OntoLex 2008 Programme

9.00 – 9.10	Introduction to the workshop

Part I: Methodology and Resources

9.10 – 9.50	Enriching Ontologies with Linguistic Content: An Evaluation Framework (A. Oltramari, A. Stellato)
9.50 –10.30	Language Resources and Tools for Ontology-based Annotation (K.Simov, P. Osenova)
10.30 – 11.00	COFFEE BREAK
11.00 – 12.00	**INVITED TALK** by P. Cimiano (AIFB, University of Karlsruhe)
12.00 – 12.40	Building a free French wordnet from multilingual resources (D. Fišer, B. Sagot)
12.40 – 13.20	Toward Estonian Ontology (N. Kahusk, K. Kerner, H. Orav)
13.20 – 14.40	LUNCH BREAK

Part II: Wiki-based projects and applications

14.40 – 15.20	Lexicon and Ontology Interplay in Senso Comune (A. Oltramari, G. Vetere)
15.20 – 16.00	Computed-Assisted Ontology Creation from Video Transcripts Using Wikipedia and WordNet (Nakhimovsky, Myers)
16.00 – 16.30	**Open Discussion: Re-thinking OntoLex?**
16.30	End of the workshop

OntoLex 2008 Organisers

Alessandro Oltramari
Institute of Cognitive Sciences and Technologies (ISTC-CNR), Trento (Italy)

Laurent Prévot
CLLE-ERSS (CNRS), Toulouse (France)

Chu-Ren Huang
Institute of Linguistics, Academia Sinica, (Taipei) Taiwan

Paul Buitelaar
DFKI GmbH
Language Technology Lab & Competence Center Semantic Web, (Saarbrücken) Germany

Piek Vossen
Faculteit der Letteren Vrije Universiteit Amsterdam De Boelelaan 1105, Amsterdam
& Irion Technologies BV, Delft
(The Netherlands)

Programme Committee

Guido Vetere, IBM Center for Advanced Studies, Rome, Italy
Armando Stellato, Università di Tor Vergata, Rome, Italy
Luisa Bentivogli, Bruno Kessler Foundation, Povo (Trento), Italy
Nancy Ide, Department of Computer Science, Vassar College – USA
Christiane Fellbaum, Princeton University, USA
Andrea Schalley, University of New England, Australia
Nigel Collier, National Institute of Informatics, Japan
Asanee Kawtrakul, NECTEC, Thailand
Sujian Li, Peking University, China
Tokunaga Takenobu, Tokyo Institute of Technology, Japan
John Bateman, University of Bremen, Germany
Philipp Cimiano, Karlsruhe University, Germany
Paola Velardi, University of Rome, Italy
Johanna Voelker, Karlsruhe University, Germany
Chris Welty, IBM, USA
Alessandro Lenci, University of Pisa, Italy
Atanas Kiryakov, Ontotext, Bulgaria
Nathalie Aussenac-Gilles, IRIT-CNRS, Toulouse, France
Wim Peters, University of Sheffield – United Kingdom

Table of Contents

Author Index

Enriching Ontologies with Linguistic Content: an Evaluation Framework

Alessandro Oltramari, Armando Stellato

Laboratory for Applied Ontology (ISTC-CNR), University of Rome, Tor Vergata
Trento, Rome
oltramari@loa-cnr.it, stellato@info.uniroma2.it

Abstract

In this paper, we present a framework for representing and evaluating integrations between ontological and linguistic resources, which originates and improves previous research reported in (Pazienza & Stellato, 2006b; Pazienza, Sguera, & Stellato, 2007) and articulates into two results: first, a set of coordinated RDF vocabularies providing descriptors for representing linguistic resources and their software counterparts, as well as offering metadata for describing the linguistic enrichment of ontologies, both on quantitative and qualitative grounds. The second result is a software library for evaluating the quality of automatic linguistic enrichment tools.

The Linguistic Watermark suite of RDF vocabularies, in the newly presented form, provides to our framework shared vocabularies for addressing the knowledge about heterogeneous linguistic resources, for accessing and managing their content on a common basis through dedicated software components and for representing the integration of this content inside ontologies. This last part constitutes the bridge towards our novel evaluation framework, which produces quality reports based on assessed evaluation metrics taken from the Information Retrieval tradition (Van Rijsbergen, 1975) and adapted to this task. We hope that this framework could provide a stable and reusable tool for evaluating the quality of competing algorithmic solutions for linguistic enrichment of ontologies.

1. Introduction

Semantic Web ontologies represent the shared vocabularies through which machines can read and access content from the Web, or even communicate between them, to exchange information or cooperate for achieving some goal. This definition implicitly assumes that in an heterogeneous scenario like the whole WWW, the same concepts will be represented by the same ontologies and that, therefore, ontological models of data will be consistent; conversely, sensible effort will be put in trying to match these "not-so-shared" vocabularies. If that general assumption may hold true for reduced-size, very specific and data-oriented ontologies (e.g. the WGS84 Geo Positioning RDF vocabulary[1], which contains only a few properties for describing latitude, longitude and point-in-space concepts), for larger domain descriptions, requiring different levels of abstraction and different perspectives depending on local needs, we expect to see several, different ontologies arise from independent organizations, often addressing overlapping domains.

Two issues then urge to be solved: first, facilitating people and automated systems in performing alignments between ontologies where they represent the same concepts and, secondly, make their vocabularies more explicit to humans, so that they can be re-used consistently in different scenarios and by different actors; in this sense, logical consistency may only help in restricting the range of possible interpretations which may be assigned to logical symbols, while common-sense human reasoning using these vocabularies may beneficiate a lot by the presence of clear and exhaustive documentation. Extensive use of Natural Language contents, providing free descriptions, synonymical expressions and translations in different idioms of the intended meaning of a vocabulary, appears thus as the most intuitive kind of documentation for data structures such as ontologies, dealing with representation of domains. Several efforts have been undertaken to cover different aspects of this problem, motivating the adoption of linguistic resources for enriching ontology vocabularies with natural language contents[2] (Pazienza & Stellato, 2006b; Prevot, Borgo, & Oltramari, 2005; Scheffczyk, Baker, & Narayanan 2006; Philpot, Hovy, & Pantel 2005; Huang 2004), showing useful applications exploiting these combined resources (Basili, Vindigni, & Zanzotto, 2003; Peter, Sack, & Beckstein 2006; Cappelli, Giovannetti & Michelassi 2004), providing standards for representing this enrichment/integration, like in SKOS[3] (Simple Knowledge Organization Systems) and in (Buitelaar, et al., 2006), and promoting the development of techniques for automating this task (Pazienza & Stellato, 2006c).

In this paper, we present an ontological and software framework for describing, referring and managing heterogeneous linguistic resources and for using their content to enrich and document ontological objects. This work, which originates ad completes previous research reported in (Pazienza & Stellato, 2006b; Pazienza, Sguera, & Stellato, 2007) articulates into two results: first, a set of coordinated RDF vocabularies providing descriptors for representing linguistic resources (ranging from lexical to frame-based ones) and their software counterparts (data structures, access libraries etc…), as well as offering metadata for describing the linguistic enrichment of ontologies, both on quantitative and qualitative grounds. The second result is a software library for evaluating the quality of automatic linguistic enrichment tools, through comparison of enriched ontologies compiled against the above vocabularies.

2. Related works

The actual practice of enriching ontologies with linguistic content basically depends on the multifariousness of lexical resources and on the explicit linguistic information

[1] http://www.w3.org/2003/01/geo/wgs84_pos

[2] The enrichment of ontologies with linguistic contents fosters the construction of peculiar kinds of semantic resources, which we could refer to as "hybrid" knowledge resources.

[3] http://www.w3.org/TR/swbp-skos-core-guide/

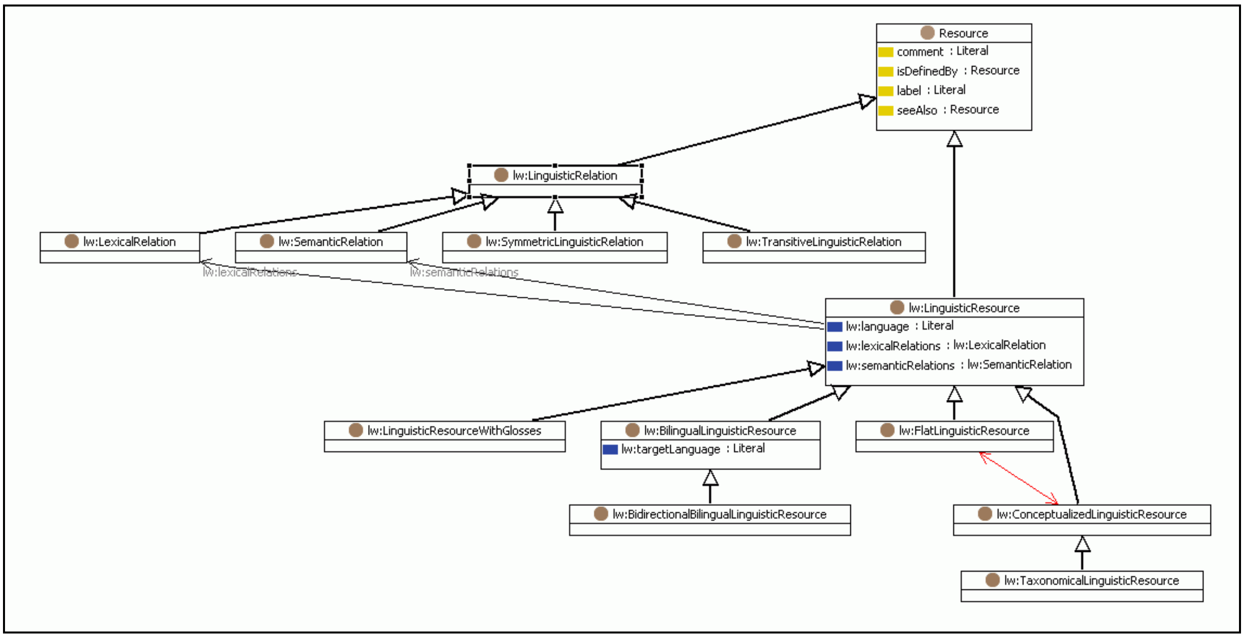

Resource
comment : Literal
isDefinedBy : Resource
label : Literal
seeAlso : Resource
lw:LinguisticRelation
lw:LexicalRelation
lw:SemanticRelation
lw:SymmetricLinguisticRelation
lw:TransitiveLinguisticRelation
lw:lexicalRelations
lw:semanticRelations
lw:LinguisticResource
lw:language : Literal
lw:lexicalRelations : lw:LexicalRelation
lw:semanticRelations : lw:SemanticRelation
lw:LinguisticResourceWithGlosses
lw:BilingualLinguisticResource
lw:targetLanguage : Literal
lw:FlatLinguisticResource
lw:BidirectionalBilingualLinguisticResource
lw:ConceptualizedLinguisticResource
lw:TaxonomicalLinguisticResource

Figure 1: An excerpt (focused on description of Linguistic Resources) from the Linguistic Watermark vocabulary

they expose (Pazienza and Stellato 2006c). Multilingual scenarios also demand for a proper lexicalization of ontological content according to different idioms and languages. From simple vocabularies of terms to wordnet-like structures, distinct lexical models need a solid and comprehensive framework of representation to enable a full-operational integration with ontologies. One example of this research trend is represented by the W3C initiative of translating WordNet to RDF/OWL, whose aim is to enable porting that kind of resource into Semantic Web infrastructure. Moreover, the integration between frame-based lexical databases and ontologies complicates the overall scenario and constitutes another important aspect of the above-mentioned process and a relatively brand-new trend in the scientific community. In a nutshell, the main rationale behind the notion of "frame semantics" (Fillmore 1968) is that meaning is represented by generalizations from stereotyped situations (frames). Berkeley FrameNet Project (Baker, Fillmore & Lowe; 1998) has been designed on the basis of that principle: nouns, verbs, and possibly modifiers (adjectives and adverbs) are clustered according to conceptual structures (e.g., the *commercial transaction* frame) and syntactic combinatory possibilities (valences). Several language-specific framenets have also emerged in the latest years according to Berkeley's model. The value of porting these kind of lexical databases into Semantic Web basically depends on the exploitation of their peculiar semantic structure for the enrichment of ontologies: this task may correspond to supply a formal semantics to frames (i.e. OWL semantics) or, besides re-engineering frame-based resources according to WWW standards, to use suitable pointers to link ontological categories and relations with frames. Similar issues arise from the task of interfacing ontologies with VerbNet (Kipper, Trang Dang, & Palmer, 2000), a project in which PropBank (Palmer, Kingsbury & Gildea, 2005) verb types are mapped to Levin Classes (Levin 1993): here the resource is organized into verb classes and alternations, without considering the role of nouns and modifiers in conceptual structures.

Despite the large interest in this area, standards for representing layered ontological-linguistic knowledge hardly finds a place in the Semantic Web stream of innovation, and while it has been shown that these processes can be handled with different levels of automation, no evaluation framework has been proposed until now.

3. The Linguistic Watermark Suite

The Linguistic Watermark suite of RDF vocabularies is composed of three ontologies:

- The *Linguistic Watermark* (*LW*) vocabulary, describing linguistic resources through their purposes and structure organization
- The *Ontological Linguistic Watermark* (*OLW*) vocabulary: a set of metadata descriptors for characterizing the linguistic expressivity of ontologies
- The *LW Linguistic Interfaces* vocabulary (*LWLI*), providing concepts for describing software libraries which grant access to specific (or ranges of) linguistic resources.

3.1. The Linguistic Watermark (LW) Vocabulary

While the Linguistic Watermark vocabulary partially covers general linguistic concepts like term, word, lexical/semantic relation, frame, agent etc... its main objective is to provide descriptors or characterizing the purpose and structure of linguistic resources: whether they represent translation vocabularies, synonyms collections, lexicons, frame based resources or terminologies, if they are organized around some kind of semantic structure or merely <entry, description> pairs etc..

Though originally conceived to cover any kind of Linguistic Resource, the first version of the Linguistic Watermark (figure 1) was limited to represent only lexical resources: by proper combination of its LW ontological

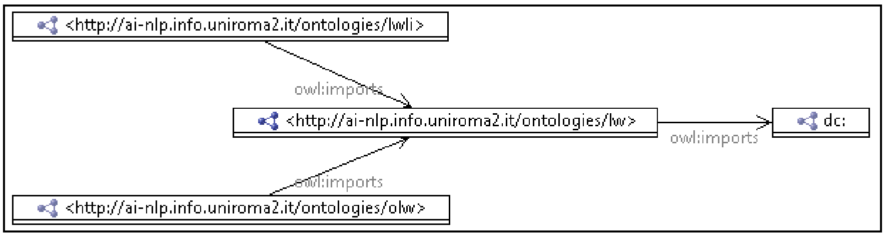

<http://ai-nlp.info.uniroma2.it/ontologies/lwli>
owl:imports
<http://ai-nlp.info.uniroma2.it/ontologies/lw>
owl:imports
dc:
owl:imports
<http://ai-nlp.info.uniroma2.it/ontologies/olw>

Figure 2: owl:imports relationships between ontologies in the Linguistic Watermark suite

descriptors, one could be able to represent very different linguistic resources, from simple synonym dictionaries, to complex resources such as WordNet (Miller et al, 1993). This provided a shared and homogeneous vocabulary upon which multilingual (and multi-resource) applications could be defined.

In this work we have extended le LW vocabulary into two main directions:

– *Instantiation*: now the vocabulary is not only used to describe linguistic resources, but even to predicate over their content (see section 4.2.2 for details)
– *Frames description*: covering frame/class based linguistic resources, such as FrameNet and VerbNet.

FrameNet and VerbNet have been modeled as distinct specializations of the newly introduced class FrameBasedResource, which is a rdfs:subClassOf of ConceptualizedLinguisticResource. This modeling choice mainly depends on the intrinsic nature of so-called "building blocks" of frame-based resources: "frames" are the organizational units of FrameNet corresponding to general schemas of specific situations. They are normally constituted by "Frame Elements", such as *Buyer* and *Seller* (in the *Commercial Transaction* frame), which are to be conceived as conceptual parts of a frame. The notion of "Frame Element" is very close to the basic notion of "Thematic Role", which is more general and domain-independent and actually adopted as the basic unit of VerbNet: some typical examples of thematic roles are *Agent*, *Patient*, *Duration*, *Destination*.

Resources of type FrameBasedResource adopt a specialization of SemanticIndex, namely Frame, which is structured according to variable sets of objects called FrameElement.

Another important issue concern relations holding between frames. Seven types of parent/child relation are used in FrameNet, namely "subframe", "inheritance", "perspective on", "using", "causative of", "inchoative of", "see also" and one type of temporal ordering relation, that is "precedes". Although is not our aim here to focus on the semantics of these relations, clearly they are not lexical ones: they pertain to the conceptual level and are used to structure the set of frames (up to date, around 1000) which compose FrameNet. Nonetheless, they can be mapped through instances of the already existing SemanticRelation class. It is relevant to notice that some frame relations are transitive (as hyponymy in wordnet-like linguistic resources); for instance, the ordering relation "precedes", which establishes a chronological nesting within frames (and frame elements too).

A crucial aspect in making the LW a vocabulary for describing instantiable linguistic resources is the link between SemanticIndex and LexicalUnit class. In general, semantic indexes can be thought as conceptual objects which can, depending on the purpose and semantics of the considered resources, be associated to simple or compound words, which are actually kinds of lexical units. According to this modeling perspective, the relation lexicalUnit has been created, holding between LexicalUnit and SemanticIndex: for instance, the verb "purchase" (simple word) is both the lexical unit of the frame *Commercial Transaction* and of the WordNet's synset <buy, purchase>[4]. This example shows how the LW model is able to capture different uses in different lexical resources of the same linguistic units. The semantics of each instantiation of the lexicalUnit property depend on the considered resource, while the LW library may offer homogeneous API for inspecting different linguistic resources, for showing their content on automatically generated GUIs or enabling its integration inside other representation formalisms, such as ontologies. This generalization thus boosts reuse and integration of several resources in several application contexts.

3.2. The Ontological Linguistic Watermark (OLW)

The characterization given by the OLW is expressed in terms of the linguistic content of the described ontology and with respect to the resources which have been adopted for enriching its concepts. As stated in (Pazienza, Sguera, & Stellato, 2007), where its adoption has been considered in a scenario involving Semantic Coordination of FIPA agents, its metadata assume great significance in all the contexts where ontologies sharing a common domain, but no explicit semantic bridging between their respective vocabularies, need to be automatically aligned or merged. Resource-based algorithms for ontology alignment and semantic coordination agents can in fact inspect the OLW data of the ontologies to be compared and configure at best the resources and facilities to be used for matching their content. This is an aspect which has often been underestimated in literature: setting up the resources to be adopted in a realistic scenario, while being not a trivial task, influences dramatically the outcome and performances of any mediation activity.

The LWLI takes its roots from the first version of the Linguistic Watermark software library[5] – developed by the University of Rome, Tor Vergata – a component providing uniform access to different and heterogeneous linguistic resources, which has been used in several resource-based tools, such as the OntoLing Protégé plug-

[4] Gloss: "obtain by purchase; acquire by means of a financial transaction"; "The family purchased a new car"; "The conglomerate acquired a new company"; "She buys for the big department store".

