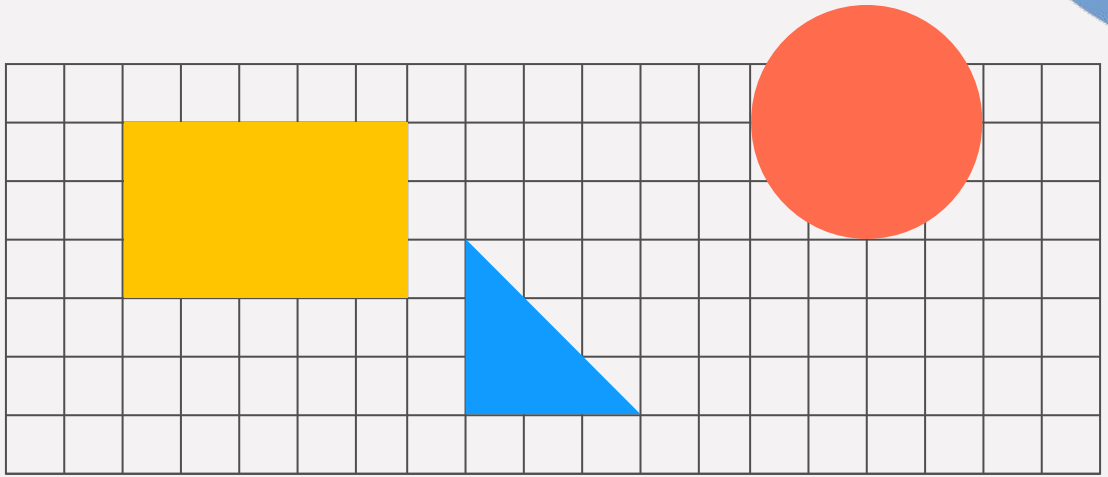
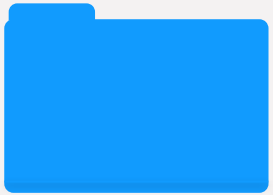
UD OVERVIEW

