# KYOTO Annual Report 2010

## “Knowledge Yielding Ontologies for Transition-based Organization”

**http://www.kyoto-project.eu/**

***KYOTO makes knowledge sharable between communities of people, across cultures and languages and it makes this knowledge understandable to computers, by assigning meaning to text and giving text to meaning.***

# 1. Introduction

The globalization of markets and communication brings with it a concomitant globalization of world-wide problems and the need for new solutions. Timely examples are global warming, climate change and other environmental issues related to rapid growth and economic developments. Environmental problems can be acute, requiring immediate support and action, relying on information available elsewhere. Knowledge sharing and transfer are also essential for sustainable growth and development on a longer term. In both cases, it is important that distributed information and experience can be re-used on a global scale. The globalization of problems and their solutions requires that information and communication be supported across a wide range of languages and cultures. Such a system should furthermore allow both experts and laymen to access this information in their own language, without recourse to cultural background knowledge.

The goal of KYOTO is a system that allows people in communities to define the meaning of their words and terms in a shared Wiki platform so that it becomes anchored across languages and cultures but also so that a computer can use this knowledge to detect knowledge and facts in text. Whereas the current Wikipedia uses free text to share knowledge, KYOTO will represent this knowledge so that a computer can understand it too. For example, the notion of environmental *footprint* will become defined in the same way in all these languages but also in such a way that the computer knows what information is necessary to calculate a *footprint*. With these definitions it will be possible to find information on footprints in documents, websites and reports so that users can directly ask the computer for actual information in their environment, e.g. what is the footprint of their town, their region or their company.

The KYOTO system works in 6 steps, as shown in Figure-1:

1. People from a domain specify the locations of diverse and distributed sources of knowledge in different languages. They can do this through a Semantic Wiki environment called Wikyplanet.
2. The text in various languages is captured from the sources and offered to the KYOTO system
3. Term yielding robots (so-called **Tybots**) automatically extract all the important terms and possible semantic relations and relate these to existing semantic networks (so-called **Wordnets**) in each language.
4. The wiki-environment (so-called **Wikyoto**) allows the domain people to maintain the terms and concepts and agree on their meaning within the community and across languages. The meanings are formalized in a domain **ontology** which can be used by computer programs.
5. Knowledge yielding robots (so-called **Kybots**), use the terms and knowledge to detect factual data in the text in various languages.
6. The factual data is indexed and can be accessed by anybody through semantic search, again in various languages, e.g. *facts on CO2 emission in Europe from 2000 to 2009*.

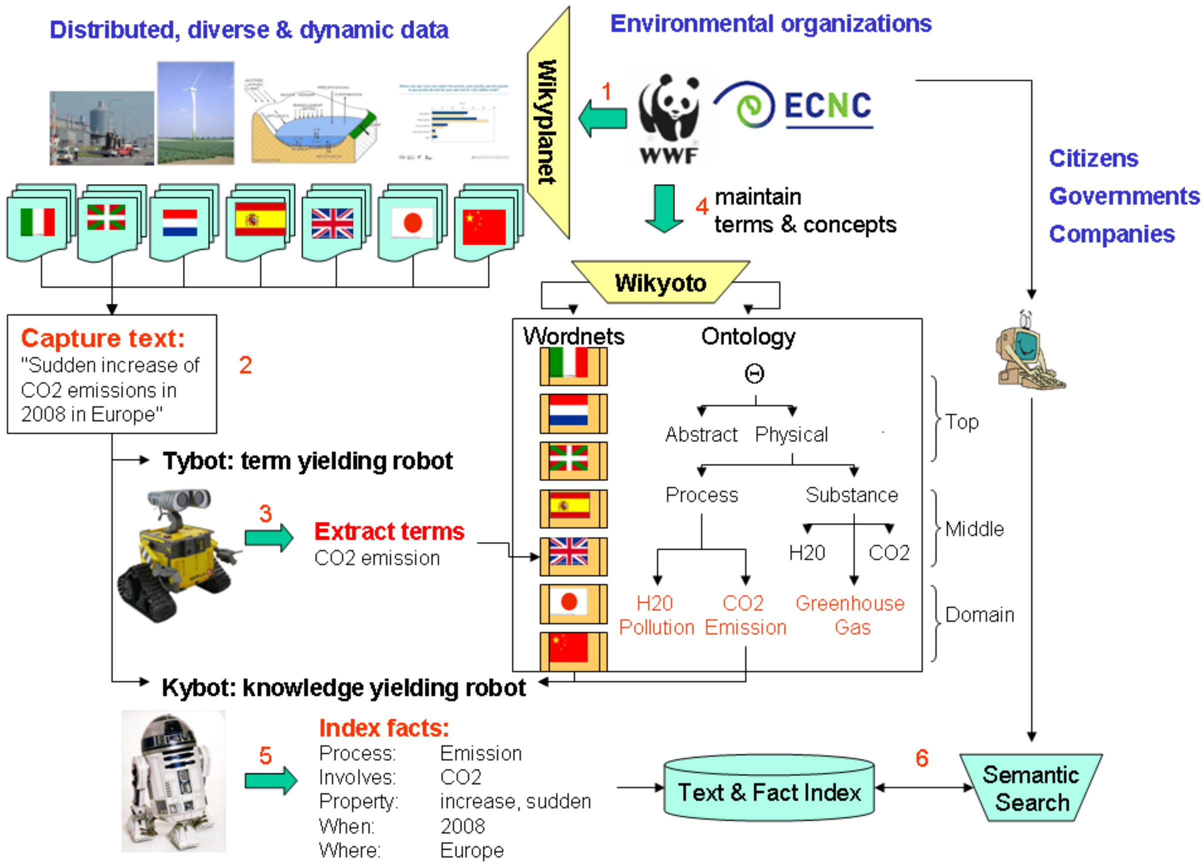


Figure-1: Overall system overview

# 2. Summary of Activities

At the end of the second year of the project, we released a first version of the integrated system. This system can apply all the analysis steps from text to facts. Different installations have been built for the languages in KYOTO that process text in the same way. Using this system, we built a database of facts for 4,625 English documents on estuaries. In the third and final year of the project we worked on evaluation and further improving the system. Four types of evaluations have been carried out:

– Environmentalist used the Wikyoto system for building a domain wordnet from the terms that were extracted from the estuary database;
– We organized and participated in the SemEval2010 task for domain-specific word-sense-disambiguation;
– We developed an evaluation frame work for evaluating the facts mined in KYOTO. This was used to evaluate the KYOTO extraction against a manually created gold-standard .
– We developed an end-user evaluation framework for comparing searches in the KYOTO output with a baseline system and a mash-up system. We organized evaluation sessions with 16 students and another one with 10 specialists to evaluate the systems;

In addition to the evaluation, we worked on improving the knowledge resources that are used for processing. In particularly, we extended and restructured the KYOTO ontology, which is the formal background for interpreting semantic relations expressed in text. Furthermore, we developed heuristics for extending the mapping relations from the English Wordnet to the ontology. Currently, all the synsets in the English WordNet have been mapped to the ontology, guaranteeing large coverage of concepts. In combination with the option to automatically derive a term database for any domain, this means that the KYOTO system can now be applied to any domain. The WordNet to ontology mapping system also generated many new role relations. These are important to interpret the roles of entities as expressed in text. Again, this adds expressive power to KYOTO as a generic system

The extraction of events is done using so-called Kybot profiles. Kybot profiles are XML files that specify patterns in the analyzed text. These patterns combine textual, structural and conceptual elements in the text. The latter are implications inserted through the above WordNet to ontology mappings and the extended ontology. We developed over 250 profiles for English to capture most of the generic relations expressed in the text. Again, these profiles are not specific to the domain and can be used to extract facts from any domain. We made a selection of 70 profiles that gave the best results on a benchmark document. From the database with 4,625 documents (3,091,842 words in size) on estuaries, the patterns yielded almost 1 million information triplets: 118,255 events with 245,563 involved participants, 317,749 dates, 271,734 place relations and 64,604 mappings to countries. The dates and places are entities mapped to ISO dates and GeoNames locations. The mappings from the text yielded 5,075 unique locations and 1,587 dates that anchor the extracted events. The following relations to participants were extracted:

| Relation | Nr. participants | Relation | Nr. participants | Relation | Nr. participants |
|---|---|---|---|---|---|
| destination-of | 11,033 | part-of | 2,464 | source-of | 5,185 |
| done-by | 37,096 | patient | 131,662 | state-of | 2,575 |
| generic-location | 15,883 | purpose-of | 8,570 | use-of | 2,093 |
| has-state | 5,278 | simple-cause-of | 23,724 | | |

The overall system and the corresponding resources represent a major system for mining facts from text in any domain. The generic system can be tuned to a domain using the automatically learned term database which is automatically linked to the generic system. In addition, users of the system can build a domain specific wordnet that represents the specific concepts that matter for them from the term database. The Wikyoto system has been improved during this year to make this process very easy for domain specialists that have no training in linguistics, language engineering and knowledge engineering. Furthermore, we extended Wikyoto with the option to map concepts in the domain wordnet to the ontology using simple interviews. Likewise, domain specialists can add the important semantic implications for the selected concepts, which are stored in a formal representation, without having to know or see these formal structures.

Through Wikyoto, the end-users in KYOTO, who are environment specialists, built a domain wordnet that can be exploited for extracting facts in their domain. The domain-wordnet adds more and more-precise implications to the processed text, which will make the above generic patterns and resources more effective for the domain. Likewise, Wikyoto allows domain-customization of the complex KYOTO processing by domain specialists not familiar with the technology.

The output of the massively mined facts can be searched by end-users in the semantic search system. This system generates structured tables for facts that express important causal relations and relate these to the places and dates in which they occur. These structured tables provide efficient knowledge and information which is mapped to time-lines to show trends or to google-maps to show regional coherence. The semantic search system was evaluated by students and domain specialists. The results of the evaluation will be available shortly after this report.

The results of KYOTO will be represented in the final workshop that will be held in Gifu (Japan) in January 2011. The workshop is co-sponsored by Toyohashi University of Technology (TUT), Japan (http://www.tut.ac.jp/english/) and The National Institute of Information and Communications Technology (NICT), Japan (http://www.nict.go.jp/index.html). The workshop lasts four days and it includes invited speakers from Industry: Google, Yahoo and the European Research Council, and several environment specialists. We will represent the evaluation results and the showcases for KYOTO at the workshop and we invited presentations from other groups.

Finally, we developed a number of tools for inspecting and accessing the knowledge that is generated by KYOTO. We developed graphical browsers that can show the network structures of the term database and show the extracted facts. Figure-2 shows a screen dump of the fact graph extracted from a single document on the Chesapeake Bay estuary in the US. The central node here is pollution. It is connected to many other concepts through colored edges, representing different relations expressed in facts. Each edge represents a unique fact in time and place. Concepts related to the same fact are connected separately through grey edges. Figure-3 shows a graph representation of terms in the term database extracted for the estuary documents, related to the term “tree”. Grey boxes are new terms, blue boxes are terms that also occur in the generic WordNet as concepts and red boxes are generic WordNet concepts that do not occur in the documents but group other terms that do.

We also developed an sql database for storing the texts annotated in the Kyoto Annotation Format (KAF) with a graphical program to inspect the data. Through the interface, developers can inspect the KAF elements and structures, run queries, and get overview data. Figure-4 shows a screen dump of the interface to the SQL database.

