Introduction to Dependency Grammar and Treebanks

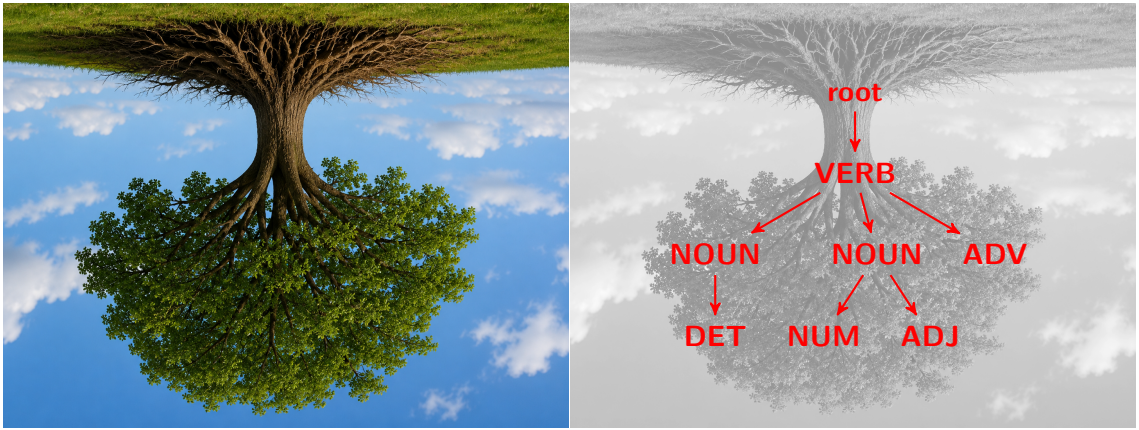
Daniel Zeman

July 20, 2026

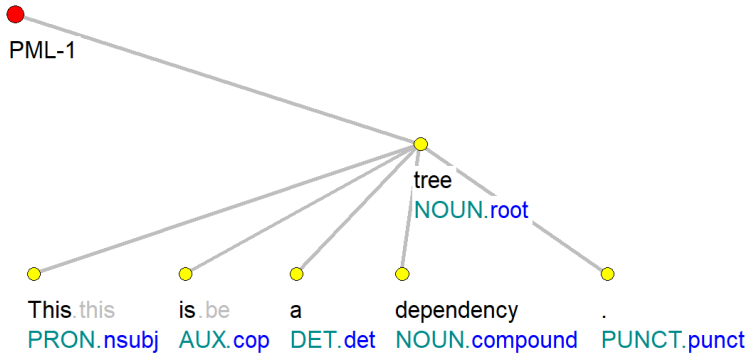
Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

https://ufal.mff.cuni.cz/courses/dz-2026-07-bh

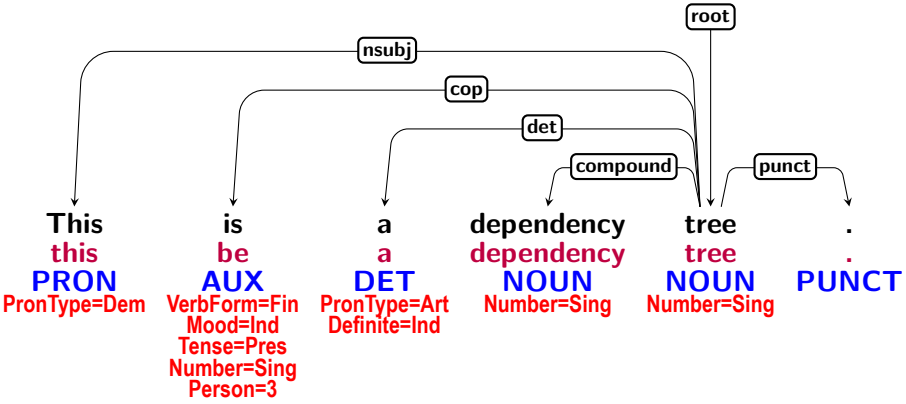
Syntactic Trees



Syntactic Trees



Syntactic Trees



Treebanks

Users (Simplified)

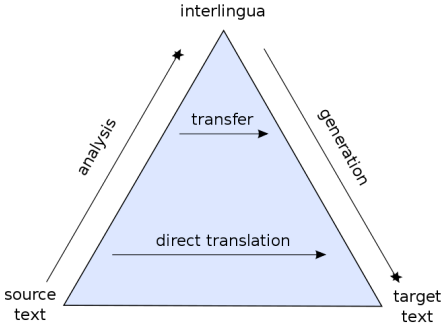
LT engineers

Users (Simplified)

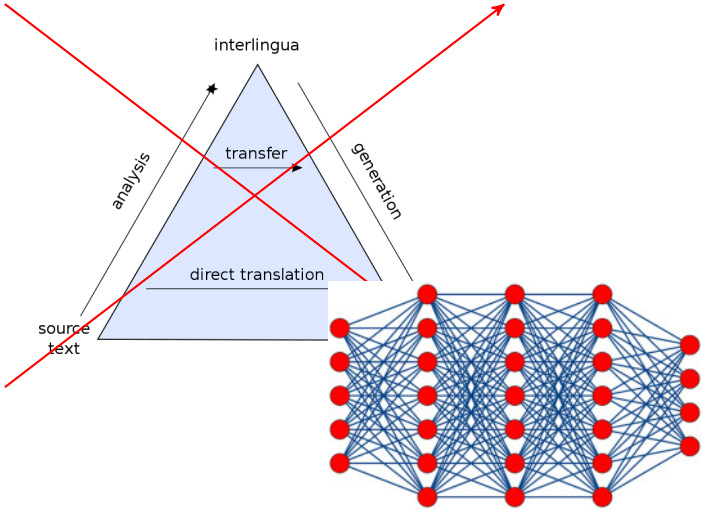
LT engineers

& linguists

Use in Language Technology?



Use in Language Technology?

