Towards an ontology based information system for Linguistic: the case study of the OTIM project

J. Seinturier, E. Murisasco, E. Bruno
LSIS UMR CNRS 6168
Université du Sud Toulon Var
Avenue de l'Université - BP20132
83957 LA GARDE CEDEX - FRANCE
name@univ-tln.fr

P. Blache
LPL UMR CNRS 6057
5 avenue Pasteur
13604 Aix-en-Provence, FRANCE
blache@lpl-aix.fr

1 Introduction

This work stands in the OTIM (Tools for Multimodal Annotation processing) project [1]. It aims at developing conventions and tools for multimodal annotation of a large conversational French speech corpus (Blache et al., 2010) (http://aune.lpl.univ-aix.fr/~otim/). OTIM can be summarized in two main steps.

The first step concerns *the multimodal annotation of a conversational speech between two persons*. It is under the responsibility of linguists; annotation is done according to different levels of linguistic analysis. Each expert has to annotate the same data flow according to its knowledge domain and the nature of the signal on which he annotates (signal transcription or signal). Experts generally use dedicated tools like PRAAT[2], ANVIL[3] or ELAN[4]. The qualifier multimodal is due to the nature of the studied corpus which is composed of text, sound, video. Within the project OTIM, linguists propose an encoding for annotating spoken language data, with the acoustic signal as well as its orthographic transcription. They have chosen to use Typed Feature Structures (Carpenter, 1992)(Copestake, 2003) (TFS) to represent in an unified view the knowledge and the information they need for annotation. Linguistic annotation tools rely on native and not often open formats which are not directly interoperable. TFS provides an abstract description using a high level formalism independent from coding languages and tools.

The second step concerns *the representation and manipulation of multimodal annotation*. We aim at providing linguists with a unique framework to encode and manipulate numerous linguistic domains (morpho-syntax, prosody, phonetics, disfluencies, discourse, gesture and posture (Blache et al., 2010)) in order to analyze and find correlations between annotated linguistic domains. For that, it has to be possible to bring together and align all the different annotations associated to a corpus.

In this paper, we focus on this last step considering semantic web technologies for the development of a Knowledge-based Information System.

2 Context and Motivation

Linguistic knowledge is captured by means of three types of information : *properties* (the set of characteristics of an object); *relations* (the set of relations that an object has with other objects); *constituents* (complex objects composed of other objects). TFS proposes a formal presentation of each annotation in terms of feature structures and type hierarchies : properties are encoded by features, constituency is implemented with complex features, and relations make use feature structure indexing; each linguistic domain is represented as a hierarchical model. TFS enables linguists to represent in an unified view the knowledge and the information they need for annotation.

Figure 1 graphically describes TFS representation of the prosodic domain; a formal definition can be found in (Carpenter, 1992) (Copestake, 2003). For sake of simplicity, we do not detail the meaning of every feature used in the example.

Due to its theoretical nature, TFS representation cannot be used within an applicative framework

[1] supported by the French ANR agency
[2] http://www.fon.hum.uva.nl/praat/
[3] http://www.anvil-software.de/
[4] http://www.lat-mpi.eu/tools/elan/

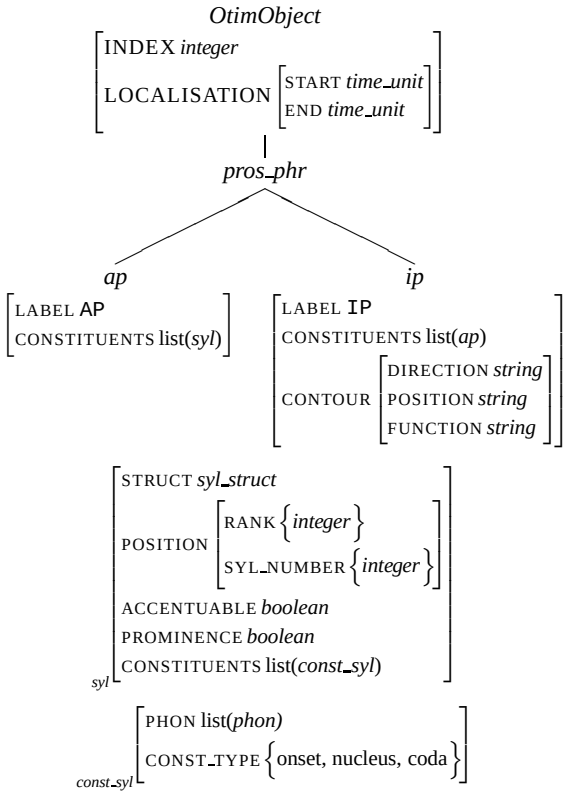

Figure 1: TFS representation of the prosodic domain

and has to be implemented into other formalisms.