*Figure-2: Graph representation of facts extracted from a single document*

*Figure-3: Graph representation of terms related to "tree"*

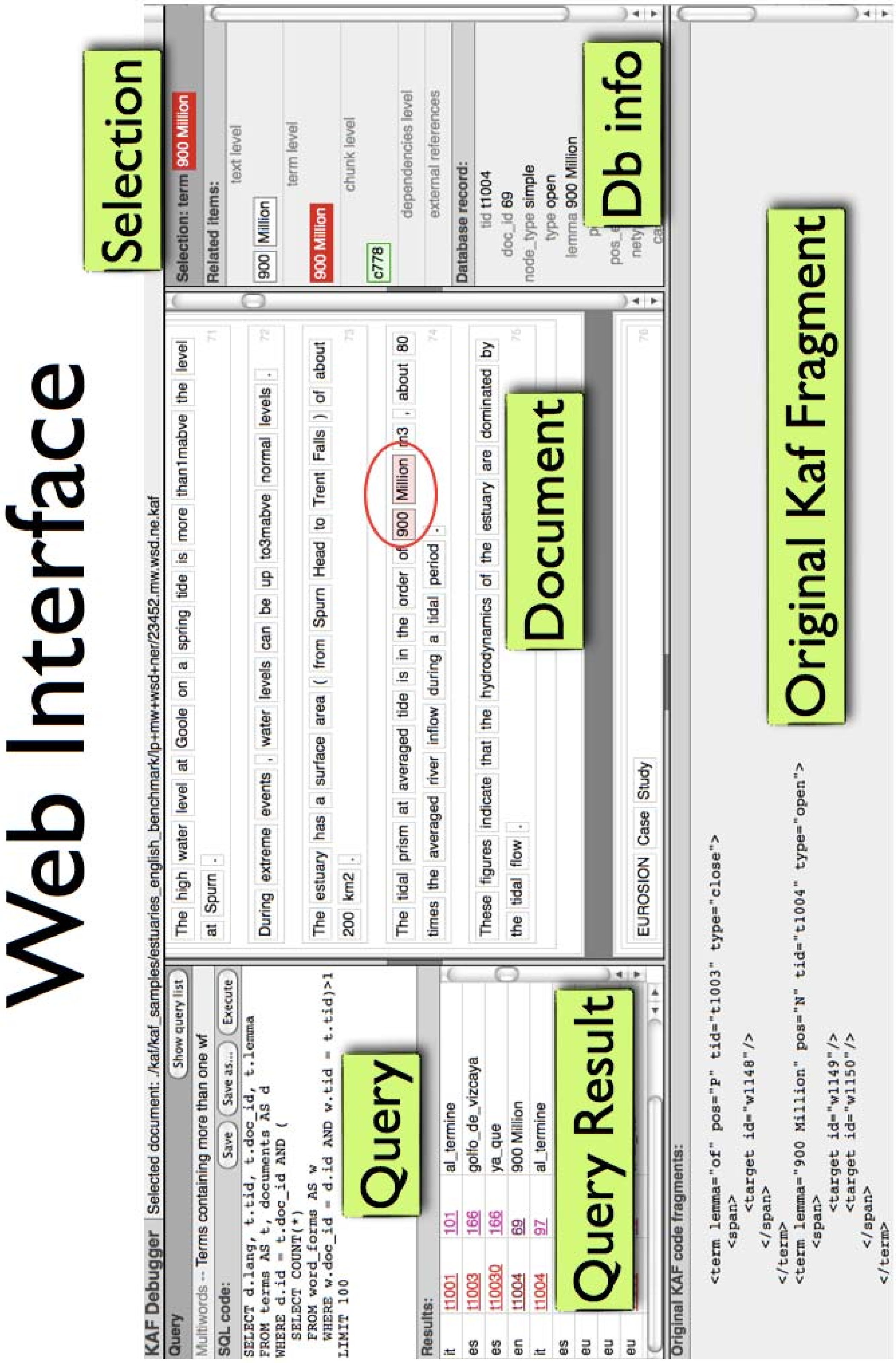


*Figure-4: Screen dump of the web interface on the KAF SQL database*

# 3. Benchmark and development

The ultimate goal of KYOTO is to mine facts from natural language text. KYOTO is an open platform that can be used to model and mine any kind of knowledge that is expressed in text. Therefore, a general representation model of text has been defined that can handle any textual structure, the so-called KYOTO Annotation Format (KAF). Also the representation of the mined facts is defined in KAF. Any evaluation of the mined facts in KYOTO thus needs to consider the way facts are presented. This makes it difficult to compare the results in KYOTO with output from other systems. We therefore developed an evaluation framework that uses a more neutral representation of factual implications. The evaluation framework consists of:

1. A triplet representation for representing factual implication in a neutral way;
2. An annotation tool that can be used to annotate facts in textual document represented in KAF and generate a triplet representation of the annotation;
3. A conversion program that converts the facts that are automatically mined by KYOTO to the triplet representation;
4. An evaluation program that can compare two triplet representations of facts from the same textual source and calculate the precision and recall;

The annotation tool can be used to create a gold-standard for any text that is represented in KAF. The minimal structure for the text-representation is a list of the word tokens that make-up the text. This evaluation scheme has been used to organize an event evaluation task in the KYOTO workshop (see below). We also used the scheme for benchmarking and evaluating our system.

For the system development, we used a gold-standard created for a single document:

www.acb-online.org\pubs\Bay Barometer 2008 Web.pdf

The document is about the Chesapeake Bay, a large estuary in the US. It has 16,145 word tokens of which 1,416 tokens have been annotated. These annotations represent 388 triplets and 204 unique events in total which are divided over the following relations:

| Relations: | Nr. | Proportion |
|---|---|---|
| TIME | 28 | 7,00% |
| destination-of | 27 | 6,00% |
| use-of | 4 | 1,00% |
| generic-location | 11 | 2,00% |
| source-of | 4 | 1,00% |
| instrument | 2 | 0,00% |
| product-of | 2 | 0,00% |
| part-of | 1 | 0,00% |
| purpose-of | 7 | 1,00% |
| LOCATION | 7 | 1,00% |
| patient | 133 | 34,00% |
| path-of | 1 | 0,00% |
| result-of | 4 | 1,00% |
| has-state | 32 | 8,00% |
| state-of | 22 | 5,00% |
| done-by | 52 | 13,00% |
| simple-cause-of | 51 | 13,00% |
| Total | 388 | |

The patient relation is most frequent (34%), followed by done-by and simple-cause-of (13% each).

As a baseline, we create a triplet for all the heads of constituents in a single sentence according to the constituent representation of the text in KAF. The baseline generates 120,134 triplets for the complete file of which 3,671 triplets are in the scope of the gold-standard triplets, i.e. at least one of the identifiers in the triplet overlaps with the identifiers of the gold-standard. The results for the baseline are shown below. Since there is no relation predicted, we can only evaluate by ignoring the relation label.

| baseline chunks in same sentence (3671 triplets) | Ignoring relation |
|---|---|
| Nr. correct | 350 |
| Precision | **0.10** |
| Recall | 0.90 |

The recall of this method is 90%, which shows that the baseline of heads of constituents matches very well the way that the triplets have been annotated. It also shows that 10% are missed, which is probably due to errors in the parsing of the text through which the wrong word is chosen as the head of the constituent (see more below). Precision is very low, even when we ignore the relation itself. This is mainly due to the fact that the baseline generates triplets that relate word tokens that are not related in the gold-standard. Note that the gold-standard only relates participants to events, whereas the baseline also relates participants to each other.

In our first evaluation of the KYOTO system, we created about 250 generic KYBOT profiles. The profiles are formulated to capture the above relations in general and do not use any specific domain concepts. Nevertheless, the profiles have been optimized to represent the relations in the gold-standard for the above document. The profiles have been created on the basis of very simple sequences of terms with part-of-speech, exhibiting basic syntactic patterns, but restricted to certain ontological classes. An example of such a profile is given below:

```
<?xml version="1.0" encoding="utf-8"?>
<Kybot id="done-by-patient">
 <variables>
   <var name="v1" type="term" pos="N" reftype="SubClassOf" reference="DOLCE-Lite.owl#physical-object | DOLCE-Lite.owl#amount-of-matter/>
   <var name="v2" type="term" pos="V" lemma="have | will | can | do | must | shall"/>
   <var name="v3" type="term" pos = "V" reftype="SubClassOf" reference="DOLCE-Lite.owl#accomplishment"/>
   <var name="v4" type="term" pos="N" reftype="SubClassOf" reference="DOLCE-Lite.owl#physical-object | DOLCE-Lite.owl#amount-of-matter”/>
 </variables>

<relations>
  <root span="v4"/>
   <rel span="v1" pivot="v2" direction="preceding" immediate="true"/>
   <rel span="v2" pivot="v3" direction="preceding" immediate="true"/>
   <rel span="v3" pivot="v4" direction="preceding"/>
</relations>

 <events>
   <event eid="" target="$v3/@tid" lemma="$v3/@lemma" pos="$v3/@pos"/>
   <role rid="" event="" target="$v1/@tid" lemma="$v1/@lemma" pos="$v1/@pos" rtype="done-by"/>
   <role rid="" event="" target="$v4/@tid" lemma="$v4/@lemma" pos="$v4/@pos" rtype="patient"/>
 </events>
</Kybot>
```

This pattern lists 4 variables with the part-of-speech N, V, V and N. The variables are defined at the term level and are restricted to certain classes in the ontology (where alternatives are separated by “|”), or in the case of v2 to certain verb lemmas. In the relations element, the profile indicates which element should precede which other element, either immediately or with intersecting elements. The pattern captures a sequence of a noun referring to a concrete object or substance, followed by one of the verbs, followed by another verb that refers to an accomplishment, finally followed by a noun referring to a concrete object or substance. Finally, the events element defines the output structure for the profile. This profile will generate an event matching the third variable v3 and two roles for the first and fourth variable. This profile generates output such as the following:


```
<event eid="e1881" target="t8694" lemma="support" pos="V" synset="eng-30-00908621-v" rank="0.0868965"/>
<role rid="r4226" event="e1881" target="t8692mw" lemma="water quality" pos="N" rtype="done-by" synset="" rank=""/>
<role rid="r4227" event="e1881" target="t8695" lemma="growth" pos="N" rtype="patient" synset="eng-30-14234074-n"
rank="0.140469"/>
```


Using the patterns, we extracted the triplets in two rounds. First we apply all the profiles to the KAF representation of the evaluation document. This generates 34,776 triplets for the complete document of which 3,100 triplets are in the scope of the gold-standard triplets. We evaluated these triplets against the above gold-standard. Besides the precision and recall for the system output, it also generates the precision for each profile. These precision figures are used to post-process the Kybot output by selecting the most-likely roles if there is a choice, or to limit the triplets to those generated by profiles with a higher score. The selection of the best role heuristics will compare triplets that involve the same events and participants but propose a different relation. It will then select the relation proposed by the profile with the highest score. The next table shows the results for the KYBOT output as is and the output after selecting the best role and using only profiles (85 out of 250) above a thresholds of 25%.

| kybot_gen7d_11767.xml (3100 triplets) | Kybot output as is (90 profiles) | | Best scoring role output, profiles with 25% confidence |
|---|---|---|---|
| 388 gold standard triplets | Ignoring relation | Matching relation | Matching relation |
| Nr. correct | 216 | 158 | 126 |
| Precision | 0.48 | 0.29 | 0.48 |
| Recall | 0.56 | 0.41 | 0.32 |
| F-measure (2.(P.R/P+R)) | 0.51 | 0.34 | 0.39 |

For the Kybot output as is, we show two columns. The first when matching partial identifiers and ignoring the relation and the second matching partial identifiers and matching the relations. The results ignoring the relation can be compared with the baseline. We see that the recall is much lower (56% compared to the baseline 90%) but the precision is much higher (48% compared to the baseline of 10%). If we consider the relation, we see that the recall drops to 41% and precision even to 29%. The last column shows the result for post-selecting highest scoring relations and using highest scoring profiles. With optimal settings, we now obtain a precision of 48%, which is equal to ignoring the relation, and a recall of still 32%. This gives the highest F-measure. We also analyzed the recall of the system for each relation. This showed that the profiles capture almost all structures that occur. Missed relations are caused by errors in parsing the text.

# 4. Knowledge structures

In the third year of the project, we extended the knowledge structure of KYOTO in two ways:

1. We provided a firm basis for the full coverage of all generic concepts by extending the ontology and creating mappings from all English WordNet concepts to the ontology;
2. We built a domain wordnet starting from an automatically built term database using the database of 4,625 documents on estuaries. This domain wordnet is linked to the generic English WordNet and through that linked to the ontology;

## 4.1 The KYOTO ontology

KYOTO should be able to accommodate changes in scientific theories as both the world and our knowledge of the world change. We, therefore, require an ontology that is not idiosyncratic but rather one that can accommodate (1) a variety of languages and their wordnets, (2) a variety of scientific domains other than ecology, (3) a variety of research communities, (4) future research in these domains, and can (4) serve as the basis of sound, formal reasoning. This means that the major role of the ontology in the KYOTO project is to provide a coherent, unified, stable frame of reference for different cultural and linguistic communities as well as different research communities. In order to do this, the ontology will deal with very general and abstract concepts in a clear, consistent manner that follows rigorous and explicit criteria.

The KYOTO ontology has three distinct levels and is based on the DOLCE-Lite-Plus (DLP). DLP is a top-level or upper ontology, which means that it models highly abstract concepts such as objects, events, and qualities. In a previous stage of the project we modified DLP to create a top-level ontology suitable for KYOTO. The middle level contains concepts that connect domain specific terms to concepts in the upper ontology. This includes:

- Base Concepts
- Units of measurement, and other qualities
- Perdurant terms (processes and states)
  - often more general than domain terms
  - often identified with a synset from WordNet 3.0,
  - not exhaustive of terms that would appear in a domain neutral ontology.
- Endurant terms (objects and substances)
- Regions and qualities that model adjectives

There are a few concepts in the middle level ontology that are slanted towards the ecology domain, e.g., “things-organisms-do”. These are present in the middle level ontology for two reasons: 1) we do not want the users to be able to edit these concepts. 2) a small number of domain specific concepts can serve as illustrations or templates for modeling choices.

The least abstract level of the ontology is the domain specific level. This contains terms and concepts that are pertinent to ecology. KYOTO-3 contains domain specific concepts that correspond to nouns, verbs, and adjectives found in the literature for ecology. Most of these terms were extracted from user documents or provided from the users themselves. Some other terms were added in the process of developing the ontology for the sake of enriching relations amongst concepts.

The KYOTO-1 Core ontology focussed mainly on mapping endurants. At the ontological level, we focussed on the distinction between types and roles. We extended DLP with Base Concepts taken from the noun portion of WordNet and likewise, the domain relevant terms were endurants: namely, species and regions.

The work done for extending KYOTO-1 to produce KYOTO-2 focussed on modelling processes and events (“perdurants”, in DOLCE lingo) as well as qualities. In this stage of ontology development, we have taken a bottom up approach, focusing on the concepts that have been identified by the users to be key concepts pertaining to estuaries. With the aim of making modelling choices that are both ontologically precise as well as relevant to domain and application concerns, we have focused on adding concepts that serve as a bridge between the domain specific concepts that ecologists have identified as key and the highly general concepts in DLP. While portions of the resulting ontology are specific to the domain of ecology, other portions are still general enough to be domain neutral. In this phase of ontology development we have separated the ontology into a top, middle and domain level in order to isolate those parts of the ontology that are too domain specific to be of general use.

