4eu+ CHARLES UNIVERSITY SORBONNE UNIVERSITÉ UNIVERSITY OF WARSAW

Introduction
Class #6, March 21 2023

Silvie Cinková cinkova@ufal.mff.cuni.cz

Information extraction

- on structured data
  - Semantic Web (standards to make Web machine-readable)
  - knowledge bases/ontologies in general
- on unstructured data (texts)
  - population of ontologies
  - dialog systems
  - ...

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

4

English | Not logged in | Talk | Contributions | Create account | Log in

WIKIDATA

Main Page | Discussion | Read | View source | View history | Search Wikidata

Main page
Community portal
Project chat
Create a new Item
Recent changes
Random Item
Query Service
Nearby
Help
Donate

Lexicographical data

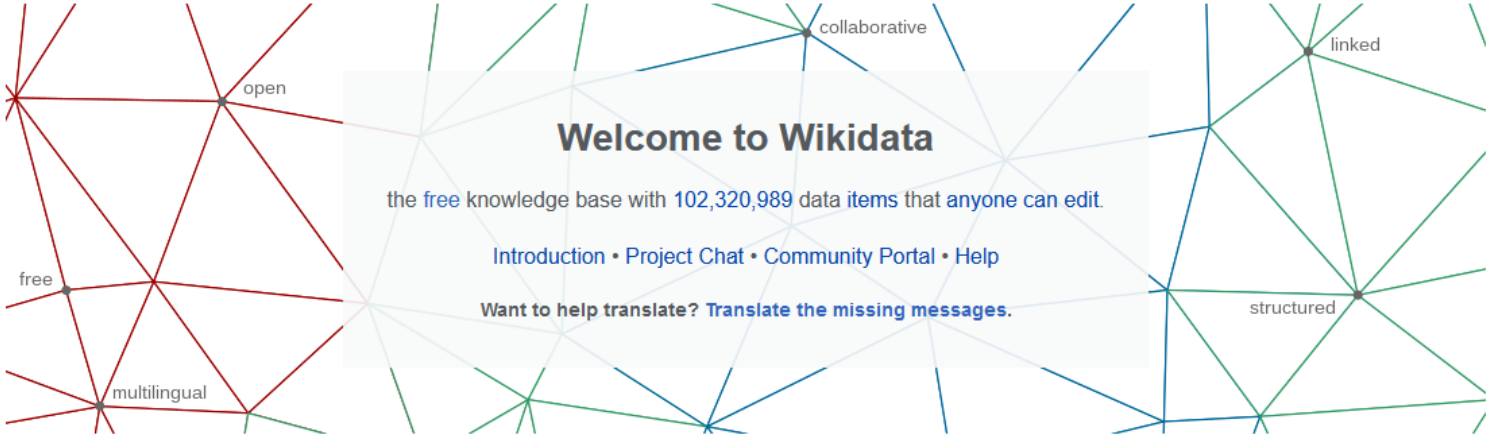
Create a new Lexeme
Recent changes
Random Lexeme

Tools

What links here
Related changes
Special pages
Permanent link
Page information
Wikidata item

In other projects

Wikimedia Commons
MediaWiki
Meta-Wiki
Multilingual Wikisource
Wikispecies
Wikibooks
Wikinews



Welcome to Wikidata

the free knowledge base with 102,320,989 data items that anyone can edit.

Introduction • Project Chat • Community Portal • Help

Want to help translate? Translate the missing messages.

Welcome!

Wikidata is a free and open knowledge base that can be read and edited by both humans and machines.

Wikidata acts as central storage for the **structured data** of its Wikimedia sister projects including Wikipedia, Wikivoyage, Wiktionary, Wikisource, and others.

Wikidata also provides support to many other sites and services beyond just Wikimedia projects! The content of Wikidata is available under a free license, exported using standard formats, and can be interlinked to other open data sets on the linked data web.

Get involved

*For a complete starters' guide, visit the **community portal**.*

Learn about data

New to the wonderful world of data? Develop and improve your data literacy through content designed to get you up to speed and feeling comfortable with the fundamentals in no time.

Item: *Earth (Q2)*

Property: *highest point (P610)*

custom value: *Mount Everest (Q513)*

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 Sa 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

André Mazon (Q181686)

French educationist, writer and professor

Andre Mazon

In more languages

Statements

instance of	human 2 references
image	 André Mazon 1934.jpg 4,185 × 4,712; 2.63 MB media legend André
date of birth	7 September 1881 *Gregorian* 8 references
place of birth	2nd arrondissement of Paris 2 references
date of death	13 July 1967 7 references
place of death	15th arrondissement of Paris 0 references

WIKIDATA

Wikidata Query Service | Examples | Help | More tools | Query Bu

```
#slavists living between 1860-1988
SELECT ?person ?personLabel  ?dob ?dod  ?placeBirthLabel ?GPS ?surnameLabel
WHERE
{
?person wdt:P101 wd:Q156864
?person wdt:P734 ?surname.
?person  wdt:P570 ?dod.
?person wdt:P569 ?dob.
?person wdt:P19 ?placeBirth.
        ?placeBirth wdt:P625 ?GPS.
FILTER("1988-01-01"^^xsd:dateTime >= ?dod && "1860-01-01"^^xsd:dateTime <= ?dob).
SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
}
ORDER BY ?surnameLabel
```

