Wordnets mapped to central ontology – revised
DELIVERABLE 6.6

Version 1
3-Apr-10
DRAFT

Piek Vossen VU University Amsterdam
Roxane Segers, VU University Amsterdam
Amanda Hicks, BBAW
Axel Herold, BBAW
German Rigau, EHU
Eneko Agirre, EHU
Ainara Estarrona, EHU
Montse Cuadros, UPC
Egoitz Laparra, EHU
Kyoko Kanzaki, NICT

KnowledgYielding Ontologies for Transition-based Organization
ICT 211423

Co-funded by EU – FP7 ICT Work Programme 2007 under Challenge 4 - Digital libraries and Content
<table>
<thead>
<tr>
<th><strong>Grant Agreement No.</strong></th>
<th>ICT 211423</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Project Acronym</strong></td>
<td>KYOTO</td>
</tr>
<tr>
<td><strong>Project full title</strong></td>
<td>Knowledge Yielding Ontologies for Transition-based Organization</td>
</tr>
<tr>
<td><strong>Funding Scheme</strong></td>
<td>FP7 – ICT</td>
</tr>
<tr>
<td><strong>Date latest version Annex I</strong></td>
<td>19-12-2007</td>
</tr>
</tbody>
</table>
| **Project Coordinator** | Prof. Dr. Piek T.J.M. Vossen  
VU University Amsterdam  
Tel. + 31 (0) 20 5986466  
Fax. + 31 (0) 20 5986500  
Email: p.vossen@let.vu.nl |
| **Project website**    | [http://www.kyoto-project.eu/](http://www.kyoto-project.eu/) |
| **Deliverable Document Number** | D6.6 |
| **Status**             | Draft     |
| **Security (distribution level)** | Public |
| **Contractual date of delivery** | February 28, 2010 |
| **Actual date of delivery** | April 1, 2010 |
| **Type**               | report    |
| **WP contributing to the deliverable** | WP6 |
| **WP responsible**     | BBAW, Amanda Hicks |
| **Authors**            | Piek Vossen  
Amanda Hicks  
Axel Herold  
German Rigau  
Roxane Segers  
Eneko Agirre  
Ainara Estarrona  
Montse Cuadros  
Egoitz Laparra  
Kyoko Kanzaki |
| **EC project officer** | Werner Janusch |
| **Keywords**           | Wordnet, ontology, mapping relations |
| **Abstract**           | This deliverable describes the work on providing a mapping from the concepts in the generic wordnet to the central ontology in KYOTO |
Table of Content

1 INTRODUCTION .................................................................................................................. 4
2 KYOTO KNOWLEDGE MODEL .......................................................................................... 6
3 MAPPING RELATIONS BETWEEN WORDNETS AND ONTOLOGY .................................. 13
4 CREATING THE GENERIC WORDNET TO ONTOLOGY MAPPINGS .................................... 19
5 RUDIFY ........................................................................................................................................ 21

   Evaluation ............................................................................................................................. 21
   Monolingual evaluation ......................................................................................................... 21
   Crosslingual combination evaluation ................................................................................. 23

   Exploitation and Rudify interviews ..................................................................................... 23
   Method 1 ................................................................................................................................... 24
   Method 2 ................................................................................................................................... 24
   Method 3 ................................................................................................................................... 25

6 EXTENDING THE MAPPING RELATIONS TO MODEL PROCESSES AND STATES FOR THE DOMAIN CONCEPTS ........................................................................................................... 26
7 FUTURE WORK ...................................................................................................................... 32
8 REFERENCES ............................................................................................................................ 33
9 APPENDIX I SET OF VERBAL MAPPINGS .......................................................................... 36
10 APPENDIX II: SET OF ADJECTIVAL MAPPINGS .................................................................. 40
1 Introduction

The goal of the KYOTO project (ICT-211423) is to develop an information and knowledge sharing system that relates text in various languages to a shared ontology in such a way that it enables the extraction of deep semantic relations and facts from text in a domain. The system should establish communication and interpretation across languages and cultures and it should support building and maintaining the system by groups of people in a shared domain and area of interest.

Most domain acquisition systems in the semantic web community model each domain separately and restrict the system to a single language or a limited set of languages. They also require knowledge engineers and language-technology experts to do the modeling. The KYOTO system is on the other hand specifically designed to build global and cross-cultural consensus on the meaning and interpretation of language. As such it is an open system that can be extended and maintained by the users themselves without requiring skills in knowledge engineering or language technology.

The system behaves similarly to Wikipedia - it allows specific social groups to agree on the interpretation and meaning of the concepts that matter for them. Nevertheless, the interpretation and meaning definitions are formalized in such a way that computer programs can use these definitions to mine the text provided by the same groups on the same matters. The process of acquiring this knowledge is further supported by automatic mining of terms and concepts from text documents that are provided by the users. A special editing environment helps the users to select and define the terms that are relevant without needing to know the formal knowledge structures that underlie these definitions. The result is a domain wordnet for their domain.

In order to further formalize the meaning of these terms and to share these concepts across languages and cultures, the editor also prompts these users for more formal constraints and relations. Again, the editor uses suggestions coming from the automatic acquisition and hides the complex knowledge structures. This leads to a domain ontology that is available to other participants, e.g. defining terms in another language, possibly based on acquisition from documents in these other languages. Cross-lingual and cross-cultural validation is established through agreeing and sharing the domain ontology.

The conceptual knowledge of the domain (both the wordnets and the shared ontology) is anchored to generic wordnets and a generic ontology. This helps building up the definitions and relations for the domain resources, i.e. not all knowledge needs to be re-defined from scratch, but this also makes it possible to share the knowledge with people outside the community. The conceptual knowledge built up in this way is thus shared across languages within the social groups and with other social groups. This is only possible if the domain knowledge is anchored to a general language and knowledge repository.

In the project, we will be working on a restricted set of languages: English, Dutch, Italian, Spanish, Basque, Simplified Mandarin Chinese and Japanese. We also will apply the system to the domain of the environment and specifically to the topic of ecosystem services. Nevertheless, the system is designed in such a way that it can be used for any language and can be applied to any domain.
KYOTO uses a model in which knowledge is distributed over 3 different layers: domain vocabularies, generic and domain wordnets and a central ontology. Terms, either from background vocabularies or acquired from text, are matched to wordnets in specific languages, which are matched to a central ontology. The matches across these resources can be partial, as long as it is possible to reach another matching concept through internal parent relations.

It is essential that the generic and most-relevant synsets of the different languages are mapped to the central ontology (described in deliverable D06.5) so that the domain experts can easily create mappings for the specific concepts in the domain. These specific concepts (mostly nouns) are edited through the Wikyoto platform through interviews generated from definitions (Segers and Vossen 2010). The content words in these definitions are used to generate simple yes-no-questions to validate the relations to processes and states in which the objects of the domain can be involved. For example, we find the following definitions for the words *prey* through Google searches:

- A prey is an animal that is **hunted** by a predator.
- A prey is an animal that is being hunted or **eaten** by other **animals**.
- A prey is an animal that is hunted for **food**.

From these definitions, we will generate questions such as the following:

<table>
<thead>
<tr>
<th>Question</th>
<th>Mapping Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is it true for [prey] that it is involved in hunting?</td>
<td>(HuntingProcess)</td>
</tr>
<tr>
<td>(yes/no)</td>
<td></td>
</tr>
<tr>
<td>Is it true for [prey] that it hunts something or someone? (yes/no)</td>
<td>(HuntingRole)</td>
</tr>
<tr>
<td>(yes/no)</td>
<td></td>
</tr>
<tr>
<td>Is it true for [prey] that it is hunted?</td>
<td>(HuntedRole)</td>
</tr>
<tr>
<td>(yes/no)</td>
<td></td>
</tr>
<tr>
<td>Is it true for [prey] that it is involved in eating?</td>
<td>(EatingProcess)</td>
</tr>
<tr>
<td>(yes/no)</td>
<td></td>
</tr>
<tr>
<td>Is it true for [prey] that it eats something/someone? (yes/no)</td>
<td>(UsedAsFoodRole)</td>
</tr>
<tr>
<td>(yes/no)</td>
<td></td>
</tr>
<tr>
<td>Is it true for [prey] that it is eaten?</td>
<td>(ServesAsFoodRole)</td>
</tr>
<tr>
<td>(yes/no)</td>
<td></td>
</tr>
</tbody>
</table>

To be able to generate these questions from content words in the definitions, it is required that processes and states are represented in the ontology and as many verbs and adjectives in languages as possible are mapped to these ontological classes. This deliverable describes the approach and work we have done in providing a first mapping from the English WordNet to the ontology.

The mapping of nominal Base Concepts (Vossen 2008 and Izquierdo, Suárez, and Rigau 2009) was described in deliverable D06.3. In this deliverable we explain how a subset of verbs and adjectives was linked to the ontology. We also describe the kind of mapping relations that have been defined (sections 3 and 4). The kind of mapping is first of all determined by the rigidity of nominal concepts for endurants (objects), as defined by Guarino and Welty (2002). We developed a tool for the automatic detection of rigidity Rudify (Hicks and Herold 2009 and also in deliverable D06.3). Rudify has now been tested for all the languages in KYOTO, focussing on the nominal Base Concepts and some domain specific terminology. This is described in section 5.