The work to extend the KYOTO-2 ontology to produce KYOTO-3 focused on enriching the way we model processes, events and qualities. This phase has three distinct parts. The first is to determine how to model antonyms amongst the adjectives that are represented in the ontology. The second is to determine how to model relations between verbs and adjectives that have been added to the ontology, and the third is to determine how to arrange the verbs and adjectives in a hierarchy. The ontology is freely available from the KYOTO website.

## 4.2 WordNet to ontology mappings

KYOTO released new WordNet to Ontology mappings. This release now includes new tables of the synset to Base Concept mappings (96.328 records), synset to ontology mappings (179.797 records), and explicit ontology mappings (27.983 records). The release is freely available from the Kyoto ftp site.

With this new release, a substantial improvement on the WordNet to Ontology mapping has been carried out. Past versions of the mapping were centered on nouns only. The new version of the mapping now includes new mappings for all WordNet verbs and adjectives. That is, the current mapping provides a complete mapping for all nominal, verbal and adjectival parts of WordNet[1].

As in previous versions we followed a semi-automatic approach to create a new version of the Ontology and to create a new WordNet to Ontology mapping. Firstly, we derived the complete mapping by exploiting the Base Concepts[2] (Izquierdo et al. 2007[3]). This time, we also used the verbal 578 Base Concepts (BCs) which have been obtained by using all relations and having at least 50 descendants. These 578 BCs have been incorporated as new types into the Kyoto ontology below its corresponding WordNet lexicographer's files. The 15 lexicographer's files have been included as classes in the Kyoto ontology as subclasses of perdurant. Secondly, all verbal WordNet concepts have been mapped to their corresponding ontological types. All verbal concepts corresponding to a verbal Base Concept have been

1 22 adjectives have not been mapped.

2 http://adimen.si.ehu.es/web/BLC

3 Izquierdo R., Suárez A. and Rigau G. Exploring the Automatic Selection of Basic Level Concepts. Proceedings of the International Conference on Recent Advances on Natural Language Processing (RANLP'07), Borovetz, Bulgaria. September, 2007.

aligned through the “sc_equivalentOf” relation. The remaining verbal concepts have been mapped as “sc_subclassOf” to the verbal class in the ontology corresponding to the its verbal Base Concept.

We also created additional mappings for those nominal concepts connected by morpho-semantic links to events. We used the WordNet morpho-semantic database[4] to create a set of rule-based heuristics to derive additional connections among nominal concepts from WordNet and event types in the Kyoto ontology. To distinguish among potential candidate connections, these rules use as a background knowledge the EuroWordNet Top Ontology[5] (Álvez et al. 2008[6]).

In order to establish appropriate connections to the Kyoto ontology for adjectives, we also use the Princeton semantic relations to connect adjectives to its more appropriate noun or verb concepts. We map each adjective as “sc_qualityOf” to the type corresponding to the selected noun or verb, and we also map each adjective as “sc_subclassOf” to the "Kyoto#quality-eng-3.0-04723816-n" type.

Together with the KYOTO ontology, the English WordNet mappings provide a extremely rich and powerful basis for semantic processing of text in any domain. Through the equivalence relations of wordnets in other languages to the English WordNet, this semantic framework can also be applied to the other languages. This provides a common framework for semantic processing of text for all the languages.

## 4.3 The domain wordnet

The KYOTO system is a multilingual tool that will be able to differentiate and bind synonyms in seven languages (English, Dutch, Spanish, French, Italian, German and Chinese) in the first stage, which may be expanded to more languages in the future. To be able to complete this complicated task, it is essential to create a wordnet covering the concepts used within the environmental domain. This wordnet will form the basis for the system to make linkages (both) between the concepts in different languages and in the context of the environmental domain, therefore strengthening the capacity of the system provide the users, in this case from the environmental domain, with the optimum search result.

During the first stage of the development of the domain wordnet, the important terms and concepts were modeled for each of the seven languages. Following the completion of what proved to be the huge task of integration, it was up to the two environmental organizations that take part in the project to manually continue with the development and extension of the domain wordnet.

For this purpose, an online tool has been developed which provides the representatives of the environmental domain in the project with a graphical user interface to conduct the manual domain wordnet extension; the KYOTO Knowledge Editor.

4 http://wordnetcode.princeton.edu/standoff-files/morphosemantic-links.xls
5 http://adimen.si.ehu.es/web/WordNet2TCO
6 Álvez J., Atserias J., Carrera J., Climent S., Oliver A., Rigau G. Consistent Annotation of EuroWordNet with the Top Concept Ontology. Proceedings of the 4th Global WordNet Conference (GWC'08), Szeged, Hungary. January, 2008.

*Figure-5: KYOTO Knowledge Editor*

*Figure-6: Definition of the new concept*

In the KYOTO Knowledge Editor, the user can search the indexed semantic resources in the window pane on the right of Figure-5 and, through a drag-and-drop system, add them to the appropriate concept in the domain wordnet, which is displayed in the left window pane. With the drag-and-drop action, the user is asked to specify the relation of the added term to the parent term. There are three options that can be chosen in regard to the relation:

- External relation; which means that the added term or concept does not belong to the edited synset in itself, but represents (part of) another hierarchy which has a relation to the edited hierarchy.
- New child concept; which means that the added term or concept is part of the edited concept.
- New synset; which means that the added term or concept represents a new synset within the domain wordnet in itself.

The development of the KYOTO Knowledge Editor has taken place in close communication with the representatives from the environmental organizations within the project. Through regular consultation with and feedback from the users, it has now been developed into an instrument which provides a streamlined and optimized methodology for the extension of the domain wordnet.

During this year, considerable effort has been put into the further development and extension of the domain wordnet. Since the optimization of the KYOTO Knowledge Editor, progress on the domain extension has been considerable and this process is being continued into the future.

The next table shows the size and coverage of the latest version of the domain wordnet (date December, 22nd, 2010).

| | |
|---|---|
| Words | 1259 |
| Concepts | 3260 |
| Mappings to the terminology database | 410 |
| Mappings to the ontology | 991 |
| Mappings to the generic WordNet | 0 |

The domain wordnet was built in a month time using the Wikyoto system. Through 209 mappings it is connected to the generic WordNet, mostly top nodes in the domain wordnet hierarchy. This provides a classification of the concepts through the ontology as well, since the generic WordNet is fully mapped. In addition, there are about 991 specific mappings from the domain wordnet to the ontology as well.

# 5. Semantic search evaluation

Now that the end of the project is approaching, we evaluated whether we have succeeded in providing better information retrieval tools to environmentalists. In the course of the project we have created three different search systems that can search within the same database:

1. The TwentyOne Conceptual Search System (http://kyoto.irion.nl/Client 3.5 Kyoto ML; described in the first Annual Report): a keyword-based system that uses natural language technology such as fuzzy matching.
2. The Kyoto Semantic Search System (http://kyoto.irion.nl/kyoto-semsearch/; described in the second Annual Report): uses the same search techniques as TwentyOne, but presents the results as facts, including properties, locations & dates found in the context.
3. The Kyoto Kybot Search System (http://kyoto.irion.nl/kyoto-kybot/): analyses the user's query to a set of concepts, then uses both those concepts and the original words from the query to search in the kybot facts, which are indexed by lemma, concept and hypernym. For instance, if someone searches for 'mammals', the system finds the concept 'mammal', then extracts all facts which contain the word 'mammal', have been tagged with the concept 'mammal' (e.g. cetacean), or with a concept which is a hypernym (meaning 'kind of') mammal (e.g. wolf). The facts are ordered by the number and the closeness of matches with the query, and are presented in an interface which is similar to the Semantic Search system. The facts are presented with cause/actor, result/patient, location and date; users can view the facts in a table, map or sorted in groups. Figure-7 shows a screenshot.

We conducted an evaluation study which compares these three systems. So far we have conducted two sessions, with a total of 16 participants, who were all students at the VU University Amsterdam. We are planning another session in early January with researchers at ECNC, which should bring the total number of participants up to 25 or even 30. At this moment, we cannot present the evaluation results before this final session has taken place; but we can describe the first day of the experiment, and give a 'preview' on the feedback we have received so far.

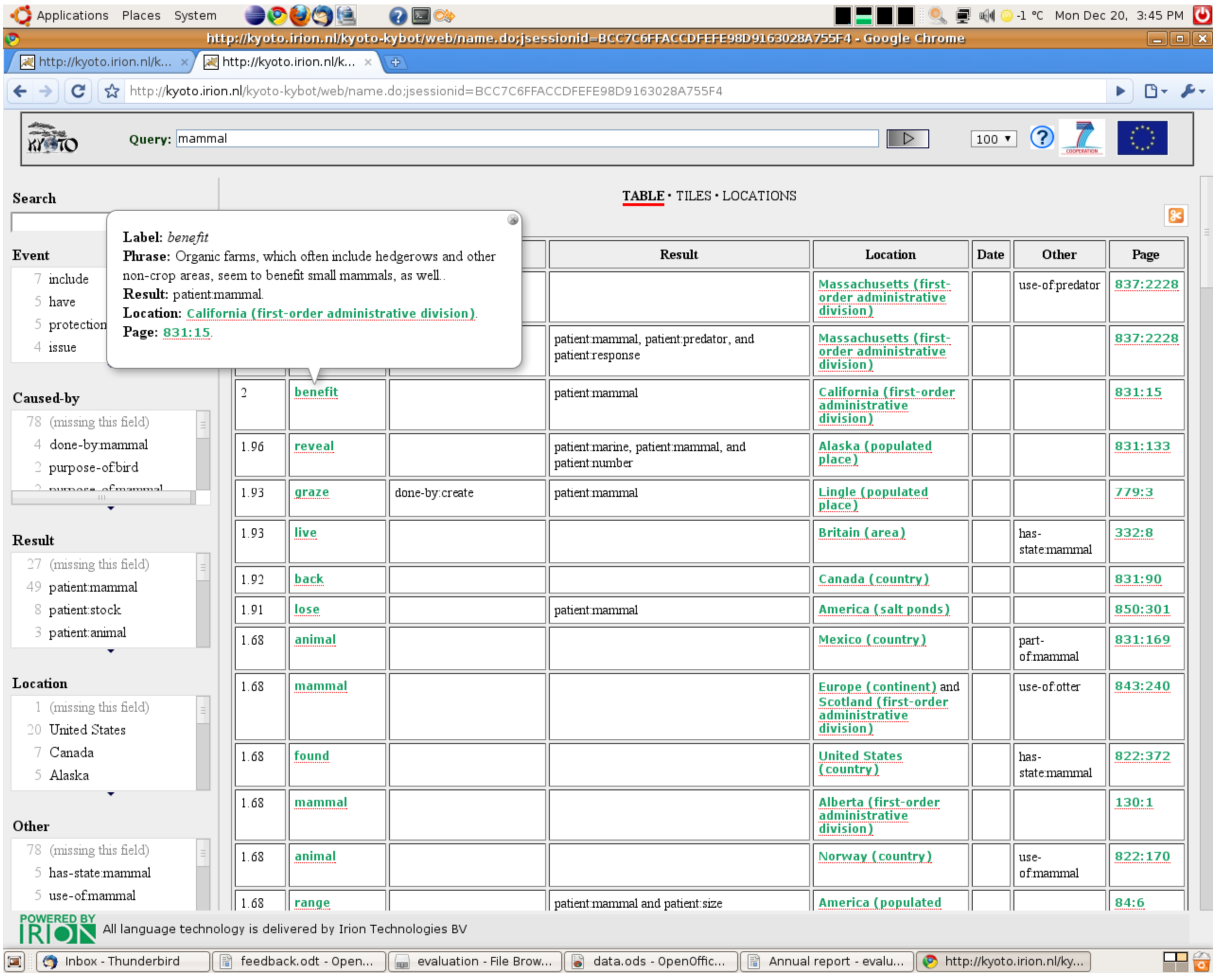


*Figure-7: The Kyoto Conceptual Search System*

## The setting

On the first day of the evaluation (December 15th, 2010) seven students were present. We asked each student to test all three systems by answering a number of questions (which were divided into three different sets). The experiment was thus divided into three one-hour slots. For each system, the student would first receive one page of instructions, so they could familiarize themselves with the system. They could then start the 'logging' tool and load the question set specified by the instructions. The tool would ask six questions, log both the answer given and the time needed to find it, and warn the student after 10 minutes to move on to the next question. When finished, the student were asked to provide feedback in the form of a standard usability questionnaire (the System Usability scale; Brooke, 1996[7]), and also to provide personal comments.

After a brief collective introduction, each student was handed a sheet with separate instructions. To avoid learning effects, or differences in the difficulty of the question sets, students were divided into three groups which received both tools and question sets in different order and different combinations. At the end of the experiment, the students were handed a final questionnaire which asked which systems they liked best and least.

7 Brooke, J. (1996). "SUS: a "quick and dirty" usability scale". In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & A. L. McClelland. *Usability Evaluation in Industry*. London: Taylor and Francis.

**Our hypothesis**

In these experiments, the null hypothesis is always that there will be no differences between the systems as regards the accuracy of the answers, time needed to find them, and the SUS-score (a number between 0 and 100). However, we are naturally hoping to find (our research hypothesis) that the Kyoto Kybot Search system enables users to find more accurate answers more swiftly, and is found to be more user friendly than the other two.