Slavic studies (Q156864)
studies of Slavic peoples, languages and culture

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	Șcheii Brașovului
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


Broch
Horten
Point(10.432778 59.420833)
4 August 1867
28 January 1961
Olaf Broch
Havránek
Prague
Point(14.421388888 50.0875)
30 January 1893
2 March 1978
Bohuslav Havránek
Alexandrov
Baytirák
Point(50.2621 55.60141)
16 April 1861
23 June 1918
Anastasiy (Aleksandrov)
Hrbek
Cedar Rapids
Point(-91.668611111 41.9833333
1 January 1878
1 January 1948
Šárka B. Hrbková
SE
EE
LV
LT
DK
BY
Nederland
Deutschland
Polska
Україна
SK
AT
HU
France
CH
HR
BA
RS
ME
BG
Italia
AD
United States of America
SHQIPËRIA
FIGURA TË SHQUARA BOTËRORE
CARL PATSCH
60
LEKË

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

```
SELECT ?wdLabel ?ps_Label ?wdpqLabel ?pq_Label {
VALUES (?company) {(wd:Q181686)}

  ?company ?p ?statement .
  ?statement ?ps ?ps_ .

  ?wd wikibase:claim ?p.
  ?wd wikibase:statementProperty ?ps.

  OPTIONAL {
  ?statement ?pq ?pq_ .
  ?wdpq wikibase:qualifier ?pq .
  }

  SERVICE wikibase:label { bd:serviceParam wikibase:language "en" }
} ORDER BY ?wd ?statement ?ps_
```

occupation	pedagogue
occupation	professor
occupation	translator

languages spoken, written or signed	Old Church Slavonic
languages spoken, written or signed	Russian
languages spoken, written or signed	Czech
languages spoken, written or signed	French

https://w.wiki/6U3V

member of	Polish Academy of Sciences
member of	Serbian Academy of Sciences and Arts

position held	vice president	of	International Committee of Slavists
position held	director	of	Institut d'études slaves
position held	academician	replaces	Henri Omont
position held	academician	of	Académie des Inscriptions et Belles-Lettres

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.

de France (1923-1951). Il dirige l'Institut d'études slaves de Paris à partir de 1937, devient vice-président du Comité international des slavistes (1958-1967).

André Mazon est cofondateur et membre du comité de rédaction de la *Revue des études slaves* (1921).

Information extraction/Text Mining with linguistic information

1. Conceptualize your research question
   - someone is a slavist/slavicist, works with Slavic studies
2. Operationalize your concepts
   - his name co-occurs with activities and works related to Slavic studies
   - teaches or translates from Slavic languages (list them)
3. Implement your operationalizations in corpus queries
   - use a corpus query language and linguistic markup

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

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%)

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:

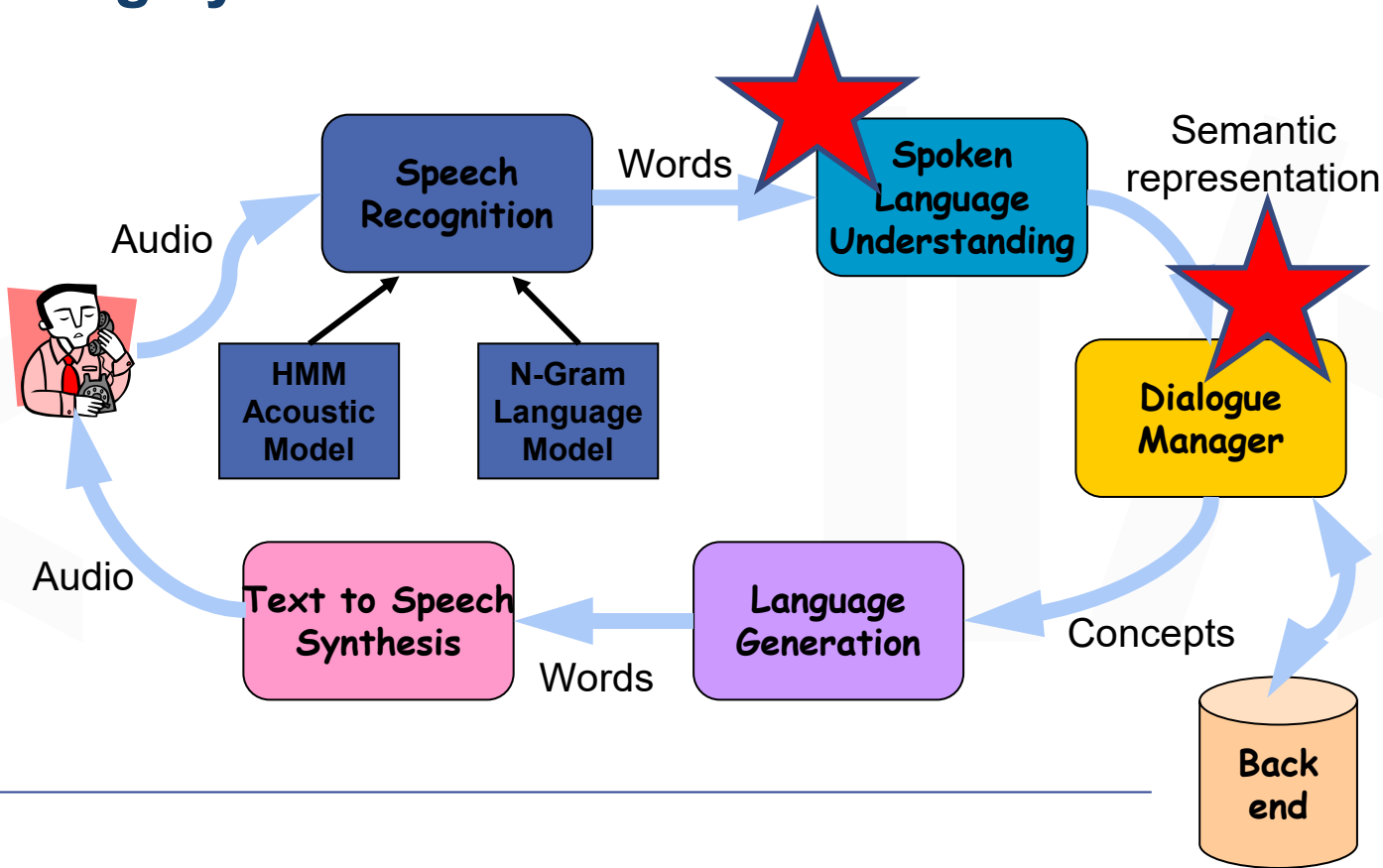
- First, it attempts to "read," or extract facts from text found in hundreds of millions of web pages (e.g., `playsInstrument(George_Harrison, guitar)`).
- Second, it attempts to improve its reading competence, so that tomorrow it can extract more facts from the web, more accurately.