3 Contributions

We have propose an knowledge representation formalism which be an alternative to TFS : an ontological approach based on Description Logics (Baader et al., 2003) (DL) and on semantic web technologies for the development of a linguistic Knowledge-based Information System(Seinturier, 2011).

Some linguistic projects have a similar objective than OTIM, for instance NITE[5], AGTK[6], PAULA[7], XStandoff (Stuhrenberg, 2009). Our approach differs from them because we focus on an ontological contribution. These proposals generally propose toolkits for multi-level annotation by means of libraries of data and annotation management. Moreover, linguistic annotation tools rely on native and not often open formats which are not directly interoperable. The multiplication of annotation schemes and coding formats is a severe limitation for interoperability. One solution consists in developing higher level approaches (Ide, 2007)(Stuhrenberg, 2009). However, these experiments still remain very programmatic.

3.1 Creating OWL ontology

Creation of the OWL ontology follows two steps. First of all, the terminological knowledge from the TFS is implemented into OWL using the Protege[8] ontology editor. The Protege framework was initially designed for biologists and biochemists. This characteristic is quite interesting because this is not a computer scientist tool and so there is no need of a specific knowledge in computer science to use it.

Figure 2 shows the ontology of the prosodic domain. This ontology is linked with two other domains: the phonetics domain, which is a part of the OTIM knowledge representation framework, and the time domain given by a standard ontology of the W3C.

3.2 Managing data and querying with SPARQL

Management and querying of OWL data relies on the standard SPARQL (Prudhommeaux, 2007) querying language. SPARQL enables to match graph pattern against the graph of RDF/OWL triple ($WHERE$ clause) and identifies values to be returned ($SELECT$ clause). The $FROM$ clause enables to identify the data sources to query.

We express within the OTIM project the linguistic inter domain queries designed on TFS by SPARQL queries on the OWL representation. A sample query expressed in natural language is:

"We need the list of phonemes that are associated with the accentual phrases stated between the second 35 and the second 55 of the speech."

This query takes into account the prosodic domain (accentual phrase), the phonetic domain (phoneme) and the time. Such a query is represented in SPARQL by:

```
1.  SELECT    ?phoneme
2.  FROM      otim − prosody.owl, otim − phonetics.owl
3.  WHERE {   ?const rdf:type prosody:SyllableConst
4.            . ?const hasPhonemes ?phoneme
5.            . ?syl rdf:type prosody:Syllable
6.            . ?sc hasConstituents ?const
7.            . ?ap rdf:type prosody:AccentualPhrase
8.            . ?ap hasSyllables ?syl
9.            . ?t rdf:type time:TemporalEntity
10.           . ?ap hasTimeLocation ?t
11.           . ?tref time:contains ?t }
```

