Development of Linguistic Linked Open Data (LLOD) Resources for Collaborative Data-Intensive Research in the Language Sciences

María Blume (Pontificia Universidad Católica del Perú)
Antonio Pareja-Lora (Universidad Complutense de Madrid)
Ted Caldwell (GORGES)

An Introduction to the Data Transcription and Analysis (DTA) cybertool example of data management format and cross-linguistic datasets encoding: Standardized and non-standardized factors

LSA Summer Institute 2015
Chicago, July 26th, 2015
Structure of this talk

1. Introduction
2. Purpose
3. The Data Transcription and Analysis Tool (DTA)
   a. Main features
   b. Structure
   c. Main sections
   d. DTA labels.
4. Comparing the DTA with other systems.
5. Ontological formalization of the DTA categories.
   a. The OntoTag and the OntoLingAnnot models & ontologies
   b. The DTA metadata ontology.
   c. The DTA labels ontology
   d. Statistics
6. Summary and future work.

+ Acknowledgments
+ References.
1. Introduction

- The result of a joint work.
- We compared and linked two different language resources.
  - OntoTag/OntoLingAnnot’s ontologies.
    - Pareja-Lora (2012c) – The basic framework (OntoTag)
    - Pareja-Lora (2012a; 2013) – Pragmatics meta-model
    - Pareja-Lora (2012b) – Linguistic relations meta-model
  - The Data Transcription and Analysis Tool
    - Blume, Flynn & Lust (2012)
    - Blume & Lust (2012)
2. Purpose

We introduce

i. the metadata and the labels used within the DTA to annotate data on language acquisition

ii. the two ontologies we have built to represent the DTA metadata and the DTA labels.

iii. Our preliminary work on linking and/or anchoring these two ontologies to the (L)LOD cloud (Chiarcos et al., 2012):

   a) Links to ISOcat categories (Windhouwer and Wright, 2012)
   b) Equivalences with some GOLD elements (Farrar and Langendoen, 2010).
3. The Data Transcription and Analysis Tool (DTA)

+ URL: [http://webdta.clal.cornell.edu](http://webdta.clal.cornell.edu)

+ A primary research web application:
  - primarily for the study of monolingual and multilingual language acquisition,
  - with a powerful relational database:
    - handles both experimental and naturalistic data.

+ We use the term *DTA* to refer to the tool itself, its experiment bank component, and its associated corpora.
3a. The DTA – Main features (1)

I. Organizes and represents metadata and data so that:
   + Data can be analyzed subsequently in a standardized and theory-neutral way.
     - This ensures data comparability within a language and across languages.

II. Structures the data creation process from its initial stages and serves both as a research and as a teaching tool for training students on language data management and analysis:
   + Researchers can create project-specific labels (codings), allowing
     1. multiple types of analyses in their own data;
     2. linking data across projects.

III. Allows for long distance collaborative research.
3a. The DTA – Main features (2)
3b. The DTA Structure (1)

* In the DTA, the data are organized in projects.

* A project contains several subjects.
  * Each subject is a participant whose language is studied in the project.

* The subjects participate in different sessions.

* The sessions are organized in groups called datasets.

* Each dataset contains the resources, transcriptions, and codings for each session.
3b. The DTA Structure (2)

- Each **project** contains the following main sections of information:
  - Project Main Info
  - References
  - Subjects
  - Datasets
  - Coding
  - Queries.
3c. DTA main sections (1)

- Project Main Info
  - Main Info
  - Results
  - Summary and discussion

- References Info
  - Publications
  - Presentations
  - Related studies
  - References
Fig. 1: References
3c. DTA main sections (2)

- Subject information
  - Subject
  - Caretaker
Fig. 3: Subject screen

- ID, name & gender
- DOB
- Nationality, ethnicity, place of birth
- Any language or cognitive impairment.
- Human Subjects required documents filled
- Multilingualism questionnaire completed
- Subject’s contact information
- Comments.
- Language(s), dialect(s), and levels of language comprehension and production for each language.
- Caretaker information:
  - Relationship with the subject, occupation, name, contact information, languages, dialects, and levels of proficiency.
3c. DTA main sections – Dataset main info

- Topic, abstract, related WebDTA project/datasets
- Hypotheses
- Subjects (a summary description of subjects in the dataset)
- Methods
- Design
  - factors and variables
  - conditions
  - controls
  - specific hypotheses
  - statistical analyses
- Stimuli and procedures
- Scoring
- Results
- Conclusions
3c. DTA Main sections – Datasets