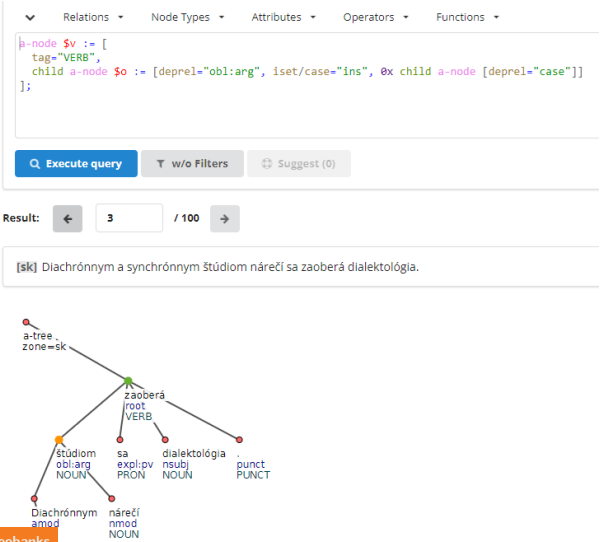


Off-the-Shelf Tools

- UDPipe (`https://lindat.mff.cuni.cz/services/udpipe/`)
  - `https://ufal.mff.cuni.cz/udpipe`
- Stanza (`https://stanfordnlp.github.io/stanza/`)

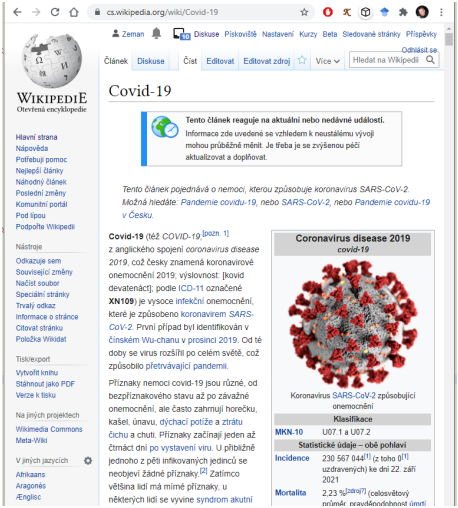
Linguists Can Search Treebanks

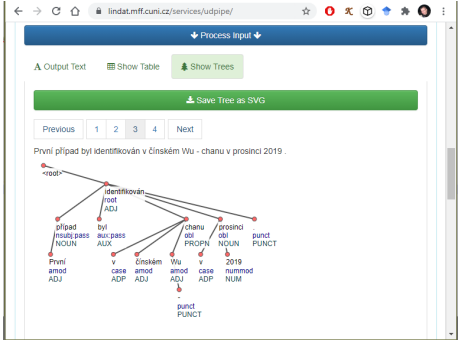
https://lindat.mff.cuni.cz/services/pmltq/



Linguists Can Parse and Search New Data

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



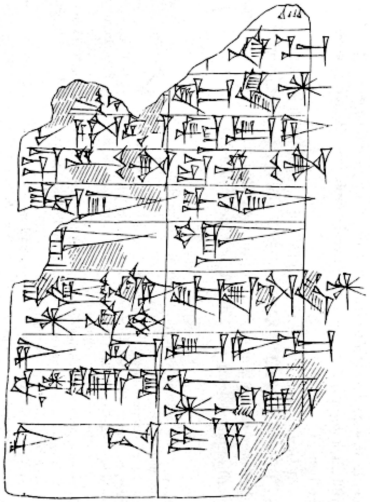


Language Learning

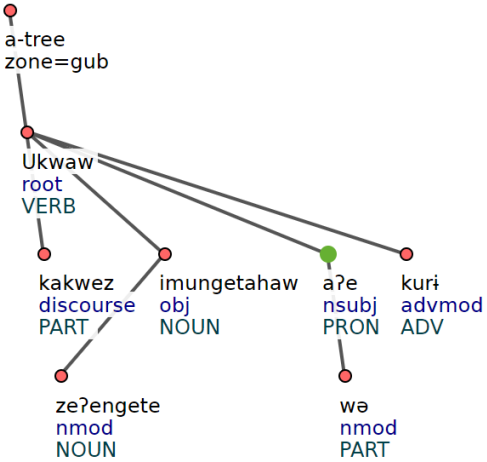
- Check grammar usage in the corpus
- Learner corpora

Historical Linguistics, Classical Languages

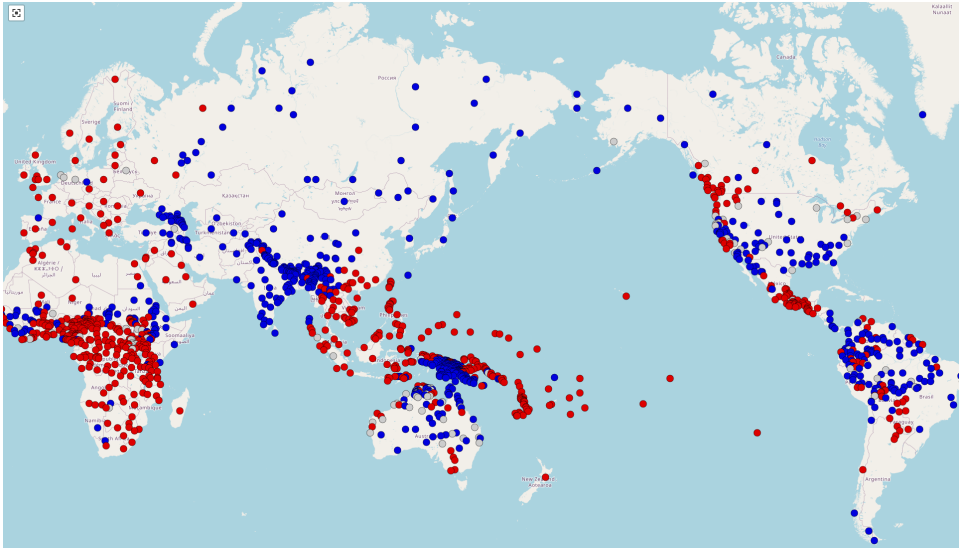
- Old Turkish
- Classical Chinese
- Sanskrit
- Akkadian
- Coptic
- Ancient Greek
- Latin
- Old French
- Gothic
- Old Church Slavonic
- Old East Slavic



