Universal Dependencies for English

Nathan Schneider
July 31, 2017

https://static.pexels.com/photos/20974/pexels-photo.jpg
Why Dependencies?

- Dependency Grammar theories are based on the observation that many syntactic relationships can be characterized as asymmetric, binary relations between head and modifier words. (Tesnière 1959, Sgall et al. 1986, ...)
  - If you learned sentence diagramming in grade school (Reed & Kellogg 1877), that is a form of dependency grammar!
  - Not all constructions fit cleanly (coordination, relative clauses, ...); different theories have different solutions. Labeling the dependencies can clarify the nature of the relationship.

- While constituency grammars work well for “well-behaved” languages like English, Turkish and other languages introduce complications.

- Because dependency parses are structurally simpler, they are computationally easier to produce. (Faster parsers!)

- Syntactic dependencies are not too far from semantic dependencies, useful for many applications.
Universal Dependencies

• PTB is a *de facto* standard for constituency syntax, at least for English.

• But despite the popularity of dependencies, conventions/label sets abound.
  ‣ Different sets of head rules for converting from PTB trees
  ‣ Different edge labels for dependency treebanks

• **Universal Dependencies** (UD) are a recent (∼2014–2016) attempt to agree on cross-linguistic conventions.
  ‣ Evolved from Stanford Typed Dependencies → Universal Stanford Dependencies
  ‣ Headedness conventions and types designed for uniformity across languages
  ‣ Also conventions for annotating morphology & POS, not discussed here
  ‣ Guidelines and corpora from dozens of languages freely available at
    [http://universaldependencies.org/](http://universaldependencies.org/)
## UD Treebanks