- Sessions
  - Transcriptions
  - Codings
  - Queries
Fig. 4: Session metadata

- Session ID
- Date
- Interviewer
- Assistants
- Session length
- Task
- Languages used
- Session location
- Subject’s
  - Current age
  - Number of siblings and position among siblings
  - Address
  - Length of residence
  - Education
  - Occupation
  - School
- Name and transcription identifier for the session participants besides the subject and interviewer.
- General activities
- Analyses performed on the data are included
3d. The DTA labels (1)

+ Labels in the DTA are called **codings**.
+ Codings and their related queries can be established at a global level or at an individual dataset level.
+ Codings are grouped in coding sets.
+ Simple coding sets were created to standardize and calibrate basic levels of linguistic analysis.
  + Utterance transcription,
  + Speech act,
  + Basic linguistic.
+ Project-specific coding sets can be created for all projects.
3d. The DTA labels (2)

- **Utterance transcription:**
  1. Includes text fields that give information to contextualize the utterances;
  2. Allows to add simple linguistic descriptions, translations and glosses of non-English utterances.

- **Speech act** relates to the pragmatic/discursive aspects of the data:
  1. Lists common speech acts and some others, common in child data;
  2. Asks about the spontaneous or responsive character of the utterance.

- **Basic linguistic** specifies:
  1. Whether the utterance is a sentence or not;
  2. Whether the sentence has an overt verb;
  3. Number of morphemes, words, and syllables of the utterance.
4. Comparing the DTA with other systems: DTA & CHILDES (1)

+ Both CHILDES and DTA focus on child language data.

+ They have common or similar labels in the codings that they adopt (about 24).

+ One main difference:
  1. The DTA provides the user with a structured interface for primary data entry and management,
  2. While CHILDES
     - lists possible metadata fields in its accompanying manual,
     - and provides no structure for the researcher.
4. Comparing the DTA with other systems: DTA & Language Archive (2)

- The DTA and the Language Archive (LA) share many of their fields:
  - Both aim at language archiving and metadata creation.

- However, they organize data differently:
  i. The LA organizes data in terms of sessions,
     - The project information is contained inside a session.
     - It has no dataset level.
  ii. Whereas the DTA organizes data in terms of projects that contain datasets which in turn contain sessions.

- The DTA can store more detailed data than the LA about:
  a. The child’s caretakers;
  b. The project and the dataset experimental design (19 add. fields);
  c. The division of references.
4. Comparing the DTA with other systems: DTA vs. Language Archive (2)

+ The LA can store more detailed data than the DTA on:
  + The different types of resources (i.e., “source”, “resource”, and “written resource”) with detailed information for type, format, encoding, access, and anonymity;
  + The type of communication context and genre of the interaction (30 add. fields), some of which would be relevant for the DTA.

+ There are some fields that the DTA has which are not child/experiment specific which the LA does not have:
  + the participant’s length of residence at the session location;
  + date of birth, nationality and place of birth;
  + levels of language or cognitive impairment, dialect, whether he/she is a native speaker of the language used in the session, and his/her levels of proficiency in the language, etc..
5. Ontological formalization of DTA categories – Introduction

+ Two different ontologies:
  - The DTA Metadata Ontology formalizes the components, concepts and objects referenced in the guide describing the DTA tool (the DTA metadata):
  - The DTA Labels Ontology includes a conceptualization of the codings (= labels) used to annotate DTA transcriptions linguistically.

+ Developed following the OntoTag & the OntoLingAnnot methodology & principles, and using their ontologies (Aguado de Cea et al., 2002; 2004; Pareja-Lora, 2012a).

+ URL: https://github.com/apareja/DTA_Ontologies.
5a. Background: OntoTag vs. OntoLingAnnot

**OntoTag:**
- Morphosyntactic, syntactic and semantic annotation of web pages.
- Linguistic annotation based on Semantic Web recommendations.
- OntoTag’s ontologies:
  - Ontological formalization of linguistic terminology.
  - Best practices for the development of annotation pipelines and annotation merge.

**OntoLingAnnot:**
- A generalization of OntoTag.
- Semantic Web + ISO/TC 37 compliant linguistic annotation.
- OntoLingAnnot’s ontologies:
  - Improvement of the semantic level.
  - Addition of the discourse and the pragmatic levels.
  - Focuses on the interoperability of annotations at any level.

