



Data Categories

and

ISOCAT

some remarks from a simple linguist

Jan Odijk

FLaReNet/CLARIN Standards Event Helsinki 30 sep 2009

Overview



- Data Categories in ISOCAT
- Interpretation of the Data Categories
- Problems with the interpretation
- Recommendations
 - For linguists
 - For the ISOCAT designers and developers
 - For the DCR Approval Committees

Data Categories in ISOCAT



- ISOCAT (www.isocat.org) provides framework for
 - Data Categories (DC)
 - Persistent identifier (PID) for each DC (in the form of a URL)
 - Definitions, examples, ... (in multiple languages)
 - Associated value domains
 - Status (candidate, private/public)
- Contributing to semantic interoperability
 - Link data category of an individual resource to the PID for the data category with same interpretation in ISOCAT (as being identical)

Interpretation of DCs



- My (Dutch) resource contains DC
 - *Overgankelijk*
- What is the corresponding DC in ISOCAT?
 - English translation is: *transitive*
- “transitive” occurs in ISOCAT
 - <http://www.isocat.org/datcat/DC-1405>
 - In profile *morphosyntax*
 - *Status: reg.:candidate, admin:private, scope:public*
- Does it mean the same as *overgankelijk*?

DC-1405 Definition



- English:
 - *A verb which takes a direct object; that is, a verb that expresses an action which directly affects another person or thing.*
 - Source:
www.southwestern.edu/~carlg/Latin_Web/glossary.html
 - Example:
 - *He **has** a horse*

DC-1405 Problems



- *Problem 1: there are two subdefinitions*
 - A verb which *takes* a direct object
 - a verb that *expresses an action* which *directly affects* another person or *thing*.
- *Problem 2:*
 - *take* in 1st subdefinition is not defined
 - Type reading
 - Obligatory reading: “must be combined with” ([Lexicon of Linguistics](#))
 - Potentiality reading: “can be combined with”
 - Token reading: “is combined with”

DC-1405 Problems



- *Problem 3: 2nd subdefinition*
 - *defines morphosyntactic DC in (apparently) purely semantic terms*
 - *This is suspect (why then is it a morphosyntactic DC?)*
- **Problem 4:**
 - *Most of the terms used in the second definition are no DC (hence have no definition) in ISOCAT*
 - *Person does occur in ISOCAT (in the sense of grammatical person, [DC-399](#)) but obviously not in the sense intended here*



- Problem 5:
 - *A plausible interpretation of 2nd definition:*
 - *→ 1st and 2nd subdefinition are different (→ internal contradiction)*
 - *→ example given is incorrect*
 - *He **has** a horse* (state v. action)
 - *He **baked** a cake* (effected .v affected thing)
 - *He **bought** a car* (nonaffected v. affected thing)
 - *He **considers** her smart* (no argument of *consider*)
 - *He **believed** her to be happy* (no argument of *believe*)
 - (the definition source distinguishes “action” from “state” in verbs)
 - *As an aside*
 - *the example sentence does not occur in its alleged source*
 - *And why is the alleged source a good source for an example? It describes the morphology of Arkian (a language not in Ethnologue!)*



- *Problem 6: (minor) should the DC be in the morphosyntax profile?*
 - *“Morphosyntax involves those elements and patterns of morphology (word formation) that reflect syntactical or grammatical functions, such as inflections and other paradigmatic elements whereby word forms change in adaptation to their usage.”*
 - *This is at least dubious*
- *Problem 7*
 - “transitive” is of course not “a verb” but “said of (a property of) a verb”



- *Definition of “direct object”* ([DC-1274](#))
 - *A direct object is a **grammatical relation** that exhibits a combination of certain independent syntactic properties, such as the following:*
 - *The usual grammatical characteristics of the patient of typically transitive verbs*
 - *A particular case marking*
 - *A particular clause position*
 - *The conditioning of an agreement affix on the verb*
 - *The capability of becoming the clause subject in passivization*
 - *The capability of reflexivization*
 - *Source:*
www.sil.org/linguistics/GlossaryOfLinguisticTerms/WhatIsADirectObject.htm
- *Example: A book is the direct object in the sentence They bought Anne a book*
 - *Source:* csli-publications.stanford.edu/LFG/4/lee/lfg99-lee.html



- Observation
 - This is a pretty reasonable definition (though not without problems)
 - It illustrates that pure definitions are often not possible for linguistic concepts
 - One can only place the DC relative to a superordinate category (*grammatical relation*), and
 - List some properties/criteria that can be used to argue for or against assigning this DC to a particular phenomenon
 - But the actual assignment is a hypothesis/theory, that usually can only be evaluated by taking into account the grammatical description of the whole range of relevant phenomena (i.e. in conjunction with other grammatical relations such as *subject*, *indirect object*, *head*, etc.)
 - **Recommendation** Therefore definitions of DCs from the same “conceptual Domain” should preferably originate from one and the same common source

