Parmenides: an opportunity for ISO TC37 SC4?

Fabio Rinaldi[1], James Dowdall[1], Michael Hess[1], Kaarel Kaljurand[1], Andreas Persidis[2],
Babis Theodoulidis[3], Bill Black[3], John McNaught[3], Haralampos Karanikas[3], Argyris Vasilakopoulos[3],
Kelly Zervanou[3], Luc Bernard[3], Gian Piero Zarri[4], Hilbert Bruins Slot[5], Chris van der Touw[5],
Margaret Daniel-King[6], Nancy Underwood[6], Agnes Lisowska[6], Lonneke van der Plas[6],
Veronique Sauron[6], Myra Spiliopoulou[7], Marko Brunzel[7], Jeremy Ellman[8],
Giorgos Orphanos[9], Thomas Mavroudakis[10], Spiros Taraviras[10]

Abstract

Despite the many initiatives in recent years aimed at creating Language Engineering standards, it is often the case that different projects use different approaches and often define their own standards. Even within the same project it often happens that different tools will require different ways to represent their linguistic data.

In a recently started EU project focusing on the integration of Information Extraction and Data Mining techniques, we aim at avoiding the problem of incompatibility among different tools by defining a Common Annotation Scheme internal to the project. However, when the project was started (Sep 2002) we were unaware of the standardization effort of ISO TC37/SC4, and so we commenced once again trying to define our own schema. Fortunately, as this work is still at an early stage (the project will last till 2005) it is still possible to redirect it in a way that it will be compatible with the standardization work of ISO. In this paper we describe the status of the work in the project and explore possible synergies with the work in ISO TC37 SC4.

[1] [1]Institute of Computational Linguistics, University of Zurich, Switzerland; [2]Biovista, Athens, Greece; [3]Centre for Research in Information Management, UMIST, Manchester, UK; [4]CNRS, Paris, France; [5]Unilever Research and Development, Vlaardingen, The Netherlands; [6]TIM/ISSCO, University of Geneva, Switzerland; [7]Uni Magdeburg, Germany; [8]Wordmap Ltd., Bath, UK; [9]Neurosoft, Athens, Greece; [10]The Greek Ministry of National Defense, Athens, Greece

1 Introduction

It is by now widely accepted that some W3C standards (such as XML and RDF) provide a convenient and practical framework for the creation of field-specific markup languages (e.g. MathML, VoiceXML). However XML provides only a common "alphabet" for interchange among tools, the steps that need to be taken before there is any real sharing are still many (just as many human languages share the same alphabets, that does not mean that they can be mutually intelligible). The necessary step to achieve mutual understanding in Language Resources is to create a common data model.

The existence of a standard brings many other advantages, like the ability to automatically compare the results of different tools which provide the same functionality, from the very basic (e.g. tokenization) to the most complex (e.g. discourse representation). Some of the NIST-supported competitive evaluations (e.g. MUC) greatly benefited by the existence of scoring tools, which could automatically compare the results of each participant against a gold standard. The creation of such tools (and their effectiveness) was possible only because the organizing institute had pre-defined and "imposed" upon the participants the annotation scheme. However, that sort of "brute force" approach might not always produce the best results. It is important to involve the community in the definition of such standards at an early stage, so that all the possible concerns can be met and a wider acceptance can be achieved.

Another clear benefit of agreed standards is that they will increase interoperability among different tools. It is not enough to have publicly available APIs to ensure that different tools can be integrated. In fact, if their representation languages (their "data vocabulary") are too divergent, no integration will be possible (or at least it will require a considerable mapping effort). For all the above reasons we enthusiastically support any concertation work, aimed at establishing common foundations for the field.

In a recently started EU project ("Parmenides") focusing on the integration of Information Extraction and Data Mining techniques (for Text Mining) we aim at avoiding the problem of incompatibility among different tools by defining a Common Annotation Scheme internal to the project. However, when the project was started (Sep 2002) we were unaware of the standardization effort of ISO TC37 SC4, and so we commenced once again trying to define our own schema. Fortunately, as this work is still at an early stage (the project will last till 2005) it is still possible to redirect it in a way that it will be compatible with the standardization work of ISO.

In this paper we will describe the approach followed so far in the definition of the Parmenides Common Annotation Scheme, even if its relation with ISO is still only superficial. In the forthcoming months our intention is to explore possible synergies between our work and the current initiatives in ISO TC37 SC4, with the aim to get at a Parmenides annotation scheme which is conformant to the approach currently discussed in the standardization committee.