[5] http://ai-nlp.info.uniroma2.it/software/LinguisticWatermark/

in (Pazienza & Stellato, 2006). The LW presented in that work, was just a class diagram offering several interfaces and abstract classes whose combination could be used to describe the main aspects of a linguistic resource: implementing the proper subset of those (software) interfaces would result in the definition of a linguistic wrapper for accessing a particular linguistic resource. The LW library thus offered a combination of descriptive (with regard to the resources to be wrapped) and operative aspects (delineating the operations which the required wrapper had to implement). Later on, the requirements which brought to developing the OLW, demanded a formal ontological representation, merely focused on resource description, to be extracted from the original class diagram, which led to the LW.

Now, the time has come to close the circle, and with the LWLI we recovered the original intent of the LW library.

3.3. The LW Linguistic Interfaces vocabulary (LWLI)

LWLI contains concepts describing parameters needed by software libraries for setting up access to their target linguistic resources. This third ontology completely migrates the original framework to RDF, thus providing a complete vocabulary at the hand of Semantic Web tools which rely on the use of linguistic resources or are even expressly dedicated to the integration of ontologies with linguistic resources.

The LWLI includes concepts like:

- LinguisticInterface: for describing a specific implementation of a wrapper for a linguistic resource
- LinguisticInterfaceConfiguration: representing instances of basic runtime configurations for a given LinguisticInterface.
- LinguisticInterfaceInstanceConfiguration: each instance of this class provides data for completing a single runtime configuration for accessing a specific linguistic resource, basing on partial configuration from a given LinguisticInterfaceConfiguration

and properties for specifying these configuration settings, among which, we list the following ones:

- configuredInterface: this property tells which LinguisticInterface is being configured through the described configuration
- interfaceableResource: tells which linguistic resources are made accessible through the described Linguistic Interface
- ConfigurationProperty: a property defining configuration parameters for accessing a linguistic resource through a dedicated linguistic interface. This property is never instantiated, though it has a few relevant subproperties for telling whether a given configuration parameter points to the file system, if a property is relevant for configuring a linguistic interface as a whole, or just for accessing specific resources etc..

As for the LW, even this vocabulary provides an upper ontology which, though extensible in principle to match the specification of each represented software library, already contains all the required descriptors for automatically driving different linguistic resources under a shared knowledge model.

To have an example, consider the following use case: we are trying to describe the fictitious YAWW (Yet Another WordNet Wrapper) library. First of all, we declare yaww as a new instance of LinguisticInterface. Then, we should consider all the parameters that the wrapper needs for its configuration, distinguishing those needed to make the interface – as a whole – work, from those which are necessary for granting access to different WordNet versions installed on the host. These parameters should be used to instantiate properties for the two configuration classes LinguisticInterfaceConfiguration (the one related to general interface configuration), and LinguisticInterfaceInstanceConfiguration, for setting up access to specific resources.

We could even add more information at conceptual level, by adding specific subclasses, YAWWInterfaceConfig and YAWWInstanceConfig, respectively, to the two configuration classes above, and binding them, through property restrictions, to ad-hoc configuration properties, like the one which is described next.

Being YAWW a wrapper for WordNet, we would probably need to define a configuration property for specifying the path to the dictionary folders of the various installed wordnets we want to access; by first, we declare the owl:DataTypeProperty wnDictPath, then we state it as being rdfs:subPropertyOf of two available subproperties of lwli:ConfigurationProperty: the first one, lwli:InstanceProperty, tells that the its instantiated value represents a parameter for accessing a given wordnet (the one installed in that path) and not for configuring the whole library (and thus, that it has to be attached to a given YAWWInstanceConfig), while the second one, lwli:FileProperty informs that this property points to a file in the file system, so that applications based on this vocabulary, could in case apply necessary filechecking mechanisms, as well as find appropriate graphical interface widgets – a file chooser dialog, for example – when interacting with the user for filling the value of this parameter.

Though we added specific subclasses and subproperties (thus extending the conceptual part of the ontology), the software *interface*, which is based on the sole LWLI, does not need any changes, and thus the same for any application software based on LWLI, which can now benefit of the new added resource wrapper, without any development effort.

4. An improved Integration Framework

In this section we describe the new libraries and tools which have been developed with the intent of providing a consistent and homogeneous layer for integrating ontologies and linguistic resources, also taking into account the variety of proposed standards and research results which have arisen in these last years

4.1. The new Linguistic Watermark library

Following the recent improvements on the LW suite, we have released a new version of the Linguistic Watermark library, which offers java API for accessing linguistic resources through dedicated Linguistic Interfaces, both entities being defined according to the LW and LWLI vocabularies. In particular, a mapping between the above ontologies and newly added java interfaces allows implemented java wrappers for linguistic resources to

```
<wn20schema:NounSynset rdf:about="wn20instances:synset-entity-noun-1" rdfs:label="entity">
      <wn20schema:synsetId>100001740</wn20schema:synsetId>
</wn20schema:NounSynset>

<rdf:Description rdf:about="wn20schema:Synset">
      <rdfs:subClassOf rdf:resource="lw:SemanticIndex"/>
</rdf:Description>

<someOntology:Noun>
      <olw:semanticDescriptor rdf:resource="wn20instances:synset-entity-noun-1">
</someOntology:Noun>
```

Figure 3: an example of resource wrapping: binding WordNet-RDF synsets to a class concept

declare themselves as new instances of the LinguisticInterface class and accept strongly typed configuration parameters, thus enabling data consistency checks and providing hooks for automatic generation of configuration user interfaces for hosting applications.

4.2. The OLW library and OLW vocabulary improvements

With the specific aim of obtaining a stable range of instruments for enriching ontologies with lexical content, and of formalizing the model and associated format for representing this information, we have developed a dedicated component which, together with the LW library, can be embedded in ontology based tools and applications needing to incorporate linguistic content.

4.2.1. Issues in representing the integrated information

So far, in tools exploiting the Linguistic Watermark framework, like the already cited OntoLing, the association between linguistic content and ontological data has been projected over standard RDFS/OWL predicates. Thus, the rdfs:label property were used for addressing short lexical objects like terms, words (used both to provide synonymical expressions as well as to provide translation for different languages) or even conceptual entities like WordNet (Miller, Beckwith, Fellbaum, Gross, & Miller, 1993) synsets, while rdfs:comment has been commonly associated to wider descriptions like those which could be extracted from word glosses and terminology definitions.

This choice, though guaranteeing a complete adherence to widely accepted standards on the one side, offered poor representation primitives: two major problems concerned the loss of information about the nature of the attached linguistic objects, which became mere strings pointed by the rdfs properties, and difficulty in the integration of artificial entities. As an example, a WordNet synset, being a kind of lw:SemanticIndex, were linked to ontology objects through the rdfs:label property, filling the xml:lang attribute of this predicate with a short namespace for indicating its association to WordNet (and the specific WordNet version), while xml:lang requires codes conforming to the official standard code ISO 3166-1-alpha-2. Clearly, a compromise between popularity, immediateness and completeness of the model needed to be found.

4.2.2. The OLW integration model

In modeling our framework for the integration of ontological and linguistic content, we have taken into consideration the following requisites, which should allow for:

1. Reporting quantitative and qualitative information on the overall process of enriching an ontology with content from a linguistic resource (this was the primary objective of the OLW metadata ontology)
2. Keeping track (at least maintain the possibility to do that) of the source used for enriching the content
3. Being able to properly map different kind of linguistic entities (words, linguistic/semantic relations etc…) with (structures of) ontological objects
4. Giving the user the possibility of adopting resources' specific objects (e.g. FrameNet frames or WordNet synsets) for enriching an ontology
5. Embedding existing models for integration of ontologies and linguistic entities, still respecting the above priorities
6. Assessing reliable links between ontological and linguistic objects as well as taking into account for probabilistic matches produced by automatic enrichment tools (which could also be used for evaluation purposes)

The first requisite has been satisfied by defining a set of meta-descriptors – represented through object properties with domain set to owl:Ontology – for providing an overview of the "linguistic expressiveness" of ontologies. These properties may prove to be helpful for services/agents which, having to map/merge/align/mediate different ontologies, may be willing to invoke the proper linguistic resources for supporting this task. These mediators can thus beneficiate of the overall statistical information provided by the OWL metadata, without inspecting the entire ontologies' content. This part of the OLW has already been described in details in (Pazienza, Sguera & Stellato; 2007).

The second, third and fourth requisites have been accomplished by extending the LW; in its first incarnation, which served solely as a conceptual driver for the software library, the LW was able to express descriptions of linguistic resources, without predicating about their specific content. Now it has been extended to make possible the instantiation of objects from the described resources. The example in Figure 3 shows fragments originating from three different ontologies: the first fragment is a description of WordNet synset 100001740

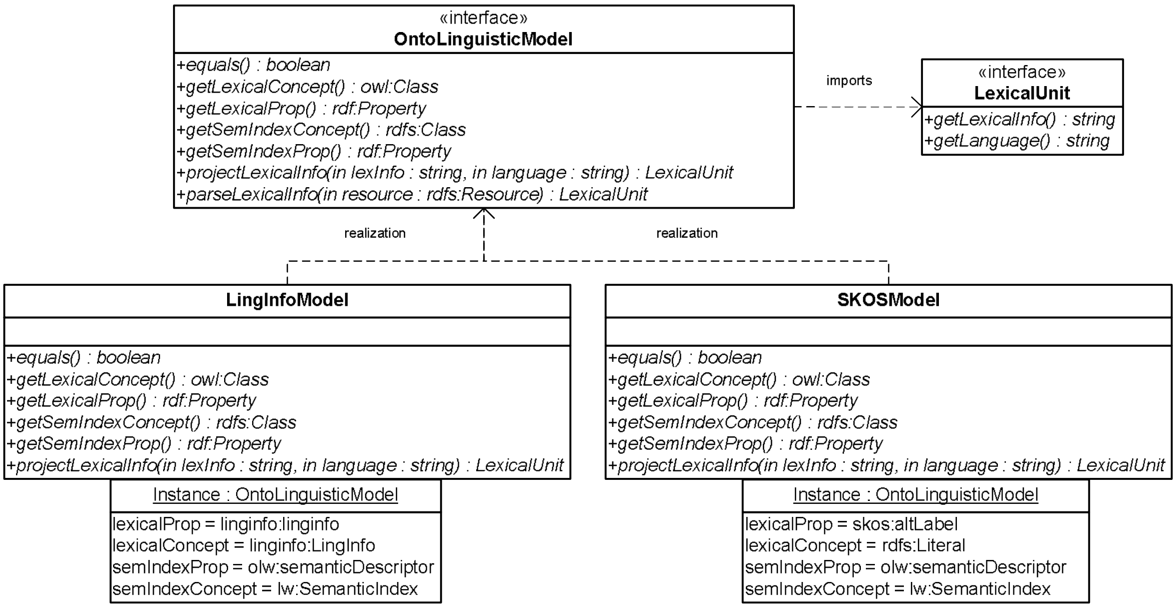

«interface»
OntoLinguisticModel
+equals() : boolean
+getLexicalConcept() : owl:Class
+getLexicalProp() : rdf:Property
+getSemIndexConcept() : rdfs:Class
+getSemIndexProp() : rdf:Property
+projectLexicalInfo(in lexInfo : string, in language : string) : LexicalUnit
+parseLexicalInfo(in resource : rdfs:Resource) : LexicalUnit
imports
«interface»
LexicalUnit
+getLexicalInfo() : string
+getLanguage() : string
realization
realization
LingInfoModel
+equals() : boolean
+getLexicalConcept() : owl:Class
+getLexicalProp() : rdf:Property
+getSemIndexConcept() : rdfs:Class
+getSemIndexProp() : rdf:Property
+projectLexicalInfo(in lexInfo : string, in language : string) : LexicalUnit
SKOSModel
+equals() : boolean
+getLexicalConcept() : owl:Class
+getLexicalProp() : rdf:Property
+getSemIndexConcept() : rdfs:Class
+getSemIndexProp() : rdf:Property
+projectLexicalInfo(in lexInfo : string, in language : string) : LexicalUnit
Instance : OntoLinguisticModel
lexicalProp = linginfo:linginfo
lexicalConcept = linginfo:LingInfo
semIndexProp = olw:semanticDescriptor
semIndexConcept = lw:SemanticIndex
Instance : OntoLinguisticModel
lexicalProp = skos:altLabel
lexicalConcept = rdfs:Literal
semIndexProp = olw:semanticDescriptor
semIndexConcept = lw:SemanticIndex

Figure 4: two examples of OntoLinguisticModel implementation

originating from the WordNet-RDF vocabulary developed by the WordNet task force of the W3C (http://www.w3.org/TR/wordnet-rdf/); the second one is the binding of concept wn20schema:Synset to the lw:SemanticIndex, through a rdfs:subClassOf relationship. Finally, a certain Noun concept coming from a fictitious ontology is enriched with the meaning expressed by the above synset, through the owl:semanticDescriptor property. With this extensible pattern, the LW+OLW offer reusable vocabularies for describing linguistic resources which drive the behavior of software applications serving the same task, while specific extensions (both in terms of ontologies and software components) can be added to describe specific lexical and semantic objects from new resources, without requiring modifications to the core vocabulary nor to the original application.

4.2.3. Compatibility with existing (proposed) models

As previously mentioned, several formats exists or have been proposed for integrating ontological content with linguistic information.

While we did not intend to propose a new one, we tried to obtain cross-compatibility with available standards and proposed models, by gearing our software library with a OntoLinguisticModel interface, consisting of a series of enrichment/retrieval operations defined upon abstract “slots” for representing linguistic information. These slots can be then implemented according to a specific onto-linguistic representation model, by specifying the properties and concepts used to map integrate linguistic information with ontological one.

Obviously, it is impossible to foresee in advance all the characteristics of each model/interface-implementation which could be integrated in the future, thus we provided a specific *project/decode* feature for projecting the linguistic information extracted from linguistic resources according to the LW ontology, towards the (possibly more fine-grained) adopted ontolinguistic model. For evaluative (see next section) and comparative purpose in general, we demand to each specific implementation the specifications of equivalence between the locally defined linguistic objects.

Implementations of OntoLinguisticModel have been developed for the traditionally adopted RDFS annotation properties (rdfs:label and rdfs:comment), for the base SKOS vocabulary (by extending the above with skos:prefLabel and skos:altLabel), for SKOS + SKOS-Mapping[6] vocabularies (thus including skos:broader/skos:narrower and skos:related, to map ontology concepts with instances of lw:SemanticIndex from the LW ontology) and, finally, for the LingInfo model, by wrapping the linginfo:linginfo property and linginfo:LingInfo class.

4.2.4. The OLW integration model

Figure 4 shows (hiding minor details) how two available linguistic models have been mapped to our meta-model and wrapped inside our library. In the reported examples, pointers to lw:SemanticIndex have been implemented by using OLW and LW descriptors, since there were no correspondence for them in the addressed models. Notice how the main mapping completely hides any information associated to more complex specifications of the concepts of the wrapped models. For example, in the LingInfo wrapper, the lexical element associated to an ontology object is bound to the linginfo:term property of the created linginfo:LingInfo object (while it is directly mapped to the value of skos:altLabel in the SKOS case); in the same manner, the language parameter of the projectLexicalInfo() method is associated to the linginfo:lang property for the same object, whereas it is directly mapped to the xml:lang attribute of the skos:altLabel property in the SKOS case.

A similar process will be carried out in the future for frame-based resources, once RDF descriptions and research about mapping of their content to ontologies will reach full maturity and stableness. The above integration model satisfied our fifth requirement, while the resolution

[6] http://www.w3.org/2004/02/skos/mapping/spec/

of the sixth one is part of the discussion presented in the next section.

5. The evaluation framework

The newly developed OLW Library provides a framework for evaluating the quality of algorithms for Linguistic Enrichment of ontologies with respect to previously defined reference standards.

Linguistic Enrichment algorithms can be evaluated by comparing the results of an Enrichment Process (E) to a reference enrichment document, which we call "the Oracle" (O). The usual approach for evaluating the results of process E is to consider them as sets of correspondences and to apply precision and recall originating from Information Retrieval (Van Rijsbergen, 1975) and adapted to the matching task. Precision and recall are thus the ratio of the number of true positive $|O \cap E|$ on that of the retrieved correspondences ($|E|$) and those expected ($|O|$) respectively.

The OLW library can accept pairs of linguistic enrichment documents (that is: ontologies with integrated linguistic content), where one is the Oracle and the other one is the result to be tested, providing that the following extensions are included in the library and properly configured:

- *Enrichment Model* and related software extension (see section 4.2.3)
- *Resource*(s) *description* (and their wrapper implementation) used for enrichment (see sections 3.1 and 4.2.2)
- *Match Specification and Evaluation (MSE)* extension, if different enrichment entries differ from simple links between ontological and linguistic objects

With the ones above, the library is able to seek the enrichment properties (at least, those which need to be considered) in the ontology documents (first extension) and to properly identify the elements used for the enrichment (second extension).

The third one is an extension needed for those cases where an algorithm produces any kind of probabilistic/quantitative result, so that the enrichment links in the tested document cannot be evaluated just in terms of correct/wrong matches versus those in the Oracle.

If this extension is included, an ontological representation for qualifying its results is to be provided (usually, it just requires a property with domain set to the adopted enrichment properties, that is olw:lexicalization olw:semanticDescriptor and range set to the description of the non-conventional link). A proper extension module for the library needs then to be plugged, with a parser for the above description and associated modifiers for adapting the precision/recall measure to the introduced range of values.

Inter-annotator agreement can as well be measured against two reports about the enrichment, compiled by human annotators (with no further requirement apart from the ones above).

6. Conclusions

In this paper we presented the Linguistic Watermark suite, a set of RDF vocabularies used to uniformly represent linguistic knowledge in heterogeneous linguistic resources and to enable shared integration-with and accessibility-from different computational ontologies. In this context the main features of LW library have been also illustrated, a set of JAVA-based software tools and interfaces developed for integrating ontologies and linguistic resources. This library exploits LW vocabularies to establish adequate mappings between linguistic resources and linguistic interfaces, helping knowledge engineers to implement their hybrid semantic systems. We expect that our work may give a contribution to the standardization of models, methodologies and tools for the effective integration of ontologies and linguistic resources; moreover, the possibly adoption by R&D communities of the general framework we presented might inspire, in the next future, new contests for the evaluation of linguistic enrichment of ontologies.

7. References

Baker, C., Fillmore, C., & Lowe, J. (1998). The Berkeley FrameNet project. *COLING-ACL*. Montreal, Canada.

Basili, R., Vindigni, M., & Zanzotto, F. M. (2003). Integrating Ontological and Linguistic Knowledge for Conceptual Information Extraction. *IEEE/WIC International Conference on Web Intelligence.* Washington, DC, USA.

Beth, L. (1993). *English verb classes and alternations: A preliminary investigation* (Vol. XVIII). Chicago: University of Chicago Press.

Buitelaar, P., Declerck, T., Frank, A., Racioppa, S., Kiesel, M., Sintek, M., et al. (2006). LingInfo: Design and Applications of a Model for the Integration of Linguistic Information in Ontologies. *Workshop on Interfacing Ontologies and Lexical Resources for Semantic Web Technologies (OntoLex2006)*, hosted by LREC Conference. Genoa, Italy.