Once the rigidity of concepts has been determined, it is possible to generate simple interviews, given the constraints in the ontology and on the mapping relations. In this approach, rigid synsets are automatically mapped to their basic types through wordnet
hyponymy relations and subclass relations in the ontology. Non-rigid synsets and terms for domain specific processes and properties need to be mapped to the relevant processes and qualities in the ontology. The mapping of English verbs and adjectives that will facilitate this is described in section 6. Finally in section 7, we will describe the future work carried out in the third year of the project to enlarge the coverage of verbal and adjectival synsets to represent all the synsets in wordnets and to extend to the mapping from the English WordNet to other languages. This will prepare the KYOTO platform for extensions to other domains than the environment. In the next section, we will first summarize the knowledge model in KYOTO.

2 KYOTO knowledge model

KYOTO uses a 3-layered knowledge model consisting of:

1. Vocabularies in different languages
2. Wordnets in different languages
3. A central ontology shared by all languages

The first layer consists of large volumes of background knowledge and new terms learned from text collections in the domain. This layer is linked to wordnets in different languages. All the wordnets are linked to the English WordNet. The wordnets represent the 2nd layer of knowledge, which is linked to the 3rd layer: the central ontology. Each of these layers has an internal semantic structure, connecting specific concepts to more general concepts and it has specific mapping relations to the next layer.

The layers in KYOTO consist of the following vocabularies:

1 Layer-1 (background and domain vocabularies):
   1.1 Generic background databases:
      1.1.1 Wikipedia: over 3 million articles in English and large volumes in other languages, by September 2009.
      1.1.2 DBPedia: 2.6 million things and 274 million pieces of information (RDF triples), by September 2009.
      1.1.3 GeoNames: 8 million geographical names and 6.5 million unique features whereof 2.2 million populated places and 1.8 million alternate names, by September 2009.
   1.2 Domain specific vocabularies:
      1.3 Species2000 database with 2.1 million species, having taxonomic relations and labels in many different languages.
      1.4 Term databases extracted from the documents provided for the domain, with about 500K concepts extracted from 1,000 documents

2 Layer-2 (wordnets):
   2.1 Generic wordnets in 7 languages with 50K to 120K synsets
   2.2 Domain wordnets in 7 languages built in the KYOTO project

3 Layer-3 (central ontology):
   3.1 DOLCE-Lite ontology ((Masolo et al 2003)
   3.2 KYOTO extension of DOLCE Lite:

1 http://www.wikipedia.org/
2 http://dbpedia.org/About
3 http://www.geonames.org/about.html
4 http://www.sp2000.org/
3.2.1 Top-layer extension
3.2.2 Middle-layer extension
3.2.3 Domain-layer extension

The 3 layered model makes it possible to apply the central and shared semantic models, as expressed in the ontology and the mapping of wordnets to the ontology to massive amounts of concepts represented in the background and domain vocabularies. Among others, the ontology expresses rich axioms for processes and qualities. It They are those processes and qualities that matter for the modeling of concepts in the domain and the knowledge that is expressed in the textual resources of a domain.

In the model we assume that the generic wordnets and the background vocabularies contain massive amounts of concepts and instances of concepts that represent the different types of things that occur in a domain. Each resource has a concept hierarchy that connects these concepts to more general concepts. Many of these concepts are rigid types as defined by Guarino and Welty (2002). Figure-1 gives a schematic representation of this knowledge, where millions of concepts for species are related through broader-than relations in the SKOS database of species and similarly through hyponymy relations in WordNet (Fellbaum 1998). Rigidity in these resources is implicit and can be formalized through the statement that each concept represents a set of essential identity criteria for instances of these concepts. This means that an instance of a *Eleutherodactylus augusti* will always be an instance as long as it exists, and will never become an instance of *Eleutherodactylus atrabracus*. In other words, these concepts are rigid and therefore disjoint. Similarly, GeoNames contains millions of instances of locations that are rigid as well: *Amsterdam* will never become *Berlin*. Likewise, the background vocabularies and the wordnets represent rich repositories of types of things.

However when we look at the documents in the domain of the environment and the terms extracted from these documents, we see that they do not include just these specific rigid concepts but mostly non-rigid concepts such as *endangered frogs, endemic frogs* and *alien frogs*. Typically, these can be applied to any type of frog and represent possibly temporary contexts for frogs which are not essential to the instances. This means that particular instances of different types of frogs can become *endemic* or *alien* and can stop having the properties denoted by these adjectives. This follows from the nature of documents on the environment. Such documents typically report trends and observations of processes and states in which these things are involved. (I took out the sentence about dictionaries and encyclopedias because they, too contain information about non-rigid entities—they have info about ALL words and concepts in a language.) Likewise, the non-rigid concepts and also the processes and states themselves have more information value for mining knowledge expressed in the text.

Consequently, we can divide the knowledge in 3 different types:

1. Abstract conceptual knowledge about the type of things in terms of the essential and rigid concepts of which their instances exist;
2. Abstract conceptual knowledge of the processes and states in which these instances can be involved, as expressed by lexicalizations of these processes and states (e.g. *endangered, alien*) and by the non-rigid terms that refer to these instances involved in these processes (e.g. *endangered species, alien frogs*);
3. The instantiation of knowledge as found in text, where reference is made to instances of the rigid things (entities would sound better) in the world being involved in processes and states;

The latter type of knowledge can be extracted by processing sentences in the text, in which participants of processes and states are collected from different constructions:

1. Sentential structures relate subjects, objects and adjuncts to verbal and adjectival expressions, e.g. “Thousands of geese migrate to the Humber Estuary each year”;
2. Noun phrases express nominalizations of process or states, e.g. “Migration of aquatic birds decreases every year”.
3. Non-rigid terms are found that implicitly relate instances to processes, e.g. “Migratory birds suffer from industrial activities in the region”.

On the basis of the above distinctions in the knowledge repositories, we can reconstruct possible mappings across instances referred to by the above constructions and group the processes and states in which they are involved, i.e. the industrial activities have an effect on migratory birds, since migratory birds (non-rigid) can refer to aquatic birds and to geese. We can thus make the inference that industrial activities may be a cause for the decrease of aquatic birds and that we can expect that there will be less than thousands of geese in the future due to these activities.

As motivated in the deliverables D06.4 and D081/D082, it is not feasible to represent all the concepts for a domain in the central ontology together with all the process and states in which they can be involved. KYOTO therefore implements a division of labour principle (comparable to Putnam 1975), where all rigid domain concepts remain in the background vocabularies and generic knowledge structures (i.e. species2000, GeoNames, and partly concepts in the generic wordnets) for the different languages, whereas the ontology is minimal, focusing on a limited set of rigid disjoint types and a main focus on the processes and states in which these types can be involved. The KYOTO architecture thus allows us to collect the processes and states for instances of things in a domain, where we can handle a wide variety of expressing these processes and states and can make inferences on the wide variety of the things that can occur in the domain. In this way, the knowledge repository has the optimal balance between expressiveness and recall of knowledge.

The large hierarchies of vocabularies in the background databases and the generic wordnets are then used to provide the recall in terms of coverage of the kind of things that can be recognized as the participants in the processes and states. In order to exploit this model, we need to have a mapping across the concepts in the different resources. However, it is not necessary to have a mapping relation between all the concepts across the resources, since we can use the internal relations in each resource to find a more general concept with a mapping. This is schematically represented in Figure-1. Whenever we come across a term such as Eleutherodactylus atrabractus in a domain text, we traverse the relations in the SKOS database of Species2000 until we find a more general concept that is matched to the generic wordnet. Similarly, the terms that we find in the documents are represented in the term database with a parent relation to a more general concept. If this parent term (or the parent of the parent term) is matched with a wordnet concept, all the more specific terms are
matched to wordnet indirectly.

The same principle can be applied to the wordnet concepts themselves. We can traverse the wordnet hypernym relations until we find a concept that is matched to the ontology. Likewise all background concepts, all terms and all wordnet concepts are ultimately related to some classification in the ontology. Using this approach, it will be sufficient to have a minimal set of mappings across the different resources to be able to apply inferences such as the above, where we assume the rigidity principle that all concepts related to more general concepts in the background vocabularies and the term database are rigid-subtypes unless there is evidence to the contrary. Furthermore, we can use the constraint that non-rigid concepts can only have non-rigid children linked to it.
We have followed the following approach to provide a mapping for every concept occurring in the domain. First, we provide a maximal mapping for each resource to a wordnet in a language:

1. Background resources are automatically aligned to WordNet on the basis of vocabulary overlap and parallel structural relations in the hierarchy.
2. GeoNames was manually aligned by mapping the location types to the corresponding synsets in Wordnet.

Secondly, we provide a mapping for the so-called Base Concepts in the wordnet to the ontology. The Base Concepts (BCs) are extracted in such a way that they represent the most important and most general concepts that cover the full hierarchy. We have selected around 2.1 million species, 100,000 synsets, and 2,000 types.

