Modelling Discourse-related terminology in OntoLingAnnot's ontologies

Antonio Pareja-Lora
Dept. Sistemas Informáticos y Computación
Universidad Complutense de Madrid

Guadalupe Aguado de Cea
Ontology Engineering Group
Universidad Politécnica de Madrid

Keywords: Discourse, ontologies, terminology, OntoLingAnnot

Abstract

Recently, computational linguists have shown great interest in discourse annotation in an attempt to capture the internal relations in texts. With this aim, we have formalized the linguistic knowledge associated to discourse into different linguistic ontologies. In this paper, we present the most prominent discourse-related terms and concepts included in the ontologies of the OntoLingAnnot annotation model. They show the different units, values, attributes, relations, layers and strata included in the discourse annotation level of the OntoLingAnnot model, within which these ontologies are included, used and evaluated.

Introduction

In recent years, discourse annotation and the development of discourse-annotated corpora have attracted growing attention within the linguistic resource community. This attraction has been motivated by (1) the matureness of other levels of annotation, such as POS tagging, syntactic annotation and also, to some extent, semantic tagging, and (2) the new results coming from the Discourse Analysis field (Mann and Thompson 1988; Polanyi 1988; Longacre 1996; Van Dijk 1997; Schriffrin *et al.* 2001; Mitkov 2002; Prévot 2004; Palmer *et al.* 2005; Prasad *et al.* 2008).

These new results in Discourse Analysis allow for a clearer definition of (a) the scope, the phenomena and the components of the discourse level of language, as well as (b) the interface of this level with the remaining ones (the morphological, the syntactic, the semantic and even the pragmatic). This, in turn, enables a more precise definition and specification of the terms that should be included in a discourse annotation schema, in order to make explicit the structure and the logical components of discourse.

Meanwhile, a growing interest has lately risen in ontologies (Gruber 1993; Borst 1997) within Computational Linguistics. Some linguistically relevant ontologies (Mahesh and Nirenburg 1995; Schalley and Zaefferer 2007; Java *et al.* 2007), such as linguistically grounded ontologies (Buitelaar *et al.* 2009), ontologies of (or for) Linguistics (Farrar and Bateman 2005), ontologies of languages (Schalley and Zaefferer 2007), ontologies of linguistic annotations (GOLD 2010; OntoNotes 2010; Farrar 2007, Chiarcos 2008) and architectures of linguistic ontologies (Buyko *et al.* 2008) have been developed thus far. Some others, such as the FunGramKB ontology (Mairal Usón and Periñán Pascual 2009), focus mainly on modelling the language-independent counterpart of semantic knowledge, thus representing (1) the universal concepts that are present in the human mind, (2) their semantic properties (into the so-called thematic frames), and (3) their meaning definitions (by means of meaning postulates). In addition, some interesting efforts have been performed to model discourse knowledge related to meetings and the associated facts, such as dialogues, agent communication and interaction, or turn-taking management (Niekrasz and Purver 2005). However, to the best of our knowledge, none of them has tried a complete formalization of the Discourse Level.

Nevertheless, a complete representation of the Discourse Level by means of ontologies seems a most convenient way to account for the terminology of this domain and also to formalize it for its use within discourse annotation schemata, as this model complies with the schemata implemented in Semantic Web languages, such as RDF(S) or

OWL, which express an annotation by means of a <*Subject, Predicate, Object*> triple. In such a triple, at least the *Subject* and the *Object* are generally specified by means of ontological terms. These languages and their associated format and style of annotation (*i.e.*, the triples) are considered the best way to develop linguistic annotation schemata in LAF/GrAF (ISO 2008a). Therefore, the need for an ontology-based formalization of the terms belonging to the Discourse Level seems rather urgent.

In this paper, we present the ontological modelling of the discourse-related terminology included in OntoLingAnnot's linguistic ontologies. First, we show the foundations of the OntoLingAnnot model, that is, the hypotheses underlying it. Then we give an overview of the whole set of OntoLingAnnot's ontologies and, as a way of example, in dedicated sections, we introduce the different discourse-related terms formalized in these ontologies, that is, the main units, such as attributes, values, relations, layers and strata that best model the discourse level. After that, we discuss the related works and the main contributions of the present research, the work to be done and the conclusions derived from this work. This paper concludes with the references and an appendix, including the main acronyms used to refer to the terms defined in the paper.

OntoLingAnnot's Foundations

The OntoLingAnnot model was devised taking into account the standardisation efforts being carried out by the ISO TC37/SC4 subcommittee on linguistic annotation. Most of these efforts have focused on delimiting either (a) separated and specific standard schemes for particular levels of annotation, such as MAF (ISO 2008c), SynAF (ISO 2006), SemAF/Time (ISO 2007) or SemAF-Dacts (ISO 2009), or (b) formal and abstract standard frameworks for the development of these standard schemes, such as LAF/GrAF (ISO 2008a) or LMF (ISO 2008b). However, the OntoLingAnnot model was developed following a comprehensive, complementary approach, which considered all these levels of annotation altogether, not separately. The levels of annotation, together with their scope (as defined in the OntoLingAnnot model), have been summarized in Table 1.

This comprehensive approach allowed comparing these different level-dependent annotation schemes and finding the differences and similarities between them, so as to bear a general and uniformed (level-independent) annotation scheme across levels. In this comparison process, some regularities and uniformities across levels were found, which help structure and formalize all of them. Hence, these regularities require a proper formalisation as well and they are referred to as layers of linguistic description. Consequently, all linguistic phenomena can be classified according to the level and the layer to which they belong. In the following paragraphs, we describe the layers of linguistic description that capture these regularities across levels and how they are delimited.

Table 1. A summary of annotation levels and their scope.

ANNOTATION LEVEL	ANNOTATION SCOPE
Morphological	Up to word structure and meaning (including morph-related annotation).
Syntactic	Up to sentential structure (including multiword token, phrasal and clausal structure).
Semantic	Up to propositional structure and sentence meaning (including phrase and clause meaning), that is, propositional meaning.
Discourse	Up to discourse structure (including coherence relation-based structure) and the supra-sentential and locutionary meaning[1] of texts (including anaphora resolution).
Pragmatic	Up to the illocutionary and perlocutionary structure and meaning[2] of texts (including deictic resolution and other pragmatic relation annotation).

