Explicit Semantics for Enriched Documents.
What Do ISOcat, RELcat and SCHEMAcat Have To Offer?

Ineke Schuurman[a], **Menzo Windhouwer**[b]

[a]Katholieke Universiteit Leuven, [a]University of Utrecht
Blijde Inkomststraat 13, Leuven, Belgium
[b]Max Planck Institute for Psycholinguistics
Wundtlaan 1, 6525 XD Nijmegen, The Netherlands
[a]ineke@ccl.kuleuven.be, [b]Menzo.Windhouwer@mpi.nl

Abstract

Most documents researched in the human and social sciences will be enriched one way or another, at least with metadata. Sometimes documents are also enriched with one or more types of annotation. Often the notions used can be interpreted in several ways, which raises the question: "What is meant in a particular case?"

ISOcat is a ISO 12620:2009 compliant registry in which such notions, in the context of this registry called data categories, are described in a concise way. Some data category descriptions will be standardized, meaning that their use is promoted. In general the descriptions are meant to be useful for as many users as possible; on the other hand specific uses might require very specific readings of a notion. This might lead to the creation of multiple data categories which are semantically close. As relationships between data categories are not part of the data model provided by ISO 12620:2009, because they could restrict data categories too much to a specific context which would hamper their reuse, a companion registry named RELcat is under construction. This registry will allow users to store and share their view on the relationships between data categories and possibly concepts from other sources. Finally a schema registry named SCHEMAcat is created which allows the reuse of schemas annotated with ISOcat data categories and typed relations in RELcat. The flexible semantic network thus created can be used to help users of enriched documents, and actually any other type of linguistic resource, to interpret the linguistic notions and to find semantically close and thus possibly interesting resources. This paper illustrates the semantic network with examples taken from the CGN PoS tagset for Dutch.

Introduction

Ideally, metadata, i.e. data about the data contained in a document, a corpus, etc. should always be available for documents used for research, be it textual, audio or visual media, or a mixture of these. They describe the nature of the document (text, recordings, images), its size, year of creation, recording date, age of interviewees, place of residence, etc. For each type of document there is a series of relevant metadata: in the digitized text of a law there will be no interviewees whereas there will be in an opinion poll, independent of its format.

However, there is no fixed set of metadata, i.e. a metadata scheme, to describe resources. As a consequence, several notions are used to describe the same state of affairs (is the country referred to as 'Nederland', 'NL', 'The Netherlands'?), or, the other way round, the same notion is used to describe different states of affairs: 'synchronic' in linguistics (describing linguistic phenomena at a specific moment in time) vs. its more general meaning (several things occurring in parallel).

Somewhere applicable meanings of specific notions are to be made clear, both for the benefit of human users, like a researcher enriching a resource with either metadata or (linguistic) annotations or using data from such a resource for research purposes, and for computer programs.

Researchers and programs can benefit from linguistic enrichments (annotations) for all kinds of documents like part-of-speech (pos) tagging, syntactic analysis and several kinds of semantic annotations (like named entity recognition, coreference labeling, semantic roles, temporal annotation, spatial annotation, sentiment analysis, etc. etc.). This could, for example, improve text mining results. In Flanders and the Netherlands, a series of smaller and larger CLARIN-related[1] projects are being executed in order to show this to the field.

However, a correct interpretation of results depends to a large extent on the correct understanding of the notions used, be it metadata or another kind of enrichment, especially when several annotations are available. [2]

In this paper we will first discuss the opportunities offered by ISOcat[3], the ISO 12620:2009 (ISO 12620, 2009) compliant Data Category Registry (DCR), and its companion registries RELcat and SCHEMAcat which are currently under development. Although the approach described can be used in a wide range of applications and domains we will make use of the CGN[4] PoS tagset (Van Eynde, 2004), a *de facto* standard for Dutch.