<table>
<thead>
<tr>
<th>Language</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ancient Greek</td>
<td>182K</td>
</tr>
<tr>
<td>Ancient Greek–PROIEL</td>
<td>198K</td>
</tr>
<tr>
<td>Arabic</td>
<td>217K</td>
</tr>
<tr>
<td>Arabic–NYUAD</td>
<td>629K</td>
</tr>
<tr>
<td>Basque</td>
<td>97K</td>
</tr>
<tr>
<td>Belarusian</td>
<td>6K</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>140K</td>
</tr>
<tr>
<td>Catalan</td>
<td>472K</td>
</tr>
<tr>
<td>Chinese</td>
<td>111K</td>
</tr>
<tr>
<td>Coptic</td>
<td>3K</td>
</tr>
<tr>
<td>Croatian</td>
<td>183K</td>
</tr>
<tr>
<td>Czech</td>
<td>1,330K</td>
</tr>
<tr>
<td>Czech–CAC</td>
<td>482K</td>
</tr>
<tr>
<td>Czech–CLTT</td>
<td>26K</td>
</tr>
<tr>
<td>Danish</td>
<td>94K</td>
</tr>
<tr>
<td>Dutch</td>
<td>197K</td>
</tr>
<tr>
<td>Dutch–LassySmall</td>
<td>93K</td>
</tr>
<tr>
<td>English</td>
<td>229K</td>
</tr>
<tr>
<td>English–ESL</td>
<td>88K</td>
</tr>
<tr>
<td>English–LinES</td>
<td>67K</td>
</tr>
<tr>
<td>English–ParTUT</td>
<td>38K</td>
</tr>
<tr>
<td>Estonian</td>
<td>34K</td>
</tr>
<tr>
<td>Finnish</td>
<td>181K</td>
</tr>
<tr>
<td>Finnish–FTB</td>
<td>143K</td>
</tr>
<tr>
<td>French</td>
<td>381K</td>
</tr>
<tr>
<td>French–ParTUT</td>
<td>17K</td>
</tr>
<tr>
<td>French–Sequoia</td>
<td>58K</td>
</tr>
<tr>
<td>Galician</td>
<td>109K</td>
</tr>
<tr>
<td>Galician–TreeGal</td>
<td>14K</td>
</tr>
<tr>
<td>German</td>
<td>277K</td>
</tr>
<tr>
<td>Gothic</td>
<td>45K</td>
</tr>
<tr>
<td>Greek</td>
<td>51K</td>
</tr>
<tr>
<td>Hebrew</td>
<td>106K</td>
</tr>
<tr>
<td>Hindi</td>
<td>316K</td>
</tr>
<tr>
<td>Hungarian</td>
<td>37K</td>
</tr>
<tr>
<td>Indonesian</td>
<td>110K</td>
</tr>
<tr>
<td>Irish</td>
<td>13K</td>
</tr>
<tr>
<td>Italian</td>
<td>195K</td>
</tr>
<tr>
<td>Italian–ParTUT</td>
<td>39K</td>
</tr>
<tr>
<td>Japanese</td>
<td>173K</td>
</tr>
<tr>
<td>Japanese–KTC</td>
<td>189K</td>
</tr>
<tr>
<td>Kazakh</td>
<td>&lt;1K</td>
</tr>
<tr>
<td>Korean</td>
<td>63K</td>
</tr>
<tr>
<td>Korean–Sejong</td>
<td>89K</td>
</tr>
<tr>
<td>Latin</td>
<td>18K</td>
</tr>
<tr>
<td>Latin–ITTB</td>
<td>280K</td>
</tr>
<tr>
<td>Latin–PROIEL</td>
<td>159K</td>
</tr>
<tr>
<td>Latvian</td>
<td>44K</td>
</tr>
<tr>
<td>Lithuanian</td>
<td>40K</td>
</tr>
<tr>
<td>Norwegian–Bokmaal</td>
<td>280K</td>
</tr>
<tr>
<td>Norwegian–Nynorsk</td>
<td>276K</td>
</tr>
<tr>
<td>Old Church Slavonic</td>
<td>47K</td>
</tr>
<tr>
<td>Persian</td>
<td>135K</td>
</tr>
<tr>
<td>Polish</td>
<td>72K</td>
</tr>
<tr>
<td>Portuguese</td>
<td>201K</td>
</tr>
<tr>
<td>Portuguese–BR</td>
<td>268K</td>
</tr>
<tr>
<td>Romanian</td>
<td>202K</td>
</tr>
<tr>
<td>Russian</td>
<td>87K</td>
</tr>
<tr>
<td>Russian–SynTagRus</td>
<td>988K</td>
</tr>
<tr>
<td>Sanskrit</td>
<td>1K</td>
</tr>
<tr>
<td>Slovak</td>
<td>93K</td>
</tr>
<tr>
<td>Slovenian</td>
<td>126K</td>
</tr>
<tr>
<td>Slovenian–SST</td>
<td>19K</td>
</tr>
<tr>
<td>Spanish</td>
<td>411K</td>
</tr>
<tr>
<td>Spanish–AnCora</td>
<td>495K</td>
</tr>
</tbody>
</table>

## Upcoming UD Treebanks

<table>
<thead>
<tr>
<th>Language</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amharic</td>
<td>–</td>
</tr>
<tr>
<td>Buryat</td>
<td>–</td>
</tr>
<tr>
<td>Cantonese</td>
<td>–</td>
</tr>
<tr>
<td>Chinese–HK</td>
<td>–</td>
</tr>
<tr>
<td>Faroese</td>
<td>–</td>
</tr>
<tr>
<td>Kurmanji</td>
<td>–</td>
</tr>
<tr>
<td>Marathi</td>
<td>–</td>
</tr>
<tr>
<td>Serbian</td>
<td>–</td>
</tr>
<tr>
<td>Somali</td>
<td>–</td>
</tr>
<tr>
<td>Sorani</td>
<td>–</td>
</tr>
</tbody>
</table>

*as of March 2017*
Manning’s Law

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

1. UD needs to be satisfactory on linguistic analysis grounds for individual languages.
2. UD needs to 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 suitable for computer parsing with high accuracy.
5. UD must be easily comprehended and used by a non-linguist, whether a language learner or an engineer with prosaic needs for language processing. We refer to this as seeking a habitable design, and it leads us to favor traditional grammar notions and terminology.
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.
Cross-linguistic Parallelism

Examples from http://universaldependencies.org/introduction.html:

- English
  - The dog was chased by the cat.