In terms of Saussure's (1916/1983) studies and theory, a **paradigm** is a class of linguistic units which are somehow exchangeable in a given piece of discourse (respecting the structure, but possibly changing its meaning). For example, nouns, verbs, adverbs, adjectives, etc., belong to the paradigm (class) of the morpho-syntactic units. The particular meaningful (ordered) context in which a linguistic unit co-occurs with other linguistic units is called a **syntagm**. Phrases, clauses and sentences are examples of syntagms (at the syntactic level). A **paradigmatic relation** (Crystal 1992) is a substitutional relationship 'that a linguistic unit has with other units in a specific context'. The relationships holding between the units which co-occur in a syntagm and between the units and the whole syntagm are called syntagmatic relations

(Crystal 1992). Both paradigmatic and **syntagmatic relationships** can be established at all levels of (linguistic) analysis and, 'constitute the statement of a linguistic unit's identity within the language system'; therefore, both types of relations characterize each linguistic unit.

Accordingly, the linguistic value (*i.e.* the meaning) of a unit can be fixed by means of two different elements, namely,

1. by contrast with the other units of its paradigm that might be replaced it in a given syntagm. Most frequently, these other units are left aside, unchosen and absent from the syntagm.

2. by combination with the value (*i.e.* the meaning) of other units that precede and follow it. Most frequently, these other units are also made explicit in the syntagm.

Thus, it seems reasonable that these two elements should be annotated somehow at each level of linguistic description, that is, for every distinguishable unit in a text. But, when comparing different schemes, levels and fashions/trends of annotation, the first issue that might attract our attention is that most of them focus *either* on

1. the annotation of the *paradigm* that a specific linguistic unit belongs to (as in POS-tagging or in Named Entity Recognition and Classification, for example), or
2. the annotation of the *syntagm* to which each unit belongs, and the *syntagmatic* relationships with its co-occurring units in that

1 According to Yule (1996), the action performed by producing an utterance consists of three related acts: the locutionary act (i.e. producing a meaningful expression, that is, an utterance,), the illocutionary act (i.e. the function which the utterance is expected to carry out or the purpose according to which it is uttered), and a perlocutionary act (i.e. the effect that the utterance is intended to produce). Therefore, the **locutionary meaning** of a text can be considered a straightforward interpretation of the text, without considering any further contextual information.

2 As pointed out in the previous footnote, the illocutionary meaning and the perlocutionary meaning of a text have to do, respectively, with the purpose and the effect that the text is expected to perform and produce, and therefore, can be considered a contextualized interpretation of the text.

syntagm (as in parsing or in syntactic annotation).

Yet, all the annotations associated to these aspects must be distributed among some uniform annotation **layers** across levels. As for the OntoLingAnnot model, the division into layers of the different levels has been inspired by the EAGLES (1996b) recommendations for the syntactic annotation of corpora. These recommendations distinguish eight different layers of annotation at the syntactic level:

1. The first EAGLES (1996b) layer, the so-called bracketing of segments, deals with the problem of delimiting the units which are recognized as having a syntactic integrity. This is a common problem for the rest of levels as well (*i.e.* the delimitation of their units). This problem has to be solved by means of an appropriate and particular segmentation process in each case (tokenisation at the morphosyntactic level, chunking or parsing at the syntactic level, etc.). Therefore, it seems most logical to generalize the application of this first syntactic layer to the rest of the levels considered in the annotation model. Accordingly, a proper and particularized bracketing layer (that is, a segmentation layer) will be distinguished at each of these levels in the present annotation scheme.

2. The second EAGLES (1996b) layer, in charge of labelling the category of the segments, is separated explicitly from the bracketing layer. Thus, finding out the paradigm of a unit (as, for example, determining of the POS tag of a morphosyntactic unit) will be considered as a different layer from the one in which the unit itself is delimited. Besides, from a conceptual point of view, this annotation part follows its identification.

3. The third and fourth EAGLES (1996b) layers, responsible for showing dependency relations and indicating functional labels, can be thought of as dealing with the annotation of syntagmatic relations. Once again, it can be regarded as a separated annotation of

a. the existence of a syntagmatic relation detected anyhow, and

b. the annotation of the type (*i.e.* the paradigm) of the syntagmatic relation detected.

   This separation might be based on and supported by the fact that it is usually easier to detect the existence of a syntagmatic relation than to find the paradigm of that syntagmatic relation; thus, annotating the former should be regarded as mandatory, whilst annotating the latter would be regarded as merely recommended. This is one of the assumptions underlying the design of the present scheme. Therefore, at each level, two additional layers will be distinguished: one for making explicit the existence of each syntagmatic relation, and another one for labelling the type (paradigm) of the relations already made explicit.

4. The subsequent EAGLES (1996b) layers can be regarded as complementary or secondary annotation facets of the paradigm of a unit, or of its syntagmatic relations with other co-occurring units. Hence, they originate no further layers of annotation within a given level. However, each of them may originate a proper (sub-)stratum within a particular Sub-Layer. This organisation of sub-layers into strata and sub-strata helps structure that particular layer internally. This is the origin of the subdivision of sub-layers into strata and sub-strata of the present model.

Apart from the layers inspired by EAGLES (1996b), there is another layer that should also be annotated at each level as well. In most cases, the syntagmatic relations that hold between two or more units at a level point out the existence of another higher-level unit or syntagm. This higher-level unit or syntagm, in turn, requires also a proper segmentation and annotation. Hence, these higher-level units or syntagms (which result from the syntagmatic interrelation of the units or syntagms of a given level) should be annotated as well.

To sum up, the OntoLingAnnot model proposes dividing every level of the annotation scheme into the following set of layers:

a) segmentation, or constituent unit recognition;

b) constituent unit paradigmatic labelling (which involves both the identification of the unit paradigm and its consequent sub-classification);
c) syntagmatic relation identification;
d) syntagmatic relation labelling (which also involves both the identification of the relation paradigm and its consequent sub-classification); and
e) resulting unit (syntagm) determination.