- Pareja- Lora & Aguado de Cea (2010)
- Pareja-Lora (2012a)
- Pareja-Lora & Aguado de Cea (2010)
- Pareja-Lora (2012b; 2012c)
- Pareja-Lora (2014)
5a. Background: OntoTag’s & OntoLingAnnot’s linguistic description levels

<table>
<thead>
<tr>
<th>LEVEL</th>
<th>ANNOTATION SCOPE</th>
<th>UNITS</th>
</tr>
</thead>
<tbody>
<tr>
<td>OntoTag</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Morphology</td>
<td>Up to word structure and meaning (including morph-related annotation).</td>
<td>• Morph</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Word</td>
</tr>
<tr>
<td>Syntax</td>
<td>Up to sentential structure (including multiword token, phrasal and clausal structure).</td>
<td>• Token</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Phrase</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Clause</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Sentence</td>
</tr>
<tr>
<td>Semantics</td>
<td>Up to propositional structure and sentence meaning (including phrase and clause meaning), that is, propositional meaning.</td>
<td>• Sense</td>
</tr>
<tr>
<td>Discourse</td>
<td>Up to discourse structure (including coherence relation-based structure) and the supra-sentential and locutionary meaning of texts and speech (including anaphora resolution).</td>
<td>• Macroproposition</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Turn, etc.</td>
</tr>
<tr>
<td>Pragmatics</td>
<td>Up to the illocutionary and perlocutionary structure and meaning of texts and speech (including deictic resolution and other pragmatic relation annotation).</td>
<td>• Macroproposition</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Pragmateme, etc.</td>
</tr>
</tbody>
</table>
5a. Background: The OntoTag annotation model

Linguistic Knowledge & Terminology

Input text

OntoTag / OntoLingAnnot

Standardised Annotation Encoding

Standardised Annotation

Standardised linguistic annotations

Linguistic Ontologies

Subject Predicate Object

Triples

Languages

OWL RDF(S) XML

Semantic Web
5a. Background: OntoTag’s / OntoLingAnnot’s ontologies

**The result**
A networked and modularised set of ontologies

- Each ontology includes only a type of linguistic object (units, features, values, etc.)
- Each ontology module refers only to the objects of one particular linguistic level

**Terminological gaps filled**
- Key concepts and terms not defined in standardised sources
- Levels not standardised yet (e.g., pragmatics)
- Bridge concepts/terms between different theories

**Methodology**
METHONTOLOGY / NeOn Methodology

**Approach followed**
- Eclectic and comprehensive
- Not sticking to any particular linguistic theory

**Main criteria adopted**
- Taxonomise as much as possible
- Avoid redundancy – share attributes and values
- Consistent and coherent modularisation
- Link modules and ontologies by suitable imports, relations and axioms

**Standardised knowledge and terminological sources**

<table>
<thead>
<tr>
<th>Morpho-syntactic level</th>
<th>Syntactic level</th>
<th>Semantic level</th>
<th>Discourse level</th>
<th>Pragmatic level</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAF (ISO, 2008c)</td>
<td>SynAF (ISO, 2006)</td>
<td>SemAF-NE (ISO, 2010)</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
5a. OntoLingAnnot’s ontologies: an overview
5a. Pragmatic units and attributes in OntoLingAnnot
5a. The pragmatic values in the Linguistic Value Ontology (LVO) [INSTANCES]