- Bulgarian
  - Кучето се преследваше от котката.

- Czech
  - Pes byl honěn kočkou.

- Swedish
  - Hunden jagades av katten.
Content vs. Functional Heads

• Between two related **content** words, deciding which is the head (the direction of the arrow) is usually easy: e.g., *catch → fish* and *cute ← puppies*.

• **Function** words like auxiliaries, copulas, and adpositions are trickier.

• Some treebanks prefer **content heads** (UD adopts this policy):

```
Little kids were always watching birds with fish
```

• Others prefer **functional heads**:

```
Little kids were always watching birds with fish
```
UD Annotation for English: A Crash Course

Adapted from the v2 Universal guidelines at http://universaldependencies.org/ with additional examples from the main English UD treebank; refer to the website for many, many additional details
Can be drawn as an unlabeled edge coming from above the sentence, or coming from a dummy ROOT node.

Any punctuation token, attached to the head of its nearest containing phrase (often the head of the clause).
Subject, Object, Oblique

In most clauses, the predicate takes the form of a verb, which may be intransitive or transitive. Regardless of whether there is an overt copula linking the predicate to the subject or not.

Some languages allow extended transitive clauses, where more than two dependents are realized as core arguments. The additional core argument roles are typically nominals, while oblique modifiers are either (oblique) nominals or adverbial modifiers. (In nonverbal clauses, both core arguments then receive the same relation despite the presence of two core arguments, one of which is typically a demoted subject in relation to the non-causative form of the verb.)

By contrast, the alternative construction where the recipient role is realized by a prepositional phrase is typical of many languages.

Intransitive and transitive clauses can be used to indicate that an object is a demoted subject in relation to the non-causative form of the verb. These may be called A in linguistic typology and are associated with different types of grammatical information such as tense, mood, aspect, voice, evidentiality, or type of subordination. Analytic constructions (e.g., "made run him") can provide important evidence in specific languages, case alignment should not be used to decide the assignment of core arguments (S) of an intranstive verb.

In ergative languages, the patient-like argument of a transitive verb (O/P) will take the ergative case and the agent-like argument (i.e., the actor) will take the absolutive case. For example, the German double object construction qualifies as an ergative construction. The UD annotation assumes the clause as one of the basic structures that we expect to find in all languages. A simple clause minimally consists of a single subject (nsubj) and a single predicate (obj).

Note, finally, that not all languages allow extended transitives (and some do only in connection with special valency-changing particles). Hence, the semantic role cannot (by itself) be used to determine whether a dependent is core or not, nor can its status as an obligatory dependent. UD does not make a distinction between (obligatory) arguments and (optional) adjuncts, and oblique dependents are realizations of core arguments.

Finally, the predicate may be associated with function words that can be used to indicate that promotion has taken place. The subtype "left" is reserved for the argument most patient-like non-promoted. The subtype "left" is reserved for the argument most patient-like non-promoted.
Subject, Object, Oblique

<table>
<thead>
<tr>
<th>subjects</th>
<th>objects</th>
<th>obliques</th>
</tr>
</thead>
<tbody>
<tr>
<td>nsubj</td>
<td>obj</td>
<td>obl</td>
</tr>
<tr>
<td>nominal subject</td>
<td>direct object</td>
<td>case-marked noun</td>
</tr>
<tr>
<td>iobj</td>
<td></td>
<td></td>
</tr>
<tr>
<td>indirect object</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

```
she
left
him a note

she
left a note to him

she
left a note on the table
```
Subject, Object, Oblique