**How it went**

We were impressed by our participants; the students worked individually, silently and with fierce concentration. Each student finished well before the four hours of the experiment ended (one hour per tool, plus another hour for introduction, feedback and breaks). There were a few questions, and some students had trouble fitting their answers in the logging tool (we had not expected them to write such long answers); but on the whole, the experiment went remarkably well.

**Feedback**

The reactions of the students to in particular the Kybot tool were very mixed. Exactly half of the students preferred the TwentyOne tool, and found the Kybot tool overly complicated:

- “I was very confused by this system”
- “The system frequently failed to produce useful results. It was unnecessarily complicated with too many confusing features”
- “The tags (patient, caused-by, purpose-of, etc.) were often assigned to words and terms that made no sense to me”
- “The results look chaotic”
- “Sometimes the answers on my query didn't have anything to do with my search words”

On the other hand, about a third of the students preferred the Kybot system because of the functionality it offered:

- “I liked that system could recognize causal relationships between events”
- “I liked this system best as it allowed me to adapt my search using the facets on the left. It was also possible to enter an entire question, this method mostly worked and provided more specific results”
- “Including causality relationships with different terms used in the papers is a great idea”
- “Being able to focus on certain locations was very helpful”

With such divided opinions we cannot say anything definite about the results of the experiment yet. We are now processing the data to measure the effectiveness of the 3 systems. The results of the evaluation sessions will be presented at the KYOTO workshop in Gifu.

# 6. Second KYOTO Workshop

The final KYOTO workshop[8] will be held in Gifu (Japan). The workshop is co-sponsored by Toyohashi University of Technology[9] (TUT), Japan and The National Institute of Information and Communications Technology[10] (NICT), Japan.

The goal of the workshop is to demonstrate the possibilities of advanced information systems for sharing information and knowledge about the environment on a global scale. The environment is a global concern crossing boundaries, cultures and languages. Whereas each regional setting is unique, there are many benefits from sharing knowledge and experiences across these regions. This workshop offers a showcase for emerging systems that can enable environmentalist to access this knowledge and information in novel ways. It hopes to bring together technology providers and environmental specialists to discuss the opportunities in the field.

The workshop will last four days and it includes invited speakers from Industry and several environment specialists:

- Marius Pasca, Google
- Jordi Atserias and Michael Matthews, Yahoo
- Ralf Steinberger, European Commission - Joint Research Centre
- Bernard Fleming, Environment Specialist
- Jim Ellsworth, environmental expert
- Lawrence Walters-Jones, ECNC
- Karin de Boom, WWF

We will represent the evaluation results and the showcases for KYOTO at the workshop and we invited presentations from other groups. We also organized an event extraction task for the workshop, where other groups can compete with the KYOTO system for mining environmental text. For that purpose, we developed a representation of events with relations to participants that is independent of the KYOTO system. Event-participant relations are presented as simple triplets. We created an annotation tool with which relations can be encoded for any text as a gold-standard. Two groups outside KYOTO participated in the task (AST from Taiwan and KAIST from Korea). KYOTO participated with a generic system that was not adapted to the domain. The results are shown in the next table:

| Overall results from relation table | Nr. triplets | Nr. of triplets in scope | Nr. of relations | Nr. correct | Recall | Precision | F-measure |
|---|---|---|---|---|---|---|---|
| **GS** | **256** | 256 | 253 | | | | |
| **Baseline** | **14902** | 1815 | 1815 | 50 | 0.20 | 0.03 | 0.05 |
| **AST** | **15** | 8 | 8 | 3 | 0.01 | 0.38 | 0.02 |
| **KAIST** | **165** | 62 | 60 | 34 | 0.13 | **0.57** | **0.22** |
| **KYOTO** | **3461** | 964 | 192 | 58 | **0.23** | 0.30 | **0.26** |

The gold-standard (GS) consists of 3 documents for which 15 different semantic relations have been used. The most dominant relations are done-by, patient and has-state. In total 256 triplets have been encoded in the gold-standard for the 3 documents on environmental news. As a baseline, we created a patient relation between any constituent in a sentence, without any

8 http://xmlgroup.iit.cnr.it/kyoto/index.php?option=com_wrapper&view=wrapper&Itemid=152
9 http://www.tut.ac.jp/english/
10 http://www.nict.go.jp/index.html

semantic processing. This generated 14,902 triplets of which 1,815 triplets overlap with the events encoded in the gold-standard (triplets in scope).

The KYOTO system has the highest F-measure (combining recall with precision): 26%. The KAIST system has the highest precision (57%) and KYOTO has the highest recall (23%). KAIST limited itself to only done-by and patient relations using the subject and object relations with verbs. Their system only considers clear structures that can be interpreted with higher precision. KAIST using semantic processing to choose the appropriate role for subjects and objects. KYOTO uses a similar semantic processing but considers all possible structures and relations. Likewise, KYOTO tries to represent all the implications encoded in the text.

The results and implications will be further discussed at the workshop.

# 7. User Involvement, Promotion and Awareness

Two environmental organizations take an active role in the project: ECNC and WWF. These organizations have however many contacts with other organizations in Europe and across the world. They will be able to organize a wide-scale validation and awareness of the system in their community, beyond the scope of the project. The Wikyplanet system is the primary platform for organizing the wider user community.

The consortium also is represented by the Global WordNet Association (GWA: http://www.globalwordnet.org), which is a non-profit organization that stimulates the development and linking of wordnets for as many languages as possible in the world. Through GWA, the KYOTO system can gain global scale and become a platform for anchoring languages and knowledge mining across the world. We proposed an extension to the LMF ISO standard for representing wordnets: WN-LMF. The wordnets for 7 languages in this format are available in a Global Wordnet Grid architecture, making it easy for other language groups to join the KYOTO framework. These wordnets are all linked to the English WordNet (http://wordnet/princeton.edu) and to a series of language neutral ontologies.

We are cooperating closely with a series of other projects as well. The ontological work is coordinated in close cooperation with the developers of major upper ontologies, e.g. DOLCE. This guarantees a solid foundation of the knowledge representation in the project that is ready for the future and not restricted to an ad-hoc project for a single domain. As for standardization of the system and knowledge repositories, we collaborate closely with the projects Clarin (http://www.clarin.eu/), Flarenet (project ECP-2007-LANG-617001) and Lirics (project 22236 – LIRICS, http://lirics.loria.fr/index.html). In the case of the latter, we have adopted the ISO proposals and recommendations for lexical resources (LMF) and terminology representation (TMF). Furthermore, the KYOTO Annotation Format (KAF) has been based on standards for linguistic annotation as proposed in ISO working groups, such as MAF, SYNAF, SEMAF, LAF and GRAF. KAF is an open representation for linguistically processed text that represents the starting point for all further processes in KYOTO. New languages can be added to the system by converting the output of their linguistic processing to the KAF format and providing a wordnet in WN-LMF that is linked to the English Wordnet.

The results of KYOTO are already used in a number of follow-up projects. The DutchSemCor project (funded by NWO) will use the word-sense-disambiguation technology of KYOTO for building a semantically tagged corpus of Dutch. The Semantics of History project (funded by the VU University Amsterdam through the Camera institute) uses KYOTO to develop the terminology and ontology for an historical archive. KYOTO is also participating in a new FP7 project proposal for the development of a platform for clinical cancer treatment in the medical domain. The Euskal Herriko Unibertsitatea in San Sebastian received funding for the project KNOW2: Language understanding technologies for multilingual domain-oriented information access. The project will start early 2010. This project will continue the open lines of predecessor project KNOW by including the outcomes of KYOTO.

The website of the project already has extensive information and demonstrations. It has 45,000 visitors from all over the world. The project has been presented at various conferences and workshops and we created posters and brochure for easy dissemination. All publications, technical papers, project deliverables and presentations can be downloaded from the public website. The website further gives access to demos, videos and tutorials for all major modules and data formats.

# 8. Future Work or Exploitation Prospects

We organized the final workshop to present the final system and the demonstration platform in the environment domain (see above). The workshop will target a wider audience and explore future exploitations and developments of KYOTO.
We will continue to release the integrated system as an open source package that can be downloaded and used by software developers outside the project.

The industrial partners IRION and SYNTHEMA will explore the exploitation of KYOTO in new products and services. The current semantic search system is a first demonstration of such an exploitation. Other systems that we foresee are:

1. a FactAlert application that will alert users for changes in factual information that is published in textual form on a specified list of online sources
2. a Dialogue system through which non-experts users can interactively explore knowledge and facts on the environment.

The Wikyplanet system will be used to embed the KYOTO system in an easy to use community platform. The users in the environment domain can see the direct benefit of KYOTO in relation to the knowledge sources that they specify for specific cases. This can be in the form of semantic search options that are directly integrated in Wikyplanet, e.g. browsing or searching facts related to the uploaded sources. ECNC is project leader of the recently granted project: Skills for local biodiversity (2010-1-NL1-LEO05-02653)[11] The proposal is to use Wikyplanet as the internet platform for the BD SKILLS website. Wikyplanet has a structure that can easily be adapted to the needs of the user.

# 9. Further Information

***Events in connection with KYOTO:***

**Workshops:**

- *1st public KYOTO Workshop, Amsterdam, the Netherlands, February 2-3, 2009*

11 http://www.adam-europe.eu/adam/project/view.htm?prj=6717

- *SemEval2010 task on domain specific word-sense-disambiguation*
- *Global Wordnet Conference 2010, Mumbai India (www.globalwordnet-iitb2010.in)*
- *2nd public KYOTO Workshop, Gifu, Japan, January 25-28, 2011*
- *OntoLex workshop, COLING 2010, Beijing, August 22nd 2010*

**Demos:**

The architecture of KYOTO is shown in the flash animation below. In the schema, the different modules of the KYOTO system are given as rectangular boxes and the data structures by blue repositories. The animation shows two cycles of processes in KYOTO which are explained in more detail below. The cycles start with documents and websites that are provided by the users in the project: ECNC and WWF. We have collected a first set of documents websites in 4 languages that focus on a series of environmental themes. If you click on the logos of the users, you will get a baseline retrieval system for these documents.

### The KYOTO Modules

The KYOTO system has the following modules:

1. **Syntactic processors**: they produce a syntactic and morphological analysis of the text
2. **Semantic processors**: they determine what the meaning is of the words in the text
3. **Tybots**: they learn the terms that are used in the documents and organize these as a hierarchy. If you click on the term extractor module, you can access a demo that gives access to term databases that have been extracted from the environment documents.
4. **Term editor** (part of the Wikyoto platform): users can edit the terms, give definitions and agree on what they mean. These users are called concept users since they are domain experts that maintain the terminology. You can click on the module to go to a demo of the term editor and try it out yourself.
5. **Kybots**: little programs that use the knowledge built up for terms to extract facts from any set of documents. If you click on this module, you will access a demo where you can design or submit a Kybot to extract facts from a sample database
6. **NL Query**: search module with which any end user (people from the domain, government, companies, students, children, etc.) can access the database of facts that is produced. If you click on this module, you can access a demo on semantic search on the English data.

### The KYOTO Repositories

In the architecture there are also 4 databases:

1. **Document base**: a database that holds all the documents after being processed by the syntactic and semantic processors. The text is represented in a special XML format called the KYOTO Annotation Format. If you click on the database, you can see examples of this format in different languages and get access to the DTD.
2. **Term database**: this database holds the output of the term extraction. The terms can be exported into XML in a special format that is called KYOTO-TMF. If you click on the database, you get more details on the term structure in TMF.
3. **Multilingual Knowledge Base**: this database holds the wordnets in all the languages and ontologies that are already given. It holds also any domain wordnet and ontology that is built by editing the term database. If you click on the database, you can view

the databases, which are represented in a special XML format for wordnets (Wordnet-LMF) and for ontologies (OWL).

4. **Fact database**: this is the database in which all the extracted facts are stored. This database still needs to be designed in the project. Further details will follow. Note that the database can also hold changing realities. It can extract a fact at some point in time and another fact related to the same things and same place at another point in time.

### The KYOTO Cycles

Documents and websites are then processed in two cycles, which is shown in the animation:

1. First cycle in which concept users upload specialized documents and sources to acquire a good term database and to enable them to build a good domain wordnet and ontology:
   1. sources are processed syntactically and semantically and the output is stored in the document base as KAF-XML
   2. the Tybots extract the terms and put the terms in the term database
   3. the domain specialists review and modify the terms, define their meaning and agree with the meanings of terms in other languages through the ontology
2. Second cycle in which the same documents or any other set of documents are sent to KYOTO to extract any facts:
   1. sources are processed syntactically and semantically and the output is stored in the document base as KAF-XML (same as in the first cycle)
   2. the Kybots extract the facts that the end-users are interested in and stores the facts in the fact database
   3. End users get alerts on new facts or can search in the database to get comprehensive and precise informations that can be organized in many different ways, e.g. per region or along time lines, to reveal trends and changes.

