Dependency parsing and tagging

Hakyung Sung
University of Hawaii at Manoa

October 7, 2026

Topics to cover

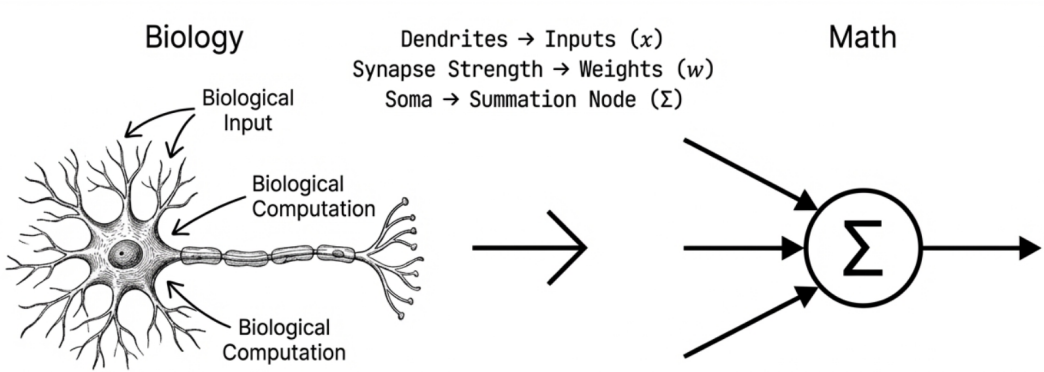
- Review
- Intro: Dependency parsing and tagging
- Basics of UD
- Some implications of UD

Review: Neural network

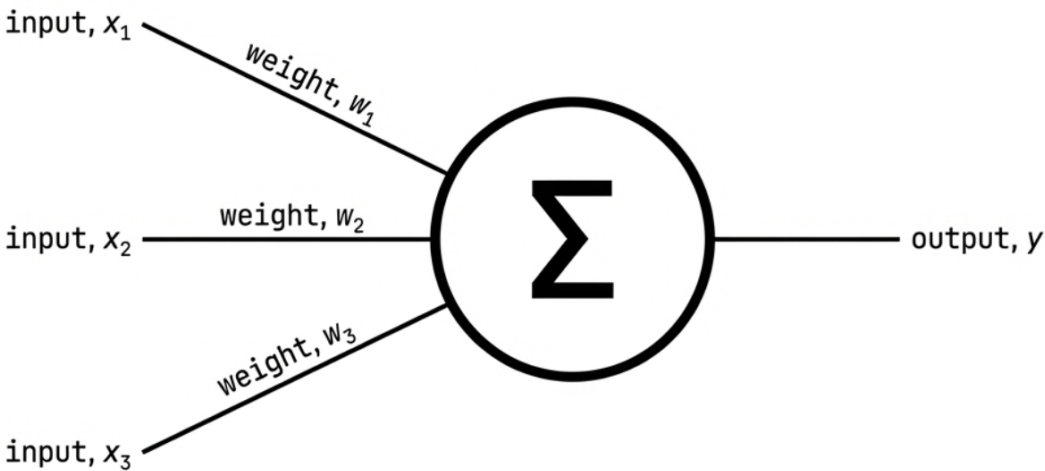
Review: Neural network

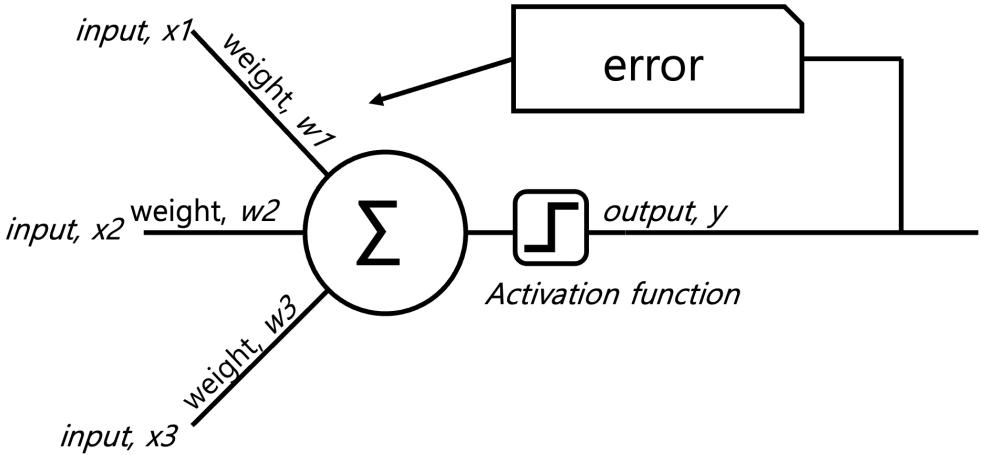
- Perceptron
- Multi-layer perceptron
- Backpropagation