The resulting classification can be summarized in a matrix (or a table). In this matrix, (1) the rows represent the different levels of linguistic description and annotation considered; (2) the columns represent the different linguistic layers which formalize the regularities and uniformities present in all the levels considered; and (3) each cell represents the set of phenomena identified for a given level and a given layer, which are grouped together into a so-called Sub-Layer. The resulting matrix is shown in Table 2. The specific characteristics and particularities that should be annotated at a Sub-Layer can be further detailed into strata and sub-strata, when needed. This division into strata helps fine-grain annotations at a Sub-Layer. Besides, it can be used to determine the degree of compulsoriness for each Stratum, as in EAGLES (1996a; 1996b) recommendations and guidelines.

Thus, after describing the hypotheses and the main pillars underlying the OntoLingAnnot model, we present the ontologies included in our model.

OntoLingAnnot's Ontologies: an Overview

OntoLingAnnot's ontologies are the result of a thorough evaluation and extension of OntoTag's ontologies (Aguado de Cea *et al.* 2004a; 2004b). More precisely, the OntoLingAnnot model reuses, restructures and extends the morphosyntactic and the syntactic modules of OntoTag's ontologies, and provides some new modules that formalize the semantic, the discourse and the pragmatic levels.
Besides, OntoLingAnnot's ontologies also inherit from OntoTag's ontologies the distribution of the concepts and terms associated to each

linguistic level among the four main ontologies, namely (1) the Linguistic Level Ontology (LLO), (2) the Linguistic Unit Ontology (LUO), (3) the Linguistic Attribute Ontology (LAO), and (4) the Linguistic Value Ontology (LVO). A fifth main ontology had to be added to these, *i.e.*, the Linguistic Relationship Ontology (LRO), in order to formalize the different kind of relations holding at the linguistic levels included in the OntoLingAnnot model. Finally, the OntoLingAnnot model reuses the Integration Ontology (IO) of OntoTag, which had to be extended and adapted to cover the new levels, layers and strata contemplated in OntoLingAnnot but not in OntoTag. So, briefly, OntoLingAnnot's ontologies can be viewed in fact as a network of six different ontologies. They are described in more detail in the following paragraphs.

Table 2. Levels, layers and sub-layers of the OntoLingAnnot model.

LAYER → / LEVEL ↓	CONSTITUENT UNIT RECOGNITION (SEGMENTATION)	CONSTITUENT UNIT PARADIGMATIC LABELLING	SYNTAGMATIC RELATION IDENTIFICATION	SYNTAGMATIC RELATION LABELLING	RESULTING UNIT
MORPHOLOGY	Sub-Layer 1.1: Morph Recognition (Morphological Segmentation)	Sub-Layer 1.2: Morph Paradigmatic Labelling	Sub-Layer 1.3: Word Formation Relation Identification	Sub-Layer 1.4: Word Formation Relation Labelling	Sub-Layer 1.5: Word Sub-Layer
SYNTAX	Sub-Layer 2.1: Syntactic Unit Recognition (Syntactic Segmentation)	Sub-Layer 2.2: Syntactic Unit Paradigmatic Labelling	Sub-Layer 2.3: Syntactic Relation Identification	Sub-Layer 2.4: Syntactic Relation Labelling	Sub-Layer 2.5: Aggregated Syntactic Unit Sub-Layer
SEMANTICS	Sub-Layer 3.1: Semantic Unit Recognition (Semantic Segmentation)	Sub-Layer 3.2: Semantic Unit Paradigmatic Labelling	Sub-Layer 3.3: Proposition Formation Relation Identification (Predication Identification)	Sub-Layer 3.4: Proposition Formation Relation Labelling (Predication Labelling)	Sub-Layer 3.5: Proposition Sub-Layer
DISCOURSE	Sub-Layer 4.1: Discourse Unit Recognition (Discourse Segmentation)	Sub-Layer 4.2: Discourse Unit Paradigmatic Labelling	Sub-Layer 4.3: Discourse Relation Identification	Sub-Layer 4.4: Discourse Relation Labelling	Sub-Layer 4.5: Macroproposition Sub-Layer
PRAGMATICS	Sub-Layer 5.1: Pragmatic Unit Recognition (Pragmatic Segmentation)	Sub-Layer 5.2: Pragmatic Unit Paradigmatic Labelling	Sub-Layer 5.3: Pragmatic Relation Identification	Sub-Layer 5.4: Pragmatic Relation Labelling	Sub-Layer 5.5: Pragmateme Sub-Layer

First, the LLO captures the stratification for linguistic annotation presented in the previous sections. As discussed above, it extends the

criteria proposed in EAGLES (1996b) for the decomposition into layers of the Syntactic Level to the rest of linguistic levels. Thus, it contains the formalisation of the terminology associated to the (sub-)layers and (sub-)strata related to the levels shown in Table 1.

Second, the LUO, the LAO and the LVO, altogether, formalize the linguistic phenomena and the terminology associated to these levels. In particular, the LUO includes all the units (categories) already identified for Morphology, Syntax, Semantics, Discourse and Pragmatics; the LAO includes the set of attributes associated to these units; and the LVO accounts for the possible values of these attributes. All these units, attributes and values have been linked by suitable axioms, distributed amongst the LUO, the LAO and the LVO. Thus, distributing the linguistic terms and concepts associated to linguistic phenomena into three ontologies facilitates the annotation of the corresponding phenomena by means of triples category-attribute-value, as promoted by EAGLES (1996a) and LAF/GrAF (ISO 2008a). In fact, these three ontologies formalize the EAGLES and MAF recommendations for morpho-syntactic (EAGLES 1996a; ISO 2008c) and syntactic annotation (EAGLES 1996b), and extend them to the semantic, the discourse and the pragmatic levels. The axioms mentioned above, which link the units in the LUO, the attributes in the LAO and the values in the LVO, constrain the way in which they can be combined and put together to make up any of these triples. This organisation and formalisation also avoids redundancy, since several values (such as SINGULAR or MASCULINE) are shared by several attributes (such as NUMBER and POSSESSOR NUMBER, or GENDER and POSSESSOR GENDER) and also a number of attributes are shared by a number of units (such as GENDER or NUMBER themselves, which are shared by NOUN, VERB, ADJECTIVE, etc.).

Third, the LRO formalizes the different relationships that can hold between linguistic units. Although this ontology was not present in the original OntoTag's ontologies, it was considered absolutely necessary to capture the knowledge associated to Discourse and Pragmatics.
Finally, a sort of upper-level (or knowledge representation) ontology, the IO, links the rest of the ontologies together, and describes the main

relations between the concepts in the other five ontologies already described.