Cappelli, A., Giovannetti, E., & Michelassi, P. (2004). Ontological Knowledge and Language in Modelling Classical Architectonic Structures. *Ontology and Lexical Resources – OntoLex 2004)*, hosted by LREC Conference. Lisboa, Portugal.

Euzenat, J. (2004). An API for Ontology Alignment. In S. A. McIlraith, D. Plexousakis, & F. van Harmelen (Ed.), *The Semantic Web - ISWC 2004: Third International Semantic Web Conference. 3298*, pp. 698-712. Hiroshima, Japan: Springer.

Euzenat, J. (2007). Semantic Precision and Recall for Ontology Alignment Evaluation. In M. M. Veloso (Ed.), *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence*, (pp. 348-353). Hyderabad, India, January 6-12.

Fillmore, C. (1968). The Case for Case. In E. Bach, R. T. Harms, & B. a. Harms (Ed.), *Universals in Linguistic Theory* (pp. 1-88). New York: Holt, Rinehart, and Winston.

Huang, C. (2004). Sinica BOW: Integrating bilingual WordNet and SUMO Ontology. *Ontology and Lexical Resources – OntoLex 2004,)*, hosted by LREC Conference. Lisboa, Portugal.

Kipper, K., Trang Dang, H., & Palmer, M. (2000). Class-Based Construction of a Verb Lexicon. *AAAI-2000 Seventeenth National Conference on Artificial Intelligence.* Austin, TX.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. Chicago: University of Chicago Press.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1993). *Introduction to WordNet: An On-line Lexical Database.*

Palmer, M., Kingsbury, P., & Gildea, D. (2005). The Proposition Bank: An annotated Corpus of Semantic Roles. *Computational Linguistics , 31* (1), 71-106.

Pazienza, M. T., & Stellato, A. (2006). An open and scalable framework for enriching ontologies with natural language content. *The 19th International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems (IEA/AIE'06), special session on Ontology & Text.* Annecy, France.

Pazienza, M. T., & Stellato, A. (2006b). Exploiting Linguistic Resources for building linguistically motivated ontologies in the Semantic Web. *Workshop on Interfacing Ontologies and Lexical Resources for Semantic Web Technologies (OntoLex2006)*, hosted by LREC Conference. Genoa, Italy.

Pazienza, M. T., & Stellato, A. (2006c). Linguistic Enrichment of Ontologies: a methodological framework. *Workshop on Interfacing Ontologies and Lexical Resources for Semantic Web Technologies (OntoLex2006).* Genoa, Italy.

Pazienza, M. T., Sguera, S., & Stellato, A. (2007). Let's talk about our "being": A linguistic-based ontology framework for coordinating agents. (R. Ferrario, & L. Prévot, Eds.) *Applied Ontology, special issue on Formal Ontologies for Communicating Agents , 2* (3-4), 305-332.

Peter, H., Sack, H. Beckstein, C. (2006). SMARTINDEXER – Amalgamating Ontologies and Lexical Resources for Document Indexing. *Workshop on Interfacing Ontologies and Lexical Resources for Semantic Web Technologies (OntoLex2006)*, hosted by LREC Conference. Genoa, Italy

Philpot, A., Hovy, E., & Pantel, P. (2005). The Omega Ontology. Ontology and Lexical Resources. *OntoLex2005 - Ontologies and Lexical Resources.* Jeju Island, South Korea.

Prevot, L., Borgo, S., & Oltramari, A. (2005). Interfacing Ontologies and Lexical Resources. *Workshop on Ontologies and Lexical resources (OntoLex2005),* hosted by IJCNLP Conference. Jeju Island, South Korea.

Scheffczyk, J., Baker, C. F., & Narayanan, S. (2006). Ontology-based Reasoning about Lexical Resources. *Workshop on Interfacing Ontologies and Lexical Resources for Semantic Web Technologies (OntoLex2006).* Genoa, Italy.

Van Rijsbergen, C. J. (1975). *Information Retrieval.* London, United Kingdom: Butterworths.

Resources and Tools for Ontology-Based Semantic Annotation

Kiril Simov and Petya Osenova
Linguistic Modeling Department,
Institute for Parallel Processing,
Bulgarian Academy of Sciences
25A Acad. G. Bonchev St.
Sofia 1113, Bulgaria
kivs@bultreebank.org, petya@bultreebank.org

Abstract

This paper presents the resources and tools, which facilitate the ontology-based semantic annotation of domain texts, and subsequently – the semantic search. Some of these resources are language independent, such as the domain ontology. Some depend on the specific language: terminological lexicons, annotation grammars, sense disambiguation rules, relation annotation rules, gold standard corpus (used in the process of ontology creation). The combination of these tools defines ontology-to-text relation. Implementing different instantiations of this relation we could achieve semantic annotation of text with different granularity and for different tasks. The ideas are based on the empirical observations within two European projects.

1. Introduction

In this paper we present our work on defining of the *ontology-to-text* relation and its instantiations for several languages and two domains. This relation is important with respect to tasks, such as ontology annotation, ontology based search (or semantic search), information extraction, ontology learning and ontology browsing. The work described here was carried out within two European projects: LT4eL[1] (*Language Technology for eLearning*) and AsIsKnown[2] (*A Semantic-Based Knowledge Flow System for the European Home Textiles Industry*).

The relation *ontology-to-text* shows how the elements of ontology (concepts, relations, instances) are realized within the text of multimedia documents. Our model of the relation comprises four components: *ontology*, *lexicon*, *grammar*, *text.* The *ontology* is a domain one mapped to an upper part. The *lexicon* contains the terms (grouped on the basis of synonymy) and associated contextual information and grammatical features. The *grammar* contains the syntactic knowledge about the forms in which the terms might be realized in the text. It also contains some disambiguation information about the term in a certain context (in case it is ambiguous). The *text* is a description of a part of the domain in question for which we would like to explicate the ontological information. In the real life the situation is more complex, because the texts usually contain other means to represent the same concept (relation, instance) out of the terms in the lexicons. In order to handle such cases the grammar needs to contain also parts devoted to such phenomena as coreferential relations, metonymy, metaphorical usages, etc. In the actual realization of the relation in the two projects we started with annotation of concepts, but we will continue with relations and instances in a follow-up project. In the paper we will be discussing mainly concept annotation.

In many respects our model of the *ontology-to-text* relation is subsumed by more general and elaborated models, such as LingInfo (see (Romanelli et al., 2007), (Buitelaar et al., 2006a) and (Buitelaar et al., 2006b)). Main differences are in: (1) the definition of the model – ontology-based model definition in LingInfo vs. XML-based resource representation oriented to particular processing tools in our work; and (2) coverage of the model – LingInfo covers all multimedia information objects like images, sounds, etc., while we focused only on the linguistic level, represented by texts. Thus, we might consider our work as an example of instantiation of some elements from the LingInfo model too. In future we envisage the incorporation of annotated images, since they - together with the texts – contribute to the better semantic search in a domain.

We would like also to stress that in practice there are many instantiations of the *ontology-to-text* relation. For example, see the ontology-based named entity annotation presented in (Kiryakov et al., 2004), among others.

The structure of the paper is as follows: first, we present in short the two projects and the role of ontology in them; then we present the ontology creation methodology employed in the projects (there are some differences in the two projects in this respect); in section 4 the elements of our model on the ontology-to-text relation are described; the last section polemizes the place of the current paper within other works, and concludes the paper.

2. The role of the ontology within the two projects

We had to construct domain ontologies for both abovementioned European projects. The main usage of these ontologies concerned, on the one hand, the annotation of domain texts for search purposes, and on the other hand, the connection among multilingual domain material or among the specific ‘views’ of the various participants in the same domain. Let us point in short to the specificities of each project. The LT4eL project aims at demonstrating the relevance of the language technology and ontology document annotation for improving the usability of learning management systems (LMS) within the learning process. Thus, a semantic search module had to be created. This module built on concept annotated documents. With the help of the domain ontology sets of learning texts have been annotated in various subdomains and in eight languages (Bulgarian, Dutch, German, English, Czech, Polish, Portuguese and Romanian). The semantic

[1] http://www.lt4el.eu/
[2] http://www.asisknown.org/

search increased the precision and the speed in finding the most relevant documents for a topic.

The AsIsKnown project is developing an architecture of interrelated modules for speeding up the process of communication among agents in the textile industry. Here the challenge is not only the cross-lingual access to the system, but also the different communication preferences of the agents in this business area. The ontology was used as follows: in the annotation of fashion magazines for search and trend analysis; as an input-output communication system among producers, retailers and clients. The languages involved in the project are Bulgarian, English, French and German.

In addition to the semantic annotation and search the ontology has to support the communication with the user for query definition and result explanation. Thus, for example, it is necessary for the users to be able to navigate over ontology in a natural for them way. In our view this task has to be done via the natural language of the user.

3. Ontology creation

In this section we briefly outline the methodology for ontology creation used within the two projects. One of the main requirements for the methodology is that the initial version of the ontology is created from existing resources. The involvement of the domain experts in the process of the ontology creation is done at a later stage. In this way we attempt to maximize their contribution. Here we present the main steps of the methodology as it was applied within the project AsIsKnown:[3]

Processing of the standards and vocabularies in the domain

We consider standards in the domain as reliable sources of conceptual information. Being created by leading experts in the domain with the goal to facilitate the whole process of production and usage of the home textile, the standards can be viewed as "expert questionnaires" usually used in the process of knowledge acquisition. Thus, we expected to find definitions of the most important concepts and relations in the domain. The definitions also helped us to establish the main relationships between the extracted concepts. As a means for the extraction of the concepts and the relations we have been using a treebank constructed semi-automatically over the text of the standards. Then we inspected manually the analysis in order to identify the relevant knowledge. The result from this step was a list of (concept) *terms* (in English), a list of *relations* (relational terms), a list of triples - (*term1 relation term2*). These lists became the backbone of the ontology. The list of relations includes general ontological relations like *is-a*, *part-of*, etc. and domain specific relations. The extracted terms in many cases were equipped with a definition. These definitions had to reflect the triples for the term and the features of the relations.

Formalization of the terms

The next step is to define formal definitions of the extracted concepts and relations in OWL-DL. We have selected OWL-DL, because there exist implemented reasoners for it. For each term in the term list we constructed a class definition in OWL-DL. We did the same for each relational term. We also encoded the additional information in the definitions of the terms and the relations. The result of this step was an initial formal version of the ontology.

Link to an upper ontology

The establishing of the connection between the upper and the domain ontology helped us to check the consistency of the domain ontology with respect to the ontology construction methodology behind the upper ontology and to inherit the knowledge encoded in the upper ontology. Also the upper ontology provided general ontological information when it was required during the usage of the ontology. We selected DOLCE Ontology (Masolo et a., 2003) as upper ontology for several reasons: (1) it is constructed on rigorous basis which reflects the OntoClean methodology (Guarino and Welty, 2002); (2) it is represented in OWL-DL; (3) the authors of the ontology provide us comments and help on the alignment of the domain ontology to DOLCE. The alignment between the two ontologies is facilitated by OntoWordNet (Gangemi et al., 2003) - a version of WordNet aligned to DOLCE. OntoWordNet ensures more understandable concepts (more specific and closer to the domain) and the mapping between the concepts is easier. The result from this step is the better structuring of the initial lists of concepts and relations. Also relations and axioms were inherited from DOLCE to the domain ontology.

Evaluation by domain experts

The evaluation of the first version of the ontology has been done in two ways:

Practical evaluation

The ontology is evaluated in the process of incorporation and integration within the overall project architecture.

Expert evaluation

The ontology is reviewed by domain experts in the project. The review is mediated by questionnaires constructed on the basis of the already constructed first version of the ontology. Here is an example from such a questionnaire on carpets:

Nr.	Question	Answer	Comment
3.	What is the difference between **Loop Column** and **Loop Row**?	a) *Loop Column* shows a product direction b) *Loop Row* shows a transverse direction	See 5.12 and 5.13, ISO 2424
4.	Does **Tuft Column** (a line of tufts essentially parallel to the direction of manufacture) consist of **Tuft**?	Yes, if the meaning of Tuft is Cut Pile in this case (q.v. ISO 2424, 5.6)	The term "tuft" describes a manufacturing technique too.

Besides the evaluation by domain experts the ontology is evaluated on the basis of annotation of a corpus of representative domain documents. In this way some adequate coverage of the ontology is ensured.

Documentation

In the process of construction of the ontology we keep track on the sources of each concept, relation, etc.

Lexicons and concept annotation grammar creation

This step is the creation of an instance of the

[3] The differences within the LT4eL project will be discussed later.

ontology-to-text relation for the given ontology. The actual model of the relation is given in the next section. In the two projects we had to create instances in several languages as it was mentioned above.

The methodology outlined in this section was successfully applied to the construction of both domain ontologies. The evaluation is still an on-going process. In case of LT4eL we did not have standards in the domain and this is why we started with the keywords annotated manually by the partners in the learning objects. Then for the keywords in the domain we collected definitions from different sources (terminological lexicons, Internet) and these definitions were the initial source for creation of the first version of the ontology.

4. Ontology-to-Text relation

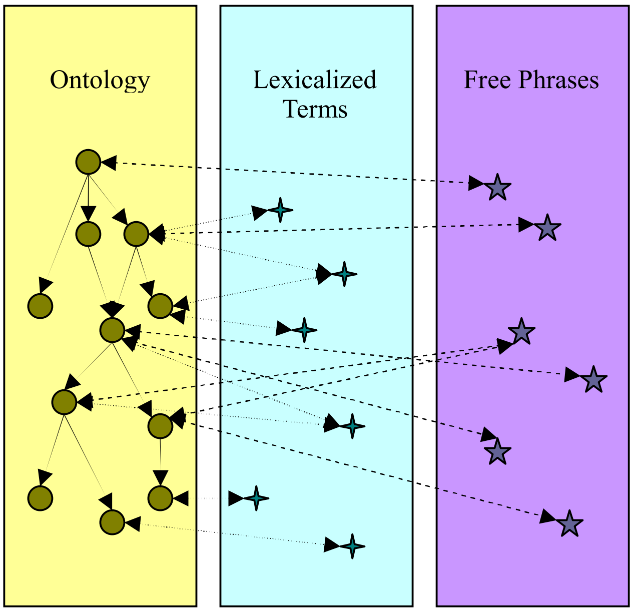
In this section we represent the two main components that define the ontology-to-text relation necessary to support the tasks within our projects. These components are: (terminological) lexicon and concept annotation grammar.

The lexicon plays twofold role in our architecture. First, it interrelates the concepts in the ontology to the lexical knowledge used by the grammar in order to recognize the role of the concepts in the text. Second, the lexicon represents the main interface between the user and the ontology. This interface allows for the ontology to be navigated or represented in a natural for the user way. For example, the concepts and relations might be named with terms used by the users in their everyday activities and in their own natural language (e.g. Bulgarian). This could be considered as a first step to a contextualized usage of the ontology in a sense that the ontology could be viewed through different terms depending on the context. For example, the color names will vary from very specific terms within the domain of carpet production to more common names used when the same carpet is part of an interior design.

Thus, the lexical items contain the following information: a term, contextual information determining the context of the term usage, grammatical features determining the syntactic realization within the text. In the current implementation of the lexicons the contextual information is simplified to a list of a few types of users (producer, retailer, etc).

With respect to the relations between the terms in the lexicon and the concepts in the ontology, there are two main problems: (1) there is no lexicalized term for some of the concepts in the ontology, and (2) there are lexical terms in the language of the domain which lack corresponding concepts in the ontology, which represent the meaning of the terms.

The first problem is overcome by writing down in the lexicon also non-lexicalized (fully compositional) phrases to be represented. Even more, we encourage the lexicon builders to add more terms and phrases to the lexicons for a given concept in order to represent as many ways of expressing the concept in the language as possible. These different phrases or terms for a given concept are used as a basis for construction of the annotation grammar. Having them, we might capture different wordings of the same meaning in the text. The picture below shows the mapping varieties. It depicts the realization of the concepts (similarly for relations and instances) in the language. The concepts are language independent and they might be represented within a natural language as form(s) of a lexicalized term, or as a free phrase. In general, a concept might have a few terms connected to it and a (potentially) unlimited number of free phrases expressing this concept in the language[4]. Some of the free phrases receive their meaning compositionally regardless their usage in the text, other free phrases denote the corresponding concept only in a particular context. In our lexicons we decided to register as many free phrases as possible in order to have better recall on the semantic annotation task. In case of a concept that is not-lexicalized in a given language we require at least one free phrase to be provided for this concept.



We could summarize the connection between the ontology and the lexicons in the following way: the ontology represents the semantic knowledge in form of concepts and relations with appropriate axioms; and the lexicons represent the ways in which these concepts can be realized in texts in the corresponding languages. Of course, the ways in which a concept could be represented in the text are potentially infinite in number, thus, we could hope to represent in our lexicons only the most frequent and important terms and phrases. Here is an example of an entry from the Dutch lexicon:

```
<entry id="id60">
   <owl:Class rdf:about="lt4el:BarWithButtons">
     <rdfs:subClassOf>
       <owl:Class rdf:about="lt4el:Window"/>
     </rdfs:subClassOf>
   </owl:Class>
   <def>A horizontal or vertical bar as a part of a window,
      that contains buttons, icons.</def>
   <termg lang="nl">
     <term shead="1">werkbalk</term>
     <term>balk</term>
     <term type="nonlex">balk met knoppen</term>
     <term>menubalk</term>
   </termg>
 </entry>
```

[4] The presence of free phrases in the lexicon is also motivated by the fact that the lexicalization is not a discrete feature. There are many different degrees of lexicalization. Thus the free phrases are the extreme end of the scale.

Each entry of the lexicons contains three types of information: (1) information about the concept from the ontology which represents the meaning for the terms in the entry; (2) explanation of the concept meaning in English; and (3) a set of terms in a given language that have the meaning expressed by the concept. The concept part of the entry provides minimum information for formal definition of the concept. The English explanation of the concept meaning facilitates the human understanding. The set of terms stands for different wordings of the concept in the corresponding language. One of the terms is the representative for the term set. Note that this is a somewhat arbitrary decision, which might depend on frequency of term usage or specialist's intuition. This representative term will be used where just one of terms from the set is necessary to be used, for example as an item of a menu. In the example above we present the set of Dutch terms for the concept *lt4el:BarWithButtons*. One of the term is non-lexicalized - attribute ***type*** with value ***nonlex***. The first term is representative for the term set and it is marked-up with attribute ***shead*** with value 1. In this way we determine which term to be used for ontology browsing if there is no contextual information for the type of users.

The second component of the ontology-to-text relation, the concept annotation grammar, is ideally considered as an extension of a general language deep grammar which is adopted to the concept annotation task. Minimally, the concept annotation grammar consists of a chunk grammar for concept annotation and (sense) disambiguation rules. The chunk grammar for each term in the lexicon contains at least one grammar rule for recognition of the term. As a preprocessing step we consider annotation with grammatical features and lemmatization of the text. The disambiguation rules exploit the local context in terms of grammatical features, semantic annotation and syntactic structure, and alsp the global context such as topic of the text, discourse segmentation, etc. Currently we have implemented chunk grammars for several languages. The disambiguation rules are under development.

