Challenges of metadata for LR interoperability and sustainability

Koenraad De Smedt

University of Bergen / CLARINO

CMDI interoperability workshop, June 4–5, 2013

Several catalogs and formats

Fixed XML schemata, RDF-based models, etc.

- META-SHARE
- VLO (CMDI, OLAC)
- TEI headers
- Europeana Data Model
- DSpace flat list
- other legacy catalogs and formats …

Costly to be compatible with everything

Need to integrate metadata in specific LR platforms

Case: META-SHARE to INESS

META-SHARE documented a large number of LRs including treebanks in CLARINO/INESS

INESS needs to maintain integrated metadata for its own purposes:

- keep track of its own resources
- present documentation to the user in a readable way
- select resources to be searched or browsed, based on metadata features
- present licensing information; user acceptance of licenses, etc.

Treebank selection with metadata

Languages: All · Abkhazian (1) · Ancient Greek (to 1453) (4) · Bulgarian (1) · Church Slavic (3) · Classical Armenian (2) · Danish (1) · English (3) · Estonian (1) · Faroese (1) · Finnish (2) · **Georgian** (8) · German (7) · Gothic (1) · **Hungarian** (6) · Icelandic (3) · Indonesian (1) · Latin (5) · Northern Sami (25) · Norwegian Bokmål (25) · Norwegian Nynorsk (5) · Old English (ca. 450-1100) (5) · Old French (842-ca. 1400) (5) · Old Norse (6) · Polish (4) · Portuguese (3) · Spanish (10) · Swedish (1) · Tamil (1) · Turkish (2) · Urdu (2) · Wolof (5)

Treebank Collections: **All** · BulTreeBank (0/1) · **GeoGram** (4) · **HunGram** (5) · ISWOC (0/23) · IcePaHC (0/1) · JRC Acquis (0/7) · Menotec (0/6) · NorGram (0/20) · PROIEL (0/15) · **ParGram** (2/10) · Sami-open (0/15) · Sami-restricted (0/7) · **Sofie** (1/8) · Test (0/6) · TiGer (0/3) · WolGram (0/1) · **XPar** (1/2)

Treebank Types: **All** · **lfg** (14/63) · dependency-proiel (0/38) · constituency (0/13) · dependency (0/6) · dependency-cg (0/32)

☐ Show only Parallel Treebanks

Reset

Chosen treebanks:

Name	*Collection*	*Type*	*Sentences*	*Words*	*Description*
Hungarian (hun)					
hun-20110919	HunGram	lfg	3 557	16 359	
hun-crow	HunGram	lfg	10	61	
hun-pargram (aligned)	HunGram, ParGram	lfg	15	72	
hun-rat	HunGram	lfg	13	111	
hun-scorpion	HunGram	lfg	12	96	
Georgian (kat)					
kat-mrs (aligned)	GeoGram	lfg	106	374	Georgian version of the mrs suite.
kat-pargram (aligned)	GeoGram, ParGram	lfg	16	57	
kat-pargram-2012 (aligned)		lfg	52	231	
kat-sofie (aligned)	Sofie, GeoGram	lfg	1 025	9 915	The Georgian part of the META-NC Sofie Parallel Treebank.
kat-xpar (aligned)	XPar, GeoGram	lfg	25	93	

CLARINO

Exporting from META-SHARE to INESS

Metadata from META-SHARE exported in XML format

- add to SVN repository
- edit
- import in INESS

Semi-automatic reformatting

- Insert <p> elements for better layout
- Extract text string from *description* field

Presenting metadata to treebank user

Resource: META-NORD Sofie Estonian Treebank

Description:

The Estonian part of the META-NORD Sofie Parallel Treebank.