<table>
<thead>
<tr>
<th>subjects</th>
<th>objects</th>
<th>obliques</th>
</tr>
</thead>
<tbody>
<tr>
<td>nsubj</td>
<td>obj</td>
<td>obl</td>
</tr>
<tr>
<td>nsubj:pass</td>
<td>iobj</td>
<td>obl:agent</td>
</tr>
<tr>
<td></td>
<td>advmod</td>
<td>obl:tmol</td>
</tr>
<tr>
<td>csubj</td>
<td></td>
<td></td>
</tr>
<tr>
<td>csubj:pass</td>
<td></td>
<td></td>
</tr>
<tr>
<td>expl</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **nsubj**: nominal subject
- **nsubj:pass**: nominal subject of passive
- **csubj**: clausal subject
- **csubj:pass**: clausal subject of passive
- **expl**: expletive subject
- **obj**: direct object
- **iobj**: indirect object
- **obl**: case-marked noun
- **obl:agent**: passive by argument
- **obl:tmol**: temporal noun (adverbal or case-marked)

Examples:
- There is a ghost in the room.
- It is clear that we should decline.
- A nonverbal predicate (nominal or adjective) takes a single argument with the
- If passivization involves the promotion of an argument to subject position, then this argument can be annotated with a special subtype.
- The existential subject and indirect object slots, as in the examples below. Note that in the analysis of these examples, we treat the expletive subject rather than overt pronouns. In languages with expletives of this sort, they can be positioned where normally a core argument appears: the
- Some languages do not have expletives of the English sort, including most languages with free pro-drop (the ability to use zero anaphora.
- The argument with respect to case marking, agreement and/or linear position in the clause. All such dependents are said to be oblique,
Auxiliaries, Copulas, Case

Remember: these are function words, so they modify content words!

- aux: auxiliary
- cop: copula
- case: preposition or case clitic modifying a nominal

aux:pass: passive auxiliary (form of be or get)

For example, the sentence "she has been happy" has the structure:

- **she** is the subject
- **has** is the auxiliary verb
- **been** is the past participle
- **happy** is the adjective

For the sentence "she is my mother":

- **she** is the subject
- **is** is the copula
- **my mother** is the object

For the sentence "there is food in the kitchen":

- **there** is the existential
- **is** is the copula
- **food** is the object
- **in the kitchen** is the prepositional phrase
Auxiliaries, Copulas, Case

Remember: these are **function** words, so they modify content words!

```latex
\begin{align*}
\textbf{aux} & \quad \text{auxiliary} \\
\textbf{cop} & \quad \text{copula} \\
\textbf{case} & \quad \text{preposition or case clitic modifying a nominal}
\end{align*}
```

\textbf{aux:pass} passive auxiliary
(form of \textit{be} or \textit{get})

\begin{align*}
\text{she} & \quad \text{has} & \quad \text{been} & \quad \text{happy} \\
\text{she} & \quad \text{is} & \quad \text{my} & \quad \text{mother} \\
\text{she} & \quad \text{is} & \quad \text{in} & \quad \text{the} & \quad \text{kitchen} \\
\text{there} & \quad \text{is} & \quad \text{food} & \quad \text{in} & \quad \text{the} & \quad \text{kitchen}
\end{align*}
Adjectives, Determiners, Nominal modifiers

- **amod**: modifying adjective
- **nmod**: modifying nominal with case (except at the clause level)
- **appos**: appositive
- **nummod**: modifying numeral
- **nmod:poss**: non-adpositional possessive
- **nmod:tmod**: modifying temporal nominal in an NP

Examples:

- Sam eats red meat
- Sam spent forty dollars
- The office of the Chair
- The Chair's office
- Would dinner
- Thursday work instead
Adjectives, Determiners, Nominal modifiers

- **amod** modifying adjective
- **nmod** modifying nominal with case (except at the clause level)
- **appos** appositive
- **det** determiner
- **nummod** modifying numeral
- **nmod:poss** non-adpositional possessive
- **det:predet** predeterminer
- **nmod:tmod** modifying temporal nominal in an NP

Diagram:

- Sam eats red meat
- Sam spent forty dollars
- the office of the Chair
- the Chair's office
- AUX# Would
- NOUN# dinner
- PROPN# Thursday
- VERB# work
- ADV instead
- PUNCT
- Sam, the manager
The syntax of comparative constructions poses various challenges for linguistic theory. For English, many of these are discussed in Bresnan.

