Coordination Structures in Dependency Treebanks

Martin Popel, David Mareček, Jan Štěpánek, Daniel Zeman, Zdeněk Žabokrtský

Charles University in Prague, Faculty of Mathematics and Physics, ÚFAL (Institute of Formal and Applied Linguistics)

September 19th 2013, Příchovice
Motivation

- Coordination and Dependency are fundamentally different relations.
- Coordinations are difficult to represent in dependency treebanks.
- Large inter-treebank differences.
Motivation

- Coordination and Dependency are fundamentally different relations
- Coordinations are difficult to represent in dependency treebanks
- Large inter-treebank differences
- Obstacle for cross-lingual parsing (evaluation)

Swedish treebank → train → delexicalized parser → parse → Danish test set
Outline

- Styles of annotating coordinations
  - Topological styles
  - Labeling styles

- Transformation of styles

- Data: HamleDT (26 languages)
Participants of coordination

- **conjunct**
- **delimiter** (separates two conjuncts)
  - Coordinating conjunction
  - Comma or other punctuation (semicolon)
- **shared modifier** (modifies two or more conjuncts)

Examples:

- **lazy** dogs, cats and rats more than two conjuncts ("multi-conjunct c.")
- **Mary** came home and cried *home* is a “private modifier”
- **John** and **Mary** or **Peter** nested (embedded) coordinations
- **big** and **cheap** apples and **oranges** coordinated shared modifier
Special cases

- Asyndetic coordination = no conjunction

Don't worry, be happy, keep smiling
Special cases

- Asyndetic coordination = no conjunction
  - Don't worry, be happy, keep smiling
- Multi-word conjunction = as well as
Special cases

- Asyndetic coordination = no conjunction
  
  Don't worry, be happy, keep smiling

- Multi-word conjunction as well as

- Single-conjunct coordination And I love her
Special cases

- Asyndetic coordination = no conjunction
  - Don't worry, be happy, keep smiling
- Multi-word conjunction as well as
- Single-conjunct coordination And I love her
- One token with more roles etc.
  - Senatus Populusque Romanus que = coord. enclitic
  (The Senate and the People of Rome)
Special cases

- Asyndetic coordination = no conjunction
  
  *Don't worry*, *be happy*, *keep smiling*

- Multi-word conjunction *as well as*

- Single-conjunct coordination *And* *I love her*

- One token with more roles *etc.*

  *Senatus Populusque Romanus* *que* = coord. enclitic

  *(The Senate and the People of Rome)*

- Paratactic vs. hypotactic means *(John with Mary)*
Special cases

- Asyndetic coordination = no conjunction
  
  Don't worry, be happy, keep smiling

- Multi-word conjunction as well as

- Single-conjunct coordination And I love her

- One token with more roles etc.

  Senatus Populusque Romanus que = coord. enclitic

  (The Senate and the People of Rome)

- Paratactic vs. hypotactic means (John with Mary)

- red and white wine = red wine and white wine
  red and white flag of Poland
Topological styles (family)

Main “family” – configuration of conjuncts

Prague
- dogs
- cats
- rats

Moscow
- dogs
- cats

Stanford
- dogs
- cats
- and
- rats
Topological styles (head)

Choice of head (which delimiter/conjunct to choose):

rightmost

leftmost
Topological styles (head)

Choice of head (which delimiter/conjunct to choose):
rightmost

leftmost

Prague
Moscow
Stanford
Topological styles (head)

Choice of head: leftmost, rightmost or mixed
Topological styles (head)

Choice of head: leftmost, rightmost or mixed

Persian treebank: rightmost for coordination of verbs leftmost otherwise
Topological styles (shared modifiers)

Attachment of shared modifiers:

below the head

below the nearest conjunct
Topological styles (shared modifiers)

Attachment of shared modifiers:

below the head

Attachment of shared modifiers:

below the nearest conjunct
Topological styles (conjunction)

Attachment of coordinating conjunctions:

“between” conjunctions

below the previous conjunct

following conjunct

Stanford, head=rightmost
Topological styles (conjunction)

Attachment of coordinating conjunctions:

“between” conjuncts

below the previous conjunct

following conjunct

Moscow, head=leftmost
Topological styles (conjunction)

Attachment of coordinating conjunctions:

“between” conjuncts

“as the head”
for Prague (the only applicable)

below the previous conjunct

following conjunct

Moscow, head=leftmost
Topological styles (punctuation)

Attachment of punctuation delimiters:

“between” conjuncts

below the previous conjunct

following conjunct

Prague
Labeling styles (dependency rel.)