2 The Parmenides *Lingua Franca*

In this section we will describe the XML-based annotation scheme proposed for the Parmenides project. In general terms the project is concerned with organisational knowledge management, specifically, by developing an ontology driven systematic approach to integrating the entire process of information gathering, processing and analysis.

The annotation scheme is intended to work as the projects' *lingua franca*: all the modules will be required to be able to accept as input and generate as output documents conformant to the (agreed) annotation scheme. The specification will be used to create data-level compatibility among all the tools involved in the project.

Each tool might choose to use or ignore part of the information defined by the markup: some information might not yet be available at a given stage of processing or might not be required by the next module. Facilities will need to be provided for filtering annotations according to a simple configuration file. This is in fact one of the advantages of using XML: many readily available off-the-shelf tools can be used for parsing and filtering the XML annotations, according to the needs of each module.

The annotation scheme will be formally defined by a DTD and an equivalent XML schema definition. Ideally the schema should remain flexible enough to allow later additional entities when and if they are needed. However the present document has only an illustrative purpose, in particular the set of annotation elements introduced needs to be further expanded and the attributes of all elements need to be verified.

There are a number of simplifications which have been taken in this document with the purpose of keeping the annotation scheme as simple as possible, however they might be put into question and more complex approaches might be required. For instance we assume that we will be able to identify a unique set of tags, suitable for all the applications. If this proves to be incorrect, a possible way to deal with the problem is the use of XML namespaces. Our assumptions allow us (for the moment) to keep all XML elements in the same namespace (and therefore ignore the issue altogether).

2.1 Corpus Development

The annotation scheme will be used to create a development corpus - a representative sample of the domain, provided by the users as typical of the documents they manually process daily. In this phase, the documents are annotated by domain experts for the information of interest. This provides the benchmark against which algorithms can be developed and tested to automate extraction as far as possible.

Of primary importance to the annotation process is the consolidation of the "information of interest", the text determined as the target of the Information Extraction modules. Given the projects' goals, this target will be both diverse and complex necessitating clarity and consensus.

2.2 Sources Used for this Document

Parmenides aims at using consolidated Information Extraction techniques, such as Named Entity Extraction, and therefore this work builds upon well-known approaches, such as the Named Entity annotation scheme from MUC7 (Chinchor, 1997). Crucially, attention will be paid to temporal annotations, with the aim of using extracted temporal information for detection of trends (using Data Mining techniques). Therefore we have investigated all the recently developed approaches to such a problem, and have decided for the adoption of the TERQAS tagset (Ingria and Pustejovsky, 2002; Pustejovsky et al., 2002).

Other sources that have been considered include the GENIA tagset (GENIA, 2003), TEI (TEI Consortium, 2003) and the GDA[1] tagset. The list of entities introduced so far is by no means complete

[1]http://www.i-content.org/GDA/tagset.html

but serves as the starting point, upon which to build a picture of the domains from information types they contain. The domain of interests (e.g. Biotechnology) are also expected to be terminology-rich and therefore require proper treatment of terminology.

To supplement the examples presented, a complete document has been annotated according to the outlined specification.[2] There are currently three methods of viewing the document which offer differing ways to visualize the annotations. These are all based on transformation of the same XML source document, using XSLT and CSS (and some Javascript for visualization of attributes). For example, the basic view can be seen in figure (1).

3 Levels of Annotation

The set of Parmenides annotations is organized into three levels:

- **Structural Annotations**
  Used to define the physical structure of the document, it's organization into head and body, into sections, paragraphs and sentences.[3]
- **Lexical Annotations**
  Associated to a short span of text (smaller than a sentence), and identify lexical units that have some relevance for the Parmenides project.
- **Semantic Annotations**
  Not associated with any specific piece of text and as such could be free-floating within the document, however for the sake of clarity, they will be grouped into a special unit at the end of the document. They refer to lexical annotations via co-referential Ids. They (partially) correspond to what in MUC7 was termed 'Template Elements' and 'Template Relations'.

Structural annotations apply to large text spans, lexical annotations to smaller text spans (sub-sentence). Semantic annotations are not directly linked to a specific text span, however, they are linked to text units by co-referential identifiers.

All annotations are required to have an unique ID and thus will be individually addressable, this allows semantic annotations to point to the lexical annotations to which they correspond. Semantic Annotations themselves are given a unique ID, and therefore can be elements of more complex annotations ("Scenario Template" in MUC parlance).