Due to the high number of terms these ontologies include, they have been divided into modules, each one corresponding to a different linguistic level formalized within the OntoLingAnnot model. Thus, OntoLingAnnot's ontologies constitute a networked and modularized set of ontologies, suitable for the annotation of most of the levels of linguistic description.

Each of the modules of OntoLingAnnot's ontologies has been developed following the METHONTOLOGY (Gómez-Pérez *et al.* 2004) methodology, and implemented within the WebODE[3] platform for ontology development. An XML, RDF(S) or an OWL version of the ontologies can be exported from the platform anytime, on the fly, for its application and reuse.

To conclude this section, OntoLingAnnot's ontologies are being evaluated by means of the OntoLing Annotizer tool (Montalvo-Martínez 2009), which reuses and extends AKTive Media[4], an ontology-based annotation tool developed within the Natural Language Processing Group of the University of Sheffield.

In the following sections, we show, as a way of example, how the terminology associated to discourse-related annotation was modelled within these ontologies.

Discourse Units in the LUO

In order to formalize discourse coherence, three different classes (units, in this case) and their corresponding subclasses had to be included in the LUO. These three classes are Proposition, Discourse Functional Unit (or DFU), and Macroproposition. They were elicited mainly from Mann and Thompson (1988), Van Dijk (1997), Hovy and Maier (1995),

3 http://webode.dia.fi.upm.es/webODE/ (accessed 9 May 2010).
4 http://www.dcs.shef.ac.uk/~ajay/html/cresearch.html (accessed 9 May 2010).

Romera (2004), Palmer *et al.* (2005) and Prasad *et al.* (2008). These works show that (1) discourse coherence can be achieved mainly by the coherence relations that hold between a particular type of discourse units, which act as the discourse bricks with which discourse is built; (2) these coherence relations are signalled or realized by means of another particular type of discourse units, which act as the discourse plaster with which the discourse bricks are stuck together; and (3) the establishment of these coherence relations might build up a new type of compound and higher-level discourse units (which, keeping the simile, are the walls made up of the discourse bricks and stuck together by the discourse plaster).

The type of discourse units between which coherence relations are established (*i.e.* the discourse bricks) are the *propositions* (or *discourse propositional units*). Accordingly, propositions (Palmer *et al.* 2005; Prasad *et al.* 2008) constitute the interface between the semantic and the discourse-related levels. They consist of the different semantic units included in sentences (or *clauses*) plus the semantic relationships holding between them. These semantic relationships are, basically, the semantic role that each of these semantic units plays within the Proposition or, in other words, the predicate-argument structure of the corresponding Sentence (or Clause). We distinguish two different classes of propositions, namely General Proposition and Instanced Proposition. The class Instanced Proposition encompasses all those *instances of* Proposition that contain at least an instanced semantic element, that is, a semantic element that refers to a particular *Instance-Of* a concept. Conversely, the class General Proposition encompasses all those *instances of* Proposition that do not contain any instanced semantic element, that is, no semantic element of the Proposition refers to a particular *Instance-Of* a concept.

The following examples show why Proposition was subclassified this way. Consider the formal differences that exist between the underlying meanings (that is, the resultant propositions) of the sentences 'Dogs can run' and 'Pluto is a cartoon'. In the first Proposition no semantic element has been instanced. In fact, a generic statement is being made: a *property* (being able to run) is attributed to a class of individuals

(dogs). In the second Proposition, on the contrary, 'Pluto' is an instanced semantic element and, hence, in this Proposition, a particular statement is being made about an individual. More concretely, the individual 'Pluto', which could be considered an *Instance-Of* Dog in some ontology, is being attributed the *property* 'being a cartoon', which cannot be attributed to all dogs in general. Consequently, the first Proposition is a General Proposition, whereas the second one is an Instanced Proposition. If a mathematic-logical formalism (first-order logic, for example) were used in order to represent each Proposition, 'Pluto is a cartoon' could be represented by the logical clause:

Cartoon(Pluto)

where *Pluto* is clearly a constant of the formalism; analogously, 'Dogs can run' could be represented by the logical clause:

"x (Dog (x) Æ CanRun(x))

where no constant appears. Actually, in this logical clause, there are only two logical predicates, *Dog* and *CanRun*, and a universally quantified variable, x.

Therefore, using a suitable inference engine, from the latter Proposition we could be infer, for example, that 'Lassie can run', provided that 'Lassie' is an *InstanceOf* Dog in some ontology. On the contrary, no further information can be inferred from the other (Instanced) Proposition. This different way of representing and using this type of propositions is the main reason that motivated the distinction between general and instanced propositions. Some other *InstancesOf* Proposition are shown in Example 1 and in Figure 1.

Example 1. An excerpt of a short dialog

Person A: 'Excuse me, can you tell me where the nearest police station is, please?'

Person B: 'Go down the street and turn left at the traffic lights. I think it's on the right.'

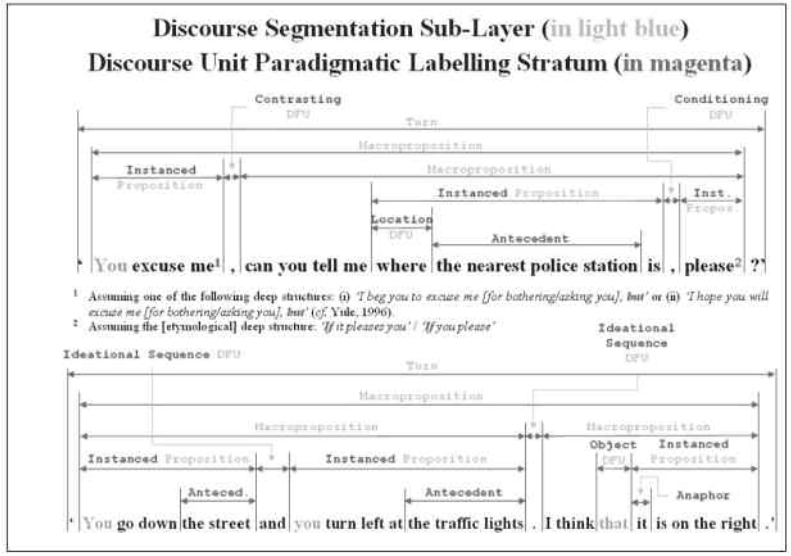
The other particular type of discourse units, which realize the coherence relations (*i.e.* the ones which constitute the discourse plaster), are termed ***DFUs*** (Romera 2004). Hence, a Discourse Functional Unit signals a ***Discourse Coherence Relation.*** In other words, a **Discourse Functional Unit** (**DFU**) is the marker of a discourse relation that holds between two (adjacent) propositions. More specifically, DFUs are "the lexical signals which make coherence relations explicit in surface text" (*cf.* Knott and Sanders 1998). DFUs can be easily identified by means of syntactic criteria: they are incarnated by conjunctions, conjunctive adverb(ial)s and punctuation marks, in appositions (*cf.* Mann and Thompson 1998; Halliday 1994)[5] . This is the easiest criterion that can be applied, provided that the annotation at the Discourse Level is performed after the syntactic annotation has been carried out. Some examples of these types of units are shown in Figure 1, such as the comma after 'Excuse me', or the one before 'please' (assuming the deep structures mentioned in the figure[6]).