3. The term database is automatically aligned through Word-Sense-Disambiguation of all the content words in the text from which the terms are derived.
300 BCs for the noun-hierarchy. Of these, 205 BC's have been added to the noun hierarchy in such a way to ensure complete coverage of WordNet.\(^5\)

In addition to this, 106 synsets that are not BCs are also in the ontology. These synsets fall into two broad categories: (1) those that directly contribute to the completeness of the abstract typologies found in the upper layers of the ontology and (2) those that are domain specific. The typology of change found in the middle layer of the ontology is an example of the first case. This typology consists largely of synsets taken directly from WN3.0. The direct subclasses of “change” include five synsets; change of integrity, change of location, change of magnitude, change of state, and motion. While this is not an exhaustive typology of change, it connects many of the domain terms with the more abstract terms in the ontology, and generates more structure in the middle layer. Another case of WN3.0 synsets that had been added to the ontology include cases such “volume unit” and “length unit”. While these are not Base Concepts in WN3.0 nor were they mentioned in the domain terms taken from estuary documents, an incomplete typology of units of measure was present at this stage. These synsets were added to fill out this typology with more units of measure that are of common interest. Domain specific synsets come from the estuary documents and mindmaps as well as the workbench extension. This process of adding synsets from the user documents is described in D6.3. A description of the workbench extension is available in D6.5.

These concepts have been represented in the ontology and a mapping has been provided for each of these BCs to the ontology class. Likewise, every concept in the domain vocabularies can thus be related to a wordnet synset and every nominal wordnet synset can be related to a BC and each BC to an ontological class. This work has been described in D06.3_Worldnets_mapped_to_central_ontology_version1.

For background vocabularies such as Species2000, we can safely assume that all concepts represent a disjoint hierarchy of rigid concepts. For GeoNames, we can assume that all concepts are instances of certain concepts in Wordnet referring to types of places and locations. The ontological class can thus be applied to the domain concept as a proper subclassOf. For the frogs in the Species2000 database, we can thus infer that they represent proper subtypes (rigid) of the most specific class in the ontology, which is represented by the BC frog, toad, toad_frog, anuran_batrachian (synset 01639765-n in WordNet3.0). The ontology represents this class as a rigid-subtype of organisms.

For some of the concepts in WordNet and for most of the concepts in the term database, this assumption does not hold. In wordnet, we typically find a mixture of rigid and non-rigid concepts and the term database, we find mostly non-rigid concepts. In order to decide on the rigidity, we use a combination of automatic and manual techniques. The automatic techniques are implemented in the Rudify tool (Hicks and Herold 2009), which uses learned linguistic patterns to decide on the rigidity. This approach works good for words that have sufficient recall in Google searches (about 80% precision for general wordnet nouns). For specific domain terms, a manual protocol is followed to decide on the rigidity through an interview.

For all concepts that are non-rigid, it is not sufficient to just link them to the WordNet

\(^5\)Side note: terms like "cake" and "flavoring" were not included in the ontology since they are not relevant, not ontologically precise, and they are subsumed by "nutrient" which is in the ontology.
hierarchy. Each non-rigid concept needs to be directly matched to the process or state in which it is involved. This is illustrated in Figure-1 for the terms in the term database, where endangered frog is matched to the endurant endangered. Although such terms are linked as hyponyms to the generic wordnet synset for frog, we still need to indicate that they are non-rigid subtypes and that they are involved in processes and states defined in the ontology.

We will define the precise mapping relations between the synsets and the ontology in the next section.
3 Mapping relations between wordnets and ontology

In order to differentiate between rigid and non-rigid concepts in the wordnets and to be able to match the non-rigid concepts to the relevant processes, we defined the following mapping relations:

For rigid synsets referring to endurants (objects):
- synset, sc_equivalenceOf, endurant;
- synset, sc_subclassOf, endurant;
- synset, sc_instanceOf, endurant;

For non-rigid synsets referring to endurants:
- synset, sc_domainOf, endurant
- synset, sc_playRole, participant-relation
- synset, sc_participantOf, perdurant
- synset, sc_hasState, quality, quality_region
- synset, sc_hasCoParticipant, endurant
- synset, sc_playCoRole, participant-relation

The following definitions are provided for these relations:

- **sc_equivalenceOf**: the synset is fully equivalent to the ontology Type & inherits all properties; the synset is Rigid
- **sc_subclassOf**: the synset is a proper subclass of the ontology Type & inherits all properties; the synset is Rigid
- **sc_domainOf**: the synset is not a proper subclass of the ontology Type & is not disjoint (therefore orthogonal) with other synsets that are mapped to the same Type either through sc_subclassOf or sc_domainOf; the synset is non-Rigid but still inherits all properties of the target ontology Type; the synset is also related to a Role with a sc_playRole relation
- **sc_playRole**: the synset denotes instances for which the context of the Role applies for some period of time but this is not essential for the existence of the instances, i.e. if the context ceases to exist then the instances may still exist (Mizoguchi et al. 2007).
- **sc_participantOf**: instances of the concept (denoted by the synset) participate in some endurant, where the specific role relation is indicated by the playRole mapping.
- **sc_hasState**: instances of the concept are in a particular state which is not essential and can be changed. There is no need to represent the role for a stative perdurant.
- **sc_hasCoParticipant**: instances of the synset participate in a process that also involves some other endurant, which is the target of this relation.
- **sc_playCoRole**: the co-participant is involved in another role for the same process, which is secondary from the perspective of the instances of the defined synset.
Using these mapping relations, we can express that the synset for duck (which has a hypernym relation to the synset bird with an equivalence relation to the ontology class bird) is thus a proper subclassOf the ontology class bird:

- wn:duck hypernym wn:bird sc_equivalenceOf ont:bird

Similarly, if the concept for the synset spider is not in the ontology, we can create the mapping: wn:spider sc_subclassOf ont:insect, which directly expresses it is a disjoint subclass.

For a concept such as migratory bird, which is also a hyponyms of bird in wordnet but not a proper subclass as a non-rigid concept, we thus create the following mapping:

- wn:migratory bird sc_domainOf ont:bird
- wn:migratory bird sc_playRole ont:done-by
- wn:migratory bird sc_participantOf ont:migration

This mapping indicates that the synset is used to refer to instances of endurants (not subclasses!), where the domain is restricted to birds. Furthermore, these instances participate in the process of migration in the role of done-by. The properties of the process migration are further defined in the ontology, which indicates that it is a active-change-of-location done-by some endurant, going from a source, via a path to some destination. The mapping relations from the wordnet to the ontology, need to satisfy the constraints of the ontology, i.e. only roles can be expressed that are compatible with the role-schema of the process in which they participate. Such constraints can be used to guide the creation of the mappings in the editing system Wikyoto (see deliverables D074.b and D075b).

Occasionally, terms occur that express multiple role relations in a single term, e.g. gas-powered vehicle, which refers to endurants in the role of a vehicle but also to the endurant gas in the role of a resource in the same process. The referent of the synset always plays a single role, whereas another endurant plays another role, which is secondary from the point of the view of the synset. In these cases, we use a special relation sc_hasCoParticipant to indicate the secondary role for another endurant, e.g.:

- wn:gas-powered vehicle sc_playRole ont:instrument
- wn:gas-powered vehicle sc_participantOf ont:active-change-of-location
- wn:gas-powered vehicle sc_hasCoParticipant ont:gas
- wn:gas-powered vehicle sc_playCoRole ont:resource

Another example is the term oyster sanctuaries. This synset is a hyponym of sanctuary; refuge which is defined as a non-rigid synset in the following way:

- wn:sanctuary;refuge sc_domainOf ont:geographical_object

We have not encountered any terms in which more than two roles are directly expressed as a part of the concept. If this turns out to be the case, we need to group sc_playCoRole and sc_hasCoParticipant relations, so that we know which participant plays what role.
Relative to this concept, the hyponym oyster sanctuary is defined as follows:

- wn:oyster sanctuaries hyponym-of wn:sanctuary;refuge
- wn:oyster sanctuaries sc_hasCoParticipant ont:bivalvia
- wn:oyster sanctuaries sc_playCoRole ont:patient

In this case, the co-role relation is defined relative to the relation expressed by the hypernym sanctuary;refuge.

For non-essential states, we can use the sc_hasState relation to express that a synset such as wild dog refers to instances of dogs that life in the wild but can stop being wild:

- wn:wild dog sc_domainOf ont:dog
- wn:wild dog sc_hasState ont:wild

Note that we can also relate rigid subtypes to essential properties. For example, if wild cherry denotes a subclass of cherries we can express this through:

- wn:wild cherry sc_subclassOf ont:cherry
- wn:wild cherry sc_hasState ont:wild

However, this is only appropriate if the state of being wild is essential.