### The Integrated System: KyotoCore

The website also contains all the modules and documentation for the integrated system, which is free for download. The source code for the modules will be released soon as well. KyotoCore is the central module of KYOTO that is used to *process textual data for generating knowledge*. It consists of a collection of separate modules that read and write text representations in the Kyoto Annotation Format (see the web page for explaining KAF). Each module adds a new layer of analysis to KAF, taking previous layers as input. The KAF representations of text are stored in separate databases that are maintained by the document base. The sequential processing of text into KAF according to a pipeline of modules is controlled by a job-dispatcher. The job-dispatcher checks the status of each document in the databases and applies the next module in a pipeline that is associated to a database. The so-called PipeT module can be used to create any pipeline of modules. The KyotoCore system thus consists of the set of modules built for KYOTO, together with the document base and the job-dispatcher, where we created specific pipelines for the KYOTO process flow. With PipeT it is possible for any developer or system integrator to build any other pipeline.

KyotoCore is embedded in the overall architecture of KYOTO, where documents and websites are collected in WikyPlanet (a Semantic Media Wiki for the environment). The Capture module can collect these sources (or any other set of documents) and will push them into a database in the document-base. The job-dispatcher will then apply the pipeline associated with that database to each document in the database.

A Multilingual Knowledge Base is used by many modules to carry out the semantic processing of text. It contains a central ontology and generic wordnets in various languages, stored in the DebVisDic platform. The Knowledge Base can be extended through the Wikyoto editor for a specific domain or application. Through one of the KYOTO pipelines, we can build a database with terms and relations from any set of documents in KAF. Wikyto can take these concepts and terms as input and combine it with background vocabularies and generic resources in the Multilingual Knowledge base. This will result in a domain wordnet (possibly in multiple languages) that is combined with the generic resources for processing KAF. Another pipeline can then generate factual data out of the KAF on the basis of the domain model and the generic resources. In the next figure, we give a detailed overview of the KyotoCore architecture and its embedding in the complete KYOTO platform:

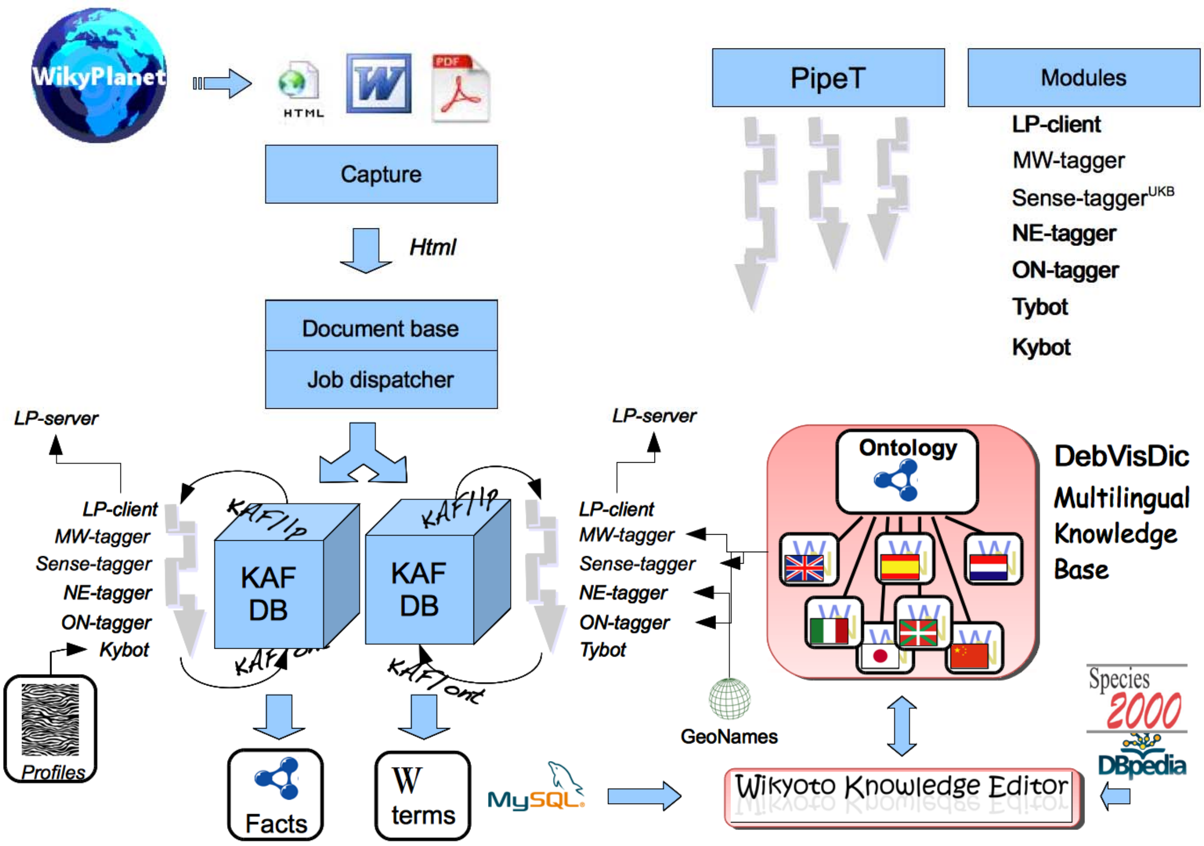


*Figure-8: The integrated system KyotoCore*

The KyotoCore system includes the following components:

1. Capture module for collecting and converting textual sources and storing them in a specified database in the KYOTO document-base;
2. The KYOTO document base, which maintains, documents, databases, and users with rights (as a usual DMS) but which also assigns pipelines of processing modules to databases;
3. The KYOTO job-dispatcher that continuously monitors the documents in databases, checks the status of the documents and tries to perform the next step in the pipeline for each document;
4. The file system for storing the KAF files that are produced and processed by the modules in KYOTO;
5. The Kyoto modules, which are combined in a pipeline architecture to produce KAF, a term databases and facts:

1. Linguistic processors (LP), which are client programs that send HTML to a LP server which returns KAF representations including tokenization, lemmatized term representation, chunks and dependencies; Servers are available for Dutch, Spanish, Basque, English, Italian. For Chinese and Japanese applications are available and server versions are being deployed.
2. Multiword (MW) tagger, which reads KAF and groups sequences of terms as multiword terms on the basis of the multiwords in generic wordnets and domain wordnets;
3. Sense tagger (UKB), which is a word-sense-disambiguation system that uses a graph of semantic relations (based on wordnets) and a personalized page-rank algorithm to detect the synsets of words in context; UKB reads KAF and generates KAF with synsets added to the term layer;
4. Named-entity (NE) tagger: which detects time points and places in KAF as named-entities. It applies named-entity disambiguation and represents the named-entities in a separate layer in KAF with GeoNames properties and Wordnet mappings for locations;
5. Ontotology (ON) tagger: reads the synsets in KAF and inserts the full set of ontological implications that apply into the term structure of KAF, where the ontological implications are drawn from the central ontology and through synset to ontology mappings;
6. Tybot: reads KAF and extracts the terms and their relations using structural, distributional and pattern-based rules. The results are stored in a MySql database that is input for the Wikyoto system for editing the domain wordnet;
7. Kybot: reads KAF and a specified set of profiles to extract events and facts from KAF, where the profiles can specify patterns at any level of KAF (wordform, terms, synsets, ontological implications, named-entities, etc.).

6. A term database in MySql with new terms that are learned from KAF representations of documents;
7. PipeT: a platform for creating pipelines of processing modules through input and outputstream connections;

The overview figure also shows the WikyPlanet module for feeding KyotoCore with textual sources and the Wikyoto system for editing the domain wordnets and the central ontology. The latter are stored in the Multilingual Knowledge Base that is implemented in the DebVisDic environment. WikyPlanet, Wikyoto and DebVisDic are external to KyotoCore. It is possible to install and run KyotoCore without these external modules. For example, documents can be captured directly through the Capture module from a list of URLs. Furthermore, although many modules in KyotoCore use external resources such as wordnets and an ontology, these can be provided as separate data files independently of the Multilingual Knowledge Base. Wikyoto and the Multilingual Knowledge Base are only needed to build the domain specific resources. Through tailoring the knowledge to a specific domain, the quality of the output of KyotoCore will be improved (both in recall and precision). The cyclic nature of KYOTO allows you to continuously improve the ontology and domain wordnet which directly feeds back into all the modules that are applied after the LPs. As shown in this overview, Wikyoto connects to large background databases such as Species2000 and DBPedia, which contain millions of concepts and terms. Any domain wordnet can thus be aligned to wordnets, the central ontology and to the background documents and it can draw from any of these for creating concepts in addition to the term database. Nevertheless, KyotoCore can also run using generic resources.

At the heart of KyotoCore lays the document-base and the job-dispatcher. The document-base keeps track of databases, users and documents. Uses get rights to work on specific databases, or to create/delete databases. Each database can contain any number of documents and the document-base keeps track of the administration, versioning and status of each document in the database. To process documents in KYOTO, a registered user needs to add a document to a database. Another important function of the document-base is the administration of the modules that should be applied to each database. Users can register which modules and pipelines of modules should run on which database. In the case of KYOTO, there are a number of standard pipelines of modules that are applied. The first step in the process is the processing of the text in any HTML file in the database to generate the KAF representation of the text with a structural analysis: tokenization, term detections, chunking and dependency structures. Once the text is represented in KAF, other KYOTO modules will apply and add further data (layers) to the KAF representation.

The processing of the documents in the databases is controlled by the job-dispatcher. The job-dispatcher permanently monitors the status of each file in the document-base and checks what module should be applied next to the file, given the modules and pipelines that are associated with the database. The KYOTO engine is thus started by pushing documents into the document-base. Adding documents to the document-base can be done directly through the document-base API or through the Capture module, which uses the document-base API. The Capture module basically pre-processes textual sources to prepare them for processing in KYOTO. For example, the KYOTO pipeline starts with the LPs and the LPs take HTML files as input streams and produce KAF/LP as output streams (as is specified in the pipeline configuration). Likewise, the Capture module crawls HTML pages from websites and converts PDF documents to HTML and adds all the HTML files (and if necessary the PDF source) to a database in the document-base. Note that it is also possible to generate KAF outside the KYOTO platform (e.g. using your own LP for your language) and add KAF representations of your document to the document-base directly, bypassing both Capture and LP. This will also kick-start the KYOTO processing but since the status of the document is already set to KAF/LP, the next step in the pipeline is immediately applied, i.e. multiword tagging.

An important aspect of this architecture is that any pipeline can be created from any set of modules using the PipeT system and that all processing is centred around KAF as an input and output streams. This makes the system very flexible and easily extendible. Furthermore, after the LPs have produced a structural representation of the text in their language, the remaining modules in KYOTO are language-neutral, except for some basic and simple patterns and the wordnets that are used for each language. Note that currently the linguistic processors (LPs) are servers that are hosted at different sites, while the KYOTO LP modules are clients that simply pass HTML files to the server and send the KAF back to the job-dispatcher for storage in the document-base. Parsers that generate KAF can be obtained for most KYOTO languages. Otherwise, developers can include their own parser in the pipeline or simply directly submit KAF to the document-base, which will start the further processing. Likewise, there is a minimal language-dependency for processing text in KYOTO.

Given the total set of modules that are developed for KYOTO, shown at the top-right side, we can create any set of pipelines reading and writing KAF back into the database. In the overview figure, we represented two different databases where each database has a different pipeline associated. The pipeline database at the right side ends with a Tybot that generates the mysql term database for the KAF documents in the database. The term database is used to build the domain wordnet in the Wikyoto editor and map the wordnet to the central ontology. The domain modelling in the multilingual knowledge base can be exported to data files that are used by the modules in the pipeline again. For example, the multiword tagging can use multiwords from the generic wordnet in combination with the multiwords from any domain wordnet. Similarly, WSD can add the relations from a domain wordnet to the relations from the generic wordnet.

The second cycle, shown for the database at the left side, ends with the Kybot that generates the final output for KYOTO. Typically, we expect domain experts first to do domain modelling for a selection of representative documents after which they can apply the domain model to another database for extracting information and knowledge (facts) from a larger database.

**Links to the demos:**

- Baseline Search
- Multilingual Knowledge Repository
- Terminology extraction: TYBOT
- Wikyoto Term Editor
- Wikyoto Kybot Editor (Fact extractor)
- Semantic Search: Fact Retrieval System

**PR:**

1. PR Brochure on KYOTO
2. *KYOTO participates in the Global WordNet Grid (*http://www.globalwordnet.org/gwa/gwa_grid.htm*) of the Global WordNet Association (*www.globalwordnet.org*).*
3. *Opening of a technology forum and an environment forum to open up discussion to a wider public at* www.kyoto-project.eu.
4. A0-poster on KYOTO, presented at LREC, Marrakech, Morocco, May 28-30, 2008.
5. KYOTO used as a promotion/demonstration of present projects combining language and ICT on the "European Day of Languages" for the Representation of the EU in Rome. "EU corner" at the Night of Researchers' exhibition, September 25, 2008
6. Factsheet on Tybots and Kybots in Wikyoto

**Presentations/Invited talks/Workshops:**

Francis Bond: invited speaker on "Cross-lingual Word Sense Disambiguation using WordNet", International Conference on Asian Language Processing 2010, Harbin, China, Dec 28-30, 2010.

Francis Bond: "New developments in the Japanese WordNet", Asian WordNet Meeting, Dec 6-9 2010, Phuket, Thailand.

Kyoko Kanzaki: "Enriching the Adjective domain in the Japanese WordNet", Asian WordNet Meeting, Dec 6-9 2010, Phuket, Thailand.

Isahara and Sornlertlamvanich: Organizer of the Asian WordNet Meeting (closed meeting attended by participants from Bhutan, Mongolia, Myanmar, Nepal, Pakistan, Singapore, Thailand, India, Indonesia, Japan, Korea, Philippines, SriLanka, Vietnam), Dec 6-9 2010, Phuket, Thailand.

