Universal Dependencies (vs. PDT): Syntactic Annotation

Markéta Lopatková

Institute of Formal and Applied Linguistics, MFF UK

lopatkova@ufal.mff.cuni.cz
Universal Dependencies (UDs)

idea:
- cross-linguistically consistent treebank annotation
- annotation for different languages as similar as possible
- support multilingual NLP (parser development, cross-lingual learning, …) and linguistic research

http://universaldependencies.org/
Universal Dependencies (UDs)

English:

Bulgarian:

Czech:

Swedish:
Syntactic annotation – main principles

• content words are related by dependency relations
Syntactic annotation – overview

• content words are related by dependency relations
• function words attach to the content word they modify
Syntactic annotation – overview

- content words are related by dependency relations
- function words attach to the content word they modify
- punctuation attach to head of phrase or clause
Main Characteristics of UDs

- **dependency annotation**
  - lexicalization

- **maximize parallelism** – but don’t overdo it:
  - don’t annotate the same thing in different ways
  - don’t make different things look the same
  - don’t annotate things that are not there

- universal taxonomy with possible language-specific features
  - languages select from a universal pool of categories
  - allow language-specific extensions

!!! NOT a new linguistic theory (but linguistically informed and relevant)
Syntactic annotation – main principles

• basic dependency representation: rooted tree
  • (surface) syntax
  • a notional ROOT
  • all other words are dependent on another word in the sentence
  • obligatory for all UD treebanks
Syntactic annotation – main principles

• basic dependency representation: rooted tree
  • (surface) syntax
  • a notional ROOT
  • all other words are dependent on another word in the sentence
  • obligatory for all UD treebanks

• enhanced dependency representation
  • in general not a tree
  • adds (change in a few cases) relations … for semantic interpretation
Syntactic annotation – overview

- content words as heads
  → maximizes parallelism between languages

**typed dependencies**

[Diagram showing syntactic structure with labels and relationships between words and tags.]

PDT: syntax in UDs

Lopatková
Syntactic annotation – overview

• content words as heads
  → maximizes parallelism between languages

  typed dependencies

• relations between function w. and content words
  → operations modifying the grammatical category of content w.

  functional relations or function word relations
  • function words (normally) have no dependents
Overview of UD Syntactic Relations

• taxonomy of 37 universal grammatical relations
  • broadly attested in language typology
  • language specific subtypes may be added
  • (incl. functional relations)

• organizing principles
  • 3 types of structures: **nominals, clauses, modifiers**
  • core arguments / oblique modifiers
  • coordination
  • multiword expressions
  • function words
  • other relations
## Nominals, Clauses, Modifiers

<table>
<thead>
<tr>
<th></th>
<th>nominal</th>
<th>clauses</th>
<th>modifiers</th>
</tr>
</thead>
<tbody>
<tr>
<td>dependents of clausal</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>predicates</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>core</td>
<td>nsubj</td>
<td>csubj</td>
<td></td>
</tr>
<tr>
<td></td>
<td>obj</td>
<td>ccomp</td>
<td></td>
</tr>
<tr>
<td></td>
<td>iobj</td>
<td>xcomp</td>
<td></td>
</tr>
<tr>
<td>non-core</td>
<td>obl</td>
<td>advcl</td>
<td>advmod</td>
</tr>
<tr>
<td></td>
<td>vocative</td>
<td>discourse</td>
<td></td>
</tr>
<tr>
<td></td>
<td>expl</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>dislocated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nominal dependents</td>
<td>nmod</td>
<td>acl</td>
<td>amod</td>
</tr>
<tr>
<td></td>
<td>appos</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>nummod</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Core Arguments / OblIQUE Modifiers

- **core arguments**
  - verbs usually only agree with core arguments
  - core args typically appear as bare nominals (= without pre/postposition)
  - certain cases (nominative, accusative, and absolutive) typically mark core args
  - core args in many languages occupy special positions in the clause
  - some syntactic phenomena are limited to core arguments in some languages.

