The RELISH MDF and GOLD crosswalk

The RELISH Project (Rendering Endangered Lexicons Interoperable through Standards Harmonization)

- The RELISH project addresses a two-pronged problem:
  - (1) the lack of harmonization between digital standards for lexical information in Europe and America;
  - (2) the lack of interoperability among existing lexicons of endangered languages, in particular those created with the Shoebox/Toolbox lexicon building software.

The cooperation partners in the RELISH project are the University of Frankfurt (FRA), the Max Planck Institute for Psycholinguistics (MPI Nijmegen), and Eastern Michigan University, the host of the Linguist List (ILIT).

The project aims at harmonizing key European and American digital standards. Focusing on 8 lexicons of endangered languages, the project establishes a unified way of referencing lexicon structure and linguistic concepts and develops a procedure for migrating these heterogeneous lexicons to a standards-compliant format.

Test Lexicons

- As test lexicons we have chosen Wichita (Caddoan), Tuvan (Turkic), Chalkan (Turkic), Udi (North Caucasian) and Batsbi (North Caucasian) on the LEXUS side and Mocovi (Mataco-Guaicuru), Western Sisaala (Niger-Congo) and Fulfulde (Niger-Congo) on the LEGO side.
- On the one hand, these languages represent pairs of relatively closely related languages (Tuvan and Chalkan, Udi and Batsbi, Western Sisaala and Fulfilde) which enables us to test various historical and comparative search options.
- On the other hand, the test lexica belong to various language types and are spoken on different continents which gives us an opportunity to use them for verifying various cognitive and typological hypotheses and perspectives.
- Moreover, some of the test languages are spoken in the contiguous areas, or in close contacts with other languages of the same affiliation, which provides for searches for lexical loans or other traces of contacts between these languages.

Referencing linguistic concepts

- Concepts used in LEXUS dictionaries are referenced to the ISOcat data categories
- Concepts used in LEGO dictionaries are referenced to the GOLD ones
- However, originally these lexicons were created by using the MDF software, supplied together with the Shoebox/Toolbox program.
- By mapping on the ISOcat data categories, a harmonization of the used semantic categories is achieved.

GOLD in ISOcat

GOLD: General Ontology for Linguistic Descriptions

- The GOLD concepts have been imported into the ISOcat data category registry and are now available as ISOcat data categories
- However, the DCR data model can only represent some of the relations in the GOLD ontology
- Currently a companion registry - Relation Registry in the DCR - is under construction which can store all the relations

MDF in ISOcat

- What is MDF and its applications?
- MDF in the RELISH project

What is MDF (Multi-Dictionary Formatter)?

MDF is a text format developed by SIL with minimal mark-up:

- The content is organized in *fields*
- Each field consists of a *marker* and the *field content*
- It is a flexible database format (optional fields, repeated fields etc.)
- Some markers contain language property: v, n , r, e
- V – vernacular language, the documented variety
- N – national language, the state language in the country where the documented variety is spoken
- R – regional – a language of broader communication
- E – the language of scientific description and glossing, normally English, but also German, Russian, etc.
- Markers with language properties form a marker "family": de, dn, dr, dv (definition in the English, national, regional, and vernacular language)
- It is quite exhaustive for standard lexicography in field research on minority languages
- Is a *de-facto* standard, although Toolbox is officially not supported by SIL any more (now replaced by FIELD / FLEX)

MDF in ISOcat Registry

- The MDF categories were introduced into the ISOcat Registry as a set and made public
- Complex categories containing language specifications were introduced both as complex categories and as simple ones, e.g.:
  - /gn „gloss in the national language“: http://www.isocat.org/datcat/DC-3711;
  - /g „gloss“: http://www.isocat.org/datcat/DC-244;
  - /n „national language“ http://www.isocat.org/datcat/DC-3702
- The category of *status* was introduced to deal with this distinction. It is a sociolinguistic category, and the ISOcat category registry does not have such a semantic domain
- We can map the data categories of individual lexicons onto the MDF DCs in the ISO DCR

