

SILT

Sustainable Interoperability for Language Technology



Fostering Language Resources Network

Grant Agreement No. ECP-2007-LANG-617001

Data Categories and their Semantics Summary

*Jan Odijk, Dan Flickinger, Jerry Hobbs,
Nancy Ide, Monica Monachini, Martha
Palmer, Anna Rumshisky, Nianwen Xue,
November 1-2, 2009
SILT-FLaReNet, Brandeis*



Interoperability & DCs

1. Goal: Achieve interoperability between data and tools
2. DCs and DC Registry are a means to achieve this
3. A subset of fine-grained DCs are put in a standardized set to serve a pivot (pivots)
4. Existing DCs and DC combinations are mapped onto and/or distinguished from DCs in the pivot
5. In this way the interpretation of the DCs in a resource is expressed in terms of DCs from the pivot set

Fundamental Questions

1. Is a DCR necessary / the best way to proceed / are there alternatives?
 - We are aware of no alternatives at this point
 - We are aware of the risks of going for a full 'interlingual' approach
 - But the only way to find out is pursue this approach and see how far we can get
 - A less ambitious approach with (multiple sub-interlinguas) interlingual DC subsets (e.g. per language, or small group of languages/per profile) is perhaps more realistic and still almost equally useful and can be used as a *fall back*



Fundamental Questions

2. How do we know that the interpretation of a DC is indeed the one that is intended?
 - No guarantees can be given but
 - risks can be minimized
 - Falsification tests can be designed (e.g. reversibility test)



Fundamental Questions

2. How do we know that the interpretation of a DC is indeed the one that is intended?
 - Required is
 - Discipline of the users
 - Active role of the TDGs
 - Some improvements in ISOCAT
 - to help users create better definitions
 - Illustration by means of real examples (e.g. annotated sentences/tokens in actual existing corpora) should perhaps be mandatory



Discussion Items

1. Definitions

- Precision (adding links to DCs)
 - Definitions too precise may be problematic
 - Maybe it is better if they are somewhat vague
 - Requiring precision will make the distinctions explicit (and these should be documented in ISOCAT)

Discussion Items

1. Definitions (cont)

- Theoretical differences (transitive) may impede standardization
 - The one that is actually used most often will become de facto standard
 - Differences must be well-documented
- ‘universal’ category set
 - Perhaps it is better to start with lg specific data categories (and add lg to a dc if it is also applicable to that lg) to avoid a English-dominated
 - Relation to versioning & PID



Discussion Items

1. Definitions (cont)
 - Add links to the sources that introduced the term. (already in source filed)
 - How does one arrive at a standardized subset? That should be clarified/ procedure should be set up

Discussion Items

2. Structures

- Some structures may be provided by LMF
 - A-v pairs,
 - a-v matrices
 - At least without recursion
 - Also with recursion?
- Explore whether these LMF structures can be used (for lexicons)
- Investigate extension to other LR types/tools
- Need feature cooccurrence constraints (e.g for PoSs) in different lgs
- For certain recursive structures, start with identifiers for them

Discussion Items

2. Structures (cont.)

- Other structures
 - Trees, lists, sets, unions, etc to be investigated
 - Can be implemented using recursive A-V matrices, but may not be optimal
 - Look at TEI Feature Structures standard!
 - Look at RDF/OWL as a candidate
 - Is the wish for a “container type” an alternative?

Discussion Items

3. Selections and Profiles

- Need to inquire / investigate creation of data category selections
 - For coherent subsets of DCs
 - E.g. Penn Treebank TagSet, STTS, CGN Tagset (Dutch), etc
- Inquire relation with profiles
- Investigate whether more grouping and selection mechanisms are required



Discussion Items

4. Interlingua is not enough
 - Even for real theoretical differences often good approximative mappings can be made
 - In many cases real differences can be avoided by mapping to duplicate entries that implement both options as an ambiguity



How to Proceed: Action Items

5. Create awareness of DCR/ISOCAT among
 - Funding agencies
 - Researchers from Europe/USA/Africa/Asia...
 - Who: FLaReNet/SILT
 - Who to: EC, national FAs: NSF, Endangered Igs initiatives, NIH, international FAs outside EU/US
 - When Near term
 - Resources needed

How to Proceed: Action Items

6. A: Explore possibilities of small scale projects showing the benefits of using DCRs/ISOCAT (only that will make it successful)
 - Either set up new projects (e.g Penn tagset+ CLAWS (<http://ucrel.lancs.ac.uk/claws7tags.html>) tagset, PPropBank/FrameNet, etc.) or German tagsets (STTS, Tiger, etc. Erhard Hinrichs), NE tagging, SPACE/TIME etc., morphosyntax, start with the less complex things. (cf. Parole/LC-STAR Mapping Pisa, ELRA ValidationC+ProductionC), also Asian etc resources