Perceptron


Biology
Dendrites → Inputs (x)
Synapse Strength → Weights (w)
Soma → Summation Node (Σ)
Math
Biological Input
Biological Computation
Biological Computation
Σ

Perceptron

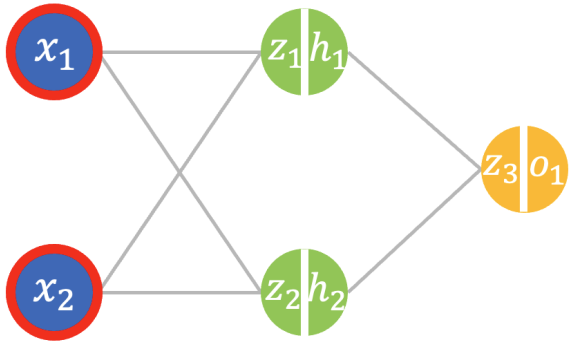

input, x_1
weight, w_1
input, x_2
weight, w_2
input, x_3
weight, w_3
Σ
output, y


input, x1
weight, w1
error
input, x2
weight, w2
Σ
output, y
Activation function
weight, w3
input, x3

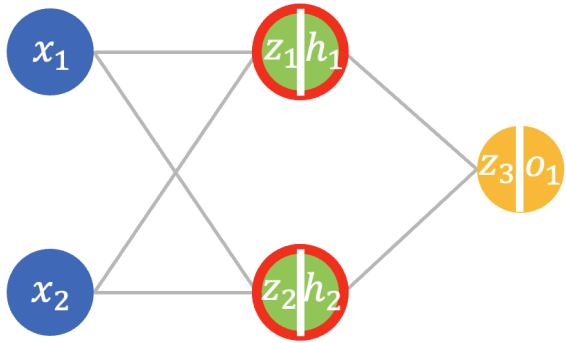
Review: Neural network

- Perceptron
- Multi-layer perceptron
- Backpropagation

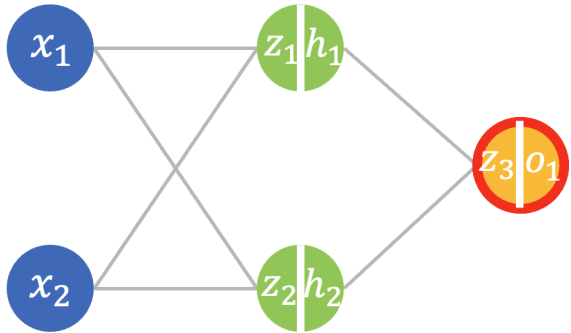
Multi-layer perceptron: Input Layer



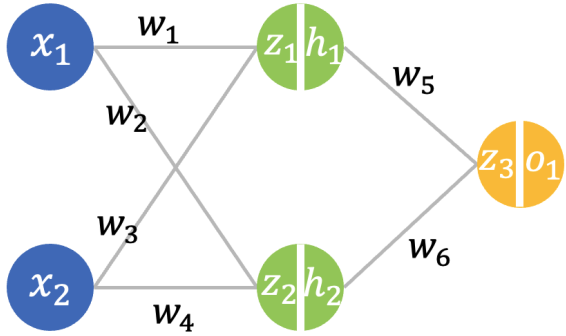
Multi-layer perceptron: Hidden Layer



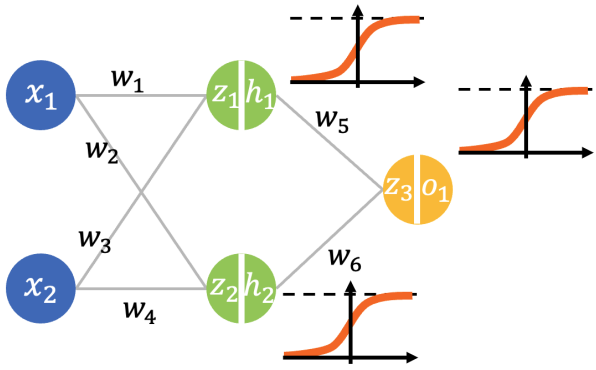
Multi-layer perceptron: Output Layer



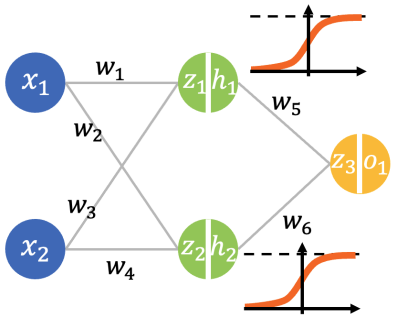
Multi-layer perceptron: Weights



Multi-layer perceptron: Activation Function



