VR MFF UK

Cross-Language Harmonization of Linguistic Resources

Daniel Zeman

November 6, 2024

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

Computational Linguistics

Multilingual:

- not just for English
- not just for resource-rich European languages (English, Czech, ...)

- ideally all languages (up to 7000?)
- including
  - endangered
  - unwritten
  - very different from Indo-European

Computational Linguistics

Multilingual:

- not just for English
- not just for resource-rich European languages (English, Czech, ...)

- ideally all languages (up to 7000?)
- including
  - endangered
  - unwritten
  - very different from Indo-European

- focus on
  - Morphology
  - Syntax

Morphology: Explaining Word Forms

English

do, does, did, done, doing

Czech

dělat, dělati, dělám, děláš, dělá, děláme, děláte, dělají, dělej, dělejme, dělejte, dělal, dělala, dělalo, dělali, dělaly, dělaje, dělajíc, dělajíce, dělán, dělána, děláno, děláni, dělány

Finnish

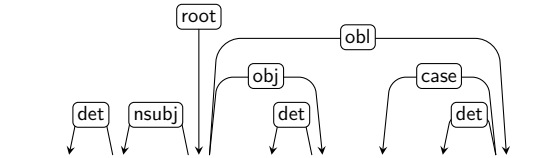
tehdä, tee, teemme, teen, teenkin, teenkö, teet, teette, tehden, tehdessä, tehdessään, tehdyillä, tehdyissä, tehdyistä, tehdylle, tehdyllä, tehdyn, tehdyssä, tehdystä, tehdyt, tehdään, tehkää, tehkäämme, tehköön, tehkööt, tehnee, tehneen, tehneensä, tehneet, tehneille, tehnemme, tehnen, tehnet, tehnette, tehnevät, tehny, tehnyt, tehnytkin, tehtiin, tehtiinkin, tehty, tehtyjen, tehtyjä, tehtynä, tehtyyn, tehtyä, tehtyäni, tehtyään, tehtäessä, tehtäisi, tehtäisiin, tehtäkö, tehtäköön, tehtämän, tehtäne, tehtäneen, tehtävien, tehtäviin, tehtäville, tehtävistä, tehtäviä, tehtävä, tehtävän, teimme, tein, teinkin, teinpä, teit, teitte, tekee, tekeekin, tekeekö, tekemiensä, tekemiin, tekeminen, tekemistä, tekemiä, tekemiäni, tekemiään, tekemä, tekemäisilläni, tekemällä, tekemän, tekemäni, tekemänsä, tekemässä, tekemästä, tekemästään, tekemät, tekemättä, tekemää, tekemään, tekemäänsä, tekevien, tekeville, tekevä, tekeväni, tekevät, tekevää, tekevään, teki, tekikin, tekis, tekisi, tekisimme, tekisin, tekisit, tekisitte, tekisivät, tekivät, tekivätkö

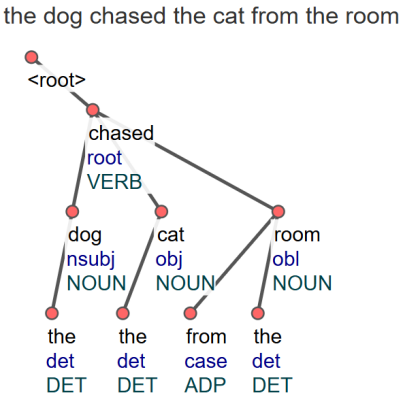
Morphology: Explaining Word Forms

word form:	*dělám* “I do”
lemma:	*dělat* “do”
word category (part of speech):	**VERB**
features:	**indicative mood** **present tense** **active voice** **first person singular**

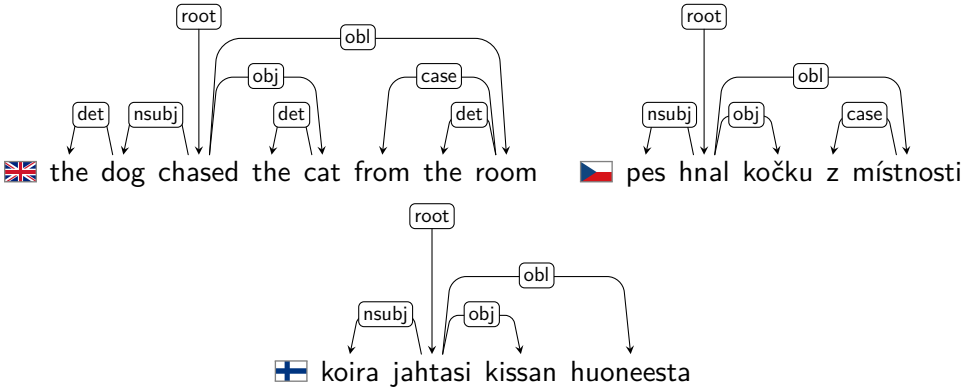
word form:	*teen* “I do”
lemma:	*tehdä* “do”
word category (part of speech):	**VERB**
features:	**indicative mood** **present tense** **active voice** **first person singular**

Syntax: Relations between Words



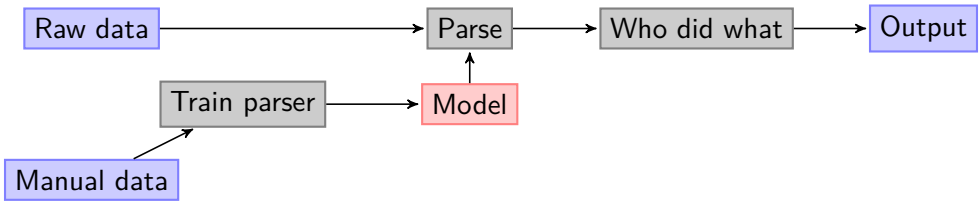


Syntax: Relations between Words



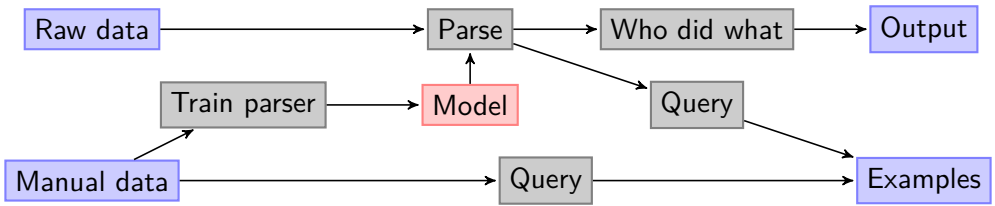
What Is It Good For?

- Downstream **language understanding** technology



What Is It Good For?

