Representing Layered and Structured Data in the CoNLL-ST Format

Jan Štěpánek, Pavel Straňák

Charles University in Prague
UFAL

ICGL 2010
Standards

Merriam-Webster's Dictionary:

- 3: something established by authority, custom, or general consent as a model or example
- 4: something set up and established by authority as a rule for the measure of quantity, weight, extent, value, or quality
Standards

Merriam-Webster's Dictionary:

- 3: something established by authority, custom, or general consent as a model or example
- 4: something set up and established by authority as a rule for the measure of quantity, weight, extent, value, or quality

cf. Henry Thompson's ad-hoc and governmental standards bodies
Easy Conversion?

- XML
  - Unicode
  - No need for (other) escape conventions
  - Ubiquity of infrastructure
- Documentation
  - Human readable

(Henry Thompson)
Various Treebank Formats

- Penn format (PTB, Penn Chinese – SGML)
  - Limited set of possible attributes and their types
- Sinica Treebank – Penn-like phrase structure with marked heads and dependency functions
- Penn Arabic – SGML + AG + Penn
- Tiger Treebank – XML
- Prague Dependency Treebank 2.0 format: PML
- Hyderabad Treebank – XML, brackets used for chunks, whitespace used to separate attributes, reference used for dependency
Hyderabad Treebank

1 (( NP <drel=k2:3>
1.1 biddalni NN ))
2 (( VGNF <drel=vmod:1/name=3>
2.1 kanetappudu VM ))
3 (( NP <drel=nmod:2>
3.1 eVMwo INTF
3.2 maMxi CL ))
4 (( NP <drel=k1:1/name=2>
4.1 wallulu NN ))
5 (( VGF <name=1>
5.1 canipowunnAru VM
5.2 . SYM ))
</Sentence>
Hyderabad Treebank

8 types of markup
Documentation

- CoNLL-ST: changes from previous year, kept at different web sites
- Sinica Treebank: 19 pages in Chinese only (MS Word DOC or PDF)
- Prague Dependency Treebank 2.0
  - Data format PML: DocBook XML (40 pages in PDF)
  - Linguistic content: Annotation manuals, DocBook XML (56 + 317 + 1287 pages in PDF)
CoNLL-ST Data Format

- Shared Task at Conferences on Computational Natural Language Learning
  - 2006-2009 dependency trees
- Used for other purposes as well:
  - e.g. ICON 2009 (parsing Indian languages), Dickinson & Ragheb (learner corpora), etc.
  - Supported by many machine learning applications
  - many treebanks have been converted into it
- De-facto standard
CoNLL-ST Data Format (2)

- **Sentence → table**
- **words → rows**
- **additional information → columns**

<table>
<thead>
<tr>
<th></th>
<th>Sentence</th>
<th>POS</th>
<th>Linenum</th>
<th>Dependency</th>
<th>Additional Info</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The</td>
<td>DT</td>
<td>4</td>
<td>NMOD</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>most</td>
<td>RBS</td>
<td>3</td>
<td>AMOD</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>troublesome</td>
<td>JJ</td>
<td>4</td>
<td>NMOD</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>report</td>
<td>NN</td>
<td>5</td>
<td>SBJ</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>may</td>
<td>MD</td>
<td>0</td>
<td>ROOT</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>be</td>
<td>VB</td>
<td>5</td>
<td>VC</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>the</td>
<td>DT</td>
<td>11</td>
<td>NMOD</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>August</td>
<td>NNP</td>
<td>11</td>
<td>NMOD</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>merchandise</td>
<td>NN</td>
<td>10</td>
<td>NMOD</td>
<td>A1</td>
</tr>
<tr>
<td>10</td>
<td>trade</td>
<td>NN</td>
<td>11</td>
<td>NMOD</td>
<td>trade.01</td>
</tr>
<tr>
<td>11</td>
<td>deficit</td>
<td>NN</td>
<td>6</td>
<td>PRD</td>
<td>deficit.01</td>
</tr>
<tr>
<td>12</td>
<td>due</td>
<td>JJ</td>
<td>13</td>
<td>AMOD</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>out</td>
<td>IN</td>
<td>11</td>
<td>APPO</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>tomorrow</td>
<td>NN</td>
<td>13</td>
<td>TMP</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>.</td>
<td>.</td>
<td>5</td>
<td>P</td>
<td></td>
</tr>
</tbody>
</table>
Problems

- Morphological information
  - Gender=Masc|Case=Nom vs. Masc|Nom
  - Same form, different representation of the same content: Number=Singular vs. num=s
Lack of Meta-Information

- Different number and meaning of the columns each year
- Meta character (easy conversion to the old form)
- Header with column description
  - # ID FORM LEMMA POS FEATS HEAD REL
  - # CoNLL-ST-2006
Identifiers

- Reference to other sentences
  - Integer (e.g. -1 = previous sentence)
  - Sentence identifiers (shuffling, cutting)
    # ID=s108
Lists

- Two ways to represent:
  - additional column per member: APRED
    - only one list per line (i.e. word)
    - preferably located in the rightmost column
  - one column with internal structure: FEATS
    - POS=N|Gen=F|Num=S
    - but Dickinson: <SUBJ, AUX, OBJ>

- List of lists
- Even more meta-characters, escaping
Multiple Layers of Annotation

- CoNLL-ST format has just a single layer
- Example: Prague Dependency Treebank 2.0
  - 4 layers, can be simplified to 2
  - relation between layer units is $M:N$ ($M,N \geq 0$)
PML – Prague Markup Language

- Not only because we are familiar with it (hopefully not NIH-syndrome)
- Rather universal: all the treebanks mentioned successfully converted
- XML
- Rich infrastructure
  - Validation tools (RNG)
  - Graphical visualization and annotation tool TrEd
  - Libraries for processing trees
  - Query language (PML-TQ) + search engines + clients
PML (2)

- **Meta-format:** PML Schema defining data types:
  - atomic – a (formatted) string
  - enumerated type – given set of possible values
  - structure – set of attribute-value pairs
  - list – (un)ordered list of units of one type
  - alternative – similar to unordered list, but with different semantics
  - sequence – similar to ordered list, but allowing members with diverse types and supporting mixed content).
PML (3)

- Roles (tree, node, order...)
- Cross-reference (e.g. coreference)
- Multi-layered
  - separated files
  - file-id#id
- Validation
  - PML Schema can be validated by a RNG Schema
  - PML Schema can be converted via XSLT to RNG Schema (validation of the data)
PDT 2.0 – Analytical and Tectogrammatical Layer

- **Analytical**: Shallow dependency syntax tree
  - One node per token, no added/deleted nodes
  - Analytical function: type of relation of a node to its parent

- **Tectogrammatical**: Deep dependency syntax tree
  - Added nodes (dropped subject, elided obligatory valency modification)
  - Deleted nodes (rather grouped together – prepositions, auxiliary verbs etc.)
  - Functor: relation to parent + many complex attributes
Which Layer as the Starting Point?

- **Analytical Layer**
  - Used in CoNLL-ST-2009
  - Includes as much of T-layer as possible (but not everything)

- **Tectogrammatical Layer**
  - Coreference – links to neighboring sentences
  - Bridging Anaphora – links between sets of nodes
  - Named Entities – hierarchical sets of nodes

- CoNLL-ST format cannot capture both structures simultaneously
Conclusion

- Simple de-facto standard format CoNLL-ST
  - A few improvements
  - Unsuitable for too complex structures
- PML for comparison
  - Complex structures (stand-off principle, various data types)
  - Rich infrastructure
- Both types useful, applications differ
Thank you.