Multi-layer perceptron: Loss Function



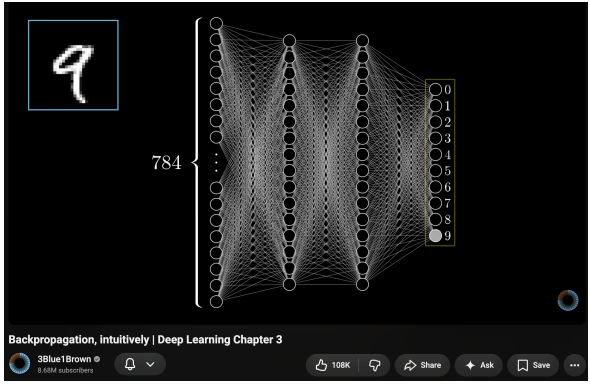
$$C = \frac{1}{n}\sum_{i}^{n}(y_i - \hat{y}_i)^2$$

Review: Neural network

- Perceptron
- Multi-layer perceptron
- Backpropagation

Learning Process

1. Feedforward: compute outputs
2. Loss calculation: evaluate error
3. **Backpropagation**: propagate errors backward
4. Update the weights, so errors can be reduced



Check here: `https://www.youtube.com/watch?v=Ilg3gGewQ5U`

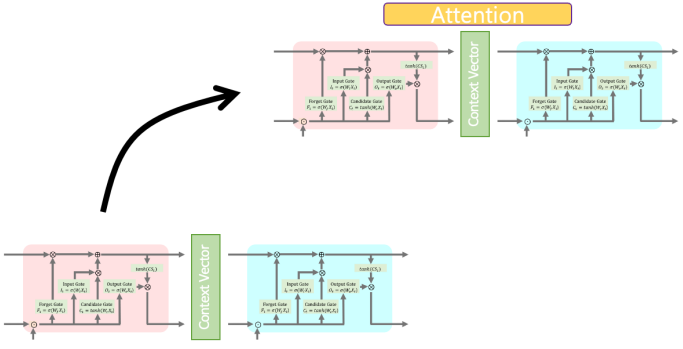


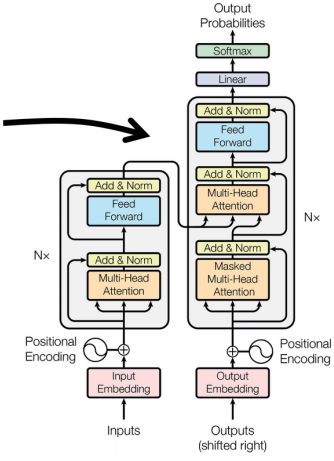
Language modeling

- Language modeling is the task of predicting the next word in a sequence.
- cf. language model, large language model
- Example:
  the students opened their ____ {*books*, *laptops*, *exams*, *minds*}
- Formally, given a sequence $x_1, x_2, \ldots, x_t$, we estimate

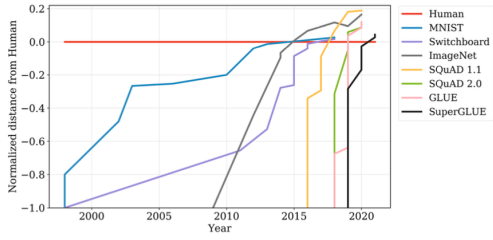
$$P(x_{t+1} \mid x_1, x_2, \ldots, x_t)$$