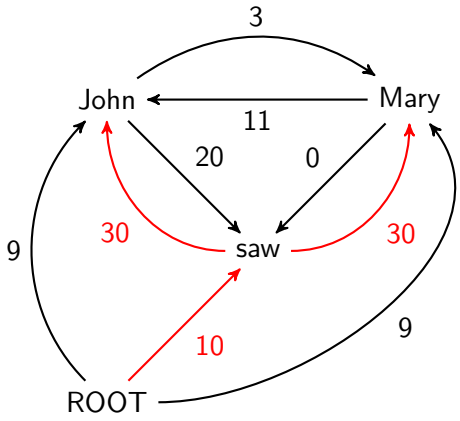
- Downstream **language understanding** technology
- **Linguistic research**



- More broadly: **digital humanities** (contrastive linguistics, typology, literary studies, cultural heritage, works in classical languages)

Parsing

Total graph + edge scores → maximum spanning tree



$$A^* = \underset{\substack{A \subseteq G \\ A \text{ is tree}}}{\text{argmax}} \sum_{e \in A} score(e)$$

$G = \langle V, E \rangle =$

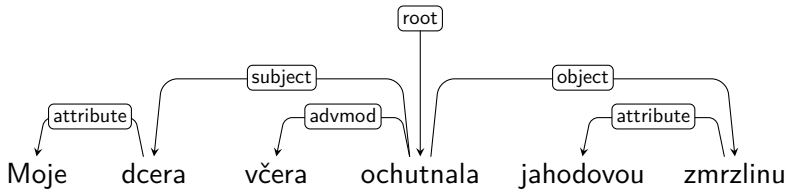
Parsing Low-Resource Languages

- Machine learning:
- Manual annotation needed to train a parser
  - Thousands of sentences – hard and expensive

- Available for a few "lucky" languages
- But there are thousands of languages – what about the less lucky ones?

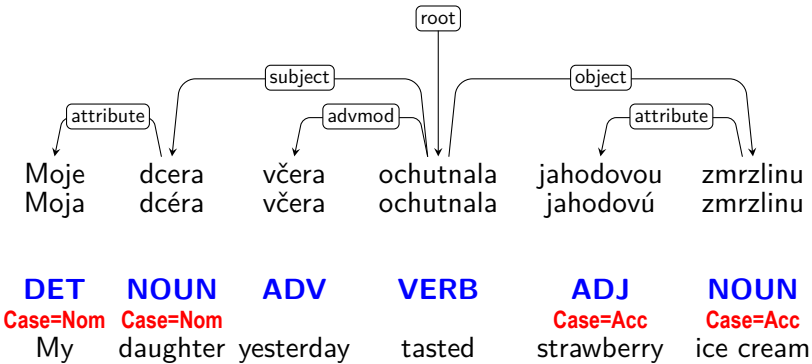
Daniel Zeman, *Philip Resnik (2008): Cross-Language Parser Adaptation between Related Languages. In: IJCNLP 2008 Workshop on NLP for Less Privileged Languages, pp. 35-42, Hyderabad, India* 265 cit.

Delexicalized Parsing

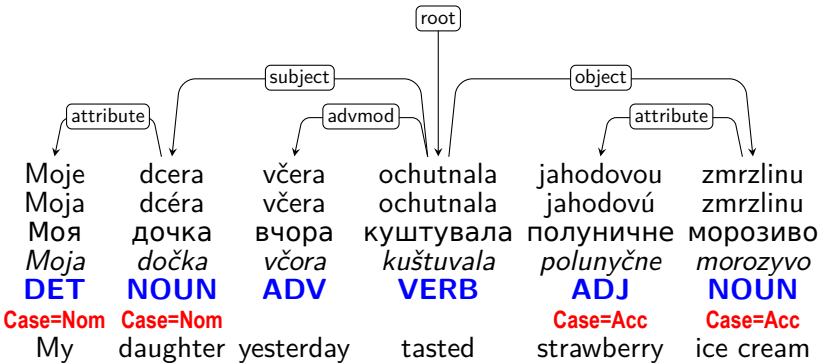


DET	NOUN	ADV	VERB	ADJ	NOUN
Case=Nom	Case=Nom			Case=Acc	Case=Acc
My	daughter	yesterday	tasted	strawberry	ice cream

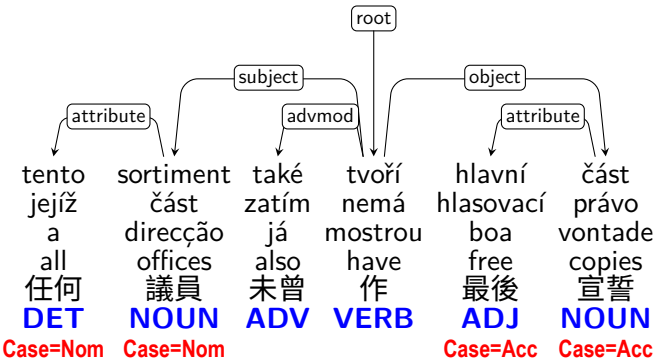
Delexicalized Parsing



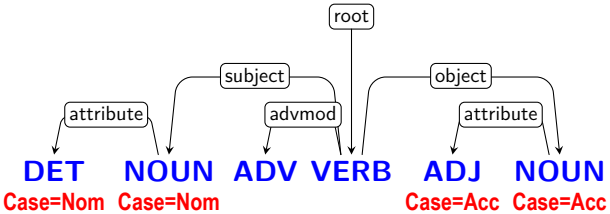
Delexicalized Parsing



Delexicalized Parsing



Delexicalized Parsing



Evaluation: Does It Work?

- Manually annotated test data needed
  ⇒ Simulate on languages for which we have test data
- Train parser on Czech
  (large training data available)

$truepos$ ... correctly identified dependencies

$$P = \frac{truepos}{truepos + falsepos}$$

$$R = \frac{truepos}{truepos + falseneg}$$

$$LAS = F_1 = \frac{2 \times P \times R}{P + R}$$

Evaluation: Does It Work?

- Manually annotated test data needed
  ⇒ Simulate on languages for which we have test data

- Train parser on Czech (large training data available)
- Evaluate on Slovak (some test data available)

$truepos$ ... correctly identified dependencies

$$P = \frac{truepos}{truepos + falsepos}$$

$$R = \frac{truepos}{truepos + falseneg}$$

$$LAS = F_1 = \frac{2 \times P \times R}{P + R}$$

Evaluation: Does It Work?

- Manually annotated test data needed
  ⇒ Simulate on languages for which we have test data

- Train parser on Czech (large training data available)
- Evaluate on Slovak (some test data available)
- If OK, assume it also works for Lower Sorbian (no data available)

$truepos$ ... correctly identified dependencies

$$P = \frac{truepos}{truepos + falsepos}$$

$$R = \frac{truepos}{truepos + falseneg}$$

$$LAS = F_1 = \frac{2 \times P \times R}{P + R}$$

PROBLEM:

The annotations must be compatible across languages!