Documentation of Endangered Languages



Linguistic Typology



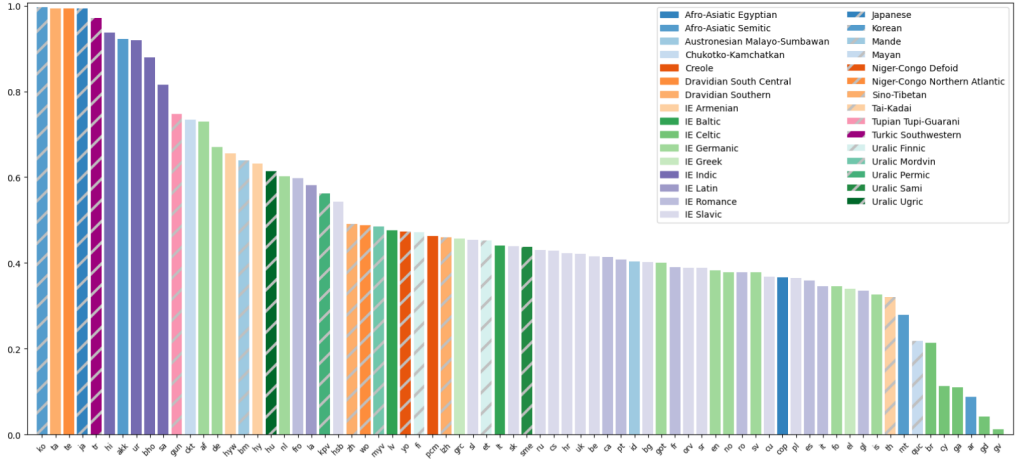


Figure 7 Percentage of head-final dependencies. Each bar is one language.

Dependency Grammar (DG)

- Dependency grammar is
  - a model developed by Lucien Tesnière (1893–1954)
  - based on structuralism
  - to describe the syntax of natural languages
- The main concern of DG is
  - dependency structure of sentence, i.e.
    - elements of the sentence
    - dependency relations between them
  - dependency = asymmetric binary relation between language units
  - governing / modified / parent unit (head)
  - dependent / modifying / child unit (dependent)
    - word (morph) grammar ... lexicalization
    - no phrase nodes
  - dependency tree: its edges are *mostly* dependency relations

Dependency-based Approaches

- Pāṇini (6^{th}–4^{th} century BC, India); Ibn Maḍāʾ (12^{th} century AD, Andalusia) … the term dependency
- Franz Kern and others († 1894) … sentence diagrams (pedagogy)
- Lucien Tesnière (1893–1954, France) … valency, "stemma" (unordered)

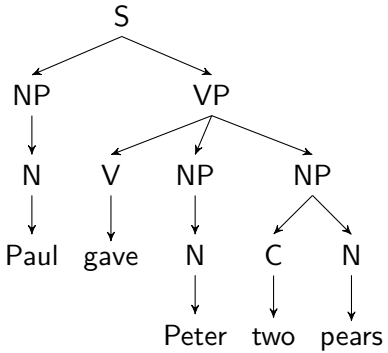
- motivation for current computational linguistics and NLP:
  - David Hays (1950–1960, machine translation ru → en)
  - Zellig Harris (since 1930; † 1992; linguistics is applied mathematics; methodology of linguistic analysis)
  - *Dependenzgrammatik* … esp. Jürgen Kunze (from 1960s; 1975)
  - *Valenzgrammatik* … esp. Gerhard Helbig (from 1960s)
  - Richard Hudson (from 1970s; 1984) … *Word Grammar*
  - Michael Halliday … *Systemic Functional Grammar*

Dependency-based Approaches

- Meaning-Text Theory (MTT) ... applied esp. in machine translation, lexicography; Igor Mel'čuk, Aleksandr Žolkovskij (since 1965)
- Functional Generative Description (FGD) ... applied in treebanks of the Prague family, used esp. for machine translation; Petr Sgall and his school (since 1967)
- Stanford Dependencies (SD) ... applied in treebanks of the Stanford family; Christopher Manning, Marie-Catherine de Marneffe and others (since 2006)
  - ⇒ Universal Dependencies (UD) ... Marie-Catherine de Marneffe, Christopher Manning, Joakim Nivre, Daniel Zeman and others (since 2014)

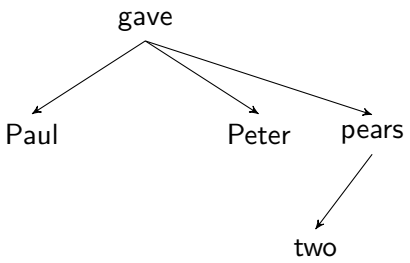
Constituent Tree

(S (NP (N Paul)) (VP (V gave) (NP (N Peter)) (NP (C two) (N pears))))



Dependency Tree

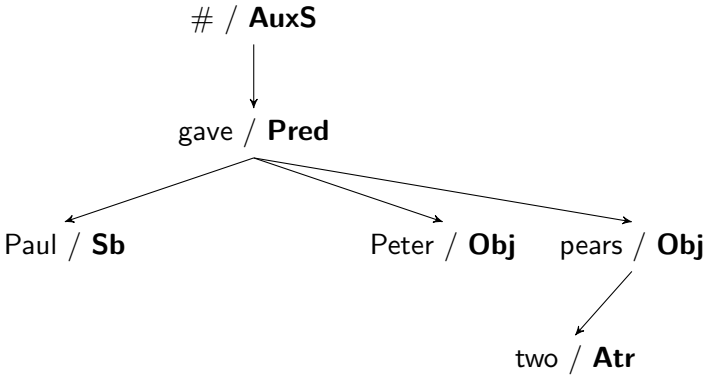
[gave,2] ([Paul,1], [Peter,3], [pears,5] ([two,4]))



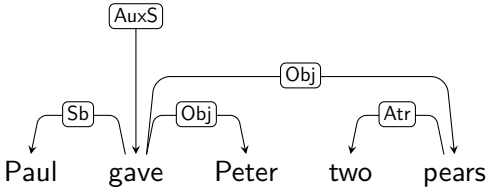
Paul gave Peter two pears.