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.

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

Dialog systems



Semantic grammar PHOENIX

- Grammar #1:

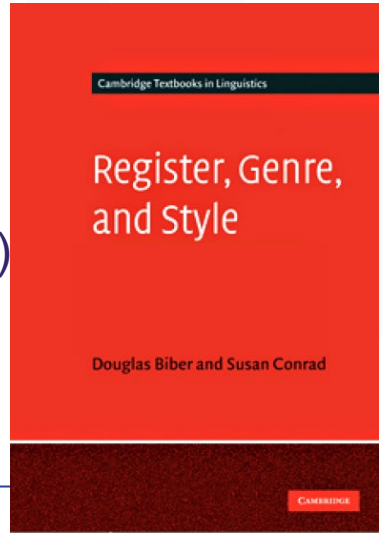
ORIGIN_CITY → [from | beginning in] [Atlanta | Pittsburgh | Boston | …]

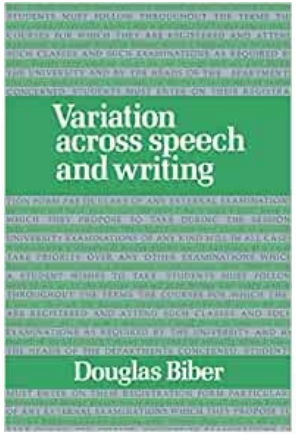
- Grammar #2:

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

Pragmatic concepts

- Social language use, Communication purpose in utterances
- Stylistic & rhetoric means
  - Described by lexical as well as grammatical features
- Genres and registers
  - Douglas Biber, since 1980s
  - Multidimensional Analysis (MDA)





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.

Narrativity

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

Descriptivity

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

Interactivity

- 2nd person
- questions
- vocatives
- imperatives

Uncertainty or distance

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

Emotionality

- + interjections
- + exclamation marks

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

Corpus Search

CQL Query: [lemma="woman"] [lemma = "of"] []{0,3} within s

Search query builder

9 results • ipm: 8.44

Tags: UD POS Tag | National POS Tag | Features | Lemma | Dependency relation | Dependency head

context	her reported to be a	**woman of an invincible spirit**	. But it shall be
context	maid's aunt , the fat	**woman of** | **Brentford , has**	a gown above . | MISTRESS
context	He cannot abide the old	**woman of** | **Brentford .**	He swears she's a witch
context	was 't not the wise	**woman of** | **Brentford ?**	| FALSTAFF | | Ay , marry ,
context	gossip Report be an honest	**woman of her** | **word .**	| SOLANIO | | I would she were
context	to desire to be a	**woman of the** | **world .**	| Enter two Pages . | | Here
context	denied , which longs | To	**women of all fashion ;**	lastly , hurried | Here to
context	to bear , | Making them	**women of good carriage .**	| This is she — ROMEO
context	man . The vows of **women** |	**Of no more bondage**	be to where they are

TEITOK

Available Corpora

DraCor - Shakespeare Drama Corpus

Home
CQL Search
PML-TQ Search
Grew Search
Search in Kontext

DEV Home

user: SC

Admin
Help
XML Files
Query Manager
UFAL admin

Powered by <TEI:TOK>
Maarten Janssen, 2014-

Grew Query

Below you can type in a Grew query that will be run on all the conll-u files of this UDWiki project

```
% search for womens characteristics (or possessors)
pattern {
womannode [lemma ="woman"];
howwoman [upos = "NOUN"];
ofnode [lemma = "of"];
womannode -[nmod]-> howwoman;
howwoman -[case]->ofnode
}
```