  ~ more-or-less Obj in PDT

- **oblique modifiers**
  - oblique args may usually or always appear marked by an adposition

  ~ more-or-less Adv in PDT

**NOT** argument / adjunct distinction !!
Coordination

- head … the first conjunct
- all other conjuncts … depend on the head via the `conj` relation
- coordinating conjunction and punctuation … attached to the immediately following conjunct via the `cc` and `punct` relations
  (different from version 1 where attached to the first one)
Multiword Expressions

<table>
<thead>
<tr>
<th>mwe</th>
<th>fixed</th>
<th>flat</th>
<th>compound</th>
</tr>
</thead>
</table>

- **fixed** … for fixed grammaticalized function-word MWEs
e.g., *in spite of*

- **flat** … for exocentric semi-fixed MWEs
e.g., *Barack Obama* (with no clear head)

- **compound** … for (headed or endocentric) compounds
  i.e., with *clear inner structure*
## Functional Words

<table>
<thead>
<tr>
<th>dependents of clausal predicates</th>
<th>nominal dependents</th>
<th>clauses</th>
<th>modifiers</th>
<th>function words</th>
</tr>
</thead>
<tbody>
<tr>
<td>core</td>
<td>nsubj</td>
<td>csubj</td>
<td></td>
<td></td>
</tr>
<tr>
<td>obj</td>
<td>ccomp</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>iobj</td>
<td>xcomp</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>non-core</td>
<td>obl</td>
<td>advcl</td>
<td>advmod</td>
<td>aux</td>
</tr>
<tr>
<td>vocative</td>
<td></td>
<td>discourse</td>
<td>cop</td>
<td></td>
</tr>
<tr>
<td>expl</td>
<td></td>
<td></td>
<td>mark</td>
<td></td>
</tr>
<tr>
<td>dislocated</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nominal dependents</td>
<td>nmod</td>
<td>acl</td>
<td>amod</td>
<td>det</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>
</tbody>
</table>
Status of Function Words

- **multiple function** words related to the same content word → as siblings

- **copula** as a function word

  *but not for clauses as non-verbal predicates*
Status of Function Words

- **multiple function** words related to the same content word ➔ as siblings

- **copula** as a function word but not for clauses as non-verbal predicates

- 4 exceptions:
  - multiword function words
  - coordinated function words
  - function word modifiers
  - promotion by head elision
Status of Function Words

- **copula** as a function word

- **multiple function** words related to the same content word → as siblings

- 4 exceptions:
  - **multiword function words** ... relation fixed
  - coordinated function words
  - function word modifiers
  - promotion by head elision
Status of Function Words

- **copula** as a function word

- **multiple function** words related to the same content word → as siblings

- 4 exceptions:
  - multiword function words
  - **coordinated function words**
  - function word modifiers
  - promotion by head elision
Status of Function Words

- **copula** as a function word

- **multiple function** words related to the same content word as siblings

- 4 exceptions:
  - multiword function words
  - coordinated function words
  - **function word modifiers** ... e.g., *modified determiners*
  - promotion by head elision

PDT: syntax in UDs
Status of Function Words

• **copula** as a function word

• **multiple function** words related to the same content word as siblings

• 4 exceptions:
  - multiword function words
  - coordinated function words
  - function word modifiers
  - promotion by **head elision**

### Others

<table>
<thead>
<tr>
<th>Category</th>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>loose</td>
<td><strong>list</strong></td>
<td>tel. numbers, addresses, … (without syntactic structure)</td>
</tr>
<tr>
<td></td>
<td><strong>parataxis</strong></td>
<td>loosely linked clauses of same rank</td>
</tr>
<tr>
<td>special</td>
<td><strong>orphan</strong></td>
<td>orphans in ellipsis linked together</td>
</tr>
<tr>
<td></td>
<td><strong>goeswith</strong></td>
<td>for badly segmented words</td>
</tr>
<tr>
<td></td>
<td><strong>reparandum</strong></td>
<td>disfluency linked to (speech) repair</td>
</tr>
<tr>
<td>other</td>
<td><strong>punct</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>root</strong></td>
<td>syntactically independent element of clause/phrase</td>
</tr>
<tr>
<td></td>
<td><strong>dep</strong></td>
<td>unspecified dependency</td>
</tr>
</tbody>
</table>
Enhanced Dependencies

