
(a) rotten (tooth)	

Slide 61

VERB vs. AUX vs. neither

(a) winning (strategy)	VERB
(a) rotting (tooth)	VERB
(a) lost (war)	VERB
(a) rotten (tooth)	neither (adjective)
Let('s dance.)	

Slide 62

4eu+ CHARLES UNIVERSITY SORBONNE UNIVERSITÉ UNIVERSITY OF WARSAW

VERB vs. AUX vs. neither

(a) winning (strategy)	VERB
(a) rotting (tooth)	VERB
(a) lost (war)	VERB
(a) rotten (tooth)	neither (adjective)
Let(‘s dance.)	VERB
(She) wants (food)	

Slide 63

4eu+ CHARLES UNIVERSITY SORBONNE UNIVERSITÉ UNIVERSITY OF WARSAW

VERB vs. AUX vs. neither

(a) winning (strategy)	VERB
(a) rotting (tooth)	VERB
(a) lost (war)	VERB
(a) rotten (tooth)	neither (adjective)
Let('s dance.)	VERB
(She) wants (food)	VERB
(She) wants (to win)	VERB
(He) became (professor)	

Slide 64

4eu+ CHARLES UNIVERSITY SORBONNE UNIVERSITÉ UNIVERSITY OF WARSAW

VERB vs. AUX vs. neither

(a) winning (strategy)	VERB
(a) rotting (tooth)	VERB
(a) lost (war)	VERB
(a) rotten (tooth)	neither (adjective)
Let('s dance.)	VERB
(She) wants (food)	VERB
(She) wants (to win)	VERB
(He) became (professor)	VERB

Slide 65

4eu+ CHARLES UNIVERSITY SORBONNE UNIVERSITÉ UNIVERSITY OF WARSAW

ADJ vs. ADV vs. neither

green	

Slide 66

ADJ vs. ADV vs. neither

green	ADJ
happily	

Slide 67

ADJ vs. ADV vs. neither

green	ADJ
happily	ADV
my	

Slide 68

ADJ vs. ADV vs. neither

green	ADJ
happily	ADV
my	neither
many	

Slide 69

4eu+ CHARLES UNIVERSITY SORBONNE UNIVERSITÉ UNIVERSITY OF WARSAW

ADJ vs. ADV vs. neither

green	ADJ
happily	ADV
my	neither
many	ADJ
oldest	

Slide 70

CHARLES UNIVERSITY
SORBONNE UNIVERSITÉ
UNIVERSITY OF WARSAW

ADJ vs. ADV vs. neither

green	ADJ
happily	ADV
my	neither
many	ADJ
oldest	ADJ
(the) third (year)	

Slide 71

ADJ vs. ADV vs. neither

green	ADJ
happily	ADV
my	neither
many	ADJ
oldest	ADJ
(the) third (year)	ADJ
(the) poor	

Slide 72

4eu+ CHARLES UNIVERSITY SORBONNE UNIVERSITÉ UNIVERSITY OF WARSAW

ADJ vs. ADV vs. neither

green	ADJ
happily	ADV
my	neither
many	ADJ
oldest	ADJ
(the) third (year)	ADJ
(the) poor	ADJ
where	

Slide 73

ADJ vs. ADV vs. neither

green	ADJ
happily	ADV
my	neither
many	ADJ
oldest	ADJ
(the) third (year)	ADJ
(the) poor	ADJ
where	ADV

Slide 74

ADJ vs. ADV vs. neither

twice	ADV
(take) off (phrasal verb)	

Slide 75

ADJ vs

twice	ADV
(take) off (phrasal verb)	neither
(write) down	

Slide 76

ADJ vs. ADV vs. neither

twice	ADV
(take) off (phrasal verb)	neither
(write) down	ADV
sometime	

Slide 77

ADJ vs. ADV vs. neither

twice	ADV
(take) off (phrasal verb)	neither
(write) down	ADV
sometime	ADV
yes	

Slide 78

4eu+ CHARLES UNIVERSITY SORBONNE UNIVERSITÉ UNIVERSITY OF WARSAW

ADJ vs. ADV vs. neither

twice	ADV
(take) off (phrasal verb)	neither
(write) down	ADV
sometime	ADV
yes	neither
none	

Slide 79

ADJ vs. ADV vs. neither

twice	ADV
(take) off (phrasal verb)	neither
(write) down	ADV
sometime	ADV
yes	neither
none	neither
how	

Slide 80

ADJ vs. ADV vs. neither

twice	ADV
(take) off (phrasal verb)	neither
(write) down	ADV
sometime	ADV
yes	neither
none	neither
how	ADV