Cluster: howwoman.lemma clear

Run Query

howwoman.lemma	Count
world	1
word	1
spirit	1
fashion	1
carriage	1
bondage	1

Export to xlsx
Export to csv
Export to txt

stored queries • store this query

Soubor Úpravy Zobrazit Historie Záložky Nástroje Nápověda

DraCor - Shakespeare Drama Corpus

https://quest.ms.mff.cuni.cz/teitok-dev/teitok/teaching/ 120 %

Available Corpora

DraCor - Shakespeare Drama Corpus

Home
CQL Search
PML-TQ Search
Grew Search
Search in Kontext

DEV Home

user: SC

Below you can type in a Grew query that will be run on
project

```
% search for womens characteristics (or possessors)
pattern {
womannode [lemma ="woman"];
howwoman [upos = "NOUN"];
ofnode [lemma = "of"];
```

Cluster: howwoman.lemma clear

Run Query

howwoman.lemma = spirit (clear)

all.conllu

1 I have heard her reported to be a woman of an invincible spirit .

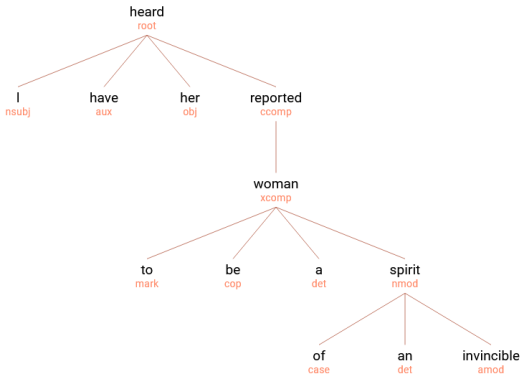
stored queries • store this query

Dependency Tree

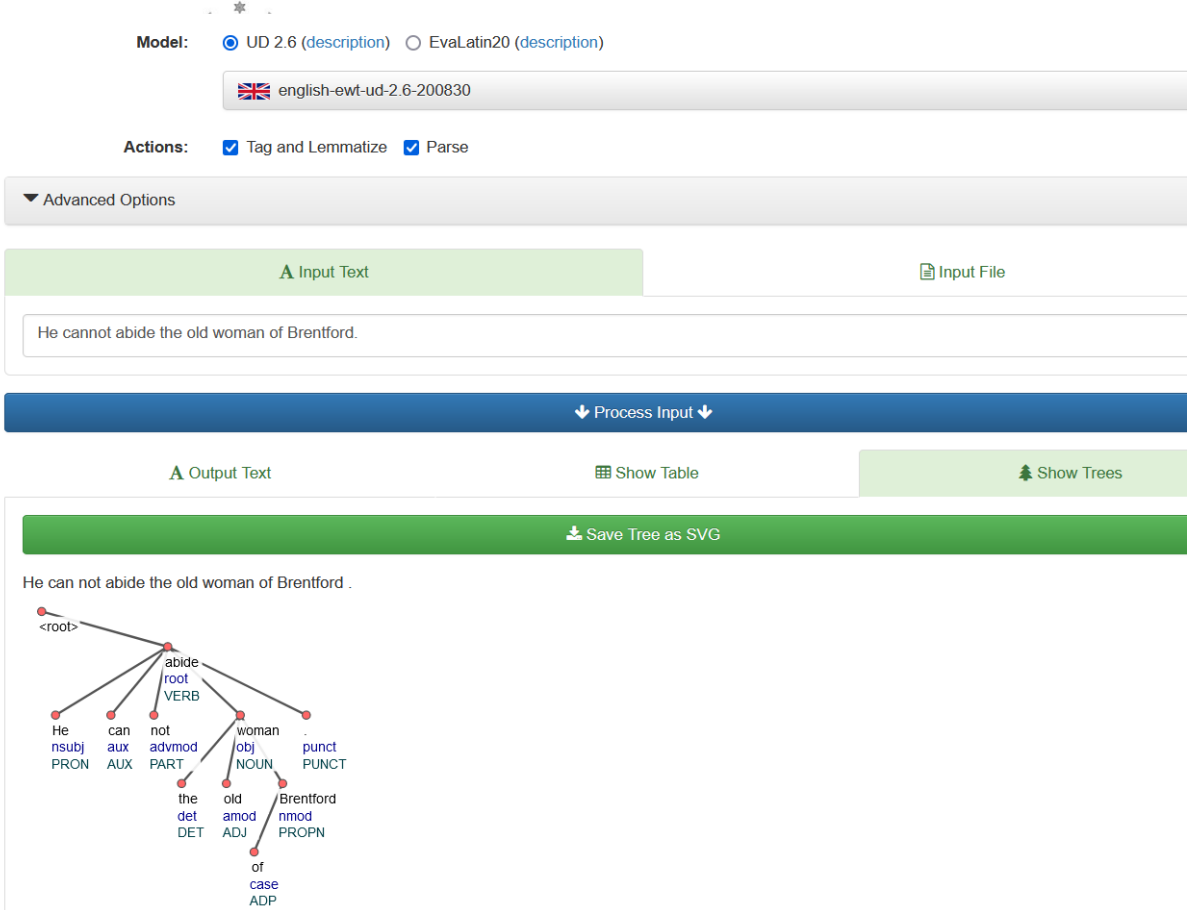
Henry VI, Part 2

s-911 < sentence s-912

I have heard her reported to be a woman of an invincible spirit .