<table>
<thead>
<tr>
<th>CONCEPTS</th>
<th>INSTANCES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common Pragmatic Value</td>
<td>{UNMARKED, MARKED}</td>
</tr>
<tr>
<td>Indirect Speech Value</td>
<td>BOOLEAN = {TRUE, FALSE}</td>
</tr>
<tr>
<td>Performative Type Value</td>
<td>{EXPLICIT, IMPLICIT}</td>
</tr>
<tr>
<td>Face Value</td>
<td>{SAVING, THREATENING, UNMARKED}</td>
</tr>
<tr>
<td>Mitigating Device Value</td>
<td>BOOLEAN = {TRUE, FALSE}</td>
</tr>
<tr>
<td>Addressing Type Value</td>
<td>{UNMARKED, OFF_RECORD, ON_RECORD, BALD_ON_RECORD}</td>
</tr>
<tr>
<td>Position Markedness Value</td>
<td>{UNMARKED, SUBINDEX, SUPERINDEX}</td>
</tr>
<tr>
<td>Format Markedness Value</td>
<td>{UNMARKED, UNDERLINED, ITALICS, BOLD, CAPITALISED, SMALL_CAPITALISED}</td>
</tr>
<tr>
<td>Size Markedness Value</td>
<td>{UNMARKED, BIGGER, SMALLER}</td>
</tr>
<tr>
<td>Style Markedness Value</td>
<td>{UNMARKED, SUBTITLE, TITLE}</td>
</tr>
<tr>
<td>Colour Markedness Value</td>
<td>{UNMARKED, MARKED}</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CONCEPTS</th>
<th>INSTANCES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Register Markedness Value</td>
<td>{UNMARKED, FAMILIAR, SLANG, FORMAL, JARGON, TECHNICAL, OFFICIAL}</td>
</tr>
<tr>
<td>Politeness Markedness Value</td>
<td>{UNMARKED, POLITE, UNPOLITE}</td>
</tr>
<tr>
<td>Abbreviational Markedness Value</td>
<td>{UNMARKED, EXPECTEDLY_ABBREVIATED, UNEXPECTEDLY_ABBREVIATED, EXPECTEDLY_NON_ABBREVIATED, UNEXPECTEDLY_NON_ABBREVIATED}</td>
</tr>
<tr>
<td>Appreciative Markedness Value</td>
<td>{UNMARKED, AUGMENTATIVE, DIMINUTIVE, PEJORATIVE}</td>
</tr>
<tr>
<td>Preferential Markedness Value</td>
<td>{UNMARKED, PREFERRED, DISPREFERRED}</td>
</tr>
<tr>
<td>Proximity Value</td>
<td>{CLOSE, INTERMEDIATE, DISTANT, UNMARKED}</td>
</tr>
</tbody>
</table>
5a. Linguistic Relationship Ontology (LRO) – Pragmatic relations

Interpersonal Coherence Relation (Participation Framework Relation)

Interpretation

Enablement

Exhortation

Antithesis

Evaluation

Background

Qualification

Concession

Support

[... (up to 6 subclasses)]

Romera (2004); Hovy & Maier (1996); Mann & Thompson (1988); Schiffrin (1987)

Interpretation

Coherence Relation

Pragmatic Realisation

Tropological Realisation

Pragmatic Coordination

Pragmatic Subordination

Pragmatic Constitution

Pragmatic Transposition

Mitigation

Hedging

Yule (1996)

SIL (2010)

Pragmatic Realisation

Pragmatic Subordination

Pragmatic Constitution

Pragmatic Transposition

Quality Hedging

Quantity Hedging

Relevance Hedging

Manner Hedging

Yule (1996)

SIL (2010)

Pragmatic Realisation

Tropological Realisation

Pragmatic Coordination

Pragmatic Subordination

Pragmatic Constitution

Pragmatic Transposition

Quality Hedging

Quantity Hedging

Relevance Hedging

Manner Hedging

SIL (2010)

Pragmatic Realisation

Tropological Realisation

Pragmatic Coordination

Pragmatic Subordination

Pragmatic Constitution

Pragmatic Transposition

Quality Hedging

Quantity Hedging

Relevance Hedging

Manner Hedging

SIL (2010)

Pragmatic Realisation

Tropological Realisation

Pragmatic Coordination

Pragmatic Subordination

Pragmatic Constitution

Pragmatic Transposition

Quality Hedging

Quantity Hedging

Relevance Hedging

Manner Hedging

SIL (2010)

Pragmatic Realisation

Tropological Realisation

Pragmatic Coordination

Pragmatic Subordination

Pragmatic Constitution

Pragmatic Transposition

Quality Hedging

Quantity Hedging

Relevance Hedging

Manner Hedging

SIL (2010)

Pragmatic Realisation

Tropological Realisation

Pragmatic Coordination

Pragmatic Subordination

Pragmatic Constitution

Pragmatic Transposition

Quality Hedging

Quantity Hedging

Relevance Hedging

Manner Hedging

SIL (2010)

Pragmatic Realisation

Tropological Realisation

Pragmatic Coordination

Pragmatic Subordination

Pragmatic Constitution

Pragmatic Transposition

Quality Hedging

Quantity Hedging

Relevance Hedging

Manner Hedging

SIL (2010)

Pragmatic Realisation

Tropological Realisation

Pragmatic Coordination

Pragmatic Subordination

Pragmatic Constitution

Pragmatic Transposition

Quality Hedging

Quantity Hedging

Relevance Hedging

Manner Hedging

SIL (2010)

