Annotation by category - ELAN and ISO DCR

Han Sloetjes, Peter Wittenburg
Max Planck Institute for Psycholinguistics
P.O. Box 310, 6500 AH Nijmegen, The Netherlands
E-mail: Han.Sloetjes@mpi.nl, Peter.Wittenburg@mpi.nl

Abstract

The Data Category Registry is one of the ISO initiatives towards the establishment of standards for Language Resource management, creation and coding. Successful application of the DCR depends on the availability of tools that can interact with it. This paper describes the first steps that have been taken to provide users of the multimedia annotation tool ELAN, with the means to create references from tiers and annotations to data categories defined in the ISO Data Category Registry. It first gives a brief description of the capabilities of ELAN and the structure of the documents it creates. After a concise overview of the goals and current state of the ISO DCR infrastructure, a description is given of how the preliminary connectivity with the DCR is implemented in ELAN.

1. Introduction

The ISO Data Category Registry is rapidly becoming a mature infrastructure for standardization in Language Resources. For researchers in linguistics to exploit the full potential of the DCR, tools need to be available that can interact with this infrastructure. Lexus [1], the lexicon tool that is being developed at the Max Planck Institute for Psycholinguistics, and GATE [2], developed at Sheffield University, were the first tools to offer such interaction.
Now ELAN, a multipurpose, multimedia annotation tool, has also been extended with this capability. Its users can now create a reference from annotations and tiers to concepts defined in the central ISO DCR. User defined complex categories can be integrated as well as custom Data Category Selections.

2. About ELAN

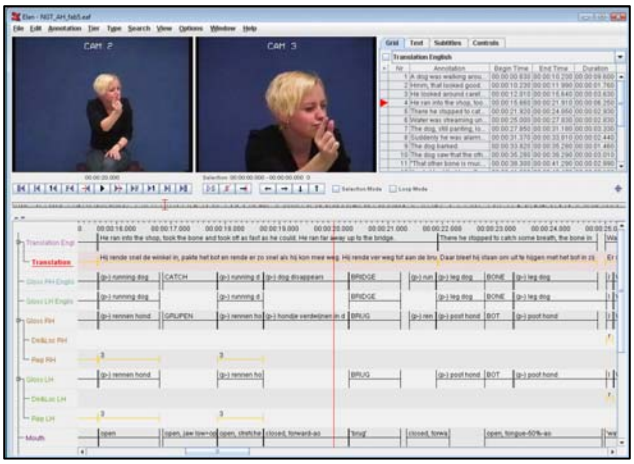

Figure 1: an ELAN window

ELAN is a tool for the manual creation of annotations to audio and/or video files [3]. In its most elementary form an annotation is a piece of text referring to a segment of the media. In the case of audio or video a segment is usually identified by a begin time and an end time, both referring to a point in the media's timeline.

2.1 Structure of an annotation document

In ELAN each annotation is part of a tier, which is a kind of layer or row. Put differently, a tier is a container for annotations. Annotations on the same tier typically refer to the same kind of phenomenon (e.g. utterances of speaker A, gestures of the left hand of participant B). The content of any annotation is (Unicode) text.
Multiple tiers can be created and these tiers can be organized hierarchically by defining parent-child relations that constitute a dependency or subordination. Annotations on a tier without a parent tier (i.e. a top level, independent tier) are always time-aligned, meaning that they are characterized by a begin time and an end time, identifying a segment in the media. Annotations on all other tiers directly or indirectly refer to an annotation on the parent tier, possibly in combination with their own temporal references to the media, thus constituting increasingly complex layers of analysis.
There is a number of predefined relations that annotations on a depending tier can have with an annotation on the parent tier and these relations are made explicit in a set of stereotypic constraints. Apart from a one-to-one relation (in ELAN identified as “Symbolic Association”), several kinds of one-to-many relations have been defined (“Time Subdivision”, “Symbolic Subdivision” and “Included In” in ELAN). (A many-to-one relation has not yet been implemented).
The constraints on time-alignment, parent-child relation and possibly the contents of annotations are collected in so called “Linguistic Type” objects, which can be defined and composed by the user. A tier is associated with one linguistic type and multiple tiers can share the same linguistic type object. Creation of tier dependencies is optional and the user is entirely free in the design of a setup of linguistic types and tiers that best suits the objectives of the research. For example, in field linguistics it is fairly common to distinguish at least 3 or 4 depending layers (text, words, part of speech or text, words, morphemes, glosses) while in gesture or sign