Ideally, all processes and states that can be applied to endurants should be defined in the ontology. If that is the case, it is sufficient to have sc_equivalentOf and perhaps sc_subclassOf mappings from synsets to ontological classes. This may hold for most verbs and adjectives in languages, which do not tend to extend in specific domains and are part of the general vocabulary (e.g. to pollute, to reduce, wild). However, domain specific text contain many new nominal terms that refer to domain-specific processes and states, e.g. air pollution, nitrogen pollution, nitrogen reduction. These terms are equally relevant as their counter-parts that refer to endurants involved in similar processes, e.g. polluted air, polluting nitrogen or reduced nitrogen. We therefore use the reverse participant and role mappings to be able to define such terms for processes as subclasses of more general processes involving specific participants in a specified role:

- synset sc_hasRole relation
- synset sc_hasParticipant endurant or perdurant
- synset sc_hasCoParticipant endurant or perdurant
- synset sc_stateOf endurant or perdurant

For the above examples, the mapping relations will then be:

---

7As indicated above, the subclassOf, equivalentOf and instanceOf relations apply also to perdurants and qualities. The domainOf relation is not used since the distinction between rigid and non-rigid does not apply to processes and states.
Note that such a specification is useful even for compositional terms and even if we can automatically derive the meaning of the elements of the term, since the specific role of the endurant may still be ambiguous (e.g. done-by or patient). Furthermore, these terms have a strong disambiguating effect on the rest of the text as well.

Since the target of the relation can either be an endurant or a perdurant, we can also express roles between processes and states. We can indicate that one process participates in another process in a particular role. This allows for the powerful expression of causal, conditional, and intentional relations between processes. Below are some examples:

- \( \text{wn:air pollution} \) sc_subclassOf \( \text{ont:pollution} \)
- \( \text{wn:air pollution} \) sc_hasParticipant \( \text{ont:air} \)
- \( \text{wn:air pollution} \) sc_hasRole \( \text{ont:patient} \)

- \( \text{wn:nitrogen pollution} \) sc_subclassOf \( \text{ont:pollution} \)
- \( \text{wn:nitrogen pollution} \) sc_hasParticipant \( \text{ont:nitrogen} \)
- \( \text{wn:nitrogen pollution} \) sc_hasRole \( \text{ont:done-by} \)

- \( \text{wn:nitrogen reduction} \) sc_subclassOf \( \text{ont:reduction} \)
- \( \text{wn:nitrogen reduction} \) sc_hasParticipant \( \text{ont:nitrogen} \)
- \( \text{wn:nitrogen reduction} \) sc_hasRole \( \text{ont:patient} \)

Finally, there is a specific mapping relation for part-whole relations:

- \( \text{wn:sea salt} \) sc_partOf \( \text{ont:sea} \)
- \( \text{wn:reduction effort} \) sc_partOf \( \text{ont:reduction} \)
- \( \text{wn:jet skiing} \) sc_partOf \( \text{ont:tourism} \)
- \( \text{wn:airborne chemical} \) sc_partOf \( \text{ont:air} \)
Through the mapping relations, we can keep the ontology relatively small and compact whereas we can still define the richness of the vocabularies of languages in a precise way. The classes in the ontology can be defined using rich axioms that model precise implications for inferencing. The wordnet to synset mappings can be used to define rather basic relations relative to the given ontology that still capture the full semantics of the terms. The term definitions capture both relevance and perspective (those relations that matter from the point of the view of the term), on the one hand, and some semantics with respect to the involved concepts and their (role) relation on the other hand. Likewise, the KYOTO platform can model the linguistic and cultural diversity of languages in domain but at the same time keep a firm anchoring to a basic and compact ontology.

If we compare the mapping relations to SUMO, the KYOTO mappings provide a much richer and explicit set of relations. Only the sc_equivalenceOf and sc_subclassOf relations are used in the SUMO to WordNet mapping, represented by the symbols ‘=’ and ‘+’ respectively. The SUMO-WordNet mapping likewise does not systematically distinguish rigid from non-rigid concepts. In our model, we separate the linguistically and culturally specific vocabularies from the shared ontology while using the ontology as a point of interface for the concepts used by the various communities. The lexicalization of the concepts can differ considerably across languages. Consider the following examples of different lexicalizations that can now be elegantly modelled:

{meat}Noun, English
- sc_domainOf ont:Cow, ont:Sheep, ont:Pig
- sc_playRole ont:patient
- sc_participantOf ont:eat

{名 肉, 食物, 餐 }Noun, Chinese
- sc_domainOf ont:Cow, ont:Sheep, ont:Pig, ont:Rat, ont:Dog, ont:Mole
- sc_playRole ont:patient
- sc_participantOf ont:eat

{طعام, لحم, غذاء }Noun, Arabic
- sc_domainOf ont:Cow, ont:Sheep
- sc_playRole ont:patient
- sc_participantOf ont:eat

In these examples, we see that words for meat in English, Chinese and Arabic are defined by the same role relation but have different ranges of domains, indicating what animals are considered as food.

Although each language in the KYOTO platform can define its own mappings to the ontology (through base concepts and through the above model for non-rigid domain concepts and for domain terms for processes), there is always the possibility to maintain the original Inter-Lingual-Index model of EuroWordNet (Vossen 1998). Through this model, the semantic specification of synsets in the English WordNet can be imposed on any language linked to the English model. Likewise, any language in the Global Wordnet...
Grid (Fellbaum and Vossen 2007) can thus benefit from the general work in KYOTO to provide a clear anchoring to a central ontology. Such a basic projection of the English WordNet based definition can then be made more precise by revising the ontology mappings according to the above model in a second step.
4 Creating the generic wordnet to ontology mappings

Given the above framework for defining the meaning of domain vocabularies through basic mappings to the ontology and through wordnet hyponymy relations, we can come to the following requirements for the KYOTO platform to be usable in any domain:

1 Generic Ontology coverage:
   1.1 Define the top and middle layer of endurants that abstract from any domain;
   1.2 Define a comprehensive set of processes and states in which the endurants can be involved;
   1.3 Define a comprehensive set of qualities and quality-regions that can apply to the endurants;
   1.4 Define the possible roles and relations that can apply to both endurants and perdurants;

2 Generic wordnet coverage (relevant to any language):
   2.1 Define the set of nominal Base Concepts (BCs) that dominate the full nominal hierarchy of a wordnet;
   2.2 Specify for each nominal BC the rigidity of the concept;
   2.3 Map all BCs to the ontology (endurants, perdurants, qualities and quality-regions), where:
      2.3.1 rigid BCs need to mapped through equivalent and subclass mappings;
      2.3.2 non-rigid BCs need to be defined in terms of their domain, participation and role or state;
   2.4 Map all verbal synsets to the comprehensive set of perdurants in the ontology;
   2.5 Map all adjectival synsets to the comprehensive set of states, qualities and quality_regions;

Since KYOTO has a main focus on the temporary contextual situations that can apply to any set of endurants, it is specifically important that the processes, states and qualities are well covered in the ontology. Furthermore, the Wikyoto editor in KYOTO is intended to be used by domain experts that are not trained in ontology and wordnet building. They will make selections of terms (mostly nouns) that express important and relevant properties that are worth mining. Furthermore, they will create mappings to processes and states in the ontology on the basis of mined definitions and simple interviews as described in the TMEKO procedure (Segers and Vossen 2010). This means that especially the verbs and adjectives in languages need to be matched to ontology beforehand.

We organized this work as follows in KYOTO:

1 Nominal synsets:
   1.1 A set of nominal BCs has been defined from the English WordNet3.0, which properly represents the full hierarchy;
   1.2 The nominal BCs have been modelled in the KYOTO ontology;
   1.3 Equivalences for the BCs have been defined in the wordnets for 6 other languages in KYOTO: Dutch, Italian, Spanish, Basque, Japanese and Chinese;
   1.4 Rigidity has been determined for all the BCs in all the languages;
1.5 Synset to ontology mappings have been provided for the BCs;

The ontology was further extended on the basis of a document on the Humber Estuary that was used by the environmental experts for creating mind maps (Deliverables D08.1/D08.2). This has resulted in mappings from the English WordNet to more specific domain synsets. The work on 1.1 and 1.2 was reported in deliverable D06.3. The other steps (1.3, 1.4 and 1.5) are described below in section 5.

In order to provide a full coverage of the domain, we focussed in the last period on modelling many verbs and adjectives that are relevant to the domain, so that the nominal domain terms can be mapped to their corresponding processes and states.

2 Verb and adjective synsets:

2.1 Three English benchmark documents have been processed in the KYOTO system and all noun, verb and adjective terms have been extracted;

2.2 Each term that also occurs in the generic English WordNet was matched with the most likely synset through the word-sense-disambiguation module (Agirre et al 2009);

2.3 We validated the synset matching for verbs and extracted mapping relations for each important verb to the corresponding ontology class. When necessary, the ontology was extended with processes and states to properly represent the meaning of the verb;

2.4 We selected the most important adjectives, both in terms of frequency and in terms of relevance for the verbals and nominal terms in the 3 documents. Whenever necessary, the ontology was extended to represent the qualities denoted by the adjectives and mappings were defined from each adjective to the ontology;

The work on the verbs and adjectives is further reported in section 6 below. These verbs and adjectives have also resulted in an extension of the KYOTO ontology. This work is reported on in deliverable D06.5.
5 Rudify

Rudify employs (often unconscious) human judgments on ontological meta-properties by harvesting vast amounts of data through the Google web search API. It helps to automatically determine for a concept whether it is rigid or non-rigid based on its lexical representation. This procedure is described in detail in D06.3 “Wordnets mapped to central ontology-1”.