To capture more fined information:
• null nodes for elided predicates

The enhanced graph is not necessarily a supergraph of the basic tree.
Enhanced Dependencies

To capture more fined information:

- null nodes for elided predicates
- additional subject relations for control and raising constructions

The enhanced graph is not necessarily a supergraph of the basic tree.
Enhanced Dependencies

To capture more fine-grained information:
- null nodes for elided predicates
- additional subject relations for control and raising constructions
- propagation of conjuncts (~ effective children/parents in PDT)
- coreference in relative clause constructions
- modifier labels that contain the preposition or other case-marking information

*The enhanced graph is not necessarily a supergraph of the basic tree.*
"Language-specific" relations

- relations annotated for specific languages (not necessary specific to the language!!)
- subtypes of universal relations
  
  e.g., \texttt{nsubj:pass}
### CONLL-U Format

- Revised and extended version of CoNLL-X format

---

*The cat drinks milk.*

<table>
<thead>
<tr>
<th>ID</th>
<th>FORM</th>
<th>LEMMA</th>
<th>UPOSTAG</th>
<th>XPOSTAG</th>
<th>FEATS</th>
<th>HEAD</th>
<th>DEPREL</th>
<th>DEPS</th>
<th>MISC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Le</td>
<td>le</td>
<td>DET</td>
<td>_</td>
<td>_</td>
<td>2</td>
<td>det</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>2</td>
<td>chat</td>
<td>chat</td>
<td>NOUN</td>
<td>_</td>
<td>_</td>
<td>3</td>
<td>nsubj</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>3</td>
<td>boit</td>
<td>boire</td>
<td>VERB</td>
<td>_</td>
<td>_</td>
<td>0</td>
<td>root</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>4-5</td>
<td>du</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>4</td>
<td>de</td>
<td>de</td>
<td>ADP</td>
<td>_</td>
<td>_</td>
<td>6</td>
<td>case</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>5</td>
<td>le</td>
<td>le</td>
<td>DET</td>
<td>_</td>
<td>_</td>
<td>6</td>
<td>det</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>6</td>
<td>lait</td>
<td>lait</td>
<td>NOUN</td>
<td>_</td>
<td>_</td>
<td>3</td>
<td>obj</td>
<td>_</td>
<td>SpaceAfter=No</td>
</tr>
<tr>
<td>7</td>
<td>.</td>
<td>.</td>
<td>PUNCT</td>
<td>_</td>
<td>_</td>
<td>3</td>
<td>punct</td>
<td>_</td>
<td>_</td>
</tr>
</tbody>
</table>
### CONLL-U Format

- Revised and extended version of CoNLL-X format

<table>
<thead>
<tr>
<th>ID</th>
<th>FORM</th>
<th>LEMMA</th>
<th>UPPOSTAG</th>
<th>XPOSTAG</th>
<th>FEATS</th>
<th>HEAD</th>
<th>DEPREL</th>
<th>DEPS</th>
<th>MISC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Le</td>
<td>le</td>
<td>DET</td>
<td>_</td>
<td>_</td>
<td>2</td>
<td>det</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>2</td>
<td>chat</td>
<td>chat</td>
<td>NOUN</td>
<td>_</td>
<td>_</td>
<td>3</td>
<td>nsubj</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>3</td>
<td>boit</td>
<td>boire</td>
<td>VERB</td>
<td>_</td>
<td>_</td>
<td>0</td>
<td>root</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>4-5</td>
<td>du</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>4</td>
<td>de</td>
<td>de</td>
<td>ADP</td>
<td>_</td>
<td>_</td>
<td>6</td>
<td>case</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>5</td>
<td>le</td>
<td>le</td>
<td>DET</td>
<td>_</td>
<td>_</td>
<td>6</td>
<td>det</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>6</td>
<td>lait</td>
<td>lait</td>
<td>NOUN</td>
<td>_</td>
<td>_</td>
<td>3</td>
<td>obj</td>
<td>_</td>
<td>SpaceAfter=No</td>
</tr>
<tr>
<td>7</td>
<td>.</td>
<td>.</td>
<td>PUNCT</td>
<td>_</td>
<td>_</td>
<td>3</td>
<td>punct</td>
<td>_</td>
<td>_</td>
</tr>
</tbody>
</table>
CONLL-U Format