The last type of discourse units used to formalize discourse coherence (the compound or higher-level ones, *i.e.*, the discourse walls) are the ***macropropositions***. Therefore, a **Macroproposition** (Van Dijk 1997) is the Linguistic Unit that results from the aggregation of some interrelated propositions. Macropropositions can also be regarded as complex discourse units that stand on the Discourse-Pragmatics interface and that serve as unitary blocks at the Pragmatic Level. As shown in Figure 1, the whole question of **Person A** in Example 1 can be regarded as an *InstanceOf* Macroproposition, whereas the answer of **Person B** can be viewed as a compound Macroproposition. This compound Macroproposition consists of other two consecutive macropropositions, associated to 'Go down the street and turn left at the traffic lights' and 'I think it's on the right', respectively.

5 Considering punctuation marks as DFUs might not be consensual at all. However, they can be viewed as an orthographic mark that points out the ellipsis or the abbreviation of an actual DFU that is implicit in the discourse. Accordingly, they are treated as such in OntoLingAnnot.

6 Derived from Yule (1996) and http://en.wiktionary.org/wiki/please (accessed 9 May 2010), respectively

Figure 1. An example of discourse annotation using OntoLingAnnot's ontologies (a)



In addition, other two main classes of discourse units have been included in the LUO, namely Discourse Reference Unit (DRU) and Turn. First, **DRUs** formalize the different units involved in anaphoric (or cataphoric) co-references. More specifically, a **Discourse Reference Unit** (**DRU**) is a linguistic unit that contributes to the cohesion of discourse by means of its participation in a co-reference mechanism or relation (such as an Anaphora or a Cataphora). A more restrictive definition, in line with Mitkov (2002) can be stated as follows: DRUs are those aggregations that are a potential Antecedent or a potential Endophor in an Anaphora or a Cataphora. In Example 1, 'the nearest police station' is the antecedent of 'it' ('I think it's on the right'), which is an anaphoric reference to 'the nearest police station'. Therefore, there is an anaphoric relation holding between them and, hence, both of them can be considered *InstancesOf* DRU.

Second, ***turns*** identify the boundaries of each speaker intervention, when a form of dialogue is involved in the discourse. Thus, a **Turn**

(Yule 1996) is a piece of discourse that is ascribable to a single participant in the discourse, within a typical, orderly arrangement, in which participants contribute to the discourse with minimal overlaps and gaps between them[7]. In Example 1, the intervention of **Person A** as a whole constitutes an instance of Turn, and so does the intervention of **Person B.**

Several other terms have been included as concepts in this LUO module. They sub-classify and specialize these five (super-)classes of discourse units, but they have not been discussed here for the sake of space. Some examples of these units are shown in Figure 1.

Discourse Attributes and Values in the LAO and the LVO

First, the discourse-related module of the LAO (dealing with attributes) consists of four concepts, the top-level concept Discourse Attribute, and the concepts Propositional Attribute, DFU Attribute and DRU Attribute, which constitute a *Disjoint-Decomposition* of Discourse Attribute. Whereas discourse attributes can be ascribed to any kind of Discourse Unit, propositional attributes, DFU attributes and DRU attributes can only be ascribed to propositions, DFUs and DRUs (respectively).

Besides, we have also included in the LAO two *InstancesOf* the concept Propositional Attribute (isQuote and hasDiscourseFunction), one *InstancesOf* DFU Attribute (isExtractable), and one *InstancesOf* DRU Attribute (hasHierarchical OrganisationFunction). The meaning and the use of these attributes can be better understood under the light of the values they can take, which are presented next.

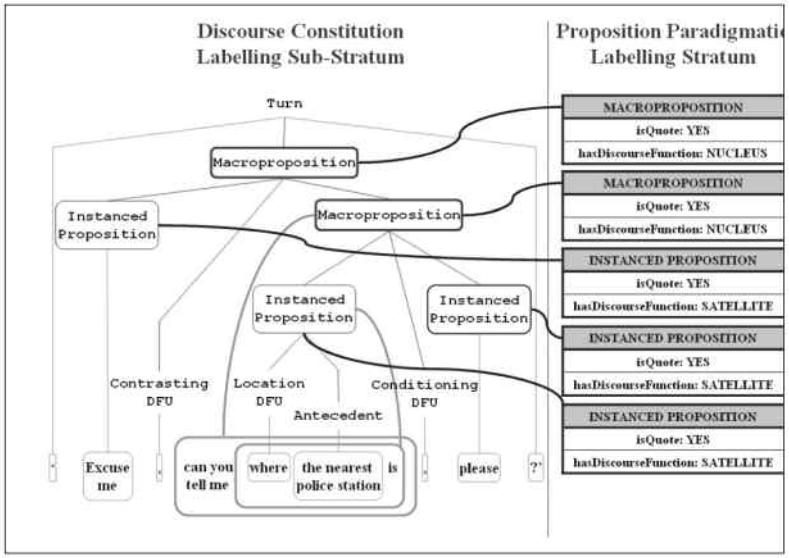
Second, the discourse-related module of the LVO (dealing with values) contains the following concepts, associated to their near-homonyms in the LAO: Discourse Value, Propositional Value, DFU Value, DRU Value, Quoting Value, Discourse Function Value, Extractability Value, and Hierarchical Organisation Function Value.