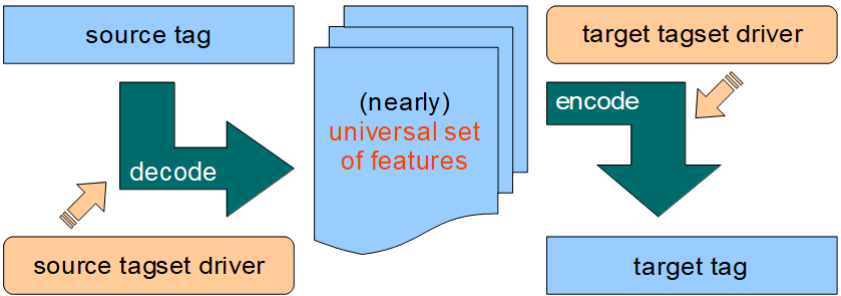
Morphology Example: SynTagRus Tags vs. PDT Tags

SynTagRus annotation	PDT annotation	Meaning
S ЕД МУЖ ИМ	NNMS1-----A----	noun masculine singular nominative
S МН РОД ОД	PSXXXXP3-------	pronoun possessive plural 3rd person
A МН ИМ	AAXP1----1A----	adjective positive plural nominative
NUM ВИН	ClXX4----------	numeral cardinal accusative
V НЕСОВ ИЗЪЯВ...	VB-P---3P-AA---	verb imperfective present indicative ...
... НЕПРОШ МН 3-Л		... 3rd person plural
ADV СРАВ	Dg-------2A----	adverb comparative
PR	RR--6----------	preposition
CONJ	J^-------------	coordinating conjunction
PART	TT-------------	particle
INTJ	II-------------	interjection

Daniel Zeman *(2008): Reusable Tagset Conversion Using Tagset Drivers. In: Proceedings of LREC, pp. 213–218, ELRA, Marrakech, Morocco* *234 cit.*

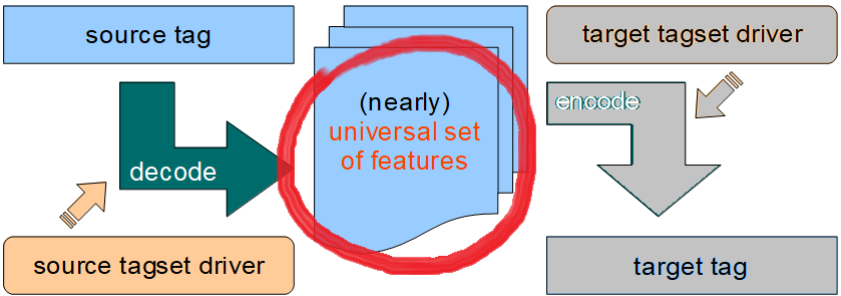
Tagset Drivers

- A module with the following functions:
  - `decode()` … converts a tag to Interset
  - `encode()` … generates a tag from Interset
  - `list()` … lists known tags in the tagset (optional)



Tagset Drivers

- A module with the following functions:
  - `decode()` … converts a tag to Interset
  - `encode()` … generates a tag from Interset
  - `list()` … lists known tags in the tagset (optional)



Interset

pos	noun	adj	num	verb	adv	adp	conj	part	int	punc	sym
nountype	com	prop	class
nametype	geo	prs	giv	sur	nat	com	pro	oth	col	sci	che	med	tec	cel	gov	jus	fin	env	cul	spo	hob
adjtype	pdt
prontype	prn	prs	rcp	art	int	rel	exc	dem	emp	neg	ind	tot
numtype	card	ord	mult	frac	sets	dist	range
numform	word	digit	roman	combi
verbtype	aux	cop	mod	light
advtype	man	loc	tim	sta	deg	cau	mod	adadj	ex
adpostype	prep	post	circ	voc
conjtype	coor	sub	comp	oper
parttype	mod	emp	res	inf	vbp
punctype	peri	qest	excl	quot	brck	comm	colo	semi	dash	root
puncside	ini	fin
poss	yes
reflex	yes
foreign	yes
abbr	yes
hyph	yes
typo	yes
echo	rdp	ech
polarity	pos	neg
definite	ind	spec	def	cons	com
gender	masc	fem	com	neut
animacy	anim	hum	nhum	inan
number	sing	dual	tri	pauc	grpa	plur	grpl	inv	ptan	coll	count
case	nom	gen	dat	acc	voc	loc	ins	abl	del	par	dis	ess	tra	com	abe	ine	ela	ill	ade	all	sub	sup	lat	per	add	tem	ter	abs	erg	cau	ben	cns	equ	cmp
prepcase	npr	pre
degree	pos	cmp	sup	abs	equ	dim	aug
person	0	1	2	3	4
clusivity	in	ex
polite	infm	form	elev	humb
possgender	masc	fem	com	neut
possperson	1	2	3
possnumber	sing	dual	plur
possednumber	sing	dual	plur
absperson	1	2	3
ergperson	1	2	3
datperson	1	2	3
absnumber	sing	dual	plur
ergnumber	sing	dual	plur
datnumber	sing	dual	plur
abspolite	infm	form	elev	humb
ergpolite	infm	form	elev	humb
datpolite	infm	form	elev	humb
erggender	masc	fem	com	neut
datgender	masc	fem	com	neut
position	prenom	postnom	nom	free
subcat	intr	tran
verbform	fin	inf	sup	part	conv	vnoun	ger	gdv
mood	ind	imp	cnd	pot	sub	jus	prp	opt	des	nec	qot	adm
tense	pres	fut	past	aor	imp	pqp
voice	act	mid	pass	rcp	cau	int	antip	dir	inv
evident	fh	nfh
aspect	imp	perf	prosp	prog	hab	iter
strength	weak	strong
variant	short	long	0	1	2	3	4	5	6	7	8	9	a	b	c
style	arch	rare	form	poet	norm	coll	vrnc	slng	expr	derg	vulg
tagset	*any*
other	*any*

