Transforming the Data Transcription and Analysis Tool Metadata and Labels into a Linguistic Linked Open Data Cloud Resource

Antonio Pareja-Lora
Univ. Complutense de Madrid
Facultad de Informática
28040 – Madrid, Spain
apareja@sip.ucm.es

María Blume
Univ. of Texas at El Paso
Dept. of Languages and Linguistics
Liberal Arts Bldg., Room 232
El Paso, Texas 79968
mblume@utep.edu

Barbara Lust
Cornell University
Dept. of Human Development
G57 Martha Van Rensselaer
Hall, Ithaca, NY 14853
bcl4@cornell.edu

Abstract

Developing language resources requires much time, funding and effort. This is why they need to be reused in new projects and developments, so that they may both serve a wider scientific community and sustain their cost. The main problems that prevent this from happening are that (1) language resources are rarely free and/or easy to locate; and (2) they are hardly ever interoperable. Therefore, the language resource community is now working to transform their most valuable assets into open and interoperable resources, which can then be shared and linked with other open and interoperable resources. This will allow data to be reanalyzed and repurposed. In this paper, we present the first steps taken to transform a set of such resources, namely the Data Transcription and Analysis Tool's (DTA) metadata and data, into an open and interoperable language resource. These first steps include the development of two ontologies that formalize the conceptual model underlying the DTA metadata and the labels used in the DTA to annotate both utterances and their transcriptions at several annotation levels.

1 Introduction

As the web evolves into the Web 2.0 and is complemented by the Web 3.0,[1] the Semantic Web and/or the Web of Data (Auer and Hellmann, 2012), the need for language resources to be transformed into open, sharable and interoperable resources becomes more urgent. Lately, this transformation has been achieved by converting language resources into linked open data sets and/or graphs. These linked data help formalize and make explicit common-sense knowledge in a way that satisfies the needs of the Web 3.0, the Semantic Web and/or the Web of Data. Indeed, computers are already using these linked data to process information "more intelligently".

In this context, many language resources may unfortunately be left aside and fade into oblivion if they fail to address this challenge (which would entail a waste of considerable data and effort for the scientific community). Making language resources easier to share and more interoperable would help researchers collaborate and build on others' work.

This is the case of the resources generated by the Data Transcription and Analysis Tool (DTA).[2] The DTA tool is a primary research web application that organizes metadata and data primarily for the study of language acquisition, either monolingual or multilingual.[3] Henceforth, we will use the term DTA to refer to the tool itself, its experiment bank component, and its associated corpora. The DTA allows for long distance collaborative research and serves as a teaching tool for training students on language data management and analysis. Besides providing a powerful relational database, which handles both experimental and naturalistic data, it also structures the primary data creation process from its initial stages. Hence, the DTA represents data so that it can be analyzed subsequently in a standardized and theory-neutral way, which ensures data comparability within a language and across languages. At the same time, it allows researchers to create project-specific codings, allowing multiple types of analyses in their own data or linking data across projects. This tool was created as part of the VCLA's[4] Virtual Linguistics Lab[5] to take advantage of the opportunities

[1] http://en.wikipedia.org/wiki/Web_2.0#Web_3.0

[2] http://webdta.clal.cornell.edu

[3] Access to the DTA cybertool is password protected due to Human Subjects confidentiality requirements and the intellectual property rights of the contributing researchers. To allow for wider dissemination, multiple levels of access must be set. The PIs are currently investigating potential funding sources for this dissemination.

[4] http://vcla.clal.cornell.edu/

[5] http://clal.cornell.edu/vll/

the digital age created for the interdisciplinary, cross-linguistic study of language acquisition (Blume and Lust, 2012; Blume et al., 2012).[6]

The research presented here is the result of a joint work in which we compared and linked two different language resources, namely OntoLingAnnot's ontologies (Pareja-Lora and Aguado de Cea, 2010; Pareja-Lora, 2012a; Pareja-Lora 2012b; Pareja-Lora, 2013) and the Data Transcription and Analysis Tool (Blume and Lust, 2012; Blume et al., 2012).

In this paper we introduce (i) the metadata and the labels that are used within the DTA to annotate data on language acquisition; and (ii) the two ontologies that we have now built to represent, respectively, the DTA metadata and the DTA labels. In some cases, these labels (such as Noun Phrase, Sentence, Statement, Question or Answer) can be linked to ISOCat categories (Windhouwer and Wright, 2012) and/or are equivalent to some GOLD element (Farrar and Langendoen, 2010).[7] These links and equivalences are being included in the ontologies as well, which should help add the DTA ontologies to the Linguistic Linked Open Data (LLOD) cloud[8] (Chiarcos et al., 2012) shortly.

