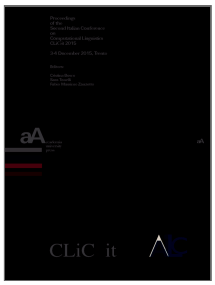

Cristina Bosco, Sara Tonelli and Fabio Massimo Zanzotto (dir.)

**Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015**
**3-4 December 2015, Trento**

Accademia University Press

# *Bolzano/Bozen Corpus*: Coding Information about the Speaker in IMDI Metadata Structure

**Marco Angster**

DOI: 10.4000/books.aaccademia.1284
Publisher: Accademia University Press
Place of publication: Torino
Year of publication: 2015
Published on OpenEdition Books: November 11, 2016
Series: Collana dell'Associazione Italiana di Linguistica Computazionale
Electronic EAN: 9788899200008

# *Bolzano/Bozen Corpus*: Coding Information about the Speaker in IMDI Metadata Structure

**Marco Angster**
Centro di Competenza Lingue
Libera Università di Bolzano
marco.angster@unibz.it

## Abstract

**English.** The paper introduces a new collection of spoken data (the *Bolzano/Bozen Corpus*) available through The Language Archive of Max Planck Institute of Nijmegen. It shows an example of the issues encountered in accommodating information of an existent corpus into IMDI metadata structure. Finally, it provides preliminary reflections on CMDI: a component-based metadata format.

**Italiano.** *Questo contributo presenta una nuova raccolta di dati di parlato (il* Bolzano/Bozen Corpus*) che è ora disponibile per la consultazione tramite il Language Archive del Max Planck Institute di Nimega. Vi si mostra un esempio dei problemi che si possono incontrare nell'inserimento all'interno della struttura di metadati IMDI delle informazioni relative a un corpus già esistente. Infine, vi si presentano alcune considerazioni preliminari riguardanti il formato di metadatazione CMDI, basato su componenti.*

## 1 Introduction

Once a Language Resource (LR) exists it should be used, and this entails several problems. First of all it must be available to the public – which may be the academic community, but also industry or institutions – and, given that producing a LR is an expensive task, it would be ideal that a LR could be exploited beyond the originally intended public. The re-usability of a LR is possible provided that it is conceived following shared standards for formats, tagging and metadata.

In this paper I focus on metadata structures, in particular I introduce a collection of spoken data (the *Bolzano/Bozen Corpus*) and I show the problems encountered in fitting the information available about the speakers sampled in the data in IMDI metadata structure.

The paper aims at providing an example of how flexible are the considered metadata structures in accommodating information of existent collections of data and in adapting to the needs of the researcher in sociolinguistics.

## 2 *Bolzano/Bozen Corpus*

The *Bolzano/Bozen Corpus* (BBC) collects and organises the language data produced during the years by the researchers of the Competence Centre for Language Studies. The common thread of the BBC is constituted by two main elements: the focus on the speech community in Alto Adige/South Tyrol, the trilingual province in Northern Italy of which Bolzano is the administrative centre; the interest on language variation, both in the social environment and in the educational context.

As a language resource the BBC is mainly destined to scholars interested in sociolinguistics and in the issue of multilingualism. Given that it collects different language varieties of the Romance and the German domain, the corpus has the function of providing original documentation for the local spoken language.

In order to give a better accessibility to the data, the corpus is made available to the public through The Language Archive (TLA), a collection of language resources hosted by the Max Planck Institute of Nijmegen (Nederlands).[1] All projects hosted by TLA must adopt a common metadata scheme on which all the structure of the database is built. The standard adopted by TLA used to be IMDI.[2]

[1] Homepage: https://tla.mpi.nl/
Corpora: https://corpus1.mpi.nl/ds/asv/?1

[2] TLA has recently made available to the users also the new, CLARIN supported CMDI metadata format. See below

The projects included in the BBC were obviously already supplied with rich information which had to fit into the metadata structure available.