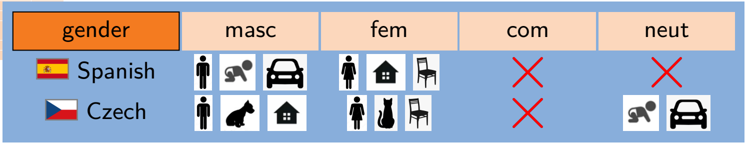
Interset

pos	noun	adj	num	verb	adv	adp	conj	part	int	punc	sym																						
nountype	com	prop	class																														
nametype	geo	prs	giv	sur	nat	com	pro	oth	col	sci	che	med	tec	cel	gov	jus	fin	env	cul	spo	hob												
adjtype	pdt																																
prontype	prn	prs	rcp	art	int	rel	exc	dem	emp	neg	ind	tot																					
numtype	card	ord	mult	frac	sets	dist	range																										
numform	word	digit	roman	combi																													
verbtype	aux	cop																															
advtype	man	loc																															
adpostype	prep	post																															
conjtype	coor	sub																															
parttype	mod	emp																															
punctype	peri	qest																															
puncside	ini	fin																															
poss	yes																																
reflex	yes																																
foreign	yes																																
abbr	yes																																
hyph	yes																																
typo	yes																																
echo	rdp	ech																															
polarity	pos	neg																															
definite	ind	spec	def	cons	com																												
gender	masc	fem	com	neut																													
animacy	anim	hum	nhum	inan																													
number	sing	dual	tri	pauc	grpa	plur	grpl	inv	ptan	coll	count																						
case	nom	gen	dat	acc	voc	loc	ins	abl	del	par	dis	ess	tra	com	abe	ine	ela	ill	ade	all	sub	sup	lat	per	add	tem	ter	abs	erg	cau	ben	cns	equ	cmp
prepcase	npr	pre																															
degree	pos	cmp	sup	abs	equ	dim	aug																										
person	0	1	2	3	4																												
clusivity	in	ex																															
polite	infm	form	elev	humb																													
possgender	masc	fem	com	neut																													
possperson	1	2	3																														
possnumber	sing	dual	plur																														
possednumber	sing	dual	plur																														
absperson	1	2	3																														
ergperson	1	2	3																														
datperson	1	2	3																														
absnumber	sing	dual	plur																														
ergnumber	sing	dual	plur																														
datnumber	sing	dual	plur																														
abspolite	infm	form	elev	humb																													
ergpolite	infm	form	elev	humb																													
datpolite	infm	form	elev	humb																													
erggender	masc	fem	com	neut																													
datgender	masc	fem	com	neut																													
position	prenom	postnom	nom	free																													
subcat	intr	tran																															
verbform	fin	inf	sup	part	conv	vnoun	ger	gdv																									
mood	ind	imp	cnd	pot	sub	jus	prp	opt	des	nec	qot	adm																					
tense	pres	fut	past	aor	imp	pqp																											
voice	act	mid	pass	rcp	cau	int	antip	dir	inv																								
evident	fh	nfh																															
aspect	imp	perf	prosp	prog	hab	iter																											
strength	weak	strong																															
variant	short	long	0	1	2	3	4	5	6	7	8	9	a	b	c																		
style	arch	rare	form	poet	norm	coll	vrnc	slng	expr	derg	vulg																						
tagset	*any*																																
other	*any*																																

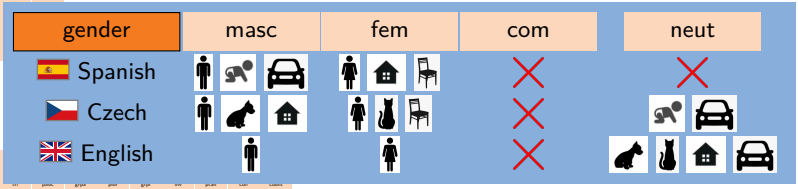
gender	masc	fem	com	neut
Spanish				

Interset

pos	noun	adj	num	verb	adv	adp	conj	part	int	punc	sym																							
nountype	com	prop	class																															
nametype	geo	prs	giv	sur	nat	com	pro	oth	col	sci	che	med	tec	cel	gov	jus	fin	env	cul	spo	hob													
adjtype	pdt																																	
prontype	prn	prs	rcp	art	int	rel	exc	dem	emp	neg	ind	tot																						
numtype	card	ord	mult	frac	sets	dist	range																											
numform	word	digit	roman	combi																														
verbtype	aux	cop																																
advtype	man	loc																																
adpostype	prep	post																																
conjtype	coor	sub																																
parttype	mod	emp																																
punctype	peri	qest																																
puncside	ini	fin																																
poss	yes																																	
reflex	yes																																	
foreign	yes																																	
abbr	yes																																	
hyph	yes																																	
typo	yes																																	
echo	rdp	ech																																
polarity	pos	neg																																
definite	ind	spec	def	cons	com																													
gender	masc	fem	com	neut																														
animacy	anim	hum	nhum	inan																														
number	sing	dual	tri	pauc	grpa	plur	grpl	inv	ptan	coll	count																							
case	nom	gen	dat	acc	voc	loc	ins	abl	del	par	dis	ess	tra	com	abe	ine	ela	ill	ade	all	sub	sup	lat	per	add	tem	ter	abs	erg	cau	ben	cns	equ	cmp
prepcase	npr	pre																																
degree	pos	cmp	sup	abs	equ	dim	aug																											
person	0	1	2	3	4																													
clusivity	in	ex																																
polite	infm	form	elev	humb																														
possgender	masc	fem	com	neut																														
possperson	1	2	3																															
possnumber	sing	dual	plur																															
possednumber	sing	dual	plur																															
absperson	1	2	3																															
ergperson	1	2	3																															
datperson	1	2	3																															
absnumber	sing	dual	plur																															
ergnumber	sing	dual	plur																															
datnumber	sing	dual	plur																															
abspolite	infm	form	elev	humb																														
ergpolite	infm	form	elev	humb																														
datpolite	infm	form	elev	humb																														
erggender	masc	fem	com	neut																														
datgender	masc	fem	com	neut																														
position	prenom	postnom	nom	free																														
subcat	intr	tran																																
verbform	fin	inf	sup	part	conv	vnoun	ger	gdv																										
mood	ind	imp	cnd	pot	sub	jus	prp	opt	des	nec	qot	adm																						
tense	pres	fut	past	aor	imp	pqp																												
voice	act	mid	pass	rcp	cau	int	antip	dir	inv																									
evident	fh	nfh																																
aspect	imp	perf	prosp	prog	hab	iter																												
strength	weak	strong																																
variant	short	long	0	1	2	3	4	5	6	7	8	9	a	b	c																			
style	arch	rare	form	poet	norm	coll	vrnc	slng	expr	derg	vulg																							
tagset	*any*																																	
other	*any*																																	