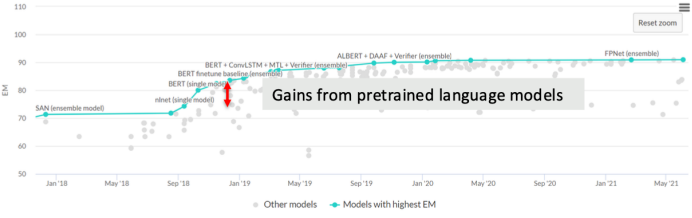
More progress





The pretraining revolution

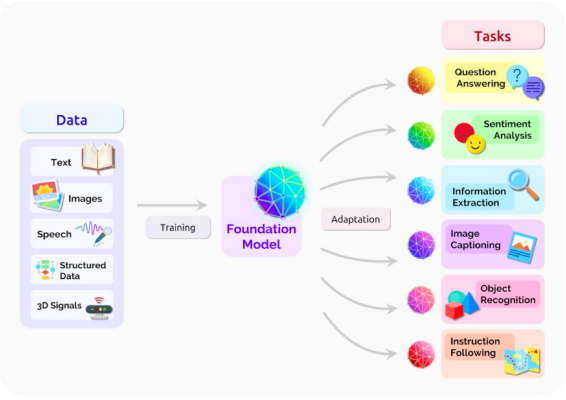




Pretraining has had a major impact on how well NLP systems work.

Images sourced from *here*

Pretraining: Scaling unsupervised learning on the internet



a LLM trained on a large amount of data that can serve as the starting point for many tasks (e.g., BERT [books, English Wiki]; GPT-2 [web corpora])

Sourced from *Bommasani et al., 2022*

Introduction



- corpus selection
- corpus compilation



- unit of analysis
- some variables (e.g., proficiency, word tokens)



Original image sourced from *here*



- Leveraging LLMs for different linguistic questions
  - Dependency parsing (this week)
  - Semantic role labeling (week 11)
  - Semantic/discourse analysis (week 13)

- Some related tasks/works
  - Measuring linguistic complexity (week 8)
  - Linguistic annotations (week 9)
  - Multimodal corpora (week 14)



- Leveraging LLMs different linguistic questions
  - Dependency parsing
  - Semantic role labeling
  - Semantic/discourse analysis

Structure over meaning

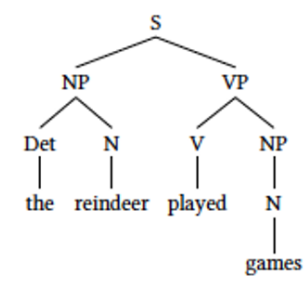
- Chomsky (1957): *"Colorless green ideas sleep furiously"*
- Nonsensical meaning, but:
  - Correct lexical and phrasal categories
  - Grammatically well-formed
- Syntax is about *structure*, not always meaning.

Understanding linguistic structure 1: Constituency

- **Constituency grammar**: a theory that analyzes sentences as nested constituents (e.g., noun phrases).

Lexicon:
N → *reindeer, dragon, lunch, game, evening, morning*
V(trans) → *play, eat*
V(intrans) → *run, swim, dance*
Adj → *fun, beautiful, interesting*
Det → *the, a, some, many*
P → *for, in, to, at*

Phrase structure rules:
S → NP VP
VP → V(trans) NP
VP → V(intrans)
NP → Det (A*) N
NP → N
NP → NP PP
PP → P NP



Understanding linguistic structure 2: Dependency

- **Dependency grammar** shows that syntactic structure consists of relationships between lexical items, normally binary asymmetric relations (“arrows”) called **dependencies**.
- Dependency structure shows which words depend on (modify, attach to, or are arguments of) which other words.