Dependency Tree with Labels

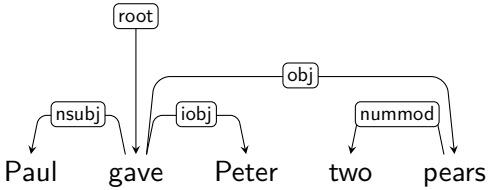
[#,0] ([gave,2] ([Paul,1], [Peter,3], [pears,5] ([two,4])), [.,6])



Dependency Tree with Labels



Dependency Tree with Labels



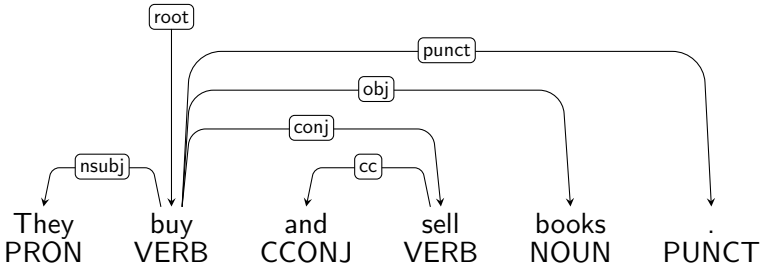
Properties of a Tree

- A directed graph (nodes + directed edges) is a tree if and only if
  - All nodes are connected
  - Every node has a unique parent (governor), except the root
- $\Rightarrow$ CoNLL-U file format

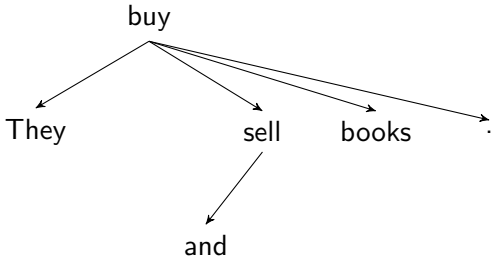
Syntactic Annotation

ID	FORM	LEMMA	UPOS	FEATS	HEAD	DEPREL
1	They	they	PRON	Case...	2	nsubj
2	buy	buy	VERB	Mood...	0	root
3	and	and	CCONJ	_	4	cc
4	sell	sell	VERB	Mood...	2	conj
5	books	book	NOUN	Number...	2	obj
6	.	.	PUNCT	_	2	punct

Dependency Tree



Dependency Tree



They buy and sell books.

Syntactic Annotation

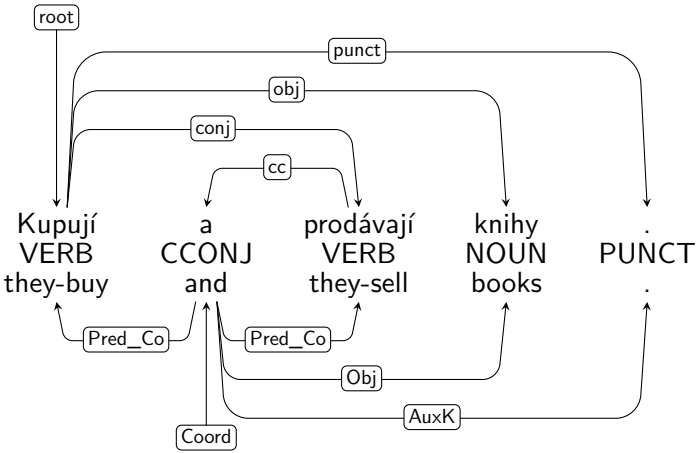
ID	FORM	LEMMA	UPOS	FEATS	HEAD	DEPREL
1	Kupují	kupovat	VERB	Mood...	0	root
2	a	a	CCONJ	_	3	cc
3	prodávají	prodávat	VERB	Mood...	1	conj
4	knihy	kniha	NOUN	Case...	1	obj
5	.	.	PUNCT	_	1	punct

Syntactic Annotation

ID	FORM	LEMMA	UPOS	FEATS	HEAD	DEPREL
1	Kupují	kupovat	VERB	Mood...	0	root
2	a	a	CCONJ	_	3	cc
3	prodávají	prodávat	VERB	Mood...	1	conj
4	knihy	kniha	NOUN	Case...	1	obj
5	.	.	PUNCT	_	1	punct

ID	FORM	LEMMA	XPOS	HEAD	DEPREL
1	Kupují	kupovat	VB-P---3P-AA---	2	Pred_Co
2	a	a	J^----------	0	Coord
3	prodávají	prodávat	VB-P---3P-AA---	2	Pred_Co
4	knihy	kniha	NNFP4-----A----	2	Obj
5	.	.	Z:----------	2	AuxK

Dependency Tree

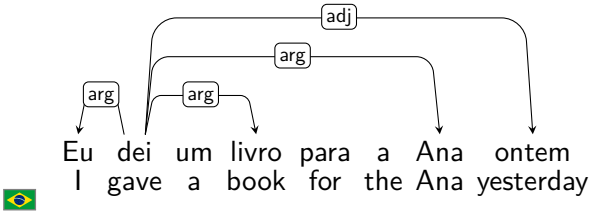


What Should be Attached to What?

- Different theories/treebanks $\Rightarrow$ different approaches
- Nevertheless, certain bits are widespread:

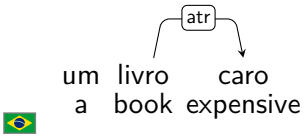
What Should be Attached to What?

- Different theories/treebanks ⇒ different approaches
- Nevertheless, certain bits are widespread:
  - Predicate (verb) governs its arguments (participants) and adjuncts (circumstances)



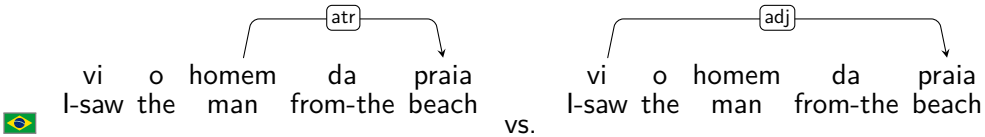
What Should be Attached to What?

- Different theories/treebanks ⇒ different approaches
- Nevertheless, certain bits are widespread:
  - Predicate (verb) governs its arguments (participants) and adjuncts (circumstances)
  - Attribute (adjective) is a child of modified noun



What Should be Attached to What?

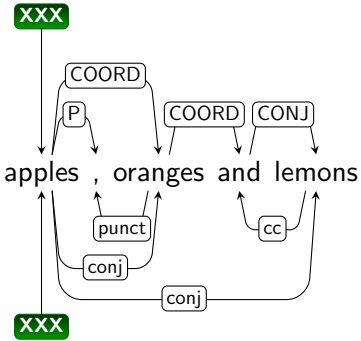
- Different theories/treebanks ⇒ different approaches
- Nevertheless, certain bits are widespread:
  - Predicate (verb) governs its arguments (participants) and adjuncts (circumstances)
  - Attribute (adjective) is a child of modified noun
  - Prepositional phrase attachment: What is the meaning?



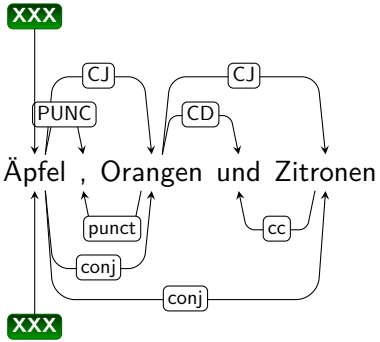
What Should be Attached to What?

