PML-TQ

Tree Query Language and software stack
PML Tree Query

- Query and report language
- Two engines
- SQL
- Perl ([bnj]tred)
- Several clients
- Browser
- TrEd
- CLI

The SQL engine requires the data to be converted to SQL and loaded into a database → stable datasets.

The Perl engine is slow, but works directly with the data files → data in progress.
PML Framework

- Multi-layered typed XML data format
- OS independent tree editor TrEd
- PML-TQ Tree Query Language and Engine(s)
- Used in development of Prague Dependency Treebank, but tested on more than ten different treebanks
Tree Query Language

- Selecting all occurrences of nodes with given properties and in given relations (tree topology, cross-referencing, surface ordering, etc.)
- (Un)bounded iteration of relations (e.g. descendant{1,3}, coref_gram.rf{1,}, sibling{-1,1}, etc.)
- Multi-layered or aligned treebanks with structured attribute values
- Quantified or negated subqueries (e.g. “find all clauses with exactly three objects but no subject”)
Tree Query Language II

- References (e.g. “find a parent and child that have the same case but different number”)
- Natural textual and graphical representation of the query
- Sublanguage for postprocessing and generating reports (filtering, grouping, aggregating, and sorting)
- Regular expressions, basic arithmetic and string operations
Reversed Dependency
Reversed Dependency

t-node [  
a/lex.rf $a\_child,  
echild t-node [ a/lex.rf  
$a\_parent ]  
];

a-node $a\_parent := [  
echild a-node $a\_child := [  
]  
];
Distribution of MWEs by their types

t-root $r := [\text{member mwes} []]$
>> \text{for } r.mwes/type
\text{give } 1, \text{count()} \text{ sort by } 2 \text{ desc}
Distribution of MWEs by their types

<table>
<thead>
<tr>
<th>lexeme</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>person</td>
<td>16708</td>
</tr>
<tr>
<td>institution</td>
<td>12243</td>
</tr>
<tr>
<td>number</td>
<td>9018</td>
</tr>
<tr>
<td>object</td>
<td>7519</td>
</tr>
<tr>
<td>time</td>
<td>6549</td>
</tr>
<tr>
<td>location</td>
<td>4434</td>
</tr>
<tr>
<td>foreign</td>
<td>256</td>
</tr>
<tr>
<td>address</td>
<td>230</td>
</tr>
<tr>
<td>biblio</td>
<td>43</td>
</tr>
</tbody>
</table>
nonterminal $p := [ * $ch := [  ] ]

>> give $p, $p.cat,
   first_defined($ch.cat,$ch.pos),
   lbrothers($ch)

>> give $2 & " → "
   & concat($3," " over $1 sort by $4)
>> for $1 give count(),$1 sort by $1 desc
nonterminal $p := [ * $ch := [ ] ]$

$>>$ give $p$, $p$.cat,
    first_defined($ch$.cat,$ch$.pos),
    lbrothers($ch$)
$>>$ give $2$ & "$ → "
    & concat($3$," " over $1$ sort by $4$)
$>>$ for $1$ give count(),$1$ sort by $1$ desc
node $p := [ 
  substr(pos, 0, 1) = "V",
  ? child node $ch := [ 
    deprel in { "SB", "OA", "OC", "OA2", "OP" } 
  ] ] 
>> give $p.xml:id,
  if($p = $ch,
    if($p.deprel = "ROOT", "V", "v"),
    substr($ch.deprel, 0, 1)),
  $ch.order
>> give distinct $1, concat($2, "" over $1 sort by $3)
>> give substitute($2, "([OS])\1+", "\1", "g")
>> filter ($1 ~ "O" and $1 ~ "S")
>> for $1
  give $1, if($1 ~ "V", count(), 0), if($1 ~ "V", 0, count())
>> give $1, if($1 ~ "V", count(), 0), if($1 ~ "V", 0, count())
  percnt(ratio($2 over all) + ratio($3 over all), 3)
  sort by $2, $3 desc
>> give $1, $3 & " %"
<table>
<thead>
<tr>
<th>Main clause</th>
<th>Number of occurrences</th>
<th>Dependent clause</th>
<th>Number of occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVO</td>
<td>52.336 %</td>
<td>SOv</td>
<td>63.942 %</td>
</tr>
<tr>
<td>VSO</td>
<td>33.031 %</td>
<td>SvO</td>
<td>19.235 %</td>
</tr>
<tr>
<td>OVS</td>
<td>10.261 %</td>
<td>vSO</td>
<td>9.419 %</td>
</tr>
<tr>
<td>VOS</td>
<td>2.903 %</td>
<td>OSv</td>
<td>5.128 %</td>
</tr>
<tr>
<td>SOV</td>
<td>0.511 %</td>
<td>OvS</td>
<td>0.922 %</td>
</tr>
<tr>
<td>OVSO</td>
<td>0.423 %</td>
<td>vOS</td>
<td>0.542 %</td>
</tr>
<tr>
<td>VOSO</td>
<td>0.297 %</td>
<td>SOvO</td>
<td>0.313 %</td>
</tr>
</tbody>
</table>
Parallel Treebanks

Prague Czech-English Dependency Treebank 2.0

t-node [
    t_lemma = "force",
    alignment/counterpart.rf t-node [
        t_lemma != "síla"
    ]
]

Tuesday, October 22, 13