MDF and GOLD subsets in the ISOcat registry and their interrelations

- GOLD and „standard“ ISOcats do not contain a language property
- GOLD and „standard“ ISOcats are thought of more as concepts, not as their concrete realizations
- Categories may be present only in GOLD (major terms in the domain “morphosyntax”)
- Categories may be present only in MDF (major terms in the domain “metadata”)
- Parallel data categories, sometimes under different names: reference (MDF) citation (GOLD) (the relation „X is equal to Y“)
- Categories that have a relation “X is partOf Y”, “X is subclassOf Y”
- Further relations

Harmonization of Terminology in the RELISH project

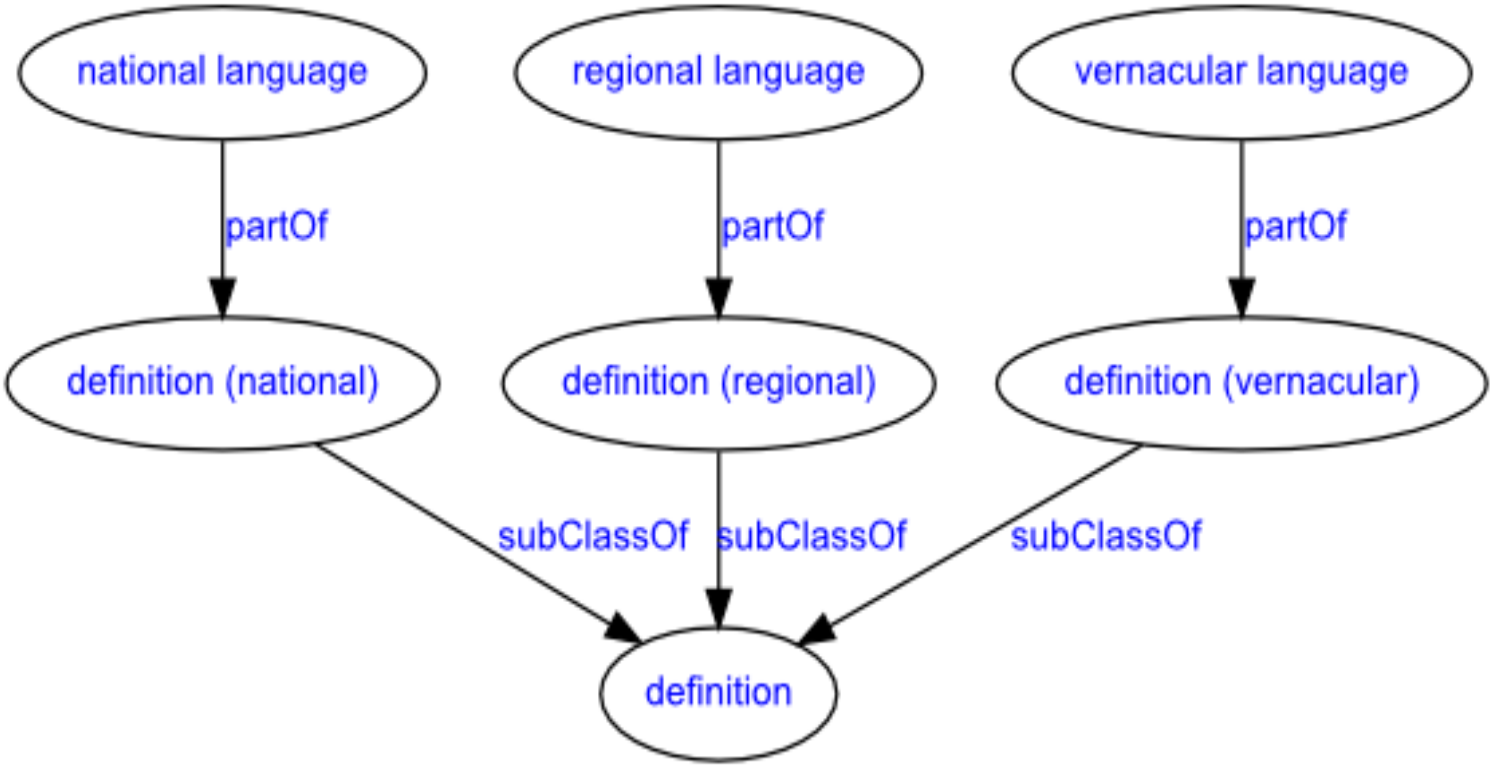
- By mapping on the ISOcat data categories, a harmonization of the used semantic categories should be achieved.
- A chart with interrelations between the MDF, GOLD and “standard” ISOcat data categories was created; types of relations between the categories were established to be implemented in the Relation Registry and in mapping the categories in the process of lexicon import into LEXUS as well as into RELISH-LIFT interchange formats.
- The Relation Registry also called RELcat being developed by the MPI will allow specification of (individual) relationships between data categories from the ISOcat DCR and possibly other concept registries.
- The chart describing the relationships between the MDF data categories, the GOLD data categories and other ISOcat data categories was imported into RELcat, which will allow tools to use these relationships in broadening and generalizing semantic searches.