- Different theories/treebanks $\Rightarrow$ different approaches
- Nevertheless, certain bits are widespread:
  - Predicate (verb) governs its arguments (participants) and adjuncts (circumstances)
  - Attribute (adjective) is a child of modified noun
  - Prepositional phrase attachment: What is the meaning?
- Common disagreements:
  - Function words (prepositions, conjunctions, auxiliaries, determiners)
  - Coordination
  - Ellipsis
  - Extra-grammatical phenomena (dates, addresses, lists etc.)

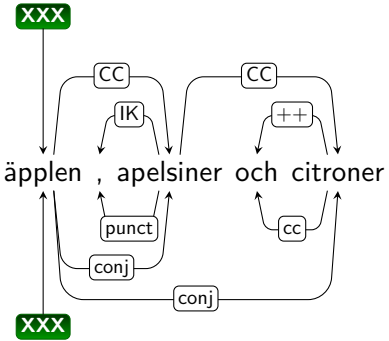
Coordination: Mel'čuk [en]



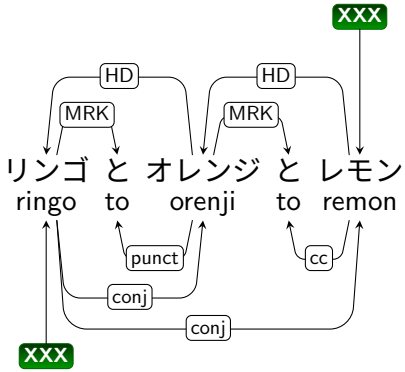
Coordination: Mel'čuk [de]



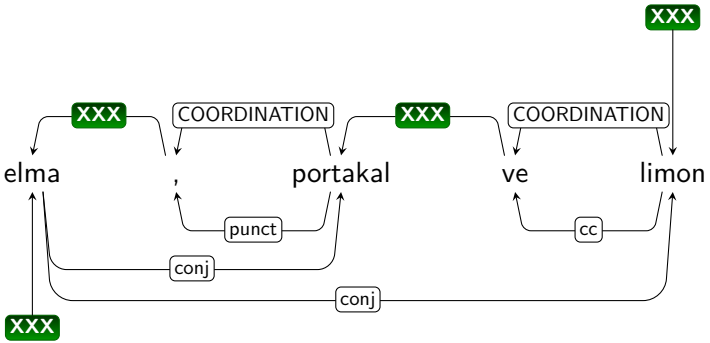
Coordination: Mel'čuk [sv]



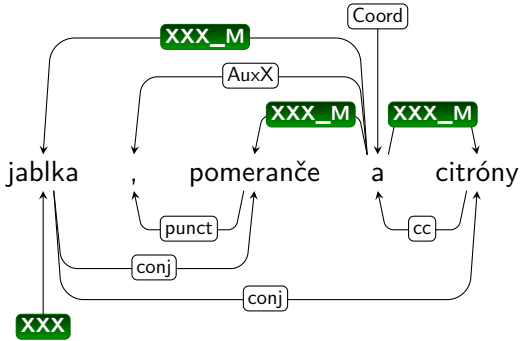
Coordination: Mel'čuk [ja]



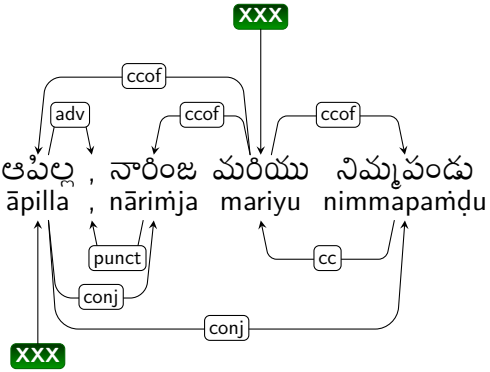
Coordination: Mel'čuk [tr]



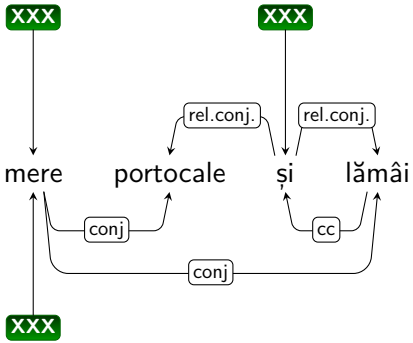
Coordination: Prague [cs]



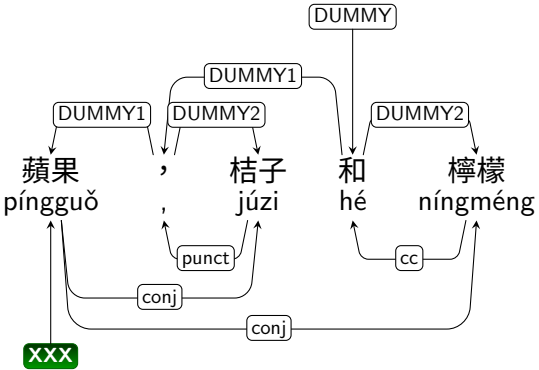
Coordination: Prague [te]



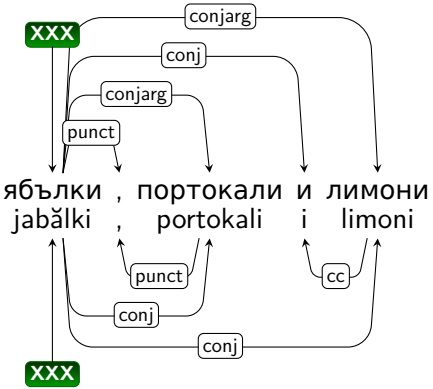
Coordination: Half-Prague [ro]



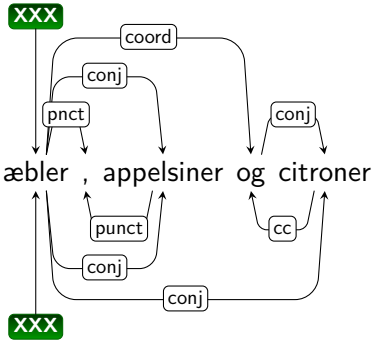
Coordination: Obligatorily Nested Prague [zh]