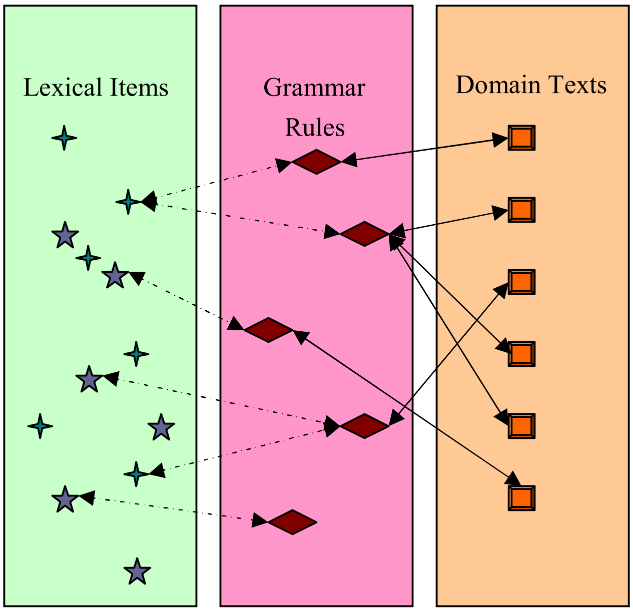
For the implementation of the annotation grammar we rely on the grammar facilities of the CLaRK System (Simov et al., 2001). The structure of each grammar rule in CLaRK is defined by the following DTD fragment:

```
<!ELEMENT line (LC?, RE, RC?, RM, Comment?) >
<!ELEMENT LC (#PCDATA)>
<!ELEMENT RC (#PCDATA)>
<!ELEMENT RE (#PCDATA)>
<!ELEMENT RM (#PCDATA)>
<!ELEMENT Comment (#PCDATA)>
```

Each rule is represented as a line element. The rule consists of regular expression (*RE*) and category (*RM* = return markup). The regular expression is evaluated over the content of a given XML element and could recognize tokens and/or annotated data. The return markup is represented as an XML fragment which is substituted for the recognized part of the content of the element. Additionally, the user could use regular expressions to restrict the context in which the regular expression is evaluated successfully. The *LC* element contains a regular expression for the left context and the *RC* for the right one. The element Comment is for human use. The application of the grammar is governed by *Xpath* expressions which provide additional mechanism for accurate annotation of a given XML document. Thus, the CLaRK grammar is a good choice for implementation of the initial annotation grammar.

The creation of the actual annotation grammars started with the terms in the lexicons for the corresponding languages. Each term was lemmatized and the lemmatized form of the term was converted into regular expression of grammar rules. Each concept related to the term is stored in the return markup of the corresponding rule. Thus, if a term is ambiguous, then the corresponding rule in the grammar contains reference to all concepts related to the term.

The following picture depicts the relations between lexical items, grammar rules and the text:



The relations between the different elements of the models are as follows. A lexical item could have more than one grammar rule associated to it depending on the word order and the grammatical realization of the lexical item. Two lexical items could share a grammar rule if they have the same wording, but they are connected to different concepts in the ontology. Each grammar rule could recognize zero or several text chunks.

The relation ontology-to-text implemented in this way provides facilities for solving different tasks, such as ontology search (including crosslingual search), ontology browsing, ontology learning. In order to support multilingual access to semantic annotated corpus we have to implement the relation for several languages using the same ontology as starting point. In this way we implement a mapping between the lexicons in these languages and also comparable annotation of texts in them.

We have been using the relations between the various elements for the task of ontology-based search. The connection from ontology via lexicon to grammars is relied on for the concept annotation of the text. In this way we established a connection between the ontology and the texts. The relation between the lexicon and the ontology is used for definition of user queries with respect to the appropriate segments within the documents. The annotation of texts in different languages on the basis of the same ontology could facilitate the definition of similarity metrics between such texts.

In AsIsKnown project we also exploited a domain independent partial grammar which supports the domain specific grammar providing additional context features.

5. Discussion and Conclusion

Our approach gains in many respects from such works as WordNet (Fellbaum, 1998), EuroWordNet (Vossen, 1998), SIMPLE (Lenci et al., 2000). The mapping between the language specific lexicons was facilitated by the ontology. Our model shares common features with other lexicon models: with WordNet-like Fellbaum, 1998; Vossen, 1998) lexicons we share the idea of grouping lexical items around a common meaning and in this respect the term groups in our model correspond to synsets in WordNet model. The difference in our case is that the meaning is defined independently in the ontology. With SIMPLE model (Lenci et al., 2000) we share the idea to define the meaning of lexical items by means of the ontology, but we differ in the selection of the ontology which in our case represents the domain of interest, and in the case of SIMPLE reflects the lexicon model. With the LingInfo model (Romanelli et al., 2007; Buitelaar et al., 2006a; Buitelaar et al., 2006b) we share the idea that grammatical and context information also needs to be presented in a connection to the ontology, but we differ in the implementation of the model and the degree of realization of the concrete language resources and tools.

In the paper we present a model for the ontology-to-text relation supporting semantic annotation. We assume the central role of the ontology on which all the other resources and tools depend. In future we envisage to implement an interaction with a general lexica and grammar. Some initial experiments are done by domain specific rules for exploiting the general analyses during domain semantic annotation. The model was successfully exploited in two EU projects for concept annotation and semantic search.

The relation annotation requires in our view much more work on the level of general language processing in tasks like coreference resolution, metonymy patterns recognition, bridging relation annotation, etc. Some of these tasks require ontology based information and our model allows for ontology centered linguistic knowledge representation as much as knowledge in the lexicon and in the grammar is always related to the ontology. When it is necessary, information from general lexicons and grammar is transferred to the domain in an appropriate form. Thus we ensure interaction between general language processing tools and resources, and the domain specific ones.

6. Acknowledgements

This work has been supported by two FP6 European projects: LT4eL (Language Technology for eLearning) (FP6-027391) and AsIsKnown (A Semantic-Based Knowledge Flow System for the European Home Textiles Industry) (FP6-028044).

We would like to thank also the three anonymous reviewers for their valuable remarks and suggestions.

7. References

Buitelaar, P., Declerck, T., Frank, A., Racioppa, S., Kiesel, M., Sintek, M., Engel, R., Romanelli, M., Sonntag, D., Loos, B., Micelli, V., Porzel, R., and Cimiano, Ph. (2006a). LingInfo: Design and Applications of a Model for the Integration of Linguistic Information in Ontologies. In: Proc. of OntoLex06, a Workshop at LREC, Genoa, Italy, May 2006.

Buitelaar, P., Sintek, M., and Kiesel, M. (2006b). A Lexicon Model for Multilingual/Multimedia Ontologies In: Proceedings of the 3rd European Semantic Web Conference (ESWC06), Budva, Montenegro, June 2006.

Fellbaum, Ch. (1998). Editor. WORDNET: an electronic lexical database. MIT Press.

Gangemi, A., Navigli, R., and Velardi, P. (2003). The OntoWordNet Project: extension and axiomatization of conceptual relations in WordNet. Meersman R, et al. (eds.), Proceedings of ODBASE03 Conference, Springer, 2003.

Guarino, N., and Welty, C. (2002). Evaluating Ontological Decisions with OntoClean. Communications of the ACM, 45(2): 61-65.

Kiryakov, A., Popov, B., Ognyanov, D., Manov, D., Kirilov, A., and Goranov, M. 2004. *Semantic Annotation, Indexing, and Retrieval.* Elsevier's Journal of Web Semantics, Vol. 1, ISWC2003 special issue (2), 2004. http://www.websemanticsjournal.org/

Lenci, A., Busa, F., Ruimy, N., Gola, E., Monachini, M. Calzolari, N., Zampolli, A., Guimier, E., Recourcé, G., Humphreys, L., Von Rekovsky, U., Ogonowski, A., McCauley, C., Peters, W., Peters, I., Gaizauskas, R., and Villegas, M. (2000). SIMPLE Work Package 2 - Linguistic Specifications. Deliverable D2.1. ILC-CNR, Pisa, Italy.

Masolo, C., Borgo, S., Gangemi, A., Guarino, N., and Oltramari, A. (2002). Ontology Library (final). WonderWeb Deliverable D18, December 2003. http://www.loa-cnr.it/Publications.html.

Romanelli, M., Buitelaar, P., and Sintek, M. (2007). Modeling Linguistic Facets of Multimedia Content for Semantic Annotation. In: Proceedings of SAMT07 (International Conference on Semantics And digital Media Technologies), Genova, Italy, Dec. 2007. pp 240-251.

Simov, K., Peev, Z., Kouylekov, M., Simov, A., Dimitrov, M., Kiryakov, A. (2001). CLaRK - an XML-based System for Corpora Development. In: *Proc. of the Corpus Linguistics 2001 Conference*. Lancaster, UK.

Vossen P. (1998). Editor. EuroWordNet General Document. Version 3, Final, July 19, 1999, http://www.hum.uva.nl/~ewn.

Building a free French wordnet from multilingual resources

Benoît Sagot[1], Darja Fišer[2]
[1]Alpage, INRIA / Université Paris 7
30 rue du Château des rentiers, 75013 Paris, France
[2]Faculty of Arts, University of Ljubljana
Aškerčeva 2, 1000 Ljubljana, Slovenia
E-mail: benoit.sagot@inria.fr, darja.fiser@guest.arnes.si

Abstract

This paper describes automatic construction of a freely-available wordnet for French (WOLF) based on Princeton WordNet (PWN) by using various multilingual resources. Polysemous words were dealt with an approach in which a parallel corpus for five languages was word-aligned and the extracted multilingual lexicon was disambiguated with the existing wordnets for these languages. On the other hand, a bilingual approach sufficed to acquire equivalents for monosemous words. Bilingual lexicons were extracted from Wikipedia and thesauri. The results obtained from each resource were merged and ranked according to the number of resources yielding the same literal. Automatic evaluation of the merged wordnet was performed with the French WordNet (FREWN). Manual evaluation was also carried out on a sample of the generated synsets. Precision shows that the presented approach has proved to be very promising and applications to use the created wordnet are already intended.

1. Introduction

The first wordnet was developed for English at Princeton University (PWN). Over time it has become one of the most valuable resources in applications for natural language understanding and interpretation, such as word-sense disambiguation, information extraction, machine translation, document classification and text summarisation and, last but not least, Semantic Web applications (Fellbaum 1998). This initiated the development of wordnets for many other languages apart from English (Vossen 1999, Tufis 2000), which was an important milestone because it enabled the developed resources to be exploited in a multilingual setting as well. Currently, wordnets for more than 50 languages are registered with the Global WordNet Association[1].

While it is true that manual construction of each wordnet produces the best results as far as linguistic soundness and accuracy are concerned, such an endeavour is too time-consuming and expensive to be feasible for most languages. This is why semi- or fully automatic approaches have been proposed. By taking advantage of the existing resources they facilitate faster and easier development of a wordnet.

Apart from the knowledge acquisition bottleneck, another major problem in the wordnet community is the availability of the developed wordnets. Currently, only a handful of them are freely available (Arabic, Hebrew, Irish and Princeton). Although a wordnet for French, the French WordNet (FREWN), has been created within the EuroWordNet project (Vossen 1999), the resource has not been widely used mainly due to licensing issues. In addition, there has been no follow-up work to further extend and improve the core FREWN since the project has ended (Jacquin et al. 2007).

This is why the goal of our experiments presented in this paper was to leverage freely available multilingual resources to automatically construct a broad-coverage open-source wordnet for French called WOLF (Wordnet Libre du Francais)[2].

The rest of the paper is organized as follows: a brief overview of the related work is given in the next section. Section 3 describes the methodology for our experiment. Sections 4 and 5 present and evaluate the results obtained in the experiment and the final section gives conclusions and work to be done in the future.

2. Related work

Automatic techniques for wordnet development can be divided in two approaches: the *merge approach* and the *extend approach* (Vossen 1999). Contrary to the merge approach, according to which an independent wordnet for a certain language is first created based on monolingual resources and then mapped to other wordnets, we have opted for the latter. This model takes a fixed set of synsets from Princeton WordNet (PWN) and translates them into the target language, preserving the structure of the original wordnet. It must be noted here that the extend model presupposes that concepts and semantic relations between them are language independent, at least to a large extent.

Apart from faster and cheaper construction of the lexical resource, the biggest advantage of this approach is that the resulting wordnet is automatically aligned to all other wordnets built on the same principle (e.g. wordnets for Swedish and Russian) and therefore available for use in multi-lingual applications, such as machine translation and cross-language information retrieval.

The cost of the expand model is that the target wordnets are biased by PWN and may, in an extreme case, become completely arbitrary (see Orav & Vider 2004 and Wong 2004).

For example, synset ENG20-09740423-n of PWN contains literals *performer* and *performing artist*. However, there is no word or phrase in French that denotes the concept describing actors, singers and other entertainers collectively. Such cases have been dealt with by providing the closest possible match for the synset and aligning the two wordnets with a near_synonym relation. In this way, the overall structure of straightforward cases remained intact and the exceptions appropriately encoded.

[1] http://www.globalwordnet.org [15.03.2008]
[2] http://wolf.gforge.inria.fr [15.03.2008]

Despite these difficulties, the approach is still attractive due to its much greater simplicity which outweighs the language difference issues This is why the expand model has been adopted in a number of projects, such as the BalkaNet (Tufis 2000) and MultiWordNet (Pianta 2002). It was also used in EWN, including for the construction of FREWN, in which a set of English synsets was automatically translated with a proprietary multilingual semantic database and later manually validated.

Research teams developing wordnets in this setting took advantage of the resources at their disposal, including machine-readable bilingual and monolingual dictionaries, taxonomies, ontologies and others (see Farreres et al. 1998). For the construction of WOLF we have leveraged three different publicly available types of resources: the JRC-Acquis parallel corpus[3], Wikipedia (and other Wikipedia-related resources)[4] and the EUROVOC thesaurus[5].

Equivalents for words that only have one sense in PWN and therefore do not require sense disambiguation were extracted from Wikipedia and the thesaurus in a way, similar to Declerck et al. (2006) and Casado et al. (2005). Roughly 82% of literals found in PWN are monosemous, which means that the bilingual approach suffices for an accurate translation. However, most of these are rather specific and do not belong to the core vocabulary[6].

The parallel corpus was used to obtain semantically relevant information from translations so as to be able to handle polysemous literals as well. The idea that semantic insights can be derived from the translation relation has already been explored by Resnik & Yarowsky (1997), Ide et al. (2002) and Diab (2004). Word-aligned parallel corpora have been used to find synonyms by van der Plas and Tiedemann (2006) and Dyvik (2002). The approach has also yielded promising results in an earlier experiment to obtain synsets for Slovene wordnet (Fišer 2007).

3. Approach

3.1 Alignment approach

In this approach we used used the SEE-ERA.NET corpus (project ICT 10503 RP), a 1.5-million-word subcorpus of JRC-Acquis (Steinberger et al. 2006) in eight languages. Apart from French, we used English, Romanian, Czech and Bulgarian. We used different tools to POS-tag and lemmatize the corpus before word-aligning it with Uplug (Tiedemann 2003). Because word-alignment was done only on single words, the approach was not able to generate any translation equivalents for multi-word expressions.

The output of the word alignment process is a file with word links between word occurrences, associated with the two related word occurrence ids and information on word link certainty.

This allowed us to build bilingual lexicons that include all translation variants of words as well as frequency, POS and word-ids information for each entry. The bilingual lexicons range from 43,024 entries for the Cz-En lexicon to 50,289 for the Cz-Bg one. These bilingual lexicons are then combined into five multilingual lexicons. They contain between 49,356 (Fr-Ro-Cz-Bg-En) to 59,019 entries (Fr-Cz-Bg-En). A few entries from the Fr-Cz-Bg-En lexicon are shown in Table 1. Obviously, not all these entries are correct; errors may appear for several reasons, such as tagging, lemmatization, or alignment problems. However, most of these errors are eliminated by the next stage of the process.

frq	**pos**	**Fr**	**Cs**	**Bg**	**En**
18	n	droit	právo	законодателство	law
56	n	droit	právo	право	law
4	n	loi	právo	закон	law
4	n	loi	právo	законодателство	law
6	n	loi	právo	право	law
33	n	loi	zákon	закон	law
8	n	loi	zákon	закона*	law
19	n	législation	právo	законодателство	law
7	n	législation	právo	право	law
4	n	législation	předpis	законодателство	law

Table 1: Translation variants of the English literal *law* from the Fr-Ro-Cs-Bg-En lexicon[7].

At the next stage the goal was to assign a synset id to each lexicon entry. To achieve this, we gathered the set of all possible synset ids assigned to each lexicon entry in all languages (apart from the French one, of course) by comparing it with the corresponding BalkaNet wordnet (Tufis 2000). This is possible because all BalkaNet wordnets use the same synset ids as PWN 2.0. We could then compute the intersection of ids for all languages. The result contains all synset ids that are shared among all non-French lexicon entries. We then assigned these synset ids to their French equivalent. Let us illustrate this by taking the French word *droit*, which is polysemous in French (possible English translation equivalents are: *right*, *law*, *droit*, *royalty*, *entitlement*, *claim*). As shown by Table 1, 56 of its occurrences were aligned with *právo* in Czech, *право* in Bulgarian and *law* in English. The intersection of all sets of synset ids containing the word in wordnets for each individual language contains only the synset id ENG20-05791721-n. It is therefore assigned to those occurrences of the French word *droit* (see Table 2). It is one of the correct synsets for this word (defined in PWN as *the branch of philosophy concerned with the law and the principles that lead courts to make the decisions they do*).

[3] http://langtech.jrc.it/JRC-Acquis.html [15.03.2008]

[4] http://www.wikipedia.org [15.03.2008]

[5] http://europa.eu/eurovoc [15.03.2008]

[6] When we refer to the core vocabulary in this paper, we have in mind all literals corresponding to concepts that are included in the BalkaNet Basic Concept Sets (Tufis 2000). There are three categories of basic synsets, BCS1 being the most fundamental one.

[7] 4-uples occurring 3 times or less are not shown. The literal marked by an asterisk comes from lemmatization errors.

Multiple languages disambiguate polysemous lexicon entries and eliminate most alignment errors. It is rather unlikely that the same polysemy occurs in many different languages or that alignment errors lead to a non-empty intersection. Therefore, the intersection of all possible senses in each language is likely to output only the correct synset.

Fr: ***droit***	**Cs:** ***právo***	**Bg:** ***право***	**En:** ***law***
droit	ENG20-06129345-n	ENG20-04893549-n	ENG20-00577416-n
	ENG20-05559593-n	ENG20-04888072-n	ENG20-05529208-n
	ENG20-05791721-n	ENG20-07928837-n	ENG20-05531141-n
	ENG20-04617988-n	ENG20-00577416-n	**ENG20-05791721-n**
	ENG20-07928837-n	**ENG20-05791721-n**	ENG20-06129345-n
		ENG20-01000872-n	ENG20-07712371-n
		ENG20-04881053-n	ENG20-07928837-n
		ENG20-04617988-n	

Table 2: Word sense disambiguation and sense assignment for French lexicon entries

Applied to the above-mentioned multilingual lexicons, this technique yielded five different sets of synsets with at least one French literal. They include between 1,338 (Fr-Ro-Cs-Bg-En) and 5,073 (Fr-Ro-En) synsets. Because the preprocessing stages, such as tagging, lemmatization and word-alignment were not perfect, it is expected that the synsets created in this way will inherit some of the errors, of course. However, the approach covers polysemous literals from the core vocabulary, which the translation approach, described in the next section, cannot handle.