## 3 IMDI and <Actors>

IMDI (ISLE/EAGLES Metadata Initiative) is a standard for metadata developed in the late '90s in the realm of standardisation initiatives ISLE (International Standard for Language Engineering) and EAGLES (Expert Advisory Group on Language Engineering Standards) – see Wittenburg et al. (2000). It provides a very rich structure in which information about a corpus, a session (i.e. a subdivision in a corpus, for example an interview), the relevant media files (the recording of an interview) and written resources (a transcription) are included. The session is the most complex substructure, because it may include a wealth of information about the interview itself: its location, its content (genre, communication context, type of task performed, languages used etc.) and its actors (interviewed, interviewee, but also transcriber, etc.).

Since BBC is a collection of data issued from sociolinguistically oriented projects, it appears clear that information about the speaker is of crucial importance and it is a fundamental concern to fit as much information about the speaker as possible in a metadata structure.

As already mentioned, part of metadata related to a session is devoted to the coding of information about people involved in the interview and in the production of the relevant resources. In this part of metadata structure the available tokens of information about a speaker involved in an interview or a language task are to be found. Some classical social variables are available: <Age>, <Sex>, <Education>, <Ethnic group>. Other useful pieces of information may be coded: <Role> ("The functional role of the person participating in the session" (IMDI, 2003); e.g. interviewer, speaker/signer, annotator, etc.), <Language> ("The language the person participating in the session is familiar with" (IMDI, 2003); more than one language may be added). A further element, <Family Social Role>, is available for coding "[t]he social or family role of the person participating in the session" and may be used "[f]or instance when interviewing part of a family group" where it can "specify the mutual relations within the group" (IMDI, 2003).

It is worth noting about the element <Language> that it is not intended to specify the language used in the session, for which another element is provided at an upper level under the node <Session> of the metadata structure. In this sense <Language> may be considered a good correspondent to the sociolinguistic concept of linguistic repertoire (Gumperz, 1964).

## 4 Speakers in Komma and Kontatto projects

I turn now back to BBC to show what information available about speakers involved in two different projects may be included in the structure sketched above.

The projects that I take into account are both focussed on South Tyrol, but with quite different perspectives, types of tasks accomplished and homogeneity of speakers involved. KOMMA (SprachKOMpetenzen von MAturandinnen und Maturanden) consists in the analysis of written and oral productions of high school graduands of the German schools of South Tyrol. It aims at studying the competence of the German standard language of young adults in mono- and multilingual settings in order to analyse linguistic phenomena, to find traces of multilingual competence or of a specific sociolinguistic background. At present the data available via TLA involve 41 students, all of German mother tongue: interviews on the language biography of the students and the renarration of a sequence of a Charlie Chaplin film (The Circus) are currently available.

More than a half of the students are female, most of them are 19 years old at the time of the interview. The picture is thus quite homogeneous, while the only variable which differentiates sets of students is the geographic area of the school they attended. This variable is coded as the location where the interaction takes place (<Location>). All students except two have both parents of German mother tongue, but this particular may not be coded in the metadata structure, unless we explicit it in the field <Description>. This is not an excellent solution, but a useful workaround to put a token of information which would be otherwise lost.

The second project considered here is Kontatto (Italiano-tedesco: aree storiche di contatto in Sudtirolo e in Trentino). The aim of the project is

section 5.

to document the present day Italian-German contacts in Bassa Atesina (the area south of Bolzano). The area is highly interesting for sociolinguistics and contact linguistics because there the interaction between German and Romance dialectal varieties dates back to a more remote time than in the rest of South Tyrol. A multilingual and multidialectal corpus of map tasks ((Anderson et al, 1991)) has been created to tackle the objective of documenting the linguistic productions of the speakers in the area.