lemmatization, morphological tagging, syntactic parsing



https://lindat.mff.cuni.cz/services/udpipe/

LINDAT Search Catalogue Education Projects Tools Services About

LINDAT/CLARIN / Services / UDPipe

UDPipe

About | Run | REST API Documentation

UDPipe is a trainable pipeline for tokenization, tagging, lemmatization and dependency parsing of CoNLL-U files. UDPipe is language-agnostic an
data in CoNLL-U format. Trained models are provided for nearly all UD treebanks. UDPipe is available as a binary for Linux/Windows/OS X, as a li
C#, and as a web service. Third-party R CRAN package also exists.

UDPipe is a free software distributed under the Mozilla Public License 2.0 and the linguistic models are free for non-commercial use and distribute
although for some models the original data used to create the model may impose additional licensing conditions. UDPipe is versioned using Sema

Copyright 2017 by Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Czech Republic.

Description of the available methods is available in the API Documentation and the models are described in the UDPipe 2 models list and UDPipe

Service

The service is freely available for testing. Respect the CC BY-NC-SA licence of the models – **explicit written permission of the authors is r
exploitation of the system**. If you use the service, you agree that data obtained by us during such use can be used for further improvements
comments and reactions are welcome.

Model: UD 2.6 (description) EvaLatin20 (description)

czech-pdt-ud-2.6-200830

Actions: Tag and Lemmatize Parse

Advanced Options

Input Text

Process Input

Output Text | Show Table | Show Trees

Model: UD 2.6 (description) EvaLatin20 (description)

czech-pdt-ud-2.6-200830

Actions:

Options

Outp

- armenian-armtdp-ud-2.6-200830
- basque-bdt-ud-2.6-200830
- belarusian-hse-ud-2.6-200830
- bulgarian-btb-ud-2.6-200830
- catalan-ancora-ud-2.6-200830
- chinese-gsdsimp-ud-2.6-200830
- chinese-gsd-ud-2.6-200830
- classical_chinese-kyoto-ud-2.6-200830
- coptic-scriptorium-ud-2.6-200830
- croatian-set-ud-2.6-200830
- czech-pdt-ud-2.6-200830
- czech-cac-ud-2.6-200830
- czech-fictree-ud-2.6-200830
- czech-cltt-ud-2.6-200830
- danish-ddt-ud-2.6-200830
- dutch-alpino-ud-2.6-200830
- dutch-lassysmall-ud-2.6-200830
- english-ewt-ud-2.6-200830
- english-gum-ud-2.6-200830
- english-lines-ud-2.6-200830
- english-partut-ud-2.6-200830

Model: UD 2.6 (description) EvaLatin20 (description)

czech-pdt-ud-2.6-200830

Actions: Tag and Lemmatize Parse

Advanced Options

A Input Text | Input File

I have heard her reported to be a woman of an invincible spirit.

Process Input

A Output Text | Show Table | Show Trees

Input Text | Input File

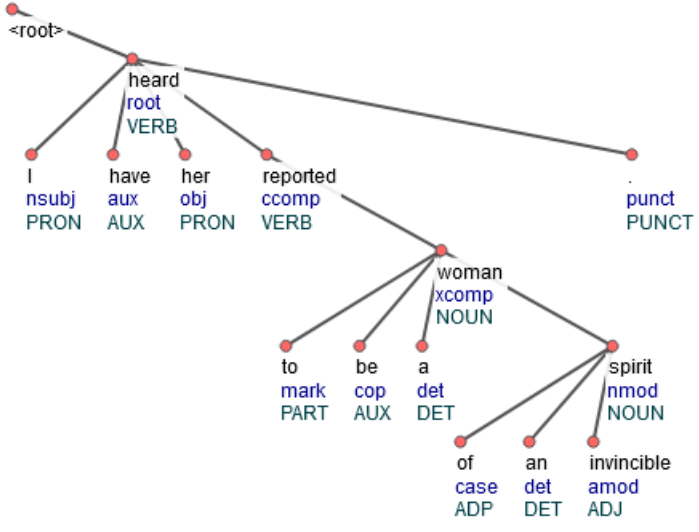
I have heard her reported to be a woman of an invincible spirit.

Process Input

Output Text | Show Table | Show Trees

Save Tree as SVG

I have heard her reported to be a woman of an invincible spirit .



Save Output File

Conll-u format

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.