3.2 Translation approach

We used the following freely available bilingual resources to translate monosemous literals from the PWN 2.0 into French:

- Wikipedia[8] is an on-line multilingual collaborative encyclopaedia. We used it to build a bilingual Fr-En lexicon (314,713 entries) by following to inter-wiki links that relate two articles on the same topic in French and English. We improved and extended this lexicon with a quick analysis of article bodies (capitalization, synonyms extraction, preliminary extraction of definitions).
- The French Wiktionary and its English counterpart[9] are lexical companions to Wikipedia that contain definitions of words as well as some additional information, including their translations into other languages. We used them to create a bilingual lexicon with 24,464 (from the English Wiktionary) and 24,873 entries (from the French Wiktionary).
- Wikispecies[10] is a taxonomy of living species which include both Latin standard names and (for common species) vernacular terms. This allowed us to identify 129,509 language-independent Latin terms as well as French equivalents for 2,648 of these Latin terms.
- Eurovoc[11] is a multilingual thesaurus that is used for classification of EU documents. Version 4.2 of the thesaurus is a structured list of 6,802 descriptors and their equivalents in 21 languages, including many multi-word expressions.

All the bilingual lexicons we extracted from these resources were used to translate monosemous PWN literals. We obtained sets of synsets of different sizes: 18,273 from Wikipedia, 6,848 from Wikispecies, 6,215 and 4,363 from the French and English Wiktionary, and 1,319 from Eurovoc. Translations of the monosemous literals are very accurate and include many multi-word expressions, which was a serious limitation of the alignment approach. Also, they mostly contain specific, non-core vocabulary.

3.3 Merging the results

In the end, synsets obtained from both approaches were merged. If the same synset was created from more than one resource (e.g. from a multilingual lexicon that was extracted from the word-aligned corpus and from a bilingual lexicon that was extracted from Wikipedia), all their unique literals were retained along with the information on the source of the generated synset. This enabled us to perform a simple heuristic filtering according to the reliability of each source, on the diversity of sources that assign a given literal to a given synset and on frequency information (for sources from the alignment approach).

Automatic induction of synsets inevitably leads to gaps in the hierarchy. Because we are aware of the importance of the conceptual density and hierarchy preservation principles for applications (Tufis 2000), we inherited the structure and relations of the missing synsets from PWN 2.0. Empty synsets will need to be addressed in the future. But for the time being, in case an application runs into an empty synset, it can still use the relation information to access a more general or more specific concept. Other language-independent information (e.g. POS, domain, semantic relations) was inherited from PWN.

4. Results

WOLF currently contains 32,351 non-empty synsets that include 38,001 unique literals (see Table 3). This is substantially more than the number of synsets present in FREWN (22,857 in the original resource, but 22,121 once FREWN synsets are mapped to PWN 2.0 synsets). This is directly related to the high number of monosemous PWN literals in non-core synsets (119,528 out of 145,627), that the translation approach was able to handle well.

WOLF contains all four parts of speech that are normally coded in wordnets, while there are only nouns and verbs in FREWN. The most common literals in WOLF are nouns (34,827 vs. 14,618 in FREWN). They are followed by adjectives (1,521 vs. 0 in FRWEN), verbs (979 vs. 3,777 FREWN), and adverbs (664 vs. 0 in FREWN).

[8] http://www.wikipedia.org [15.03.2008]

[9] http://www.wiktionary.org [15.03.2008]

[10] http://species.wikimedia.org [15.03.2008]

[11] http://europa.eu/eurovoc [15.03.2008]

	PWN 2.0	WOLF	WOLF/PWN	FREWN	FREWN/PWN
All synsets	115,424	32,351	**28.0%**	22,121	**19.2%**
BCS1	1,218	870	71.4%	1,211	99.4%
BCS2	3,471	1,668	48.0%	3,022	87.1%
BCS3	3,827	1,801	47.1%	2,304	60.2%
non-BCS	106,908	28,012	26.2%	15,584	14.6%
nominal	79,689	25,559	35.8%	17,381	21.8%
verbal	13,508	1,544	11.5%	4,740	35.1%
adjectival	18,563	1,562	8.4%	0	0.0%
adverbial	3,664	676	18.4%	0	0.0%

Table 3: Quantitative data about WOLF in comparison to PWN and FRWN.

Average polysemy in WOLF is 1.21 synsets per literal (10.5% of literals are polysemous, including 1.2% of multiword literals). In PWN 2.0, average polysemy stands at 1.74 synsets per literal, and 1.39 in FREWN. Coverage of the core vocabulary in WOLF was checked on Base Concept Sets and then compared to FREWN. As Table 3 shows, the core vocabulary in FREWN is denser that in WOLF but the latter has a reasonable coverage of BCS senses as well (71.4% of BCS1, 51.0% of all BCS). It also shows, unsurprisingly, that the more basic the synset, the more likely it is to have been built with the alignment approach.

5. Evaluation

The quality of the resource we created was evaluated automatically as well as manually. In automatic evaluation we compared the resulting wordnet to FREWN and computed f-measure. For a better insight into the problems of our techniques we took a closer look at a representative sample of literals that were not assigned a 100% precision in automatic evaluation. The errors we identified in manual evaluation were classified into several categories.

5.1 Automatic evaluation

FREWN was used as a gold standard to compute precision and recall of sense assignment in WOLF. The most straightforward approach for evaluation of the quality of the obtained wordnet would be to compare the generated synsets with the corresponding synsets from FREWN. But in this way we would be penalizing the automatically induced wordnet for missing literals, which are not part of the vocabulary of the corpus or the bilingual resources that were used to generate the synsets. Instead we opted for a somewhat different approach by comparing literals in the gold standard and in the automatically induced wordnet with regard to which synsets they appear in. This information was used to calculate precision, and recall. Precision gives the number of synset ids assigned to a literal by both wordnets according to the number of synset ids assigned by WOLF. Recall gives the number of synset ids assigned to a literal by both wordnets according to the number of synset ids assigned by FREWN. Results are shown in Table 4.

It must be noted here, however, that literals translated with Wikipedia have a 93,0% precision compared to FREWN. Since the majority of non-BCS synsets are populated from Wikipedia, most synsets that go beyond the coverage of FREWN are of very high quality. Moreover, if a literal appears in a particular synset in WOLF whereas it does not in FREWN, this does not necessarily mean that there is an error in WOLF but it is also possible that FREWN may be incomplete. We therefore selected a sample of 100 literals that were not assigned a 100% precision in automatic evaluation and looked at them by hand as described below.

POS	WOLF/align		WOLF/transl		WOLF/total	
	Prec	Rec	Prec	Rec	Prec	Rec
n	77.2%	68.7%	82.6%	74.9%	**80.4%**	**74.5%**
v	65.8%	54.7%	54.8%	35.8%	**63.2%**	**52.5%**
n+v	74.6%	65.4%	78.8%	69.6%	**77.1%**	**70.3%**

Table 4: Precision and recall of WOLF compared to FREWN for nominal and verbal synsets[12].

5.2 Manual evaluation

A set of randomly selected 100 literals for which WOLF and FREWN show discrepancies was checked by hand. They correspond to 183 literal-synset pairs. We checked manually whether the generated literal-synset pairs are correct or not. We classified errors into several categories, according to the relationship between the literal and the synset it is associated with:

- it is semantically close to the synset (hypernym, hyponym, near-synonym; e.g. *absence* in the synset {*lack, deficiency, want*}),
- it is semantically related (any other kind of semantic relation; e.g. *abri* in the synset {*penthouse*}),
- it is morphologically related (it is part of a compound which would have been correctly assigned to the synset if word alignment was not restricted to single words, or it is a morphologically different form of an otherwise correct literal; e.g. *affaire* in the synset {*things*}, whereas the plural form *affaires* would be correct; *aisance* in the synset {*toilet, lavatory, lav, can, john, privy, bathroom*} whereas the compound *cabinet d'aisances* would

[12] FREWN does not contain any adjectives or adverbs which could therefore not be evaluated automatically.

have been correct),

- it is not related at all (because of alignment and/or disambiguation error; e.g. *abattre* in the synset {*excavate, dig up, turn up*}).

POS	**n**	**v**	**adj**	**adv**	**all**
in FREWN	76 68%	33 46%	0 0%	0 0%	109 60%
not in FREWN					
correct	16	18	*4*	*0*	38
sem. close	10	6	*0*	*0*	17
sem. related	2	6	*0*	*0*	7
morph. related	2	0	*0*	*0*	2
not related	5	5	*0*	*0*	10
total	**111**	**68**			**183**
total correct (WOLF prec.)	**92 83%**	**51 75%**	***4***	***0***	**147 80%**

Table 5. Manual evaluation of WOLF[13].

The results for different POS are shown in Table 5. Approximately 50% of discrepancies are literals that are missing in FREWN synses rather than errors in WOLF. Unsurprisingly, the least problematic synsets are those lexicalizing specific concepts (such as *hippopotamus*, *kitchen*) and the most difficult ones were those containing highly polysemous words describing vague concepts (e.g. *face* which as a noun has 13 different senses in PWN or *place* which as a noun has 16 senses). For a more detailed evaluation, including the resource-by-resource evaluation and resource confidence ranking, see Fišer and Sagot (submitted).

6. Conclusions and future work

The paper has presented a methodology to combine several freely available resources in order to generate a wordnet for a new language. The evaluation of the results shows that the proposed approach is promising from quantitative as well as qualitative aspects. However, precision of the automatically generated synsets drops as ambiguity of words increases, thus affecting the core vocabulary in the developed resource the most. This means that a systematic manual revision of the automatically generated synsets is necessary in order increase the overall quality of WOLF and turn it into a useful resource for NLP applications. Synsets from Base Concept Sets are already being edited by our students.

In addition to this, we intend to extend automatic techniques in order to improve the coverage of WOLF. In particular, we plan to use word sense disambiguation techniques such as those described in Ruiz (2005) to assign synset ids to polysemous Wikipedia entries.

We also plan to extend the scope of WOLF's use and evaluation. In particular, we want to use it for parsing disambiguation and information retrieval purposes. Not only will this validate the usefulness of the resource, it will also enable a more application-oriented evaluation of its relevance and the necessary refinement.

7. References

Casado, R. M., E. Alfonseca, and P. Castells (2005): Automatic Extraction of Semantic Relationships for WordNet by Means of Pattern Learning from Wikipedia. In: *Natural Language Processing and Information Systems: 10th International Conference on Applications of Natural Language to Information Systems*, NLDB 2005, Alicante, Spain, June 15-17, 2005.

Christine Jacquin, Emmanuel Desmontils, Laura Monceaux (2007): French EuroWordNet Lexical Database Improvements. In: *Proceedings of CICLing 2007*, pp. 12—22.

Declerck, Thierry, Asunción Gómez Pérez, Ovidiu Vela, Zeno Gantner, David Manzano-Macho (2006): Multilingual Lexical Semantic Resources for Ontology Translation. In: *Proceedings of the 5th International Conference on Language Resources and Evaluation*. Genoa, Italy, 24-26 May 2006.

Diab, Mona (2004): The Feasibility of Bootstrapping an Arabic WordNet leveraging Parallel Corpora and an English WordNet. In: *Proceedings of the Arabic Language Technologies and Resources*, NEMLAR, Cairo 2004.

Dyvik, Helge (2002). *Translations as semantic mirrors: from parallel corpus to wordnet*. Revised version of paper presented at the ICAME 2002 Conference in Gothenburg.

Farreres, Xavier, G. Rigau, H. Rodrguez (1998): Using WordNet for Building WordNets. In: *Proceedings of COLING-ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, Montreal, Canada.

Fellbaum, Christiane (1998): *WordNet: An Electronic Lexical Database*. MIT Press.

Fišer, Darja (2007). Leveraging parallel corpora and existing wordnets for automatic construction of the Slovene wordnet. In: *Proceedings of the 3rd Language and Technology Conference*, LTC07, Poznan, Poland, October 3-5 2007.

Fišer, Darja, Benoît Sagot (submitted): *Combining multiple resources to build reliable wordnets*.

Ide, Nancy, Tomaž Erjavec, Dan Tufis (2002): Sense Discrimination with Parallel Corpora. In: *Proceedings of ACL'02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, Philadelphia, pp. 54--60.

Orav, Heili and Kadri Vider (2004): Concerning the Difference Between a Conception and its Application in the Case of the Estonian WordNet. In: *Proceedings of the Second Global WordNet Conference*, pp. 285--290, Brno, Czech Republic, January 20-23, 2004.

[13] Figures in italics have to be considered with caution, given the small amount of corresponding data.

Pianta, Emanuele, L. Bentivogli, C. Girardi: MultiWordNet (2002): developing an aligned multilingual database. In: *Proceedings of the First International Conference on Global WordNet*, Mysore, India, January 21-25, 2002.

Resnik, Philip, David Yarowsky (1997): A perspective on word sense disambiguation methods and their evaluation. In: *ACL-SIGLEX Workshop Tagging Text with Lexical Semantics: Why, What, and How?* April 4-5, 1997, Washington, D.C., pp 79--86.

Steinberger Ralf, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, Dániel Varga (2006): The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In: *Proceedings of the 5th International Conference on Language Resources and Evaluation*. Genoa, Italy, 24-26 May 2006.

Tiedemann, Jörg (2003): *Recycling Translations - Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing*, Doctoral Thesis. Studia Linguistica Upsaliensia 1.

Tufis, Dan (2000): BalkaNet - Design and Development of a Multilingual Balkan WordNet. In: *Romanian Journal of Information Science and Technology Special Issue* (Volume 7, No. 1-2).

van der Plas, Lonneke, Jörg Tiedemann (2006): Finding Synonyms Using Automatic Word Alignment and Measures of Distributional Similarity. In: *Proceedings of ACL/COLING 2006*.

Vossen, Piek (ed.) (1998): *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*. Kluwer, Dordrecht.

Wong, Shun Ha Sylvia (2004): Fighting Arbitrariness in WordNet-like Lexical Databases - A Natural Language Motivated Remedy. In: *Proceedings of the Second Global WordNet Conference*, pp. 234--241, Brno, Czech Republic, January 20-23, 2004.

Towards Estonian Ontology

Neeme Kahusk, Kadri Kerner, Heili Orav

Institute of Estonian and General Linguistics, University of Tartu,
J. Liivi 2, 50409 Estonia
{neeme.kahusk,kadri.kerner,heili.orav}@ut.ee

Abstract

This paper describes the project of creating ontologies in OWL for Estonian. The initiative for the Project comes from State Information System Department of Estonian Government. The purpose of this project is to streamline semantic interoperability of different information systems with the help of X-Road — the modernization program of national databases with the aim to change national databases into a common public, service-rendering resource, which would enable agencies, legal and private persons to search data from national databases over the Internet. We are going to make use of Estonian WordNet and EuroVoc as semantic resources. There will be two different ontologies, later the two resources are going to be merged together. Estonian WordNet is built since 1998 and is still in progress. Nowadays there are more than 18,000 synsets, nearly 31,500 lexical entries and more than 20,000 semantic relations in the lexical-semantic database. It is built entirely as a general thesaurus. Also there is available a database of legislative terms built according to wordnet principles that we hope to integrate into Estonian WordNet as well.

1. Introduction

Software applications are creating the need for a complete set of precise concepts. Web searching is limited, because users must specify their queries in terms of keywords. Automated natural language understanding is limited by the ambiguity of language.

In order to enable continued progress in e-commerce, public e-services and software integration, we must give computers a common language with a richness that more closely approaches that of human language. Unfortunately, there is now a trade-off between precision and expressiveness. Computer-readable languages permit computers to represent only very specific and limited things. Human languages can state anything one would ever want to say. However many of the terms and structures of human languages are ambiguous, so that these languages are not very useful for specifying meanings to a computer. They can be so ambigious that people from one organisation do not understand people from another organisation.

We need semantic interoperability both for organisations and technical systems. As for people, they have a “built-in device” for semantic interoperability — dialogue in natural language. They can discuss terms and meanings of words and — hopefully — find common ground. There is not such a mechanism for software systems. Scene where two office-bots are discussing whether database field “Surname” is the same as field “FamilyName” from antother database or not, belongs to science fiction, not reality yet. Common ground would improve efficiency of information systems integration.

This is the part where ontology comes handy. The term “ontology” itself is a good example of ambiguity. Search “define:ontology” in Google gives us 28 answers. Still, mostly they fall into one of the two broader senses: (1) ontology as ’study of being (philosophy’ or (2) ontology as ’a set of concepts within a domain and the relationships between those concepts’[1]. The second meaning is relevant in computer science and artifical intelligence.

There is introduced the concept of Semantic Interoperability Assets in Michard and Rizk (2005), which gathers under one term all the resources involved in ensuring semantic interoperability. Terminologies, thesauri ontologies and mapping rules are all considered as Semantic Interoperability Assets (SIA). They define the agreed meaning of terms and the relationship between terms, control the vocabularies used within data or XML elements, and ensure the element contents are interpreted in the same way by communicating parties.

What is the difference between ontologies and thesauri, then? According to Michard and Rizk (2005) the differences are not major, they differ only in span of domain and set of relations between concepts.

Still, we would treat thesauri as more lexical-oriented assets, containing lexical information that ontology might not have. In our view ontology in computer science sense is a way of representing knowlede about world, not language, although many thesauri contain world-knowledge as well. When we are talking about interopability of software systems, there is more use of an ontology than a thesaurus.

1.1. Motivation

The initiative for the Project came from State Information System Department of Estonian Government. The purpose of this project is to streamline semantic interoperability of different information systems with the help of X-Road.

X-Road is the modernization program of national databases with the aim to change national databases into a common public, service-rendering resource, which would enable agencies, legal and private persons to search data from national databases over the Internet.

The X-Road was launched in 2001. At the beginning, it was developed as an environment that would facilitate making queries to different databases. By now, a number of standard tools have been developed for the creation of eServices capable of simultaneously using the data of different databases. These services enable to read and write data, develop business logic based on data etc.

The X-Road must enable to do any common data process-

[1] http://en.wikipedia.org/wiki/Ontology_(computer_science)

ing operation. Proceeding from this principle, several extensions have been developed for the X-Road: writing operations to databases, transmission of huge data sets between information systems, successive search operations of data in different data sheets, possibility to provide services via web portals, and more (Kalja, 2003).

Nowadays X-Road consists of several databases, security and adapter servers. Still, there is a lack of interopability, as there is no semantic-aware service. Adding semantic layer to the system would help improve cross-usage of current services and make new ones more easily.

2. Thesauri

There are two thesauri available for Estonian. First and most well-known is wordnet-type thesaurus of Estonian (Estonian Wordnet, EstWN) (Vider and Orav, 2005).

