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

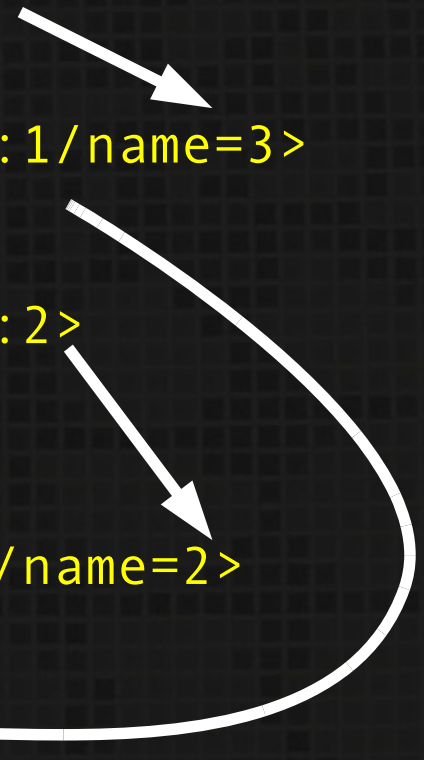
- Penn format (PTB, Penn Chinese – SGML)
  - Limited set of possible attributes and their types
- Sinica Treebank – Penn-like phrase structure with marked heads and dependency functions
- Penn Arabic – SGML + AG + Penn
- Tiger Treebank – XML
- Prague Dependency Treebank 2.0 format: PML
- Hyderabad Treebank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

Hyderabad Treebank

```
<Sentence id="8">
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

```
<Sentence id="8">
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>
```



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

1	The	the	DT	4	NMOD	_	_	_	
2	most	most	RBS	3	AMOD	_	_	_	
3	troublesome	troublesome	JJ	4	NMOD	_	_	_	
4	report	report	NN	5	SBJ	_	_	_	
5	may	may	MD	0	ROOT	_	_	_	
6	be	be	VB	5	VC	_	_	_	
7	the	the	DT	11	NMOD	_	_	_	
8	August	august	NNP	11	NMOD	_	_	AM-TMP	
9	merchandise	merchandise	NN	10	NMOD	_	A1	_	
10	trade	trade	NN	11	NMOD	trade.01	_	A1	
11	deficit	deficit	NN	6	PRD	deficit.01	_	A2	
12	due	due	JJ	13	AMOD	_	_	_	
13	out	out	IN	11	APPO	_	_	_	
14	tomorrow	tomorrow	NN	13	TMP	_	_	_	
15	.	.	.	5	P	_	_	_	

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 ≥ 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 (formated) 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.