The founder of modern DG and VT is the Frenchman Lucien Tesnière (1893–1954) whose main work *Éléments de syntaxe structurale* was published posthumously (1959; 1976). The *Éléments* had essentially been completed in the 1940s. A short summary under the title of *Esquisse d'une syntaxe structurale* was published in 1953. There is a shortened translation of the 670-page *Éléments* into German (Tesnière 1980), but none into English. Tesnière's theory of structural syntax—only his followers named it "dependency grammar"—has the following basic features:

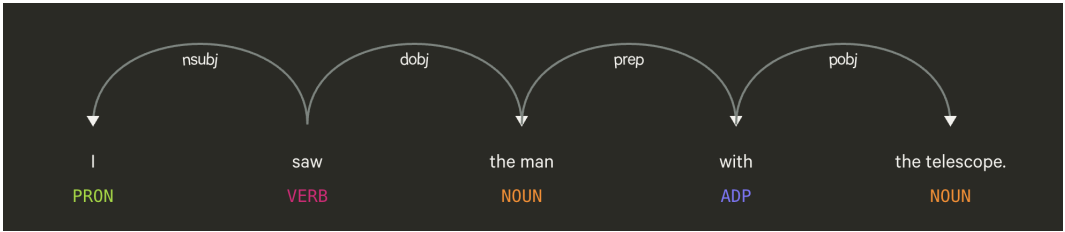
1. It combines a *typological* with a *universal* outlook.
2. It accounts for the *infinity* of theoretically producible and analyzable sentences.
3. It is *autonomous*.
4. It is *functional*.

Tesnière, who spoke more than 20 languages and used examples from more than 60 languages in the *Éléments*, placed great emphasis on applying his descriptive apparatus to different languages and linguistic types. He realized this typological-

Sourced from *Ágel & Fischer (2010)*

Why do we need dependency structure?

e.g., *I saw the man with the telescope.*



- Readers/Listeners/NLP models need to work out what modifies (attaches to) what.

Basics of UD

The rise of annotated data

What is a treebank? An annotated corpus that includes syntactic or morphological structure, often in the form of parse trees.

Milestones in treebank development:

- *Brown corpus* (1967): First general-purpose corpus; part-of-speech (PoS) tagged in 1979
- *Lancaster-IBM treebank* (Late 1980s): One of the first syntactically annotated corpora
- *The Penn treebank* (Marcus et al., 1993): Influential constituency-based treebank for English
- ***Universal Dependencies (UD)***: A multilingual, cross-linguistically consistent treebank project using dependency grammar

The rise of annotated data

Starting off, building a treebank seems a lot slower and less useful than writing a grammar (by hand)

But a treebank gives us many things:

- Reusability of the labor
  - Many parser, POS taggers, and built on it
  - Valuable resource for linguistics
- Broad coverage, not just a few samples
- General frequencies and distributional information
- A way to evaluate NLP systems (work as a benchmark for empirical science)

A shared framework for comparing languages

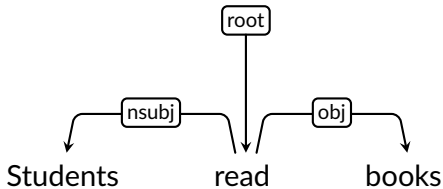
Universal Dependencies (UD) represents *morphosyntactic* structure using a shared annotation framework.

- Languages express comparable meanings through different word orders, morphology, and function words.

Dependency structures

- Dependency grammar describes relations between words.
- A **head** is structurally central to a relation.
- A **dependent** receives its grammatical role relative to that head.
- The **root** is the topmost word in the structure.
- Treats words as basic units ▸ See Appendix

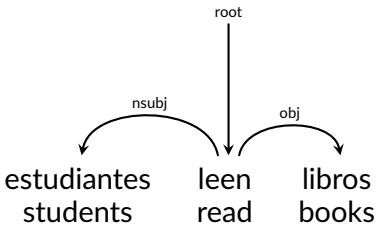
Subjects and objects depend on the predicate



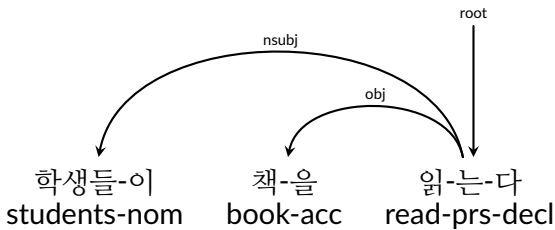
- nsubj: nominal subject
- obj: object

Comparable relations across word orders

Spanish



Korean



Universal part-of-speech categories (UPOS)