The WordNet[2] (WN) created by G.A. Miller and others at Princeton University in the 1980s already existed, and we followed suit. Work on compiling the Estonian wordnet started in 1997 and is still in progress. By now, there are all together approximately 18000 synsets in EstWN. The Estonian wordnet currently includes 11634 noun synsets, 3881 verb synsets, 1580 adjective synsets, 550 adverbs synsets and 440 proper names. Parallel works with thesaurus are increasing in size, adding new semantic relations and specification of concrete domains (for example vocabulary of character traits, transportation etc). Every synset has to have different Language-Internal relation and one InterLingual Index (ILI) relation in English. EstWN has an online version called TEKSaurus as well.[3]

There is a subset of EstWN created as another project. This is a thesaurus of juridical terms which follows same principles as EstWN. There are more than 4200 concepts all enriched with lots of semantical relations.

Second, there is a multilingual thesaurus called Eurovoc. It exists in 21 official languages of the European Union, including Estonian. Eurovoc provides the means of indexing the documents in the documentation systems of the European institutions and of their users. Like any thesaurus, Eurovoc continually has to be adapted to take account, on the one hand, of developments in the fields in which the Community institutions are active and, on the other, of changes in the language.

2.1. Estonian WordNet

Estonian WordNet started about the some time as EuroWordNet[4] (Vossen, 1998). The Estonian team joined the project supported by European Union in 1998 together with Czech, French and German languages. The main idea and basic design of all wordnets in the project come from Princeton WordNet. Each wordnet is structured along the same lines: synonyms (sharing the same meaning) are grouped into synonym sets (synsets). Synsets are connected to each other by semantic relations, like hyperonymy (is-a) and meronymy (is-part-of). There are 43 semantic relations used in Estonian WordnNet version kb53b, most of them are reciprocated (e.g. if 'koer' (dog) `has_hyperonym` 'loom' (animal) then 'loom' (animal) `has_hyponym` 'koer' (dog)).

There is an Inter-Lingual-Index (ILI). Each wordnet is connected to ILI by special ili-relations (called eq-relations). Princeton WordNet ver. 1.5 serves ILI records. ILI concepts themselves do not have intra-language relations, this allows handling lexicalization and knowledge (ontology) separately: see (Vossen, 2004) for futher details.

2.2. EuroVoc

The Eurovoc thesaurus covers all fields which are of importance for the activities of the European institutions: politics, international relations, European Communities, law, economics, trade, finance, social questions, education and communications, science, business and competition, employment and working conditions, transport, environment, agriculture, forestry and fisheries, agri-foodstuffs, production, technology and research, energy, industry, geography, international organizations. Some fields are more highly developed than others because they are more closely involved with the Community's centres of interest. Thus, for example, the names of the regions of each Community Member State are in Eurovoc but not those of non-Community countries.

The Eurovoc thesaurus is published in the official languages of the European Community. Eurovoc 4.2 exists in 21 (of 23 total) official languages of the European Union (Spanish, Czech, Danish, German, Greek, English, French, Italian, Latvian, Lithuanian, Hungarian, Dutch, Polish, Portuguese, Slovak, Slovenian, Finnish, Swedish, Bulgarian, Romanian, and Estonian), and Croatian.

All these languages have equal status: each descriptor in one language necessarily matches a descriptor in each of the other languages. However, there is no equivalence between the non-descriptors in the various languages, as the richness of the vocabulary in each language varies from field to field. The Eurovoc thesaurus has been compiled in accordance with the standards of the International Standards Organization: ISO 2788-1986 — Guidelines for the establishment and development of monolingual thesauri; ISO 5964-1985 — Guidelines for the establishment and development of multilingual thesauri.

At generic level Eurovoc has a two-tier hierarchical classification; fields, identified by two-digit numbers and titles in words, e.g.: 10 EUROPEAN COMMUNITIES microthesauri, identified by four-digit numbers — the first two digits being those for the field containing the microthesaurus — and by titles in words, e.g.: 1011 COMMUNITY LAW

At the specific level of descriptors and non-descriptors, the structure of Eurovoc depends on semantic relationships. They are: scope note, microthesaurus relationship, equivalence relationship, hierarchical relationship, associative relationship.

Some descriptors are accompanied by notes, introduced by the abbreviation SN (Scope note), containing: either a definition, if this clarifies the meaning of the descriptor; or guidance on how to use the descriptor when indexing documents and formulating queries.

All descriptors are accompanied by a reference to a mi-

[2] http://wordnet.princeton.edu/

[3] http://www.cl.ut.ee/ressursid/teksaurus

[4] http://www.illc.uva.nl/EuroWordNet/

crothesaurus, introduced by the abbreviation MT to show to which microthesaurus or microthesauri they belong.

The equivalence relationship between descriptors and non-descriptors is shown by the abbreviations: ”UF” (Used For), between the descriptor and the non-descriptor(s) it represents; ”USE” between a non-descriptor and the descriptor which takes its place. The equivalence relationship in fact covers relationships of several types: genuine synonymity, or identical meanings; near-synonymity, or similar meanings; antonymy, or opposite meanings; inclusion, when a descriptor embraces one or more specific concepts which are given the status of non-descriptors; because they are not often used.

The hierarchical relationship between descriptors is shown by the abbreviations: ”BT” (Broader Term) between a specific descriptor and a more generic descriptor, together with a number showing the number of hierarchical steps between the specific descriptor and each broader term. ”NT” (Narrower Term) between a generic descriptor and a more specific descriptor, together with a number showing the number of hierarchical steps between the generic term and each narrower term.

The associative relationship between descriptors is shown by the abbreviation RT (Related Term) between two associated descriptors. The associative relationship can be of various kinds: cause and effect, agency or instrument, hierarchy, sequence in time or space, constituent elements, characteristic feature, object of an action, process or discipline, location, similarity, antonymy. Attention should also be drawn to the essential features of the associative relationship: it is symmetrical; it is incompatible with the hierarchical relationship: if two descriptors are linked by a hierarchical relationship there cannot be an associative relationship between them, and inversely; descriptors under the same top term cannot be linked by an associative relationship. (Eurovoc, 2005)

3. The Suggested Upper Merged Ontology

The Suggested Upper Merged Ontology (SUMO) and its domain ontologies form the largest formal public ontology in existence today. They are being used for research and applications in search, linguistics and reasoning. SUMO is the only formal ontology that has been mapped to WordNet lexicon. SUMO is written in the SUO-KIF language. SUMO is free and owned by the IEEE. The ontologies that extend SUMO are available under GNU General Public License.

The Suggested Upper Merged Ontology (SUMO) is one of the largest freely available formal ontologies in the world. SUMO is said to be stable, because the structure has not changed remarkably during recent years. SUMO is also language independent; terms have been successfully translated also to other languages. Although SUMO terms were created in English, the labels are not linguistically dependent. (Niles and Pease, 2001)

So it is possible to translate SUMO terms also to Estonian language; to adapt the SUMO top-level ontology. An upper ontology is limited to concepts that are meta, abstract and philosophical. SUMO is also combined and associated with domain-ontologies and with 20,000 terms and 70,000 axioms. Domain specific ontologies have been created that extend SUMO in the fields of finance and investment, country almanac information, terrain modeling, distributed computing, endangered languages description, biological viruses, engineering devices, weather and a number of military applications including terrorist events, army battlefield planning and air force mission planning. (Pease and Fellbaum, 2004)

Having a formal ontology like SUMO, which consists of an upper level concepts, can really be helpful for the creation of a domain specific ontology. It allows the modeller to focus on the content of the domain specific ontology without having to worry on the exact higher structure that gives his ontology a rigid backbone. A formal ontology (SUMO for example) can act a great crossmapping hub if a complete distinction between the content and structure of the external information sources and the formal ontology itself is maintained. This is possible by specifying a mapping relation between concepts from a chaotic external information source and a concept in the formal ontology that corresponds with the meaning of the former concept.

For natural language processing applications it is meaningful to map the human language to a formal ontology. (Niles and Pease, 2003) mapped the synsets from WordNet to terms in SUMO. This task was done manually over a year; all noun, verb, and adjective synsets were linked.

SUMO has also a translation into OWL.

4. From Thesauri to Ontology

Mapping from thesaurus to ontology is carried out via ILI links. As different languages may have different lecicalisations of concepts, there are several kinds of equal-relations. The most common one is eq-synonym, that denotes exact match. There should not be problems in mapping synsets with that equal relation. Table 1 gives overview of most ILI links.

Number	Eq_Relation
8424	eq_synonym
2159	eq_near_synonym
889	eq_has_hyperonym
415	eq_has_hyponym
239	eq_involved
128	eq_be_in_state
122	eq_is_caused_by
121	eq_role
188	other

Table 1: Number of ILI links in Estonian WordNet

Estonian WordNet is not used very much for any application or natural language processing task. There are a couple of games that make use of wordnet’s semantic relations and definitions: an on-line *Scrabble* clone and an ‘Alias’ simulation (‘Alias’ is a word explanation game worked out by Finnish company “Tactic”).

More serious attempt has been done with word sense disambiguation (Vider and Kaljurand, 2002), and it has influenced futher developement of Estonian WordNet (Kahusk and Vider, 2002). We are looking eagerly forward to test EstWN in some real-life application.

One promising topic is the ability to make queries about concepts in natural language. Kaljurand (2007) in his Ph.D theses has developed a system that maps OWL into Attempto Controlled English.

5. References

Eurovoc. 2005. Eurovoc thesaurus. Internet Resource. Available at http://europa.eu/eurovoc/.

N. Kahusk and K. Vider. 2002. Estonian wordnet benefits from word sense disambiguation. In *Proceedings of the 1st International Global WordNet Conference*, pages 26–31, Mysore, India. Central Institute of Indian Languages.

A. Kalja. 2003. System integration process of government information systems. Technical report. Working paper. Available at: http://www.ria.ee/27309.

K. Kaljurand. 2007. *Attempto Controlled English as a Semantic Web Language*. Ph.d. diss., Faculty of Mathematics and Computer Science, University of Tartu, Tartu, Estonia. http://hdl.handle.net/10062/4876.

A. Michard and A. Rizk. 2005. Idabc content interoperability strategy. Technical report, IDABC European eGovernment Services, Paris. Working paper. Available at: http://europa.eu.int/idabc.

I. Niles and A. Pease. 2001. Towards a standard upper ontology. In Chris Welty and Barry Smith, editors, *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, Ogunquit, Maine, October 17–19.

I. Niles and A. Pease. 2003. Linking lexicons and ontologies: Mapping wordnet to the suggested upper merged ontology. In *Proceedings of the 2003 International Conference on Information and Knowledge Engineering (IKE '03)*, Las Vegas, Nevada, June 23–26.

A. Pease and C. Fellbaum. 2004. Language to logic translation with phrasebank. In Petr Sojka, Karel Pala, Pavel Smrz, Christiane Fellbaum, and Piek Vossen, editors, *Proceedings of the Second International WordNet Conference (GWC 2004)*, pages 187–192, Brno. Masaryk University.

K. Vider and K. Kaljurand. 2002. Automatic wsd: Does it make sense of estonian? In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 159–162.

K. Vider and H. Orav. 2005. Estonian wordnet and lexicography. In H. Gottlieb, J. E. Mogensen, and A. Zettersten, editors, *Symposium on Lexicography XI. Proceedings of the Eleventh International Symposium on Lexicography. May 2-4, 2002 at the University of Copenhagen*, volume 115 of *Lexicographica, Series Maior*, pages 549–555, Tbingen. Max Niemeyer Verlag.

P. Vossen, editor. 1998. *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*. Kluwer, Dordrecht.

P. Vossen. 2004. Eurowordnet: a multilingual database of autonomous and language-specific wordnets connected via an inter-lingual-index. *Semi-special issue on multilingual databases. International Journal of Linguistics*.

Lexicon and Ontology Interplay in *Senso Comune*

Alessandro Oltramari, Guido Vetere

ISTC-CNR, IBM Italia, Center for Advanced Studies
via alla Cascata 56/c Trento (Italy), via Sciangai 53 Rome (Italy)
oltramari@loa-cnr.it, gvetere@it.ibm.com

Abstract

Following a fashionable recent trend in the scientific community, computational lexicons are often said to incorporate or even correspond to linguistic ontologies, whose purpose is to describe semantic constructs of language (bound to grammatical units). Nevertheless there's a big debate on whether the categorial structures of computational lexicons could be acknowledged as ontologies or not. We think that the most effective approach is to keep those layers separated, as the philosophy underlying *Senso Comune* suggests. *Senso Comune* is a collaborative platform to build and maintain an open (hybrid) knowledge base of Italian language. As linguistic knowledge base here we mean a machine-readable dictionary that provides semantic information in a formal way. The knowledge base will be initially populated with a suitable formalization of basic Italian lexicon (2K lemmas, about 10K senses) (De Mauro, 1965), then it will be integrated with other existing linguistic resources, as well as user supplied information. The project is backed by an association of Italian scientists, under the supervision of Prof. Tullio De Mauro, which includes as emeritus member Padre Roberto Busa, and is being supported by Fondazione IBM Italia.

1. An introduction to Senso Comune

Senso Comune is a collaborative platform to build and maintain an open knowledge base of Italian language. As linguistic knowledge base here we mean a machine-readable dictionary that provides semantic information in a formal way. The knowledge base will be initially populated with a suitable formalization of basic Italian lexicon (2K lemmas, about 10K senses) (De Mauro, 1999), then it will be integrated with other existing linguistic resources, as well as user supplied information. The project is backed by an association of Italian scientists chaired by Prof. Tullio De Mauro, and is being supported by Fondazione IBM Italia.

The idea at the basis of Senso Comune is that natural languages consist in their concrete use. In the line of Saussure's linguistics (de Saussure, 1949), natural languages are seen as social products, based on users' consensus. At the same time, language users pursue specific goals, with respect to entities that belong to their world (be them physical or not), within social contexts where expressions are creatively produced and understood. This is the reason why physical and cultural realities can be regarded to as the dimension in which speakers' consensus takes shape. Ontologies, as conceptualizations of such realities, and languages, though clearly distinct, are therefore significantly related.

The interplay of linguistic expressions with that kind of abstractions of physic and social situations which we call 'concepts', is subject of a lasting philosophical debate that we won't introduce here. Nevertheless, Senso Comune aims at collecting lexicographic information and put it into relation with corresponding conceptualizations, which raises the non trivial question: how to model such a relationship?

2. The General Model

Ontologies represent an essential link between Knowledge Representation and Computational Lexical Semantics. The most relevant areas of interest in this context are represented by Semantic Web and Human-Language Technologies (HLT): they converge in the task of providing a semantic description of content, although concerning two different dimensions: the conceptual and lexical one. Implemented ontologies and computational lexicons aim at digging out the basic elements of a given semantic space (domain-dependent or general), characterizing the different relations holding among them. Nevertheless, they differ with respect to some relevant aspects:

- the polymorphic nature of lexical knowledge can't be straight off related to ontological categories;
- the widespread phenomenon of polysemy bears upon the lexicon but doesn't affect ontologies at all;
- the architectural features of computational lexicons are far from being easily coded in a logic-based language;
- considering foundational ontologies, a major distinction appears with respect to computational lexicons, the former focusing on high-level concepts (endurant, amount of matter, quality, perdurant) while the latter affect basic-level categories (dog, gold, red, walk).

Following a recent trend in the scientific community, computational lexicons are often said to incorporate or even correspond to linguistic ontologies, whose purpose is to describe semantic constructs of language (they are bound to grammatical units). Nevertheless, there's a big debate on whether categorial structures of computational lexica could be acknowledged as ontologies or not. We think that the most effective approach is to keep the two layers separated. Separating linguistic senses and relationships (e.g. synonymy, hyponymy, and antinomy) from their ontological counterparts (concept, inclusion, and disjointness) is therefore at the basis of our model. This separation prevents linguistic facts to be directly mapped to logic propositions, thus relieves linguistic meanings the burden of embodying ontological commitments. Still, of course, we want the

two layers to be somehow interlinked: in fact, interfacing implemented ontologies and computational lexicons is the key-goal for the new generation of knowledge systems. The model we describe here provides an account of this linkage. By separating linguistic information from conceptualization, we allow language users to manifest their knowledge in a free, incremental, natural, and collaborative way. Of course, this kind of knowledge elicitation is potentially conflicting. As Wikipedia demonstrates, collaborative projects produce huge amount of knowledge, which is continuously updated, amended and extended by wiki-editors. We think that this dynamic approach can be also adapted to Semantic Web frameworks, exploiting human common-sense and linguistic knowledge.
In the rest of this paper we will present in details the features of the ontological and the lexical model underlying Senso Comune, together with the survey of a tutoring methodology for interactive cooperative building of knowledge resources.

2.1. The Metamodel

The *metamodel* at the basis of Senso Comune is a description logics called DL-Lite (Calvanese et al., 2004). With respect to the typical applications of lexical ontologies, we analyzed that DL-Lite provides an appropriate computability and tractability trade-off. UML [1] (Class Diagrams, in particular) has been adopted ad concrete diagrammatic syntax to develop the model, based on a known correspondence with DL-Lite constructs (Table 1).
Basically, DL-Lite is a tractable description logics to specify ontologies and to query large knowledge bases with the same efficiency as relational DBMS. To obtain such efficiency, DL-Lite limits the use of constructs such as universal quantification, disjunction, and enumeration. In fact, the use of these constructs in data-intensive systems would lead to bad computational properties, as Calvanese et al. (Calvanese et al., 2007) have shown.
As any description logics, DL-Lite provides means to define *concepts* (i.e. classes) and *roles* (i.e. binary relations), inclusion dependencies, existential quantification on roles, and negation. Furthermore, syntactic restrictions are adopted to limit the language expressiveness. These are based on distinguishing:

AtomicConcept : atomic concepts (*A*)

BasicConcept : basic concepts (*B*)

GeneralConcept : general concepts (*C*)

AtomicRole : atomic roles (*P*)

BasicRole : basic roles (*Q*)

GeneralRole : general roles (*R*)

ValueDomain : attribute domain (*D*)

These elements are interlinked by the following rules:

- Concepts:
$$\begin{array}{c} B \leftarrow A \mid \exists R \\ C \leftarrow B \mid \neg B \end{array}$$
- Roles:
$$\begin{array}{c} Q \leftarrow P \mid P^{-} \\ R \leftarrow R \mid \neg R \end{array}$$

where the construct P^{-} is used to represent inverse roles (e.g. $love^{-}$ = loved-by). Moreover, roles can be marked as *functional*, that is, of range cardinality equals to 1.
DL-Lite allows *inclusion axioms* of the form:

$$B \sqsubseteq C \quad Q \sqsubseteq R$$

In practice, it is possible to set inclusion dependencies involving base concepts (roles) on the left-side, and general concepts (roles) on the right side. This limitation is crucial to improve tractability of ontology-based data access.
Membership axioms are specified as usual:

$$A(a) \quad D(a) \quad P(a,b)$$

Finally, DL-Lite formal semantics is given by a standard first-order interpretation structure like other description logics (Baader et al., 2003).

2.2. The Ontology

Linguistic resources like WordNet are generally built by lexicographers on the basis of analysis of language. The main taxonomic structure of these resources consists in a hierarchy of hyponyms derived from a comprehensive enquiry of the lexicon. In general, this approach does not deal with ontology-based distinctions, namely with the categorial structure of concepts (synsets). The ontological rearrangement of these resources is possibly made *a posteriori*, as in the case of OntoWordNet (Gangemi et al., 2003).
Senso Comune starts from a different perspective. A small number of concepts is taken *a priori* as a reference ontological structure that constrains the other semantic constructs to be defined in the resource. This reference ontology has been designed according to DOLCE basic distinctions (Gangemi et al., 2002)[2]. In the following list we provide some informal descriptions of the main basic categories:

Entity (∈ Atomic Concept) : the most general category.