[2]available at `http://www.ifi.unizh.ch/Parmenides`

[3]Apparently the term 'structure' is used with a different meaning in the ISO documentation, referring to morpho-syntactical structure rather than document structure.

Structural Annotations The structure of the documents will be marked using an intuitively appropriate scheme which may require further adaptations to specific documents. For the moment, the root node is `<ParDoc>` (Parmenides Document) which can contain `<docinfo>`, `<body>`, `<ParAnn>`. The `<docinfo>` might include a title, abstract or summary of the documents contents, author information and creation/release time. The main body of the documents (`<body>`) will be split into sections (`<sec>`) which can themselves contain sections as well as paragraphs (`<para>`). Within the paragraphs all sentences will be identified by the `<sentence>` tag. The Lexical Annotations will (normally) be contained within sentences. The final section of all documents will be `<ParAnn>` (Parmenides Annotations) where all of the semantic annotations that subsume no text are placed. Figure (2) demonstrates the annotation visualization tool displaying the documents structure (using nested boxes).

Lexical Annotations Lexical Annotations are used to mark any text unit (smaller than a sentence), which can be of interest in Parmenides. They include (but are not limited to):

1. Named Entities in the classical MUC sense
2. New domain-specific Named Entities
3. Terms
4. Temporal Expressions
5. Events
6. Descriptive phrases (chunks)

The set of Lexical Annotations described in this document will need to be further expanded to cover all the requirements of the project, e.g. names of products (Acme Arms International's KryoZap (TM) tear gas riot control gun), including e.g. names of drugs (Glycocortex's Siderocephalos).

When visualizating the set of Lexical Tags in a given annotated document, clicking on specific tags displays the attribute values (see figure (3)).

Semantic Annotations The relations that exist between lexical entities are expressed through the semantic annotations. So lexically identified people can be linked to their organisation and job title, if this information is contained in the document (see figure (4)). In terms of temporal annotations, it is the explicit time references and events which are identified lexically, the temporal relations are then captured through the range of semantic tags.