Id	Form	Lemma	UPosTag	XPosTag	Feats	Head	DepRel	Deps	Misc
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

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

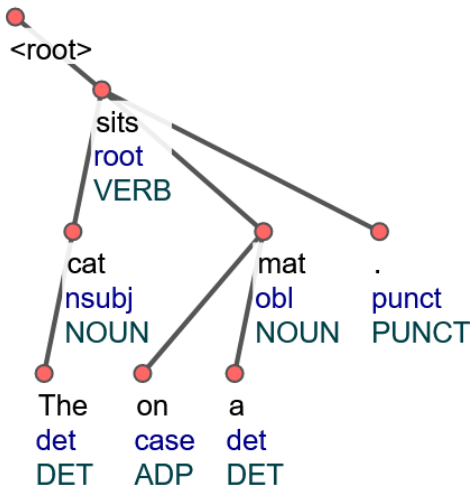
newpar

sent_id = 1

text = The cat sits on a mat.

Id	Form	Lemma	UPosTag	XPosTag	Feats	Head	DepRel		
1	The	the	DET	DT	Definite=Def|PronType=Art	2	det		
2	cat	cat	NOUN	NN	Number=Sing	3	nsubj		
3	sits	sit	VERB	VBZ	Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin	0	root	_	TokenRange=8:12
4	on	on	ADP	IN	_	6	case	_	TokenRange=13:15
5	a	a	DET	DT	Definite=Ind|PronType=Art	6	det	_	TokenRange=16:17
6	mat	mat	NOUN	NN	Number=Sing	3	obl	_	SpaceAfter=No|TokenRange=18:21
7	.	.	PUNCT	.	_	3	punct	_	TokenRange=21:22

The cat sits on a mat .



4eu+ CHARLES UNIVERSITY SORBONNE UNIVERSITÉ UNIVERSITY OF WARSAW

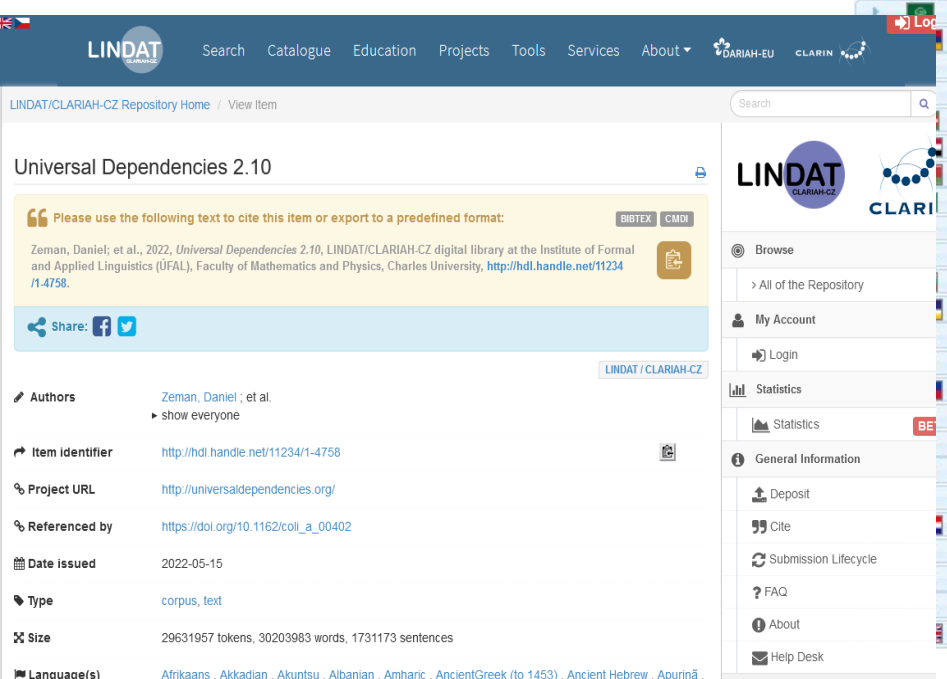
CLSINFRA

Universal Dependencies

universaldependencies.org

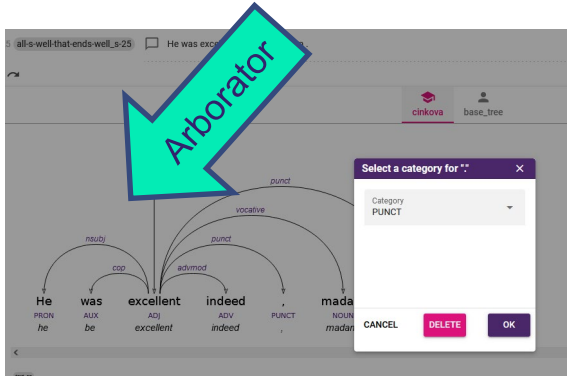
Consistent grammar annotation across languages

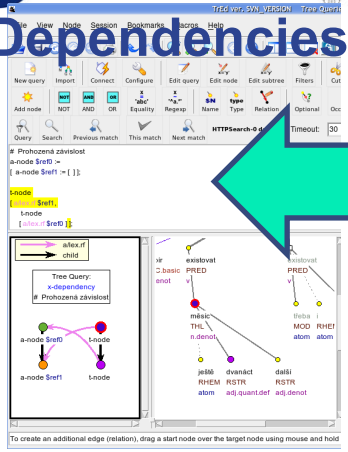
- over 300 contributors
- nearly 200 treebanks (corpora w. syntax annotation)
- over 100 languages
- publicly available

