Zagreb

Universal Dependencies

Where do we stand and where do we want to go from here?

Daniel Zeman

July 3, 2023


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

What is Universal Dependencies?

"Universal Dependencies (UD) is a project developing cross-linguistically consistent treebank annotation for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective."

(`https://universaldependencies.org/introduction.html`)

What is Universal Dependencies?

"Universal Dependencies (UD) is a project developing cross-linguistically consistent treebank annotation for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective."

(`https://universaldependencies.org/introduction.html`)

A project – but also an annotation scheme, a data repository and a community.

What is Universal Dependencies?

"Universal Dependencies (UD) is a project developing cross-linguistically consistent treebank annotation for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective."

(`https://universaldependencies.org/introduction.html`)

A project – but also an annotation scheme, a data repository and a community.

What is Universal Dependencies?

"Universal Dependencies (UD) is a project developing cross-linguistically consistent treebank annotation for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective."

(`https://universaldependencies.org/introduction.html`)

A project – but also an annotation scheme, a data repository and a community.

What is Universal Dependencies?

"Universal Dependencies (UD) is a project developing cross-linguistically consistent treebank annotation for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective."

(`https://universaldependencies.org/introduction.html`)

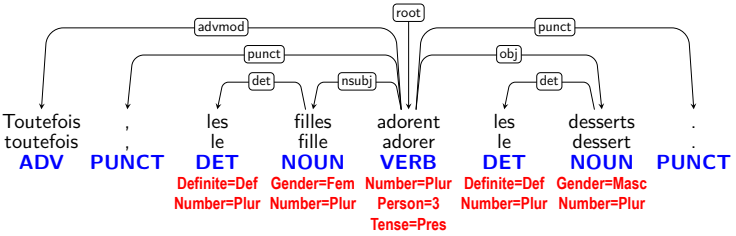
A project – but also an annotation scheme, a data repository and a community.

The UD Annotation Scheme

Basic design principles:

- Words as basic units
- Binary syntactic relations

Built on de facto standards



J. Nivre, M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajič, C. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty, and D. Zeman (2016) Universal Dependencies v1: A Multilingual Treebank Collection. In Proceedings of LREC, 1659–1666.

J. Nivre, M.-C. de Marneffe, F. Ginter, J. Hajič, C. D. Manning, S. Pyysalo, S. Schuster, F. Tyers, and D. Zeman (2020) Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In Proceedings of LREC, 4034–4043.

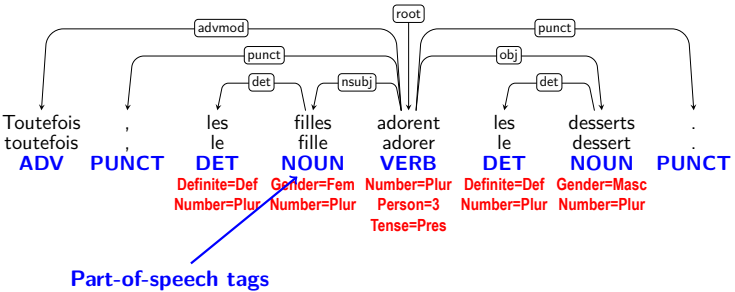
M.-C. de Marneffe, C. D. Manning, J.Nivre, and D. Zeman (2021) Universal Dependencies. Computational Linguistics, 47(2):255–308, 2021.

The UD Annotation Scheme

Basic design principles:

- Words as basic units
- Binary syntactic relations

Built on de facto standards



J. Nivre, M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajič, C. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty, and D. Zeman (2016) Universal Dependencies v1: A Multilingual Treebank Collection. In Proceedings of LREC, 1659–1666.

J. Nivre, M.-C. de Marneffe, F. Ginter, J. Hajič, C. D. Manning, S. Pyysalo, S. Schuster, F. Tyers, and D. Zeman (2020) Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In Proceedings of LREC, 4034–4043.

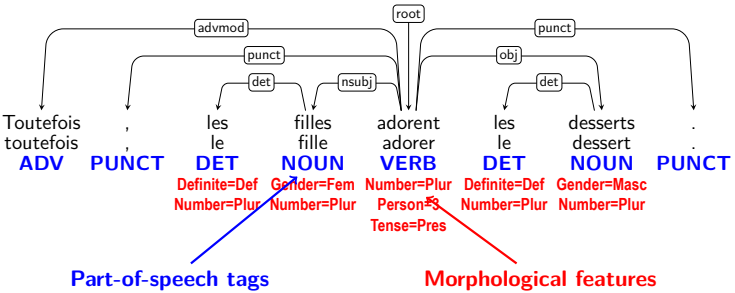
M.-C. de Marneffe, C. D. Manning, J.Nivre, and D. Zeman (2021) Universal Dependencies. Computational Linguistics, 47(2):255–308, 2021.

The UD Annotation Scheme

Basic design principles:

- Words as basic units
- Binary syntactic relations

Built on de facto standards



J. Nivre, M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajič, C. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty, and D. Zeman (2016) Universal Dependencies v1: A Multilingual Treebank Collection. In Proceedings of LREC, 1659–1666.

J. Nivre, M.-C. de Marneffe, F. Ginter, J. Hajič, C. D. Manning, S. Pyysalo, S. Schuster, F. Tyers, and D. Zeman (2020) Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In Proceedings of LREC, 4034–4043.

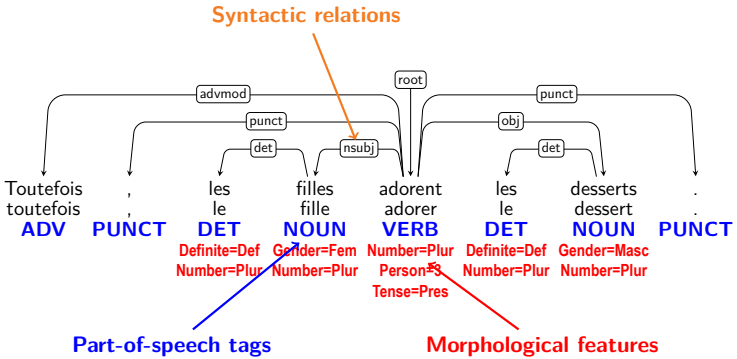
M.-C. de Marneffe, C. D. Manning, J.Nivre, and D. Zeman (2021) Universal Dependencies. Computational Linguistics, 47(2):255–308, 2021.