The Rudify tools currently support the following Kyoto languages besides English: Dutch, Spanish, Italian, Basque, and Japanese. The software is available freely at http://rudify.sorceforge.net/ and will be available from the Kyoto home page.

Evaluation

Monolingual evaluation

The results of the evaluation for the English data sets are reported in D6.3. They are also published in Hicks and Herold (2009). For methodological considerations on relying on Google’s web service also see Hicks and Herold (2009).

The initial English training data was translated into Dutch, Spanish, Italian, Basque, and Japanese and was used for the monolingual evaluation. As there is no linguistic module for Japanese implemented in Rudify as of the time of writing this language is excluded from this deliverable. Basque is excluded from the evaluation as well because there is too little data available through the Google API. An update will be provided at a later time.

There are two main observations:

- Not all patterns developed for the different languages work equally well. In order to measure the individual contribution of patterns to the classification task information gain was computed per pattern on the basis of the training data sets and the manual evaluation of the base concept data sets for English, Spanish and Dutch. For Italian information gain was computed on the basis of the training set.

Overview of the contributions of individual patterns per language:

<table>
<thead>
<tr>
<th>Language</th>
<th>Contributing Patterns</th>
<th>Non-Contributing Patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>became (a</td>
<td>an) X</td>
</tr>
<tr>
<td></td>
<td>is now (a</td>
<td>an) X</td>
</tr>
<tr>
<td></td>
<td>would be (a</td>
<td>an) (good</td>
</tr>
<tr>
<td></td>
<td>would make (a</td>
<td>an) (good</td>
</tr>
<tr>
<td>Dutch</td>
<td>is nu (een) X (wordt</td>
<td>worden) (een) X</td>
</tr>
<tr>
<td></td>
<td></td>
<td>hield op (met) (een) X te zijn</td>
</tr>
<tr>
<td></td>
<td></td>
<td>is geen X meer terwijl (hijzij</td>
</tr>
<tr>
<td></td>
<td></td>
<td>zou een (goede</td>
</tr>
<tr>
<td></td>
<td></td>
<td>wordt beschouwd als (een) X</td>
</tr>
<tr>
<td>Spanish</td>
<td>ahora es (el</td>
<td>la</td>
</tr>
<tr>
<td></td>
<td>se convirtio en (el</td>
<td>la</td>
</tr>
<tr>
<td></td>
<td></td>
<td>mientras fue (el</td>
</tr>
</tbody>
</table>
For every language performance in classification is notably worse than for English.

### Performance on the English gold standard data set for different classifiers (50 rigid terms, 50 non-rigid terms):

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Mis-class. R+/R-</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>5/10</td>
</tr>
<tr>
<td>Logistic</td>
<td>0.78</td>
<td>0.77</td>
<td>0.77</td>
<td>6/17</td>
</tr>
<tr>
<td>NNge</td>
<td>0.72</td>
<td>0.69</td>
<td>0.68</td>
<td>24/7</td>
</tr>
<tr>
<td>LWL</td>
<td>0.78</td>
<td>0.76</td>
<td>0.76</td>
<td>6/18</td>
</tr>
<tr>
<td>SVM, linear</td>
<td>0.81</td>
<td>0.69</td>
<td>0.66</td>
<td>0/31</td>
</tr>
<tr>
<td>SVM, quadr.</td>
<td>0.77</td>
<td>0.73</td>
<td>0.72</td>
<td>3/24</td>
</tr>
</tbody>
</table>

### Performance on the Dutch gold standard data set (50 rigid terms, 50 non-rigid terms):

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Mis-class. R+/R-</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48</td>
<td>0.61</td>
<td>0.54</td>
<td>0.45</td>
<td>3/43</td>
</tr>
<tr>
<td>Logistic</td>
<td>0.64</td>
<td>0.57</td>
<td>0.51</td>
<td>4/39</td>
</tr>
<tr>
<td>NNge</td>
<td>0.51</td>
<td>0.51</td>
<td>0.51</td>
<td>27/22</td>
</tr>
<tr>
<td>LWL</td>
<td>0.58</td>
<td>0.54</td>
<td>0.47</td>
<td>5/41</td>
</tr>
<tr>
<td>SVM, linear</td>
<td>0.67</td>
<td>0.59</td>
<td>0.53</td>
<td>3/38</td>
</tr>
<tr>
<td>SVM, quadr.</td>
<td>0.65</td>
<td>0.58</td>
<td>0.53</td>
<td>4/38</td>
</tr>
</tbody>
</table>

### Performance on the Italian gold standard data set (50 rigid terms, 50 non-rigid terms):

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Mis-class. R+/R-</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48</td>
<td>0.74</td>
<td>0.61</td>
<td>0.55</td>
<td>1/38</td>
</tr>
<tr>
<td>Logistic</td>
<td>0.69</td>
<td>0.60</td>
<td>0.55</td>
<td>3/36</td>
</tr>
<tr>
<td>NNge</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>21/19</td>
</tr>
<tr>
<td>LWL</td>
<td>0.71</td>
<td>0.57</td>
<td>0.48</td>
<td>1/42</td>
</tr>
<tr>
<td>SVM, linear</td>
<td>0.75</td>
<td>0.62</td>
<td>0.56</td>
<td>1/37</td>
</tr>
<tr>
<td>SVM, quadr.</td>
<td>0.70</td>
<td>0.59</td>
<td>0.53</td>
<td>2/39</td>
</tr>
</tbody>
</table>

### Performance on the Spanish gold standard data set (36 rigid terms, 34 non-rigid terms):
There is room for improving the individual classification results by elaborating the language specific patterns. As the classification concerns a judgment on an inherently non-linguistic property it should be possible to improve the accuracy of the procedure by a cross lingual combination of the data.

Crosslingual combination evaluation

As a preliminary result of an evaluation of the classification performance for combined data sets from different languages we report on the combination of English and Dutch.

Mis-classifications on the combined data sets for English and Dutch:

<table>
<thead>
<tr>
<th>Classifier</th>
<th>English (+R/-R)</th>
<th>Dutch (+R/-R)</th>
<th>Combined (+R/-R)</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48</td>
<td>5/10</td>
<td>3/43</td>
<td>3/12</td>
</tr>
<tr>
<td>Logistic</td>
<td>6/17</td>
<td>4/39</td>
<td>10/20</td>
</tr>
<tr>
<td>NNge</td>
<td>24/7</td>
<td>27/22</td>
<td>30/7</td>
</tr>
<tr>
<td>LWL</td>
<td>6/18</td>
<td>5/41</td>
<td>1/20</td>
</tr>
<tr>
<td>SVM, linear</td>
<td>0/31</td>
<td>3/38</td>
<td>1/28</td>
</tr>
<tr>
<td>SVM, quadr.</td>
<td>3/24</td>
<td>4/38</td>
<td>1/26</td>
</tr>
</tbody>
</table>

There is in general no positive effect on classifier performance obtained by the combination of the data sets. Only support vector machines with linear kernels perform slightly better in this setting than on individual data sets. Due to the poor performance on data sets for any language except English this is an expected observation.

Exploitation and Rudify interviews

There are some concerns with regard to the direct application of the Rudify software in the interactive Kyoto system:

- On-line queries take up too much time for an interactive system, even more so when launched for different languages. This would cause unacceptable delays for users that want to add a synset to the wordnet domain extension. The tool could still be used off-line, i.e. asynchronous with users editing the wordnet instead.
- We expect sparse data being an increasing problem with more specific domain terms that are rare in non-expert discourses.
- Currently classification performance is acceptable only for English data. Therefore the Rudify procedure will be reversed in the Kyoto system. Users are actively
exposed questions and the decision between rigid and non-rigid concepts will be based on their answers in this question answering task. A similar approach has been adopted by the Senso comune project, (Oltramari and Vetere 2008). There is no production version of their interview system available yet, though.

In the following sections we outline three different approaches to the user based rigidity decision task.

**Method 1**
The Kyoto ontology can be traversed top-down and used as a decision tree. If the user edits a non-rigid synset he will construct a path leading to the [ROLE] hierarchy of the ontology. The major disadvantage of this method is that the users are exposed to a high level philosophical decision task even if the questions are reformulated somewhat colloquially. Furthermore it is not clear that the underlying questions can be formulated in a way so that philosophically untrained users can easily and correctly answer them. The number of questions that have to be answered for a single synset addition is at least equal to the dimension of the shortest rooted path within the ontology. This roughly resembles the idea described by Oltramari and Vetere.