Dependency relation at “**upper level**” = with the head node

Dependency relation at “**lower level**” = with the conjuncts
Labeling styles (dependency rel.)

Dependency relation at “upper level” = with the head node

Dependency relation at “lower level” = with the conjuncts

Allows different labels of conjuncts.
Labeling styles (other)

- **Are conjuncts annotated?**
  - additional attribute (*is_member*) or
  - encoded into the dependency label: Sb\_M, Obj\_M, Atr\_M,...

- **Are shared modifiers annotated?**
  - In PDT not explicitly, but it can be deduced.

- Proposed, but unseen in treebanks: co-indexation attributes or bubbles for nested coordinations and shared modifiers
Annotation styles – overview

How many treebanks (out of 26 in HamleDT 1.0) use a given style?

- **Family** (Prague=14, Moscow=5, Stanford=6)
- **Head** (Leftmost=10, Rightmost=14, Mixed=1)
- **Shared modifiers** (below Head=11, Nearest conjunct=15)
- **Conjunctions** (Previous=2, Following=1, Between=8, as Head=14)
- **Punctuation** (Previous=7, Following=1, Between=15, Missing=2)
- **Dependency relation** (Upper=17, Lower=9)
- **Annotated conjuncts** (yes=21, no=5)
- **Annotated shared modifiers** (yes=8, no=18)
Annotation styles – overview

How many possible styles?
2*3*2*3*3+1*3*2*1*3 = 126 topological

* 8 labeling variants = 1008

How many styles really found?
16 (in 26 treebanks)
Transformations of styles

Subtasks

1. Detect coordinations in a sentence (esp. boundaries of nested coordinations)

2. Classify participants of coordinations (conjunct, commas, conjunctions, shared m.)

3. Transform each coordination to the target style (depth-first recursion, start with inner coord.)
Problematic cases

big and cheap
apples and oranges

and

apples
oranges

and

Prague

big
cheap

apples
oranges

Moscow

big
and
cheap
Problematic cases

"Save money, don't phone, use fax."

PDT 2.0
HamleDT v1.0 collection of treebanks

- HArmonized Multi-LanguagE Dependency Treebank
  http://ufal.mff.cuni.cz/hamledt/

- **Sources:** CoNLL, ICON, other
- We tried to harmonize also:
  prepositions, determiners, subordinated clauses, punctuation
- We plan to harmonize:
  verb groups, tokenization, …
- Recent “competitor”: Google Universal Treebanks
# HamleDT v1.0 statistics