The UD Annotation Scheme

Basic design principles:

- Words as basic units
- Binary syntactic relations

Built on de facto standards



J. Nivre, M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajič, C. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty, and D. Zeman (2016) Universal Dependencies v1: A Multilingual Treebank Collection. In Proceedings of LREC, 1659–1666.

J. Nivre, M.-C. de Marneffe, F. Ginter, J. Hajič, C. D. Manning, S. Pyysalo, S. Schuster, F. Tyers, and D. Zeman (2020) Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In Proceedings of LREC, 4034–4043.

M.-C. de Marneffe, C. D. Manning, J.Nivre, and D. Zeman (2021) Universal Dependencies. Computational Linguistics, 47(2):255–308, 2021.

The UD Repository

UD v2.12:

- 30 language families
- 141 languages
- 245 treebanks
- 553 contributors
- 1.8 million sentences
- 31 million words

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

The UD Community

Treebank Users

Treebank Developers

Universal Guidelines Group

Word Segmentation

Let's go to the sea.

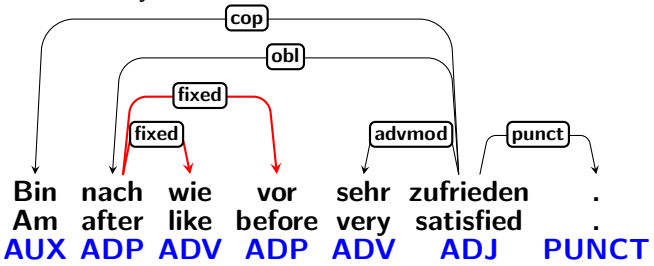
Vámonos	**al**	**mar**	**.**	**Vamos**	**nos**	**a**	**el**	**mar**	**.**
VERB?	**X**	**NOUN**	**PUNCT**	**VERB**	**PRON**	**ADP**	**DET**	**NOUN**	**PUNCT**

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

Fixed Expressions

One syntactic word spans several orthographic words?

I am still very satisfied.



Part-of-Speech Tags

http://universaldependencies.org/u/pos/index.html

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

- Taxonomy of 17 universal POS tags
- All languages use the same inventory
  - Not all tags have to be used by all languages
  - Need extensions? Use features!

Features

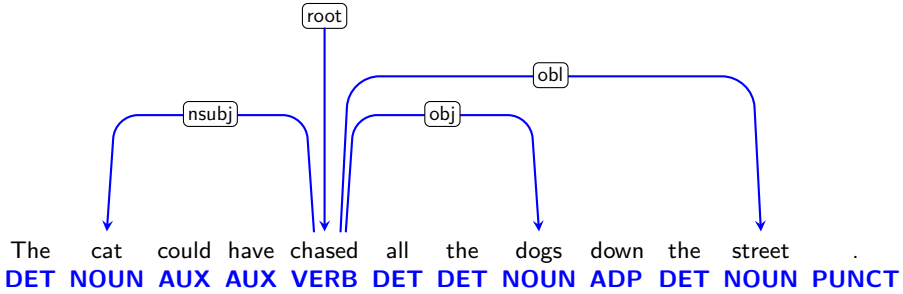
Lexical	Inflectional ("Nominal")	Inflectional ("Verbal, Pronominal")
PronType	Gender	VerbForm
NumType	Animacy	Mood
Poss	NounClass	Tense
Reflect	Number	Aspect
Foreign	Case	Voice
	Definite	Evident
	Degree	Polarity
Abbr		Person
Typo		Polite
		Clusivity

- 24 features, each with a number of possible *values*
- Languages select relevant features
- May add language-specific features or values

Syntax

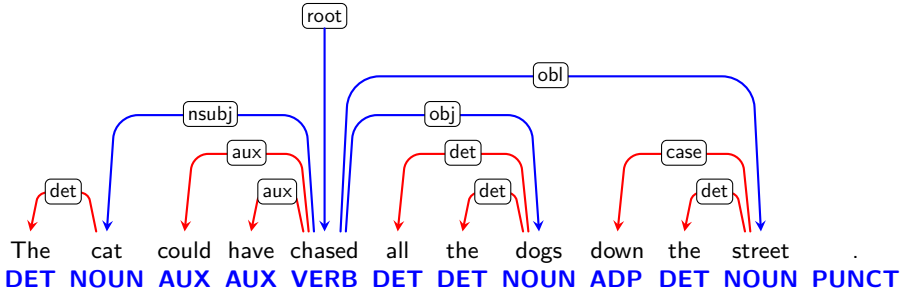
The	cat	could	have	chased	all	the	dogs	down	the	street	.
DET	**NOUN**	**AUX**	**AUX**	**VERB**	**DET**	**DET**	**NOUN**	**ADP**	**DET**	**NOUN**	**PUNCT**

Syntax



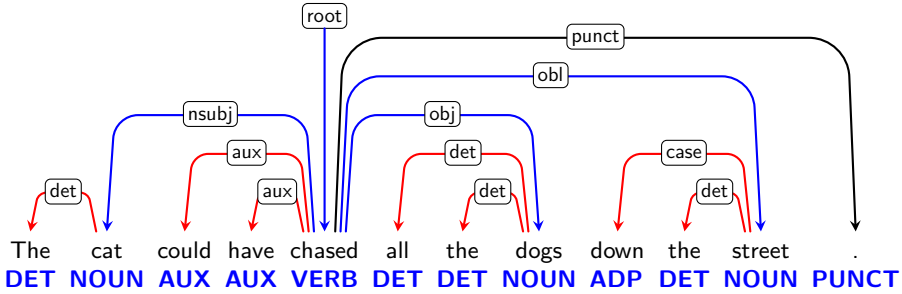
- Content words are related by dependency relations