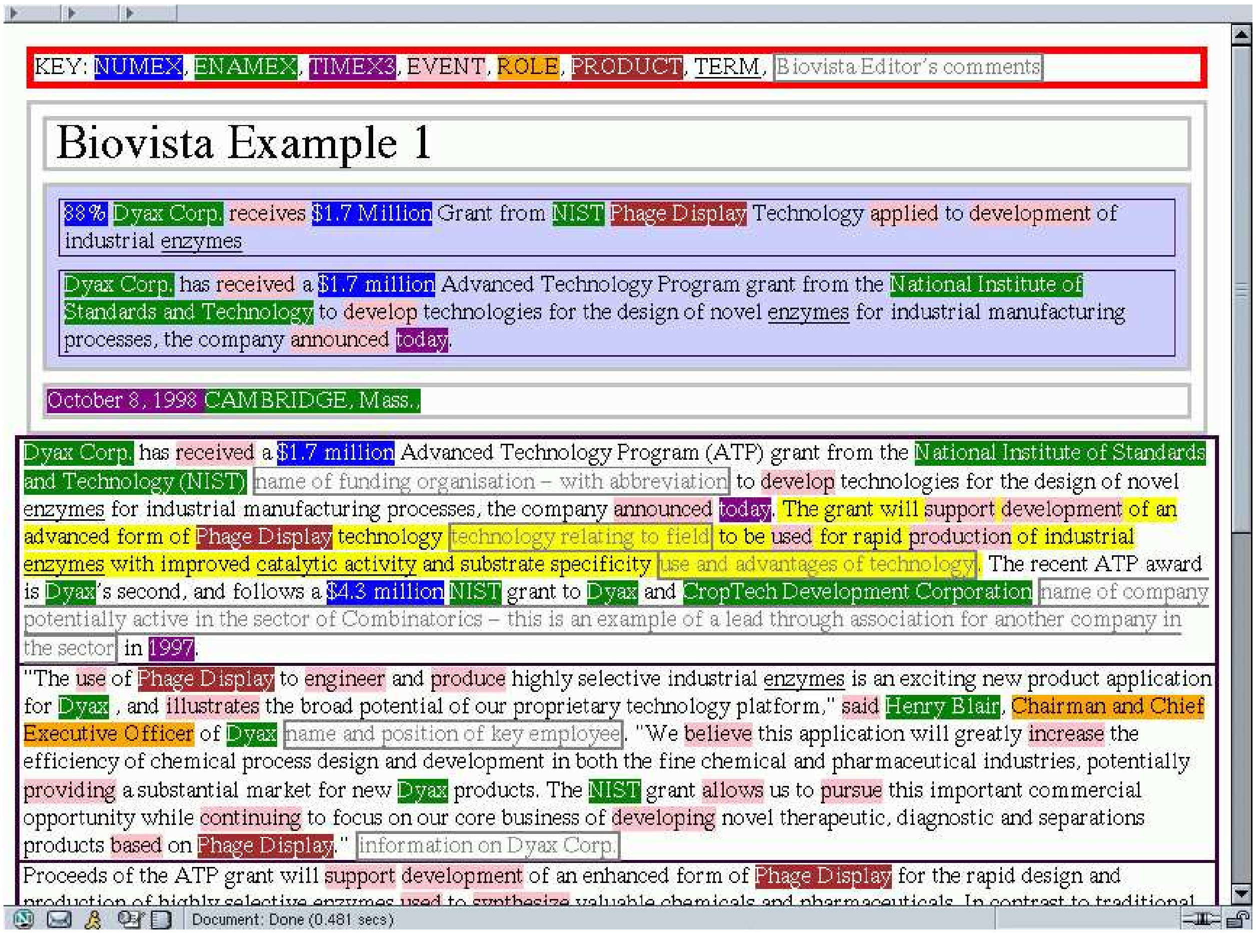


Figure 1: Basic Annotation Viewing

3.1 Example

While the structural annotations and lexical annotations should be easy to grasp as they correspond to accepted notions of document structure and of conventional span-based annotations, an example might help to illustrate the role of semantic annotations.

(1) The recent ATP award is
<ENAMEX id="e8" type="ORGANIZATION">
Dyax
</ENAMEX>
's second, and follows a
<NUMEX id="n5" type="MONEY">
$4.3 million
</NUMEX>
<ENAMEX id="e9" type="ORGANIZATION">
NIST
</ENAMEX>
grant to
<ENAMEX id="e10" type="ORGANIZATION">
Dyax
</ENAMEX>
and
<ENAMEX id="e11" type="ORGANIZATION">
CropTech Development Corporation
</ENAMEX>
in
<TIMEX3 tid="t4" type="DATE" value="1997">
1997
</TMEX3>

There are two occurrences of Dyax in this short text: the two Lexical Entities `e8` and `e10`, but clearly they correspond to the same Semantic Entity. To capture this equivalence, we could use the syntactic notion of co-reference (i.e. Identify the two as co-referent). Another possible approach is to make a step towards the conceptual level, and create a semantic entity, of which both `e8` and `e10` are lexical expressions (which could be different, e.g. "Dyax", "Dyax Corp.", "The Dyax Corporation"). The second approach can be implemented using an empty XML element, created whenever a new entity is mentioned in text. For instance, in (2) we can use the tag