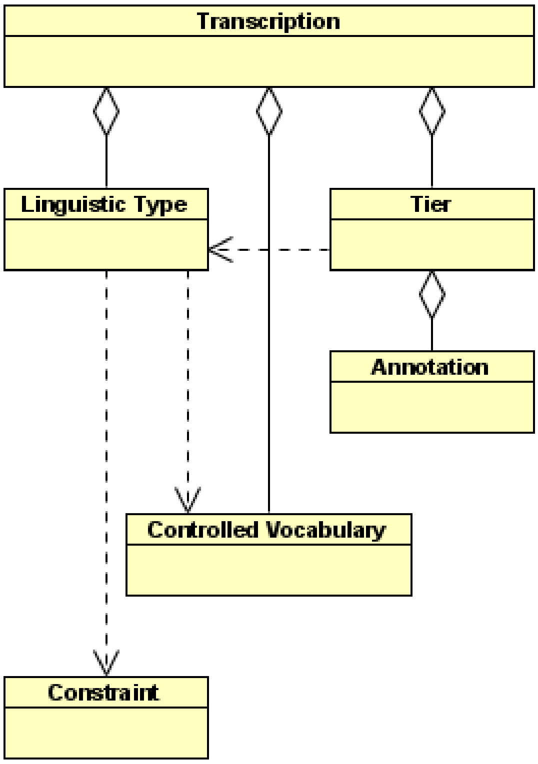

Figure 2: simplified model of an ELAN transcription

language research often a lot of independent tiers are used with an occasional occurrence of a 2-level deep dependency.

2.2 Controlled Vocabularies

Also part of the linguistic type object can be a reference to a controlled vocabulary, which puts a constraint on the content of the annotations on the referencing tier. A controlled vocabulary is a user definable list of values that are likely to be related in some way and that the user plans to apply to annotations on one or more tiers. These vocabularies are completely under the control of the user: entries can be added, modified or deleted as needed. When creating or editing annotations on a tier that has been associated with a controlled vocabulary, a listing of the entries of that CV is presented to the user, who can then select the appropriate value. E.g. a user could create a vocabulary containing the Part of Speech tags that are needed in a set of related transcriptions.

2.3 Template files

As with the linguistic type and tier setup, the design and creation of controlled vocabularies and the annotation coding scheme applied, is fully under the user's control. This setup can be stored and re-used for other transcriptions by means of the ELAN template files; a template file contains the definitions of the tiers and linguistic types as well as user defined controlled vocabularies. Template files can be shared and modified and existing annotation files can import changes made to the template. Coherent sets of annotation resources can thus be created by teams working on the same project or even across projects.

This freedom to create one's own tier setup and to follow one's own ideas about classification schemes is not just a luxury, it is often a necessity. Especially in the study of indigenous languages, where native speakers often are involved in the transcription process, it is important not to be forced into the use of an international (English) encoding scheme. The same holds for deaf studies where signers are involved who may not be fluent in English. As a result, heterogeneous corpora emerge that are difficult or impossible to compare or to be subject of cross corpora search. In the best case there is consistency within the single corpus.

3. Flexibility vs. interoperability or the best of both worlds?

Even when merely considering language resources that are available in digital form, an enormous variety can be seen. For both lexical and annotation resources, a whole range of tools have come into existence, each with its own specialization and each targeted at a certain audience or a certain kind of task. These tools often come with their own way of structuring and storing the data, including an own format. Added to the variety of existing coding conventions the result is a tangle of diverse resources.

But, awareness of the need for standardization and interoperability seems to be growing in the linguistics community. This holds for the level of tool formats but also for the level of the use of linguistic concepts. There is always a conflict between care for interoperability and adherence to standards on the one hand and the need for freedom and originality on the other (Rumble et al., 2005). Infrastructures like the DCR and related ontologies bear the promise to pacify this conflict.

While leaving practitioners a great amount of freedom to stick to their own encodings, interoperability and compatibility would be assured through references to standardized categories.

4. About ISO Data Category Registry

The ISO Data Category Registry (ISO 12620) is being developed as part of the ISO TC 37/SC 4 [4] efforts, which are aimed at the establishment of standards for Language Resource management, creation and coding. The DCR specifies the names and definitions for data categories relevant to the language resource domain as well as the management procedures for a Data Category Registry for Language Resources.

4.1 Registry requirements

The ISO DCR is required to be accessible online and free of charge. Data categories that are publicly available, i.e. those categories with registration status “standard”, should be conveniently browsable. The registry must provide tools for administration and registration tasks, including a personal working space for involved experts.