Syntax



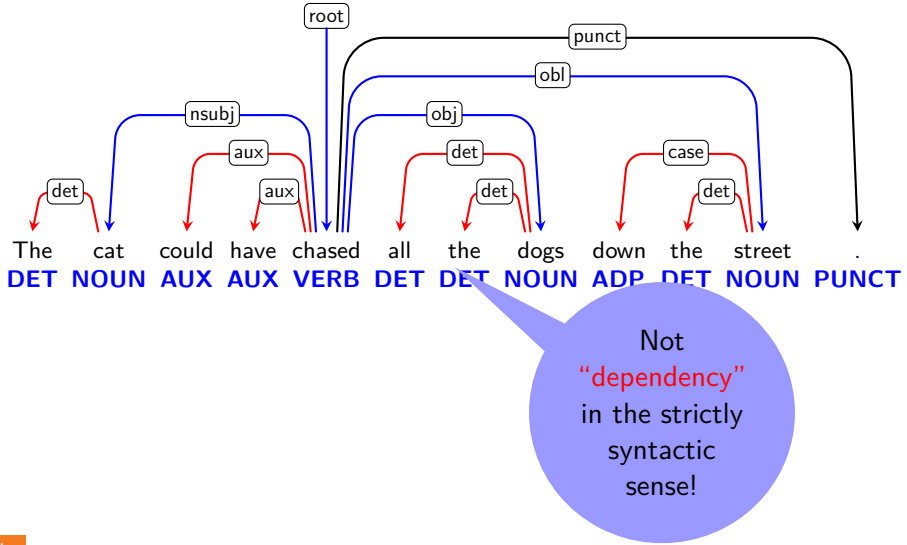
- Content words are related by dependency relations
- Function words attach to closest content words

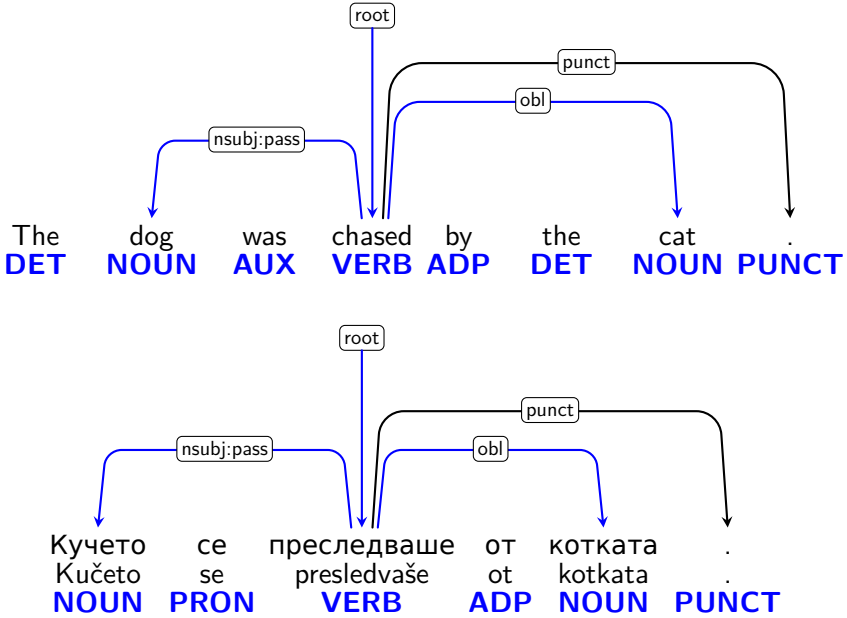
Syntax

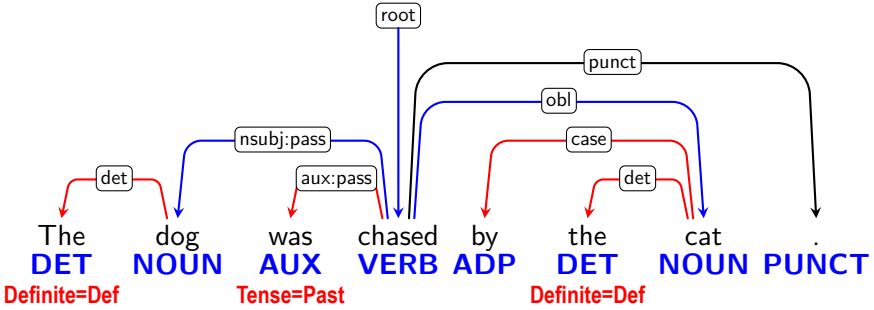


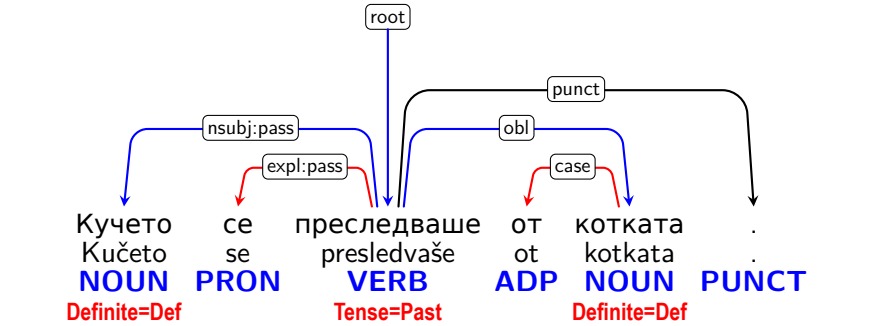
- Content words are related by dependency relations
- Function words attach to closest content words
- Punctuation attach to head of phrase or clause