<table>
<thead>
<tr>
<th>Language</th>
<th>Orig. type</th>
<th>Dataset</th>
<th>Sents.</th>
<th>Tokens</th>
<th>Original CS style code</th>
<th>CSs / 100 tok.</th>
<th>CJs / CS</th>
<th>SMs / CS</th>
<th>Nested CS [%]</th>
<th>RT UAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ancient Greek</td>
<td>dep</td>
<td>prim.</td>
<td>31 316</td>
<td>461 782</td>
<td>fP hR sH cH pB dL m11</td>
<td>6.54</td>
<td>2.17</td>
<td>0.16</td>
<td>10.3</td>
<td>97.86</td>
</tr>
<tr>
<td>Arabic</td>
<td>dep</td>
<td>C07</td>
<td>3 043</td>
<td>116 793</td>
<td>fP hL sH cH pB dL m00</td>
<td>3.76</td>
<td>2.42</td>
<td>0.13</td>
<td>10.6</td>
<td>96.69</td>
</tr>
<tr>
<td>Basque</td>
<td>dep</td>
<td>prim.</td>
<td>11 225</td>
<td>151 593</td>
<td>fP hR sN cH pP dU m00</td>
<td>3.37</td>
<td>2.09</td>
<td>0.03</td>
<td>5.1</td>
<td>99.32</td>
</tr>
<tr>
<td>Bengali</td>
<td>dep</td>
<td>I10</td>
<td>1 129</td>
<td>7 252</td>
<td>fP hR sH cH pP dU m11</td>
<td>4.87</td>
<td>1.71</td>
<td>0.05</td>
<td>24.1</td>
<td>99.97</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>phr</td>
<td>C06</td>
<td>13 221</td>
<td>196 151</td>
<td>fS hL sN cB pB dU m10</td>
<td>2.99</td>
<td>2.19</td>
<td>0.00</td>
<td>0.0</td>
<td>99.74</td>
</tr>
<tr>
<td>Czech</td>
<td>dep</td>
<td>C07</td>
<td>25 650</td>
<td>437 020</td>
<td>fP hR sH cH pB dL m11</td>
<td>4.09</td>
<td>2.16</td>
<td>0.20</td>
<td>14.6</td>
<td>99.42</td>
</tr>
<tr>
<td>Danish</td>
<td>dep</td>
<td>C06</td>
<td>5 512</td>
<td>100 238</td>
<td>fS* hL sN cP pB dU m10</td>
<td>3.68</td>
<td>1.93</td>
<td>0.13</td>
<td>7.5</td>
<td>99.76</td>
</tr>
<tr>
<td>Dutch</td>
<td>phr</td>
<td>C06</td>
<td>13 735</td>
<td>200 654</td>
<td>fP hR sH cH pB dL m10</td>
<td>2.06</td>
<td>2.17</td>
<td>0.05</td>
<td>3.3</td>
<td>99.47</td>
</tr>
<tr>
<td>English</td>
<td>phr</td>
<td>C07</td>
<td>40 613</td>
<td>991 535</td>
<td>fP hR sH cH pB dL m10</td>
<td>2.07</td>
<td>2.33</td>
<td>0.05</td>
<td>6.3</td>
<td>99.84</td>
</tr>
<tr>
<td>Finnish</td>
<td>dep</td>
<td>prim.</td>
<td>4 307</td>
<td>58 576</td>
<td>fS hL sN cB pB dU m10</td>
<td>4.06</td>
<td>2.41</td>
<td>0.00</td>
<td>6.4</td>
<td>99.70</td>
</tr>
<tr>
<td>German</td>
<td>phr</td>
<td>C09</td>
<td>38 020</td>
<td>680 710</td>
<td>fM hL sN cP pP dU m10</td>
<td>2.79</td>
<td>2.09</td>
<td>0.01</td>
<td>0.0</td>
<td>99.73</td>
</tr>
<tr>
<td>Greek</td>
<td>dep</td>
<td>C07</td>
<td>2 902</td>
<td>70 223</td>
<td>fP hR sH cH pB dL m11</td>
<td>3.25</td>
<td>2.48</td>
<td>0.18</td>
<td>7.2</td>
<td>99.43</td>
</tr>
<tr>
<td>Hindi</td>
<td>dep</td>
<td>I10</td>
<td>3 515</td>
<td>77 068</td>
<td>fP hR sH cH pP dU m11</td>
<td>2.45</td>
<td>1.97</td>
<td>0.04</td>
<td>10.3</td>
<td>98.35</td>
</tr>
<tr>
<td>Hungarian</td>
<td>phr</td>
<td>C07</td>
<td>6 424</td>
<td>139 143</td>
<td>fT hX sN cX pX dL m00</td>
<td>2.37</td>
<td>1.90</td>
<td>0.01</td>
<td>2.2</td>
<td>99.84</td>
</tr>
<tr>
<td>Italian</td>
<td>dep</td>
<td>C07</td>
<td>3 359</td>
<td>76 295</td>
<td>fS hL sN cB pB dU m10</td>
<td>3.32</td>
<td>2.02</td>
<td>0.03</td>
<td>3.8</td>
<td>99.51</td>
</tr>
<tr>
<td>Latin</td>
<td>dep</td>
<td>prim.</td>
<td>3 473</td>
<td>53 143</td>
<td>fP hR sN cB pB dL m11</td>
<td>6.74</td>
<td>2.24</td>
<td>0.41</td>
<td>12.3</td>