Our paper is organized as follows. The DTA metadata categories are presented in section 2. Section 3 introduces the labels used in the DTA to annotate the data linguistically. A comparison of the DTA labels and metadata with those of other related projects, such as CHILDES and the Language Archive (LA), is provided in section 4. In Section 5, we show the two ontologies built to conceptualize, the DTA metadata and the DTA labels, each one in a dedicated subsection. Finally, section 6 discusses the conclusions of this research and gives an overview of our future work.

2 The DTA metadata categories

The DTA is based on 10 tables with the following basic markup categories: project, dataset, subject, session, recording, transcription, utterance, coding set, coding, and utterance coding. Metadata codings involve the project and subject levels and the dataset level leading to transcribed utterances and related linguistic codings. 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 subject screen is where all the participant info is stored. The application uses the UTF-8 encoding to store text and adopts the ISO 639-3 standard language codes, which cover over 7000 languages. It links with GeoNames.org for geographic reference.

Each project contains the following main sections of information: Project Main Info, References, Subjects, Datasets, Coding, and Queries.

2.1 The Project Main Info Section

Under *Project Main Info* there are three tabs: *Main Info, Results, Summary and Discussion*. *Main info* provides an overview of the main information on a project so that other users may decide whether they choose to continue reading about this project or move to another one. The text fields include title (project's name), principal investigator, additional and assisting investigators, acknowledgments, dates, purpose, leading hypotheses, and comments. *Results* and *summary and discussion* allow one to enter the results and the discussion for the whole project (from all the datasets).

2.2 The References Section

References include *publications*, *presentations*, *related studies*, and *references*. They all have the same basic structure based on an APA format. The only variation is on the items one can select under "type". Under *publications, related studies, and references*, "type" includes book, chapter, article, web page, thesis, dissertation, and other. All types include the basic fields title, authors, and date, and other needed fields according to the reference type. *Presentations* contains the following types: conference, invited speaker, colloquium, and other (which are also included under *related studies*) with the same basic fields as publications plus "place of publication/presentation", "URL" and "notes".

2.3 The Subjects Section

The subject data are not session-specific; i.e., permanent characteristics of the subject are recorded in this screen. The subject screen has two sections: *Subject* and *Caretakers*. *Subject* in-

[6] The DTA capabilities extend to other areas of language knowledge and use, such as language deterioration in adult dementia. Although the VLL and its cybertools were created with the study of language acquisition and multilingualism in mind, they can not only be expanded to other language areas but also used as a prototype for data management and linking in other areas of scientific investigation

[7] The cross-linguistic data in the DTA should also be a good test for how well GOLD categories work across languages, an issue central to the 2005 E-Meld workshop (*cf.* http://www.emeld.org/workshop/2005/).

[8] http://linguistics.okfn.org/resources/llod/

cludes ID, name, gender, DOB, nationality, ethnicity, place of birth, whether the subject has any language or cognitive impairment, whether Human Subjects required documents have been filled in and a multilingualism questionnaire has been completed, the subject's contact information, and comments. Information on the language(s), dialect(s), and levels of language comprehension and production for each language are also indicated. When subjects are children, information on their caretakers is also stored including relationship with the subject, occupation, name, contact information, languages, dialects, and levels of proficiency.

2.4 The Sessions and Datasets Section

The subjects participate in different recording sessions (tests, observations, or surveys). The sessions are organized in groups called *datasets*. Each subject has at least one session, but they can have more than one. All sessions for a subject may be part of one dataset but they can be divided into more datasets.[9] Each dataset contains the recordings, transcriptions, and codings for each session.

Datasets contain two main sections: *Main Info* and *Sessions*. *Main Info* includes title, type (investigation or experiment), topic, abstract, and related WebDTA project/datasets, hypotheses, subjects (a summary description of subjects in the dataset), methods (production, comprehension, perceptual discrimination, or grammaticality judgment), design (factors, variables, conditions, controls, specific hypotheses, statistical analyses), stimuli, procedures, scoring, results, and conclusions. *Sessions* include the following information fields: Session ID, date, interviewer, assistants, session length, task, languages used, and session location. Information on the subject characteristics at each particular session is also included: Current age (calculated by the DTA), number of siblings, position among siblings, address, length of residence, education, occupation, and school. Fundamental information (name and transcription identifier) on the session participants besides the subject and interviewer is also created. Information on the general activities carried out during the session and the analyses performed on the data are included.