Traditional POS	UPOS tag	Category
noun	NOUN	common noun
	PROPN	proper noun
verb	VERB	main verb
	AUX	auxiliary verb or other tense, aspect, or mood particle
adjective	ADJ	adjective
	DET	determiner (including article)
	NUM	numeral (cardinal)
adverb	ADV	adverb
pronoun	PRON	pronoun
preposition	ADP	adposition (preposition/postposition)
conjunction	CCONJ	coordinating conjunction
	SCONJ	subordinating conjunction
interjection	INTJ	interjection
–	PART	particle (special single-word markers in some languages)
–	X	other (e.g., words in foreign-language expressions)
–	SYM	non-punctuation symbol (e.g., a hash (#) or emoji)
–	PUNCT	punctuation

Language-specific part-of-speech categories (XPOS)

e.g., English: Penn Treebank tags

1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle

Language-specific part-of-speech categories (XPOS)

e.g., Korean: Sejong tag set

Tag	Description	Tag	Description
NNG	Noun, common	JKB	Postposition, adverbial
NNP	Noun, proper	JKV	Postposition, vocative
NNB	Noun, common bound	JKQ	Postposition, quotative
NR	Numeral	JC	Postposition, conjunctive
NP	Pronoun	JX	Postposition, auxiliary
VV	Verb, main	EP	Ending, prefinal
VA	Adjective	EF	Ending, closing
VX	Verb, auxiliary	EC	Ending, connecting
VCP	Copula, positive	ETN	Ending, nominal
VCN	Copula, negative	ETM	Ending, determinative
MM	Determiner	XPN	Prefix, nominal
MAG	Common adverb	XSN	Suffix, noun derivative
MAJ	Conjunctive adverb	XSV	Suffix, verb derivative
IC	Exclamation	XSA	Suffix, adjective derivative
JKS	Postposition, nominative	XR	Root
JKC	Postposition, complement	NF	Undecided (consider a noun)
JKG	Postposition, prenominal	NV	Undecided (consider a verb)
JKO	Postposition, accusative	NA	Undecided

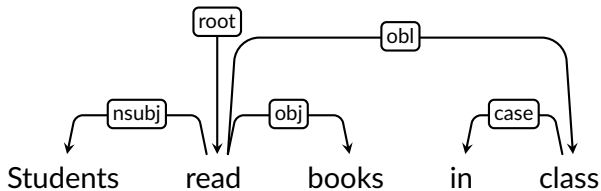
Grammatical relations

UD provides 37 basic relation labels connecting dependents to heads.

Relation	Function
nsubj	Nominal subject
obj	Primary non-subject nominal core argument
iobj	Indirect object or secondary nominal core argument
obl	Non-core nominal dependent of a predicate
case	Case-marking element, including an adposition

See lecture note's page 58 (or de Marneffe et al., 2021) for a complete list

Core arguments and oblique dependents



- *students* and *books* are core arguments of *read*.
- *class* is an oblique dependent expressing location.
- *in* depends on *class* through case.

Let’s play

`https://demos.explosion.ai/displacy`

Token IDs and heads in UD

She likes chocolate very much.

ID	Token	HEAD	Relation
0	She	1	`nsubj`
1	likes	0	`root`
2	chocolate	1	`obj`
3	very	4	`advmod`
4	much	1	`advmod`
5	.	4	`punct`

- *She*: HEAD = 2, so its head is *likes*.
- *very*: HEAD = 5, so its head is *much*.
- *likes*: HEAD = 0, so it is the root.

Evaluation

- Gold Standard: Hand-annotated syntactic structure used for evaluating parser output.
- Metrics:
  - UAS (Unlabeled Attachment Score): Correct head only
  - LAS (Labeled Attachment Score): Correct head and label

Example:

ID	Word	Gold Head	Gold Label	Pred Head	Pred Label
0	She			1	`nsubj`
1	likes			0	`root`
2	chocolate			1	`nmod`
3	very			4	`advmod`
4	much			4	`advmod`

Evaluation:

- **Total dependencies:** 5
- **Correct heads (UAS)**
- **Correct heads + labels (LAS)**