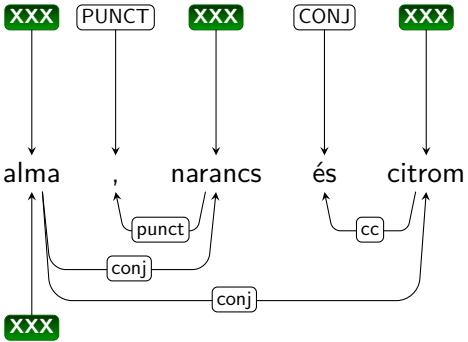
Coordination: Stanford [bg]



Coordination: Half-Stanford [da]

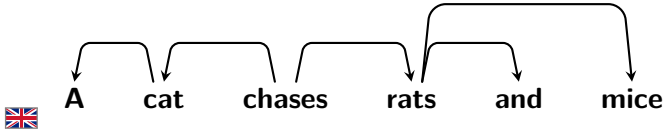


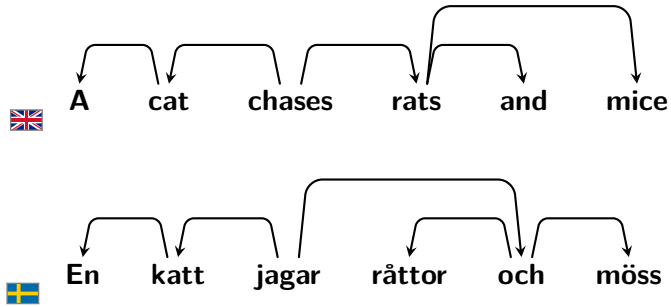
Coordination: Tesnière [hu]

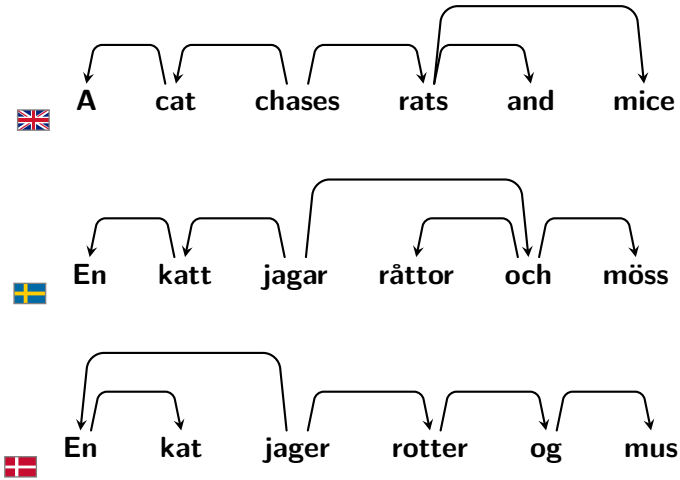


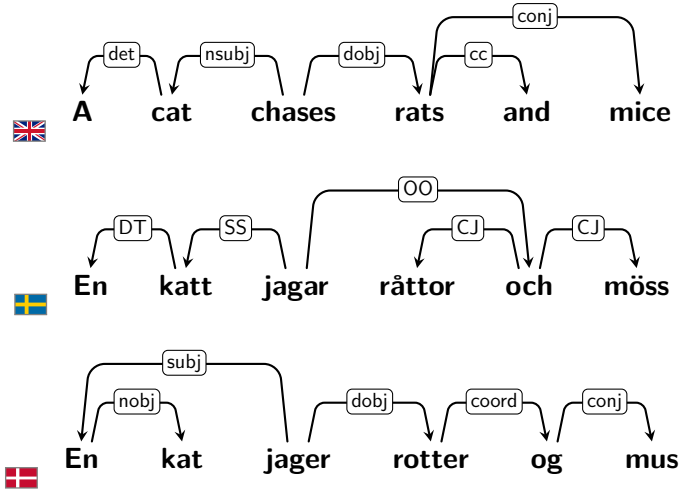
Digital Treebanks

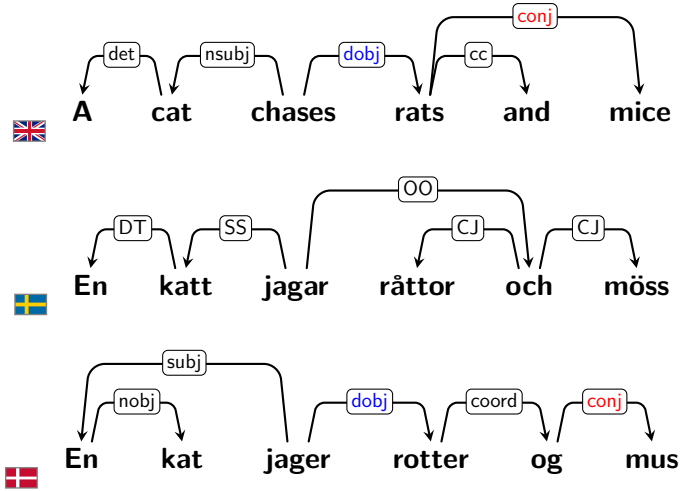
- 1970–1976, 2005–2006 Talbanken (Swedish)
- 1971–1985, 2007–2008 Czech Academic Corpus (Czech)
- 1989–1996 Penn TreeBank (English)
- 1995–2025 Prague Dependency Treebank (Czech)
- 1997–2004 Negra/Tiger Treebank (German)
- 2000–2007 Floresta Sintá(c)tica / Bosque (Portuguese)
- ~2005 Dependency parsing becomes dominant
- 2005 Stanford parser (2002) generates both constituents and dependencies
- 2006 CoNLL-X parsing shared task ⇒ CoNLL(-X) format, treebanks available for 10+ languages
- 2006–2011 Interset, universal part of speech tags
- 2012–2014 HamleDT (harmonized dependency treebanks, up to 30 languages)
- 2013–2014 Google universal dependency treebanks (11 languages)
- 2014–now Universal Dependencies (now almost 200 languages)


A
cat
chases
rats
and
mice


A cat chases rats and mice
En katt jagar råttor och möss


A cat chases rats and mice
En katt jagar råttor och möss
En kat jager rotter og mus


det
nsubj
dobj
cc
conj
A
cat
chases
rats
and
mice
DT
SS
OO
CJ
CJ
En
katt
jagar
råttor
och
möss
subj
nobj
dobj
coord
conj
En
kat
jager
rotter
og
mus