[9] Each project user can determine what constitutes a dataset; they are usually divided in terms of the experimental task used (each different task used in one project constitutes an independent dataset) or by participant characteristics (e.g. Spanish-speakers vs. English-speakers).

3 The DTA Labels

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. Global codings can be used by all projects and can only be established by users with administrator access. Codings are grouped in coding sets. Simple coding sets were created to standardize and calibrate basic levels of linguistic analysis. They may also introduce students to linguistic coding with increasing levels of analysis/difficulty. Experimental projects create their project-specific coding sets in addition to basic ones, as do researchers who work with natural speech.

There are three basic coding sets: *Utterance transcription*, *speech act*, and *basic linguistic*. *Utterance transcription* includes text fields that give information to contextualize the utterances[10] and allows one to add simple linguistic descriptions, translations and glosses of non-English utterances.[11] *Speech act* lists common speech acts with some additional ones common in child data[12] and asks about the spontaneous or responsive character of the utterance, and therefore relates more to the pragmatic/discursive aspects of the data.[13] Finally, *basic linguistic* asks whether the utterance is a sentence or not, and whether the sentence has an overt verb,[14] as well as for the number of morphemes, words, and syllables of the utterance, to calculate the Mean Length of Utterance (MLU), an important developmental measure in child language acquisition. Additional basic linguistic codings are now being created.

[10] *General context* (a description of the participants, their location and activities throughout the session), *utterance context* (the context necessary to understand the contents of a particular utterance, for example, what the speaker is referring to or who they are addressing), and *comment*s.

[11] M*orphological coding*, *word-by-word gloss*, *general gloss* (a translation into English that conveys the meaning of the utterance regardless of structure), and *phonetic transcription*.

[12] Declarative/assertive, question, imperative, promise, wish/request, expressives/exclamations, yes/no/OK, naming, counting, singing, politeness, greetings, unclear, and other.

[13] Spontaneous, self-repetition, other repetition, answer-Y/N, answer-Wh, other answer (i.e., when the subject answers a question which is not a Wh-question or a Y/N-question), unclear, other.

[14] Verbless sentences are common in early child speech (e.g. *Me [ə] cookie from mommy*). The corresponding labels are "is this a sentence?" and "is the verb overt?".

4 Comparing the DTA with other systems

Certain databases, such as CHILDES and the Language Archive, share some of the purposes of the DTA. Given that both CHILDES and DTA have focused on child language data, they obviously have common or similar labels in the codings that they adopt (about 24[15]). However, one main difference is that the DTA provides the user with a structured interface for primary data entry and management, while CHILDES lists possible metadata fields in its accompanying manual, and provides no structure for the researcher. The information on what to fill in when archiving data is provided in the CHILDES manual in a narrative form.[16]

One label in CHILDES may be covered by more than one label in the DTA. For example, the "creator" label corresponds to three labels in the DTA, namely "Principal Investigator", "Additional Investigators", and "Assisting investigators." A "How was data collected" label is covered by the DTA's more specific fields under the Dataset Main Info: "type, method type, method details, design, and stimuli". Some identical labels refer to different things.[17] CHILDES asks for information on the funding for the project which is not included in the DTA, but could easily be incorporated, and on some other aspects which the DTA creators did not consider relevant, e.g., "religion", "interests", "friends", "layout of child's home and bedroom" and whatever is included under "and so forth".

Although researcher compliance in filling the required fields cannot be assured, the main advantage of the DTA is its structured format, which helps researchers in the primary data creation process.[18]

To compare the DTA and the Language Archive (LA), we looked at the metadata fields in Brugman et al. (2003). For clarification of the LA field definitions we consulted IMDI Metadata 3.0.4. The DTA and the LA share many of their fields since both have language archiving and metadata creation purposes in mind. The main differences are related to content organization. While the LA organizes data in terms of sessions, with project information contained inside a session and no dataset level, the DTA organizes it in terms of projects that contain datasets which in turn contain sessions.[19]

The main differences between the systems stem from their partially divergent purposes. The DTA was developed mainly for child language acquisition so it asks for detailed information on the child's caretakers and it was intended for experimental as well as observational data; thus it has much more detailed fields related to project and dataset experimental design (19) which do not exist in the LA. The LA has a much more detailed information section on the different types of resources (it distinguishes "source", "resource", and "written resource" with detailed information for type, format, encoding, access, and anonymity for all), and on the type of communication context and genre of the interaction (30), some of which would be relevant for the DTA. Surprisingly, there are more than a few fields that the DTA has which are not child/experiment specific which the LA does not have, such as the participant's length of residence at the session location, date of birth, nationality, 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. The DTA also has a more detailed division of references as explained in section 2.2 above.