Types of relations between the MDF, GOLD and “standard” ISOcat data categories

- sameAs: MDF-ISOcat “gloss” (DC-3711) is sameAs the ISOcat “gloss” http://www.isocat.org/datcat/DC-244
- subclassOf: MDF-ISOcat “gloss national” (DC-3711) is subclassOf the ISOcat “gloss” http://www.isocat.org/datcat/DC-244
- partOf: the ISOcat “gloss” http://www.isocat.org/datcat/DC-244 and the MDF-ISOcat „national language” http://www.isocat.org/datcat/DC-3702

  have the relation “part of” to the MDF-ISOcat “Gloss (national)” http://www.isocat.org/datcat/DC-3711

A Fragment of the RELcat



Problems to find a match

The names can be completely different, but the same phenomena are meant:

- The MDF „Bibliography“ (DC-3687) is sameAs the standard ISOcat “external reference” (DC-1975)
- The MDF “Borrowed word (loan)“ (DC-3688) is sameAs the standard ISOcat “source language” (DC-2494)

The names can (partially) coincide, but different phenomena are meant:

- The MDF ISO category “citation form (vernacular)”, defined as “a form for representing a lexeme” (DC-3716) should not be confused with the GOLD category “citation”.

The GOLD category is defined as “The action of citing or quoting any words or written passage from a publication that allows others to locate and identify the original source. Typical details include the title, author's name, the journal title (for articles), publication date and page numbers used in research.”

The MDF ISO category is almostSameAs the ISOcat "sort key" (DC-469), defined as “a form for sorting a printed dictionary”.

No generic term in the standard set of ISO categories

- In the ISOcats registry, there is no generic term “date”, but a number of more specific “date” terms:

- “origination date”:
- http://www.isocat.org/datcat/DC-166
- “check date”:
- http://www.isocat.org/datcat/DC-126
- “creation date”
- http://www.isocat.org/datcat/DC-2251
- “importation date”:
- http://www.isocat.org/datcat/DC-265
- “modification date”:
- http://www.isocat.org/datcat/DC-365
- etc.
- The MDF category “date” (DC-365) marks both the date of the data creation and the date of the data modification.

Polysemy of the categories

MDF-ISOcat “gloss” (DC-3707) is sameAs the ISOcat “gloss” http://www.isocat.org/datcat/DC-244

only in its first meaning: “In TEI: A phrase or word used to provide a gloss or definition for some other word or phrase. In 1951: Any editorial comment.”

Synonymy of the categories

- The ISOcat “etymology”
  http://www.isocat.org/datcat/DC-221

is sameAs to the ISOcat “etymological root”

http://www.isocat.org/datcat/DC-1987

- The ISOcat „comment“
  http://www.isocat.org/datcat/DC-1846

is sameAs to the ISOcat “note”
  http://www.isocat.org/datcat/DC-382