Referring to data categories in ISOcat

To make the semantics of the tags in CGN explicit they are linked to the data categories, i.e. (linguistic) notions and the values associated with them, available in ISOcat. For these references ISO 12620:2009 provides a small XML vocabulary in Annex A[5] consisting mainly of the core attributes `dcr:datcat` or `dcr:valueDatcat` to be

[1] Cf. http://www.clarin.nl, http://www.ccl.kuleuven.be/CLARIN

[2] Either different types, like pos and co-reference, or the same type, using different tagsets (for one or more languages).

[3] Cf. http://www.isocat.org/ hosted by The Language Archive

[4] 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.

[5] This XML vocabulary is also online available at: http://www.isocat.org/12620/schemas/DCR.html

used in any XML document or XML schema (Windhouwer, Kemps-Snijders and Wright, 2010), e.g. a Relax NG or a W3C XML Schema document. However, the CGN grammar, which can be seen as the schema of the tagset, isn't XML but text-based. To still embed the references to the data categories a small annotation vocabulary was created. The following snippet[6] shows the main tag structure and some annotated part-of-speech tags and features.

```
(* @dcr:datcat pos  DC-1345 *)
tag    = pos '(' feat* ')' ;
(* @dcr:datcat 'WW'  DC-1424 *)
(* @dcr:datcat 'TW'  DC-1334 *)
(* @dcr:datcat 'VG'  DC-1260 *)
pos    = 'WW' | 'TW' | 'VG' | … ;
(* @dcr:datcat PVTIJD DC-1286 *)
feat   = PVTIJD | … ;
(* @dcr:datcat 'verleden' DC-1347 *)
(* @dcr:datcat 'conjunctief' DC-1843 *)
PVTIJD = 'verleden' |
         'conjunctief' | … ;
```

Figure 1: CGN schema snippet

Using the fully annotated schema it can be distilled that the following CGN tag `WW(persoonsvorm, verleden, enkelvoud)`[7] uses the data categories:

- */verb/* (DC-1424) associated with the CGN part-of-speech tag `'WW'`;
- */partOfSpeech/* (DC-1345) associated with the CGN grammar non-terminal symbol `pos` and containing the simple data category */verb/* (DC-1424) in its value domain;
- */finite/* (DC-1287) associated with CGN value `'persoonsvorm'` for the feature `WVORM`;
- */finiteness/* (DC-1893) associated with the CGN feature `WVORM` and containing the simple data category */finite/* (DC-1287) in its value domain;
- */past/* (DC-1347) associated with CGN value `'verleden'` for the feature `PVTIJD`;
- */tense/* (DC-1286) associated with the CGN feature `PVTIJD` and containing the simple data category */past/* (DC-1347) in its value domain;
- */singular/* (DC-1387) associated with CGN value `'enkelvoud'` for the feature `GETAL`;
- */number/* (DC-2709) associated with the CGN feature `GETAL` and containing the simple data category */singular/* (DC-1387) in its value domain.

Notice that this already gives a low level of explicit semantics and thus semantic interoperability. For any other tagset linked to the same data categories semantic equivalences can be established.

The data categories used until now were all taken from the set created by the ISOcat Morphosyntax Thematic Domain Group (TDG) (Francopoulo et al., 2008) and were checked by CGN experts on semantic equivalence. Unfortunately reusing existing data categories isn't always possible. Specific design choices or language features influence the semantics of the data category concepts used in the tagset.

In CGN, for example, the notion *noun* (part-of-speech tag N) is used in a specific way. Whereas the data category */noun/* (DC-1333) is formulated in a neutral way, in CGN a more specific semantic interpretation is used. The CGN interpretation is that from a morphosyntactic perspective only the word type is of importance, not its use. So, although a word like *Monday* (as in *Should I go see him Monday?*) is often used as an adverb, it is still considered to be a noun in a form-driven approach (as advocated by CGN). In a function-driven approach, however, it is considered an adverb. Something similar holds for *poor* in *That day the poor were obliged to …* . In a form-driven approach *poor* is analysed as a nominalized adjective, i.e. an ADJ(nom), whereas in a function-driven approach, as used the Dutch PAROLE tagset (Dutilh and Kruyt, 2002), it is considered a noun. Therefore the CGN notions *N* and *ADJ(nom)* both maintain a *part of* relation with */noun/* (DC-1333) (cf the section on RELcat).