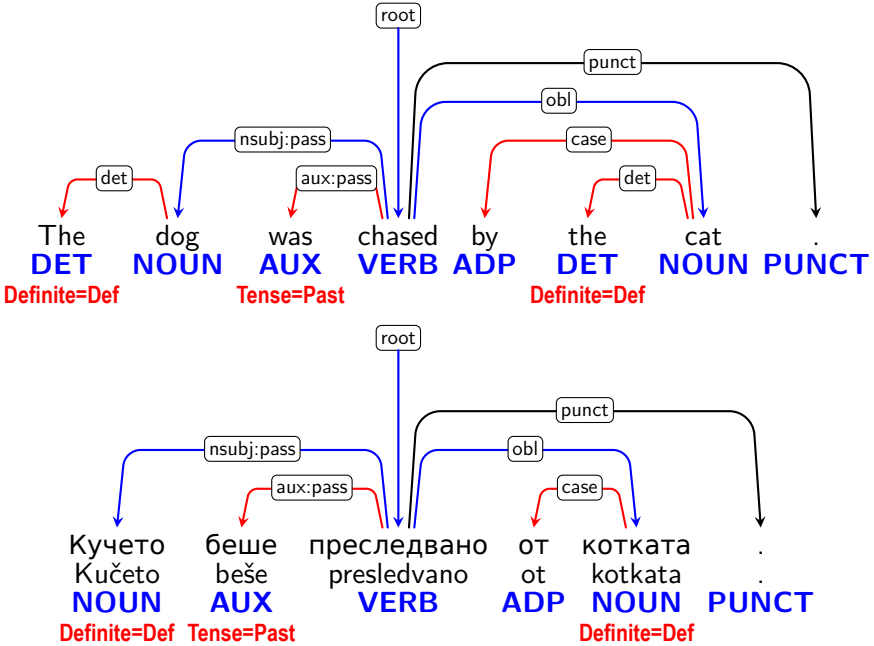
Syntax

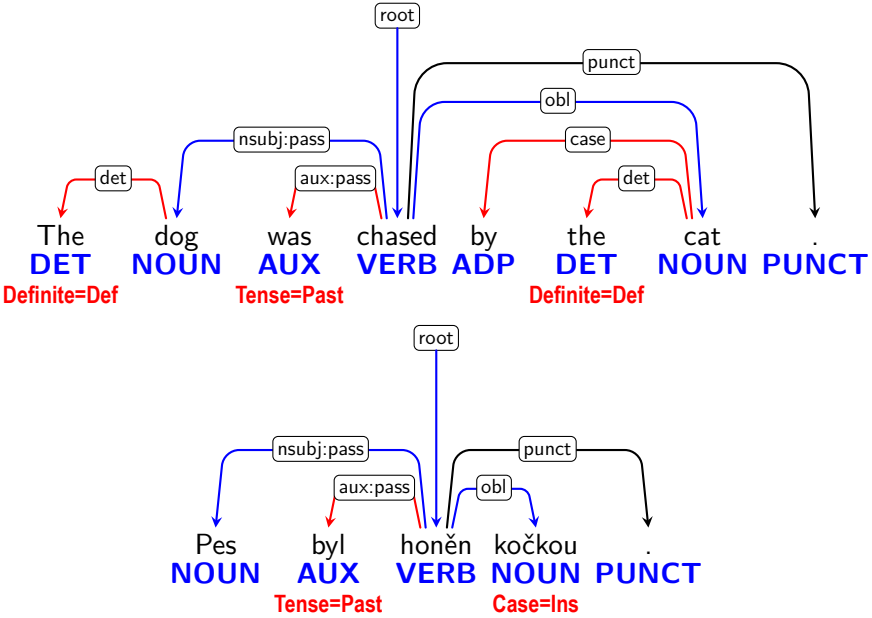

root
punct
obl
nsubj
obj
aux
det
case
det
aux
det
det
The cat could have chased all the dogs down the street .
DET NOUN AUX AUX VERB DET DET NOUN ADP DET NOUN PUNCT
Not "dependency" in the strictly syntactic sense!


root
punct
obl
nsubj:pass
The dog was chased by the cat .
DET NOUN AUX VERB ADP DET NOUN PUNCT
root
punct
nsubj:pass
obl
Кучето се преследваше от котката .
Kučeto se presledvaše ot kotkata .
NOUN PRON VERB ADP NOUN PUNCT


root
punct
obl
nsubj:pass
case
det
aux:pass
det
The dog was chased by the cat .
DET NOUN AUX VERB ADP DET NOUN PUNCT
Definite=Def
Tense=Past
Definite=Def


root
punct
nsubj:pass
obl
expl:pass
case
Кучето се преследваше от котката .
Kučeto se presledvaše ot kotkata .
NOUN PRON VERB ADP NOUN PUNCT
Definite=Def
Tense=Past
Definite=Def


root
punct
obl
nsubj:pass
case
det
aux:pass
det
The dog was chased by the cat .
DET NOUN AUX VERB ADP DET NOUN PUNCT
Definite=Def
Tense=Past
Definite=Def
root
punct
nsubj:pass
obl
aux:pass
case
Кучето беше преследвано от котката .
Kučeto beše presledvano ot kotkata .
NOUN AUX VERB ADP NOUN PUNCT
Definite=Def
Tense=Past
Definite=Def


root
punct
obl
nsubj:pass
case
det
aux:pass
det
The dog was chased by the cat .
DET NOUN AUX VERB ADP DET NOUN PUNCT
Definite=Def
Tense=Past
Definite=Def
root
nsubj:pass
punct
aux:pass
obl
Pes byl honěn kočkou .
NOUN AUX VERB NOUN PUNCT
Tense=Past
Case=Ins

Dependents of Clauses (Verbal or Not)

	Nominal	Clausal	Modifier	Function
Core	**nsubj**	csubj		
Non-Core	**obl**	advcl	advmod	aux
	vocative		discourse	cop
	dislocated			mark
	expl			

Dependents of Verbs, Adjectives and Adverbs

	Nominal	Clausal	Modifier
Core	**obj**	ccomp	
	iobj	xcomp	
Non-Core	**obl**	advcl	advmod
	expl		

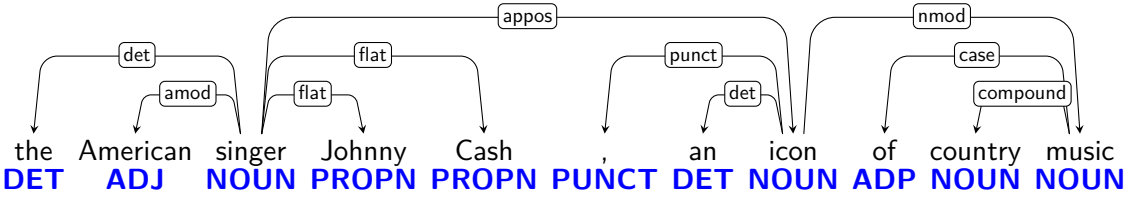
Dependents of Nominals

Nominal	Clausal	Modifier	Function
nmod	acl	amod	det
appos		nummod	case

Noun Phrase

Dependents of Nominals

Nominal	**Clausal**	**Modifier**	**Function**
nmod	acl	amod	det
appos		nummod	case
compound			clf
flat			



Language-specific Relation Subtypes

- Language-specific relations are subtypes of universal relations added to capture important phenomena
- Subtyping permits us to "back off" to universal relations