Slide 81

4eu+ CHARLES UNIVERSITY SORBONNE UNIVERSITÉ UNIVERSITY OF WARSAW

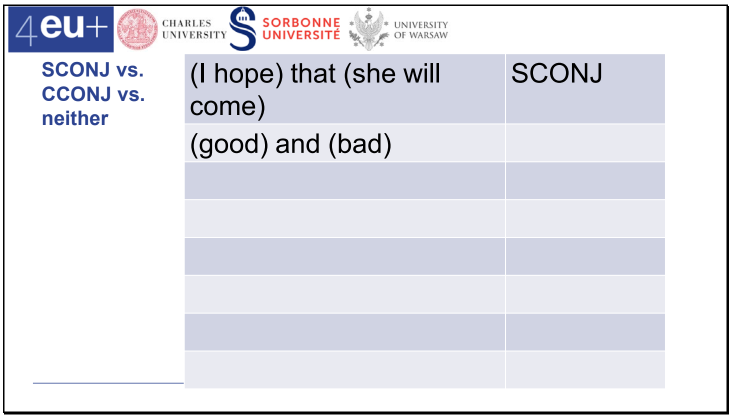
ADJ vs. ADV vs. neither

twice	ADV
(take) off (phrasal verb)	neither
(write) down	ADV
sometime	ADV
yes	neither
none	neither
how	ADV
twice	ADV

Slide 82

Subordinating conjunctions link constructions by making one of them a constituent of the other (e. g. an attribute, an adverbial...). Coordinating conjunctions links words or larger constituents and expresses a semantic relation between them (and, but, or)

Slide 83



Subordinating conjunctions link constructions by making one of them a constituent of the other (e. g. an attribute, an adverbial...). Coordinating conjunctions links words or larger constituents and expresses a semantic relation between them (and, but, or)

Slide 84

SCONJ vs. CCONJ vs. neither

(I hope) that (she will come)	SCONJ
(good) and (bad)	CCONJ
(nobody) but (you)	

Subordinating conjunctions link constructions by making one of them a constituent of the other (e. g. an attribute, an adverbial...). Coordinating conjunctions links words or larger constituents and expresses a semantic relation between them (and, but, or)

Slide 85

SCONJ vs. CCONJ vs. neither

(I hope) that (she will come)	SCONJ
(good) and (bad)	CCONJ
(nobody) but (you)	CCONJ
(this) or (that)	

Subordinating conjunctions link constructions by making one of them a constituent of the other (e. g. an attribute, an adverbial...). Coordinating conjunctions links words or larger constituents and expresses a semantic relation between them (and, but, or)

Slide 86

4eu+ CHARLES UNIVERSITY SORBONNE UNIVERSITÉ UNIVERSITY OF WARSAW

SCONJ vs. CCONJ vs. neither

(I hope) that (she will come)	SCONJ
(good) and (bad)	CCONJ
(nobody) but (you)	CCONJ
(this) or (that)	CCONJ
(this or) that	

Subordinating conjunctions link constructions by making one of them a constituent of the other (e. g. an attribute, an adverbial...). Coordinating conjunctions links words or larger constituents and expresses a semantic relation between them (and, but, or)

Slide 87

SCONJ vs. CCONJ vs. neither

(I hope) that (she will come)	SCONJ
(good) and (bad)	CCONJ
(nobody) but (you)	CCONJ
(this) or (that)	CCONJ
(this or) that	neither
(I know) which (to take)	

Subordinating conjunctions link constructions by making one of them a constituent of the other (e. g. an attribute, an adverbial...). Coordinating conjunctions links words or larger constituents and expresses a semantic relation between them (and, but, or)

Slide 88

SCONJ vs. CCONJ vs. neither

(I hope) that (she will come)	SCONJ
(good) and (bad)	CCONJ
(nobody) but (you)	CCONJ
(this) or (that)	CCONJ
(this or) that	neither
(I know) which (to take)	neither
(He left,) which (made her sad)	

Subordinating conjunctions link constructions by making one of them a constituent of the other (e. g. an attribute, an adverbial...). Coordinating conjunctions links words or larger constituents and expresses a semantic relation between them (and, but, or)

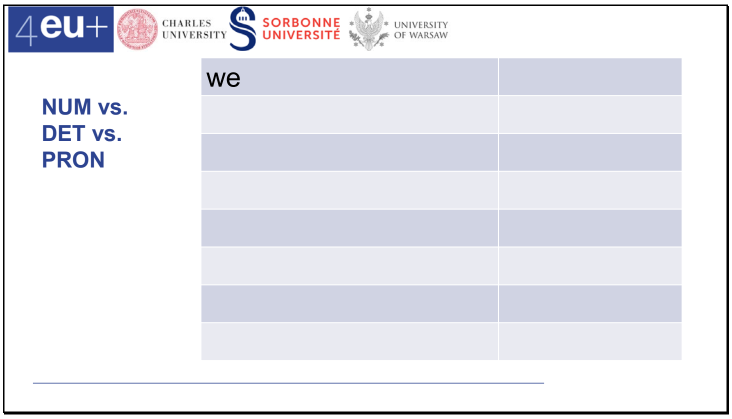
Slide 89

SCONJ vs. CCONJ vs. neither

(I hope) that (she will come)	SCONJ
(good) and (bad)	CCONJ
(nobody) but (you)	CCONJ
(this) or (that)	CCONJ
(this or) that	neither
(I know) which (to take)	neither
(He left,) which (made her sad)	neither
(Ask) whether (we may leave)	SCONJ

