CGN, asking for collaborating cats: ISOcat, SCHEMAcat and RELcat

Ineke Schuurman
KU Leuven and U.Utrecht
ineke@ccl.kuleuven.be

Menzo Windhouwer
The Language Archive – DANS
menzo.windhouwer@dans.knaw.nl

Corpus Gesproken Nederlands
http://lands.let.kun.nl/cgn/ and available via http://tst-centrale.org/

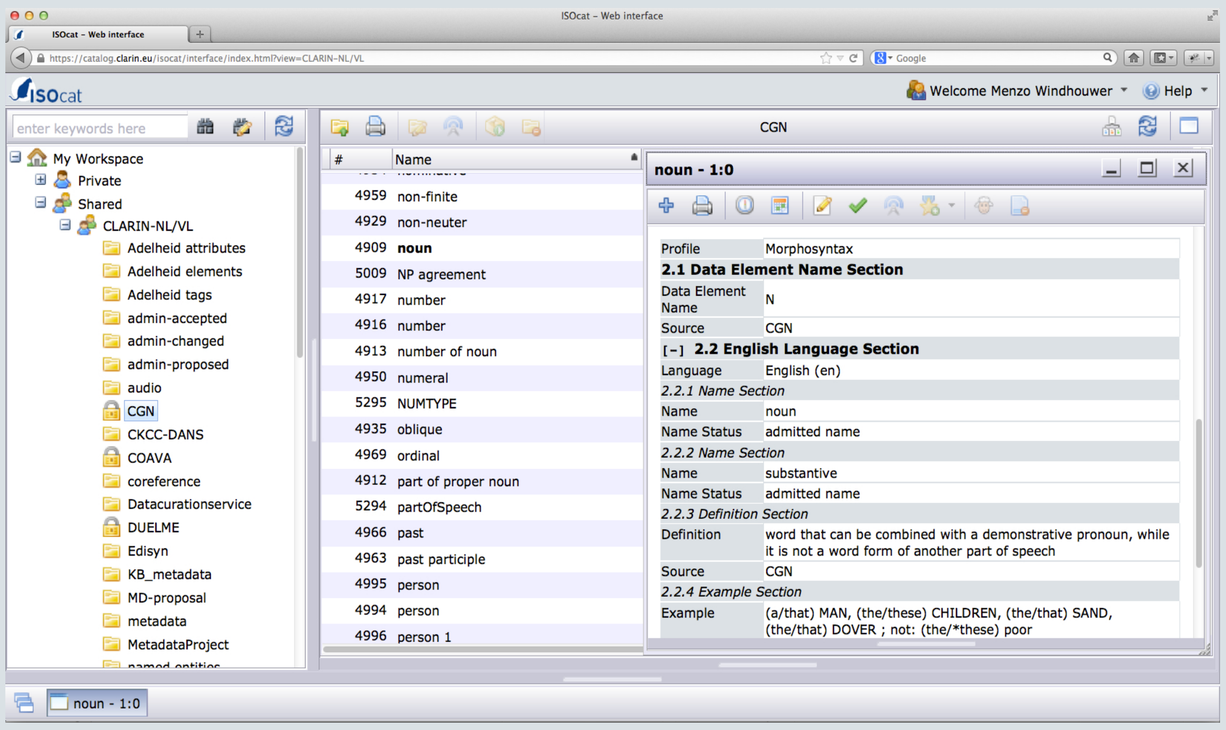
CGN (Spoken Dutch Corpus): a 10M corpus of contemporary Dutch as spoken in the Netherlands and Flanders. The tagset was reused in large corpora of written Dutch. Many CLARIN-NL projects use CGN.

ISOcat: a Data Category Registry
http://www.isocat.org/

CLARIN-NL defines metadata and linguistic concepts in ISOcat. For both types good definitions are needed. ISOcat as such doesn't enforce strict rules although it provides guidelines. CLARIN-NL has much stricter rules, which are enforced by the content coordinator. Important rule:

As general as possible, as specific as necessary.

As a consequence full complex CGN tags like `N(soort,mv,dim)` are *not* as such to be included in ISOcat, but the components are.



http://www.isocat.org/interface/index.html?view=CLARIN-NL/VL

SCHEMAcat: a Schema Registry
http://lux13.mpi.nl/schemacat/ (alpha)

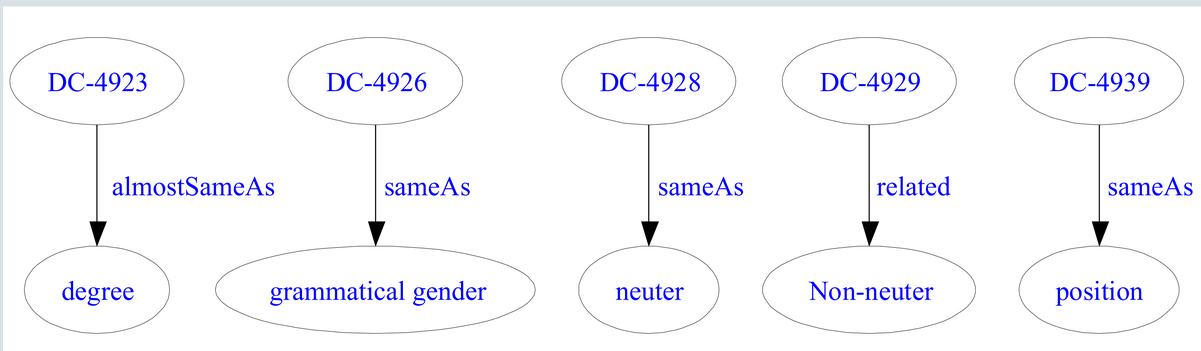
In SCHEMAcat schemata for language resources are stored. A wide variety of schema languages can be used. The CGN tagset is described with an ISO 14977:1994 compatible EBNF annotated with ISOcat data category references. Using this EBNF a parser can be generated to return the ISOcat data categories used by a complex CGN tag.

```
(* @dcr:datcat 'N' http://www.isocat.org/datcat/DC-4909 *)
tag = 'N',    '(', NTYPE,    ',', GETAL,   ',', GRAAD ')'... ;
...
(* @dcr:datcat NTYPE   http://www.isocat.org/datcat/DC-4908 *)
(* @dcr:datcat 'soort' http://www.isocat.org/datcat/DC-4910 *)
(* @dcr:datcat 'eigen' http://www.isocat.org/datcat/DC-4911 *)
NTYPE = 'soort' | 'eigen' ;
...
```

http://hdl.handle.net/1839/00-SCHM-0000-0000-004A-A (snippet)

RELcat: a Relation Registry
http://lux13.mpi.nl/relcat/ (alpha)

In RELcat ontological relationships between ISOcat data categories can be defined. These relationships can be between data categories from one selection, like CGN, or between several selections, i.e., to provide crosswalks between CGN and GOLD or STTS.



http://lux13.mpi.nl/relcat/set/cgn (snippet)

The semantic network created by these collaborating cats can be exploited in various ways. For example:
- semantic documentation
- fuzzy semantic search
- crosswalks to other models

In the future the proper place in this network for CGN T and U tag abbreviations will be determined. This place might be the CLAVAS Vocabulary Service currently under construction by CLARIN-NL.