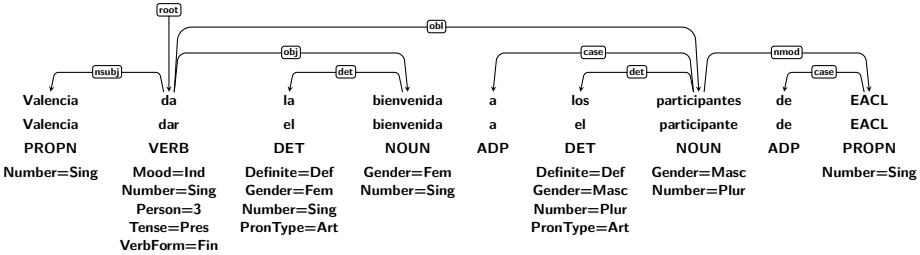

conj
det
nsubj
dobj
cc
A
cat
chases
rats
and
mice
OO
DT
SS
CJ
CJ
En
katt
jagar
råttor
och
möss
subj
nobj
dobj
coord
conj
En
kat
jager
rotter
og
mus

Why was this a problem?

- Hard to compare empirical results across languages
- Hard to usefully do cross-lingual structure transfer
- Hard to evaluate cross-lingual learning
- Hard to build and maintain multilingual systems
- Hard to make comparative linguistic studies
- Hard to validate linguistic typology
- Hard to make progress towards a universal parser

Universal Dependencies

http://universaldependencies.org



- Lemmas
- Part-of-speech tags
- Morphological features
- Syntactic dependencies

Design Principles

- Dependency
  - Widely used in practical NLP systems
  - Available in treebanks for many languages
- Lexicalism
  - Basic annotation units are words – syntactic words
  - Words have morphological properties
  - Words enter into syntactic relations
- Recoverability
  - Transparent mapping from input text to word segmentation

Universal Dependencies

- Same things annotated same way across languages...
- ... while highlighting different coding strategies

Manning's Law

The secret to understanding UD is to realize that the design is a very subtle compromise between approximately 6 things:

1. UD must be satisfactory on linguistic analysis grounds for individual languages.

It's easy to come up with a proposal that improves UD on one of these dimensions. The interesting and difficult part is to improve UD while remaining sensitive to all these dimensions.

Manning's Law

The secret to understanding UD is to realize that the design is a very subtle compromise between approximately 6 things:

1. UD must be satisfactory on linguistic analysis grounds for individual languages.
2. UD must be good for linguistic typology, i.e., providing a suitable basis for bringing out cross-linguistic parallelism across languages and language families.

It's easy to come up with a proposal that improves UD on one of these dimensions. The interesting and difficult part is to improve UD while remaining sensitive to all these dimensions.

Manning's Law

The secret to understanding UD is to realize that the design is a very subtle compromise between approximately 6 things:

1. UD must be satisfactory on linguistic analysis grounds for individual languages.
2. UD must be good for linguistic typology, i.e., providing a suitable basis for bringing out cross-linguistic parallelism across languages and language families.
3. UD must be suitable for rapid, consistent annotation by a human annotator.

It's easy to come up with a proposal that improves UD on one of these dimensions. The interesting and difficult part is to improve UD while remaining sensitive to all these dimensions.

Manning's Law

The secret to understanding UD is to realize that the design is a very subtle compromise between approximately 6 things:

1. UD must be satisfactory on linguistic analysis grounds for individual languages.
2. UD must be good for linguistic typology, i.e., providing a suitable basis for bringing out cross-linguistic parallelism across languages and language families.
3. UD must be suitable for rapid, consistent annotation by a human annotator.
4. UD must be easily comprehended and used by a non-linguist, whether a language learner or an engineer with prosaic needs for language processing. … it leads us to favor traditional grammar notions and terminology.

It's easy to come up with a proposal that improves UD on one of these dimensions. The interesting and difficult part is to improve UD while remaining sensitive to all these dimensions.

Manning's Law

The secret to understanding UD is to realize that the design is a very subtle compromise between approximately 6 things:

1. UD must be satisfactory on linguistic analysis grounds for individual languages.
2. UD must be good for linguistic typology, i.e., providing a suitable basis for bringing out cross-linguistic parallelism across languages and language families.
3. UD must be suitable for rapid, consistent annotation by a human annotator.
4. UD must be easily comprehended and used by a non-linguist, whether a language learner or an engineer with prosaic needs for language processing. … it leads us to favor traditional grammar notions and terminology.
5. UD must be suitable for computer parsing with high accuracy.

It's easy to come up with a proposal that improves UD on one of these dimensions. The interesting and difficult part is to improve UD while remaining sensitive to all these dimensions.

Manning's Law

The secret to understanding UD is to realize that the design is a very subtle compromise between approximately 6 things:

1. UD must be satisfactory on linguistic analysis grounds for individual languages.
2. UD must be good for linguistic typology, i.e., providing a suitable basis for bringing out cross-linguistic parallelism across languages and language families.
3. UD must be suitable for rapid, consistent annotation by a human annotator.
4. UD must be easily comprehended and used by a non-linguist, whether a language learner or an engineer with prosaic needs for language processing. … it leads us to favor traditional grammar notions and terminology.
5. UD must be suitable for computer parsing with high accuracy.
6. UD must support well downstream language understanding tasks (relation extraction, reading comprehension, machine translation, …)

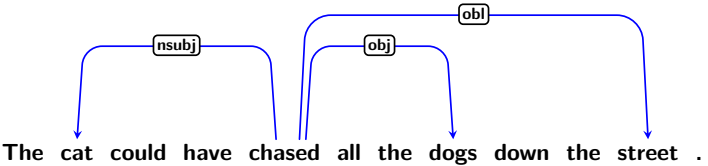
It's easy to come up with a proposal that improves UD on one of these dimensions. The interesting and difficult part is to improve UD while remaining sensitive to all these dimensions.