[5]http://groups.inf.ed.ac.uk/nxt/

[6]http://weblex.ens-lsh.fr/projects/xitools/logiciels/AGTK/agtk.htm

[7]http://www.sfb632.uni-potsdam.de/ d1/paula/doc/

[8]http://protege.stanford.edu/

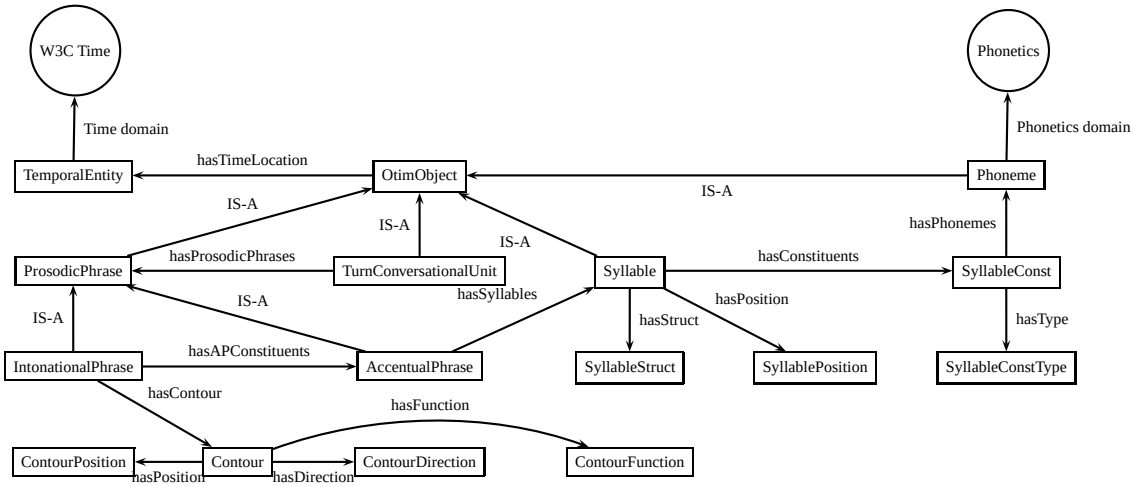

Figure 2: Ontological representation of the prosodic domain

We assume ithat the time bounds given are represented as a $TemporalEntity$ named $tref$. The $SELECT$ clause specifies that the result to build is made of phonemes. The clause $FROM$ contains the two data sources on which the query is processed (the two target domains prosody and phonetics). The $WHERE$ clause describes the patterns for a phoneme to match.

The OTIM framework for linguistic multimodal annotations management has been implanted within a Java/OWL framework. The OWL standard used is OWL-DL as this is the specification that gives all the expressiveness we need and guarantees some calculability results that are critical for querying data.

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

J. Allen. 1991. Time and time again : The many ways to represent time. *International Journal of Intelligent Systems*, 6(4):341–355, july.

Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider. 2003. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003.

Philippe Blache, Roxane Bertrand, Brigitte Bigi, Emmanuel Bruno, E Cela, Robert Esperrer, Gaelle Ferre, Marion Guardiola, Daniel Hirst, Ep Magro, Martin JC, Meunier C, Morel Ma, Elisabeth Murisasco, Nesterenko I, Nocera P, Pallaud B, Prevot Laurent, Priego-Valverde J, Julien Seinturier, Tan N, Tellier Marion, and Rauzy Stephane. 2010. Multimodal annotation of conversational data. In *Proceedings of the fourth linguistic annotation workwhop (LAW)*, pages 186–191. Assocation for computational Lunguistics (ACL), 15 july 2010.

Robert L. Carpenter. 1992. *The Logic of Typed Feature Structures*, volume 32 of *Cambridge Tracts in Theoretical Computer Science*. Cambridge University Press, The Edinburgh Building, Shaftesbury Road, Cambridge CB2 8RU, United Kingdom, c. j. van rijsbergen edition, 1992.

Ann Copestake. 2003. *Collaborative Language Engineering: A Case Study in Efficient Grammar-based Processing*, chapter Definitions of Typed Feature Structures. CSLI Publications, Ventura Hall, Stanford University, Stanford, CA 94305-4115, 2003.

N. Ide and K. Suderman. 2007. Graf: A graph-based format for linguistic annotations. In *Linguistic annotation workshop (LAW)*. Association for computational Lunguistics (ACL), 2007.

Eric Prud'hommeaux and Andy Seaborne. 2007. Sparql query language for rdf (working draft). Technical report, W3C, March 2007.

J. Seinturier, H. Rouine, E. Murisasco, and E. Bruno andP. Blache. 2011. Knowledge-based multimodal data represnetation and querying. In *Int. conf. on Knowledge Engineering and Ontology Developmement (KEOD 2011)*, Paris, France, October 2011.

M. Stührenberg and D. Jettka. 2009. A toolkit for multi-dimensional markup - the development of sgf to xstandoff. In *Proceedings of Balisage: The Markup Conference 2009*. Assocation for computational Lunguistics (ACL), Balisage Series on Markup Technologies, vol. 3, 2009.