By contrast, compounds, flat names, and fixed expressions are used to analyze (endocentric) compounds like

\[ \text{I put in flour.} \]
\[ \text{I put in as much flour as the recipe called for.} \]

Get the cash to him as soon as possible.

Because many of these constructions are fixed grammatical units, they are annotated to show their modification structure, including a regular concept of head:

- **compound**: verb particle
- **compound:prt**: verb particle
- **compound:svc**: serial verb construction

With **fixed** and **flat**, the first word heads all other words in the expression.
Example
Example
Coordinating constructions

- **conj** non-initial conjunct
- **cc** coordinating conjunction (attaches to successive conjunct)
- **cc:preconj** preconjunction

Example:

- He came home, took a shower and immediately went to bed.

Example with annotations:

- Both Tina and Vicky are excellent astronauts.
Complement Clauses

**ccomp**: clausal complement

**mark**: subordinator, complementizer, or infinitive marker

**xcomp**: a predicate's clausal (or predicate A/N) complement that shares an argument with the matrix predicate

- He says you like to swim
- He says that you like to swim
- The boss said to start digging
- Please let us know
- Sue persuaded Fred to accept the job.
**Modifier Clauses**

**advcl** adverbial clause (e.g. expressing time, purpose, reason, condition...)

**acl** adjectival clause

**acl:relcl** relative clause

---

The accident happened as night was falling

He talked to him in order to secure the account

If you know who did it, you should tell the teacher

Sam spent everything he had

Sam spent everything that he had
Depictives, Resultatives, Secondary Predicates

**Depictive**, not a dependent of verb

Obligatory argument of verb which is understood as **predicating** one of the verb’s nominal arguments

**Resultative**, predicate indicating an outcome of the verbal event on one of its nominal arguments
Example

I really have n't thought about writing a book.
N.B. This is an example from the English treebank, but it is debatable whether advcl is correct.
Activity!

https://static.pexels.com/photos/20974/pexels-photo.jpg
**Discourse Stuff™ 1**

- **vocative**: addressee
- **dislocated**: topicalized noun phrase
- **discourse**: expression functioning as an interjection, filler, or similar conversational marker

**Examples:***

- **Guys**, take it easy!
- **Yes**, we should apply for membership!
- **Please**, join our growing family!!!
**Discourse Stuff™**  2

**parataxis** independent clauses/fragments forming a larger sentence, ideally separated with punctuation (but no conjunction); includes parentheticals, reported speech, tag questions

**list** items that do not form a syntactic sentence

---

Calafia has great fries (they are to die for!)

Steve

Jones Phone: 555-9814 Email: jones@abc.edf

The guy, John said

left early in the morning

John said: “The guy left early in the morning.”

Bearded dragons are sight hunters; they need to see the food to move.
Speech Errors and Overtokenization

**reparandum** superfluous word or phrase, such as a speech error

**goeswith** superfluous space between words (would normally be written as a single word). As with *fixed* and *flat*, the first word heads all other words in the expression.
Questions

• There are no special dependency types for questions (or, for that matter, imperatives, which simply lack an overt subject).

• For yes/no questions, try rephrasing as a confirmation question. The dependencies will be the same.
  ‣ Do you like my hat? ⟹ You do like my hat?
  ‣ Is this a hat? ⟹ This is a hat?

• For WH-questions, rephrase with an in situ WH-word.
  ‣ Why do you like my hat? ⟹ You do like my hat why?
  ‣ What did you eat? ⟹ You did eat what?
  ‣ Who do you think wants my hat? ⟹ You do think (that) who wants my hat?
Ellipsis

From http://universaldependencies.org/u/overview/specific-syntax.html#ellipsis:

The UD approach to ellipsis can be summarized as follows:

1. If the elided element has no overt dependents, we do nothing.

2. If the elided element has overt dependents, we promote one of these to take the role of the head.

3. If the elided element is a predicate and the promoted element a core argument, we use the orphan relation when attaching other non-functional dependents to the promoted head.