Morphological Annotation

Le	chat	chasse	les	chiens	.
le	chat	chasser	le	chien	.
the	cat	chases	the	dogs	.
DET	NOUN	VERB	DET	NOUN	PUNCT
Definite=Def Gender=Masc Number=Sing	Gender=Masc Number=Sing	Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm=Fin	Definite=Def Gender=Masc Number=Plur	Gender=Masc Number=Plur	

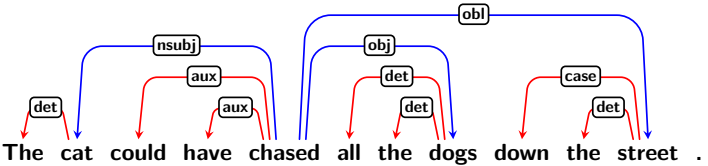
- Lemma representing the semantic content of a word
- **Part-of-speech tag** representing its grammatical class
- **Features** representing lexical and grammatical properties of the lemma or the particular word form

Syntactic Annotation



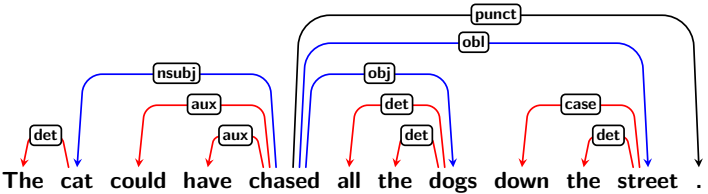
- Content words are related by dependency relations
- Function words attach to the content word they modify
- Punctuation attach to head of phrase or clause

Syntactic Annotation



- Content words are related by dependency relations
- Function words attach to the content word they modify
- Punctuation attach to head of phrase or clause

Syntactic Annotation



- Content words are related by dependency relations
- Function words attach to the content word they modify
- Punctuation attach to head of phrase or clause

CoNLL-U Format

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1	Es	es	PRON	_	_	2	nsubj	_	_
2	unterscheidet	unterscheiden	VERB	_	_	0	root	_	_
3	sich	sich	PRON	_	_	2	expl:pv	_	_
4-5	vom	_	_	_	_	_	_	_	_
4	von	von	ADP	_	_	7	case	_	_
5	dem	der	DET	_	_	7	det	_	_
6	westlichen	westlich	ADJ	_	_	7	amod	_	SpaceAfter=No
7	Teil	Teil	NOUN	_	_	2	obl	_	_
8	des	der	DET	_	_	9	det	_	_
9	Landes	Land	NOUN	_	_	7	nmod	_	SpaceAfter=No
10	.	.	PUNCT	_	_	2	punct	_	_

- Revised and extended version of CoNLL-X format
- Two-level segmentation and enhanced dependencies

Universal Dependencies

- 36 language families
- 193 languages
- 353 treebanks
- 791 contributors
- 2.3 million sentences
- 38 million words

(as of release 2.18)

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
Abkhaz	1	9K		Northwest Caucasian
Afrikaans	1	49K		IE, Germanic
Akkadian	2	25K		Afro-Asiatic, Semitic
Akuntsu	1	1K		Tupian, Tupari
Albanian	2	4K		IE, Albanian
Alemannic	1	1K		IE, Germanic
Amharic	1	10K		Afro-Asiatic, Semitic
Ancient Greek	3	456K		IE, Greek
Ancient Hebrew	1	90K		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
Azerbaijani	1	<1K		Turkic, Southwestern
Bambara	1	13K		Mande
Basque	1	121K		Basque
Bavarian	1	15K		IE, Germanic
Beja	1	11K		Afro-Asiatic, Cushitic
Belarusian	1	305K		IE, Slavic
Bengali	1	<1K		IE, Indic
Bhojpuri	1	6K		IE, Indic
Bokota	1	2K		Chibchan, Guaymiic
Bororo	1	160K		Bororoan
Breton	1	10K		IE, Celtic
Bulgarian	1	156K		IE, Slavic
Buryat	1	10K		Mongolic
Cantonese	1	13K		Sino-Tibetan, Chinese
Cappadocian	2	4K		IE, Greek
Catalan	1	553K		IE, Romance
Cebuano	1	1K		Austronesian, Central Philippine
Chinese	7	309K		Sino-Tibetan, Chinese
Chukchi	1	6K		Chukotko-Kamchatkan
Classical Armenian	1	99K		IE, Armenian
Classical Chinese	2	433K		Sino-Tibetan, Chinese
Coptic	2	89K		Afro-Asiatic, Egyptian
Croatian	1	199K		IE, Slavic
Czech	6	4,284K		IE, Slavic
Danish	1	100K		IE, Germanic
Dutch	2	506K		IE, Germanic
Egyptian	1	21K		Afro-Asiatic, Egyptian

Basic Universal Dependencies: 193 (189) Languages and Growing

▪ **I.-E.:** Armenian (+West +Class. +Mid.), Greek (+Ancient), Cappadocian, Phrygian, Albanian, Gheg, Hittite, Breton, Irish (+Old), Manx, Scottish, Welsh, Afrikaans, Alemannic, Danish, Dutch, English (+Old), Faroese, Frisian, German, Gothic, Icelandic, Low Saxon, Norwegian, Swedish, Yiddish, Catalan, French (+Old +Mid.), Galician, Italian, Latin, Ligurian, Neapolitan, Occitan (+Old), Portuguese, Romanian, Sicilian, Spanish, Umbrian, Belarusian, Bulgarian, Church Slavonic, Croatian, Czech, Macedonian, Polish, Pomak, Russian (+Old), Serbian, Slovak, Slovenian, Ukrainian, Upper Sorbian, Latvian, Lithuanian, Kurdish (North+Centr.+South), Zazaki, Persian, Khunsari, Nayini, Soi, Pashto, Urdu, Hindi, Punjabi, Kangri, Bhojpuri, Bengali, Assamese, Marathi, Odia, Sindhi, Sinhala, Sanskrit, Nepali ▪ **Dravidian:** Malayalam, Tamil, Telugu, Brahui ▪ **Uralic:** Erzya, Estonian, Finnish, Hungarian, Karelian, Livvi, Komi Permyak+Zyrian, Moksha, Nenets, Sámi North+Skolt, Veps ▪ **Turkic:** Kazakh, Kyrgyz, Old Turkish, Tatar, Turkish, Uyghur, Uzbek, Yakut ▪ Buryat ▪ Xibe ▪ Korean ▪ Japanese ▪ **Sino-T.:** Cantonese, Chinese (+Class.), Shanghainese, Chintang, Naga ▪ **Tai-K.:** Thai ▪ **Aus.-As.:** Vietnamese ▪ **Austron.:** Indonesian, Javanese, Gorontalo, Tagalog, Cebuano ▪ **Pama-Nyu.:** Warlpiri ▪ **Chu.-Kam.:** Chukchi ▪ **Esk.-Al.:** Yupik ▪ **Na.-De.:** Gwichin ▪ **U.-Az.:** Nahuatl

che ▪ **Chibchan:** Bokota, Ika, Pesh ▪ **Arawakan:**