For tools to interact with the ISO DCR an Application Programming Interface (API) has been designed

(Kemps-Snijders et al., 2006). The DCR host implements this API and makes it available as a web service. The first host of the ISO DCR is the Inria Syntax Server [5], its successor ISOcat [6] is expected to continue this task from the middle of 2008.

4.2 Data Categories

The categories managed by the registry are elementary descriptors of linguistic concepts (e.g. "part of speech", "grammatical number") and a category can be either simple (simple, atomic data category, e.g. "masculine") or complex (complex data category, e.g. "gender", with a value range "masculine", "feminine", "neuter"). The value range, on the concept level referred to as the conceptual domain, of a complex data category consists of simple data categories. The DCR distinguishes the “working language” and the “object language” of data categories: the working language is the language used for the description of a category, while the object language is the language being described. The value range of a complex category can be dependent on the object language (some languages may only have the values “masculine” and “feminine” for “gender”).

The registry defines a set of thematic profiles, such as “Terminology”, “Syntax”, “Morpho-Syntax” etc. and each category must be associated with at least one profile. A category may also refer to a more general data category by means of a broader concept generic attribute (the "is_a" relation, e.g. "common noun" is a "noun"). Although the registry is essentially a flat list of categories, the categories can be grouped per profile or represented as a tree structure based on the "is_a" references.