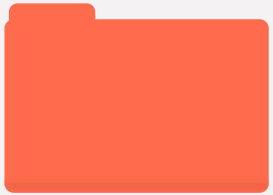




Definition



History

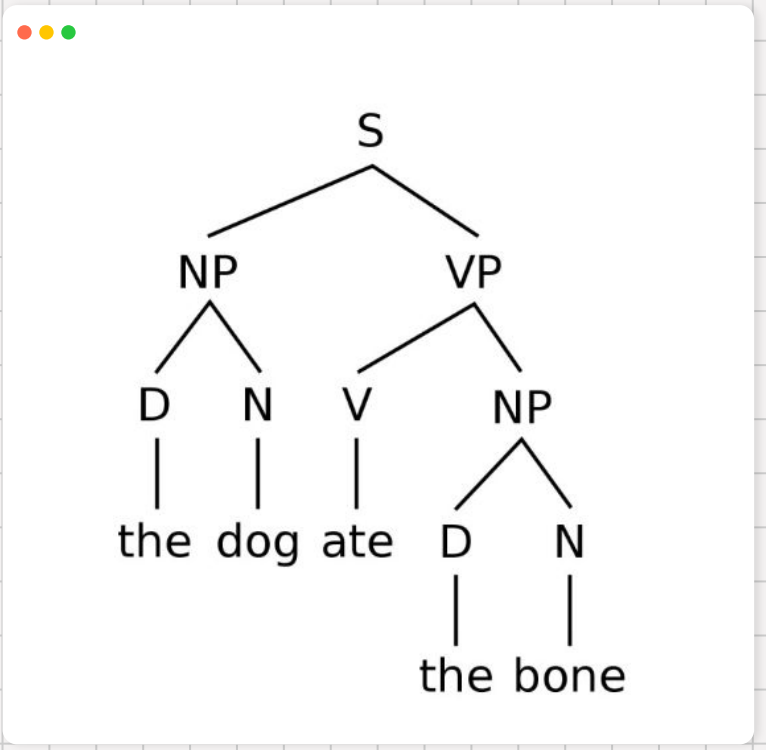


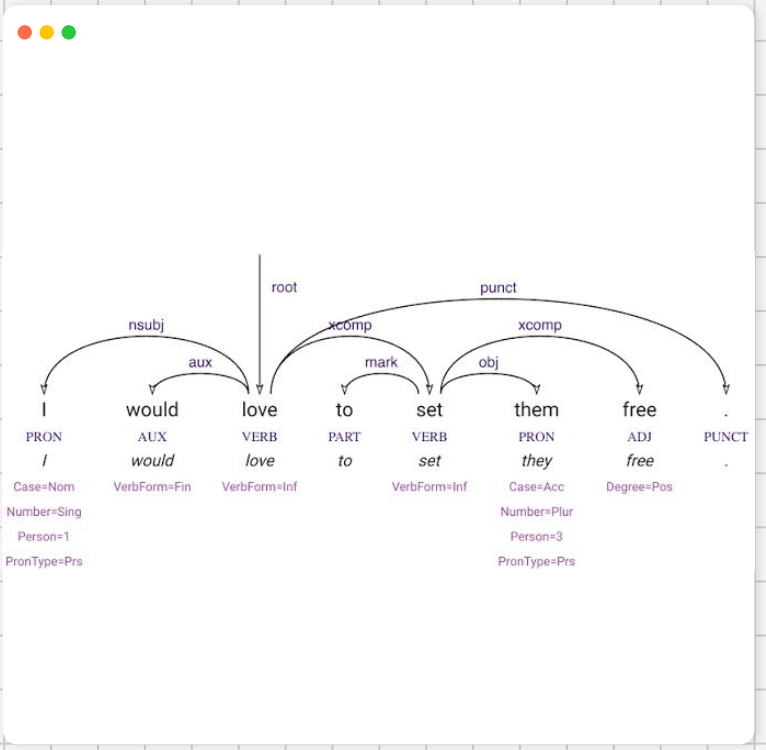
What is it?



Applications

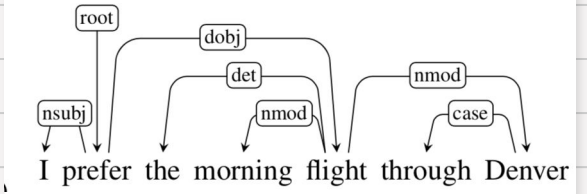
Dependency Grammar

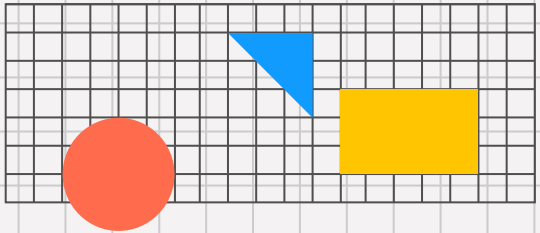




Dependency Grammar

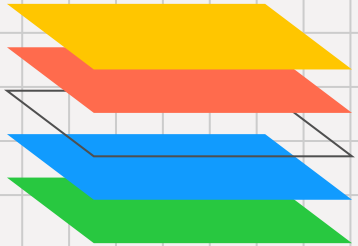
- Focuses on relationships between the words
- Centers the ‘root’
- Made of ‘heads’ and ‘dependents’
  - **‘Heads’ - provide basic meaning (lexical or grammatical)**
  - **‘Dependents’- modify or are arguments of the head**
- Handles free-order languages better
- https://universaldependencies.org/u/dep/





About UD

Words are basic units with morphology and syntactic relationships



- Made over time with contributions from many people
  - Marie-Catherine de Marneffe, Christopher Manning, Joakim Nivre, and Daniel Zeman
- Combination of **Stanford dependencies**, **Google's universal pos tags**, and the **Interset interlingua for morphosyntactic tagsets**
- Based on Dependency Grammar
- Including format for these trees: CoNLL-U
- https://universaldependencies.org

Definition

Universal Dependencies (UD) is a framework for consistent annotation of grammar (parts of speech, morphological features, and syntactic dependencies) across different human languages. UD is an open community effort with over 600 contributors producing over 200 treebanks in over 150 languages.