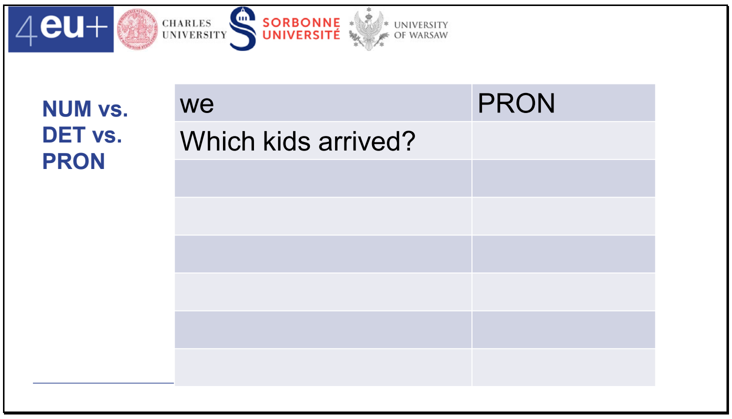
Subordinating conjunctions link constructions by making one of them a constituent of the other (e. g. an attribute, an adverbial...). Coordinating conjunctions links words or larger constituents and expresses a semantic relation between them (and, but, or)

Slide 90



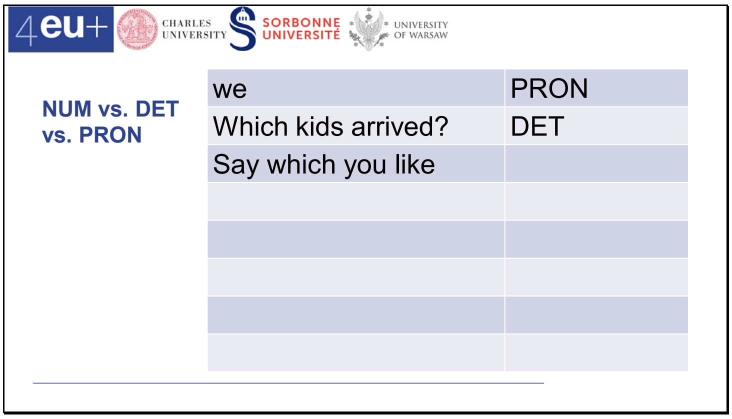
Pronouns are substitutes for nouns our noun phrases, so they should function like nouns. NOT those functioning like adjectives. These are tagged as determiners. Uhm... English breaks this. Possessive pronouns are PRON.
Pronouns do not act as adjectives, when they substitute a noun, even if they are relative pronouns (many languages use adjectival pronouns there, such as *welcher, kotoryj*)

Slide 91



Pronouns are substitutes for nouns our noun phrases, so they should function like nouns. NOT those functioning like adjectives. These are tagged as determiners. Uhm... English breaks this. Possessive pronouns are PRON.
Pronouns do not act as adjectives, when they substitute a noun, even if they are relative pronouns (many languages use adjectival pronouns there, such as *welcher, kotoryj*)

Slide 92



Pronouns are substitutes for nouns our noun phrases, so they should function like nouns. NOT those functioning like adjectives. These are tagged as determiners. Uhm... English breaks this. Possessive pronouns are PRON.
Pronouns do not act as adjectives, when they substitute a noun, even if they are relative pronouns (many languages use adjectival pronouns there, such as *welcher, kotoryj*)

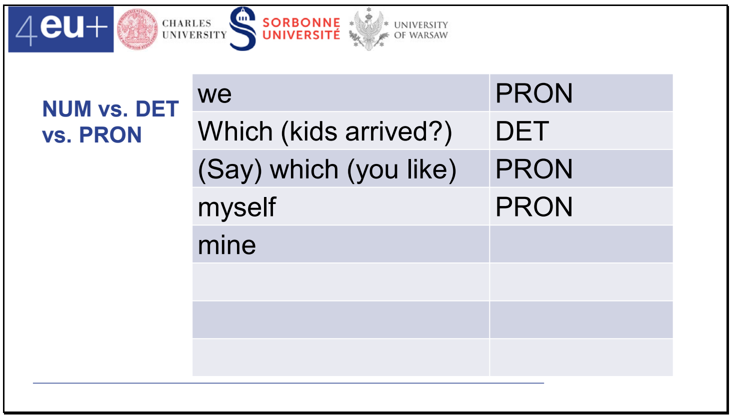
Slide 93

NUM vs. DET vs. PRON

we	PRON
Which (kids arrived?)	DET
(Say) which (you like)	PRON
myself	

Pronouns are substitutes for nouns our noun phrases, so they should function like nouns. NOT those functioning like adjectives. These are tagged as determiners. Uhm... English breaks this. Possessive pronouns are PRON.
Pronouns do not act as adjectives, when they substitute a noun, even if they are relative pronouns (many languages use adjectival pronouns there, such as *welcher, kotoryj*)