Pragmatic Realisation

Tropological Realisation

Pragmatic Coordination

Pragmatic Subordination

Pragmatic Constitution

Pragmatic Transposition

Quality Hedging

Quantity Hedging

Relevance Hedging

Manner Hedging

SIL (2010)

Pragmatic Realisation

Tropological Realisation

Pragmatic Coordination

Pragmatic Subordination

Pragmatic Constitution

Pragmatic Transposition

Quality Hedging

Quantity Hedging

Relevance Hedging

Manner Hedging

SIL (2010)

Pragmatic Realisation

Tropological Realisation

Pragmatic Coordination

Pragmatic Subordination

Pragmatic Constitution

Pragmatic Transposition

Quality Hedging

Quantity Hedging

Relevance Hedging

Manner Hedging

SIL (2010)

Pragmatic Realisation

Tropological Realisation

Pragmatic Coordination

Pragmatic Subordination

Pragmatic Constitution

Pragmatic Transposition

Quality Hedging

Quantity Hedging

Relevance Hedging

Manner Hedging

SIL (2010)

Pragmatic Realisation

Tropological Realisation

Pragmatic Coordination

Pragmatic Subordination

Pragmatic Constitution

Pragmatic Transposition

Quality Hedging

Quantity Hedging

Relevance Hedging

Manner Hedging

SIL (2010)

Pragmatic Realisation

Tropological Realisation

Pragmatic Coordination

Pragmatic Subordination

Pragmatic Constitution

Pragmatic Transposition

Quality Hedging

Quantity Hedging

Relevance Hedging

Manner Hedging

SIL (2010)

Pragmatic Realisation

Tropological Realisation

Pragmatic Coordination

Pragmatic Subordination

Pragmatic Constitution

Pragmatic Transposition

Quality Hedging

Quantity Hedging

Relevance Hedging

Manner Hedging

SIL (2010)

Pragmatic Realisation

Tropological Realisation

Pragmatic Coordination

Pragmatic Subordination

Pragmatic Constitution

Pragmatic Transposition

Quality Hedging

Quantity Hedging

Relevance Hedging

Manner Hedging

SIL (2010)

Pragmatic Realisation

Tropological Realisation

Pragmatic Coordination

Pragmatic Subordination

Pragmatic Constitution

Pragmatic Transposition

Quality Hedging

Quantity Hedging

Relevance Hedging

Manner Hedging

SIL (2010)

Pragmatic Realisation

Tropological Realisation

Pragmatic Coordination

Pragmatic Subordination

Pragmatic Constitution

Pragmatic Transposition

Quality Hedging

Quantity Hedging

Relevance Hedging

Manner Hedging

SIL (2010)

Pragmatic Realisation

Tropological Realisation

Pragmatic Coordination

Pragmatic Subordination

Pragmatic Constitution

Pragmatic Transposition

Quality Hedging

Quantity Hedging

Relevance Hedging

Manner Hedging

SIL (2010)

Pragmatic Realisation

Tropological Realisation

Pragmatic Coordination

Pragmatic Subordination

Pragmatic Constitution

Pragmatic Transposition

Quality Hedging

Quantity Hedging

Relevance Hedging

Manner Hedging

SIL (2010)

Pragmatic Realisation

Tropological Realisation

Pragmatic Coordination

Pragmatic Subordination

Pragmatic Constitution

Pragmatic Transposition

Quality Hedging

Quantity Hedging

Relevance Hedging

Manner Hedging

SIL (2010)

Pragmatic Realisation

Tropological Realisation

Pragmatic Coordination

Pragmatic Subordination

Pragmatic Constitution

Pragmatic Transposition

Quality Hedging

Quantity Hedging

Relevance Hedging

Manner Hedging

SIL (2010)

Pragmatic Realisation

Tropological Realisation

Pragmatic Coordination

Pragmatic Subordination

Pragmatic Constitution

Pragmatic Transposition

Quality Hedging

Quantity Hedging

Relevance Hedging

Manner Hedging

SIL (2010)

Pragmatic Realisation

Tropological Realisation

Pragmatic Coordination

Pragmatic Subordination

Pragmatic Constitution

Pragmatic Transposition

Quality Hedging

Quantity Hedging

Relevance Hedging

Manner Hedging

SIL (2010)