Some implications

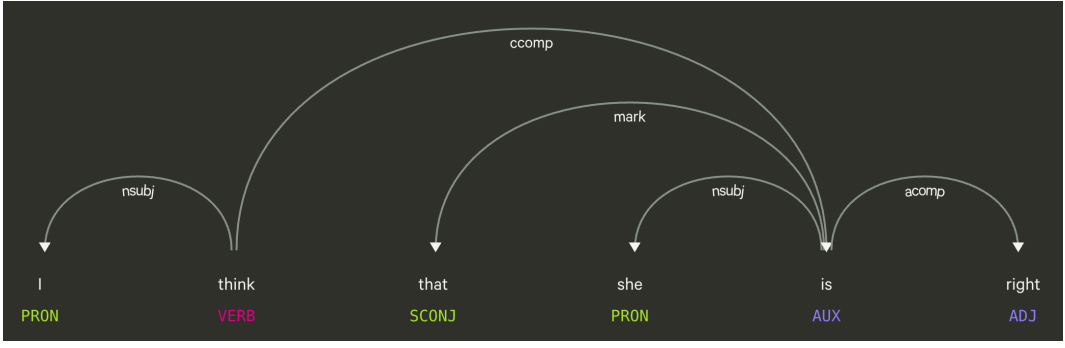
1. UD treebanks as corpus resources

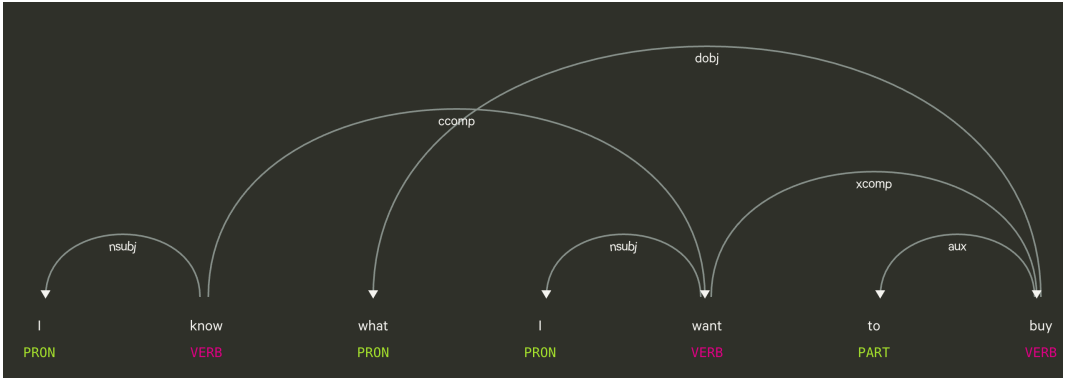
- Treebanks provide annotated POS information and dependency trees.
- The source text describes (1) gold annotation and/or (2) manually verified silver annotation.

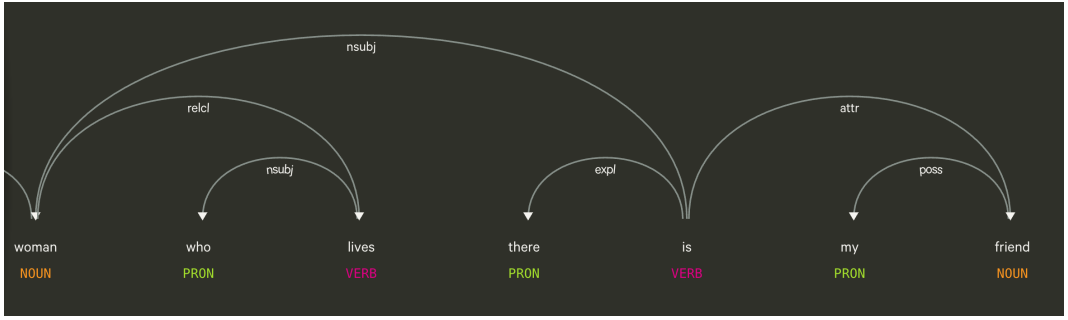
2. Research questions supported by dependencies

e.g., Relative clauses

- Retrieve relevant dependency structures.
- Option 1. Reliance on lexical cues such as *who*, *which*, or *that*.