Slide 94



Pronouns are substitutes for nouns our noun phrases, so they should function like nouns. NOT those functioning like adjectives. These are tagged as determiners. Uhm... English breaks this. Possessive pronouns are PRON.
Pronouns do not act as adjectives, when they substitute a noun, even if they are relative pronouns (many languages use adjectival pronouns there, such as *welcher, kotoryj*)

Slide 95

NUM vs. DET vs. PRON

we	PRON
Which (kids arrived?)	DET
(Say) which (you like)	PRON
myself	PRON
mine, yours	PRON
my, your, his	

Pronouns are substitutes for nouns our noun phrases, so they should function like nouns. NOT those functioning like adjectives. These are tagged as determiners. Uhm... English breaks this. Possessive pronouns are PRON.
Pronouns do not act as adjectives, when they substitute a noun, even if they are relative pronouns (many languages use adjectival pronouns there, such as *welcher, kotoryj*)

Slide 96

NUM vs. DET vs. PRON

we	PRON
Which (kids arrived?)	DET
(Say) which (you like)	PRON
myself	PRON
mine, yours	PRON
my, your, his	PRON
every	

Pronouns are substitutes for nouns our noun phrases, so they should function like nouns. NOT those functioning like adjectives. These are tagged as determiners. Uhm... English breaks this. Possessive pronouns are PRON.
Pronouns do not act as adjectives, when they substitute a noun, even if they are relative pronouns (many languages use adjectival pronouns there, such as *welcher, kotoryj*)

Slide 97

NUM vs. DET vs. PRON

we	PRON
Which (kids arrived?)	DET
(Say) which (you like)	PRON
myself	PRON
mine, yours	PRON
my, your, his	PRON
every	DET
no (man)	

Pronouns are substitutes for nouns our noun phrases, so they should function like nouns. NOT those functioning like adjectives. These are tagged as determiners. Uhm... English breaks this. Possessive pronouns are PRON.
Pronouns do not act as adjectives, when they substitute a noun, even if they are relative pronouns (many languages use adjectival pronouns there, such as *welcher, kotoryj*)

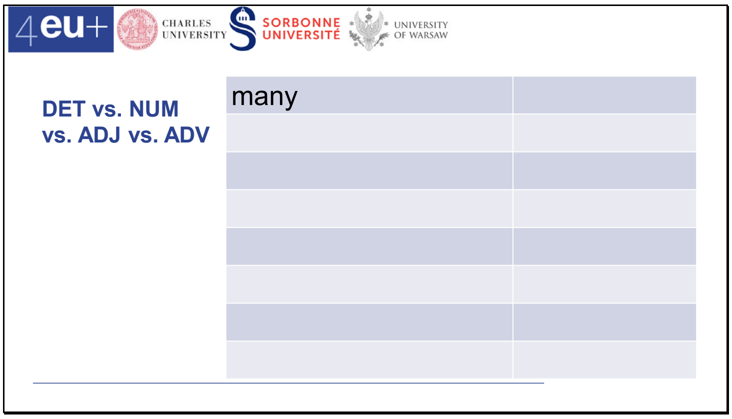
Slide 98

DET vs. PRON

we	PRON
Which (kids arrived?)	DET
(Say) which (you like)	PRON
myself	PRON
mine, yours	PRON
my, your, his	PRON
every	DET
no (man)	DET

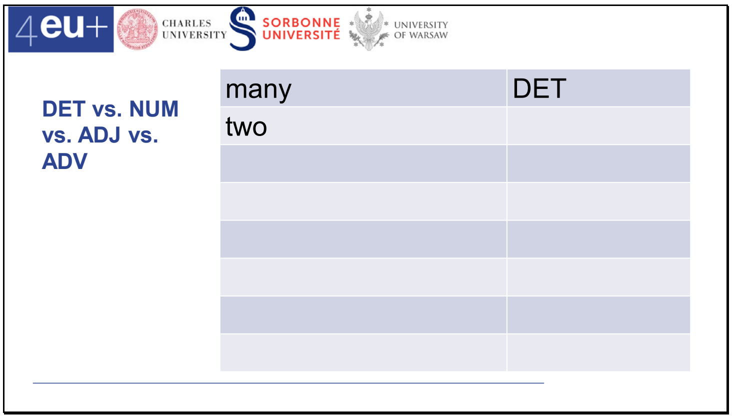
Pronouns are substitutes for nouns our noun phrases, so they should function like nouns. NOT those functioning like adjectives. These are tagged as determiners. Uhm... English breaks this. Possessive pronouns are PRON.
Pronouns do not act as adjectives, when they substitute a noun, even if they are relative pronouns (many languages use adjectival pronouns there, such as *welcher, kotoryj*)