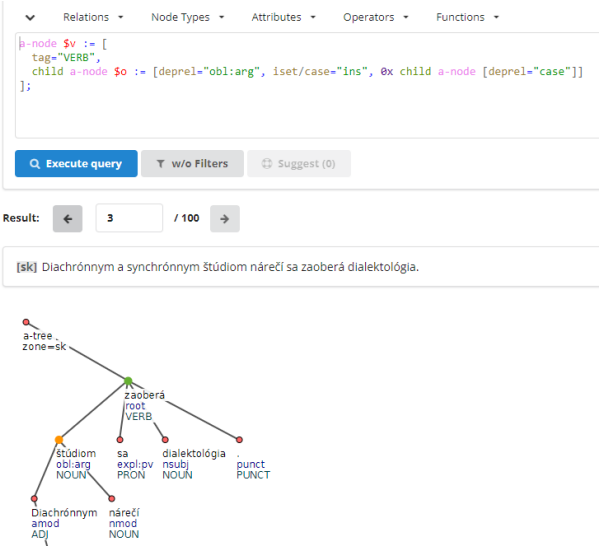
Language-specific Relation Subtypes

Relation	Explanation
acl:relcl	Relative clause (the boy who lived)
compound:prt	Verb particle (dress up)
nmod:poss	Possessive nominal (Mary 's book)
obl:agent	Agent in passive (saved by the bell)
cc:preconj	Preconjunction (both ... and)
det:predet	Predeterminer (all those ...)

Use in Digital Humanities

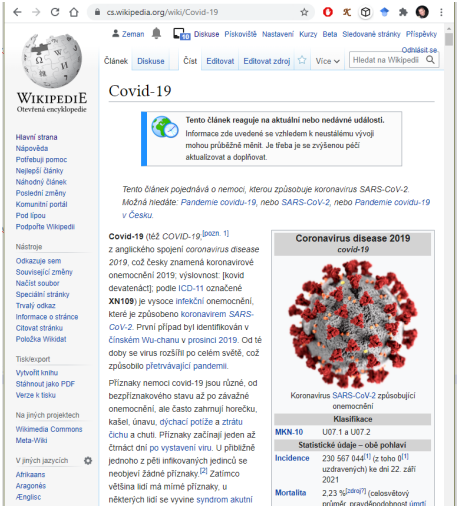
Linguists Can Search Treebanks

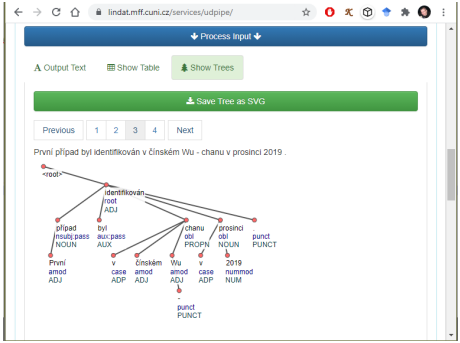
https://lindat.mff.cuni.cz/services/pmltq/



Linguists Can Parse and Search New Data

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



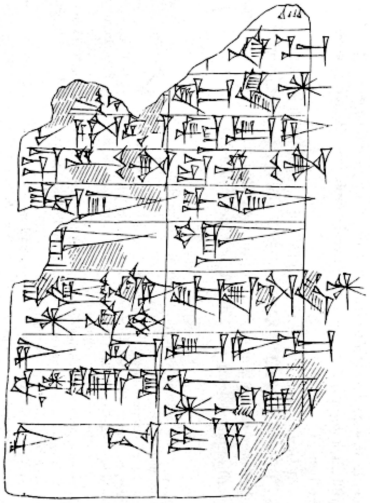


Language Learning

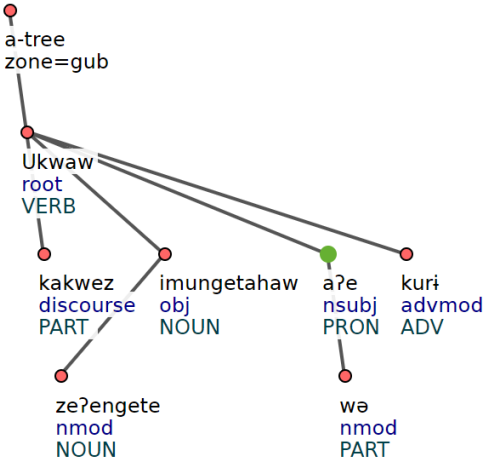
- Check grammar usage in the corpus
- Learner corpora

Historical Linguistics, Classical Languages

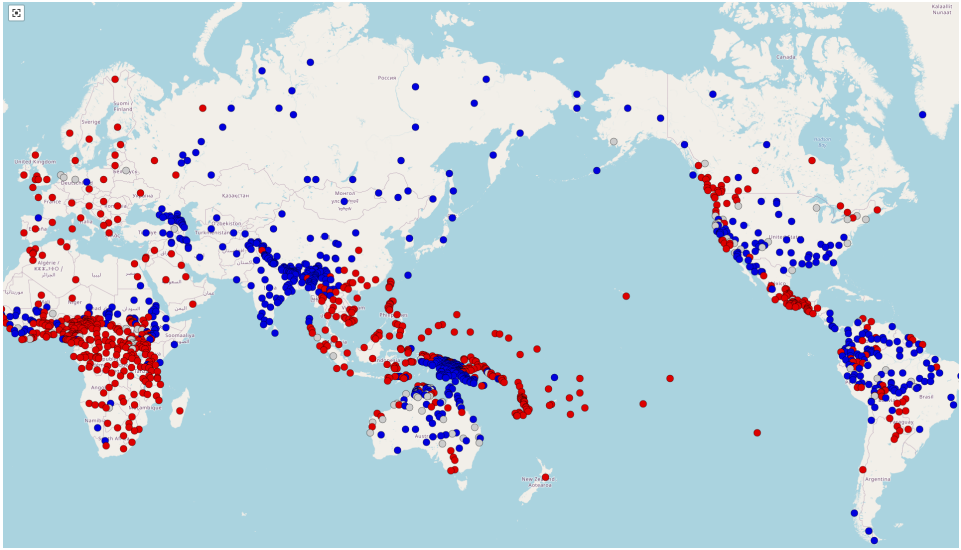
- Old Turkish
- Classical Chinese
- Sanskrit
- Akkadian
- Ancient Hebrew
- Coptic
- Ancient Greek
- Latin
- Old French
- Old Irish
- Gothic
- Old Church Slavonic