• Revised and extended version of CoNLL-X format
CONLL-U Format

- Revised and extended version of CoNLL-X format

List of morphological features (universal as well as lang. specific)

<table>
<thead>
<tr>
<th>ID</th>
<th>FORM</th>
<th>LEMMA</th>
<th>UPOSTAG</th>
<th>XPOSTAG</th>
<th>FEATS</th>
<th>HEAD</th>
<th>DEPREL</th>
<th>DEPS</th>
<th>MISC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Le</td>
<td>le</td>
<td>DET</td>
<td>_</td>
<td>_</td>
<td>2</td>
<td>det</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>2</td>
<td>chat</td>
<td>chat</td>
<td>NOUN</td>
<td>_</td>
<td>_</td>
<td>3</td>
<td>nsubj</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>3</td>
<td>boit</td>
<td>boire</td>
<td>VERB</td>
<td>_</td>
<td>_</td>
<td>0</td>
<td>root</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>4-5</td>
<td>du</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>4</td>
<td>de</td>
<td>de</td>
<td>ADP</td>
<td>_</td>
<td>_</td>
<td>6</td>
<td>case</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>5</td>
<td>le</td>
<td>le</td>
<td>DET</td>
<td>_</td>
<td>_</td>
<td>6</td>
<td>det</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>6</td>
<td>lait</td>
<td>lait</td>
<td>NOUN</td>
<td>_</td>
<td>_</td>
<td>3</td>
<td>obj</td>
<td>_</td>
<td>SpaceAfter=No</td>
</tr>
<tr>
<td>7</td>
<td>.</td>
<td>.</td>
<td>PUNCT</td>
<td>_</td>
<td>_</td>
<td>3</td>
<td>punct</td>
<td>_</td>
<td>_</td>
</tr>
</tbody>
</table>
CONLL-U Format

- Revised and extended version of CoNLL-X format

<table>
<thead>
<tr>
<th>ID</th>
<th>FORM</th>
<th>LEMMA</th>
<th>UPOSTAG</th>
<th>XPOSTAG</th>
<th>FEATS</th>
<th>HEAD</th>
<th>DEPREL</th>
<th>DEPS</th>
<th>MISC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Le</td>
<td>le</td>
<td>DET</td>
<td>_</td>
<td>_</td>
<td>2</td>
<td>det</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>2</td>
<td>chat</td>
<td>chat</td>
<td>NOUN</td>
<td>_</td>
<td>_</td>
<td>3</td>
<td>nsubj</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>3</td>
<td>boit</td>
<td>boire</td>
<td>VERB</td>
<td>_</td>
<td>_</td>
<td>0</td>
<td>root</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>4-5</td>
<td>du</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>4</td>
<td>de</td>
<td>de</td>
<td>ADP</td>
<td>_</td>
<td>_</td>
<td>6</td>
<td>case</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>5</td>
<td>le</td>
<td>le</td>
<td>DET</td>
<td>_</td>
<td>_</td>
<td>6</td>
<td>det</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>6</td>
<td>lait</td>
<td>lait</td>
<td>NOUN</td>
<td>_</td>
<td>_</td>
<td>3</td>
<td>obj</td>
<td>_</td>
<td>SpaceAfter=No</td>
</tr>
<tr>
<td>7</td>
<td>.</td>
<td>.</td>
<td>PUNCT</td>
<td>_</td>
<td>_</td>
<td>3</td>
<td>punct</td>
<td>_</td>
<td>_</td>
</tr>
</tbody>
</table>
CONLL-U Format

- Revised and extended version of CoNLL-X format
Universal Dependencies (UDs)

idea:
• cross-linguistically consistent treebank annotation
• annotation for different languages as similar as possible
• support multilingual NLP (parser development, cross-lingual learning, …) and linguistic research

people:
• J. Nivre, D. Zeman, F. Ginter, …………. (more than 160 contributors)
• open community effort – anyone can contribute!

main facts:
• from 2014
• version 2 in December 2016 (guidelines) and March 2017 (treebanks)
• next to 80 treebanks for 49 languages

http://universaldependencies.org/