Slide 99



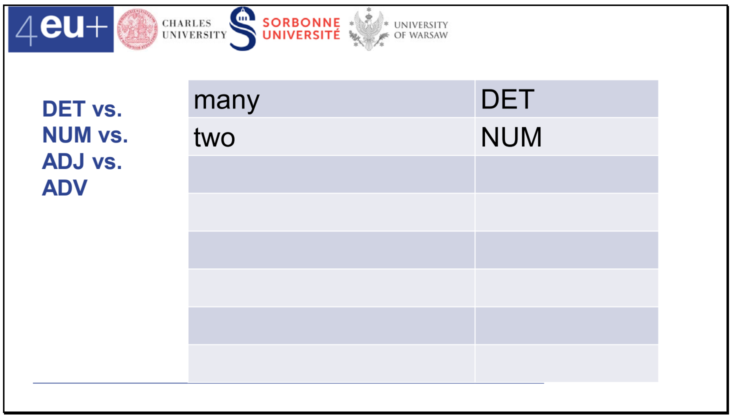
Numerals express numbers and a relation to the number, e.g. quantity, sequence, frequency, or fraction. Cardinal numbers: NUM. Ordinal numbers: ADJ

Slide 100



Numerals express numbers and a relation to the number, e.g. quantity, sequence, frequency, or fraction. Cardinal numbers: NUM. Ordinal numbers: ADJ

Slide 101



Numerals express numbers and a relation to the number, e.g. quantity, sequence, frequency, or fraction. Cardinal numbers: NUM. Ordinal numbers: ADJ

Slide 102

DET vs. NUM vs. ADJ vs. ADV

many	DET
two	NUM
first (minute)	

Numerals express numbers and a relation to the number, e. g. quantity, sequence, frequency, or fraction. Cardinal numbers: NUM. Ordinal numbers: ADJ

DET vs. NUM vs. ADJ vs. ADV

many	DET
two	NUM
first (minute)	ADJ
last (minute)	

Numerals express numbers and a relation to the number, e. g. quantity, sequence, frequency, or fraction. Cardinal numbers: NUM. Ordinal numbers: ADJ

Slide 104

DET vs. NUM vs. ADJ vs. ADV

many	DET
two	NUM
first (minute)	ADJ
last (minute)	ADJ
one (man)	

Numerals express numbers and a relation to the number, e. g. quantity, sequence, frequency, or fraction. Cardinal numbers: NUM. Ordinal numbers: ADJ

Slide 105

DET vs. NUM vs. ADJ vs. ADV

many	DET
two	NUM
first (minute)	ADJ
last (minute)	ADJ
one (man)	ADJ
(Charles) IV	

Numerals express numbers and a relation to the number, e. g. quantity, sequence, frequency, or fraction. Cardinal numbers: NUM. Ordinal numbers: ADJ

DET vs. NUM vs. ADJ vs. ADV

many	DET
two	NUM
first (minute)	ADJ
last (minute)	ADJ
one (man)	ADJ
(Charles) IV	NUM
both (men)	

Numerals express numbers and a relation to the number, e. g. quantity, sequence, frequency, or fraction. Cardinal numbers: NUM. Ordinal numbers: ADJ

Slide 107

DET vs. NUM vs. ADJ vs. ADV

many	DET
two	NUM
first (minute)	ADJ
last (minute)	ADJ
one (man)	ADJ
(Charles) IV	NUM
both (men)	DET
twice	

Numerals express numbers and a relation to the number, e. g. quantity, sequence, frequency, or fraction. Cardinal numbers: NUM. Ordinal numbers: ADJ

Slide 108

4eu+ CHARLES UNIVERSITY SORBONNE UNIVERSITÉ UNIVERSITY OF WARSAW

DET vs. NUM vs. ADJ vs. ADV

many	DET
two	NUM
first (minute)	ADJ
last (minute)	ADJ
one (man)	ADJ
(Charles) IV	NUM
both (men)	DET
twice	ADV

Numerals express numbers and a relation to the number, e. g. quantity, sequence, frequency, or fraction. Cardinal numbers: NUM. Ordinal numbers: ADJ

Slide 109

CHARLES UNIVERSITY SORBONNE UNIVERSITÉ UNIVERSITY OF WARSAW

ADP vs. ADV vs. SCONJ

for (you)	

Slide 110

ADP vs. ADV vs. SCONJ

for (you)	ADP
(forgive me), for (I have done wrong)	