7 *Cf.* http://www.sil.org/linguistics/GlossaryOfLinguisticTerms/WhatIsATurn.htm (accessed 9 May 2010).

Both Quoting Value (*cf.* Yule 1996) and Extractability Value (Knott and Sanders, 1998) are *SubclassOf* the concept Boolean Value and, hence, they have two unique instances associated (TRUE and FALSE). Additionally, we have included in the LVO two *Instances-Of* the concept Discourse Function Value[8] (NUCLEUS and SATELLITE) and three *InstancesOf* Hierarchical Organisation Function Value[9] (ENCAPSULATING, PROSPECTING and NOT_APPLICABLE).

No other type or example of Discourse Attribute (or Discourse Value) has been found in these sources. Some examples of these attributes and their values have been included in Figure 2.

Figure 2. An example of discourse annotation using OntoLingAnnot's ontologies (b)



Discourse Relations in the LRO

As far as the concepts in the LRO are concerned, most of them have been

8 *Taken from Mann and Thompson (1988).*

9 *Taken from Álvarez de Mon y Rego (2001).*

extracted or derived from a taxonomy included in Hovy and Maier (1995). This taxonomy is the result of an extensive study of nearly other thirty different sources, which were combined and interlinked. Therefore, both the concepts and the taxonomy have been reused in the LRO. Nevertheless, when specifying the core sub-concepts of Discourse Coherence Relation, we have followed Mann and Thompson (1988) even though the underlying taxonomy was not present in these authors.

Accordingly, the main discourse-related concepts included in the LRO are the following:

- Discourse Coordination, Discourse Subordination and Discourse Constitution, which hold between two (or more) discourse units (some examples are presented in Figure 2);
- Presentational Cohesion Relation and Ideational Coherence Relation (whose main subclasses are: Cause-Result Relation, Elaboration, Circumstance Relation, Ideational Sequence, Comparative Relation, and General Condition Relation), which hold between two (or more) propositions;
- Discourse Role (or Discourse Function) and its *subclasses*: Cause-Result Role, Elaboration Role, Circumstantial Role, Ideational Sequence Role, Comparative Role, General Condition Role, and Other Discourse Role, which characterize the meaning contribution of a Proposition to a given discourse; and
- Endophora and its two *subclasses*, namely Anaphora and Cataphora, which hold between two DRUs.

Most of these main classes have been sub-specified by means of suitable *subclasses* and characterized by several (ontological) attributes, but they have not been included here for space restrictions.

The Discourse Level in the LLO

The linguistic level concerning discourse has been formalized in the LLO by means of a *SubclassOf* the concept Linguistic Level, that is, the

concept Discourse Level. This concept has been decomposed into the Discourse Unit Recognition Sub-Layer (also referred to as the Discourse Segmentation Sub-Layer – see Figure 1), the Discourse Unit Paradigmatic Labelling Sub-Layer, the Discourse Relation Identification Sub-Layer, the Discourse Relation Labelling Sub-Layer and the Macroproposition Sub-Layer. All of these sub-layers are *PartOf* the concept Discourse Level (see also Table 2).

The following concepts are PartOf the Discourse Unit Paradigmatic Labelling Sub-Layer in the LLO (shown in Figure 1): the Proposition Paradigmatic Labelling Stratum (shown in detail in Figure 2), the Discourse Functional Unit Paradigmatic Labelling Stratum, the Discourse Reference Unit Paradigmatic Labelling Stratum and the Turn Paradigmatic Labelling Stratum. Each of these strata deals with (1) the subclassification of its corresponding type of Discourse Unit according to the concepts included in the LUO; and (2) its characterisation according to the discourse attributes and values included in the LAO and the LVO, respectively.

In turn, the following strata are *PartOf* the Discourse Relation Labelling Sub-Layer: the Discourse Composition Labelling Stratum, the Discourse Coherence/Cohesion Relation Labelling Stratum, the Discourse Role Labelling Stratum and the Endophora Labelling Stratum. Each of them deals with the subclassification and characterisation of its corresponding type of Discourse Relation according to the concepts, attributes and values included in the LRO.

Related Work and Contributions of the Present Research

As far as the studies on discourse are concerned, the main approaches followed thus far lack some degree of generality. For example, Mann and Thompson (1988), Hovy and Maier (1995), Romera (2004) and Prévot (2004) focus on the study and classification of coherence relations, and indeed they provide a complete explanation of discourse coherence and cohesion phenomena. However, they fail to address other Discourse phenomena, such as anaphoric references and anaphora resolution (Mitkov 2002) and other discourse devices, such as

encapsulation or prospection (Álvarez de Mon y Rego 2001), which also contribute significantly to the coherence and cohesion of discourse. Besides, the main studies mentioned above dealing with coherence and cohesion (1) do not address the way macropropositions and discourse superstructures (Van Dijk 1997) are created and used, and (2) do not provide a set of additional syntactic features that can help characterize and subclassify systematically discourse coherence relations (or, equivalently, DFUs), which is precisely the problem addressed by Knott and Sanders (1998). The present research is the result of a thorough study and analysis of the works mentioned previously and, hence, one of its main contributions is that it offers a comprehensive, global view of discourse-related phenomena and terminology.

Regarding ontologies, most of the ontology-based approaches to (Computational) Linguistics mentioned in the Introduction (Mahesh and Nirenburg 1995; Farrar and Bateman 2005; Schalley and Zaefferer 2007; Farrar 2007; Java *et al.* 2007, Chiarcos 2008; Buyko *et al.* 2008; Buitelaar *et al.* 2009; Mairal Usón and Periñán Pascual 2009; OntoNotes 2010; GOLD 2010) focus on the discourse level. Some others, such as Niekrasz and Purver (2005) have modelled the discourse knowledge related to particular phenomena also included in OntoLingAnnot's ontologies, such as the discourse associated to dialogues and turns, but they do not aim at a comprehensive and global modelling of the discourse phenomena and terminology. Thus, the ontological modules presented here can be considered the first attempt to structure, formalize and model coherently this linguistic level in an ontology.