**Method 2**
The second method closely resembles the idea behind Rudify and narrowly focuses on the rigidity property without concerning the higher level context of the ontology: a set of sentences is constructed using the lexical patterns that the Rudify software tries to query from web search engines with lexical representations from the target synset as filler. The task for user then is to judge whether those sentences are plausible or nonsensical. The advantage of this method compared to the first one is that is essentially a rating task over general language material rather then a delicate question answering task. This method requires the synset to be anchored with a superordinate synset in order to enable the question generator to automatically construct appropriate questions as typically a base concept will be part of the sentences. Two possible sets of questions are:

Please mark the phrases with [European Bee-eater] in the sense of [gloss(European Bee-eater)] that are clearly nonsensical or implausible:
- John knows an animal that would be a good/prototypical European Bee-eater
- Some animals are regarded as European Bee-eaters
- Gradually some animals became European Bee-eaters
- Given these problems it is understandable when an animal stops being a European Bee-eater

- John knows a bird/animal that would be a good/prototypical migratory bird
- Some birds/animals are regarded as migratory birds
- Gradually some birds/animals became migratory birds
- Given these problems it is understandable when a bird/animal stops being a migratory bird

The first set is generated for the target lexical representation *European Bee-eater* with *animal* as the associated base concept. Within the given domain these sentences sound strange and
implausible. The second set targets the domain term *migratory bird* with either *bird* or *animal* as superordinate base concept. The constructed sentences are plausible and understandable within the environmental domain.

**Method 3**

This method directly exploits the properties necessity and duration underlying rigidity which the second method tries to determine indirectly. Thus only two questions are necessary in order to decide between rigid and non-rigid concepts. Examples are:

1. If a thing/being is a European Bee-eater: could it be something else instead as well? [yes/no]
2. If an animal/a bird is a European Bee-eater: can it cease being a European Bee-eater? [yes/no]

1. If a thing/being is a migratory bird: could it be something else instead as well? [yes/no]
2. If a bird/an animal is a migratory bird: can it cease being a migratory bird? [yes/no]

By answering the first question the user decides on the property of necessity for *European Bee-eater* and *migratory bird* respectively; the second questions targets duration. Only if both questions are positively confirmed it is safe to assume that the user models a rigid concept.

This last method has the advantage of being philosophically sound and comparatively fast and easy to perform by untrained users. The actual questions need to be carefully reformulated in order to be more easily understandable for the domain users. It will be implemented in the Wikyoto editor.
6 Extending the mapping relations to model processes and states for the domain concepts

As explained, the KYOTO platform focuses on non-rigid domain terms that need to mapped to processes and properties that may apply, whereas rigid terms are related to basic types through the wordnet hyponymy relations. To get an initial definition of the processes and properties that matter in this domain, we extracted the terms from 3 English documents using the Tybot system (Deliverable D05.1, Bosma and Vossen 2010). Two documents are about the Humber Estuary in the UK and one document is on the Chesapeake Bay in the US.

We extracted all the nominal, verbal and adjectival terms with a confidence score of 0.4 or higher. The next table gives an overview of the extracted terms. In total, 3950 terms have been detected, most of which are nouns.

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>WN mapping</th>
<th>Proportion</th>
<th>New terms</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nouns</td>
<td>2818</td>
<td>1130</td>
<td>40.1%</td>
<td>1688</td>
<td>59.9%</td>
</tr>
<tr>
<td>Verbs</td>
<td>534</td>
<td>525</td>
<td>98.3%</td>
<td>9</td>
<td>1.7%</td>
</tr>
<tr>
<td>Adjectives</td>
<td>598</td>
<td>513</td>
<td>85.8%</td>
<td>85</td>
<td>14.2%</td>
</tr>
<tr>
<td>Total</td>
<td>3950</td>
<td>2168</td>
<td>54.9%</td>
<td>1782</td>
<td>45.1%</td>
</tr>
</tbody>
</table>

Table 1: Overview of the terms extracted for the benchmark documents on English Estuaries

Since the documents have been tagged with synsets through WSD, also the terms have been tagged with the most likely synsets. In the case of the nouns, 40% of the terms was matched with WordNet synsets, whereas for verbs and adjectives this was 98% and 85% respectively. This clearly shows that new terminology is mainly restricted to nouns.

The WordNet mapping allowed us to create a WordNet hierarchy on top of the extracted terms. Such hierarchies represent a conceptual clustering of the concepts that play a role in the documents. A conceptual clustering is more efficient for deciding on the relevant concepts. In these hierarchies, new terms are linked to their structural head as child concepts (e.g. estuary birds is linked to bird) and terms that matched with synsets are linked to the hypernyms in WordNet. If the hypernym did not occur in the documents, it was introduced into the term set. Likewise, we have the following three types of concepts as shown in the screen dumps of the hierarchies below:

- new terms from the documents that do not occur in WordNet (blue in the screen dumps below)
- terms from the documents that also occur in WordNet (black in the screen dumps below)
- terms that do not occur in the documents but are hypernyms of concepts/synsets that do occur (red in the screen dump below)
Drawing 2: Hierarchy fragment extracted for nominal terms
**Drawing 3: Hierarchy fragments extracted for verbal terms**

<table>
<thead>
<tr>
<th>sc_subClassOf</th>
<th>sc_subClassOf</th>
<th>sc_subClassOf</th>
<th>sc_subClassOf</th>
<th>sc EquivalentOf</th>
<th>sc EquivalentOf</th>
<th>scEquivalentOf</th>
<th>sc EquivalentOf</th>
<th>sc EquivalentOf</th>
<th>sc EquivalentOf</th>
</tr>
</thead>
<tbody>
<tr>
<td>consequence_effect_outcome</td>
<td>consequence_effect_outcome</td>
<td>improvement-eng-3.0-002</td>
<td>improvement-eng-3.0-002</td>
<td>clean&amp;improvement-eng-1</td>
<td>empty&amp;change_of_state-e</td>
<td>drain_drainage-eng-3.0-002</td>
<td>change of Integrity-eng-3.0</td>
<td>enable verb conf:0.409384</td>
<td>empower verb conf:0.409384</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>nDescendants:2 synset:eng-30-00512877-v-1.0</td>
<td>nDescendants:2 synset:eng-30-00316787-v-1.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>strengthen verb conf:0.409384</td>
<td>steady verb conf:0.409384</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>nDescendants:2 synset:eng-30-00328669-v-0.36568</td>
<td>nDescendants:1 synset:eng-30-02472339-v-0.523002</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>clean verb conf:0.409384</td>
<td>wash verb conf:0.409384</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>nDescendants:2 synset:eng-30-01532593-v-0.168059</td>
<td>nDescendants:1 synset:eng-30-01533246-v-0.105453</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>empty verb conf:0.409384</td>
<td>drain verb conf:0.532495</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>nDescendants:4 synset:eng-30-00496962-v-0.252472</td>
<td>nDescendants:1 synset:eng-30-00351648-v-0.27331</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>hollow out verb conf:0.0</td>
<td>undermine verb conf:0.409384</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>nDescendants:2 synset:eng-30-01282545-v-1.0</td>
<td>nDescendants:1 synset:eng-30-01282545-v-1.0</td>
</tr>
</tbody>
</table>
We can clearly observe that the nominal hierarchy has most structure (5 tops), the verbal hierarchy is flatter (118 top concepts) and the adjectival terms form a flat list (513 top concepts), due to the WordNet structure. For nouns, 322 Wordnet synsets were added to represent hypernym layers not in the document. For verbs 220, synsets were introduced and for adjectives none.

To provide the mapping to the ontology and suggest new extensions, the tables with the term hierarchies and lists were extended with two additional columns with text fields, as is shown for verbs and adjectives (the 3rd text field for adjectives has not been used). We used the text field to specify the mapping relation between the term and the ontology. The first column was used to specify the mapping relation, and the second column holds the ontology concept to which the term should be matched. For example for the adjective “environmental”, we indicate that it is a subclass of the concept environmental_condition which was already in the KYOTO ontology:

```
sc_equivalentOf
environmental_condition-eng-3.0-14516501-n
environmental adjective conf:0.58094 nDescendants:1 synset:eng-30-02943303-a:0.500193
```

**Drawing 4: Hierarchy fragment extracted for adjectival terms**

We can clearly observe that the nominal hierarchy has most structure (5 tops), the verbal hierarchy is flatter (118 top concepts) and the adjectival terms form a flat list (513 top concepts), due to the WordNet structure. For nouns, 322 Wordnet synsets were added to represent hypernym layers not in the document. For verbs 220, synsets were introduced and for adjectives none.

To provide the mapping to the ontology and suggest new extensions, the tables with the term hierarchies and lists were extended with two additional columns with text fields, as is shown for verbs and adjectives (the 3rd text field for adjectives has not been used). We used the text field to specify the mapping relation between the term and the ontology. The first column was used to specify the mapping relation, and the second column holds the ontology concept to which the term should be matched. For example for the adjective “environmental”, we indicate that it is a subclass of the concept environmental_condition which was already in the KYOTO ontology:

```
sc_equivalentOf
environmental_condition-eng-3.0-14516501-n
environmental adjective conf:0.58094 nDescendants:1 synset:eng-30-02943303-a:0.500193
```
Similarly, we mapped “recent” to the existing class temporal-region through a subClassOf relation:

```
sc_subclassOf
temporal-region
recent adjective conf:0.58094 nDescendants:1 synset:eng-30-01730444-a:0.515613
```

In other cases, we considered a term so relevant that we propose an extension to the ontology. This is for example the case for the next two examples:

```
sc_equivalentOf
*clean&improvement-eng-3.0-00248977-n
clean verb conf:0.409384 nDescendants:2 synset:eng-30-01532589-v:0.168059
sc_equivalentOf
*ecological&indefinite_quality
ecological adjective conf:0.523495 nDescendants:1 synset:eng-30-02906478-a:0.505865
```

The target of these terms is proceeded by a ‘*’, indicating that it is a new class, followed by the superclass of the new concept following ‘&’.