[15] Numbers in parentheses refer to number of fields.

[16] "**7. Biographical data.** Where possible, extensive demographic, dialectological, and psychometric data should be provided for each informant. There should be information on topics such as age, gender, siblings, schooling, social class, occupation, previous residences, religion, interests, friends, and so forth.[…]" (MacWhinney, 2012: 23)

[17] E.g. "acknowledgments" in the DTA refers to acknowledgments of the persons who made the project possible, and in CHILDES it refers to the rules for citing data used by a researcher who did not create such data.

[18] In CHILDES, the requested information is not completed in several of the available corpora. To take one relevant case, the CHILDES corpus does not have all the requested information and includes several pieces of information (related to OLAC and IMDI), which are not mentioned in the manual. To get more complete information on a corpus, readers are directed to the Database Manuals in which each corpus is described. Length of descriptions varies from a short paragraph to two or three pages.

[19] The DTA and the LA share very few fields at the different levels (i.e., project description (3), session description (4) and transcription/annotation (1)). Several fields have similar names in the two systems (20). Nine fields in the LA are divided into more than one field in the DTA (e.g., *task* in the LA corresponds to *dataset method type*, *dataset method details*, and *session task* in the DTA, *annotator* in the LA corresponds to *transcriber* and *checker* in the DTA).

5 Ontological Formalization of DTA Categories

As shown in the previous section, the DTA provides the most detailed and exhaustive repertoire developed so far with metadata and labels for child language analysis and annotation. Therefore, it seems reasonable to formalize this repertoire by means of some ontologies. This formalization will help to compare, integrate and link DTA annotations with the annotations resulting from CHILDES or the LA later on.[20]

As noted above, the DTA language acquisition data are annotated with extensive metadata, such as the time and place where they were collected, and the data (e.g. transcriptions) are annotated linguistically. At this time, these linguistic annotations pertain mostly to the pragmatic and the phonological levels, in order to calibrate incoming data, but also, to a lesser extent, to the morphosyntactic and the syntactic levels.

Thus, the first ontology built for DTA (namely the DTA Metadata Ontology) contains a formalization of the DTA metadata, which is particular of this initiative and, hence, had to be built mostly from scratch. The second ontology (that is, the DTA Labels Ontology) includes a conceptualization of the labels used to annotate DTA transcriptions linguistically. Accordingly, it reuses other linguistic resources and ontologies. In particular, the OntoLingAnnot set of ontologies (Pareja-Lora and Aguado de Cea, 2010; Pareja-Lora, 2012a; Pareja-Lora 2012b; Pareja-Lora, 2013) has been reused to formalize the DTA pragmatic level labels, [21] including convenient links to ISOCat[22] categories and OWL equivalences with GOLD elements. This will help make the DTA ontologies become part of the Linguistic Linked Open Data (LLOD) cloud. Each of the ontologies is described below.

5.1 The DTA Metadata Ontology

The DTA Metadata Ontology contains the different elements described in section 2. In its development, we have followed as faithfully as possible the categorizations applied in developing the DTA. The top-level classes of this ontology are shown in Figure 1.

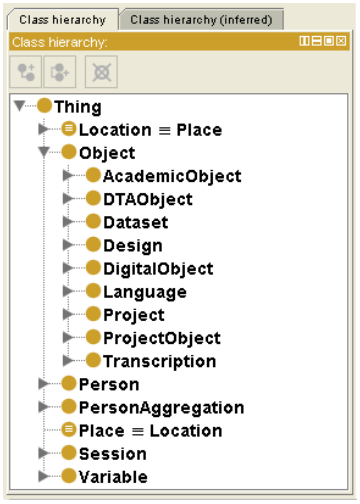

Figure 1: DTA Metadata Ontology – Main classes

These top-level classes include the formalization of some of the ten DTA basic categories presented in section 2 (namely `Project`, `Dataset`, `Session` and `Transcription`). The ones not shown in the figure are subclasses of one or several of the classes shown: `Subject` is `rdfs:subClassOf Person`; the classes formalizing recording, coding and coding set are subclasses of `DTAObject` and of `ProjectObject`; and `Utterance` and `UtteranceCoding` have been included in the DTA Labels Ontology (cf., next section). Other relevant items in the DTA, i.e. languages, are also represented at this level, by means of the class `Language`.

The `Project` and `ProjectObject` classes have two main subclasses respectively, i.e., `DTAProject` and `DTAProjectObject`. They are the most prominent subclasses of this ontology, as shown in Figure 2. Indeed, as shown in the figure, most of the concepts presented in sections 2.1-2.4 have been represented as subclasses of these two concepts.