The speakers involved in Kontatto are less homogeneous: they differ for age, occupation, own linguistic repertoire and linguistic background (parents' mother tongue, variety spoken where they live), place of origin of the parents, place of residence (as opposed to <Location>). This wealth of data – with the exception of the variables already mentioned above for KOMMA – would all be included in a <Description> field if one desires to keep this information available to the user interested in correctly interpreting the relevant data.

As for the case of KOMMA this could be a workaround, but a much more expensive one, from the point of view of future information retrieval. A metadata element is, let's say, a box where information is stored, but it is a box with an own particular tag, which indicates what is in. In addition this tag gives sense to the content and makes possible and easier to find the content itself among all information available. Putting information in a <Description> field corresponds to give up the possibility to exploit its classifying potential at a later time, thus making the information almost unusable.

## 5 CMDI: a very customisable, but closed structure

The limits of IMDI as a metadata structure are nonetheless well-known as we can read in the User Guide of the CLARIN-D infrastructure (Váradi et al, 2008):

> "Most existing metadata schemas for language resources seemed to be too superficial (e.g. OLAC) or too much tailored towards specific research communities or use cases (e.g. IMDI)." (CLARIN-D User Guide, 2012)

This words express the need of a new, more comprehensive standard for metadata description which could give to the researchers the possibility to tailor metadata profiles on the needs of their sub-disciplines. The new standard should display the following crucial features:

> 1. allow users to define their own components resulting in tailored profiles,
> 2. the components need to make use of categories the definitions of which are registered in ISOcat (see the section called "ISOcat, a Data Category Registry"), and
> 3. semantic interoperability and interpretability [must be] guaranteed by fine-grained semantics.
>
> (CLARIN-D User Guide, 2012)

At present CLARIN-D supports a new standard for metadata: CMDI. It is more flexible in that it allows the researcher to create own components rejecting profiles (for example <Session> or <Actor(s)>) which may be too restrictive or too fine-grained for their specific needs and modifying existing ones by adding or removing elements or by creating brand new profiles.

It is difficult for me to judge how open is CMDI for creating new profiles and how much flexible it is. In fact the possibility of creating new components and profiles is restricted to the accredited users of CLARIN centres.

In any case I try to imagine how should for instance a new CMDI-compliant component be structured in order to hold all information needed to give a complete description of a student of the KOMMA project. As shown above, the main problem is the impossibility to include information about parents' mother tongue. The solution of this lack would be to attribute to an actor involved in an interview a relation to another person – described as father or mother using the field <Family Social Role> – which is nonetheless not present in the interaction. Another possibility would be to code under the <Language> node one or more <Family Social Role> items pointing at the people with whom the relevant actor has a language in common. However solved, the problem apparently may be overcome.

It is worth noting that CMDI components are still based on the same elements on which IMDI is based. More precisely CMDI elements must point to a trusted data category registry (DCR), among

which ISOcat used to be one of the most used in IMDI structure.[3] In Kontatto, as we have seen, speaker profiles are very complex, but a wealth of information is available to the researcher. To characterise some of the interactions sampled in the project it may be useful to explicit both the "mutual relations within the group" as can be done through the field <Family Social Role> and the social background of the same speaker, for example its occupation, beyond the other social features he or she has. If an actor is the father of another actor, this should be independent from the fact that he is a boss, a doctor, a mayor, a teacher or a shaman/priest – just to cite some of the values of the open vocabulary category <Family Social Role> that are nonetheless suggested in IMDI Guidelines.

This fact highlights two different kinds of problems. The first one is a limit of IMDI: in its structure only one value for <Family Social Role> was allowed leading to the odd conclusion that one cannot be at the same time a father and a doctor. The second problem is more critical and significantly it is inherited by CMDI: <Family Social Role> is a category which is useful only to provide an explanation of the consequences for the interaction of the fact that a boss rather than a shaman/priest or a brother interacts with another actor. The category is instead simply unsatisfactory to accommodate background information, maybe irrelevant for the interaction but crucial to evaluate speaker's choices, such as what is the occupation of an actor, feature which contributes to the definition of the classic sociolinguistic variable of social class (Ash, 2003). However the unsatisfactory category <Family Social Role> appears to have no better alternative in ISOcat DCR, which is quite disappointing, because if I want to create my brand new <Actor> profile within CMDI I need to point to some existent data category and uses which contradict the meaning of a category are rightly deprecated.