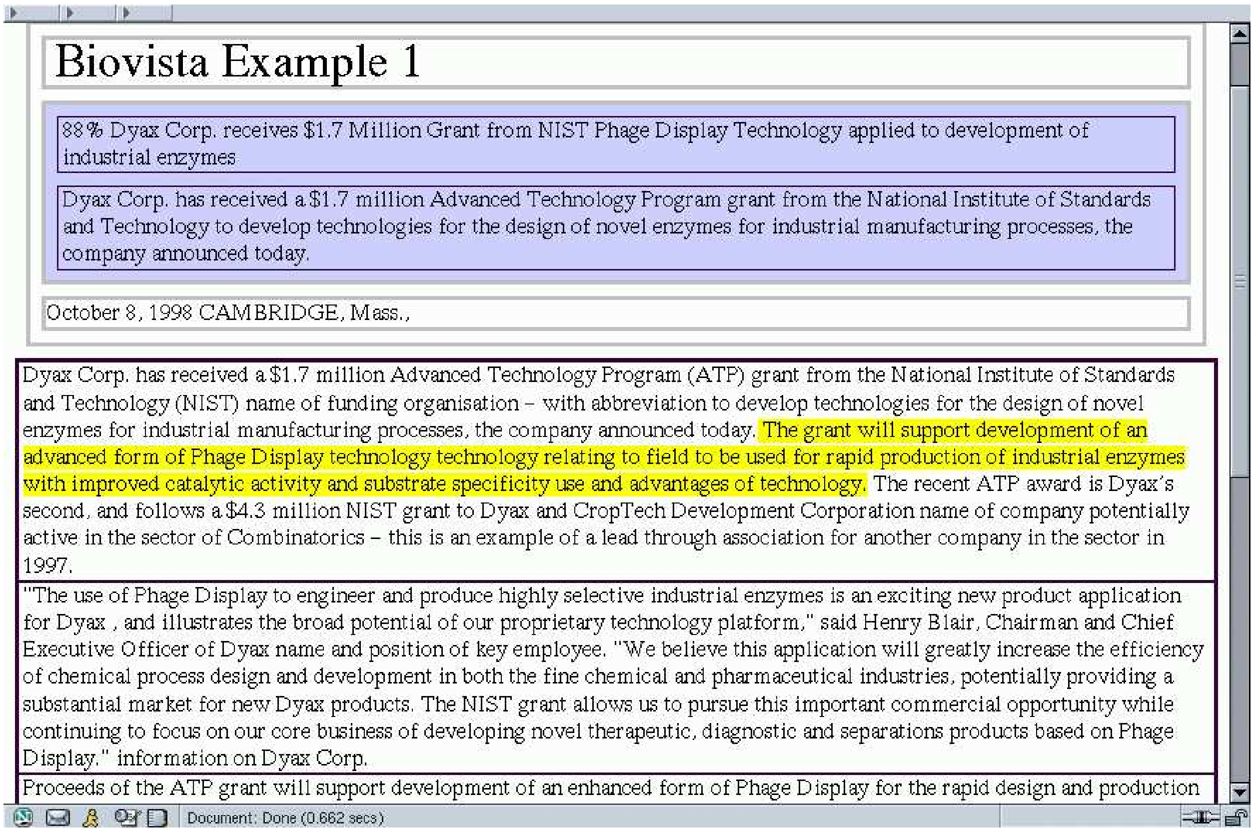


Figure 2: Visualization of Structural Annotations

<PEntity> (which stands for Parmenides Entity).

(2)
```
<PEntity peid="obj1" type="ORGANIZATION"
mnem="Dyax" refid="e1 e3 e6 e8 e10 e12"/>
```

The new element is assigned (as usual) a unique identification number and a type. The attribute `mnem` contains just one of the possible ways to refer to the semantic entity (a mnemonic name, possibly chosen randomly). However, it also takes as the value of the `refid` attribute as many coreferent ids as are warranted by the document. In this way all lexical manifestations of a single entity are identified. All the lexical entities which refer to this semantic entity, are possible ways to 'name' it (see also fig. 4).

Notice that the value of the 'type' attribute has been represented here as a string for readability purposes, in the actual specification it will be a pointer to a concept in a domain-specific Ontology.

Other semantic entities from (1) are:

(3)
```
<PEntity peid="obj2" type="ORGANIZATION"
mnem="NIST" refid="e2 e4 e7 e9"/>
<PEntity peid="obj3" type="ORGANIZATION"
mnem="CropTech" refid="e11"/>
```

The newly introduced semantic entities can then be used to tie together people, titles and organizations on the semantic level. Consider for example the text fragment (4), which contains only Lexical Annotations.

(4)
```
... said
<ENAMEX id="e17" type="PERSON">
Charles R. Wescott
</ENAMEX>
, Ph.D.,
<ROLE type='x' id="x5">
Senior Scientist
</ROLE>
at
<ENAMEX id="e60" type="ORGANIZATION">
Dyax Corp
</ENAMEX>
```

The Lexical Entity `e17` requires the introduction of a new semantic entity, which is given the arbitrary identifier 'obj5':

(5)
```
<PEntity peid="obj5" type="PERSON"
mnem="Charles R. Wescott" refid="e17"/>
```