The classes `DTAInformationSection` and `DTAInfoTab` are related by means of the object property `HasPart` in the ontology, that is, `DTAInformationSection HasPart DTAInfoTab`. Thus, each of the tabs associated to the different sections of information have been straightforwardly formalized as subclasses of one of the subclasses of `DTAInfoTab`, namely `ProjectMainInfoTab`, `ReferencesTab`, `SubjectsTab` and `DatasetTab`. They are not exhaustively described here to avoid redundancy with section 2. However, it is important to note that (1) the formalization of the ReferencesTab entailed the inclusion of a whole sub-ontology of academic objects, shown in Figure 3.

[20] The resulting ontologies have been published under a 3-clause BSD license at https://github.com/apareja/DTA_Ontologies.
[21] For more information about OntoLingAnnot (including the code of its ontological modules), please contact the first author of this paper.
[22] http://www.isocat.org

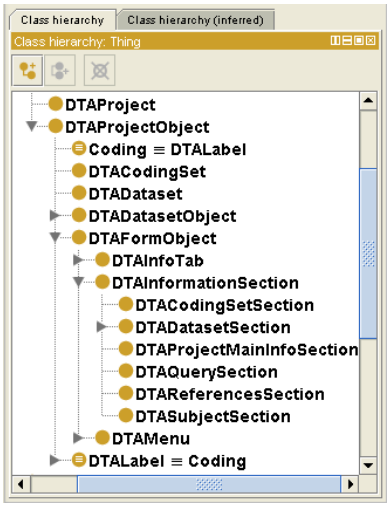

Figure 2: DTA Metadata Ontology – `DTAProjectObject` main subclasses

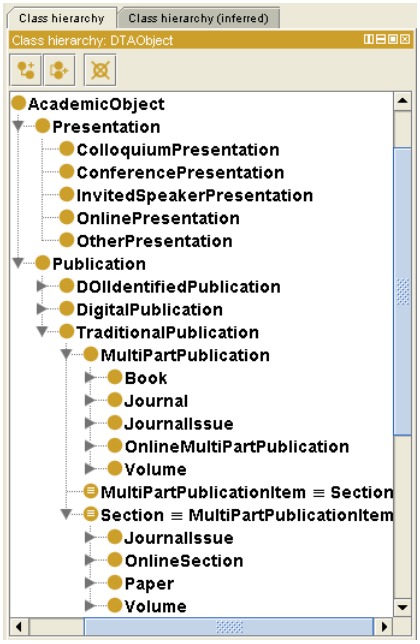

Figure 3: DTA Metadata Ontology – the `AcademicObject` sub-ontology

All these classes have corresponding data properties attached, which represent the different text and menu fields used in DTA to assign values and annotations (*cf.* section 2). The resulting hierarchy of properties is partially shown in Figure 4. Also a number of object properties have been formalized in this ontology, but they are not described due to space limitations.

5.2 The DTA Labels Ontology

The DTA Labels Ontology includes the DTA elements discussed in section 3. They are used in the annotation of utterances in the DTA. We decided to develop a separate ontology for these elements due to their more general nature and, hence, their higher reusability in all kinds of linguistic annotation projects.

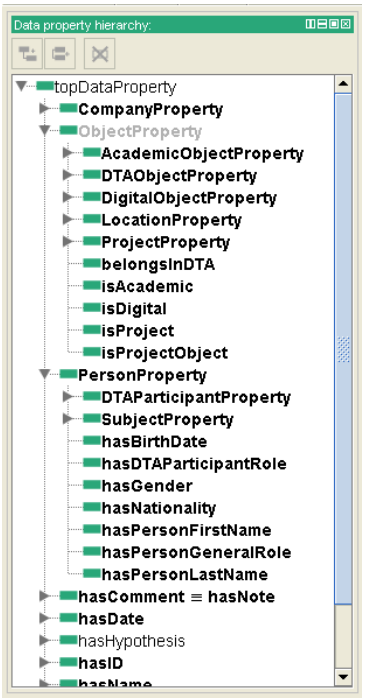

Figure 4: DTA Metadata Ontology – The hierarchy of data properties

In this case, since the DTA labels are a particular case of linguistic annotation, we reused other existing ontologies and repositories of categories for linguistic annotation, such as GOLD, DCR/ISOCat, OntoTag (Aguado de Cea et al., 2002; 2004; Pareja-Lora, 2012c), and OntoLingAnnot. We kept the same **criteria and methodologies of classification and subdivision** applied in these other linguistic resources, making the DTA Labels Ontology more interoperable with them.[23] For example, we developed three separate taxonomies within the ontology, one for linguistic units, one for linguistic attributes (or features), and another one for the linguistic values that these attributes can take. The super-classes of these taxonomies are, respec-