As said, adding new data categories implies adding them to a Data Category Registry (DCR). Max Planck Institute for Psycholinguistics ceased in December 2014 to be the Registration Authority for ISOcat DCR. Now the new DCR for CMDI is CCR (CLARIN Concept Registry) which is nonetheless closed to changes. To add or change categories in the CCR the national CCR coordinators must be contacted, because only they are able to input new concepts and edit already existent ones.[4] This means that, in order to include a reasonable field <Occupation> instead of <Family Social Role> I have to operate outside CMDI and propose a new category to CCR national coordinators.

## 6 Conclusion

In this paper, I have shown an example of the difficulty of using a metadata structure to accommodate information on speaker's linguistic background. I have taken into account the case of *Bolzano Bozen Corpus* and two sociolinguistically oriented projects (KOMMA, Kontatto) hosted on The Language Archive.

IMDI, the former standard of TLA, is now an outdated tool and is too rigid to adapt to specific purposes. The new standard CMDI provides huge possibilities to the research community to define metadata formats tailored on specific needs. However CMDI does not provide until now satisfactory profiles and components for sociolinguistic studies, especially as far as background information about the speaker is concerned. Furthermore, direct contribution to CMDI components is restricted to CLARIN centres and in some crucial cases even categories available in CMDI are unsatisfactory and must be proposed to the relevant (and closed) DCR. The case I have proposed shows on the one hand the possibilities of CMDI. However, on the other hand, the difficulty to contribute to CMDI profiles and components from outside CLARIN may lead to the uncomfortable condition of having huge amounts of data with unsatisfactory metadata, which have low possibilities to be re-used, failing one of the main objectives of a standardisation initiative.

## Acknowledgments

I thank the project leaders of KOMMA (Rita Franceschini) and Kontatto (Silvia Dal Negro) and their collaborators for their support during the elaboration of the data which have been loaded on TLA platform. I also thank Roberto Cappuccio for technical support on many occasions. Finally I thank three anonymous reviewers for their useful comments.

[3] The list of data categories of ISOcat is available for consultation at `http://www.isocat.org/`.

[4] I thank an anonymous reviewer for pointing me out this possibility.

## References


Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry Thompson, and Regina Weinert. 1991. The Hcrc Map Task Corpus. *Language and Speech*, 34(4):351-366.

Sharon Ash. 2003. Social Class. In: J. K. Chambers, Peter Trudgill and Natalie Schilling-Estes. *The Handbook of Language Variation and Change*. Malden/Oxford: Blackwell Publishing. 402–422.

CLARIN-D User Guide. 2012. Version: 1.0.1. `http://media.dwds.de/clarin/userguide/userguide-1.0.1.pdf`.

John J. Gumperz. 1964. Linguistic and social interaction in two communities. *American Anthropologist*, 66/(6/2): 137–53.

IMDI Metadata Elements for Session Descriptions. 2003. Version 3.0.4. MPI Nijmegen. `https://tla.mpi.nl/?attachment_id=4532`.

Tamás Váradi, Peter Wittenburg, Steven Krauwer, Martin Wynne and Kimmo Koskenniemi. 2008. CLARIN: Common language resources and technology infrastructure. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*. 1244-1248.

P. Wittenburg, D. Broeder and B. Sloman. 2000. EAGLES/ISLE: A Proposal for a Meta Description Standard for Language Resources, White Paper. LREC 2000 Workshop, Athens. `http://www.mpi.nl/ISLE/documents/papers/white_paper_11.pdf`.