Slide 111

for (you)	ADP
(forgive me), for (I have done wrong)	SCONJ
ago	

Slide 112

ADP vs. ADV vs. SCONJ

for (you)	ADP
(forgive me), for (I have done wrong)	SCONJ
ago	ADV
in	

Slide 113

ADP vs. ADV vs. SCONJ

for (you)	ADP
(forgive me), for (I have done wrong)	SCONJ
ago	ADV
in	ADP
towards	

Slide 114

ADP vs. ADV vs. SCONJ

for (you)	ADP
(forgive me), for (I have done wrong)	SCONJ
ago	ADV
in	ADP
towards	ADP
upwards	ADV
as/like (a teacher)	

Slide 115

ADP vs. ADV vs. SCONJ

for (you)	ADP
(forgive me), for (I have done wrong)	SCONJ
ago	ADV
in	ADP
towards	ADP
upwards	ADV
as/like (a teacher)	ADP
(call) as (you go)	

Slide 116

ADP vs. ADV vs. SCONJ

for (you)	ADP
(forgive me), for (I have done wrong)	SCONJ
ago	ADV
in	ADP
towards	ADP
upwards	ADV
as/like (a teacher)	ADP
(call) as (you go)	SCONJ

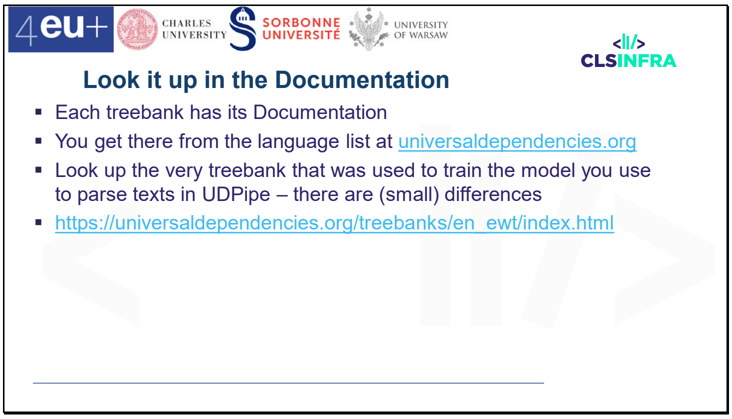
Slide 117

A trash bin in most languages. English and other Germanic languages: not particles from phrasal verbs!

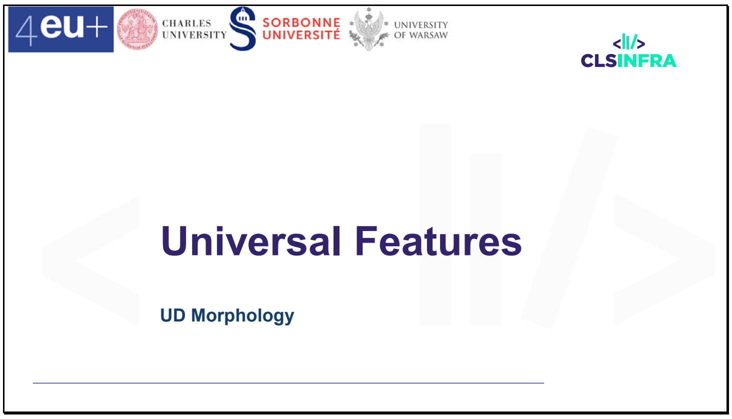
Slide 118

Exclamations, performative expressions, but not nouns: God, Thanks

Slide 119


4eu+
CHARLES UNIVERSITY
SORBONNE UNIVERSITÉ
UNIVERSITY OF WARSAW
CLSINFRA
Look it up in the Documentation
Each treebank has its Documentation
You get there from the language list at universaldependencies.org
Look up the very treebank that was used to train the model you use to parse texts in UDPipe – there are (small) differences
https://universaldependencies.org/treebanks/en_ewt/index.html

Slide 120



Slide 121

Universal features - feats (English EWT corpus)

- lexical & grammatical properties of words beyond upos tags
- Table: the most common feats, each feature has a set of possible values
- Feature labels should be consistent across languages, but each language can add theirs if not covered
- feats: alphabetically concatenated, separated by | (vertical bar)

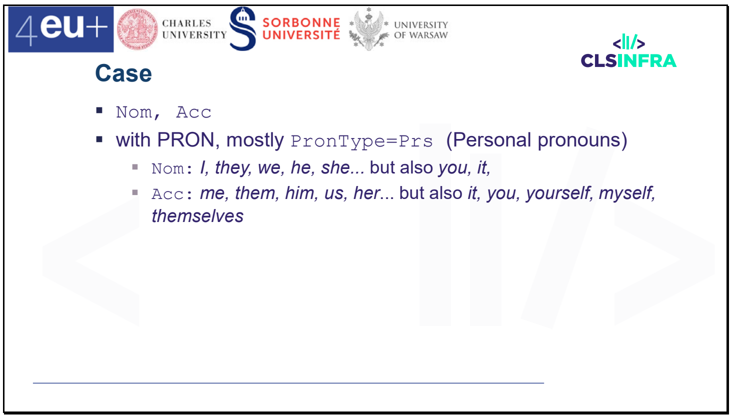
Lexical features*	Inflectional features*	
	*Nominal**	*Verbal**
PronType	Gender	VerbForm
NumType	Animacy	Mood
Poss	NounClass	Tense
Reflex	Number	Aspect
Foreign	Case	Voice
Abbr	Definite	Evident
Typo	Degree	Polarity
		Person
		Polite
		Clusivity