[23] However, the formalization of the links and the equivalences with e.g. GOLD and ISOCat is still ongoing. Whereas GOLD entities are linked by means of `owl:equivalentClass` statements, ISOCat categories are linked by means of an ad-hoc defined data property, namely `correspondsToISOCatDataCategory`, whose value is an `xsd:anyURI` pointer to the category's ISO persistent identifier. A matching between the DTA ontologies with the FOAF (Friend Of A Friend) vocabulary (http://www.foaf-project.org) and with the Dublin Core Metadata (http://dublincore.org/) is planned as well. All the matches found will be added subsequently to the DTA ontologies.

tively, LinguisticUnit, LinguisticAttribute and Linguistic-Value, which have been imported from the OntoLingAnnot ontologies.

Each of these taxonomies is linked to each other by the corresponding relation of the OntoLingAnnot model, namely: LinguisticUnit **hasFeature** LinguisticAttribute, LinguisticAttribute **takesValue** LinguisticValue, LinguisticValue **isValueTakenBy** LinguisticAttribute, LinguisticAttribute **isAttachedTo** LinguisticUnit.

We created a **DTALabel class**, which is a rdfs:subClassOf LinguisticAttribute. Most DTA labels are subclasses of DTALabel. We have only classified DTA glosses differently, since they are in fact the aggregation of a label (namely WordByWordGlossLabel or GeneralGlossLabel, which are the subclasses of DTAGlossLabel – see below) and a value (the actual text provided as a gloss).

Each DTALabel is a GlobalCoding or a ProjectSpecificCoding. The main **subclasses of GlobalCoding** are BasicLinguisticLabel (which has only one subclass, i.e. DTASyntacticLabel), UtteranceTranscriptionLabel (whose subclasses are Context, DTAGlossLabel, MorphologicalCodingLabel and PhoneticTranscriptionLabel) and SpeechActLabel, whose subclasses detail the attributes that can be applied to Searle's types of speech acts (luo:Assertive, [24] luo:Commissive, luo:Declaration, luo:Directive and luo:Expressive[25]) and have been classified accordingly.

The main **subclasses of ProjectSpecificCoding** are isAdjectivalPhrase, isAdverbialPhrase, isFragment, isNounPhrase, isPrepositionalPhase, isRelativePronoun, isSentence and isWh-Word.

The **linguistic units included and/or imported into the DTA Labels Ontology** are the following: luo:PhonologicalUnit (whose main subclasses are luo:Phoneme, luo:ProsodicUnit, luo:Syllable and luo:Utterance), luo:MorphoSyntacticUnit (whose main subclasses are luo:MorphologicalUnit, luo:SyntacticUnit and luo:Word), luo:SemanticUnit, luo:DiscourseUnit, luo:PragmaticUnit (which is one of the superclasses of luo:SpeechAct in this ontology, together with luo:SpeechUnit), and luo:TextUnit (whose main subclasses relevant to DTA, are MorphologicalCoding, PhoneticTranscription, PhoneticTranscriptionSymbol, UtteranceTranscription and luo:Text).

We have also imported the luo:Morpheme class, which is an rdfs:subClassOf luo:MorphologicalUnit, and several subclasses of luo:SyntacticUnit, such as luo:Clause, luo:Phrase (and some of its subclasses, i.e. luo:AdjectivalPhrase, luo:AdverbialPhrase, luo:NounPhrase and luo:PrepositionalPhrase) and Sentence (together with some of its subclasses, i.e., ComplexSentence, CompoundSentence and SimpleSentence). We have also added a particular DTA rdfs:subClassOf luo:SyntacticUnit (Fragment), which represents the syntactic projection of those transcribed utterances that cannot be considered an instance of any of the other syntactic units.

The main **individuals** of the DTA Labels Ontology are members of the subclasses of SpeechActLabel; for example, CountingLabel, GreetingLabel, NamingLabel, PolitenessLabel, SingingLabel, PromiseLabel, QuestionLabel and YesOrNoOrOKLabel formalize the particular types of speech act labels available within the DTA (see footnotes 12 and 13). They are used for the subclassification and/or annotation of utterances as speech acts, for instance.

Briefly, the DTA Label Ontology entities were categorized as LinguisticUnit, LinguisticAttribute or LinguisticValue subclasses or individuals, and they were also linked among them with suitable **object properties**, such as Has/PartOf, Labels/isLabelledWith, hasSyntacticProjection/isSyntacticProjectionOf, or hasTranscription/isTranscriptionOf. As shown in these examples, we declared an inverse property for each direct object property identified, in order to facilitate inferences.