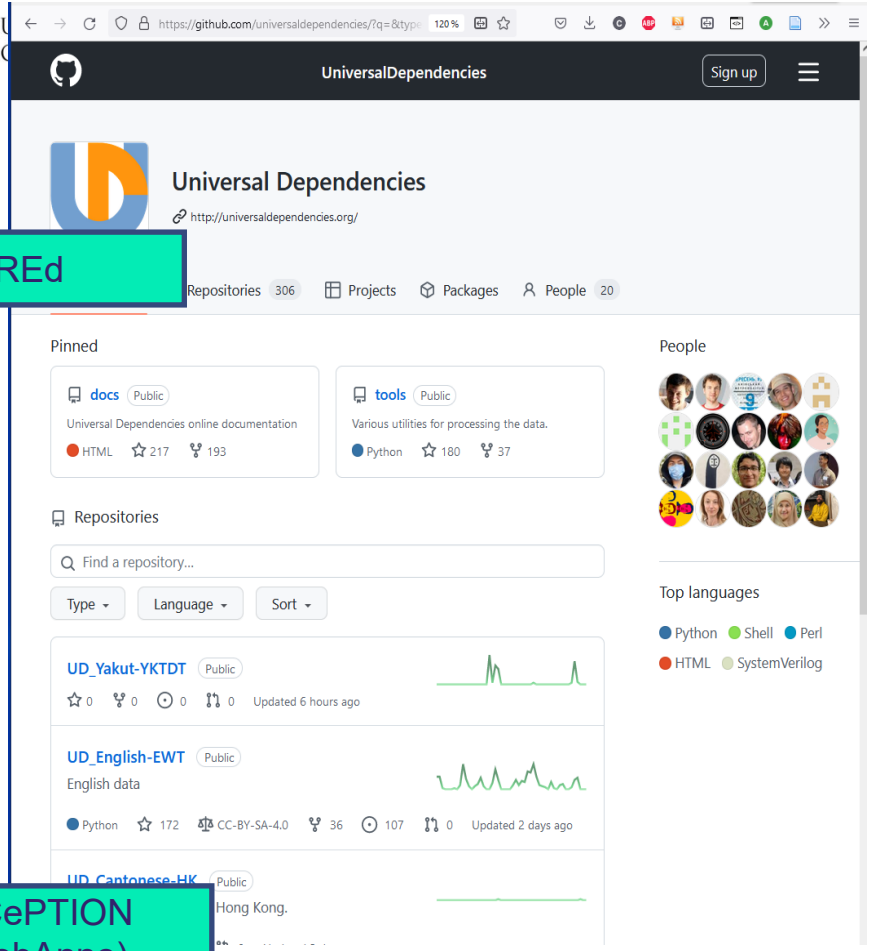


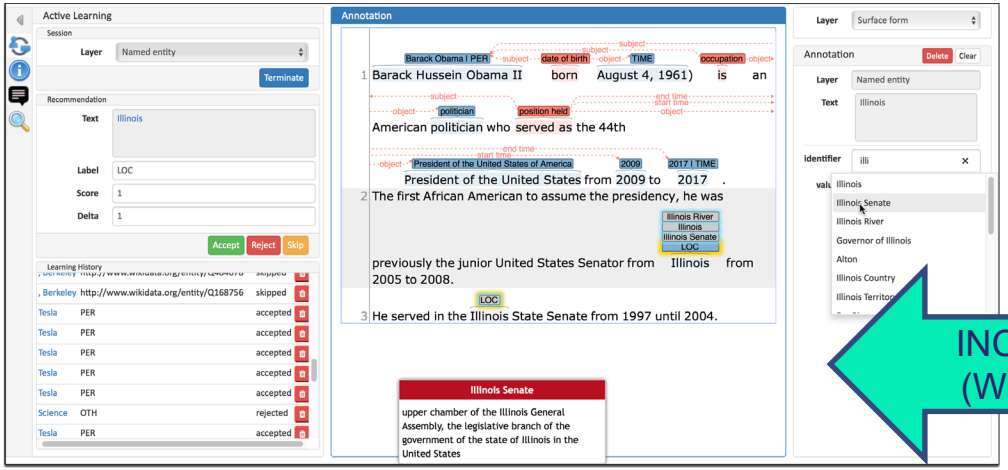
Akkadian	2	25K		Afro-Asiatic, Semitic
Akuntsu	1	1K		Tupian, Tupari
Albanian	1	<1K	W	IE, Albanian
Amharic	1	10K		Afro-Asiatic, Semitic
Ancient Greek	2	416K		IE, Greek
Ancient Hebrew	1	39K		Afro-Asiatic, Semitic
Apurina	1	<1K		Arawakan
Arabic	3	1,042K	W	Afro-Asiatic, Semitic
Armenian	2	94K	W	IE, Armenian
Assyrian	1	<1K		Afro-Asiatic, Semitic
Bambara	1	13K		Mande
Basque	1	121K		Basque
Beja	1	<1K		Afro-Asiatic, Cushitic
Belarusian	1	305K	W	IE, Slavic
Bengali	1	<1K		IE, Indic
Bhojpuri	1	6K		IE, Indic
Breton	1	10K	W	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	W	Sino-Tibetan
Chukchi	1	6K		Chukotko-Kamchatkan
Classical Chinese	1	289K		Sino-Tibetan
Coptic	1	52K		Afro-Asiatic, Egyptian
Croatian	1	199K	W	IE, Slavic
Czech	5	2,227K	W	IE, Slavic
Danish	1	100K		IE, Germanic
Dutch	2	306K	W	IE, Germanic
English	9	762K	W	IE, Germanic