We read the 3 documents carefully and we evaluated each verb and adjective with respect to its relevance and correctness in interpretation. The latter was limited to verbs, since the adjectives do not have a semantic hierarchy. Most of the verb synsets have been assigned correctly, in the sense that the hypernym of the verbal synset matches with the interpretation of the text.

Table-2 gives an overview of the mappings. We can see that 42% of the verbs and 38% of the adjectives have been selected as being relevant. Of these, 35% verbs have an equivalent relation to an ontology class and 65% has a subclass relation. For adjectives this is almost reversed: 70% versus 30%. This difference is due to the fact that verbs could very often be mapped to an equivalent nominal concept (e.g. improve and improvement) that was already present. In the case of the adjectives, this was mostly not the case and we had to add many concepts to the ontology for adjectives denoting various qualities. For the verbal synsets, we added 43 new classes to the ontology and for the adjectives 119 additions were created. Whenever a new class was created, we could map the verbs and adjectives directly through an equivalenceOf relation.

**Table 2: Mapping of subset of verbs and adjectives to the KYOTO ontology**

<table>
<thead>
<tr>
<th></th>
<th>Total terms</th>
<th>Total mapped</th>
<th>% of total terms</th>
<th>sc_equivalentOf</th>
<th>% of total mapped</th>
<th>sc_subClassOf</th>
<th>% of total mapped</th>
</tr>
</thead>
<tbody>
<tr>
<td>verbs</td>
<td>534</td>
<td>225</td>
<td>42.13%</td>
<td>79</td>
<td>35.11%</td>
<td>146</td>
<td>64.89%</td>
</tr>
<tr>
<td>adjectives</td>
<td>598</td>
<td>227</td>
<td>37.96%</td>
<td>160</td>
<td>70.48%</td>
<td>67</td>
<td>29.78%</td>
</tr>
</tbody>
</table>

The current selection represents just a small proportion of the total set of adjetival (30,004) synsets and verbal (25,047) synsets in WordNet. However, many of the parent concepts will also be applicable to other more specific synsets (in the case of verbs), and to so-called satellite concepts (in the case of adjectives).

In order to extend the coverage of of the mapping for verbs and adjectives we developed a program to automatically expand the verbal and adjectival mapping.
Regarding the **verbal mapping** containing manually created links between a verbal synset and a class in the ontology, the program also assigns to the ontology class all hyponymy descendants of the verbal synset. Consider for instance the following connection between a verbal synset and a class of the ontology:

\[
\text{eng-30-00151689-v sc\_equivalentOf Kyoto\#decrease\_in\_magnitude}
\]

The synset `eng-30-00151689-v` corresponds to the verbal concept in WordNet:

*decrease\_1 diminish\_1 lessen\_1 fall\_11 decrease in size, extent, or range: The amount of homework decreased towards the end of the semester; The cabin pressure fell dramatically; her weight fell to under a hundred pounds; his voice fell to a whisper;*

Following its hyponymy hierarchy in WordNet, this synset has sixty descendant verbal concepts. All of them has been automatically assigned to the same ontological class in the form with a `sc\_subclassOf` relation in the form:

\[
\begin{align*}
\text{eng-30-00093979-v sc\_subclassOf Kyoto\#decrease\_in\_magnitude} \\
\text{eng-30-00107604-v sc\_subclassOf Kyoto\#decrease\_in\_magnitude} \\
\text{eng-30-00152558-v sc\_subclassOf Kyoto\#decrease\_in\_magnitude} \\
\text{eng-30-00152762-v sc\_subclassOf Kyoto\#decrease\_in\_magnitude}
\end{align*}
\]

Currently, the total number of verbal mappings is 202 between 189 synsets from WordNet and 86 classes from the ontology. After a validation and expansion process, the total number of verbal mappings is 5989 between 5978 synsets from WordNet and 78 classes from the ontology. That is, after the expansion, the verbal mapping augmented more that 29 times.

Regarding the **adjectival mapping**, we performed in a similar way. Having a manually created mapping between an adjectival synset and a class in the ontology, the program also assigns to the ontology class all satellites of the adjectival synset with the same polarity (adjectives are organized by groups of antonyms). Consider for instance the following connection between an adjectival synset and a class of the ontology:

\[
\text{eng-30-00124077-a sc\_equivalentOf Kyoto\#aquatic}
\]

The synset `eng-30-00124077-a` corresponds to the adjectival concept in WordNet:

*aquatic\_2 operating or living or growing in water: boats are aquatic vehicles;*

Following its similar-to relations in WordNet, this synset has four related adjectival concepts. All of them has been automatically assigned to the same ontological class in the form with a `sc\_subclassOf` relation in the form:

\[
\begin{align*}
\text{eng-30-00124353-a sc\_subclassOf Kyoto\#aquatic} \\
\text{eng-30-00124493-a sc\_subclassOf Kyoto\#aquatic} \\
\text{eng-30-00124685-a sc\_subclassOf Kyoto\#aquatic}
\end{align*}
\]

Currently, the total number of verbal mappings is 224 between 222 synsets from
WordNet and 160 classes from the ontology. After a validation and expansion process, the total number of verbal mappings is 1081 between 1081 synsets from WordNet and 157 classes from the ontology. That is, after the expansion, the verbal mapping augmented more than 4 times.

Consider also that the verbal and adjectival coverage is augmented not only because of the potential number of synonyms of each synset.

7 Future work

The current mappings have focused on the verbs and adjectives that play an important role in the 3 English documents that have been selected. This is fine as far as we want to model to the relations that matter for these 3 documents and it will also yield many useful data for any other document from the same domain. The current ontological model and the mapping of the WordNet to this model will be further extended in the next year of the project to provide the relations that are relevant for the environment as a whole. In the next phase, we will extract the verbs and adjectives for complete databases including hundreds up to thousands of documents. This will give a comprehensive list of the semantic relations and properties. We will use the complete database to achieve a full coverage of processes and properties denoted by verbs and adjectives for the domain.

However, KYOTO also intends to provide a generic platform for modelling the knowledge that can be used for any other domain, along the same approach. This implies that we need to represent all processes and states in an abstract way as they are represented by the verbs and adjectives in WordNet. Whereas for nouns, we have been able to define the BCs that dominate the full hierarchy, this is more difficult for verbs and adjectives. The verbal hierarchy does not have a well-defined top structure as the nominal hierarchy and the adjectives have no hierarchy at all. We will explore a number of different strategies though to define an abstract and general layer for mapping the verbal and adjectival synsets, among which:

1. Wordnets in other languages do have a structured hierarchy for verbs and adjectives that can be imposed on the English WordNet;
2. Cross-part-speech relations (e.g. to act and action) can be used to impose the nominal hierarchy on verbs and partially also on adjectives (e.g. colour pertains to red, blue, yellow, etc.);
3. The SUMO ontology with its wordnet mappings can be exploited to provide a basic classification of verbals and adjectival synsets in the English wordnet;
4. FrameNet provides another classification of verbs and adjectives that can be used. FrameNet is being mapped to WordNet both manually (Baker and Fellbaum 2008, 2009; Fellbaum and Baker in press) and automatically (Burchardt, Erk and Frank 2005; Chow and Webster 2007; Ferrandez et al. 2010; Laparra, Rigau and Cuadros 2010; Oltramari 2006; Tonelli and Pianta 2009, inter alia).
5. The satellite structure of adjectives in WordNet can be used to expand ontology classes to more adjectives.

The above approaches will provide a comprehensive modelling of abstract relations and properties in the ontology and their corresponding mapping to a minimal set of nominal, verbal and adjectival synsets in the English WordNet. Another major effort for the last year of KYOTO will be to impose this on the wordnets in other languages. For this, we will
first use the equivalence relations between the synsets in these languages and the English WordNet and secondly use the internal semantic relations in each wordnet to either find more concepts to be covered or the validate the ontological mappings imposed through the English WordNet.

Finally, we have not discussed the mapping of nominal domain terms to the ontology. This is because we assume that this work can be done by the domain experts once the there is a sufficient coverage through the generic wordnets to the abstract set of endurants and the comprehensive set of perdurants, qualities and quality-regions. As was shown in table-1, most of the domain specific terms are nouns that somehow depend on concepts referred to by verbs and adjectives from the general vocabulary. First work on editing such domain terms is reported in the deliverable D08.1/D08.2, which involves a subset of nominal terms selected by the environment experts.

8 References

KYOTO (211423) Deliverables available at: www.kyoto-project.eu:
D05.1 Concept miners version 1
D06.3 Wordnets mapped to central ontology version 1
D06.4 Automatic deduction and inferencing techniques
D06.5 Central ontology revised
D07.4b Wiki environment for wordnet editing revised
D07.5b Wiki environment for ontology editing revised
D08.1 Domain extions of central ontology – version 1
D08.2 Domain extions of wordnets – version 1


Ferrández, Oscar; Ellsworth, Michael; Munoz, Rafael & Baker, Collin F. Forthcoming. A Graph-based Measure of FrameNet-WordNet Alignment. *Proceedings of ICGL 2010*


Oltramari, Alessandro. 2006. LexiPass methodology: a conceptual path from frames to senses and back. Proceedings of LREC.