Overall, the most relevant characteristic of this categorization is that it allows for a formalization of DTA annotations as linguistic RDF triples, as in the OntoLingAnnot model. This will allow for

[24] The luo namespace stands for OntoTag's and OntoLingAnnot's Linguistic Unit Ontology (LUO).

[25] This classes are subclasses of luo:SpeechAct, see below.

a fairly straightforward conversion of DTA annotations into RDF triples and, therefore, into linked (open) data. Some statistics about the number of classes, properties, data types, individuals and axioms included in these ontologies have been included in Table 1.

Table 1: Some statistic about the elements included in the DTA ontologies

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

6 Summary and future work

In this paper, we have presented the first steps in the transformation of the DTA metadata and labels into a Linguistic Linked Open Data resource. The main results of this work are the two ontologies presented in Section 5, which formalize the DTA elements, described in Sections 2 and 3. We have also provided a comparison in Section 4 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.

A suitable integration and linking of DTA annotations with the annotations resulting from CHILDES or the LA is still pending. This would first require the formalization of the label mappings between DTA and CHILDES and the LA (already identified in Section 4) in the two ontologies presented here.

Other future work might include 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.

Even though it is in its initial stages, this collaboration has already produced two immediate outcomes: (i) the evaluation of the categories included in OntoLingAnnot's ontologies against the resources in the DTA[26] and (ii) the detection of inconsistencies and gaps in the annotations of linguistic elements in the DTA, with the definitions in other linguistic resources.[27] This two-way evaluation follows an interdisciplinary approach (computational and linguistic) and will allow for the transformation of the existing DTA data into linked (open) data, using the items now formalized in the DTA Metadata Ontology and the DTA Labels Ontology, allowing future linked-data-based, data-intensive research. Moreover, since the OntoLingAnnot model is ISO conformant and aims at the interoperability of linguistic resources and annotations, it will lead to the standardization of the DTA in order to make it more interoperable.

Acknowledgments

The authors thank 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. We also thank the anonymous reviewers for their many useful suggestions for this paper.

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).

[26] For example, the inclusion of `rdfs:subClassOf luo:SyntacticUnit (Fragment)`; *cf.* section 5 and, in particular, Figure 3.

[27] For example, the DTA classifies sentences according to their structure into two types: complex and simple; and then subdivides complex sentences into those involving coordination and those involving subordination. This classification does not correspond to how sources such as the SIL Glossary (http://www-01.sil.org/linguistics/GlossaryOflinguisticTerms/) or OntoTag and OntoLingAnnot classify them. In these other resources, (1) complex sentence refers to sentences including at least one main clause and at least one subordinate clause; and (2) compound sentence refers to sentences that consist of two or more coordinate clauses.

We gratefully acknowledge the collaboration of the Virtual Center for Language Acquisition's other founding members: Suzanne Flynn (MIT), Claire Foley (Boston College), Marianella Casasola, Claire Cardie, James Gair, and Qi Wang (Cornell University); Elise Temple (NeuroFocus); Liliana Sánchez (Rutgers University at New Brusnwick); Jennifer Austin (Rutgers University at Newark); YuChin Chien (California State University at San Bernardino); and Usha Lakshmanan (Southern Illinois University at Carbondale). We are grateful for the collaboration of scholars who are VCLA affiliates including Sujin Yang (Korea), Gita Martohardjono, Valerie Shafer, and Isabelle Barrière (City University of New York); Cristina Dye (Newcastle University); Yarden Kedar, (the Center for Academic Sudies, Israel), Joy Hirsch (Columbia University); Ellen Courtney and Alfredo Urzúa (University of Texas at El Paso); Sarah Callahan (University of California at San Diego); Jorge Iván Pérez Silva (Pontificia Universidad Católica Del Perú), Kwee Ock Lee (Kyungsung University); R. Amritavalli (Central Institute of English and Foreign Languages); A. Usha Rani (Osmania University).

We thank 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; and the many students at all the participating institutions who helped us with comments and suggestions. In particular, we thank 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.

References

Guadalupe Aguado de Cea, Asunción Gómez-Pérez, Inmaculada Álvarez de Mon, Antonio Pareja-Lora, and Rosario Plaza-Arteche. 2002. OntoTag: A semantic web page linguistic annotation model. In *Semantic Web Meets Language Resources*. AAAI Technical Report WS-02-16, pp. 20–29. Menlo Park, California, USA, 2002. AAAI Press.