Carlo Aliprandi: presentation on "Kyoto and industrial applications of NLP technologies " at the workshop: "WS3: Human Language Technologies for Italian",
11th AI*IA Symposium on Artificial Intelligence (AIIA 2010), Brescia (Italy), December 1-3, 2010.

Piek Vossen: Invited speaker on "Kyoto: multilinguale terminologie op basis van Wordnets" at 2nd Day of "Terminologie in het Nederlandse Taalgebied" (TiNT Day 2010), December 3, 2010, VU University Amsterdam.

Monica Monachini: invited talk on "Semantic Search Systems: applicative experiences". TAL Conference (under the high patronage of the Italian Republic President), October 28-29, 2010, Rome, Italy.

Piek Vossen: Invited speaker on "KYOTO: a platform for anchoring textual meaning across languages" at W3C workshop "The Multilingual Web - Where Are We?", October 26-27, 2010, Madrid, Spain.

Piek Vossen: Invited speaker on "Semantics of History" at the CATCH meeting "Finding, Linking and Organizing Sources for Media Research", organized by "Bridge", October 22, 2010, Boijmans van Beuningen, Rotterdam, the Netherlands.

Piek Vossen: Invited speaker on "KYOTO: Open knowledge and text mining environment" at Xerox Seminar, September 16, 2010, Geneva, Switzerland.

Christiane Fellbaum: Invited speaker on "KYOTO: Knowledge for Everyone ", at 13th International Conference on Text, Speech and Dialogue (TSD2010), Brno, Czech Republic, September 6-10, 2010.

Piek Vossen e.a.: Co-organizer of OntoLex 2010: 6th Workshop on "Ontologies and Lexical Resources", COLING, August 21-22, 2010, Beijing, China.

Piek Vossen: presentation on "KYOTO: an Open Platform for Mining Facts", on OntoLex 2010: 6th Workshop on "Ontologies and Lexical Resources", COLING, August 21-22, 2010, Beijing, China.

Agirre e.a.: Organizer of SemEval-2010: 5th International Workshop on Semantic Evaluations on Kyoto's subtask WSD17: All-words Word Sense Disambiguation on a Specific Domain, workshop collocation: ACL2010, July 11-16, Uppsala, Sweden.

Piek Vossen: invited speaker on KYOTO: A Community Platform for Knowledge Modeling and Text Mining" at the NWO Catch-SIKS-TiCC symposium Text Mining in the Real World", Tilburg, June 30, 2010.

Piek Vossen: Invited speaker on "KYOTO: Open knowledge and text mining environment", Linguist Day at Q-go, Amsterdam, June 11, 2010.

Piek Vossen: Invited speaker at forum discussion on "Bijdrage van ICT aan bronnenonderzoek in humaniora", VU University Amsterdam, June 9, 2010.

Piek Vossen: Invited speaker on "World of Words" at the Minisymposium "Caught in your own web", VU University Amsterdam, the Netherlands, March 18, 2010. 'Caught in your own web' provides a first glimpse into the new and fascinating field of complex networks. Participants will be exposed to a multilayered perspective on complex networks from disciplines spanning from computer science to neurology and text analysis to sociology.

Piek Vossen: Lecture Course Masters on the Computational Lexicon with reference to Kyoto, Faculty of Sciences (Artifical Intelligence) and Faculty of Arts, VU University Amsterdam, February - June, 2010.

Savas: Presentation on "A LMF-based RESTful Web Service for Accessing WordNet-type Semantic Lexicons", in: Proceedings of the 7th international conference on Language Resources and Evaluation (LREC2010), Malta, May 17-23, 2010.

Bosma: Presentation on "Bootstrapping language neutral term extraction", in: Proceedings of the 7th international conference on Language Resources and Evaluation, (LREC2010), Malta, May 17-23, 2010.

Rigau: Presentation on "Integrating a large domain ontology of species into WordNet", in: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC2010), Malta, May 17-23, 2010.

Segers: Presentation on "Facilitating Non-expert Users of the KYOTO Platform: the TMEKO Editing Protocol for Synset to Ontology Mappings", in: Proceedings of the 7th international conference on Language Resources and Evaluation (LREC2010), Malta, May 17-23, 2010.

Piek Vossen: Invited speaker on "Kyoto" at the Semantic Web Meeting, VUA Amsterdam, the Netherlands, February 15, 2010.

Piek Vossen: Invited speaker on "How to keep distributed efforts for developing semantic vocabularies together?", in: S4 - Social Networking and Web 2.0 Methods for Language Resources, at the FLaReNet Forum 2010, February 11-12, 2010, Barcelona, Spain.

Piek Vossen/Christiane Fellbaum organized 5th Global WordNet Conference, Mumbai, India, January 31-February 4, 2010

Vossen: Presentation on the "Division of semantic labor in the Global Wordnet Grid", in Proceedings of the 5th Global WordNet Conference (GWC2010), Mumbai, India, January 31-February 4, 2010, Narosa Publishing House (cover by Sam Vossen), ISBN 979-81-8487.

Rigau: presentation on "Semantic Services In Freeling 2.1: WordNet And UKB", in: Proceedings of the 5th Global WordNet Conference, (GWC2010), Mumbai, India. January 31- February 4, 2010.

Laparra: Presentation on "Exploring the integration of WordNet and FrameNet", in: Proceedings of the 5th Global WordNet Conference (GWC2010), Mumbai, India. January 31- February 4, 2010.

Ronzano F., Maurizio Tesconi, Salvatore Minutoli, Andrea Marchetti: Presentation on "Collaborative management of KYOTO Multilingual Knowledge Base: the Wikyoto Knowledge Editor", in: Proceedings of the 5th Global WordNet Conference (GWC2010), Mumbai, India. January 31- February 4, 2010.

Vossen: Presentation on "Wordnets as intermediate between lexicons and ontologies" at the 5th Joint ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation, in conjunction with the Second International Conference on Global Interoperability for Language Resources (ICGL 10), Hong Kong, January 15-17, 2010.

Vossen: Presentation on "A full knowledge cycle for semantic interoperability" in: Proceedings of the 5th Joint ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation, in conjunction with the Second International Conference on Global Interoperability for Language Resources (ICGL 10), Hong Kong, January 15-17, 2010.

Marchetti A., Salvatore Minutoli, Francesco Ronzano and Maurizio Tesconi: Presentation on "KAFnotator: a Multilingual Semantic Text Annotation Tool" at the 5th Joint ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation, in conjunction with the Second International Conference on Global Interoperability for Language Resources (ICGL 10), Hong Kong, January 15-17, 2010.

Del Gratta : presentation on the paper of Toral A., M. Monachini, C. Soria, M. Cuadros, G. Rigau, W. Bosma, P. Vossen: presentation on "Linking a domain thesaurus to WordNet and conversion to WordNet-LMF", in: Proceedings of the Second International Conference on Global Interoperability for Language Resources, (ICGL 2010) Hong Kong, January 15-17, 2010.

D'Onofrio: presentation on "A Web-based Architecture for interoperability of lexical" in: Proceedings of the Second International Conference on Global Interoperability for Language Resources, (ICGL 2010) Hong Kong, January 15-17, 2010.

Marchetti A., Salvatore Minutoli, Francesco Ronzano and Maurizio Tesconi: Poster on "A Wiki Framework for cooperatively building multilingual language resources", in: Proceedings of the Second International Conference on Global Interoperability for Language Resources, (ICGL 2010) Hong Kong, January 15-17, 2010.

Marchetti, A., S. Minutoli, F. Ronzano, M. Tesconi: Presentation on "WIKYOTO Knowledge Editor: The collaborative web environment to manage KYOTO Multilingual Knowledge Base", in: Proceedings of the ICKM2009: the 6th International Conference on Knowledge Management, Hong Kong, China, December 3-4, 2009.

Piek Vossen: invited keynote speaker: Presentation on "Kyoto" at eLexicography in the 21st century: new challenges, new applications, Louvain-la-Neuve, Belgium, October 22-24, 2009

Piek Vossen: Invited Speaker on "Division of semantic labor over vocabulary and ontology layers", FLaReNet/SILT Workshop on Lexicon-Ontology Relationship in connection with GL 2009 - 5th International Conference on Generative Approaches to the Lexicon, Pisa, Italy, September 17-19, 2009.

Piek Vossen: Presentation on "KAF: a generic semantic annotation format" at the 5th International Conference on Generative Approaches to the Lexicon (GL2009), September 17-19, 2009, Pisa, Italy

Toral: Presentation on "Studying the role of Qualia Relations for Word Sense Disambiguation" at the 5th International Conference on Generative Approaches to the Lexicon (GL2009), September 17-19, 2009, Pisa, Italy

Piek Vossen: Keynote speaker: Presentation on "Kyoto" at the 25th Annual Conference of the Spanish Society for Natural Language Processing 2009 (SEPLN´09), San Sebastian, Spain, September 8-10, 2009.

Rigau: Poster on Kyoto the 25th Annual Conference of the Spanish Society for Natural Language Processing 2009 (SEPLN´09), San Sebastian, Spain, September 8-10, 2009.

Rigau: Presentation on Kyoto the 25th Annual Conference of the Spanish Society for Natural Language Processing 2009 (SEPLN´09), San Sebastian, Spain, September 8-10, 2009.

Vossen: Lecture Course Masters (College Wordnets, College Ontologies, Colllege Kyoto): Theme Language Acquisition & Documentation: Multilingual Lexical Framing, MA, Faculty of Arts, VU University Amsterdam, February 2 - June 10, 2009.

Nicoletta Calzolari: invited speaker at the PAROLE Consortium Workshop: “New horizons for Linguistic Resources in a Global Context” , Barcelona 7-8 July 2009, presented Kyoto in her talk “From PAROLE to FLaReNet and beyond”.

Monica Monachini: invited speaker at the PAROLE Consortium Workshop “New horizons for Linguistic Resources in a Global Context”, Barcelona 7-8 July 2009, gave a talk on Standards for Lexixal Resources and presented the Kyoto WordNet-LMF and the Kyoto WordNet Grid as a model for a future PAROLE Grid.

Piek Vossen: Invited speaker: Presentation on Kyoto at EU-Japan ICT Cooperation Forum on ICT, Brussels, Belgium, July 2, 2009.

Nicoletta Calzolari presented Kyoto during the ISO TC47 SC4 Plenary Meeting held in Boulder (Colorado), June 2009.

Nicoletta Calzolari presented Kyoto during the Institutional Evaluation of CNR-ILC research activities by an International Panel of experts of the field (10 June 2009).

Piek Vossen: Invited speaker on "Kyoto" at the EU Information and Networking Event Info-day of promising projects for the 5th Call (ICT FP7 SO 4.3) on Intelligent Information Management, Luxemburg, May 12, 2009

Nicoletta Calzolari: Presentation on Kyoto at the PhD courses at the Pisa University, Italy, May 4, 2009.

Nicoletta Calzolari: Invited speaker and Presentation on Kyoto in “Language Resources: from local initiatives to priorities and challenges in the International scenario” at the Endinburg University, Edinburg, April 17, 2009.

Nicoletta Calzolari: Presentation of the Kyoto architecture as an exemplary instantiation of interoperability of language resources and tools, Athens, CLARIN Meeting, April 4-5, 2009.

Nicoletta Calzolari: Presentation of the Kyoto Annotation Format used for NLP analysis, as an instantiation of an ISO-conformant format for Language Resources to be used as input for the CLARIN infrastructure, CLARIN meeting, Barcelona, April 4-5, 2009.

Piek Vossen: invited keynote speaker and Presentation on "Kyoto" at the Asian Language Resource Summit (ALRS), Phuket, Thailand, March 20-21, 2009.

Piek Vossen: invited guest Lecture "From WordNet to Global WordNet for language technology", Course Language Engineering Applications, Center for Computational Linguistics, University of Leuven, Belgium, February 26, 2009.

Claudia Soria: Presentation on Kyoto at IWIC 2009, Stanford, United States of America, February 20-21, 2009.

German Rigau: Brief summary on the Kyoto-project at III JORNADAS PLN-TIMM, Modelos y técnicas para el acceso a la información multilingüe y multimodal en la web, Colmenarejo, Spain. February 4-5, 2009.

Bosma: Poster on Kyoto at CAMeRA@VU, VU University, January 30, 2009, Amsterdam

Wauter Bosma: Presentation on Kyoto: November 10, 2008, University of Twente, the Netherlands

Piek Vossen: invited keynote speaker on Kyoto at Lustrum of NL TERM, October 25, 2008, Amsterdam, the Netherlands

Piek Vossen: invited keynote speake on Kyoto at DART 2008: 2nd Workshop on Distributed Agent-based Retrieval Tools, 10 September 2008, Cagliari, Italy

Brochure on Kyoto

Vossen: Presentation on KYOTO: "A System for Mining, Structuring and Distributing Knowledge Across Languages and Cultures" at LREC 2008, May 28-30, 2009, Marrakech, Morocco.

Chu-Ren Huang: Invited speaker on the KYOTO Project: "Introducing a multilingual view towards global infrastructure in semantic computing and knowledge engineering". "Semantic Computing" Panel. 2008 AI Forum, Taipei, Taiwan, May 17-18, 2008

Vossen: Presentation on Kyoto at the 4th Global WordNet Conference (GWC2008), Szeged, Hungary, January 22-25, 2008.