Appendix I Set of verbal mappings

sc_equivalentOf absorb eng-30-01539063-v
sc_equivalentOf accomplishment eng-30-01640855-v
sc_equivalentOf accomplishment eng-30-02526085-v
sc_equivalentOf accomplishment eng-30-02526085-v
sc_equivalentOf action eng-30-01645601-v
sc_equivalentOf action eng-30-01719302-v
sc_equivalentOf active-change-of-location eng-30-01835496-n
sc_equivalentOf absorb eng-30-01539063-v
sc_equivalentOf accomplishment eng-30-01640855-v
sc_equivalentOf accomplishment eng-30-02526085-v
sc_equivalentOf accomplishment eng-30-02526085-v
sc_equivalentOf action eng-30-01645601-v
sc_equivalentOf action eng-30-01719302-v
sc_equivalentOf active-change-of-location eng-30-01835496-n
sc_equivalentOf absorb eng-30-01539063-v
sc_equivalentOf accomplishment eng-30-01640855-v
sc_equivalentOf accomplishment eng-30-02526085-v
sc_equivalentOf accomplishment eng-30-02526085-v
sc_equivalentOf action eng-30-01645601-v
sc_equivalentOf action eng-30-01719302-v
sc_equivalentOf active-change-of-location eng-30-01835496-n
sc_equivalentOf absorb eng-30-01539063-v
sc_equivalentOf accomplishment eng-30-01640855-v
sc_equivalentOf accomplishment eng-30-02526085-v
sc_equivalentOf accomplishment eng-30-02526085-v
sc_equivalentOf action eng-30-01645601-v
sc_equivalentOf action eng-30-01719302-v
sc_equivalentOf active-change-of-location eng-30-01835496-n
sc_equivalentOf absorb eng-30-01539063-v
sc_equivalentOf accomplishment eng-30-01640855-v
sc_equivalentOf accomplishment eng-30-02526085-v
sc_equivalentOf accomplishment eng-30-02526085-v
sc_equivalentOf action eng-30-01645601-v
sc_equivalentOf action eng-30-01719302-v
sc_equivalentOf active-change-of-location eng-30-01835496-n
sc_equivalentOf absorb eng-30-01539063-v
sc_equivalentOf accomplishment eng-30-01640855-v
sc_equivalentOf accomplishment eng-30-02526085-v
sc_equivalentOf accomplishment eng-30-02526085-v
sc_equivalentOf action eng-30-01645601-v
sc_equivalentOf action eng-30-01719302-v
sc_equivalentOf active-change-of-location eng-30-01835496-n
sc_equivalentOf absorb eng-30-01539063-v
sc_equivalentOf accomplishment eng-30-01640855-v
sc_equivalentOf accomplishment eng-30-02526085-v
sc_equivalentOf accomplishment eng-30-02526085-v
sc_equivalentOf action eng-30-01645601-v
sc_equivalentOf action eng-30-01719302-v
sc_equivalentOf active-change-of-location eng-30-01835496-n
sc_equivalentOf absorb eng-30-01539063-v
sc_equivalentOf accomplishment eng-30-01640855-v
sc_equivalentOf accomplishment eng-30-02526085-v
sc_equivalentOf accomplishment eng-30-02526085-v
sc_equivalentOf action eng-30-01645601-v
sc_equivalentOf action eng-30-01719302-v
sc_equivalentOf active-change-of-location eng-30-01835496-n
sc_equivalentOf absorb eng-30-01539063-v
sc_equivalentOf accomplishment eng-30-01640855-v
sc_equivalentOf accomplishment eng-30-02526085-v
sc_equivalentOf accomplishment eng-30-02526085-v
sc_equivalentOf action eng-30-01645601-v
sc_equivalentOf action eng-30-01719302-v
sc_equivalentOf active-change-of-location eng-30-01835496-n
sc_equivalentOf absorb eng-30-01539063-v
sc_equivalentOf accomplishment eng-30-01640855-v
sc_equivalentOf accomplishment eng-30-02526085-v
sc_equivalentOf accomplishment eng-30-02526085-v
sc_equivalentOf action eng-30-01645601-v
sc_equivalentOf action eng-30-01719302-v
sc_equivalentOf active-change-of-location eng-30-01835496-n
sc_equivalentOf absorb eng-30-01539063-v
sc_equivalentOf accomplishment eng-30-01640855-v
sc_equivalentOf accomplishment eng-30-02526085-v
sc_equivalentOf accomplishment eng-30-02526085-v
sc_equivalentOf action eng-30-01645601-v
sc_equivalentOf action eng-30-01719302-v
sc_equivalentOf active-change-of-location eng-30-01835496-n
sc_equivalentOf absorb eng-30-01539063-v
sc_equivalentOf accomplishment eng-30-01640855-v
sc_equivalentOf accomplishment eng-30-02526085-v
sc_equivalentOf accomplishment eng-30-02526085-v
sc_equivalentOf action eng-30-01645601-v
sc_equivalentOf action eng-30-01719302-v
sc_equivalentOf active-change-of-location eng-30-01835496-n
Appendix II: set of adjectival mappings

sc_equivalentOf *little eng-30-01391351-a
sc_equivalentOf *native eng-30-01036083-a
sc_equivalentOf abundant eng-30-00013887-a
sc_equivalentOf adult eng-30-02134297-a
sc_equivalentOf agriculture eng-30-0082241-a
sc_equivalentOf agriculture eng-30-01865807-a
sc_equivalentOf agriculture eng-30-02790222-a
sc_equivalentOf airborne eng-30-01522895-a
sc_equivalentOf appropriate eng-30-00134701-a
sc_equivalentOf aquatic eng-30-00124077-a
sc_equivalentOf artifact-role eng-30-001573568-a
sc_equivalentOf availability eng-3.0-04718999-n eng-30-00183053-a
sc_equivalentOf average eng-30-00486290-a
sc_equivalentOf bad eng-30-01125429-a
sc_equivalentOf big eng-30-01382086-a
sc_equivalentOf blue eng-30-00425002-a
sc_equivalentOf breath&activity eng-30-00267452-a
sc_equivalentOf change eng-3.0-00191142-n eng-30-00808822-a
sc_equivalentOf chemical eng-30-02692624-a
sc_equivalentOf clean eng-30-00417413-a
sc_equivalentOf clear eng-30-00431447-a
sc_equivalentOf coarse eng-30-01950198-a
sc_equivalentOf color__colour__coloring__colouring eng-3.0-04956594-n eng-30-00394135-a
sc_equivalentOf commerce__commercialism__mercantilism eng-3.0-01090446-n eng-30-00483146-a
sc_equivalentOf compulsory eng-30-00848466-a
sc_equivalentOf concentration eng-3.0-001903893-n engaging-30-001157762-a
sc_equivalentOf considerable eng-30-00624026-a
sc_equivalentOf creation_1__creative_activity_1 eng-3.0-00908492-n eng-30-01288690-a
sc_equivalentOf critical eng-30-00650577-a
sc_equivalentOf damage eng-30-00679147-a
sc_equivalentOf death eng-30-00095280-a
sc_equivalentOf decedent eng-30-00914104-a
sc_equivalentOf decrease_in_magnitude eng-30-00881735-a
sc_equivalentOf degrade eng-30-00154956-a
sc_equivalentOf depth eng-30-00445937-a
sc_equivalentOf detrimental eng-30-001161984-a
sc_equivalentOf development eng-3.0-00205029-n eng-30-00741867-a
sc_equivalentOf development eng-3.0-00205029-n eng-30-01302544-a
sc_equivalentOf diffuse eng-30-00157887-a
sc_equivalentOf diminished eng-30-00882890-a
sc_equivalentOf dirty eng-30-00419289-a
sc_equivalentOf dissolved eng-30-001506258-a
sc_equivalentOf drain__drainage eng-3.0-00396029-n eng-30-001088478-a
sc_equivalentOf dry eng-30-02551380-a
sc_equivalentOf ecological eng-30-02906478-a
sc_equivalentOf economy eng-30-00840212-a
sc_equivalentOf education eng-3.0-00883297-n eng-30-02946221-a
sc_equivalentOf effective eng-30-00510644-a
sc_equivalentOf environmental_condition eng-3.0-14516501-n eng-30-02943303-a
sc_equivalentOf essential eng-30-00900616-a
sc_equivalentOf female eng-30-001477806-a
sc_equivalentOf flat eng-30-02277078-a
sc_equivalentOf fluvial eng-30-02729812-a
sc_equivalentOf fragile eng-30-02164913-a
sc_equivalentOf frequency eng-30-001066542-a
sc_equivalentOf fresh eng-30-001067694-a
sc_equivalentOf full eng-30-001083157-a
sc_equivalentOf good eng-30-001123148-a
sc_equivalentOf grazing eng-3.0-00841091-n eng-30-02445394-a