History of UD

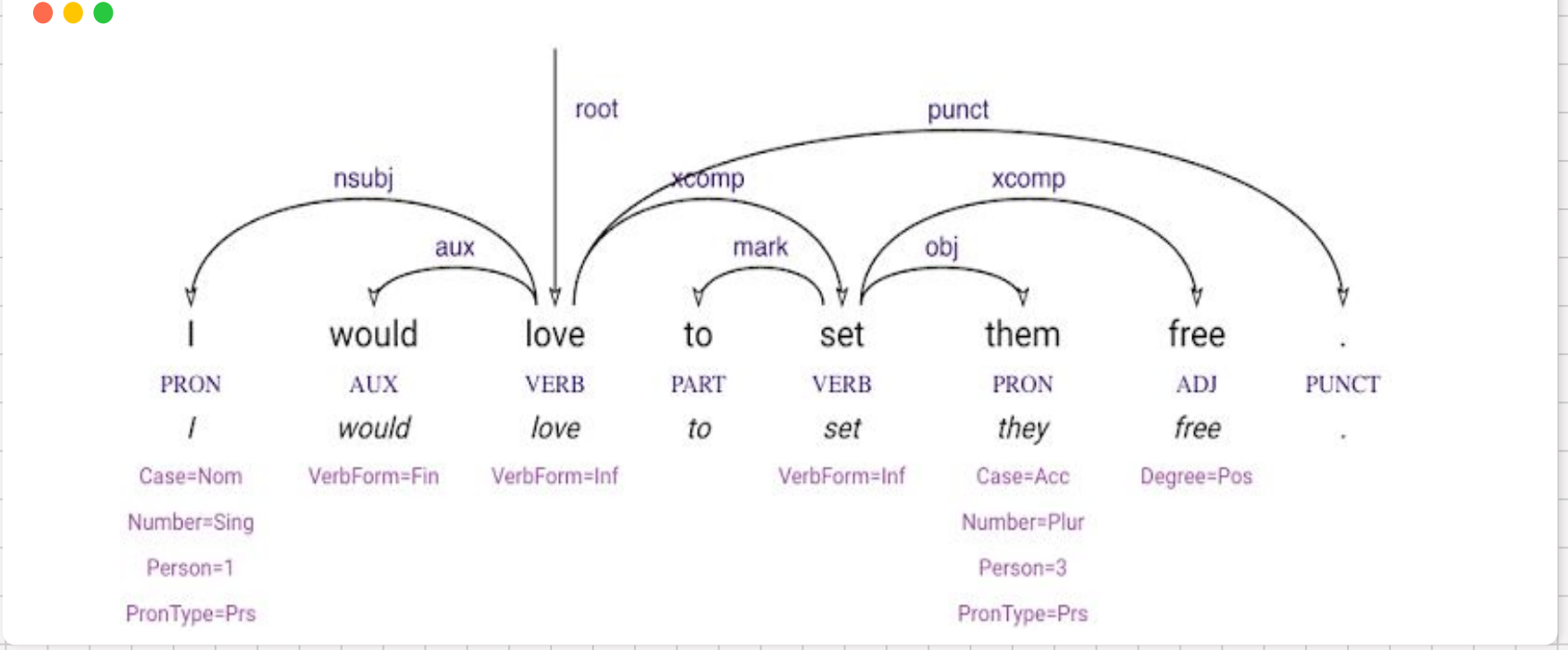
Why?

- Universal
- Many layers of annotation
- Easily parsable
- Constantly being updates
- Over 150 languages supported
- Allows for language-specific additional
- Better analysis of MWEs
- Free word order languages

DO
- Common categories for cross-linguistic analysis
- Keep same things same

DON'T
- Force languages into rigid boxes
- Annotate things that don't exist in the language

CoNLL-U


root
punct
nsubj
xcomp
xcomp
aux
mark
obj
I
would
love
to
set
them
free
.
PRON
AUX
VERB
PART
VERB
PRON
ADJ
PUNCT
I
would
love
to
set
they
free
.
Case=Nom
VerbForm=Fin
VerbForm=Inf
VerbForm=Inf
Case=Acc
Degree=Pos
Number=Sing
Number=Plur
Person=1
Person=3
PronType=Prs
PronType=Prs

CoNLL-U

```
(S
  (NP (NNP Isaac) (NNP Asimov))
  (VP
    (VBD was)
    (NP
      (NP (DT an) (JJ American) (NN writer) (CC and) (NN professor))
      (PP (IN of) (NP (NN biochemistry)))
      (PP (IN at) (NP (NNP Boston) (NNP University)))))
  (. .))
```

CoNLL-U

```
1   1   This    this    DET DT0 _   5   nsubj   _   _
2   2   is  be  AUX VBZ _   5   cop _   _
3   3   my  my  DET DPS _   5   det _   _
4   4   witness witness NOUN    NN1 _   5   compound    _   _
5   5   account account NOUN    NN1 _   0   root    _   _
6   6   of  of  ADP PRF _   8   case    _   _
7   7   the the DET AT0 _   8   det _   _
8   8   accident    accident    NOUN    NN1 _   5   nmod    _   _
9   9   ,   ,   PUNCT   PUN _   5   punct   _   _
10  10  regarding   regarding   ADP PRP _   13  case    _   _
11  11  file    file    NOUN    NN1 _   13  compound    _   _
12  12  number  number  NOUN    NN1 _   13  compound    _   _
13  13  F16 F16 PROPN   NP0 _   5   nmod    _   _
14  14  .   .   PUNCT   SENT    _   5   punct   _   _
```

Pros and Cons

Pros

- Many tools can be made to encode language specific data
  - **Does this take away from the U of UD?**
- Generally accessible for non-programmers
- Community based and open source

Cons

- How Universal can it really be?
  - **Cross linguistic comparisons vs losing language specific details**
- Not all languages are supported to the same extent
- Multi-Word Expressions
  - **The suspects remain *at large***

Applications

https://github.com/Data-Science-for-Linguists-2025/DEU-ENG-Mono-and-Billingual-Speakers

- Comes with more annotation schemes
  - **POS (Universal + Language-Specific)**
  - **Morphology**
  - **Lemma**
- POS analysis
  - **Syntactic entropy**
  - **Data Science for Linguistics projection → BPhil**
- Dependency analysis
  - **Comparing children + adult's usage of phrasal verbs**

- Multilayer annotations (CoNLL-U format)
- Many applications

Tips