Piek Vossen: invited guest Lecture on Corpusgebaseerd tekstonderzoek for the ICT Onderwijscentrum of the Vrije University Amsterdam, December 13, 2007.

## Publications:

Aliprandi C., Ronzano F., Marchetti A., Tesconi M., Minutoli S.: "Extracting events from Wikipedia as RDF triples linked to widespread Semantic Web Datasets", to appear in: Proceedings of HCI International 2011, 14th International Conference on Human-Computer Interaction, 9-14 July 2011, Florida, USA.

Vossen P.: "WordNet: principles, developments and applications", in: Handbook of Linguistics and Communications (HSK), volume Dictionaries, an International Encyclopedia of Lexicography, Supplementary volume: Recent developments with special focus on computational lexicography. Eds. R.H. Gouws, U. Heid, W. Schweickard, H.E. Wiegand, Mouton de Gruyter, Berlin, Germany.

Vossen P., E. Agirre, F. Bond, W. Bosma, C. Fellbaum, A. Hicks, S. Hsieh, H. Isahara, Ch. Huang, K. Kanzaki, A. Marchetti, G. Rigau, F. Ronzano, R. Segers, M. Tesconi
"KYOTO: a Wiki for Establishing Semantic Interoperability for Knowledge Sharing across Languages and Cultures", in: Handbook of Research on Culturally-Aware Information Technology: Perspectives and Models. Ed. Dr. E. Blanchard (Mc Gill University (Canada)) and Dr. D. Allard (Dalhousie University), Publ. IGI Global USA, p. 265-294, ISBN 978161520883.

Horák A., P. Vossen P., A. Rambousek
"A Distributed Database System for Developing Ontological and Lexical Resources in Harmony", chapter in Lecture Notes in Computer Science, 2010, Volume 4919,Computational Linguistics and Intelligent Text Processing, Pages 1-15, Ed. A. Gelbukh, ISBN 978-3-540-78135-6_1,

Oltramari A., P. Vossen, Q. Lu (Eds.)
Proceedings of the OntoLex 2010: 6th Workshop on "Ontologies and Lexical Resources", COLING, August 21-22, 2010, Beijing, p. 1-10, Ed. A. Oltramari, P. Vossen & Q. Lu, Publ. Coling 2010 Organizing Committee.

Vossen P., G. Rigau, E. Agirre, A. Soroa, M. Monachini, R. Bartilini
"KYOTO: an Open Platform for Mining Facts", in: Proceedings of OntoLex 2010: 6th Workshop on "Ontologies and Lexical Resources", COLING, August 21-22, 2010, Beijing, China, p. 1-10, Ed. A. Oltramari, P. Vossen & Q. Lu, Publ. Coling 2010 Organizing Committee.

Fellbaum Ch. & Piek Vossen: "Challenges for a Multilingual Wordnet", in: Lexical Resources and Evaluation.

Barrón-Cedeño A., Rosso P., Agirre E., Labaka. G.: "Plagiarism Detection across Distant Language Pairs" in: COLING, August 21-22, 2010, Beijing, China, pages 37–45, Ed. A. Oltramari, P. Vossen & Q. Lu, Publ. Coling 2010 Organizing Committee.

Soroa A., E. Agirre, O. López de Lacalle, W. Bosma, P. Vossen, M. Monachini, J. Lo, S. Kai Hsieh: "Kyoto: An Integrated System for Specific Domain WSD", in: Proceedings of SemEval-2010: 5th International Workshop on Semantic Evaluations on Kyoto's subtask WSD17: All-words Word Sense Disambiguation on a Specific Domain, workshop collocation: ACL2010, July 11-16, 2010, Uppsala, Sweden, p. 75-80, Ed. K. Erk & C. Strapparava, Publ. The Association for Computational Linguistics (ACL), ISBN 978-1-932432-70-1.

Agirre E., O. Lopez de Lacalle1, Ch.Fellbaum, S. Hsieh, M. Tesconi4, M. Monachini, P. Vossen, R. Segers: "SemEval-2010 Task 17: All-words Word Sense Disambiguation on a Specific Domain", in: Proceedings of SemEval-2010: 5th International Workshop on Semantic Evaluations on Kyoto's subtask WSD17: All-words Word Sense Disambiguation on a Specific Domain, workshop collocation: ACL2010, July 11-16, 2010, Uppsala, Sweden, p. 123-128, Ed. K. Erk & C. Strapparava,, Publ. The Association for Computational Linguistics (ACL), ISBN 978-1-932432-70-1.

Izquierdo R., Suárez A. and Rigau G. GPLSI-IXA: "Using Semantic Classes to Acquire Monosemous Training Examples from Domain Texts", in: in: Proceedings of SemEval-2010: 5th International Workshop on Semantic Evaluations on Kyoto's subtask WSD17: All-words Word Sense Disambiguation on a Specific Domain, workshop collocation: ACL2010, July 11-16, 2010, Uppsala, Sweden, Ed. K. Erk & C. Strapparava,, Publ. The Association for Computational Linguistics (ACL), ISBN 978-1-932432-70-1.

Zapirain B., Agirre E., Màrquez L., Surdeanu M.: "Improving Semantic Role Classification with Selectional Preferences", in: Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010), Los Angeles, California.

Calzolari, N., Monachini, M., Quochi, V., Soria, C., Toral, A. (fc): "Lexicons, Terminologies, Ontologies: "Reflections from Experiences in Resource Construction". To be published in N. Dershowitz and E. Nissan (eds.): Volume in honor of Yaakov Choueka, LNAI Festschrift Volume of the Lecture Notes in Computer Science series, Springer, Berlin.

Hicks, A., Herold, A. (fc.): "Cross-lingual Evaluation of Ontologies with Rudify", in: Knowledge Discovery, Knowledge Engineering and Knowledge Management, "Communications in Computer and Information Science" series, ed. Ana Fred, Jan L. G. Dietz, Kecheng Liu and Jaoquim Fillipe. Springer-Verlag GmbH Berlin Heidelberg.

Alvez J., Paqui Lucio, German Rigau: "Providing First-Order Reasoning Support to Large and Complex Ontologies", in: the Journal of Web Semantics (also Kyoto: Technical Report KYOTO T007/WP6).

Segers R., P. Vossen: "Facilitating Non-expert Users of the KYOTO Platform: the TMEKO Editing Protocol for Synset to Ontology Mappings", in: Proceedings of the 7th international conference on Language Resources and Evaluation (LREC2010), Malta, May 17-23, 2010.

Vossen P., W. Bosma, M. Cuadros, G. Rigau: "Integrating a large domain ontology of species into WordNet", in: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC2010), Malta, May 17-23, 2010.

Agirre E., Cuadros M., Rigau G. and Soroa A. Exploring Knowledge Bases for similarity. Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC2010), Malta, May 17-23, 2010.

Bosma W., P. Vossen: "Bootstrapping language neutral term extraction", in: Proceedings of the 7th international conference on Language Resources and Evaluation, (LREC2010), Malta, May 17-23, 2010.

Savas B., Hayashi Y., Monachini M., Soria C., Calzolari N.: "A LMF-based RESTful Web Service for Accessing WordNet-type Semantic Lexicons", in: Proceedings of the 7th international conference on Language Resources and Evaluation (LREC2010), Malta, May 17-23, 2010.

Bond F., H. Isahara, K. Uchimoto, T. Kuribayashi, K. Kanzaki: "Japanese WordNet 1.0", (website) in: Proceedings of the 16th Annual Conference of The Japanese Society for Natural Language Processing (NLP2010), Tokyo, Japan, March 2010.

Lluís Padró, Samuel Reese, Eneko Agirre and Aitor Soroa: "Semantic Services In Freeling 2.1: WordNet And UKB", in: Proceedings of the 5th Global WordNet Conference, (GWC2010), Mumbai, India. January 31- February 4, 2010.

Laparra E., Rigau G. and Cuadros M.: "Exploring the integration of WordNet and FrameNet", in: Proceedings of the 5th Global WordNet Conference (GWC2010), Mumbai, India. January 31- February 4, 2010.

Vossen P. ,G .Rigau: "Division of semantic labor in the Global Wordnet Grid", in Proceedings of the 5th Global WordNet Conference ((GWC2010), Mumbai, India. January 31- February 4, 2010.

Toral, A.Monachini, M., Soria, C., Bracale, S.: "Rejuvenating the ItalianWordNet: upgrading, standardising, extending", in: Proceedings of the 5th Global WordNet Conference (GWC2010), Mumbai, India. January 31- February 4, 2010.

Ronzano F., Maurizio Tesconi, Salvatore Minutoli, Andrea Marchetti: "Collaborative management of KYOTO Multilingual Knowledge Base: the Wikyoto Knowledge Editor", in: Proceedings of the 5th Global WordNet Conference (GWC2010), Mumbai, India. January 31- February 4, 2010.

Kuroda K., F. Bond, K. Torisawa: "Why Wikipedia needs to make Friends with WordNet" in: Proceedings of the 5th Global WordNet Conference ((GWC2010), Mumbai, India. January 31- February 4, 2010.

Maks I., Vossen P.: "Modeling Attitude, Polarity and Subjectivity in Wordnet", in Proceedings of the 5th Global WordNet Conference ((GWC2010), Mumbai, India. January 31- February 4, 2010.

Horak A., A. Rambousek: "Using DEB Services for Knowledge Representation within the KYOTO Project", in: Proceedings of the 5th Global WordNet Conference ((GWC2010), Mumbai, India. January 31- February 4, 2010.

D'Onofrio L., Del Gratta R., Toral A., Monachini M., Soria C., Bartolini R., Quochi V.,N. Calzolari: "A Web-based Architecture for interoperability of lexical" in: Proceedings of the 5th Joint ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation, in conjunction with the Second International Conference on Global Interoperability for Language Resources, (ICGL 2010) Hong Kong, January 15-17, 2010.

Toral A., Monachini M., Soria C., Cuadros M., Rigau G., Bosma W. and Vossen P.: "Linking a domain thesaurus to WordNet and conversion to WordNet-LMF", in: Proceedings of the Second International Conference on Global Interoperability for Language Resources, (ICGL 2010) Hong Kong, January 15-17, 2010.

Vossen P., W. Bosma, E. Agirre, G. Rigau, A. Soroa: "A full Knowledge Cycle for Semantic Interoperability", in: Proceedings of the 5th Joint ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation, in conjunction with the Second International Conference on Global Interoperability for Language Resources, (ICGL 2010) Hong Kong, January 15-17, 2010.

Marchetti A., Salvatore Minutoli, Francesco Ronzano and Maurizio Tesconi: "A Wiki Framework for cooperatively building multilingual language resources", in: Proceedings of the Second International Conference on Global Interoperability for Language Resources, (ICGL 2010) Hong Kong, January 15-17, 2010.

Tesconi M., Francesco Ronzano, Salvatore Minutoli, Carlo Aliprandi and Andrea Marchetti: "KAFnotator: a multilingual semantic text annotation tool", in: Proceedings of the 5th Joint ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation, in conjunction with the Second International Conference on Global Interoperability for Language Resources, (ICGL 2010) Hong Kong, January 15-17, 2010.

Vossen P., Fellbaum C.: "Universals and Idiosyncracies in Multilingual WordNets",in: Multilingual FrameNets in Computational Lexicography, Methods and Applications, p. 319-346, Ed. H.C. Boas, Mouton de Gruyter, Berlin, 2009, ISBN 978-3-11-021296-9.

Marchetti A., S. Minutoli, F. Ronzano and M. Tesconi : "Wikyoto Knowledge Editor: The Collaborative Web Environment To Manage Kyoto Multilingual Knowledge Base", in: Proceedings of the ICKM2009: the 6th International Conference on Knowledge Management, Hong Kong, China, December 3-4, 2009.

Volkhart Wildermuth: Wo steckt der Salamander im Netz? (Report on the KYOTO project.) Yearbook of the Berlin-Brandenburg Academy of Sciences 2009/2010, 16-21.

Herold, A., Hicks, A: "Evaluating Ontologies with Rudify", in: Proceedings of Knowledge International Conference on Knowledge Engineering and Ontology Development, Madeira, Portugal, October 6-8, 2009.

Agirre E, Arantxa Otegi, Hugo Zaragoza, "Using Semantic Relatedness and Word Sense Disambiguation for (CL)IR", in: Working Notes of the Cross-Lingual Evaluation Forum, (CLEF 2009), Corfu, Greece, September 30-October 2, 2009.

Agirre E., Giorgio Maria Di Nunzio, Thomas Mandl and Arantxa Otegi: "CLEF 2009 Ad Hoc Track Overview: Robust-WSD Task, in: Working Notes of the Cross-Lingual Evaluation Forum, (CLEF 2009), Corfu, Greece, September 30-October 2, 2009.

Wauter Bosma, Piek Vossen, Aitor Soroa, German Rigau, Maurizio Tesconi, Andrea Marchetti, Monica Monachini and Carlo Aliprandi: "KAF: a generic semantic annotation format", in Proceedings of the 5th International Conference on Generative Approaches to the Lexicon GL 2009, Pisa, Italy, September 17-19, 2009

Toral A., M. Monachini, A. Soroa, G. Rigau: "Studying the role of Qualia Relations for Word Sense Disambiguation", in: Proceedings of GL 09, Pisa, Italy, September 17-19, 2009.

