Extensions

Corpus Query Language and Sketch Grammar

Pavel Rychlý

February 4, 2010

Outline

Within keyword

within works with any subquery not only a structure

```
[lemma="dream"] within ([word="my"] [lemma="dream"])

MU (meet [lemma="dream"] [word="my"] -1 -1)

[word="the"] []{0,3} [lemma="dream"]
        within ([tag="AT."][tag="AJ."]{0,4}[tag="NN."])
```

containing keyword

New containing keyword

- inverts within keyword
- matches results of the first subquery which contains matches of the second subquery

```
<phr/> containing [lemma="dream"]

MU (meet [lemma="dream"] [word="my"] -1 -1)

[word="the"] []{1,3} [lemma="dream"]
        containing [lemma="wild"]
```

Combinations of containing/within

Both keyword forms a query which can be used as subquery, they can be nested.

```
[lemma="break"] within (<s/> containing [lemma="rule"])

[lemma="student"] within
      (<s/> containing [lemma="break"]
            containing [lemma="rule"])

[lemma="break"] within ([]{5} containing [lemma="rule"])
```

Separate page

New *SEPARATEPAGE directive

- follwing TRINARY relation will be displayed on a separate page with only links from the main wordsketch page
- optional parameter is the name of the aggregated gramrel name, it defaults to the relation name with %s substituted to '*':
  `pp_%s` is displayed as `pp_*`

```
*SEPARATEPAGE pp_X
*TRINARY
=pp_%s
    1:[tag="N.."|tag="AJ."] 3:"PR." 2:"N.."
```

Constructions

New *CONSTRUCTION directive

- *CONSTRUCTION indicates that the following gramrel should be displayed in the 'Constructions' list (if salient)
- displayed always at the beginning of a word sketch
- both unary and binary relations supported

```
*CONSTRUCTION -n$
*UNARY
=wh
        1:any_noun adv_string wh_not_when

*CONSTRUCTION
=PP_Ving
        1:any_noun 2:any_prep adv_string ing_verb
```

Multi-word collocations

New *COLLOC directive

- specifies a created value for the collocation
- it could contain '%' substitution strings, in the form '%(n.attr)', where n is the numeric label used in the query, and attr is the attribute name
- it use a created value for the collocation instead of the attribute given by the WSATTR option

```
=PP_PP
*COLLOC "%(2.lemma)_%(3.lemma)-p"
        1:any_noun 2:any_prep short_np0 3:any_prep
        1:any_noun 2:any_prep any_pro 3:any_prep
```

Order of gramrels

New *FIXORDER directive

- list of gramrel names
- different order for different PoS

Multi-level tokenization

How to handle different needs for tokenization.

- tokenize into the smallest elements
- define other levels
  - delete selected tokens (matching a query)
  - join several tokens into one token
    (matching a query, covered by a structure)
- choose one (or more) levels for making queries, computing frequencies and collocations