**orphan** dependent of elided material
Comparatives 1

- *more + Adj or comparative Adj*

- *more as a quantity*
Comparatives 2

• as-as

• enough
flat:foreign

- E.g., *ad hoc*
nmod:npmod and obl:npmod

• These subtypes are (IMO, unintuitively) applied to rates, compounds where only one of the words is a noun, and a few other postmodifier-of-a-noun constructions.

• Details: https://github.com/UniversalDependencies/docs/issues/478
The 37 “universal” relations (omits subtypes; **clf** – *classifier* not used for English)

<table>
<thead>
<tr>
<th></th>
<th>Nominals</th>
<th>Clauses</th>
<th>Modifier words</th>
<th>Function Words</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Core arguments</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nsubj</td>
<td></td>
<td>csubj</td>
<td></td>
<td>aux</td>
</tr>
<tr>
<td>obj</td>
<td></td>
<td>ccomp</td>
<td></td>
<td>cop</td>
</tr>
<tr>
<td>iobj</td>
<td></td>
<td>xcomp</td>
<td></td>
<td>mark</td>
</tr>
<tr>
<td><strong>Non-core dependents</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>obl</td>
<td>advcl</td>
<td>advmod*</td>
<td>discourse</td>
<td></td>
</tr>
<tr>
<td>vocative</td>
<td></td>
<td></td>
<td></td>
<td>aux</td>
</tr>
<tr>
<td>expl</td>
<td></td>
<td></td>
<td></td>
<td>cop</td>
</tr>
<tr>
<td>dislocated</td>
<td></td>
<td></td>
<td></td>
<td>mark</td>
</tr>
<tr>
<td><strong>Nominal dependents</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nmod</td>
<td>acl</td>
<td>amod</td>
<td>det</td>
<td></td>
</tr>
<tr>
<td>appos</td>
<td></td>
<td></td>
<td>clf</td>
<td></td>
</tr>
<tr>
<td>nummod</td>
<td></td>
<td></td>
<td>case</td>
<td></td>
</tr>
<tr>
<td><strong>Coordination</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>conj</td>
<td>fixed</td>
<td>list</td>
<td>orphan</td>
<td>punct</td>
</tr>
<tr>
<td>cc</td>
<td>flat</td>
<td>parataxis</td>
<td>goeswith</td>
<td>root</td>
</tr>
<tr>
<td></td>
<td>compound</td>
<td></td>
<td>reparandum</td>
<td>dep</td>
</tr>
</tbody>
</table>

* The **advmod** relation is used for modifiers not only of predicates but also of other modifier words.
UD Treebank Search

http://bionlp-www.utu.fi/dep_search
If you see problems in the online guidelines/data

https://github.com/UniversalDependencies/docs/docs/issues
• The development of parsing algorithms is a major topic of NLP research.
  ‣ Tradeoffs: accuracy, speed, complexity (constituency parse more complex than dependency parse)
  ‣ For Wall Street Journal news, state-of-the-art accuracies are in the low-to-mid 90% range!
  ‣ But HUGE variation in accuracy for other genres and languages

• Many parsers are open source. E.g. Stanford Parser, TurboParser, spaCy
  ‣ May require you to use a command line interface or a programming language

• Web demos that sometimes work: Stanford (http://corenlp.run/—currently UDv1), TurboParser (http://demo.ark.cs.cmu.edu/parse—Stanford Dependencies, not quite UD!)
Arborator

https://github.com/kimgerdes/arborator/
Thanks

Marie Catherine de Marneffe

Chris Manning

Sebastian Schuster

Amir Zeldes

Yi Zhu

Students in the Corpus Linguistics course at the 2017 Linguistic Institute, Lexington, KY