Slide 122

Features mostly describe only grammatical categories explicitly indicated by morphemes

- *he* ***writes*** `Person=3`, but *they* ***write*** does not have `Person`!
- *is sleeping* ≠ present progressive tense, but 2 verbs
  - *is* `Mood=Ind|Number=Sing|Person=3|Tense=Present|VerbForm=Fin`
  - *sleeping* `Tense=Pres|VerbForm=Part`
- Many inconsistencies:
  - e. g. *be*: parser tries to assign person beside 1^{st} and 3^{rd} singular present tense, other verbs not so much.

Slide 123



Case

- `Nom, Acc`
- with PRON, mostly `PronType=Prs` (Personal pronouns)
  - `Nom:` *I, they, we, he, she...* but also *you, it,*
  - `Acc:` *me, them, him, us, her...* but also *it, you, yourself, myself, themselves*

Slide 124

Gender

- `Fem, Masc,  Neut`
- with PRON, `PronType=Prs`
- usually also co-occurs with `Number, Person, Case, Poss`

Slide 125

Person

- `1,2,3`
- with VERB and AUX, mostly with `VerbForm=Fin, Mood=Ind, Number=Sing, Tense=Pres`
- with PRON, mostly with `PronType=Pers, Case, Poss,` and `Number` (any values)

Slide 126

Number

- `Plur, Sing`
- with NOUN and PROPN
- with PRON, mostly with `PronType=Prs, Case, Gender, Poss`
- with DET, mostly with `PronType=Dem`

Slide 127

Tense

- `Past, Pres`
- with VERB and AUX, mostly with `VerbForm=Fin, Mood=Ind, Number, Person`
- with SCONJ – `Past`: *given, based, provided*

Slide 128

Mood

- `Imp, Ind, Sub`
- with VERB and AUX, mostly with `VerbForm=Fin, Number, Person, Tense`

Slide 129

Voice

- `Pass`
- with VERB, mostly with `VerbForm=Part, Tense=Past`
- This is quite a weird feature in English. It occurs systematically in past participles, when they are combined with be as AUX (*I was invited*). In this case, it considers the context. Cf. (the invited experts: `Voice=Pass` is not there, just `Tense=Past|VerbForm=Part`.
- Perhaps the parser just decided to do this, based on input from some other data?

Slide 130

VerbForm

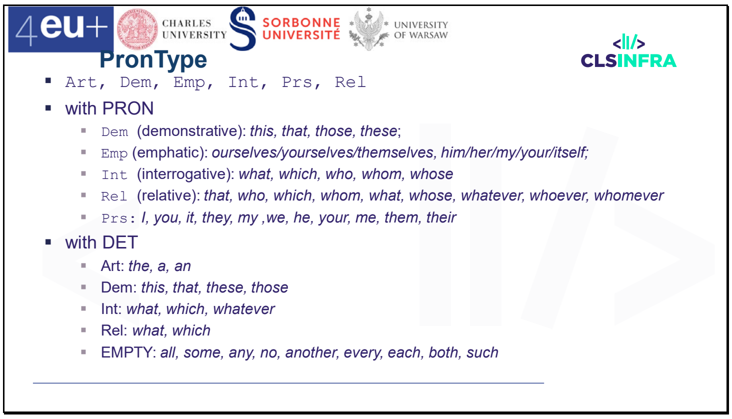
- `Fin, ~~Ger,~~ Inf, Part`
- with VERB and AUX
- with SCONJ (very little cases, maybe annotation errors)

Slide 131

Playtime!

https://quizlet.com/_bkoupi?x=1jqt&i=c5q4t
https://quizlet.com/_bkoqmz?x=1jqt&i=c5q4t

Slide 132


4eu+
CHARLES UNIVERSITY
SORBONNE UNIVERSITÉ
UNIVERSITY OF WARSAW
<|l/>
CLSINFRA
PronType
Art, Dem, Emp, Int, Prs, Rel
with PRON
Dem (demonstrative): this, that, those, these;
Emp (emphatic): ourselves/yourselves/themselves, him/her/my/your/itself;
Int (interrogative): what, which, who, whom, whose
Rel (relative): that, who, which, whom, what, whose, whatever, whoever, whomever
Prs: I, you, it, they, my ,we, he, your, me, them, their
with DET
Art: the, a, an
Dem: this, that, these, those
Int: what, which, whatever
Rel: what, which
EMPTY: all, some, any, no, another, every, each, both, such