Also the data model specified in ISO 12620:2009 and implemented in ISOcat requires a specific data category type to be assigned to a data category. The basic types are:

- *complex data categories* which have a, possibly enumerated, conceptual domain,
- *simple data categories* which represent an value in a conceptual domain of one ore more complex data category, and
- *container data categories*[8] which do not have a conceptual domain but are used in linguistic resources to group other data categories.

A tagset or another type of linguistic resource might use the same linguistic notion but due to its realization in the resource might require a different type. For example, in a tagset *noun* has to be a simple data category as it's a possible value of the complex data category */partOfSpeech/* but in a feature structure representation of a parse tree *noun* might be a complex data category which gets assigned the noun phrase of the parsed text. Although the same underlying linguistic notion or data category concept is meant their usage in the linguistic resource enforces the creation of two different data categories of different types in the ISOcat registry.

This means that due to theoretical or technical reasons new data categories might have to be created. But notice that the aim should still be to specify them in ISOcat in such a manner that reuse is encouraged. This means, for example, that the definitions should be

- as neutral as possible, but still reflecting the characteristic properties of a specific annotation scheme,
- maximally language-neutral, and

[6] Due to space limitations the data category references have been abbreviated. All these references should be a full URL using http://www.isocat.org/datcat/ as base. For example the reference DC-1424 should be read as http://www.isocat.org/datcat/DC-1424.

[7] Spaces have been added for readability. For the same reason abbreviations have been spelled out.

[8] The container data category type is not part of the ISO 12620:2009 DCR data model, but a recent addition sanctioned by ISO TC37. The CGN EBNF grammar in Figure 1 has not been annotated with container data categories yet. In that case also the non-terminal symbols `tag` and `feat` would be associated with data categories.

- not be in a circular argument with other definitions.

The CGN part-of-speech tag `VNW` which represents the linguistic notion *pronoun* is an example where the creation of a new data category is required for theoretical reasons. In CGN *pronoun* is used in broader sense than the definition of the /*pronoun*/ (DC-1370) data category, as created by the TDG, allows. To describe this tag a new data category /*pronoun*/ (DC-4109)[9] was created.
In CGN the tag for a specific type of pronoun, e.g., an indefinite pronoun, consists of the combination of a series of data categories. Therefore the definition of *pronoun* should be such that it holds in all cases, one way or another. When *indefinite pronoun* is defined as a unit, cf /*indefinitePronoun*/ (DC-1309), it can by-pass it. Note, however, that although /*indefinitePronoun*/ (DC-1309) is said to be in an 'Is A' relation with /*pronoun*/ (DC-1370), it can also be considered a kind of contradiction.

Relating data categories in RELcat

As described in the previous section specific needs might lead to the creation of different data categories which can be considered semantically very close. These ontological relationships between data categories are very valuable as they provide means to, for example, enable a semantic search algorithm to broaden its scope and also return close matches.
Early in the design of ISOcat it was decided that ontological relationships can not be a part of the standardized core of the registry (Kemps-Snijders et al., 2009), i.e., these kind of relationships are too specific for an application or resource and might thus hinder reusability.[10] A companion registry to ISOcat, a Relation Registry called RELcat, is currently under construction and does support the storage of (user-specific) sets of relations.
The following snippet shows the relationship between the two /*pronoun*/ data categories mentioned in the previous section.

```
…
relcat:cgn {
 …
 isocat:DC-4109 rel:almostSameAs isocat:DC-
1370.
 …
}
```

Figure 2: CGN relation set snippet

This example also shows that the relations are semantically typed and this allows the registry to do a limited amount of reasoning with them, e.g., the reverse of the shown *almost same as* relation is also considered a valid relationship.
Currently the following relationship types are supported:

1. *related*
   - 1.1. *same as* (a symmetric and transitive relationship)
   - 1.2. *almost same as* (a symmetric relationship)
   - 1.3. *narrower* (the inverse of the 'broader' relationship)
     - 1.3.1. *super class of* (the inverse of the 'sub class of' relationship)
   - 1.4. *broader* (the inverse of the 'narrower' relationship)
     - 1.4.1. *sub class of* (the inverse of the 'super class of' relationship)
   - 1.5. part of
     - 1.5.1. *direct part of*
     - 1.5.2. *indirect part of*

These relationship types are placed in an extensible taxonomy. This means that other relationship types from other vocabularies, e.g., OWL or SKOS, can be inserted at their proper place in this hierarchy and sets of relations using these vocabularies can be loaded in RELcat. For example, *owl:sameAs*, *owl:equivalentClass* and *owl:equivalentProperty* can be assigned as subtypes of *same as* in the RELcat taxonomy. Linguistic ontologies expressed in OWL, like the GOLD ontology[11] (Farrar and Langendoen, 2010) or the OLiA Reference Model (Chiarcos, 2010), can then be directly imported into RELcat. The generic part of the taxonomy can then be used by generic algorithms to traverse the graph created by the relationships from possibly mixed origins.
Until now only simple binary relationships were used, but to enable more complex mappings between tagsets *n*-ary relationships will be needed. Just like many semantic web and linked data approaches RELcat's basic building block are typed binary relationships so to enable *n*-ary relationships reification is needed. This means that the *n*-ary relationship itself becomes a resource in the registry and is able to participate in the needed *n* binary relationships with data categories or other resources. First experiments with these complex mappings are also started in the context of the CGN tagset, where they are expected to be needed to deal with shortcuts assigned to valid tags, e.g., `T304` is equivalent with the tag `WW(persoonsvorm, verleden, enkelvoud)`.

Publishing a schema in SCHEMAcat

There are already many (multimedia) resources for Dutch which make use of the CGN tagset. To enable sharing the version of the schema which is annotated with data category references, as illustrated with a snippet in Figure **1**, another registry is taking shape: the Schema Registry SCHEMAcat. This registry will become a place to persistently store schemata of various kinds, e.g., XML schema's but also text-based (EBNF) grammars like CGN. In the case where variants of a schema exist they will also be stored in SCHEMAcat. For CGN there is, for example, the D-COI/SONAR variant (Van Eynde, 2005) which allows some tags which are considered invalid in CGN and there also exists a shortened version of the

[9] In ISOcat mnemonic identifiers do not need to be unique, as long as the PID of the data category is unique. This allows various domains to use the same term, and using the PID to resolve ambiguity.
[10] Due to legacy reasons the DCR data model does support *is a* relations between simple data categories. This relationship type is equivalent to the *sub class of* type in RELcat. Notice, that a simple data category can only take part in one subsumption hierarchy in ISOcat but in a theoretically unlimited number in RELcat.
[11] Cf. http://linguistics-ontology.org/

CGN tagset,[12] used for example as input for the syntactic annotation (Schuurman et al., 2003). Instead of the tag associated with `T304`, a tag `WW1` is used which is equivalent with `WW(persoonsvorm, enkelvoud)`. Note that these tags generalize over the present and past tense forms.

This schema registry has a fundamental right to exist on its own. It will support long term storage of schemata, i.e., decoupling a schema from a relatively temporary resource creating tool or service and safeguarding its existence for archived resources based on this schema.

From the viewpoint of RELcat it is also a place to harvest (untyped) relationships between data categories, which then can be fine tuned by a user to her specific needs.

In the context of exploiting the semantic network created by the combination of ISOcat, RELcat and SCHEMAcat for semantic searches the function of SCHEMAcat will be on two levels:

1. identify candidate matching archived resources based on their SCHEMAcat schema containing patterns of specific relationships between specific data categories, and
2. check that the resource actually instantiates this pattern by running a specific validation method.