dobj
ccomp
xcomp
nsubj
nsubj
aux
I
PRON
know
VERB
what
PRON
I
PRON
want
VERB
to
PART
buy
VERB


nsubj
relcl
nsubj
expl
attr
poss
woman
NOUN
who
PRON
lives
VERB
there
PRON
is
VERB
my
PRON
friend
NOUN

2. Research questions suppored by dependencies

verb argument construction	Example (register)
subject—verb—nominal complement	***It is*** *also an* ***indication*** *of the ways …* (academic)
subject—verb—adjective complement	***They are*** *very* ***discerning*** … (news)
subordinator—subject—verb—adjective complement	She hears ***that he is arrogant***. (news)
clausal complement—subject—verb	***[“Andy is an amalgamation of all the douchebags that I ’ve dealt with in my life”], Helms says***. (magazine)

Source: Kyle et al. (2021). Measuring longitudinal writing development using indices of syntactic complexity and sophistication. Table 7.

3. UD and learner corpora

UD annotation supports research on morphology and syntax in L2 production.

Target language (modality)	Reference
Chinese (written)	Lee et al. (2017)
English (written)	Berzak et al. (2016)
English (spoken)	Kyle et al. (2022)
Italian (written)	Di Nuovo et al. (2019)
Korean (written)	Sung and Shin (2024)
Russian (written)	Rozovskaya (2024)
Spanish (written)	Pulido et al. (2025)
Swedish (written)	Volodina et al. (2025)

Table 6: Examples of UD-annotated L2 learner corpora

Today's presentation: L2 English - Berzak et al. (2016)

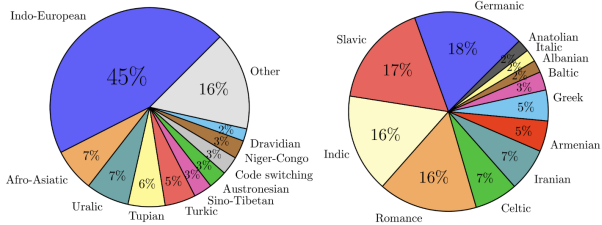
4. Comparability and representativeness

UD coverage is broad, but its collection does not constitute a balanced sample:

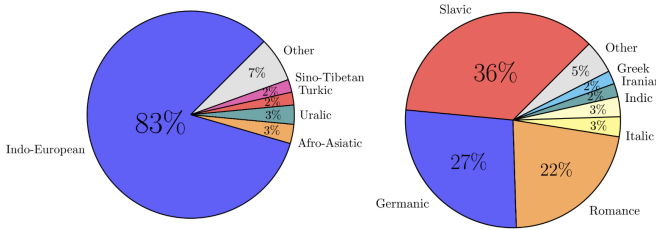
1. Grammatical preferences across languages
2. Different genres or registers
3. Corpus size or composition
4. Annotation practices

Diversity of the Latest Release

Languages per Family (Genus)



Words per Family (Genus)



See also separate pages with a list of UD languages and a list of scripts used in UD treebanks.

Choose a language that interests you and select one of its treebanks in the UD repository. Prepare a brief report covering the following (not all):

1. Language overview: Where is the language spoken? What is its basic word order?
2. Treebank overview: Give the treebank name, text types (e.g., news, conversation), size, and/or genre.
3. An interesting feature: Describe one dependency or word-order feature that differs from another language you know.
4. A question: Note something you found unclear or would like to investigate further.

Put this information into shared deck: `https://docs.google.com/presentation/d/13eh739isXzY_KW2NkA0Ozs6P6Al5kqWJZdoXDp-r0J4/edit?usp=sharing`

More resources

- `https://ufal.mff.cuni.cz/udpipe/2/models`
- `https://lindat.mff.cuni.cz/services/udpipe/`

Appendix

Words as the basic units

- Syntactic words carry morphological properties and enter dependency relations.
- cf. Internal morphemes normally receive morphological analysis rather than separate syntactic dependencies.

Lexical integrity principle: word-internal organization belongs primarily to morphology. Syntax relates words to other words.

Identifying syntactic words

- Orthographic tokens do not automatically correspond to syntactic words.
- Boundaries among function words, clitics, and affixes can be difficult to establish.
- Orthographic traditions may segment comparable material differently.

UD annotation requires explicit segmentation decisions, especially for languages with rich morphology.