Interset



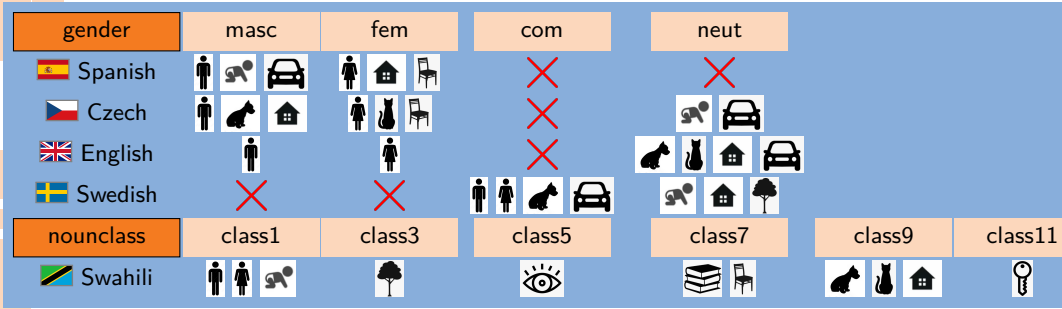
Interset

pos	noun	adj	num	verb	adv	adp	conj	part	int	punc	sym
morftype	com	prop	class							
nomtype	geo	pro	giv	ser	nat	com	pro	oth	col	sci	che	med	tec	cel	gov	jus	fin	env	cul	spo	hob
adjtype	pdt																				
prontype	prs	rcp	art	int	rel	exc	dem	emp	neg	ind	tot										
numtype	card	ord	mult	frac	sets	dist	range														
numform	word	digit	roman	combi																	
verbtype	aux	cop																			
advtype	man	loc																			
adpostype	prep	post																			
conjtype	coor	sub																			
parttype	mod	emp																			
punctype	peri	ques																			
poscode	inc	fin																			
poss	yes																				
nfrm	yes																				
foreign	yes																				
abbr	yes																				
hyph	yes																				
typo	yes																				
echo	rdp	ech																			
polarity	pos	neg																			
definite	ind	spec																			
gender	masc	fem																			
animacy	anim	hum																			
number	sing	dual																			
case	nom	gen																			
prepcase	npr	pre																			
degree	pos	cmp	sup																		
person	0	1	2																		
clusivity	in	ex																			
polite	infm	form	elev	humb																	
possgender	masc	fem	com	neut																	
possperson	1	2	3																		
possnumber	sing	dual	plur																		
possdualhead	sing	dual	plur																		
abspernum	1	2	3																		
ergpernum	1	2	3																		
datpernum	1	2	3																		
absnumber	sing	dual	plur																		
ergnumber	sing	dual	plur																		
datnumber	sing	dual	plur																		
absgcase	infm	form	elev	humb																	
ergpolite	infm	form	elev	humb																	
datpolite	infm	form	elev	humb																	
absgender	masc	fem	com	neut																	
ergpolite	masc	fem	com	neut																	
datgender	masc	fem	com	neut																	
position	prenom	postnom	nom	free																	
subcat	intr	tran																			
voiceyn	fin	inf	sup	part	conv	vnoun	ger	gdv													
mood	ind	imp	cnd	pot	sub	jus	prp	opt	des	nec	qot	adm									
tense	pres	fut	past	aor	imp	pup															
voice	act	mid	pass	rcp	cau	int	antip	dir	inv												
evident	fh	nfh																			
aspect	imp	perf	prosp	prog	hab	iter															
strength	weak	strong																			
variant	short	long	0	1	2	3	4	5	6	7	8	9	a	b	c						
style	arch	rare	form	poet	norm	coll	vrnc	sing	expr	derg	valg										
tagset	any																				
other	any																				

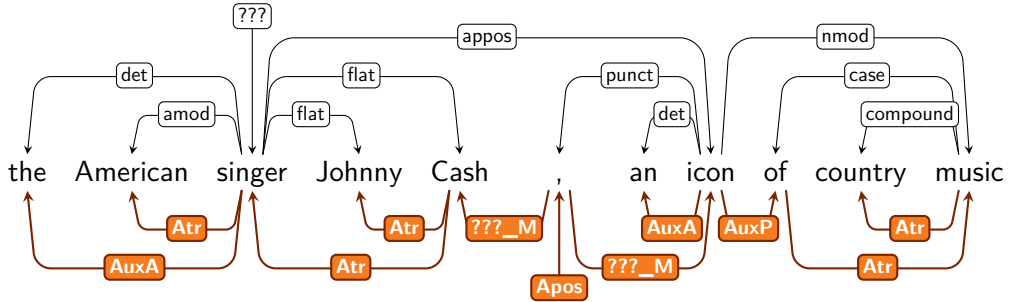
gender	masc	fem	com	neut
Spanish				
Czech				
English				
Swedish				

abe	erg	cau	ben	cns	equ	cmp

Interset


gender
masc
fem
com
neut
Spanish
Czech
English
Swedish
nounclass
class1
class3
class5
class7
class9
class11
Swahili

Syntax: Stanford vs. Prague Dependencies



- Change relation labels
- Bottom-up tree transformations

Daniel Zeman, *Ondřej Dušek, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zdeněk Žabokrtský, Jan Hajič (2014): HamleDT: Harmonized Multi-Language Dependency Treebank. In: Language Resources and Evaluation, ISSN 1574-020X, vol. 48, no. 4, pp. 601–637* 88 cit.

Universal Dependencies

- Joined our (Prague) forces with similar efforts at Stanford, Google, and Uppsala
- Defined universally-applicable annotation standard
  - Morphology from Interset
  - Syntax adapted from "Stanford Dependencies"

Marie-Catherine de Marneffe, Christopher Manning Joakim Nivre, ***Daniel Zeman*** *(2021): Universal Dependencies. In: Computational Linguistics, vol. 47, no. 2, pp. 255–308* 610 cit. WoS Q1

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, ... ***Daniel Zeman*** *(2016): Universal Dependencies v1: A Multilingual Treebank Collection. In: Proceedings of LREC, pp. 1659–1666* 1694 cit.

Universal Dependencies

- Joined our (Prague) forces with similar efforts at Stanford, Google, and Uppsala
- Defined universally-applicable annotation standard
  - Morphology from Interset
  - Syntax adapted from "Stanford Dependencies"
- Converted existing datasets to UD

Marie-Catherine de Marneffe, Christopher Manning Joakim Nivre, ***Daniel Zeman*** *(2021): Universal Dependencies. In: Computational Linguistics, vol. 47, no. 2, pp. 255–308* *610 cit. WoS Q1*

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, ... ***Daniel Zeman*** *(2020): Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In: Proceedings of LREC, pp. 4034–4043* *574 cit.*

Universal Dependencies

- Joined our (Prague) forces with similar efforts at Stanford, Google, and Uppsala
- Defined universally-applicable annotation standard
  - Morphology from Interset
  - Syntax adapted from "Stanford Dependencies"
- Converted existing datasets to UD

- Collecting new datasets for new languages
- Building infrastructure to lower the entry barrier for new data contributors
- Organizing shared tasks (evaluation campaigns for parsers)

Marie-Catherine de Marneffe, Christopher Manning Joakim Nivre, ***Daniel Zeman*** *(2021): Universal Dependencies. In: Computational Linguistics, vol. 47, no. 2, pp. 255–308* *610 cit. WoS Q1*

Daniel Zeman, *Jan Hajič, Martin Popel, ... Slav Petrov (2018): CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In: Proceedings of CoNLL. pp. 1–21* *671 cit.*

