Querying Diverse Treebanks in a Uniform Way

Jan Štěpáněk, Petr Pajáš
Charles University in Prague, MFF ÚFAL

What is “syntax”? • Different names of categories and their values • Various data formats • Different tree encoding (by structure ⇒ by reference)

Word-order typology (German CoNLL)
node $p := \{ \text{substr}(\text{pos},0,1) = 'V', \}$ node $\text{ch} := \{ \text{deprel} \in ['SB', 'OB', 'BC', 'OC', 'OP'] \}$ \[ \begin{array}{l} \text{give $p$.mmlid,} \\
\text{if($p$.deprel = 'ROOT', 'V', 'v'),} \\
\text{substr($ch$.deprel,1)},} \\
\text{$ch$.order} \\
\text{give distinct $\}$,} \\
\text{concat($2$, ' over $1$ sort by $3$)} \\
\text{give substitute($2$,'$[0][1][2]' '1', '');} \\
\text{filter ($1$ - 'O' and $1$ - 'S')} \\
\text{for $1$ give $1$.count()} \text{sort by $2$ desc} \end{array} \]

Main Num. of clause occurrences | Dependent Num. of clause occurrences
--- | ---
STO 11280 | SVO 7111
OSV 2209 | vSO 1113
OVOS 625 | OV 606
SOV 91 | OSv 109
OVOS 64 | vOS 64
OVOS 31 | OSOv 34

Grammar extraction (constituent trees)
nonterminal $p := \{ \}$
\[ \begin{array}{l} \text{give $p$, $p$.cat,} \\
\text{first\_defined($ch$.cat,$ch$.pos),} \\
\text{index\_children($ch$)} \\
\text{give $2$ & "$" over $1$ sort by $4$} \\
\text{& concat($3$, ' over $1$ sort by $4$)} \\
\text{for $1$ give $1$.count()} \text{sort by $1$ desc} \end{array} \]

Verb clause without a subject
189856  PP -- IN NP
1228140 S -- NP VP
87482 NP -- NP PP
72186 NP -- DT NN
65588 S -- NP VP
45995 NP -- NONE
36918 NP -- DT JJ NW
31916 VP -- TO VP
28796 NP -- NNP NNP
23727 SBAR -- IN S

Non-projective edges
node supervised (order-precedes Supernode and order-follows Bowser) or (order-follows Supernode and order-precedes Bowser)

Treebank | NPE/edge | NPE/tree | NPTrees
--- | --- | --- | ---
German CoNLL 2.38% 41.99% 28.19%
PDT 2.0 English CoNLL 1.89% 30.36% 22.98%
PDT 2.0 English CoNLL 0.39% 9.48% 7.63%

Prague Markup Language
PML Schema can define the following types:
- Atomic: a string, its value can further be restricted to a specific format (e.g. any, integer, date...)
- Enumerated: atomic type with a given set of possible values.
- Structure: set of attribute-value pairs.
- List: ordered or unordered list of constructs 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-TQ
- selecting all occurrences of nodes from the treebanks with given properties and in given relations w.r.t. the tree topology, cross-referencing, surface ordering, etc.
- bounded or unbounded iteration (i.e. transitive closure) of relations
- multi-layered or aligned treebanks with structured attribute values
- quantified or negated subqueries
- referencing among nodes
- natural textual and graphical representation of the query (the structure of the query corresponds to the structure of the matched subtree)
- sublanguage for postprocessing and generating reports (filtering, grouping, aggregating, and sorting)
- support for regular expressions, basic arithmetic and string operations

PML-Q

Conversion to PML

Conversion to SQL

Querying to PML