Slide 133

PronType - continuation

- with ADV
  - `Dem`: *then, there, here*
  - Int: *how, why, where, when, whenever, however*
  - Rel: *when, where, how, wherein*
  - EMPTY: *so, just, very, also, now, even, only, as, back, well*
- with SCONJ
  - Int: *when, how, where, why, whenever, wherever, who*
  - Rel: *where, when, why*
  - EMPTY: *that, if, as, because, for, of, since, before, like, with*

Slide 134

Definite

- `Def, Ind`
- with DET
  - `Def`: *the*
  - `Ind`: *an, a*
  - `EMPTY`: *this, all, some, any, no, that, these, another, every, such*

Slide 135

NumType

- `Card, Frac, Mult, Ord`
- with NUM:
  - `Card`: *one, two, 1,30...*
- with ADJ:
  - `Frac`: *half*
  - `Ord`: *first, second, third,* 16^{th}, *...*
- with ADV:
  - `Frac`: *half*
  - `Mult`: *once, twice*

Slide 136

Degree

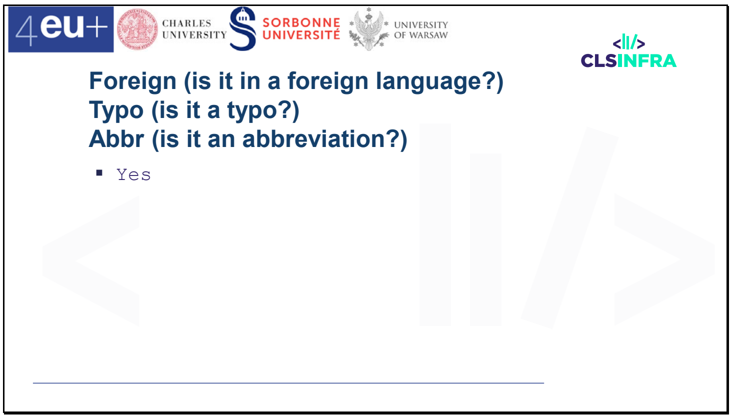
- `Cmp, Pos, Sup`
- with ADJ and ADV:
  - `Cmp`: *more, better, less, bigger...*
  - `Pos`: *good, great, new, far, well, soon, late, little, close...*
  - `Sup`: *best, most, least, worst, cheapest, largest...*

Slide 137

Poss (is it possessive?)
Reflex (is it reflexive?)

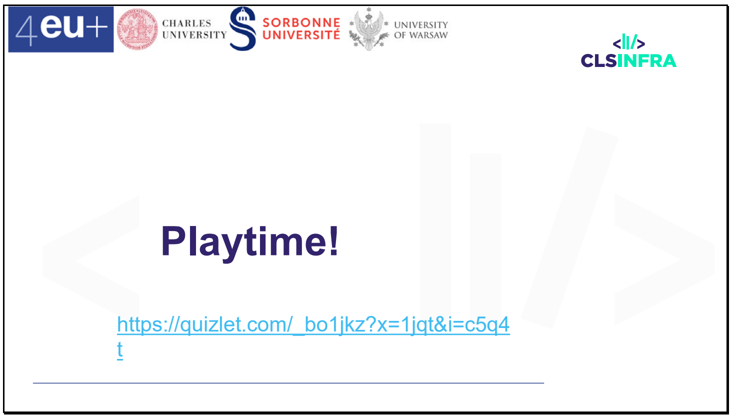
- `Yes`
- with PRON, mostly with `PronType=Prs, Gender, Number, Person`

Slide 138



4eu+
CHARLES UNIVERSITY
SORBONNE UNIVERSITÉ
UNIVERSITY OF WARSAW
</I/>
CLSINFRA
Foreign (is it in a foreign language?)
Typo (is it a typo?)
Abbr (is it an abbreviation?)
Yes

Slide 139


CHARLES UNIVERSITY
SORBONNE UNIVERSITÉ
UNIVERSITY OF WARSAW
CLSINFRA
Playtime!
https://quizlet.com/_bo1jkz?x=1jqt&i=c5q4
t

Slide 140

Feats and their values in your languages!

- A mind map of features (mainly of verbs) across languages is here:

https://www.orgpad.com/o/DfIEIyUSIBzY6YTaK-pUDf?token=Dp_2WHU1pHFKcAmAsmqLeC&open=all

- The UD documentation page on feats is here:

https://universaldependencies.org/u/feat/all.html

- Create groups and set up a list of words from your languages that would combine features and values not present in English.
- Are there word forms with ambiguous upos, such as participles in adjectival positions? Show us!
- You can consult UDPipe: **https://lindat.mff.cuni.cz/services/udpipe/**
  - Select an appropriate language model
  - Create an example sentence with the candidate and check out the markup.
  - If there are several models for your language, do they disagree?