4.3 Overview of the DCR model

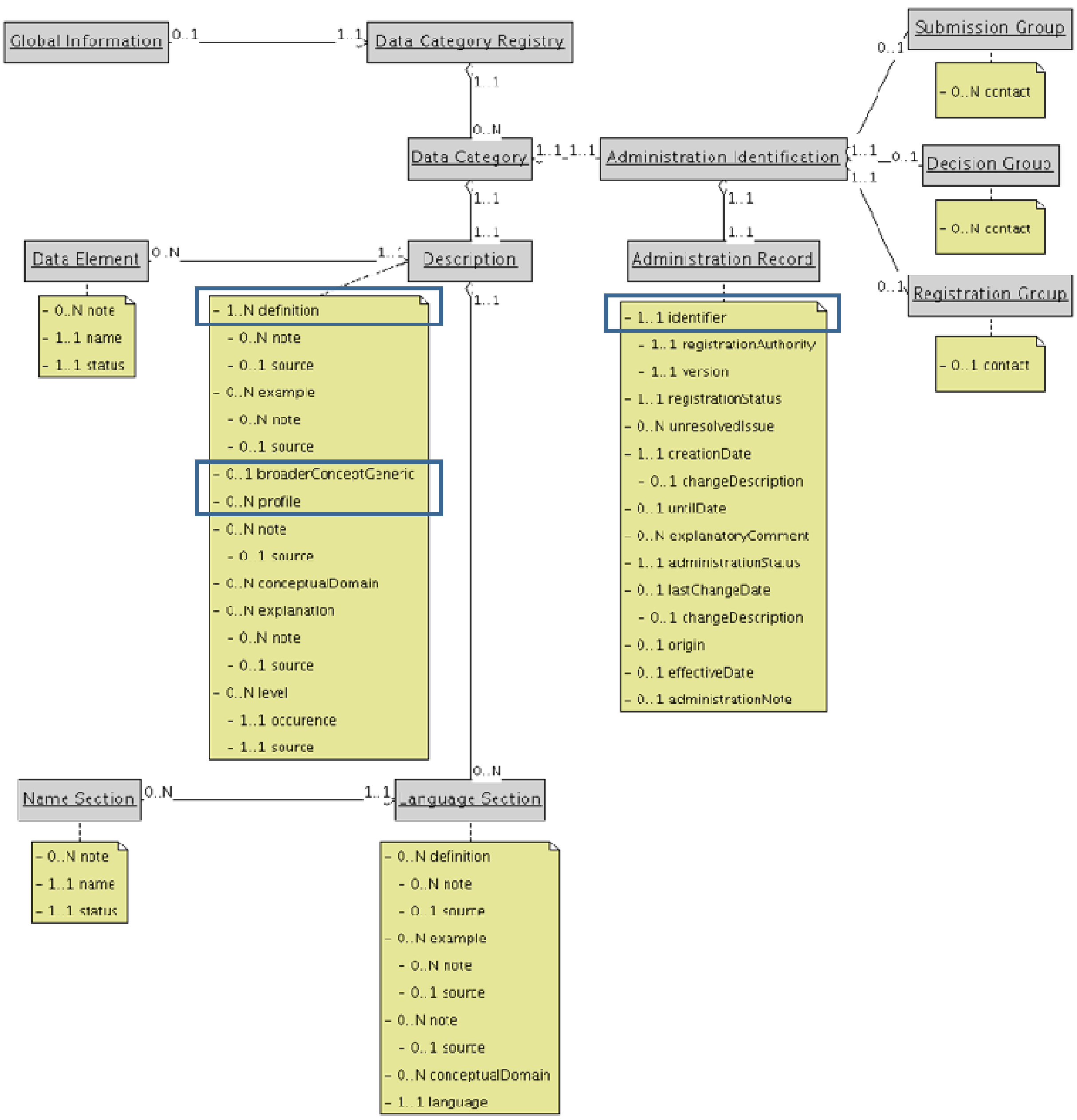

Figure 3: the DCR model. Marked in blue the information currently accessed and cached by ELAN.

Figure 3 shows an overview of the structure of the DCR. On the level of the Data Category a distinction can be seen between administrative information (Administration Identification) and the linguistically more relevant information (Description).

5. Implementation of DCR connectivity in ELAN

5.1 DCR Connector

Building on a DCR Connector that already had been developed at the MPI for Psycholinguistics for other tools, like Lexus [1], ELAN has been extended with a module to contact the DCR service and retrieve information by calling one of the methods defined by the API. Three methods of the API are currently used by ELAN:

- **getProfiles()** - to present a list of available profiles. The user can select the profile of interest, after which the method
- **getDataCategories(profile, status)** - is called and a list of returned data categories belonging to that profile is presented (if no status is provided only "standard" data categories are returned)
- **getDataCategory(id)** - if necessary all information of a category can be fetched, but typically a user will at this point create a reference from an annotation or a tier to the concept represented by that category

What in the end will be stored is the unique ID of a category together with a human readable short description, for reasons of convenience. What the preferred, unified way of referring to a category ID will be, is still to be established.

5.2 A local Data Category Selection

The API methods mentioned above are also used to cache an XML representation of (a part of) the DCR for local, offline use. There are 2 major reasons for doing so:

- ELAN is not a web application but a local tool. Once downloaded and installed on a computer, there no longer needs to be an internet connection to be able to create transcriptions. In fact ELAN is very often used in circumstances where there is no possibility to access internet at all. Therefore it is desirable, if not necessary, to be able to create references from elements (annotations, tiers) to a concept in the DCR from a cached version of (the relevant parts of) the DCR.

- it would be very time consuming, ineffective and annoying for the user to go through these steps (connect to the DCR, select profile, select category) every time she/he wishes to create a reference from an element in the transcription (e.g. annotation, tier) to a registered category.

This local DCR or DCS (Data Category Selection) is used in the transcription process to instantly present a short pick list of categories. This personal profile can consist of references to the unique ID's of (simple) data categories in the registry, but can eventually also contain complex user defined categories. These profiles could then be used and re-used for multiple transcriptions and could be shared between collaborators or even with other research teams.