How to Proceed: Action Items

6. A: (cont.)
 - Who
 - Who to: LDC (Ann Bies) also Asia Josen, Tsujii, Takenobu Tokunaga, Chu Ren,...
 - When: Mid term
 - Resources needed



How to Proceed: Action Items

6. B: Explore possibilities of small scale projects showing the benefits of using DCRs/ISOCAT (only that will make it successful)
 - Adjust direction of existing projects (e.g. KYOTO towards ISOCAT)
 - Who: Pisa (Monica)
 - Who to: Piek Vossen+Pisa
 - When Near Term
 - Resources needed



How to Proceed: Action Items

6. C: Explore possibilities of small scale projects showing the benefits of using DCRs/ISOCAT (only that will make it successful)
 - Projects in the context of national CLARIN funding
 - Who Flarenet (JO)
 - Who to: CLARIN National Coordinators + CLARIN Board
 - When Near term
 - Resources needed



How to Proceed: Action Items

7. Stimulate Programme Committees to encourage project / article that work with/ on DCR/ISOCAT
 - Who: FLaReNet/SILT
 - Who to: LREC, ACL, COLING, DH, ..
 - When; Near Term
 - Resources needed



How to Proceed: Action Items

8. Stimulate others to acquire/make available funding to work in DCR
 - Also for European research groups currently not involved in/financed by (EU or national) CLARIN, e.g. Yorick Wilks, Lancaster, etc.
 - Who FLaReNet
 - Who to relevant non-CLARIN researchers
 - When Near Term
 - Resources needed



How to Proceed: Action Items

9. Clarification of a number of technical issues / request for extension of functionality
 - Grouping sets of DCs/data category selections/profiles
 - Structured objects / structured values / container type; use other standards for this? (TEI feature Structs, LMF feature cooccurrence constraints, etc.)
 - Relations among DCs
 - Support approximative mappings for overcoming real theoretical differences



How to Proceed: Action Items

9. Clarification of a number of technical issues (cont.)
 - Who JO
 - Who to: the ISOCAT designers/developers (MPI, S-EW,...)
 - When Near Term
 - Resources Needed

Recommendations (Linguist)

- Test ISOCAT against your resource / tool / project and report on virtues / problems/ omissions / questions!!
- Be hesitant in adding new DCs and definitions. It is not easy to do this correctly.
- If you add a new DC, provide 1 and only 1 definition!
- Clarify and illustrate your definition
- For DCs and their definitions: Use reliable and respected sources that are relevant to the point
- Stick to the [DCR Style Guidelines](#)
- Try to find DCs that have other names but the same meaning (spelling variants, synonyms)
- Ensure that all data categories used in your definition are defined as a DC and refer to their PIDs
- Take good care of polysemy shifts in meaning of one and the same expression. Make the conventions you adopt explicit!
- Use the “Conceptual Domain” facility much more than is done so far (it is currently underutilized)
- Select definitions of DCs preferably from one and the same common RELIABLE source to increase overall coherence of the DC set in ISOCAT, and certainly do this for DCs from the same “conceptual Domain” .

Recommendations (designers)

- Add [a system](#) that checks whether all terms used in a definition refer to a DC in ISOCAT and if not mark the DC as “incomplete”
- Add facilities that allow the user to deal with systematic cases of polysemy in a systematic manner
- Warn for DCs in the same “Conceptual Domain” that originate from different sources
- Provide means to group closely related concepts together (to find synonyms, and to prevent creation of new synonyms)
- Add relations between DCs (I understand this is already in the plan), esp. to implement an “is very similar to” relation
- Warn if all terms in the definition come from grammatical level (profile?) A when defining a term in grammatical level (profile?) B
- Add a formal link to the superordinate DC (can be done with relations)
- Add automatic notification to TD groups upon change
- Make the search better (search on **comparative** does not find **degree**)
- Make (the whole interface but esp.) text search in the web interface much faster!!!
 - **Google is faster on the whole WWW than the search facility is on ISOCAT!**
- Correct the source references (which are often wrong)

Recommendations (TD Groups)

- Provide more documentation on the definition of fields and their possible values in ISOCAT DC entries
 - The website has **DCR Style Guidelines and Web Interface — Overview** (and some presentations)
- Check sticking to the [DCR Style Guidelines](#) (currently often violated)
- Be organized for many additions/modifications/questions in 2010Q2!
- Respond quickly and adequately to newly added/modified DCs!
- Study and make recommendations on grouping of related concepts
- Study and make recommendations on how to deal with (systematic) metonymic shifts in meaning in linguistic data categories
- Be very critical before approving proposed DCs!
- Be very critical for “alternative definitions” in the same DC-entry
- Do not accept differences in definitions between different languages (adapt them or split the DC up into two DCs)