Documentation of Endangered Languages



Linguistic Typology



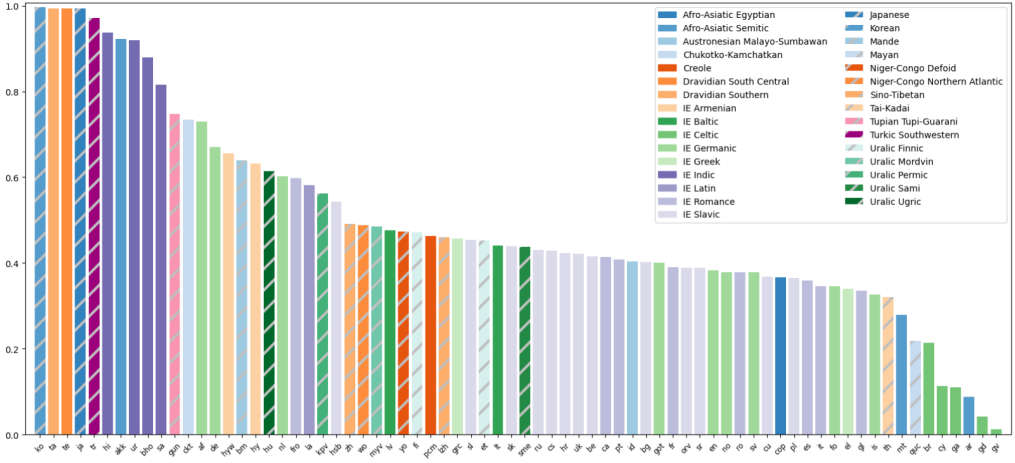


Figure 7 Percentage of head-final dependencies. Each bar is one language.

So Where Are We Now?

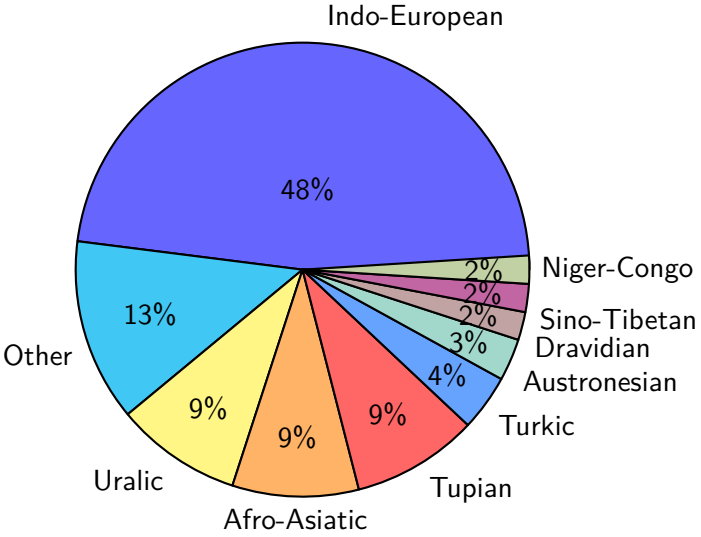
Achievements:

- Amazing growth in terms of contributors and treebanks
- Data sets widely used for research in NLP and linguistics

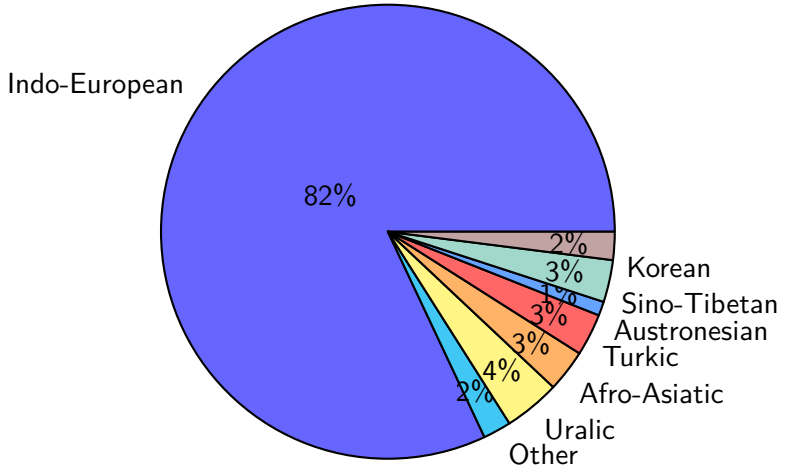
Challenges:

- Annotation scheme and documentation can be improved
- Data sets are uneven in terms of quality and quantity
- Extensions: what to do with demand to annotate extra stuff?

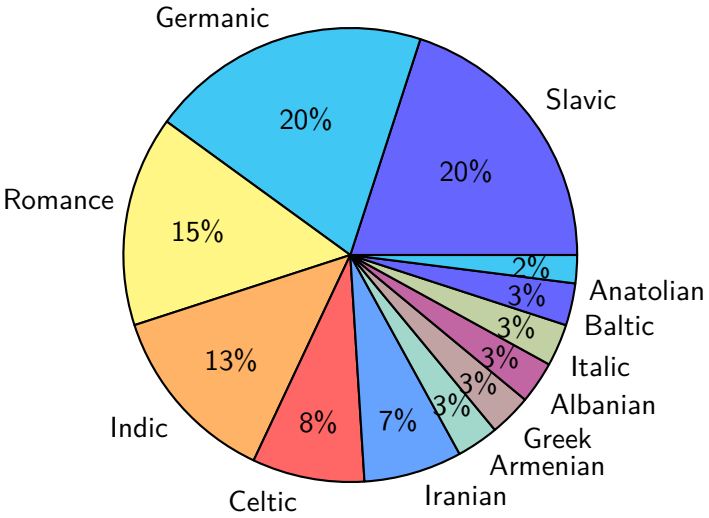
Languages per Families



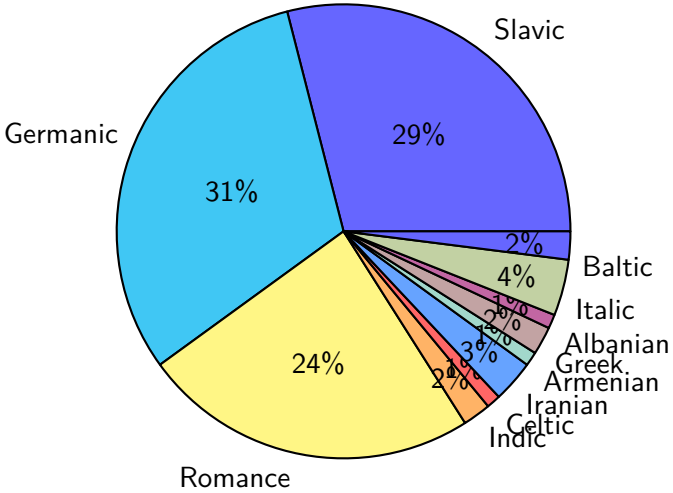
Words per Families



Languages per Indo-European Genuses



Words per Indo-European Genuses



Extensions

- Enhanced Universal Dependencies 

Extensions

- Enhanced Universal Dependencies
- Deep Universal Dependencies

Extensions

- Enhanced Universal Dependencies
- Deep Universal Dependencies
- Universal Proposition Banks