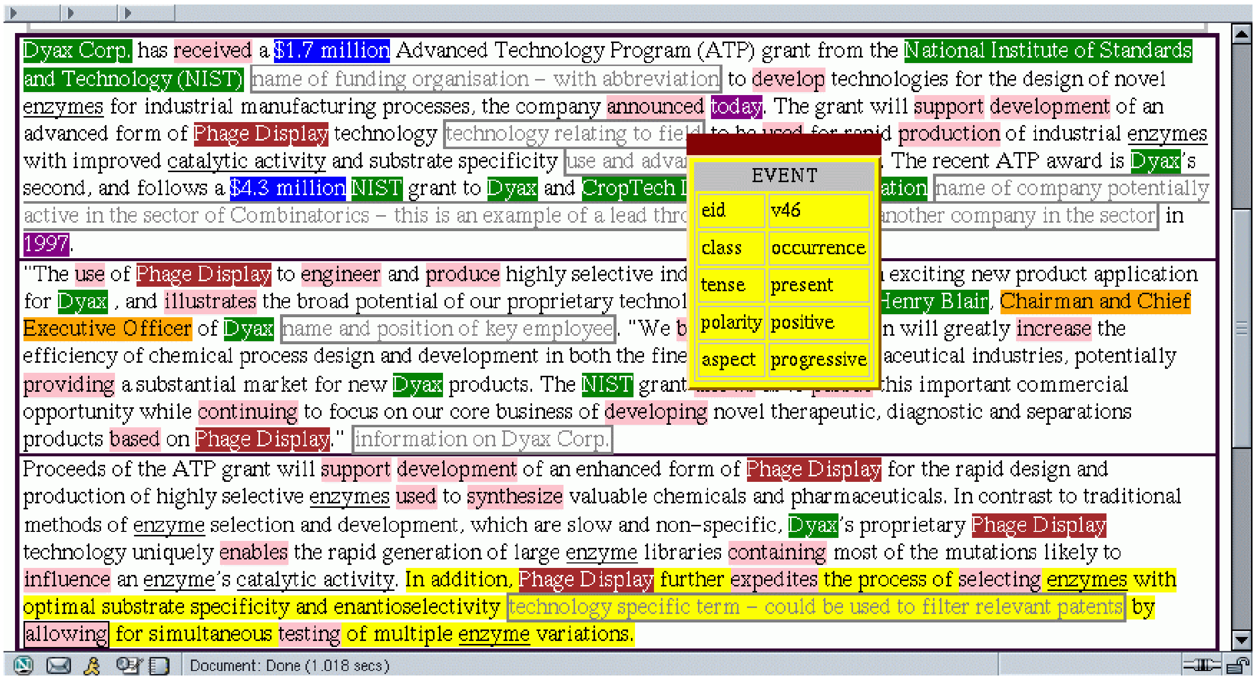


Figure 3: Visualization of Lexical Annotations and their attributes

In turn, this entity is linked to the entity `obj1` from (1) by a relation of type 'workFor' (PRelation stands for Parmenides Relation):

(6)
```
<PRelation prid="rel2" source="obj5"
target="obj1" type="worksFor" role="Senior
Scientist" evidence="x5"/>
```

4 Discussion

As the status of the Parmenides annotation scheme is still preliminary, we aim in this section to provide some justification for the choices done so far and some comparison with existing alternatives.

4.1 Named Entities

One of the purposes of Named Entities is to instantiate frames or templates representing facts involving these elements. A minor reason to preserve the classic named entities is so that we can test an IE system against the MUC evaluation suites and know how it is doing compared to the competition and where there may be lacunae. As such, the MUC-7 specification (Chinchor, 1997) is adopted with the minor extension of a non-optional identification attribute on each tag.

4.2 Terminology

A term is a means of referring to a concept of a special subject language; it can be a single wordform, a multiword form or a phrase, this does not matter. The only thing that matters is that it has *special reference*: the term is restricted to refer to its concept of the special domain. The act of (analytically) defining fixes the special reference of a term to a concept. Thus, it makes no sense to talk of a term not having a definition. A concept is described by defining it (using other certain specialised linguistic forms (terms) and ordinary words), by relating it to other concepts, and by assigning a linguistic form (term) to it.

If we are interested in fact extraction from densely terminological texts with few named entities apart from perhaps names of authors, names of laboratories, and probably many instances of amounts and measures, then we would need to rely much more on prior identification of terms in the texts, especially where these are made up of several word forms.

A term can have many variants: even standardised terms have variants e.g. singular, plural forms of a noun. Thus we should perhaps more correctly refer to a termform, at least when dealing with text. Among variants one can also include acronyms and reduced forms. You therefore find a set of variants, typically, all referring to the same concept in a special domain: they are all terms (or termforms). Again this problem pinpoints the need for a separation of the lexical annotations (the surface variants within the document) and semantic annotations (pointing abstractly to the underlying concept).

4.3 Approaches to Temporal Annotations

TIDES (Ferro et al., 2001) is a temporal annotation scheme that was developed at the MITRE Corporation and it can be considered as an extension of the MUC7 Named Entity Recognition (Temporal Entity Recognition - TIMEX Recognition) (Chinchor, 1997). It aims at annotating and normalizing explicit temporal references. STAG (Setzer, 2001) is an annotation scheme developed at the University of Sheffield. It has a wider focus than TIDES in the sense that it combines explicit time annotation, event annotation and the ability to annotate temporal relations between events and times.