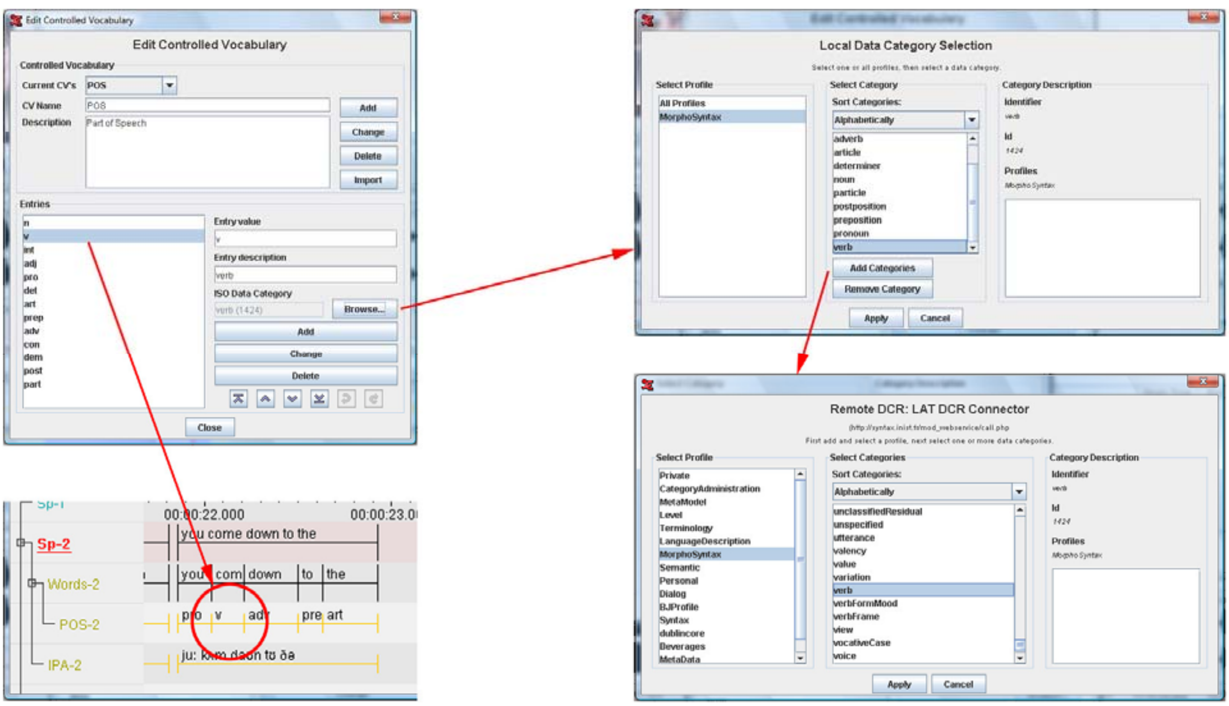

Figure 4: creating a reference from an entry in a Controlled Vocabulary to a data category

5.3 Extension of Controlled Vocabularies

Extending the existing Controlled Vocabularies facility, by allowing the user to associate entries in a CV to one or more categories from the DCR or a personal selection, seemed to be an obvious and efficient approach. In doing so the vocabulary is now capable of mimicking a data category selection.

Moreover, this mechanism is very well suited to add category references to existing annotations in a batch; once the references have been added to the vocabulary annotations on tiers referring to that vocabulary can be updated accordingly.

5.4 Extension of tiers and annotations

As mentioned before constraints on a tier are collected in a linguistic type object. This object has now been extended with a reference to a data category making it possible to designate a tier as e.g. a Part of Speech tier. Although this is at the moment not much more than a kind of label, it still improves interoperability (because there is no other way to indisputably ascertain what kind of tier it is).

In the future it might be possible to use this reference to e.g. the “part of speech” category to create a Controlled Vocabulary based on the value range of that category for the language of choice.

On the annotation level it is now possible to create a reference from individual annotations to any data category in the registry.

5.5 Usage example

The lower left corner of Figure 4 shows a part of an ELAN transcription, with four inter-depending tiers. A phrase has been annotated on a tier named “Sp-2” with as contents the text “you come down to the”. On the depending tier “Words-2” annotations for the individual words of the phrase level have been created, constituting a subdivision or decomposition. The words are individually time-aligned within the boundaries of the parent annotation. On the next depending level “POS-2”, part of speech tags have been added in a one-to-one relation to the words. In this case the POS labels come from a Controlled Vocabulary that has been created by the user. The entries of the vocabulary consist of an abbreviated label or code and a more verbose, explanatory description. Here the labels and descriptions appear to be language specific. E.g. one of the entries has the label “v” and description “verb”; a Dutch user might choose “ww” and “werkwoord” for annotating the same concept. By adding a reference to (the ID of) a category with identifier “verb”, in the MorhoSyntax profile of the DCR, interoperability of resources becomes independent of the particular coding scheme applied.

6. Conclusion

ELAN has taken the first steps in allowing its users to add references from transcription elements to concepts in the central ISO data category registry.

Adding these references requires some extra efforts from the side of the researcher, even though the referencing process has been implemented as efficient as possible. But the benefits of these efforts are improved interoperability and standardization. The equation of costs and benefits may eventually lead to a positive "return on investment".

7. References

Kemps-Snijders, M., Ducret, J., Romary, L., Wittenburg, P. (2006). An API for accessing the Data Category Registry. In *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation.*

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H. (2006). ELAN: a Professional Framework for Multimodality Research. In *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation.*

Rumble, J., Carroll, B., Hodge, G., Bartolo, L. (2005). Developing and Using Standards for Data and Information in Science and Technology.

[1] http://www.mpi.nl/lexus
[2] http://gate.ac.uk
[3] http://www.lat-mpi.eu/tools/tools/elan
[4] http://www.tc37sc4.org
[5] http://syntax.inist.fr
[6] http://www.isocat.org