The UD Community

Daniel Zeman, *Joakim Nivre,* *Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Salih Furkan Akkurt, Gabriele Aleksandravičiūtė, Ika Alfina, Avner Algom, Khalid Alnajjar, Chiara Alzetta, Erik Andersen, Lene Antonsen, Tatsuya Aoyama, Katya Aplonova, Angelina Aquino, Carolina Aragon, Glyd Aranes, Maria Jesus Aranzabe, Bilge Nas Arcan, Þórunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Katla Ásgeirsdóttir, Deniz Baran Aslan, Cengiz Asmazoğlu, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Mariana Avelãs, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Starkaður Barkarson, Rodolfo Basile, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Shabnam Behzad, Juan Belieni, Kepa Bengoetxea, Ibrahim Benli, Yifat Ben Moshe, Ansu Berg, Gözde Berk, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Ésma Fatıma Bilgin Taşdemır, Kristín Bjarnadóttir, Verena Blaschke, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Johnatan Bonilla, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Anouck Braggaar, António Branco, Kristina Brokaitė, Aljoscha Burchardt, Marisa Campos, Marie Candito, Bernard Caron, Gauthier Caron, Catarina Carvalheiro, Rita Carvalho, Lauren Cassidy, Maria Clara Castro, Sérgio Castro, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Neslihan Cesur, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Liyanage Chamila, Shweta Chauhan, Yifei Chen, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Bermet Chontaeva, Jayeol Chun, Juyeon Chung, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Claudia Corbetta, Daniela Corbetta, Francisco Costa, Marine Courtin, Benoît Crabbé, Mihaela Cristescu, Vladimír Cvetkoski, Ingerid Løyning Dale, Philemon Daniel, Elizabeth Davidson, Leonel Figueiredo de Alencar, Mathieu Dehouck, Martina de Laurentiis, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Roberto Antonio Díaz Hernández, Carly Dickerson, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Hoa Do, Kaja Dobrovoljc, Caroline Döhmer, Adrian Doyle, Timothy Dozat, Kira Droganova, Magali Sanches Duran, Puneet Dwivedi, Christian Ebert, Hanne Eckhoff, Masaki Eguchi, Sandra Eiche, Roald Eiselen, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Soudabeh Eslami, Farah Essaidi, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Federica Favero, Jannatul Ferdaousi, Marília Fernanda, Hector Fernandez Alcalde, Amal Fethi, Jennifer Foster, Theodorus Fransen, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Edith Galy, Federica Gamba, Marcos Garcia, Moa Gärdenfors, Tanja Gaustad, Efe Eren Genç, Fabrício Ferraz Gerardi, Kim Gerdes, Luke Gessler, Filip Ginter, Gustavo Godoy, Jakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Kirian Guiller, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Hinrik Hafsteinsson, Jan Hajič, Jan Hajič, jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Muhammad Yudistira Hanifmuti, Takahiro Harada, Sam Hardwick, Kim Harris, Naïma Hassert, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Diana Hoefels, Peter Hohle, Yidi Huang, Marivel Huerta Mendez, Jena Hwang, Takumi Ikeda, Anastasia Iliadou, Anton Karl Ingason, Radu Ion, Elena Irimia, Olájídé Ishola, Artan Islamaj, Kaoru Ito, Federica Iurescia, Sandra Jagodzińska, Siratun Jannat, Tomáš Jelínek, Apoorva Jha, Katharine Jiang, Mayank Jobanputra, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Hüner Kaşıkara, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, Ritván Karahóǧa, Andre Kåsen, Tolga Kayadelen, Sarveswaran Kengatharaiyer, Václava Kettnerová, Lilit Kharatyan, Jesse Kirchner, Elena Klementieva, Elena Klyachko, Petr Kocharov, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Mehmet Köse, Alexey Koshevoy, Natalia Kotsyba, Barbara Kovačić, Jolanta Kovalevskaitė, Simon Krek, Parameswari Krishnamurthy, Sandra Kübler, Adrian Kuqi, Oğuzhan Kuyrukçu, Asli Kuzgun, Sookyoung Kwak, Kris Kyle, Kăbi Laan, Veronika Laippala, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phương Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Lauren Levine, Cheuk Ying Li, Josie Li, Keying Li, Yixuan Li, Yuan Li, KyungTae Lim, Bruna Lima Padovani, Yi-Ju Jessica Lin, Krister Lindén, Yang Janet Liu, Nikola Ljubešić, Irina Lobzhanidze, Olga Loginova, Lucelene Lopes, Stefano Lusito, Anne-Marie Lutgen, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Menel Mahamdi, Jean Maillard, Ilya Makarchuk, Aibek Makazhanov, Francesco Mambrini, Michael Mandl, Christopher Manning, Ruli Manurung, Büşra Marşan, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Stella Markantonatou, Héctor Martínez Alonso, Lorena Martín Rodríguez, André Martins, Cláudia Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Alessandro Mazzei, Ryan McDonald, Sarah McGuinness, Maitrey Mehta, Pierre-André Ménard, Gustavo Mendonça, Tatiana Merzhevich, Paul Meurer, Niko Miekka, Emilia Milano, Aaron Miller, Karina Mischenkova, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Judit Molnár, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Giovanni Moretti, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhlé, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Manuela Nevaci, Lương Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Victor Noroman, Alireza Nourian, Maria das Graças Volpe Nunes, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Hulda Óladóttir, Adédayọ̀ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Noam Ordan, Petya Osenova, Robert Östling, Annika Ott, Lilja Øvrelid, Şaziye Betül Özates, Merve Özçelik, Arzucan Özgür, Balkız Öztürk Başaran, Teresa Paccosi, Alessio Palmero Aprosio, Anastasia Panova, Thiago Alexandre Salgueiro Pardo, Hyunji Hayley Park, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Giulia Pedonese, Angelina Peljak-Łapińska, Siyao Peng, Siyao Logan Peng, Rita Pereira, Silvia Pereira, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Andrea Peverelli, Jason Phelan, Claudel Pierre-Louis, Jussi Piitulainen, Yuval Pinter, Clara Pinto, Rodrigo Pintucci, Tommi A Pirinen, Emily Pitler, Magdalena Plamada, Barbara Plank, Alistair Plum, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Rigardt Pretorius, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Robert Pugh, Tiina Puolakainen, Christoph Purschke, Sampo Pyysalo, Peng Qi, Andreia Querido, Andriela Rääbis, Alexandre Rademaker, Mizanur Rahoman, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Joana Ramos, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Mathilde Regnault, Georg Rehm, Arji Riabi, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Putri Rizqiyah, Luisa Rocha, Eirikur Rögnvaldsson, Ivan Roksandic, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Davide Rovati, Ben Rozonoyer, Olga Rudina, Jack Rueter, Paolo Ruffolo, Kristján Rúnarsson, Shoval Sadde, Pegah Safari, Aleksi Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Xuîa Sánchez-Rodríguez, Manuela Sanguinetti, Ezgi Sanıyar, Dage Särg, Marta Sartor, Albina Sarymsakova, Mitsuya Sasaki, Baiba Saulīte, Agata Savary, Yanin Sawanakunanon, Shefali Saxena, Kevin Scannell, Salvatore Scarlata, Emmanuel Schang, Nathan Schneider, Sebastian Schuster, Lane Schwartz, Djamé Seddah, Wolfgang Seeker, Sven Sellmer, Mojgan Seraji, Syeda Shahzadi, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Yana Shishkina, Muh Shohibussirri, Maria Shvedova, Janine Siewert, Einar Freyr Sigurðsson, João Silva, Aline Silveira, Natalia Silveira, Sara Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Haukur Barri Símonarson, Kiril Simov, Dmitri Sitchinava, Ted Sither, Aaron Smith, Isabela Soares-Bastos, Per Erik Solberg, Barbara Sonnenhauser, Shafi Sourov, Rachele Sprugnoli, Vivian Stamou, Steinþór Steingrímsson, Antonio Stella, Abishek Stephen, Milan Straka, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Shingo Suzuki, Daniel Swanson, Zsolt Szántó, Chihiro Taguchi, Dima Taji, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Dipta Tanaya, Mirko Tavoni, Samson Tella, Isabelle Tellier, Marinella Testori, Guillaume Thomas, Tarık Emre Tıraş, Sara Tonelli, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sveinbjörn Þórðarson, Vilhjálmur Þorsteinsson, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Elena Vagnoni, Sowmya Vajjala, Socrates Vak, Rob van der Goot, Martine Vanhove, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Uliana Vedenina, Giulia Venturi, Eric Villemonte de la Clergerie, Veronika Vincze, Aníshka Vissamsetty, Natalia Vlasova, Eleni Vlígouridou, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, John Wang, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Shira Wigderson, Sri Hartati Wijono, Vanessa Berwanger Wille, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Qishen Wu, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Arife Betül Yenice, Enes Yılandiloğlu, Olcay Taner Yıldız, Zhuoran Yu, Arlisa Yuliawati, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, He Zhou, Hanzhi Zhu, Yilun Zhu, Anna Zhuravleva, Rayan Ziane* *(2024): Universal Dependencies 2.14. LINDAT/CLARIAH Digital Repository.*