As for the field of linguistic annotation and its standardisation, only two of the standard proposals under development seem to tackle the phenomena and the terminology presented here, namely the Data Category Registry (DCR[10]) and the Semantic Annotation Framework – Dialogue Acts (SemAF-Dacts) drafts. First, almost none of the terms included in OntoLingAnnot's ontologies relating discourse (or

10 http://www.isocat.org/

pragmatics) have been included in the Data Category Registry (DCR) yet. Indeed, the DCR includes a significant number of terms relating morphosyntax, syntax and semantics up to date, but other linguistic levels, such as the two mentioned previously have not been considered yet. Thus, the ontological modelling of discourse and pragmatics phenomena and the terminology contained in OntoLingAnnot's ontologies could be a most suitable starting point for the inclusion of Discourse-related and pragmatic categories and terminology into the DCR.

Second, the SemAF-Dacts standard draft (ISO 2009) deals currently only with dialogue segmentation and turn definition and management. It also deals with the standardisation of speech acts (re-termed as dialogue acts) which traditionally have been studied as pragmatic phenomena (and so they have been considered in the OntoLingAnnot model). The modelling of coherence relations and referential entities and relations (already modelled within OntoLingAnnot's ontologies) is envisaged but still pending.

In addition, we should mention that the general uniform view of linguistic annotation included in the OntoLingAnnot model, based on their subdivision into (sub-)layers and (sub-)strata, conjoins and encompasses all the main levels of linguistic annotation. Besides, it also allows for a rather flexible instantiation of this general scheme at each level being annotated. This could be regarded as a perfect complement to the LAF/GrAF (ISO 2008a) standard proposal, which only defines an abstract framework for linguistic annotation, which is too independent from the level and the phenomena being annotated.

Further Work

As commented, the modules of OntoLingAnnot's ontologies have been developed together with the OntoLingAnnot annotation model. This model is being evaluated by means of the OntoLingAnnot Annotizer tool (Montalvo-Martínez 2009), which reuses and extends AKTive Media , from the Natural Language Processing Group of the University of Sheffield. Thus, OntoLingAnnot's ontologies are currently under

evaluation as for the suitability of their content and its formalization. This evaluation phase is expected to help find inconsistencies and gaps in the representation of both the terms and their associated linguistic phenomena considered in OntoLingAnnot so far.

This should lead to an (expectedly minor) update of these ontological modules and a new evaluation phase, possibly with a different annotation tool, in order to measure, at least to some extent, their degree of interoperability and reusability.

Conclusions

In this paper, we have presented the ontological modules dealing with the discourse level of the OntoLingAnnot (linguistic) annotation model. These ontological modules formalize the different discourse units, features and relationships identified in the literature so far, as well as a coherent distribution and structuring of these discourse terms for their use in (discourse) annotation.

Briefly, the modules relating discourse in OntoLingAnnot's ontologies contain

- 68 discourse units (in the LUO);
- 8 discourse attributes – 4 concepts and 4 instances (in the LAO);
- 15 discourse values – 8 concepts and 7 instances (in the LVO);
- 106 discourse relations – 70 concepts and 36 instances (in the LRO); and
- 24 discourse concepts relating levels, layers and strata (in the LLO).

They amount to 221 discourse-related terms in OntoLingAnnot's ontologies: 174 concepts and 47 instances. There are also several other ontological terms concerning discourse (attributes, *SubclassOf, PartOf* and *ad hoc relations, rules* and *axioms*) but, as mentioned, they are not

11 http://www.dcs.shef.ac.uk/~ajay/html/cresearch.html (accessed 9 May 2010)

discussed here for space reasons.
As shown in the previous sections, thus far, this is the first global ontological conceptualization of linguistic annotation in general and of discourse (Annotation) in particular and, hence, it is an important contribution per se to the areas of Ontological Engineering, Discourse and Linguistic Annotation. Besides, no other discourse model accounts globally and coherently for such a number of discourse terms as those included in OntoLingAnnot's ontologies, which is another important contribution to the areas aforementioned.

Acknowledgements

We would like to thank Mr. Martín Montalvo for his help at the implementation of the OntoLing Annotizer tool mentioned in this paper. This work has also been partly supported by the European Project *Monnet* (FP7-248458).

References

Aguado de Cea, G., Gómez-Pérez, A., Álvarez de Mon y Rego, I. and Pareja-Lora, A. (2004a) 'OntoTag's Linguistic Ontologies: Improving Semantic Web Annotations for a Better Language Understanding in Machines', in *Proceedings of ITCC 2004,* Las Vegas: IEEE Publications, 2: 124-128.

Aguado de Cea, G., Gómez-Pérez, A., Álvarez de Mon y Rego, I. and Pareja-Lora, A. (2004b) 'OntoTag's linguistic ontologies: Enhancing higher level and Semantic Web annotations', in *Proceedings of LREC 2004,* Lisboa, Portugal, VI: 1905–1908.

Álvarez de Mon y Rego, I. (2001) 'Encapsulation and prospection in written scientific English', in *Estudios ingleses de la Universidad Complutense* Madrid: Universidad Complutense, 2001, 9: 81-102. Available online at [http://dialnet.unirioja.es/servlet/articulo?codigo=174471] (accessed 9 May 2010).

Arrizabalaga-Hernández, F.J. (2004) *OntoTagger: Herramienta de anotación lingüístico-ontológica,* [OntoTagger: A linguistic and ontological tool], M.Sc. Thesis, Universidad Politécnica de Madrid.

Borst, W.N. (1997) *Construction of Engineering Ontologies,* PhD Thesis, University of Twente, Enschede.

Buitelaar, P., Cimiano, P., Haase, P. and Sintek, M. (2009) 'Towards Linguistically Grounded Ontologies', in *The Semantic Web: Research and Applications* (Lecture

Notes in Computer Science, vol. 5554/2009), Berlin/Heidelberg: Springer.

Buyko, E., Chiarcos, C. and Pareja-Lora, A. (2008) 'Ontology-Based Interface Specifications for an NLP pipeline architecture', in *Proceedings of LREC 2008.* Marrakech, Morocco.

Chiarcos, C. (2008) 'An Ontology of Linguistic Annotations', in *LDV Forum (GLDV-Journal for Computational Linguistics and Language Technology),* 23(1): 1-16.

Crystal, D. (1992) *A Dictionary of Linguistics and Phonetics* (3rd edition), Oxford: Blackwell.