Pragmatic Realisation

Tropological Realisation

Pragmatic Coordination

Pragmatic Subordination

Pragmatic Constitution

Pragmatic Transposition

Quality Hedging

Quantity Hedging

Relevance Hedging

Manner Hedging

SIL (2010)
5a. Some pragmatics-related statistics of OntoLingAnnot’s ontologies

<table>
<thead>
<tr>
<th>PRAGMATIC TERMS IN ONTOLINGANNOT’S ONTOLOGIES</th>
<th>UNITS (Luo)</th>
<th>ATTRIBUTES (LAO)</th>
<th>VALUES (LVO)</th>
<th>RELATIONS (LRO)</th>
<th>LEVEL, LAYERS, etc. (LLO)</th>
<th>TOTAL PRAGMATIC TERMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONCEPTS</td>
<td>192</td>
<td>10</td>
<td>27</td>
<td>86</td>
<td>24</td>
<td>339</td>
</tr>
<tr>
<td>INSTANCES</td>
<td>0</td>
<td>16</td>
<td>54</td>
<td>0</td>
<td>0</td>
<td>70</td>
</tr>
<tr>
<td>TOTAL TERMS</td>
<td>192</td>
<td>26</td>
<td>81</td>
<td>86</td>
<td>24</td>
<td>409</td>
</tr>
</tbody>
</table>

There are also several other ontological terms concerning pragmatics (attributes, SubclassOf, PartOf and ad hoc relations, rules and axioms) not mentioned here for the sake of brevity (Pareja-Lora, 2012a; 2014)
5a. Some discourse-related statistics of OntoLingAnnot’s ontologies

<table>
<thead>
<tr>
<th>DISCOURSE TERMS IN ONTOLINGANNOT’S ONTOLOGIES</th>
<th>UNITS (LUO)</th>
<th>ATTRIBUTES (LAO)</th>
<th>VALUES (LVO)</th>
<th>RELATIONS (LRO)</th>
<th>LEVEL, LAYERS, etc. (LLO)</th>
<th>TOTAL DISCOURSE TERMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONCEPTS</td>
<td>68</td>
<td>4</td>
<td>8</td>
<td>70</td>
<td>24</td>
<td>174</td>
</tr>
<tr>
<td>INSTANCES</td>
<td>0</td>
<td>4</td>
<td>7</td>
<td>36</td>
<td>0</td>
<td>47</td>
</tr>
<tr>
<td>TOTAL TERMS</td>
<td>68</td>
<td>8</td>
<td>15</td>
<td>106</td>
<td>24</td>
<td>221</td>
</tr>
</tbody>
</table>

There are also several other ontological terms concerning discourse(*attributes, SubclassOf, PartOf and ad hoc relations, rules and axioms*) not mentioned here for the sake of brevity (Pareja-Lora & Aguado de Cea, 2010).
## 5a. Discourse level of OntoLingAnnot’s ontologies – Linguistic Theories Coverage

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Coherence/cohesion relations</td>
<td>GLOBAL</td>
<td>GLOBAL</td>
<td>GLOBAL</td>
<td>GLOBAL</td>
<td>PARTIAL</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>GLOBAL</td>
</tr>
<tr>
<td>Syntactic features for coherence relation and DFU classification</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>YES</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>YES</td>
</tr>
<tr>
<td>Anaphoric reference and anaphora resolution</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>YES</td>
<td>-</td>
<td>-</td>
<td>YES</td>
</tr>
<tr>
<td>Encapsulation and prospection</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>YES</td>
<td>-</td>
<td>YES</td>
</tr>
<tr>
<td>Discourse constitution (macropropositions, superstructures, etc.)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>DISCOURSE GLOBAL VIEW</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>YES</td>
</tr>
</tbody>
</table>
## 5a. Discourse level of OntoLingAnnot’s ontologies – Standards Coverage

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>YES</td>
<td>MINIMALLY</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>DISCOURSEANNOT.GLOBALVIEW</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
</tbody>
</table>
5b. The DTA Metadata Ontology (1)

+ Classes:

  • The components of the forms included in the tool:
    o Info tabs (e.g., DatasetTab, SubjectsTab, ReferencesTab);
    o Information sections (e.g., DTADatasetSection, DTAQuerySection)
    o Menus (e.g., PresentationsMenu, BibliographyMenu).
  • A taxonomy of academic objects (presentations and publications, basically), as they are used/understood within the DTA;
  • A subclassification of the Person concept, according to the different roles that a person can play in a DTA project and/or session (e.g., Subject, Assistant, Interviewer, Researcher)
5b. The DTA Metadata Ontology (2)

Object Properties:

- Structure:
  - Hierarchically arranged, whenever possible:
    - Ready for deeper levels of inference (e.g., through inheritance).
- Contents:
  - Relationships between the classes included in the ontology:
    - E.g.: `hasPart` → `hasInvestigator` → `hasPrincipalInvestigator`.
  - A taxonomy of detailed interpersonal and family relationships, in order to formalize the possible relationships between subjects and caretakers in DTA sessions:
    - E.g.: `hasInterpersonalRelationshipWith` → `isRelativeOf` → `isDescendantOf` → `isChildOf` → `isDaughterOf`. 
5b. The DTA Metadata Ontology (3)

**Data Properties:**

- **Structure:**
  - Hierarchically arranged, whenever possible. This helps
    - avoid redundancy;
    - improve the class hierarchy (e.g. Reference subclasses).

- **Contents:** Attributes within DTA of the classes formalized:
  - **Of Person:**
    - Generic attributes (e.g., `hasGender`, `hasNationality`, `hasPersonFirstName`, `hasBirthDate`);
    - Subject attributes (e.g., `hasNumberOfSiblings`, `hasPositionAmongSiblings`, `hasCognitiveImpairment`);
    - Caretaker attributes (e.g., `hasEducation`, `hasOccupation`)
  - **Of academic objects** (e.g., `isDigital`, `hasAcademicObjectType`)
  - **Other top-level attributes**, which help to build the hierarchy (e.g., `hasDate`)
5b. The DTA Metadata Ontology (4)

+ Datatypes:
  
  - The values displayed in the menus of the application:
    - `PresentationDataType = \{"COLLOQUIUM", "CONFERENCE", "INVITED SPEAKER", "ONLINE", "OTHER PRESENTATION"\}`;
  
  - The different sets of enumerated values that the data properties can take:
    - E.g.: `PersonRoleDataType = \{"ASSISTANT", "AUTHOR", "CHECKER", "DTA PARTICIPANT", "EDITOR", "INTERVIEWER", "INVESTIGATOR", "SUBJECT", "TRANSCRIBER"\}`;
5c. The DTA Labels Ontology (1)

+ Main components – Classes:
  - The top-level concepts dealing with linguistic annotations, according to the OntoLingAnnot framework:
    - LinguisticAnnotation, LinguisticFeature, LinguisticTriple, LinguisticUnit, LinguisticAttribute, LinguisticValue, etc.
  - The units to which DTA labels can be attached, reused from the OntoTag and the OntoLingAnnot ontologies whenever possible:
    - E.g.: Fragment, Sentence, Phrase, Word.
  - A partial taxonomy of linguistic units, reused from the OntoTag and the OntoLingAnnot ontologies:
    - They anchor and structure the units that are annotated with DTA labels.
    - E.g.: PhonologicalUnit, MorphoSyntacticUnit, SyntacticUnit, DiscourseUnit, PragmaticUnit.
  - The subclassification of DTA labels (codings) included in the guide describing the DTA tool:
    - E.g.: SpeechActLabel, Context, MorphologicalCoding, etc.
Main components – Properties:

- Links to ISOcat categories ([http://www.isocat.org/](http://www.isocat.org/)) will be formalized by means of the `correspondsToISOcatDataCategory` data property.
  - Its value is an `xsd:anyURI` pointer to the category’s ISO persistent identifier.
  - The problem:
    - ISOcat data categories are undergoing no real standardization process yet.
    - There are several alternative categories for a given DTA label/unit (e.g. phrase/Phrase – 4 entries; word – 2 entries; verb – 3 entries).
    - Which one should be linked? Decisions have been deferred for the moment.
5c. The DTA Labels Ontology (3)

Main components – Logical axioms:

- Equivalences being established with OntoTag, OntoLingAnnot and GOLD items (ongoing process):
  - Implemented by means of `owl:equivalentClass` statements.
  - Semantic and/or terminological mismatches arise:
    - `ComplexSentence_{DTA} \supset ComplexSentence_{OntoLingAnnot}`
    - `ComplexSentence_{DTA} \supset CompoundSentence_{OntoLingAnnot}`
    - `CoordinateSentence_{DTA} = CompoundSentence_{OntoLingAnnot}`
    - `ComplexSentence_{DTA} \not\subset CoordinateSentence_{DTA}`
    - `CoordinateSentence_{DTA} \not\subset ComplexSentence_{DTA}`
5d. The DTA ontologies – Statistics