Contribute to Universal Dependencies









Universal Parts of Speech (upos)

UD Morphology

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

NOUN vs. PROPN vs. neither

strawberries	

NOUN vs. PROPN vs. neither

strawberries	NOUN
cat	

NOUN vs. PROPN vs. neither

strawberries	NOUN
cat	NOUN
small	

NOUN vs. PROPN vs. neither

strawberries	NOUN
cat	NOUN
small	neither
Peter	

NOUN vs. PROPN vs. neither

strawberries	NOUN
cat	NOUN
small	neither
Peter	PROPN
butter	

NOUN vs. PROPN vs. neither

strawberries	NOUN
cat	NOUN
small	neither
Peter	PROPN
butter	NOUN
beer	

NOUN vs. PROPN vs. neither

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

NOUN vs. PROPN vs. neither

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

NOUN vs. PROPN vs. neither

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

VERB vs. AUX vs. neither

are	

VERB vs. AUX vs. neither

are	AUX
can	

VERB vs. AUX vs. neither

are	AUX
can	AUX
(He) did (it)	

VERB vs. AUX vs. neither

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

VERB vs. AUX vs. neither

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

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 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

VERB vs. AUX vs. neither

(a) winning (strategy)	

VERB vs. AUX vs. neither

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

VERB vs. AUX vs. neither

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

VERB vs. AUX vs. neither

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

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 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 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 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

ADJ vs. ADV vs. neither

green	

ADJ vs. ADV vs. neither

green	ADJ
happily	

ADJ vs. ADV vs. neither

green	ADJ
happily	ADV
my	

ADJ vs. ADV vs. neither

green	ADJ
happily	ADV
my	neither
many	

ADJ vs. ADV vs. neither

green	ADJ
happily	ADV
my	neither
many	ADJ
oldest	

ADJ vs. ADV vs. neither

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

ADJ vs. ADV vs. neither

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

ADJ vs. ADV vs. neither

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

ADJ vs. ADV vs. neither

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

ADJ vs. ADV vs. neither

twice	ADV
(take) off (phrasal verb)	

ADJ vs

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

ADJ vs. ADV vs. neither

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

ADJ vs. ADV vs. neither

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

ADJ vs. ADV vs. neither

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

ADJ vs. ADV vs. neither

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

ADJ vs. ADV vs. neither

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

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

SCONJ vs. CCONJ vs. neither

(I hope) that (she will come)	

SCONJ vs. CCONJ vs. neither

(I hope) that (she will come)	SCONJ
(good) and (bad)	

SCONJ vs. CCONJ vs. neither

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

SCONJ vs. CCONJ vs. neither

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

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	

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)	

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)	

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

NUM vs. DET vs. PRON

we	

NUM vs. DET vs. PRON

we	PRON
Which kids arrived?	

NUM vs. DET vs. PRON

we	PRON
Which kids arrived?	DET
Say which you like	

NUM vs. DET vs. PRON

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

NUM vs. DET vs. PRON

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

NUM vs. DET vs. PRON

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

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	

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)	

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

DET vs. NUM vs. ADJ vs. ADV

many	

DET vs. NUM vs. ADJ vs. ADV

many	DET
two	

DET vs. NUM vs. ADJ vs. ADV

many	DET
two	NUM

DET vs. NUM vs. ADJ vs. ADV

many	DET
two	NUM
first (minute)	

DET vs. NUM vs. ADJ vs. ADV

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

DET vs. NUM vs. ADJ vs. ADV

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

DET vs. NUM vs. ADJ vs. ADV

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

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 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	

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

ADP vs. ADV vs. SCONJ

for (you)	

ADP vs. ADV vs. SCONJ

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

A

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

ADP vs. ADV vs. SCONJ

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

ADP vs. ADV vs. SCONJ

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

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 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)	

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

Particles (PART)

- *not, n’t*
- *to* (infinitive marker)
- *‘s* (genitive ending)

Interjections (INTJ)

- yes, no
- please
- well
- hi
- ok, bravo
- like
- lol
- hey
- oh, ouch

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

Universal Features

UD Morphology

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

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 1st and 3rd singular present tense, other verbs not so much.

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*

Gender

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

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)

Number

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

Tense

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

Mood

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

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?

VerbForm

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

Playtime!

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

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*

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*

Definite

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

NumType

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

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...*

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

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

Foreign (is it in a foreign language?)
Typo (is it a typo?)
Abbr (is it an abbreviation?)

- Yes

Playtime!

https://quizlet.com/_bo1jkz?x=1jqt&i=c5q4t

Feats and their values in your languages!

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

https://www.orgpad.com/o/DfIEIyUSlBzY6YTaK-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:
  - 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?

https://lindat.mff.cuni.cz/services/udpipe/