EAGLES (1996a) *EAGLES: Recommendations for the Morphosyntactic Annotation of Corpora.* EAGLES Document EAG—TCWG—MAC/R.

EAGLES (1996b) *EAGLES: Recommendations for the Syntactic Annotation of Corpora.* EAGLES Document EAG—TCWG—SASG/1.8.

Farrar, S. (2007) 'Using 'Ontolinguistics' for language description', in Schalley, A.C. and Zaefferer, D. (eds) *Ontolinguistics: how ontolinguistic status shapes the linguistic coding of concepts.* Berlin/New York: Mouton de Gruyter.

Bateman, J.A. and Farrar, S. (2005) *OntoSpace Project Reports - DeliverableD3 - Linguistic Ontology Baseline,* University of Bremen, Germany. Available online at [http://www.ontospace.uni-bremen.de/pub/FarrarBateman05-i1-d3.pdf] (accessed 21 April 2010).

GOLD (2010) Available online at [http://linguistics-ontology.org/] (accessed 9 May 2010).

Gómez-Pérez, A., Corcho, Ó., and Fernández-López, M. (2004) *Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web,* London: Springer-Verlag London Limited.

Gruber, T.R. (1993) 'A Translation Approach to Portable Ontologies', in *Journal on Knowledge Acquisition,* 5(2): 199-220.

Hovy, E.H. and Maier. E. (1995) *Parsimonious or Profligate: How Many and Which Discourse Structure Relations?* Available online at [http://www.isi.edu /naturallanguage/people/hovy/papers/93discproc.pdf] (accessed 9 May 2010).

Knott, A. and Sanders, T. (1998) 'The Classification of Coherence Relations and their Linguistic Markers: An Exploration of Two Languages', in *Journal of Pragmatics* 30:135-175.

ISO (2006) *Language resource management –Syntactic Annotation Framework (SynAF).* ISO/TC 37/SC 4, ISO/WD 24615.

ISO (2007) *Language resource management – Semantic annotation framework (SemAF) — Part 1: Time and events (SemAF/Time).* ISO/TC 37/SC 4 N412 – ISO/CD 24617-1.

ISO (2008a) *Language resource management – Linguistic Annotation Framework (LAF/GrAF).* ISO/TC 37/SC 4 N522 – ISO/CD 24612.

ISO (2008b) *Language resource management – Lexical Markup Framework (LMF).* ISO/TC 37/SC 4 N453 – ISO FDIS 24613:2008.

ISO (2008c) *Language resource management – Morpho-Syntactic Annotation Framework (MAF).* ISO/TC 37/SC 4 N225 – ISO/CD 24611.

ISO (2009) *Language resource management – Semantic annotation framework (SemAF) — Part 2: Dialogue Acts (SemAF-Dacts).* ISO/TC 37/SC 4 N442 rev 05 –ISO/CD 24617-2-2009-10-05.

Java, A., Nirenburg, S. McShane, M. Finin, T., English, J., and Joshi, A. (2007) 'Using a Natural Language Understanding System to Generate Semantic Web Content', *International Journal on Semantic Web and Information Systems* 3(4).

Longacre, R.E. (1996) *The grammar of discourse,* New York: Plenum Press.

Mahesh, K. and Nirenburg, S. (1995) 'A Situated Ontology for Practical NLP', in *Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing,* International Joint Conference on Artificial Intelligence (IJCAI-95), Montreal, Canada.

Mairal Usón, R. and Periñán Pascual, J.C. (2009) 'The anatomy of the lexicon within the framework of an NLP knowledge base', in RESLA: *Revista española de lingüística aplicada,* 22: 217-244.

Mann, W. C. and Thompson, S.A. (1988) 'Rhetorical Structure Theory: Toward a Functional Theory of Text Organization', in *Text* 8(3): 243-281.

Mitkov, R. (2002) *Anaphora Resolution,* London: Longman.

Montalvo-Martínez, M. (2009) *OntoLing Annotizer: Una herramienta de ayuda a la anotación* [OntoLing Annotizer: An annotation-aiding tool], M.Sc. Thesis, Universidad Complutense de Madrid.

Niekrasz, J. and Purver, M. (2005) 'A Multimodal Discourse Ontology for Meeting Understanding', in *Machine Learning for Multimodal Interaction: Second International Workshop (MLMI 2005).* Edinburgh: Springer Verlag. Available online at [http://godel.stanford.edu/twiki/pub/Public/SemlabPublications/mlmi05.pdf] (accessed 9 May 2010).

OntoNotes (2010) Available online at [http://www.bbn.com/ontonotes/] (accessed 9 May 2010).

Palmer, M., Gildea, D., and Kingsbury, P. (2005) 'The Proposition Bank: an annotated corpus of semantic roles', in *Computational Linguistics,* 31(1).

Polanyi, L. (1988) 'A formal Model of the Structure of Discourse', in *Journal of Pragmatics* 12: 601-638.

Prasad, R., Dinesh, N., Lee, L., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008) 'The Penn Discourse Treebank 2.0', in *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008),* Marrakech, Morocco.

Prévot, L. (2004) *Structures sémantiques et pragmatiques pour la modélisation de la cohérence dans des dialogues finalisés* [Semantic and pragmatic structures for modelling coherence in finalized dialogues]. Thèse de doctorat de l'université Paul Sabatier, Toulouse, France.

Romera, M. (2004) *Discourse functional units: the expression of coherence relations in spoken Spanish,* Munich: LINCOM.

Saussure, F. de (1916/1983) *Course in General Linguistics,* trans. by Roy Harris (1983), London: Duckworth.

Schalley, A.C. and Zaefferer, D. (2007) 'Ontolinguistics – An outline', in Schalley, A.C. and Zaefferer, D. (eds) *Ontolinguistics: how ontolinguistic status shapes the linguistic coding of concepts.* Berlin/New York: Mouton de Gruyter.

Schiffrin, D., Tannen, D. and Hamilton, H.E. (eds) (2001) *Handbook of Discourse Analysis.* Oxford: Blackwell.

Van Dijk, T.A. (ed) (1997) *Discourse Studies.* 2 vols. London: Sage.

Yule, G. (1996) *Pragmatics* (Oxford Introductions to Language Study). Oxford: Oxford University Press.