TimeML (Ingria and Pustejovsky, 2002) stands for "Time Markup Language" and represents the integration and consolidation of both TIDES and STAG. It was created at the TERQAS Workshop[4] and is designed to combine the advantages of the previous temporal annotations schemes. It contains a set of tags which are used to annotate events, time expressions and various types of event-event, event-time and time-time relations. TimeML is specifically targeted at the temporal attributes of events (time of occurrence, duration etc.).

As the most complete and recent, TimeML should be adopted for the temporal annotations. Broadly, its organization follows the Parmenides distinction between lexical/semantic annotations. Explicit temporal expressions and events receive an appropriate (text subsuming) lexical tag. The temporal relations existing between these entities are then captured through a range of semantic (non-text subsuming) tags.

For example, each event introduces a corresponding semantic tag. There is a distinction between event "tokens" and event "instances" motivated by predicates that represent more than one event. Accordingly, each event creates a semantic `<MAKEINSTANCE>` tag that subsumes no text. Either, one tag for each realised event or a single tag with the number of events expressed as the value of the cardinality attribute. The tag is introduced and the event or to which it refers is determined by the attributes eventID.

5 Conclusion

We believe that ISO TC37/SC4 provides a very interesting framework within which specific research concerns can be addressed without the risk of reinventing the wheel or creating another totally new and incompatible annotation format. The set of annotations that we have been targeting so far in Parmenides is probably a small subset of what is targeted by ISO TC37/SC4. Although we had only limited access to the documentation available, we think our approach is compatible with the work being done in ISO.

It is, we believe, extremely important for a project like ours, to be involved directly in the ongoing discussion. Moreover we are at precisely the right stage for a more direct 'exposure' to the ISO TC37/SC4 discussion, as we have completed the exploratory work but no irrevocable modeling commitment has so far been taken. Therefore we would hope to become more involved in order to make our proposal fit exactly into that framework. The end result of this process might be that Parmenides could become a sort of "Guinea Pig" for at least a subset of ISO TC37 SC4.

[4] http://www.cs.brandeis.edu/~jamesp/arda/time

Acknowledgments

The Parmenides project is funded by the European Commission (contract No. IST-2001-39023) and by the Swiss Federal Office for Education and Science (BBW/OFES). All the authors listed have contributed to the (ongoing) work described in this paper. Any remaining errors are the sole responsibility of the first author.

References

Nancy Chinchor. 1997. MUC-7 Named Entity Task Definition, Version 3.5. http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/ne_task.html.

Lisa Ferro, Inderjeet Mani, Beth Sundheim, and George Wilson. 2001. Tides temporal annotation guidelines, version 1.0.2. Technical report, The MITRE Corporation.