Concrete (⊑ Entity) : spatio-temporal entities (i.e., objects, events).

Abstract (⊑ Entity) : non spatio-temporal entities (i.e., propositions, numbers).

Object (⊑ Concrete) : spatial concrete entities with autonomous existence. Objects don't have temporal parts but their properties can change in time (i.e. a ship, a rock, a person).

Event (⊑ Atomic Concept) : temporal concrete entities. Events depend on suitable participants (those objects which take part to a particular event) and can have temporal parts (i.e. a race)[3]

[1] http://www.uml.org/

[2] The latest release of DOLCE consists in thirty concepts and twenty binary relations.

[3] Atomic events don't have, stricly speaking, temporal parts, since they are considered as instantaneous.

Table 1: UML and DL-Lite

UML	DL-Lite
Class	A
Association, Attribute ($\neq PrimitiveType$)	P,P^-
Attribute (PrimitiveType)	D
InstanceSpecification	$A(a)$
LiteralString	$D(d)$
Slot (definingFeature.type $\neq PrimitiveType$)	$P(a,b)$
Slot (definingFeature.type $= PrimitiveType$)	$D(a,b)$
Generalization	$B \sqsubseteq C$
cardinality = 1	$funtc(P)$

Quality (⊑ Entitiy) : qualifying characteristics of entities; the existence of qualities is bound to the existence of the correspondent entities (i.e. the colour of a particular rose), although they are not parts of them.

In *Senso Comune*, the association between linguistic senses and the reference ontology is based on a genuinely naive assumption, namely that objects are commonly lexicalised by nouns, qualities by adjectives and kinds of events by verbs[4]. Nevertheless, the relation holding between the previous list of ontological categories and suitable parts of speech (nouns, verbs, adjectives and adverbs) is not as simple as it could appear: those correspondences are not stable across languages and case exceptions are frequent in linguistic practice.

2.3. The Lexicon

Lexical information in managed in Senso Comune by means of a suitable extension of the base ontology, which consists in a set of abstract concepts to represent linguistic notions. During the analysis phase, the need of representing and integrating classic lexicographic structures along with user-collected data emerged. This lead us to a representational model which is more complex than other state-of the art ones, e.g. the Lexical Markup Framework (Francopoulo et al., 2006). In any case, our model shares with LMF most of the basic structures, making it easy to map them if needed.

Besides representing morphological structures, Senso Comune lexical model provides classes and relations to represent meanings and semantic relationships.

2.3.1. Meanings

The class diagram in 1 shows how word meanings are modeled.

Meaning (⊑ Abstract) : reified relation that represents the fundamental semantic structure (*sign*), independently from any description (**MeaningDescription**). The meaning relation brings together a word form (or multi-word) to the concept in an ontology and (possibly) the contexts (which, in turn, are concepts) where the meaning occurs.

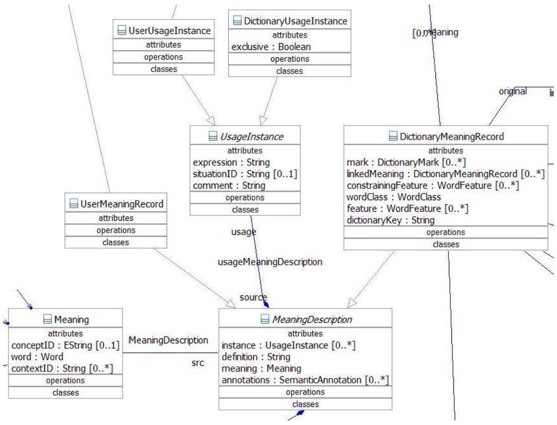


Figure 1: Linguistic Model: meanings

MeaningDescription (⊑ Abstract) : descriptive structure associated to **Meaning**, including a phrase (glossa), a set of usage instances, and a set of semantic annotations.

UserMeaningRecord (⊑ MeaningDescription) : **MeaningDescription** provided by users.

DictionaryMeaningRecord (⊑ MeaningDescription) : **MeaningDescription** coming from to dictionary lexicographic structures.

UsageInstance (⊑ Abstract) : usage instances which are part of **MeaningDescription**.

UserUsageInstance (⊑ UsageInstance) : usage instances provided by users.

DictionaryUsageInstance (⊑ UsageInstance) : usage instances coming from dictionary lexicographic structures.

Note that **Meaning** represents a linguistic acceptation in form of association between linguistic expressions and conceptual content. The latter consists in a `URI` pointer to a single concept, so that it is possible to define a function:

$$\sigma : Meaning \rightarrow Concept$$

In particular, σ is neither injective (different meanings could point to the same concept), nor surjective (not all concepts must be mapped with lexical counterparts). We just require each meaning to be mapped to a unique concept.

[4] We avoid here to consider the ontological counterparts of adverbs, which however could be preliminarily conceived as "modes" of events, like in the example "John was running *fastly*".

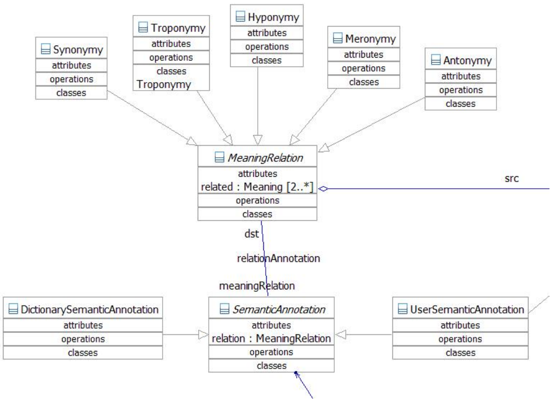


Figure 2: Linguistic Model: meaning relations

2.3.2. Lexical relations

The diagram in 2 shows binary relations involving meanings. In particular, relationships taken into account include: synonymy, troponymy, hyponymy, antonymy, and meronymy. Corresponding classes are:

MeaningRelation (⊑ **Abstract**) : reified relation that associates meanings pairwise.

Synonymy (⊑ **MeaningRelation**) : represents synonymy, i.e. meaning equivalence in all contexts. Differences in connotation (e.g. *child* vs. *kid*) may determine fuzziness in users' perception.

Troponymy (⊑ **MeaningRelation**) : represents troponymy, i.e. different ways for an action to take place (e.g. *walk* vs. *crawl*). It is in question whether troponymy can always maps to conceptual inclusion.

Hyponymy (⊑ **MeaningRelation**) : represents specialization (e.g. *dog* vs. *canine*). As for synonymy, hyponymy is subject to fuzzy perception by users.

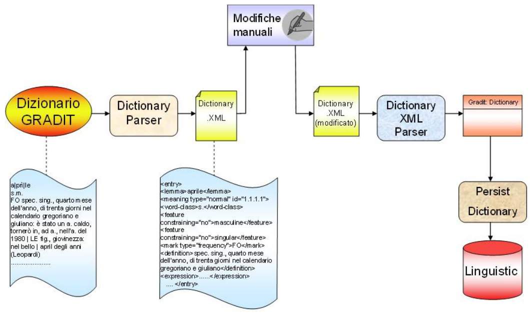
Antonymy (⊑ **MeaningRelation**) : represents contrariety, typically for adjectives (e.g. *bad* vs. *good*). Whether antonymy implies conceptual disjointness should be evaluated case by case.

Meronymy (⊑ **MeaningRelation**) : represents part-whole relationships. Conceptually, this relation may be in correspondence to a number of different parthood notions.

In sum, semantic relationships elicited by users cannot be directly mapped in logic relationships within the framework of formal theories of linguistic meanings as lexical ontologies are. Instead, these theories must be constructed by carefully analyzing linguistic perceptions declared by users or condensed by dictionaries.

3. The Development Process

In the initial stage of the project, *Senso Comune* knowledge base will be populated with approximately 10000 senses associated to 2075 lemmas of De Mauro's core dictionary (De Mauro, 1965); for each of these senses a DOLCE-based conceptual counterpart (see 2.2.) will be provided. Suitable conversions of ontological linguistic resources for Italian, such as EuroWordNet (Vossen, 1998), will enable an integration with *Senso Comune*[5]. Starting from a core set of fundamental senses, *Senso Comune* knowledge base is going to be developed by supervised contribute of speakers through a cooperative open platform.



a|prì|le
s.m.
FO spec. sing., quarto mese dell'anno, di trenta giorni nel calendario gregoriano e giuliano: è stato un a. caldo, tornerò in, ad a., nell'a. del 1980 | LE fig., giovinezza: nel bello | april degli anni (Leopardi)
....

```
<entry>
<lemma>aprile</lemma>
<meaning type="normal" id="1.1.1.1">
<word-class>s.</word-class>
<feature
constraining="no">masculine</feature>
<feature
constraining="no">singular</feature>
<mark type="frequency">FO</mark>
<definition>spec. sing., quarto mese
dell'anno, di trenta giorni nel calendario
gregoriano e giuliano</definition>
<expression>......</expression>
    .... </entry>
```

Figure 3: Acquiring the basic lexicon

3.1. Acquiring the Basic Lexicon

The acquisition of the basic lexicon is under completion. Starting from plain textual lemmas extracted from the dictionary, the main goal is to build the correspondent instances of the **LexicalEntry** class. The overall strategy of conversion depends on the exploitation of an intermediate format: an XML file is created with suitable identifiers for the lexical contents of the dictionary. The population of the knowledge base is then obtained through a compiling process.

Manual annotation of lemmas has been discarded on the basis of a feasibility study, estimating approximately 3-years working period for the complete annotation of De Mauro's core 2075 entries. Moreover, the analysis of the main structures of the dictionary revealed that textual formats in De Mauro's resource can be hardly tractable with a fully automatic methodology of extraction: this study prevented from developing an *ad-hoc* parser.

In this context, the employment of a semi-automatic approach emerged as the most adequate solution: first, a suitable parser is used to produce an approximation of an XML desired format, which is then adapted and amended by linguists, who are also responsible for solving uncertainties and deciding for the best candidate entry (see 3). In particular, the distinction between use cases and "nuances" of meaning cannot be regularly extracted from the syntactic structures of the textual formats of the dictionary.

3.2. The Cooperative Platform

After the acquisition of the basic terminology, *Senso Comune* computational lexicon will be extended through a cooperative platform mirroring the main characteristcs of the so-called 'wiki'. Wiki is a web-based software that allows visitors to edit the content of a given website. This open platform is particularly appropriate and easy-to-use for cooperative tasks related to texts and hypertexts. Currently,

[5] Currently, we are evaluating how part of proprietary resources like EuroWordNet could be made available as Open Source through *Senso Comune* model, interface and format.

a large number of wiki systems is available on the web; although wikis are usually task-oriented and designed according to specific user requirements, they share some common essential features:

- Editing through browser: contents are usually inserted through web-browsers with no need of specific software plug-ins.
- Rollback mechanism: versioning of saved changes is available, so that an incremental history of the same web page is mantained.
- Non restrictive access: in most cases, wikis are free access resources and visitors have the same 'privileges'[6] in the editing process.
- Collaborative editing: many wiki systems provide support for editing through discussion forums, change indexes, and so on and so forth.
- Emphasis on *linking*: wiki pages are usually strongly connected with other hypertexts.
- Search functions: in practice, every wiki system allows for search over internal contents.
- Upload of non-textual contents: many wikis allow visitors to upload multimedia data (images, audio files, videos).

There are mainly three critical aspects in wiki-systems:

1. Difficulty of keeping neutral perspective on information[7]. It's diffult to represent the neutral view on wiki contents, since total agreement on topics is almost impossibile to be reached. In general, the moderators of a wiki are responsible for monitoring contents and sensibilize visitors.
2. Quality of contents. This aspect share a similar scenario with the previous issue but focuses on 'bad' or low-level contents.
3. Exposure to 'malevolent attacks': Attacks aim at damaging contents or to introduce offensive (or out of scope) information.

On the basis of wiki philosophy and architecture, Wiktionary project has been initiated, aiming at building an open multilingual dictionary with meanings, etimologies, pronunciations. Although Wiktionary could be seen as the closest initiative to *Senso Comune*, the strong limitations of the resource[8] lead Senso Comune association to develop a brand new original system.

The current prototype version of *Senso Comune* computational lexicon is grounded on a relational database resulting from the linguistic model (see 2.3.). The database has been also integrated with a suitable DL-Lite reasoner, designed and implemented to operate on large ontologies. After visualising the information linked to a searched meaning, a user will be able to decide whether to insert a new lemma, a new sense, a new lexical relation or simply to leave a 'feedback' (i.e., her familiarity with available senses and lexical relations). On the contrary, the deep conceptual part of the lexicon (the ontology) won't be made accessible to users: when a new sense of a lemma is added, the system semi-automatically creates a corresponding specific concept to be positioned with respect to the ontological layer of the database. This semi-automatical procedure will be initially driven by an interactive Q/A system, by means of what we have called a **T**utoring **M**ethodology for the **E**nrichment of **O**ntologies (**TMEO**).

4. The TMEO Methodology...in a nutshell

Senso Comune depends on two core aspects: 1) a top-down direction, where top-level ontological categories and relations are introduced and maintained by ontologists to constrain lexicalised concepts; 2) a bottom-up direction, where non-expert users are asked to enrich the semantic resource with linguistic information through a wiki-like platform. In this building-up process, visitors are allowed only to access to the lexical level of the resource (therefore, explicit ontological choices are kept 'opaque' to ease users' task). These access-restrictions produce an *epistemological spread* between dimensions 1) and 2), a necessary requirement if we want to keep the deep technical aspects of the ontological layer aside from wiki-users. Conversely, to make dimension 2) plainly effective, those lexical concepts and relations which are introduced by users must fit the intended ontological choices underlying the system. For this reason, we are designing a tutoring methodology to support linguistic enrichment of ontologies, towards the creation of comprehensive hybrid semantic resources. **TMEO** is an interactive Q/A system based on general distinctions embedded in DOLCE. We present here some preliminary characteristics of the methodology[9].
First, a given lemma and the corresponding gloss is visualised by *Senso Comune* wiki-user interface: for instance, the word 'glass' defined as "a container for holding liquids while drinking"(sense 2 of WordNet). Afterwards, the system *asks* natural language questions to the user, aiming at

[6] using the Internet jargon

[7] One could object that 'neutrality' of information is an utopia. Although it's not our aim to deepen this issue in the present paper, we could say that, under a context-driven treshold, as far as multiple views and heterogeneity of information are mantained, the richness of a wiki can be assessed. On the other hand, it's trivial that when wiki contents reflect partial and distorted views on facts and knowledge, the danger for the overall Internet community is extremely high.

[8] The most important one concerns the model of the dictionary, which depends on the couple lemma-page, where different senses of a lemma are coded as free text without specific identifiers in the same page. This feature almost completely hides the conceptual information associated to lemmas.

[9] For the sake of simplicity and to make the example readable for non-Italian speakers, we exploits available mappings between *Senso Comune* lexical nodes for Italian and Princeton WordNet synsets, providing examples in English.

specializing the intended meaning of the submitted lemma. In the following we report some examples[10]:

1. Would you consider [glass] in the sense of ["a container for holding liquids while drinking"] as something concrete, namely which has a spatial and/or temporal nature?
2. Does [glass] refer to something tangible, namely that a human can sense?
3. Could you count [glass]-es[11]?
4. Is [glass] produced/built by hand/machines?

The typical answers to those questions would be: yes/no/yes/yes/yes; however, the method will optimize the way questions are posed to the user by navigating ontology inclusions and disjunctions. In particular, terms like *glass* would commonly be interpreted as referring to some concrete tangible object. The second-last question aims at helping the user to discriminate between unitary entities (artifacts like tables and coins, natural entities like trees and animals, etc.) from scattered and unbound entities like substances (liquids like water, materials like gold, etc.). The conclusive result is that in the macro-world of human senses - which is the actual domain of *Senso Comune* - the selected sense of the word *glass* can be modeled by the class 'Artifact', which is a specialization of DOLCE-based top-node 'Concrete'. Although this information might appear trivial, in case of a different sense of the term *glass*, namely "a brittle transparent solid with irregular atomic structure" (Wordnet sense 1), the final output would have been different: since here the lexicalised concept refers to the material and not to the object, the answer to question 3. should have been negative, cognitively 'evoking' the ontological category 'Substance'.

The internal algorithm of **TMEO** automatically selects the most adequate category of the reference ontology as the super-class of the given lexicalised concept: difference sequences of answers induce different mappings between the lexicon and the (hidden) ontological layer. In this context it's important to notice that **TMEO** list of questions does not have a flat organization: a conditional chain based on "if..then' clauses[12] rules the logical structure of the tutoring system. Moreover, the system makes automatic storage of each Q/A interaction, building a sort of dynamic reference manual to be exploited as help documentation by wiki-users[13]. Of course there may be cases where a user does not know how to answer to **TMEO** questions: we will adopt two solutions to overcome the stall. In the short-term, we are creating an open forum where expert modelers will periodically answer *vis-à-vis* to specific questions posited by users; in the long-term, we are going to include uncertainty in **TMEO** algorithm, allowing for a third optional answer ("I don't know") by the user. Although this enhancement is going to make the general heuristics of the tutoring system more complex, it will fasten the interactive process with respect to the forum solution.

5. Conclusions

We have presented Senso Comune, an open, cooperative project to build a knowledge base for Italian language. Basing on a simple and yet powerful metamodel (the DL-Lite description logic), a minimal foundational ontology (DOLCE), a specific representation model for linguistic knowledge, and a core lexical resource (De Mauro's fundamental lexicon), Senso Comune will be built and continuously updated by collecting input from users. One of the major features of our approach is the way linguistic meanings and ontological concepts are put into relation. Meanings are not modeled as concepts, but rather as signs. Accordingly, lexical relationships such as synonymy or hyponymy are not mapped into formal relations such as equivalence or inclusion, but rather are taken as input for the construction of ontological theories.

Future research will include modeling situations by means of frame-like structures, consistently with the formal model developed so far. Lexical relationships to capture thematic roles will be therefore introduced. Another research direction is toward algorithms for automating the introduction ontology axioms (e.g. equivalence, inclusion, disjointness, participation) based on linguistic information, by taking both quantitative and qualitative aspects into account.

Finally, we think that Senso Comune as an open source of knowledge of Italian language can make a long way as key enabling factor for business, Web communities, and public services in Italy. The resource will be distributed under Creative Commons license and made available for any kind of use.

6. Acknowledgements

We would like to thank the President of *Senso Comune* Tullio De Mauro, the board of directors Angelo Failla, Aldo Gangemi, Nicola Guarino, Maurizio Lenzerini and Malvina Nissim for their important feedbacks and comments to this paper, and Ilaria Gorga (IBM Italia), Rita Plantera (Università di Roma I), Fabrizio Smith (Università di Roma III) for making available par of the material we presented here. Alessandro Oltramari's participation to this work was supported by the Italian FIRB project TOCAI.it (Knowledge-based technologies of enterprise integration).

7. References

F. Baader, D. Calvanese, D. L. Mcguinness, D. Nardi, and P. F. Patel-Schneider, editors. 2003. *The Description Logic Handbook : Theory, Implementation and Applications*. Cambridge University Press, January.