DC-1274 Problems



- Problem 8:
 - *Grammatical Relation* is **not** a DC in ISOCAT
 - *Grammatical Function* **is** a DC in ISOCAT ([DC-1296](#))
 - With only a French definition (name: *fonction grammaticale*)
 - In profile Morphosyntax
 - And its “conceptual domain” data type “String”, but no list of possible grammatical functions
 - Name status: admitted name
 - *syntacticFunction* **is** a DC in ISOCAT ([DC-1507](#))
 - With only an English definition
 - In profiles Syntax and MorphoSyntax
 - Its “conceptual domain” datatype “String” and 3 possible values are distinguished
 - Subject ([DC-1391](#)), indirectobject ([DC-1310](#)), directobject ([DC-1274](#))
 - Twice: once for profile “Syntax”, once for profile “Morphosyntax”
 - Name status: nothing specified

DC-1274 Problems



- Problem 8: (cont.)
 - “Syntactic function” **is** a DC in ISOCAT ([DC-2244](#))
 - With an English and a French definition (French name: *fonction syntaxique*)
 - In profile Syntax
 - Conceptual Domain: data type string, no values listed
 - Name status: standardized name
 - Dependency **is** a DC in ISOCAT ([DC-2323](#)) (related, not identical)
 - Should DC-1296, DC-1507, DC-2244 be considered as synonymous/equivalent?
- Problem 9
 - How do I find alternative names for the same concept in ISOCAT?
 - How do I find closely related DCs in ISOCAT?
 - It currently requires a linear manual search... , even across different profiles!!
 - Is there a way of grouping closely related concepts together?

DC-1274 Problems



- Problem 10 “direct object”:
 - grammatical relation (relation-reading)
 - word or phrase that bears the grammatical relation “direct object” (phrase-reading)
 - Definition of *transitive* uses the phrase-reading
- Problem 11: Even stronger, an “alternative” definition is given:
 - *A noun, pronoun, or noun phrase whose referent receives the direct action of a verb*
 - Where the *phrase-reading* is given as the only one possible
 - Such metonymic shifts occur very often in natural language
 - E.g. morphosyntax = (1) a particular level of grammatical description, (2) study of this level of grammatical description, and (3) properties at this level of grammatical descriptions (e.g. “*the morphosyntax of this word is quite interesting*”, and idem for other levels of grammatical description)
 - **Recommendation**: Conventions on how to deal with such cases (preferably in a systematic manner) must be agreed upon and applied
 - **Recommendation** ISOCAT should provide means for supporting a systematic use of metonymic variants.

DC-1274 Problems



- Problem 12:
 - The example given for English is OK, but raises many questions that are not even mentioned
 - “the book” meets some criteria in the definition but not all
 - “Anne” meets some criteria in the definition but not all
 - The source given is curious
 - It is a study of adverbial case marking in Korean
 - In the LFG-framework, which does not accept the notion “direct object”(!), but uses OBJ and OBJ1 instead
 - » They bought OBJ/a book
 - » They bought OBJ/Anne OBJ1/a book
 - The example sentence does **not** occur in the article
 - The term “direct object” only occurs in the title of an article by others

DC-1274 Problems



- Problem 13 (Minor)
 - “direct object” is also in the profile “morphosyntax”
 - Even more dubious than “transitive”
- Problem 14
 - There are often also definitions and examples in other languages, e.g. French.
 - These have similar problems as the English definitions, and
- Problem 15
 - they usually differ from the English definitions (i.e the English and French definitions are not just translations of each other), e.g. in the case of “transitive”

Answer?



- Can I map my DC *overgankelijk* to [DC-1405](#) (E. *transitive*, Fr. *transitif*)?

I really do not know!

- So I will probably be forced to make a new DC in ISOCAT (which would be bad for semantic interoperability!)

Recommendations (Linguist)



- 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 metonymic 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 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 metonymy 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 DCR committees upon change
- 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!

Recommendations (DCR committees)



- 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)



Thanks for your Attention

Questions?

Do not go beyond this slide!



Are all definition terms in ISOCAT?



- Use a list of function words (stop list) for a language to identify the content words in a definition
- Each content word in a definition must
 - Be part of a span marked as *definiendum*, or
 - Be part of a span marked as *connector*, or
 - Be part of a span hyperlinked to an ISOCAT DC, or
 - Be part of a span marked as “intentionally not defined”
- Otherwise the definition is marked as incomplete

Are all definition terms in ISOCAT?



[BACK](#)

- Example:
- a `<definiendum>` transitive verb `</definiendum>`
- `<connector>` is defined as `</connector>`
- a `` verb `` that
- `` syntactically selects ``
- a `` noun phrase ``
- with the
- `` grammatical function ``
- `` direct object ``