<td>97.45</td>
</tr>
<tr>
<td>Persian</td>
<td>dep</td>
<td>prim.</td>
<td>12 455</td>
<td>189 572</td>
<td>fM*hM sN cB pP dU m00</td>
<td>4.18</td>
<td>2.10</td>
<td>0.18</td>
<td>3.7</td>
<td>99.82</td>
</tr>
<tr>
<td>Portuguese</td>
<td>phr</td>
<td>C06</td>
<td>9 359</td>
<td>212 545</td>
<td>fS hL sN cB pB dU m10</td>
<td>2.51</td>
<td>1.95</td>
<td>0.26</td>
<td>11.1</td>
<td>99.16</td>
</tr>
<tr>
<td>Romanian</td>
<td>dep</td>
<td>prim.</td>
<td>4 042</td>
<td>36 150</td>
<td>fP* hR sN cH p0 dU m10</td>
<td>1.80</td>
<td>2.00</td>
<td>0.00</td>
<td>0.0</td>
<td>100.00</td>
</tr>
<tr>
<td>Russian</td>
<td>dep</td>
<td>prim.</td>
<td>34 895</td>
<td>497 465</td>
<td>fM hL sN cB p0 dU m10</td>
<td>4.02</td>
<td>2.02</td>
<td>0.07</td>
<td>3.9</td>
<td>99.86</td>
</tr>
<tr>
<td>Slovene</td>
<td>dep</td>
<td>C06</td>
<td>1 936</td>
<td>35 140</td>
<td>fP hR sH cH pB dL m00</td>
<td>4.31</td>
<td>2.49</td>
<td>0.00</td>
<td>10.8</td>
<td>98.87</td>
</tr>
<tr>
<td>Spanish</td>
<td>phr</td>
<td>C09</td>
<td>15 984</td>
<td>477 810</td>
<td>fS hL sN cB pB dU m10</td>
<td>2.79</td>
<td>1.98</td>
<td>0.14</td>
<td>12.7</td>
<td>99.24</td>
</tr>
<tr>
<td>Swedish</td>
<td>phr</td>
<td>C06</td>
<td>11 431</td>
<td>197 123</td>
<td>fM hL sN cF pF dU m10</td>
<td>3.94</td>
<td>2.19</td>
<td>0.13</td>
<td>0.7</td>
<td>99.66</td>
</tr>
<tr>
<td>Tamil</td>
<td>dep</td>
<td>prim.</td>
<td>600</td>
<td>9 581</td>
<td>fP hR sH cH pB dL m11</td>
<td>1.66</td>
<td>2.46</td>
<td>0.22</td>
<td>3.8</td>
<td>99.67</td>
</tr>
<tr>
<td>Telugu</td>
<td>dep</td>
<td>I10</td>
<td>1 450</td>
<td>5 722</td>
<td>fP hR sH cH pP dU m11</td>
<td>3.48</td>
<td>1.59</td>
<td>0.06</td>
<td>5.0</td>
<td>100.00</td>
</tr>
<tr>
<td>Turkish</td>
<td>dep</td>
<td>C07</td>
<td>5 935</td>
<td>69 695</td>
<td>fM hR sN cB pB dL m10</td>
<td>3.81</td>
<td>2.04</td>
<td>0.00</td>
<td>34.3</td>
<td>99.23</td>
</tr>
</tbody>
</table>
HamleDT v1.0

The chart compares the token counts of various languages. The languages are ranked from the highest to the lowest token count, showing the prominence of languages like English, German, Spanish, and Russian. The languages are grouped into categories such as Prague, Moscow, and Stanford, which likely represent different datasets or methods used to measure the token counts. The token count is on a logarithmic scale, ranging from $10^5$ to $10^6$. The x-axis lists the languages, and the y-axis represents the token count.
CoNLL (2006-2010)
Google Universal Treebank v1.0

The diagram shows the number of tokens in different languages. The Y-axis represents the number of tokens, ranging from 0 to 1x10^6. The X-axis lists the languages: English, Spanish (*Spanish), Swedish (*Spanish), French (*Spanish), Korean (*Spanish), German (*German). The English language has significantly more tokens than the other languages.
Current / Future work

- HamleDT 1.5 (29 languages, done)
- HamleDT 2.0 (Rudolf Rosa, Jan Mašek)
  - More consistent, bigger, more languages
    (Hebrew, Polish, Korean, French, Northern Sami,... )
  - Stanford dependencies instead Afun
  - English translations and alignments (Google Translate)
- Experiments with parsers and learnability
  Different styles may be better for different parsers.
Thank you

Questions?