The second level might not be possible for all schema types, but SCHEMAcat should allow to plugin these specific validation methods. For example, an EBNF plugin would be able to parse the CGN tag `WW(persoonsvorm, verleden, enkelvoud)` and return the used set of data categories, i.e., {/*verb*/ (DC-1424), /*partOfSpeech*/ (DC-1345), /*finite*/ (DC-1287), /*finiteness*/ (DC-1893), /*past*/ (DC-1347), /*tense*/ (DC-1286), /*singular*/ (DC-1387), /*number*/ (DC-2709)}, which is just a small subset of all the data categories associated with the CGN tagset. Also only some of these data categories and the ontological relationships they have among each other maybe specified in RELcat. Notice, however, that within the context of CLARIN the support of semantic context search and the possible use of ISOcat, RELcat and SCHEMAcat is a task of an individual CLARIN centre. In order to synchronize the efforts of linguists and technologists curating resources and these centres a series of workshops and tutorials[13] are being organized.

Conclusion and future work

This paper discussed some steps taken to make the semantics of enriched documents explicit by relating them with data categories in the context of ISOcat. To achieve a higher level of semantic interoperability, i.e., not only based on finding shared data categories, ontological and contextual information will have to be taken into account. RELcat and SCHEMAcat will provide the means to harvest and specify this information in the form of relationships and allow (search) algorithms to traverse the semantic graph thus made explicit. Using this combination of various registries it has become possible to fully describe a tagset, which was basically impossible in ISOcat alone as the experiment for the Polish NKJP tagset indicated (Patejuk and Przepiórkowski, 2010). Future work includes extending the model of RELcat to deal with *n*-ary relationships and also SCHEMAcat will need more functionality, e.g., metadata profiles to state specific characteristics of a tagset.

[12] Whereas the full tagset employs 320 tags, the reduced one contains only some 50 more generic tags.

[13] We are also in the process of making available a booklet on how ISOcat entries should be constructed in order to make them as reusable as possible without neglecting the theoretical characteristics of a specific annotation scheme.

References

Chiarcos, C. (2010). Grounding an Ontology of Linguistic Annotations in the Data Category Registry. Language Resource and Language Technology Standards workshop at LREC 2010, Malta.

Dutilh, T. and T. Kruyt (2002). Implementation and Evaluation of PAROLE PoS in a National Context. Third International Conference on language and Evaluation (LREC). Las Palmas - Canary Islands, Spain: 1615-1621.

Farrar, S. and D. T. Langendoen (2010). An OWL-DL implementation of GOLD: An ontology for the Semantic Web. Linguistic Modeling of Information and Markup Languages: Contributions to Language Technology. A. W. Witt and D. Metzing. Dordrecht, Springer.

Francopoulo, G., T. Declerck, et al. (2008). Data Category Registry: Morpho-syntactic and Syntactic Profiles. Uses and usage of language resource-related standards workshop at LREC. Marrakech, Marocco.

ISO 12620 (2009). Terminology and other language and content resources - Specification of data categories and management of a Data Category Registry for language resources, International Organization for Standardization.

Kemps-Snijders, M., M. Windhouwer, et al. (2009). ISOcat: remodelling metadata for language resources. International Journal on Metadata, Semantics and Ontologies 4(4).

Patejuk, A. and A. Przepiórkowski (2010). ISOcat Definition of the National Corpus of Polish Tagset. Language Resource and Language Technology Standards workshop at LREC 2010, Malta.

Schuurman, I., M. Schouppe, et al. (2003). CGN, an annotated corpus of Spoken Dutch. 4th International Workshop on Linguistically Interpreted Corpora (LINC-03). A. Abeillé, S. Hansen-Schirra and H. Uszkoreit. Budapest, Hungary: 101-112.

Van Eynde, F. (2004). Part of Speech Tagging en Lemmatisering van het CGN Corpus, Centrum voor Computerlinguïstiek, KU Leuven.

Van Eynde, F. (2005). Part of Speech Tagging en Lemmatisering van het D-Coi Corpus, Centrum voor Computerlinguïstiek, KU Leuven.

Windhouwer, M., M. Kemps-Snijders, et al. (2010). Referencing ISOcat Data Categories. Language Resource and Language Technology Standards workshop at LREC 2010, Malta.