Universal
PropBank

Extensions

- Enhanced Universal Dependencies
- Deep Universal Dependencies
- Universal Proposition Banks
- CorefUD
- Named Entities

Extensions

- Enhanced Universal Dependencies
- Deep Universal Dependencies
- Universal Proposition Banks
- CorefUD
- Named Entities
- Surface Syntactic UD (SUD)

Extensions

- Enhanced Universal Dependencies
- Deep Universal Dependencies
- Universal Proposition Banks
- CorefUD
- Named Entities
- Surface Syntactic UD (SUD)
- PARSEME

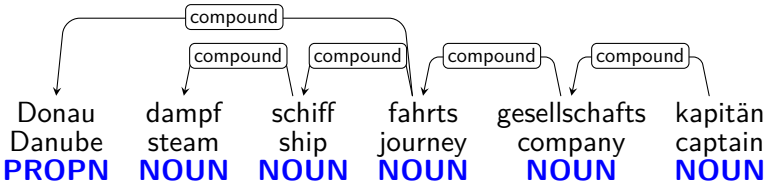
Universal
PropBank

PARSEME

Extensions

- Enhanced Universal Dependencies
- Deep Universal Dependencies
- Universal Proposition Banks
- CorefUD
- Named Entities
- Surface Syntactic UD (SUD)
- PARSEME
- UniMorph+UD, sub- and superwords

Subword Relations: Compounds



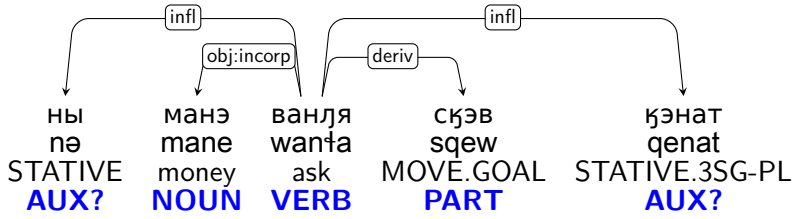
"captain of the Danube steamship company"

Compounds in Parallel UD

PUD sentence id w05003050

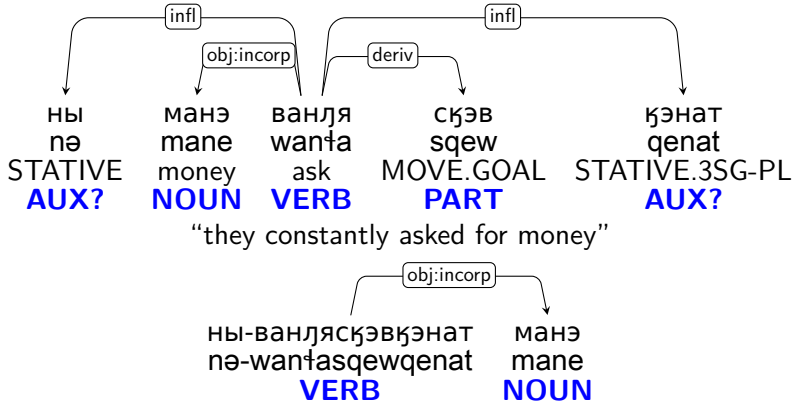
- [en] Under his administration, the female vote was approved.
- [sv] Under hans administration godkändes den kvinnliga rösträtten.
- [de] Unter seiner Regierung wurde das Frauenwahlrecht eingeführt.
- [es] En su gestión se decretó el sufragio femenino.
  - sufragio = suffrage
- [tr] Yönetimi sırasında kadınların oy kullanması kabul edilmiş.
  - = Administration during women-of vote to-use acceptance been-taken
- [cs] Pod jeho správou směly volit ženy.
  - (Pod jeho správou bylo zavedeno volební právo pro ženy.)

Subword Relations: Incorporation

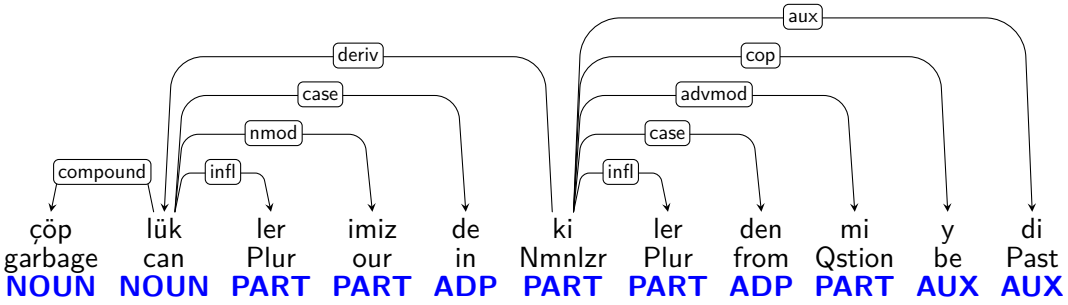


"they constantly asked for money"

Subword Relations: Incorporation

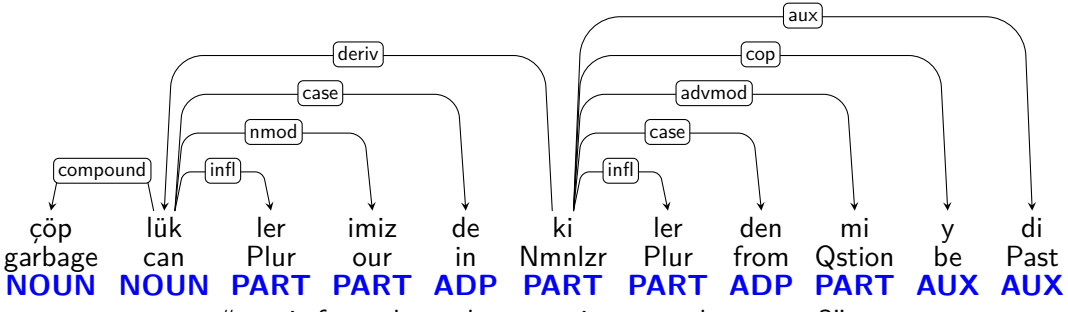


Subword Relations: Agglutinative Languages

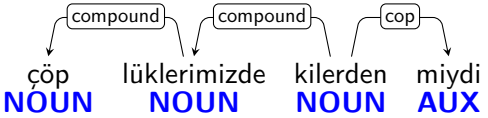


"was it from those that were in our garbage cans?"