This is a syntactically annotated parallel corpus based on the first chapters of the novel "Sofies verden" (Sophie's World) by Jostein Gaarder, published by Aschehoug forlag. The treebank consists of grammatical annotations of extracts from the Estonian translation of the novel, originally created as part of the Nordic Treebanking Network and now included in the extended META-NORD Sofie Parallel Treebank. The Estonian translation is published by Koolibri Publishing House. For more information, see the metadata description of the META-NORD Sofie Parallel Treebank.

ACCESS TO THE TREEBANK

The following terms hold for the use of the treebank:

The IPR holdership remains with Jostein Gaarder, who kindly permits INESS to distribute the "Sofie analyses" outside the project under the following terms of use:

a. The "Sofie analyses" can only be used for language technology research and development.

b. The users of the "Sofie analyses" are not allowed to redistribute or to publish the "Sofie analyses", only the knowledge and work that has been made on the basis of the "Sofie analyses",

c. The users of the "Sofie analyses" will ensure appropriate acknowledgement/references to the author of the original text, Jostein Gaarder, to Aschehoug Publishing House, Koolibri Publishing House and to the project INESS.

The alignments in the META-NORD Sofie Parallel Treebank are available under a CC-BY license (http://creativecommons.org/licenses/by/3.0/).

Attribution text:

"Alignments provided by the project INESS (www.iness.uib.no) in cooperation with META-NORD (http://www.meta-nord.eu/)."

Availability: available-restrictedUse

License: See «Description» for details

Restrictions of use: Treebank creators and IPR holders must be attributed, see «Description» for details.

Access medium:

Can be used for academic and commercial purposes.

By accepting the terms of the license you will be granted access to the resource.

Accept

Funding project (Nordic Treebank Network)

Project URL: ***http://w3.msi.vxu.se/~nivre/research/nt.html***

Project funded by: the Nordic Language Technology Program

Project start date: , end date:

Language: Estonian (***ISO code:*** et)

Alternative route

Treebanks added to INESS are documented using META-SHARE compatible templates.

The records may then be validated and uploaded again to a META-SHARE node using a simple http-based communication protocol developed at Tilde, with necessary webservices:

`MetaShare_getCollectionList(string)` – to return a catalog of all LRs

`MetaShare_getCollection(string)` – invoked with a parameter collectionID to return an XML description

Interoperability or migration?

Interoperability only for defined META-SHARE schemata

One-time migration to CLARIN CMDI seems like a more sustainable solution when new profiles are developed which are incompatible with META-SHARE schemata

Issues of granularity will arise as new needs become obvious

Documentation challenge: the case of complex resources

Definition: A resource is complex if

- it has several components
- it is multilingual
- several tools or methods have been applied in the process of creating it

Parallel treebanks

- a combination of monolingual treebanks
- an extra layer of annotation: alignment

Complex resources: preliminary solution

In META-SHARE:

- One metadata record for whole complex resource
- One metadata record for each (monolingual) component
- Language components listed both in the ‘sizeInfo’ part and the ‘relations’ part
- Linked with ‘relation’ feature:
  - ‘hasPart’ to children
  - ‘partOf’ relation to mother
  - ‘alignedWith’ relation to sisters

Parallel treebanks: better solution

- One metadata record for whole complex resource with subrecords for each component
- For each (monolingual) component, individual descriptions of a variable number of layers of annotation
- Alignment as a special annotation layer (standoff or not)
- Some description of evaluation (correctness/coverage) of the analyses

Conclusion: heterogeneous landscape

On the one hand, many existing schemata. On the other hand, not the ideal schemata.

Several co-existing frameworks. Need to define boundaries and possible cooperation.

Work out modest but concrete goals for interoperability

Conclusion: towards best practice

- Establish metadata requirements based on real documentation needs
- Share conversion tools to (and from) CMDI, with easy interfaces or web services
- Gradual ingest of data and metadata from users at deposit time
- Support CMDI compatibility in repository systems (e.g. DSpace)
- Coordinate creation of CMDI profiles and components, limit proliferation
- National coordinators for metadata and data categories