The UD Repository

UD v2.14:

- 31 language families
- 161 languages
- 283 treebanks
- 616 contributors
- 1.9 million sentences
- 32 million words
- 200 thousand downloads (all versions)

Current UD Languages

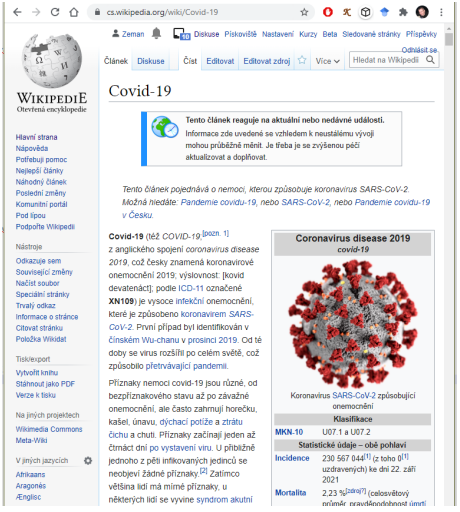
Information about language families (and genera for families with multiple branches) is mostly taken from WALS Online (IE = Indo-European).

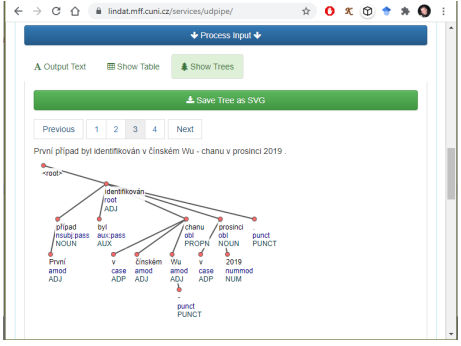
Abaza	1	<1K		Northwest Caucasian
Afrikaans	1	49K		IE, Germanic
Akkadian	2	25K		Afro-Asiatic, Semitic
Akuntsu	1	1K		Tupian, Tupari
Albanian	1	<1K		IE, Albanian
Amharic	1	10K		Afro-Asiatic, Semitic
Ancient Greek	2	416K		IE, Greek
Ancient Hebrew	1	39K		Afro-Asiatic, Semitic
Apurina	1	<1K		Arawakan
Arabic	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
Bororo	1	<1K		Bororoan
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 Philippine
Chinese	6	287K		Sino-Tibetan
Chukchi	1	6K		Chukotko-Kamchatkan
Classical Chinese	1	433K		Sino-Tibetan
Coptic	1	55K		Afro-Asiatic, Egyptian
Croatian	1	199K		IE, Slavic
Czech	5	2,247K		IE, Slavic
Danish	1	100K		IE, Germanic
Dutch	2	306K		IE, Germanic
English	10	726K		IE, Germanic
Erzya	1	20K		Uralic, Mordvin

Use in Digital Humanities

Linguists Can Parse and Search New Data

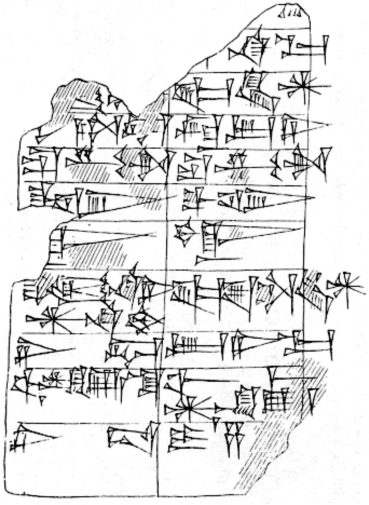
https://lindat.mff.cuni.cz/services/udpipe/



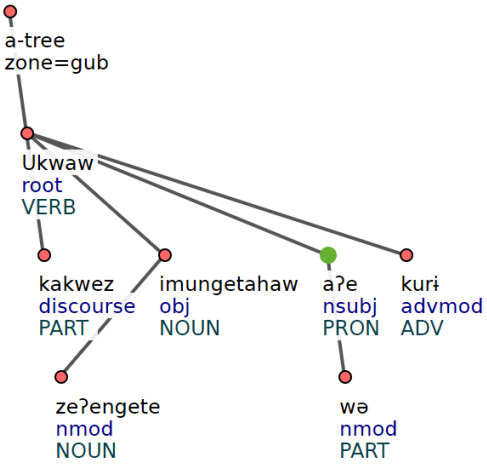


Historical Linguistics, Classical Languages

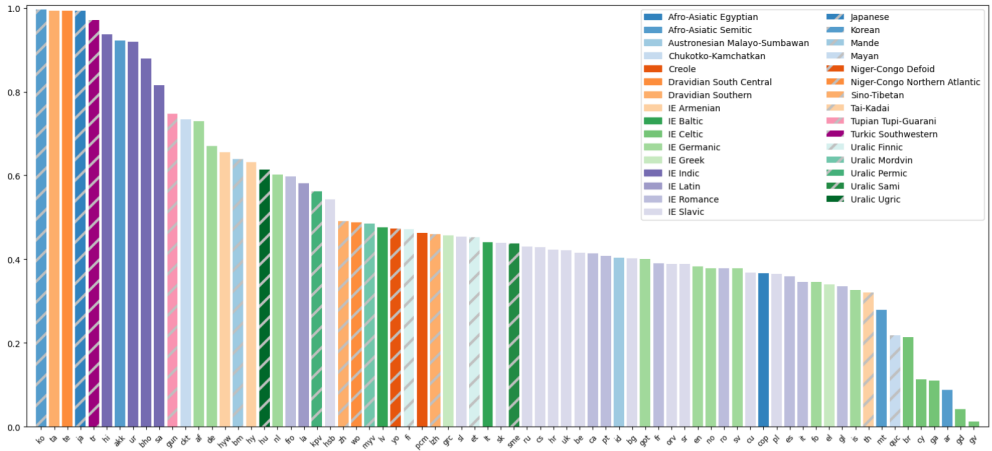
- Old Turkish
- Classical Chinese
- Sanskrit
- Akkadian
- Ancient Hebrew
- Coptic
- Ancient Greek
- Latin
- Old French
- Old Irish
- Gothic
- Old Church Slavonic