<table>
<thead>
<tr>
<th>DTA Ontologies Statistics</th>
<th>DTA Metadata Ontology</th>
<th>DTA Labels Ontology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classes</td>
<td>169</td>
<td>137</td>
</tr>
<tr>
<td>Object properties</td>
<td>139</td>
<td>12</td>
</tr>
<tr>
<td>Data properties</td>
<td>188</td>
<td>9</td>
</tr>
<tr>
<td>Annot. properties</td>
<td>61</td>
<td>5</td>
</tr>
<tr>
<td>Datatypes</td>
<td>32</td>
<td>7</td>
</tr>
<tr>
<td>Individuals</td>
<td>2</td>
<td>66</td>
</tr>
<tr>
<td>Axioms</td>
<td>2222</td>
<td>698</td>
</tr>
<tr>
<td>Logical axioms</td>
<td>1406</td>
<td>350</td>
</tr>
<tr>
<td>Subclass axioms</td>
<td>486</td>
<td>193</td>
</tr>
</tbody>
</table>

(Extracted from the NeOn ToolKit)
6a. Summary (1)

We have presented:

i. The first steps in the transformation of the DTA metadata and labels into a LLOD resource.

ii. A comparison that shows that this is, to the best of our knowledge, one of the most relevant and detailed initiatives in the study and annotation of child language.
6a. Summary (2)

Results:

i. Two ontologies, which formalize the DTA elements, being linked to LLOD cloud items.

ii. These ontologies will help us to:
   • Transform the DTA into an LLOD cloud item.
   • Standardize the annotations of the DTA and make them more interoperable.

iii. The detection of inconsistencies and gaps in the annotations of linguistic elements in the DTA, with the definitions in other linguistic resources.
6b. Future work

+ A suitable integration and linking of DTA annotations with the annotations resulting from CHILDES or the LA.

+ A re-engineering of the DTA to convert it into a semantic portal, using Semantic Web technologies.
  - This would allow us to produce automatically open linked data annotations in the future, instead of (1) storing the annotations first in a database; and then (2) transforming them into linked data.

+ Matching and linking the DTA ontologies with other LOD cloud resources, such as FOAF and/or Dublin Core (currently undergoing).

+ Inclusion of DTA labels for pragmatic annotation and speech transcription into ISOcat.
Acknowledgments

- The organizing committee of the first Linked Data in Linguistics workshop for helping us know of each other’s projects and therefore initiate this collaboration.

- The DTA project was supported by several funding sources
  - “Transforming the Primary Research Process through Cybertool Dissemination: An Implementation of a Virtual Center for the Study of Language Acquisition.” National Science Foundation grant to María Blume and Barbara Lust, 2008, NSF OCI-0753415
  - “Planning Grant: A Virtual Center for Child Language Acquisition Research. National Science Foundation grant to Barbara Lust, 2003, NSF BCS-0126546
  - “Planning Information Infrastructure Through a New Library-Research Partnership.” National Science Foundation Small Grant for Exploratory Research to Janet McCue and Barbara Lust, 2004-2006
  - Cornell University Faculty Innovation in Teaching Awards, Cornell Institute for Social and Economic Research (CISER)
  - New York State Hatch grant
  - Grant Number T32 DC00038 from the National Institute on Deafness and Other Communication Disorders (NIDCD).
Acknowledgments

The Virtual Center for Language Acquisition’s other founding members and affiliates:

- Suzanne Flynn (MIT)
- Claire Foley (Boston College)
- Marianella Casasola, Claire Cardie, James Gair, and Qi Wang (Cornell University)
- Liliana Sánchez (Rutgers University at New Brunswick) and Jennifer Austin (Rutgers University at Newark)
- YuChin Chien (California State University at San Bernardino)
- Usha Lakshmanan (Southern Illinois University at Carbondale).
- Cristina Dye (Newcastle University) and Yarden Kedar (Ben Gurion University at the Negeb).

Janet McCue of Cornell University Library and her collaborators at Cornell A. Mann Library for their assistance on integration of metadata standards and structure to our emerging DTA tool and their assistance in developing formal relations between research labs and University Libraries.
Acknowledgments

+ Application developers Ted Caldwell and Greg Kops (GORGES)
+ Consultants Cliff Crawford and Tommy Cusick
+ Student research assistants Darlin Alberto, Gabriel Clandorf, Natalia Buitrago, Poornima Guna, Jennie Lin, Marina Kalashnikova, Martha Rayas Tanaka, Lizzeth Jensen, María Jiménez, and Mónica Martínez
+ The many students at all the participating institutions who helped us with comments and suggestions.
References


References


References


Thank you!

Questions?