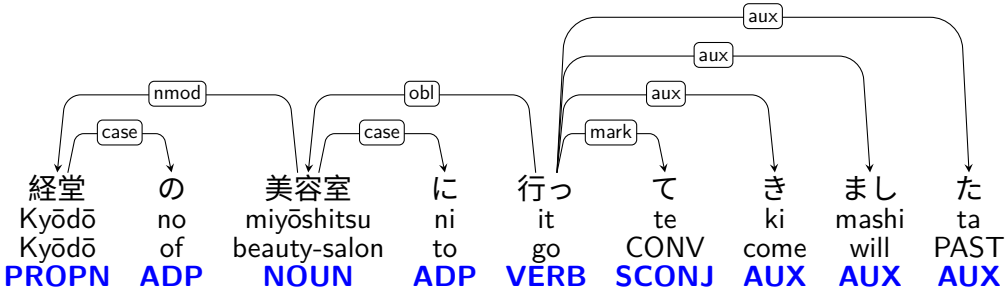
Subword Relations: Agglutinative Languages



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

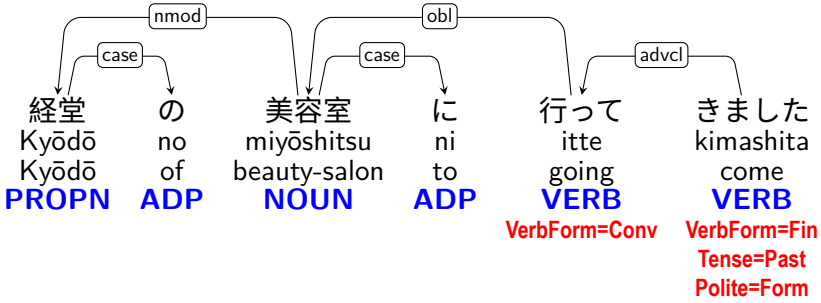


Subword Relations: Long and Short Words in Japanese



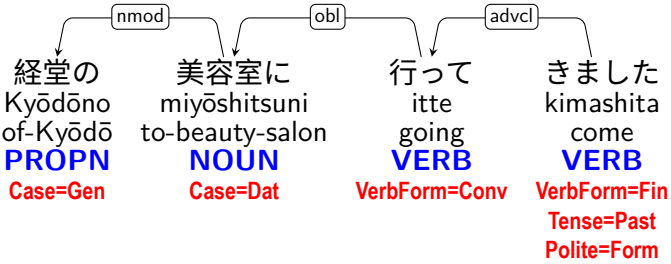
"I went to the beauty salon of Kyōdō [, Beyond-R.]"

Subword Relations: Long and Short Words in Japanese



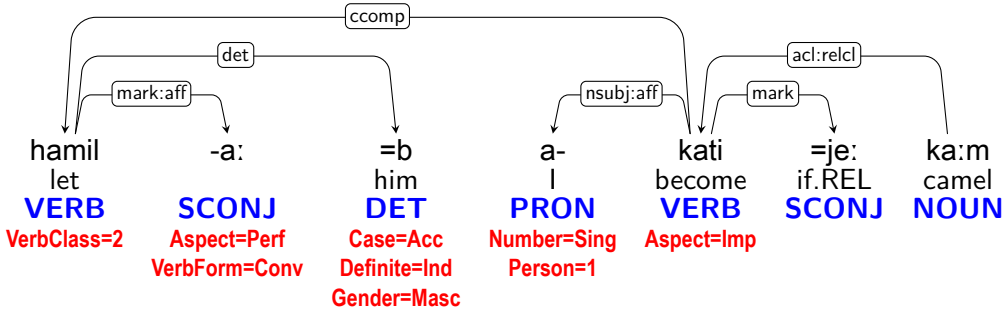
"I went to the beauty salon of Kyōdō [, Beyond-R.]"

Subword Relations: Long and Short Words in Japanese



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

Subword Relations: Fieldwork Glosses



“a camel that I had let loose”


Superword Features

- In UD, morphological features apply to individual words.
  - Future tense in Spanish and German: no **Tense=Fut** in German!

Dormirá	Er	wird	schlafen
He-will-sleep	He	will	sleep
VERB	**PRON**	**AUX**	**VERB**
VerbForm=Fin	PronType=Prs	VerbForm=Fin	VerbForm=Inf
Mood=Ind	Number=Sing	Mood=Ind	
Tense=Fut	Person=3	Tense=Pres	
Number=Sing	Gender=Masc	Number=Sing	
Person=3	Case=Nom	Person=3	


Superword Features

- In UD, morphological features apply to individual words.
  - Future tense in Spanish and German: no **Tense=Fut** in German!
- Can we also add features to "phrases"?

Dormirá	Er	wird schlafen
He-will-sleep	He	will sleep
VERB	**PRON**	**VERB**
VerbForm=Fin	PronType=Prs	VerbForm=Fin
Mood=Ind	Number=Sing	Mood=Ind
Tense=Fut	Person=3	Tense=Fut
Number=Sing	Gender=Masc	Number=Sing
Person=3	Case=Nom	Person=3

Summary

- Universal Dependencies
  - Unified annotation for all languages
    - Language-specific extensions
  - Content words higher than function words … better parallelism
  - Words as the main unit (tree nodes)
- Extension of UD: above and below word level
  - UD-like relations between subword units
  - UD-like features for groups of words (“phrases”)

https://universaldependencies.org/

Thanks!
Hvala!

https://universaldependencies.org/