Documentation of Endangered Languages



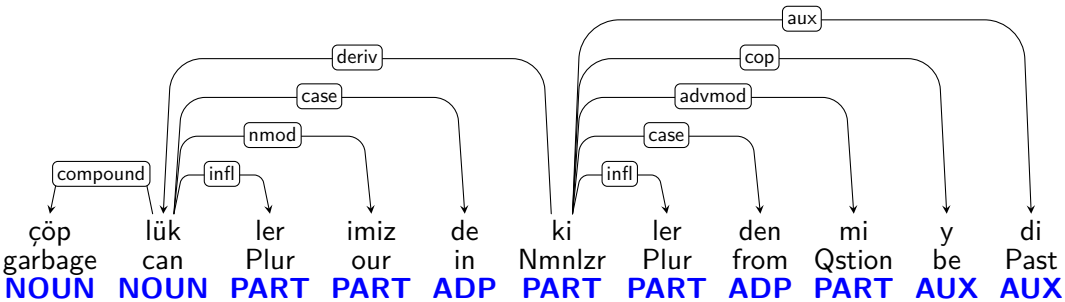
Linguistic Typology



Marie-Catherine de Marneffe, Joakim Nivre, ***Daniel Zeman*** *(2024): Function Words in Universal Dependencies. In: Linguistic Analysis, vol. 43, no. 3–4, pp. 549–588*

What's Next?

Subword Relations: Agglutinative Languages



"was it from those that were in our garbage cans?"

Universal Semantics?

Kira Droganova

Universal Semantic Roles

Dima Taji

Coreference in Deep UD

Federica Gamba

Uniform Meaning Representation for Latin

Minoo Nassajian

Uniform Meaning Representation for Persian

Universal Semantics?

Kira Droganova

Universal Semantic Roles

Dima Taji

Coreference in Deep UD

Federica Gamba

Uniform Meaning Representation for Latin

Minoo Nassajian

Uniform Meaning Representation for Persian

Diego Alves (defended 2023)

Computational Typological Analysis of Syntactic Structures

Parsing Accuracy (for Discussion)

Treebank	Model	UPOS	Feats	Lem	UAS	LAS
Czech PDT (cs_pdt)	UDPipe	99.18	97.23	**99.02**	93.33	91.31
	Lang	99.18	96.87	98.72	94.35	92.41
	UDify	99.18	96.85	98.56	94.73	92.88
	UDify+Lang	**99.24**	**97.44**	98.93	**95.07**	**93.38**
German GSD (de_gsd)	UDPipe	94.48	90.68	**96.80**	85.53	81.07
	Lang	94.77	91.73	96.34	87.54	83.39
	UDify	94.55	90.65	94.82	87.81	83.59
	UDify+Lang	**95.29**	**91.94**	96.74	**88.11**	**84.13**
English EWT (en_ewt)	UDPipe	96.29	97.10	**98.25**	89.63	86.97
	Lang	**96.82**	**97.27**	97.97	**91.70**	**89.38**
	UDify	96.21	96.17	97.35	90.96	88.50
	UDify+Lang	96.57	96.96	97.90	91.55	89.06
Spanish AnCora (es_ancora)	UDPipe	**98.91**	**98.49**	**99.17**	92.34	90.26
	Lang	98.60	98.14	98.52	92.82	90.52
	UDify	98.53	97.84	98.09	92.99	90.50
	UDify+Lang	98.68	98.25	98.68	**93.35**	**91.28**
French GSD (fr_gsd)	UDPipe	97.63	**97.13**	**98.35**	90.65	88.06
	Lang	**98.05**	96.26	97.96	92.77	90.61
	UDify	97.83	96.59	97.48	**93.60**	**91.45**
	UDify+Lang	97.96	96.73	98.17	93.56	91.45
Russian SynTagRus (ru_syntagrus)	UDPipe	**99.12**	**97.57**	**98.53**	93.80	92.32
	Lang	98.90	96.58	95.16	94.40	92.72
	UDify	98.97	96.35	94.43	94.83	93.13
	UDify+Lang	99.08	97.22	96.58	**95.13**	**93.70**
Belarusian HSE (be_hse)	UDPipe	93.63	73.30	87.34	78.58	72.72
	Lang	95.88	76.12	84.52	83.94	79.02
	UDify	**97.54**	**89.36**	85.46	**91.82**	**87.19**
	UDify+Lang	97.25	85.02	**88.71**	90.67	86.98
Buryat BDT (bxr_bdt)	UDPipe	40.34	32.40	58.17	32.60	18.83
	Lang	52.54	37.03	54.64	29.63	15.82
	UDify	**61.73**	**47.86**	**61.06**	**48.43**	**26.28**
	UDify+Lang	61.73	42.79	58.20	33.06	18.65
Upper Sorbian UFAL (hsb_ufal)	UDPipe	62.93	41.10	68.68	45.58	34.54
	Lang	73.70	46.28	58.02	39.02	28.70
	UDify	84.87	48.63	**72.73**	**71.55**	**62.82**
	UDify+Lang	**87.58**	**53.19**	71.88	71.40	60.65
Kazakh KTB (kk_ktb)	UDPipe	55.84	40.40	63.96	53.30	33.38
	Lang	73.52	46.60	57.84	50.38	32.61
	UDify	**85.59**	**65.14**	**77.40**	**74.77**	**63.66**
	UDify+Lang	81.32	60.50	67.30	69.16	53.14
Lithuanian HSE (lt_hse)	UDPipe	81.70	60.47	**76.89**	51.98	42.17
	Lang	83.40	54.34	58.77	51.23	38.96
	UDify	**90.47**	**68.96**	67.83	**79.06**	**69.34**
	UDify+Lang	84.53	56.98	58.21	58.40	39.91

Treebank		UPOS	Feats	Lem	UAS	LAS
Breton KEB	**br_keb**	63.67	46.75	53.15	63.97	40.19
Tagalog TRG	**tl_trg**	61.64	35.27	75.00	64.73	39.38
Faroese OFT	fo_oft	77.86	35.71	53.82	69.28	61.03
Naija NSC	pcm_nsc	56.59	52.75	97.52	47.13	33.43
Sanskrit UFAL	sa_ufal	40.21	18.45	37.60	41.73	19.80

Table 4: Test set results for zero-shot learning, i.e., no UD training annotations available. Languages that are pretrained with BERT are bolded.

Dan Kondratyuk, Milan Straka (2019): 75 Languages, 1 Model: Parsing Universal Dependencies Universally. In: Proceedings of EMNLP, Hong Kong