D. Calvanese, G. De Giacomo, M. Lenzerini, R. Rosati, and G. Vetere. 2004. Dl-lite: Practical reasoning for rich dls. In *Proc. of the 2004 Description Logic Workshop (DL 2004)*, volume 104 of *CEUR Electronic Workshop Proceedings, http://ceur-ws.org/Vol-104/*.

[10] The form of **TMEO** questions is generally fixed: words in square brackets (lemma + gloss) change every time a new lemma is submitted to the wiki-user by the system.

[11] The plural syntactical form is automatically generated by the system.

[12] *IF answer = Yes THEN (term IS-A ontological categoryA) ELSE (term IS-A ontological categoryB).*

[13] For instance, a user that has to model sense 1 of glass might want to look up how sense 2 has been treated by previous visitors.

D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, and R. Rosati. 2007. Tractable reasoning and efficient query answering in description logics: The dl-lite family. *J. Autom. Reasoning*, 39(3):385–429.

T. De Mauro. 1965. *Introduzione alla semantica*. Laterza, Bari.

T. De Mauro, editor. 1999. *Grande Dizionario dell'Italiano dell'Uso*. UTET, Torino.

F. de Saussure. 1949. *Cours de linguistique generale*. Payot, Paris.

G. Francopoulo, N. Bel, M. George, N. Calzolari, M. Monachini, M. Pet, and C. Soria. 2006. Lexical markup framework (lmf) for nlp multilingual resources. In *Proceedings of the COLING-ACL Workshop on Multilingual Lexical Resources and Interoperability*, pages 1–8.

A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, and L. Schneider. 2002. Sweetening ontologies with dolce. In *Proceedings of EKAW*, pages 21–29.

A. Gangemi, R. Navigli, and P. Velardi. 2003. The ontowordnet project: extension and axiomatization of conceptual relations in wordnet.

P. Vossen, editor. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht.

Computer-Assisted Ontology Creation and Maintenance Using Calais, Wikipedia, and WordNet

Alexander Nakhimovsky, Tom Myers
Computer Science Department, Colgate University
13 Oak Drive Hamilton NY 13346 USA
N-Topus Software
56 Payne Street Hamilton NY 13346 USA
tommyers@dreamscape.com, adnakhimovsky@colgate.edu

Abstract

This paper describes a system under development that consists of the following components:

- Repository: a collection of historical and ethnographic materials, including texts, images, and multimedia.
- Glossary: a master list of disambiguated words and phrases, with definitions from WordNet and/or Wikipedia articles.
- Ontology of people, places, organizations, political roles, and events, with links to their occurrences in the Repository.
- An Exhibit (2008) display with Timeline (2008) and Google maps, derived from the Ontology, with links to the Repository.
- A multimedia player that can be invoked to play specific time-aligned segments of video and text, linked to annotations.

We also describe interactive procedures to create and maintain the system, making use of Reuters' Calais ontology-generator, WordNet and Wikipedia. One of the goals of the system is to provide links to historical and anthropological background for news stories.

1. Introduction

This paper describes the structure and functionality of an emerging repository of materials (text and multimedia) on the history and ethnography of Pashtuns. We visualize history as a sequence of events, most of which are actions by human agents, occurring at specific times and places. We thus wish to create and maintain an ontology of people, places and events, with links back from concepts in the ontology to the contexts of their occurrence. (The contexts include time-aligned and annotated segments of video transcripts that each has a URL and can thus be located and replayed by a multimedia player.)

The main components of the system include:

1. Repository: a collection of primary materials: texts, images, multimedia
2. Multimedia player that can link to a segment of time-aligned annotated video and replay that segment.
3. Glossary: a master list of disambiguated words and phrases, with definitions from WordNet and/or Wikipedia articles, including summaries from the Yahoo Search service.
4. Ontology of people, organizations, political roles, places and events
5. An Exhibit (2008) presentation with Timeline and Google maps, derived from the Ontology, with links to the Repository.

The main use cases we intend to support are as follows:

1. Display events in the ontology as an Exhibit. By using appropriate filters, display biographies and storylines.
2. Produce a report on a specific item (a person, a place, an event), either within the Timeline-with-maps context or as a text page with links.
3. Expose the contents of the repository via an RDF feed for integration with other resources.
4. Extract information from a new item, such as a news story: identify the concepts in our ontology that it makes references to; add new relevant concepts, if any, to the repository.
5. Ingest a new item into the repository. The details depend on the type of the item.

The rest of the paper presents a report on the current stage of development of the system, concentrating primarily on the last two use cases. They follow the same dataflow, usually resulting in additions to the ontology and the Exhibit display. The main difference is that for ingested documents we generate links back from ontological concepts to those contexts in the new document in which they occur.

What constitutes the context depends on the kind of document being processed. If it is an annotated video transcript then the context is a segment within the structure of time-aligned video-transcript segments. (See section 5 for details.) For other HTML and XML documents it is the lowest element in the DOM tree that contains the lexical item and enough words in it to serve as a context. Here "enough" is a settable parameter currently set to 13.

The dataflow for use cases 3 and 4 is shown below. Steps 1, 3, 5 and 6 are automated; steps 2 and 4 are a lot of work.

1. Run the document through Calais.
2. Interactively process the results.
3. Tokenize and create a KWIC (Key Word in Context) index of the text, excluding 1000 most common words but including "complex words" such as Federal Reserve or Abdul Zahir.
4. Interactively process the index.
5. Generate a set of triples in a custom-designed JSON format from the KWIC material.

6. Update the Exhibit display and RDF feed.

2. Processing steps

In this section we briefly characterize the tools and algorithms of the six processing steps.

2.1 Run the text through Calais

This invokes the Calais web service. For testing purposes, we use a short text that we tweak in various ways to see how the service responds. Here is a sample:

Text:

This is our testing text containing references to Milton Friedman, Abdul Zahir, and Afghanistan. Instead of Federal Reserve it talks about the bank of the River Oxus. The date of this testing text is April 1, 1836. The location is obviously Lewis Carroll's *Through the Looking Glass*, where the Red Queen rules. Bobby Fischer, of Reykjavik, may also be involved.

Calais output:

Organization: Federal Reserve System
NaturalFeature: River Oxus
IndustryTerm: bank (sic! adn)
Country: Afghanistan
Person: Lewis Carroll, Milton Friedman, Bobby Fischer, Abdul Zahir
City: Reykjavik

The results of Calais make two kinds of useful suggestions: "complex words" for the KWIC and possible triples to add to our ontology. Both kinds of suggestions are checked by the human editor using Wikipedia and WordNet.

2.2 Process the results

The editor checks Calais output against the accumulated ontology and the current text content. If the subject of a Calais triple already occurs in the ontology, the editor uses the context to verify that it is indeed the same subject. For instance, Wikipedia lookup for Abdul Zahir leads to a disambiguation page with four options, one of which, Afghanistan's prime minister in 1971-72, is in our ontology. The context will most likely help establish whether the current occurrence refers to the prime minister or one of the other three options.

If the subject is not in the ontology, then a new triple is added to the ontology, containing the generic information from Calais. The new subject is also looked up in WordNet and Wikipedia for more precise categorization: "prime minister" rather than "person." For some categories we have templates (HTML forms) to fill in, such as: a person holding a rank has a term in office with a start date and an end date.

Complex words recognized by Calais are recorded to make sure they are added to the KWIC as it is created.

2.3 Tokenize the text and build a KWIC index

In this step, we split the text into words and punctuation marks, and then group the words into phrases, including proper names. We have a variety of capitalization-based heuristics for recognizing new phrases or names, but it has to be subject to human correction. Consider this example:

> When the 1964 Civil Rights Act was debated in the Senate, the audience included Martin Luther King, Jr. and Malcolm X. Lyndon B. Johnson signed the bill three months later.

We want to identify phrases in the text as corresponding to Wikipedia articles:

"1964Civil Rights Act" as wk:Civil_Rights Act of1964
"Senate" as wk:United_States Senate
"Martin Luther King, Jr. "as wk:Martin_Luther_King_Jr.
"Malcolm X" as wk:Malcolm_X
"Lyndon B. Johnson" as wk:Lyndon_B._Johnson

If these identifications have already been entered into our ontology, they will be matched correctly. If not, we will still do pretty well: on the first item, our algorithm will find the disambiguation page which points to the correct answer among others; the last three will be identified correctly, and only "Senate" will point to a specific but wrong article.

Naturally, this approach does poorly with names that are not famous enough for a Wikipedia listing, but these would need the editor's attention in any case, so little is lost.

What we are doing here is information extraction: skim the text "looking for occurrences of a particular class of object or event and for relationships among those objects and events." (Russell and Norvig, 2002:848) Probably the best tool for this task is FASTUS (Appelt et al. 1995), and we are in discussions about the possibility of using it in our project. In the meantime, we use Calais and Wikipedia checking.

Another feature of our KWIC index that deserves mention is that the context is not defined as a fixed number of words on both sides of the key word, but rather as a structural unit defined by the document markup. If the document is ingested into the system then each index entry contains a link back to the unit in which the key word of the entry occurs.

2.4 Process the KWIC

This is a human editor task. (In this application, we do automatically prune all tokens that do not have a verb or a noun sense in WordNet, but this still leaves a big number of tokens to prune.)

The framework tries to be helpful in two ways. First, we make common tasks easy: it takes one click to delete a row in the index. Second, for disambiguation purposes we try to provide selection lists and reasonable defaults so that the editor selects the first choice. We do that by consulting the offline versions of both WordNet and Wikipedia. While the offline use of WordNet is well-familiar, the Wikipedia option we are using (Tsiodras 2007) deserves to be better known.

2.4.1. Offline Wikipedia lookup

Tsiodras's central insight is that the Wikipedia downloads are compressed in the bzip2 format, storing data in compressed chunks of less than a megabyte each which are then concatenated. Accordingly, the huge download file can easily be broken into several thousand files, each containing many articles although there will usually be an article which crosses the boundary from one file to the next. The author then combined standard tools to index these by title, so a query by title expands just the chunk (or at most two chunks) that contain the target page (which is often a disambiguation page). This seems to be much easier than setting up the local database version, and it is really reasonably quick.

As the author explains, this approach does not allow wiki updates, only retrieval. Otherwise, it provides a useful alternative to the database version [18] and to DBPedia [19] which is more limited in coverage, and sometimes gives only summaries in English.

2.4.2. Ordering of disambiguation options

Wikipedia searches often return a disambiguation page, and WordNet in most cases returns a number of synsets. We use a simple similarity metric to compare the context of the entry with disambiguation options, whether WordNet definitions or Wikipedia pages. Specifically, we compute the vector of case-insensitive word-counts for the context; we then compute its dot product with such vectors for each disambiguation option. We divide the dot-product by the length of each option's vector, and get what might be considered a density. The options are presented to the editor ordered by that value.

2.4.3 Action on each entry

If an index entry is not pruned (editorial judgement is used here), the editor takes the following actions:

- Check if the entry is already in the ontology. This may or may not require going through the process of disambiguation, including calls to WordNet and Wikipedia. If yes and the text is ingested, add link from the ontology to the new context.
- If the entry is not in the ontology, add a concept to the ontology and a lexical entry to the dictionary.

2.4.4 Updating the dictionary

As we mentioned, in addition to the ontology, we maintain a global glossary of disambiguated lexical items, each with a part-of-speech tag and definition. The part-of-speech tag also indicates which WordNet sense is intended: *sound_n7* is "a narrow channel of the sea joining two larger bodies of water." The definitions come from WordNet or Wikipedia article summaries as supplied by Yahoo Search. AJAX calls from the editor's processing page run the appropriate web services and display the results for copying and pasting.

We should clarify that we use WordNet sense numbers and Wikipedia article URLs as identifiers for our ontological concepts. We thus often associate a word sense from WordNet with an ontological concept, and we may use WordNet's hyperonym and hyponym information in computing our similarity metric. However, we do not use WordNet itself as an ontology. As Gangemi et al (2003) show, this would require a significant separate effort with unclear payoff, tangential to our goals.

2.5 Generate a set of triples in JSON

Our JSON format (unlike, e.g., the JSON format of the Simile Exhibit) is designed for maximum generality: we want to be able to express any RDF and RDFS content in it, including subclassing and reification. It is patterned after the Turtle format: namespace declarations followed by triples that use the declared prefixes. In the sample below, the disInto predicate indicates that Wikipedia or WordNet disambiguates the Subject as Object:

```
{  prefix_mapping:
 [
   {url: "http://n-topus.com/name/", prefix: "nt"},
   {url: "http://n-topus.com/name/rel/", prefix: "ntr"},
   {url: "http://purl.org/dc/elements/1.1/", prefix: "dc"},
   {url: "http://en.wikipedia.org/wiki/", prefix: "wk"},
    . . .
 ],
 actual_triples: [
   ["nt:Ghazni","rdf:type","nt:city"],
   ["nt:abdication","rdfs:subClassOf","nt:event"],
   ["nt:Abdul_Zahir","ntr:disInto","wk:Abdul_Zahir
_%28Afghan_Prime_Minister%29"],
   ["nt:Abdul_Zahir","rdf:type","nt:primeMinister"],
   ["nt:primeMinister","rdfs:subClassOf","nt:person"],
   ["nt:city","rdfs:subClassOf","nt:place"],
   ["nt:Afghan_King_Shuja","rdf:type","nt:king"],
   . . .
  ]
}
```

Obviously, this text cannot be created manually, nor can it be directly passed to Exhibit for display. As mentioned, we use HTML forms to generate this RDF-looking JSON. We use a script to convert it to JSON that can be submitted to Exhibit.

2.6 Exhibit JSON

Exhibit's JSON looks very much like Marvin Minsky's frames, or as database records. (They can be automatically generated from an Excel spreadsheet.)

```
{
   "items" : [
      {   type :                  "Nobelist",
          label :                "Burton Richter",
          discipline :            "Physics",
          shared :               "yes",
          "last-name" :           "Richter",
          "nobel-year" :          "1976",
          relationship :         "alumni",
          "co-winner" :           "Samuel C.C. Ting",
          "relationship-detail":      "MIT      S.B.1952,Ph.D.
1956",
          imageURL :             "a long URL"
      },
      {   type :                  "Nobelist",
      . . .
      many more items
   ]
}
```

In RDF, this comes out as:

```
<rdf:RDF xmlns:rdf='well-known URI'
   xmlns:exhibit='http://simile.mit.edu/2006/11/exhibit#'
   xmlns:a='http://www.w3.org/2000/01/rdf-schema#'
   xmlns:b='http://www.w3.org/...-rdf-syntax-ns#'
   xmlns:c='http://simile.mit.edu/2006/11/exhibit#'
   xmlns:d='file:///C:/.../exhibit/property#'>
<rdf:Description rdf:about='file:///...'>
   <a:label>Burton Richter</a:label>
   <b:type>Nobelist</b:type>
   <d:discipline>Physics</d:discipline>
   <d:shared>yes</d:shared>
   <d:last-name>Richter</d:last-name>
   <d:nobel-year>1976</d:nobel-year>
   <d:relationship>alumni</d:relationship>
   <d:co-winner>Samuel C.C. Ting</d:co-winner>
   . . .
```

We will establish two paths from our JSON records to Exhibit, one via RDF transformed into Exhibit RDF/XML, the other directly from JSON to JSON.

3. Ingesting a video clip

An important part of our repository is a collection of time-aligned video clips, with transcripts and annotations. This section describes how they are brought into the system. An earlier version and more detail for this part of the paper can be found in Nakhimovsky et al. (2005).

The initial input is usually a large minimally-compressed AVI file that is often the result of digitizing a miniDV tape, although hard-drive camcoders are becoming dominant.

The steps we go through are as follows:

1. Cut up the video into clips of no more than 15 minutes long.
2. Compress each clip into a deliverable format. We prefer Adobe Flash because it is most widely available.
3. Transcribe and time-align the compressed clip using either ELAN (2008) or Transcriber (2006).
4. Convert the time-aligned transcript into XHTML format and add annotations.
5. Ingest the XHTML file into the system as described in the preceding section.

While the video processing steps are of little interest here, the process of time alignment is important because it provides the basis for adding annotations to segments of time-based media, including semantic annotations. By time-alignment we mean the following task: given a unit of time-based media (e.g., a video file) and an associated text (usually the transcript of the media) divide media and text into matching text-media segments, so that, given a text segment, the corresponding video segment can be accessed and replayed. (The time complexity of access will depend on the implementation of the underlying media-player and its Application Programming Interface or API.) The converse task of finding the text segment corresponding to a given media point can be done, with proper indexing, in constant time.

Given time-alignment, annotation becomes trivial: to attach an annotation to a media segment, simply attach it to the matching text segment. Furthermore, since time-alignment creates, in effect, an associative array of media segments indexed by text segments, media search is reduced to text search. Since annotations can contain metadata organized into an ontology, multimedia collections can be searched using the full array of Semantic Web technologies.

3.1 Tools for time alignment

To create and use time alignment, we need two programs: annotator and player. Two most commonly used programs for creating time alignment are ELAN and Transcriber. Both are desktop programs that create XML files in custom-designed formats (EAF and TRS, respectively). ELAN is a much more complex and powerful program that can work with both video and audio; it can also import TRS files created in Transcriber. The latter is a simpler program for transcribing audio files; it creates time- aligned segments to help with transcription.

For playback, we use our own program called MannX. It is a browser-based program that uses an XHTML format with div's, span's and class attributes. We have converters to create MannX XHTML from either TRS or a restricted subset of EAF. After the conversion, annotations can be added to time-aligned transcript segments using an HTML editor. Each annotated segment is a DIV element that has a URL and can be used as a context in creating the KWIC index, as described in the preceding section.

4. Conclusion

In this paper, we presented the structure of an online repository of text and multimedia, and the procedures used to create and maintain it. We believe that the structure is quite general and applicable to a variety of fields of knowledge. We hope that our procedures can be adopted and further developed in other online repositories.

5. Acknowledgements

Work on this project has been partially supported by the NSF grant #0553546, Five Languages of Eurasia, under the Documenting Endangered Languages program; and by the NEH Digital Humanities Fellowship.

6. References

Appelt, Douglas et al. (1995). SRI International FASTUS system. In *Proceedings of the 6th conference on Message understanding*. Columbia, Maryland.

Calais (2008) http://opencalais.com/.

ELAN (2008) http://www.lat-mpi.eu/tools/elan/, v.3.4.0.

Exhibit (2008) http://simile.mit.edu/exhibit/, v.2.0.

Gangemi, Aldo, Nicola Guarino, Claudio Masolo, Alessandro Oltramari (2003). Sweetening WORDNET with DOLCE. *AI Magazine* 24(3) pp. 13 - 24.

Nakhimovsky, A., Chris Helmuth, Tom Myers (2005). Semantic Annotations for Digital Video. In *Proceedings of the 2nd Italian Semantic Web Workshop*. Trento, Italy

Russell, Stuart and Peter Norvig. (2002). *Artificial Intelligence: a Modern Approach*. 2nd Edition.

Prentice Hall New York.

Timeline (2008). http://simile.mit.edu/timeline/

Transcriber (2006) A tool for segmenting and transcribing speech http://trans.sourceforge.net/en/presentation.php.

Tsiodras (2007) http://users.softlab.ece.ntua.gr/~ttsiod/