Toral A., O. Ferrández, E. Agirre, R. Muñoz: "A study on Linking and disambiguating Wikipedia categories to Wordnet using text similarity", in: Proceedings of RANLP 2009, Borovets, Bulgaria, September 14-16, 2009.

Egoitz Laparra and German Rigau: Integrating WordNet and FrameNet using a knowledge-based Word Sense Disambiguation algorithm. In: Proceedings of RANLP 2009, Borovets, Bulgaria, September 14-16, 2009.

W. Bosma: "Contextual salience in query-based summarization":, in Proceedings of RANLP 2009, Borovets, Bulgaria, September 14-16, 2009

Casillas A,, Arantza Diaz de Illarraza, Kike Fernandez, Koldo Gojenola, Egoitz Laparra, German Rigau, Aitor Soroa: "the Kyoto Project", in: Proceedings of the 25th Annual Conference of the Spanish Society for Natural Language Processing 2009 (SEPLN´09), San Sebastian, Spain, September 8-10, 2009.

Eric Yeh, Daniel Ramage, Chris Manning, Eneko Agirre and Aitor Soroa: "WikiWalk: Random walks on Wikipedia for Semantic Relatedness", in: Proceedings of the workshop "TextGraphs-4: Graph-based Methods for Natural Language Processing", Singapore, August 7, 2009

Yamada I., K. Torisawa, J. Kazama, K. Kuroda, M. Murata, S. de Saeger, F. Bond, A. Sumida: "Hypernym discovery based on distributional similarity and hierarchical structures", in: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2009), Singapore, August 6-7, 2009.

Bond F., H. Isahara, S. Fujita, K. Uchimoto, T. Kuribayashi, K. Kanzaki: "Enhancing the Japanese WordNet" in: Proceedings of the 7th Workshop on Asian Language Resources, in conjunction with ACL-IJCNLP2009, Singapore, August 6-7, 2009.

Zapirain B., E. Agirre, L. Màrquez: "Generalizing Over Lexical Features: Selectional Preferences for Semantic Role Classification", in: Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing, ACL-IJCNLP 2009, August 2-7, 2009, Singapore.

Eneko Agirre, Oier Lopez de Lacalle and Aitor Soroa: "Knowledge-based WSD andspecific domains: performing over supervised WSD", in: Proceedings of IJCAI2009, July 14-17, 2009, Pasadena, USA.

Agirre E., O. Lopez de Lacalle, Ch. Fellbaum, A. Marchetti, A. Toral, Vossen P. All-words Word Sense Disambiguation on a Specific Domain (SemEval-2010 Task 17), in: Proceedings of the workshop NAACL HLT 2009 conference, June 4, 2009, Boulder, Colorado, USA

Agirre E., E. Alfonseca, K. Hall, J. Kravalova, M. Pasca, A. Soroa: ""A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches", in: Proceedings of NAACL-HLT 09. Boulder, Colorado, May 31- June 5, 2009.

Agirre E., I. Castellón, L. Padró: "KNOW: Developing large-scale multilingual technologies for language understanding", in: Actas de la Jornada de Seguimiento de Proyectos en Tecnologías Informáticas.

Agirre E., L. Màrquez, R. Wicentowski: Special Issue: Computational Semantic Analysis of Language: SemEval-2007 and Beyond", Language Resources and Evaluation Journal, Volume 43, Number 2, May 14, 2009, pages 97-208. ISSN 1574-020X.

Agirre E., A. Soroa: "Personalizing Page Rank for Word Sense Disambiguation", in: Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL-2009), Athens, Greece, March 30-April 3, 2009.

Agirre E., Lopez de Lacalle: "Supervised Domain Adaption for Supervised WSD Systems", in: Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL-2009). Athens, Greece, March 30-April 3, 2009.

Izquierdo R., A. Suárez, G. Rigau: "An Empirical Study on Class-based Word Sense Disambiguation", in: Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL-2009). Athens, Greece, March 30-April 3, 2009.

Bond F., H. Isahara, K. Uchimoto, T. Kuribayashi, K. Kanzaki: "Extending the Japanese WordNet", in Proceedings of the 15th Annual Conference of The Japanese Society for Natural Language Processing (NLP2009), Tottori, Japan, March 2009.

Yamada I., A. Sumida, J. Kazama, K. Kuroda, M. Murata, F. Bond, K. Torisawa: "Extending a hypernym dictionary using syntactic characteristics, in: Proceedings of the 15th Annual Conference of The Japanese Society for Natural Language Processing (NLP2009), Tottori, Japan, March 2009.

Soria C., M. Monachini, P. Vossen "Wordnet-LMf: fleshing out a standardized format for wordnet interoperability", in: Proceedings of IWIC2009, Stanford, USA, February 20-21, 2009.

Bond F., E. Nichols, D. Scott Appling, M. Paul: "Improving Statistical Machine Translation by Paraphrasing the Training Data, in: Proceedings of International Workshop on Spoken Language Translation(IWSLT-2008), Haiwaii, October 21-22, 2008.

Cuadros M. and Rigau G.: " KnowNet: a proposal for building knowledge bases from the web". First Symposium on Semantics in Systems for Text Processing, STEP'08. Venice, Italy. September 22-24, 2008.

Agirre E., G. M. Di Nunzio, N. Ferro, T. Mandl, C. Peters, "CLEF 2008: Ad Hoc Track Overview 2009", in: Proceedings of the 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008. Also available as Springer Journal ""Evaluating Systems for Multilingual and Multimodal Information Access", Lecture Notes in Computer Science Volume 5706, pp. 118-125. Springer Berlin / Heidelberg. ISSN 0302-9743 ISBN 978-3-642-04446-5.

Cuadros M. and Rigau G.: "Multilingual Evaluation of KnowNet", in: Proceedings of the 24th Annual Meeting of Sociedad Española para el Procesamiento del Lenguaje Natural, SEPLN2008, Madrid, Spain, Procesamiento del Lenguaje Natural. Vol. 41. ISSN: 1135-5948, September 10-12, 2008.

Cuadros, M., G. Rigau. "KnowNet: Building a Large Net of Knowledge from the Web", in: Proceedings of COLING2008, Manchester, 18-22 August, 2008.

Pease A., C. Fellbaum, P. Vossen: "Building the Global WordNet Grid", in: Proceedings of the 18th International Congress of Linguists (CIL18), Seoul, Republic of Korea, July 21-26, 2008.

Vossen P., E. Agirre, N. Calzolari, C. Fellbaum, S. Hsieh, C. Huang, H. Isahara, K. Kanzaki, A. Marchetti, M. Monachini, F. Neri, R. Raffaelli, G. Rigau, M. Tescon (2008). "KYOTO: A system for Mining, Structuring and Distributing Knowledge Across Languages and Cultures", in: Proceedings of LREC 2008, Marrakech, Morocco, May 28-30, 2008

Agirre E. and Soroa A., "Using the Multilingual Central Repository for Graph-Based Word Sense Disambiguation", in: Proceedings of LREC 2008, Marrakech, Morocco, May 28-30, 2008.

Álvez J., Atserias J., Carrera J., Climent S., Laparra E., Oliver A. and Rigau G. "Complete and Consistent Annotation of using the Top Concept Ontology", in: Proceedings of LREC 2008, Marrakech, Morocco, May 28-30, 2008.

Pociello E., Gurrutxaga A., Agirre E., Aldezabal I. and Rigau G. "WNTERM: Enriching the MCR with a terminological dictionary", in: Proceedings of LREC2008, Marrakech, Morocco, May 28-30, 2008.

K. Kanzaki, F. Bond, N. Tomuro, H. Isahara: "Extraction of Attribute Concepts from Japanese Adjectives", in: Proceedings of LREC2008, Marrakech, Morocco, May 28-30, 2008.

Isahara H., F. Bond, K. Uchimoto, M. Utiyama, K. Kanzaki: "Development of Japanese WordNet", in: Proceedings of LREC2008, Marrakech, Morocco, May 28-30, 2008.

Bond F., H. Isahara, K. Kanzaki, K. Uchimoto: "Boot-strapping a WordNet using Multiple Existing WordNets", in: Proceedings of LREC2008, Marrakech, Morocco, May 28-30, 2008.

Bond F., Z. Chang, K. Uchimoto: "Extracting Bilingual Terms from Mainly Monolingual Data", in: Proceedings of the 14th Annual Meeting of The Japanese Association for Natural Language Processing, Tokyo, March 2008.

Vossen, P., E. Agirre, N. Calzolari, C. Fellbaum, S. Hsieh, C. Huang, H. Isahara, K. Kanzaki, A. Marchetti, M. Monachini, F. Neri, R. Raffaelli, G. Rigau, M. Tescon, J. van Gent (2008): " KYOTO: A System for Mining, Structuring, and Distributing Knowledge Across Languages and Cultures ", in: Proceedings of the Fourth International Global Word Net Conference - GWC 2008, Szeged, Hungary, January 22-25, 2008

Fellbaum, C., Vossen, "Challenges for a Global WordNet". In: Online Proceedings of the First International Workshop on Global Interoperability for Language Resources(ICGL 2008), City University of Hongkong, January 8-12, 2008, p. 75-82, online published by Ed. Webster J., N.Ide, A.Chengyu Fang.

Baldwin T. S. Nam Kim, F. Bond, S. Fujita, D. Martinez, T. Tanaka: "MRD-based Word Sense Disambiguation: Further Extending Lesk", in: Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP 2008), Hyderabad, India, January 7-12, 2008.

## Deliverables :

| | | | | |
|---|---|---|---|---|
| D 11.1 | Project Public Website | Final | P | March 2008 |
| D 1.1 | Data types and sources | 1.0 | P | May 2008 |
| D 7.1 | XML Schema for Wordnet and Ontology | 2.0 | P | May 2008 |
| M03 | M03 Quarterly Management Report | Final | R | June 2008 |
| D 1.2 | Knowledge types and questions | 1.0 | P | August 2008 |
| D 2.1 | Database models and data formats | 3.2 | P | August 2008 |
| D 2.2 | Overall system design | 2.0 | P | August 2008 |
| M06 | M06 Quarterly Management Report | Final | R | September 2008 |
| D 9.1 | User-scenarios | | P | October 2008 |
| D 7.2a | Knowledge Base Server API version 1 | 1.0 | P | November 2008 |
| D 3.1 | Capture module | 1.0 | P | November 2008 |
| D 6.1 | Accumulated knowledge | 1.0 | P | November 2008 |
| AR | Annual public report 2008 | Final | P | November 2008 |
| M09 | M09 Quarterly Management Report | Final | R | December 2008 |
| D 1.3 | User Requirements (draft) | 2.0 | P | March 2009 |
| D 11.2 | 1st Kyoto Workshop | Final | P | March 2009 |
| D 7.4a | Wiki Environment for WordNet Editing (draft) | 1.0 | P | March 2009 |
| D 4.1 | Indexing Module (draft) | 1.0 | P | March 2009 |
| D 11.5 | Design of Showcase DVD (draft) | 1.0 | P | March 2009 |

| | | | | |
|---|---|---|---|---|
| D 6.2 | Central Ontology (draft) | 3.0 | P | March 2009 |
| D 11.4 | Early Dissemination Plan & Strategy (draft) | 4.0 | P | March 2009 |
| M12 | M12 Quarterly Management Report | Final | R | March 2009 |
| D 6.3 | Wordnets mapped to Central Ontology (draft) | 2.0 | R | March 2009 |
| D 5.1 | Concept Miners version 1 (draft) | 2.0 | P | March 2009 |
| D 7.3a | Multilingual Wordnet Services (draft) | 2.0 | P | March 2009 |
| D 7.5a | Wiki Environment for Ontology Editing (draft) | 1.0 | P | March 2009 |
| M15 | M15 Quarterly Management Report | 1.0 | R | June 2009 |
| D 5.2a | Fact Miners Version 1 | 1.0 | P | August 2009 |
| D 9.2 | Search Engine Client | 1.0 | P | August 2009 |
| M18 | M18 Quarterly Management Report | 1.0 | R | August 2009 |
| D 6.4 | Automatic Deduction and Inferencing Technques | 1.0 | P | August 2009 |
| D 8.1 | Domain Extension of Central Ontology - version 1 | 1.0 | P | August 2009 |
| D 8.2 | Domain Extension of Wordnets - version 1 | 1.0 | P | August 2009 |
| D 9.3 | Benchmark Test of 1st Indexes | 1.0 | P | September 2009 |
| | Annual Report 2009 | 1.0 | P | November 2009 |
| D 7.2b | Knowledge Base Server API - revised | 1.0 | P | November 2009 |
| D 9.4 | End-user Test of 1st Indexes | 1.0 | P | November 2009 |
| D 7.3b | Multilingual Wordnet Services - revised | 1.0 | P | December 2009 |
| M21 | M21 Quarterly Management Report | 1.0 | R | December 2009 |

**Working Papers:**

7. TR002/WP02 KYOTO LMF WordNet Representation Format, v.04
8. TR006/WP02 Fact Annotation Format for KYOTO
9. TR002/WP02 KYOTO LMF WordNet Representation Format, v.03
10. TR004/WP02 KYOTO: The representation of terms, v.01
11. TR003/WP02 Formalizing Knowledge by Ontologies: OWL and KIF, v.02
12. TR001/WP01 User Scenarios Wikyoto, v.01
13. TR005/WP05 Storyboard: "to mine by example" for building Kybots, v.01