Guadalupe Aguado de Cea, Asunción Gómez-Pérez, Inmaculada Álvarez de Mon, Antonio Pareja-Lora. 2004. OntoTag's linguistic ontologies: Improving semantic web annotations for a better language understanding in machines. In *Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'04)*, vol. 2, pp. 124–128, Washington, DC, USA, 2004. IEEE Computer Society.

Sören Auer and Sebastian Hellmann. The Web of Data: Decentralized, collaborative, interlinked and interoperable In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012)*, Istanbul, Turkey, May 2012.

María Blume and Barbara Lust. 2012. First steps in transforming the primary research process through a Virtual Linguistic Lab for the study of language acquisition and use: Challenges and accomplishments. *Journal of Computational Science Education*, vol. 3 (1): 34-46.

María Blume, Suzanne, Flynn, and Barbara Lust. 2012. Creating linked data for the interdisciplinary international collaborative study of language acquisition and use: Achievements and challenges of a new Virtual Linguistics Lab. In Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann (eds.) *Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata*, pp. 85-96. Heidelberg: Springer.

Hennie Brugman, Daan Broeder, and Gunter Senft. 2003. Documentation of language and archiving of language data at the Max Planck Institute for Psycholinguistics in Nijmegen. Paper presented at the *Ringvorlesung "Bedrohte Sprachen" Sprachenwert – Dokumentation – Revitalisierung*. Fakultät für Linguistik und Literaturwissenschaft. Universität Bielefeld. 05/02/2003. [http://www.mpi.nl/IMDI/documents/articles/BI-EL-PaperA2.pdf]

Christian Chiarcos, Sebastian Hellmann and Sebastian Nordhoff. 2012. Linking linguistic resources: Examples from the Open Linguistics Working Group, In Christian Chiarcos, Sebastian Nordhoff and Sebastian Hellmann (eds.) *Linked Data in Linguistics. Representing Language Data and Metadata*, pp. 201-216. Heidelberg: Springer.

Scott Farrar and D. Terence Langendoen. 2010. An OWL-DL implementation of GOLD: An ontology for the Semantic Web. In A. Witt and D. Metzing (eds.) *Linguistic Modeling of Information and Markup Languages*, pp. 45-66. Dordrecht:Springer.

IMDI. 2003. Isle Metadata Initiative (IMDI) Part 1. Metadata elements for session descriptions. Version 3.0.4. October 2003. [http://www.mpi.nl/IMDI/documents/Proposals/IMDI_MetaData_3.0.3.pdf]

Brian MacWhinney. 2012. *The CHILDES Project. Tools for analyzing talk-Electronic edition. Part 1.*

The CHAT transcription format. August 6, 2012. [http://childes.psy.cmu.edu/manuals/CHAT.pdf]

Antonio Pareja-Lora. 2012a. OntoLingAnnot's Ontologies: Facilitating Interoperable Linguistic Annotations (Up to the Pragmatic Level). In Christian Chiarcos, Sebastian Nordhoff and Sebastian Hellmann (eds.) *Linked Data in Linguistics. Representing Language Data and Metadata*, pp. 117-127. Heidelberg: Springer.

Antonio Pareja-Lora. 2012b. OntoLingAnnot's LRO: An Ontology of Linguistic Relations. In *Proceedings of the 10th Terminology and Knowledge Engineering Conference (TKE 2012)*. Madrid, June 2012, pp. 49-64. [http://www.oeg-upm.net/tke2012/proceedings, paper 04]

Antonio Pareja-Lora. 2012c. *Providing Linked Linguistic and Semantic Web Annotations – The OntoTag Hybrid Annotation Model*. Saarbrücken: LAP – LAMBERT Academic Publishing.

Antonio Pareja-Lora. 2013. The pragmatic level of OntoLingAnnot's ontologies and their use in pragmatic annotation for language teaching. In J. Arús, M.E., Bárcena, and T. Read (eds.) *Languages for Special Purposes in the Digital Era.* Springer [IN PRESS].

Antonio Pareja-Lora and Guadalupe Aguado de Cea. 2010. Modeling Discourse-related terminology in OntoLingAnnot's ontologies. In *Proceedings of the TKE 2010 workshop "Establishing and using ontologies as a basis for terminological and knowledge engineering resources"*. Dublin, August 2010.

Menzo Windhouwer and Sue Ellen Wright. 2012. Linking to linguistic data categories in ISOcat. In Christian Chiarcos, Sebastian Nordhoff and Sebastian Hellmann (eds.) *Linked Data in Linguistics. Representing Language Data and Metadata*, pp. 99–107. Heidelberg: Springer.