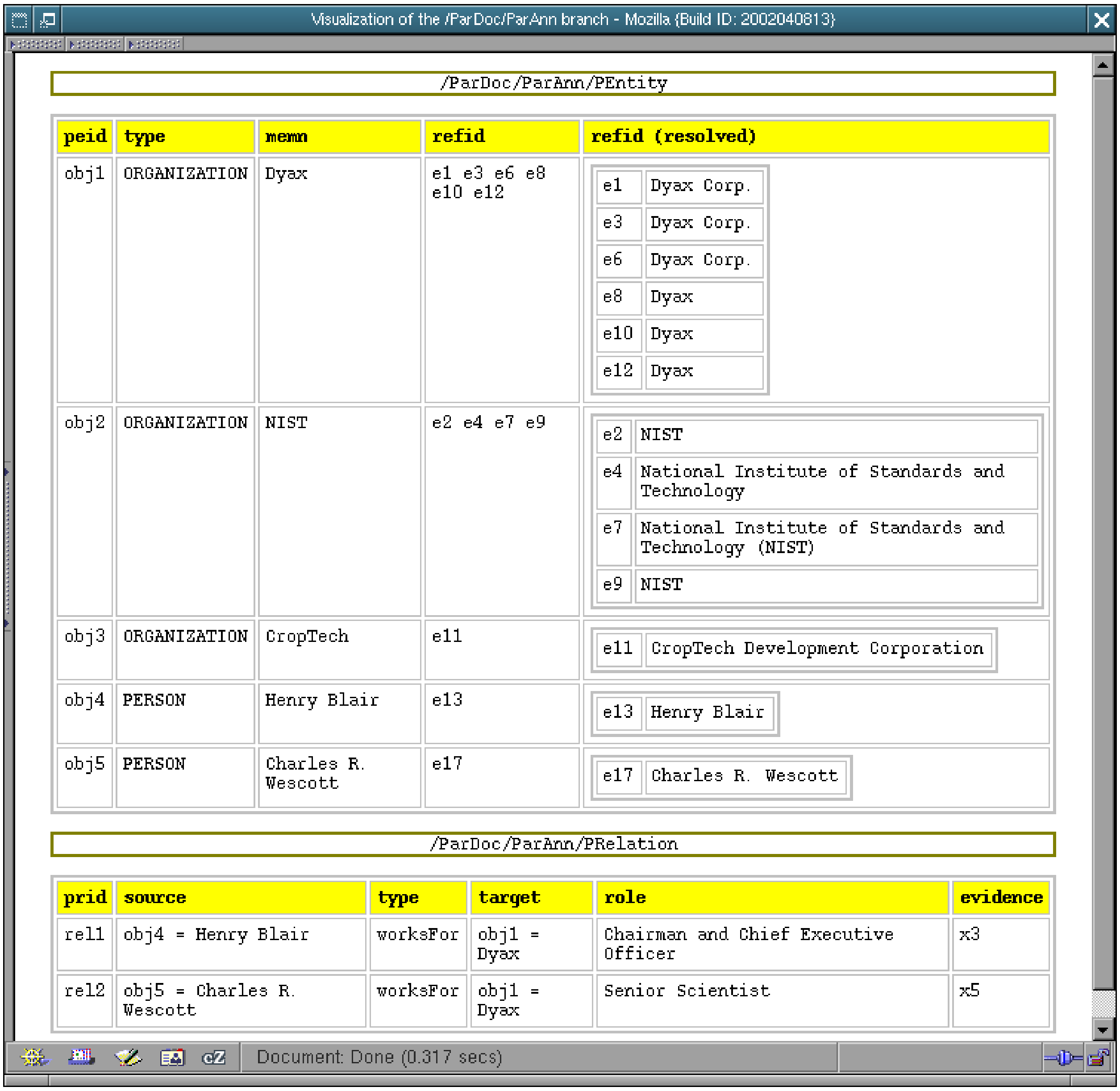
GENIA. 2003. Genia project home page. http://www-tsujii.is.s.u-tokyo.ac.jp/~genia.

Bob Ingria and James Pustejovsky. 2002. TimeML Specification 1.0 (internal version 3.0.9), July. http://www.cs.brandeis.edu/%7Ejamesp/arda/time/documentation/TimeML-Draft3.0.9.html.

James Pustejovsky, Roser Sauri, Andrea Setzer, Rob Giazauskas, and Bob Ingria. 2002. TimeML Annotation Guideline 1.00 (internal version 0.4.0), July. http://www.cs.brandeis.edu/%7Ejamesp/arda/time/documentation/TimeML-Draft3.0.9.html.

Andrea Setzer. 2001. *Temporal Information in Newswire Articles: An Annotation Scheme and Corpus Study*. Ph.D. thesis, University of Sheffield.

TEI Consortium. 2003. The text encoding initiative. http://www.tei-c.org/.



Visualization of the /ParDoc/ParAnn branch - Mozilla {Build ID: 2002040813}

/ParDoc/ParAnn/PEntity

peid	type	memn	refid	refid (resolved)
obj1	ORGANIZATION	Dyax	e1 e3 e6 e8 e10 e12	e1 Dyax Corp. e3 Dyax Corp. e6 Dyax Corp. e8 Dyax e10 Dyax e12 Dyax
obj2	ORGANIZATION	NIST	e2 e4 e7 e9	e2 NIST e4 National Institute of Standards and Technology e7 National Institute of Standards and Technology (NIST) e9 NIST
obj3	ORGANIZATION	CropTech	e11	e11 CropTech Development Corporation
obj4	PERSON	Henry Blair	e13	e13 Henry Blair
obj5	PERSON	Charles R. Wescott	e17	e17 Charles R. Wescott

/ParDoc/ParAnn/PRelation

prid	source	type	target	role	evidence
rel1	obj4 = Henry Blair	worksFor	obj1 = Dyax	Chairman and Chief Executive Officer	x3
rel2	obj5 = Charles R. Wescott	worksFor	obj1 = Dyax	Senior Scientist	x5

Document: Done (0.317 secs)

Figure 4: Visualization of Semantic Annotations