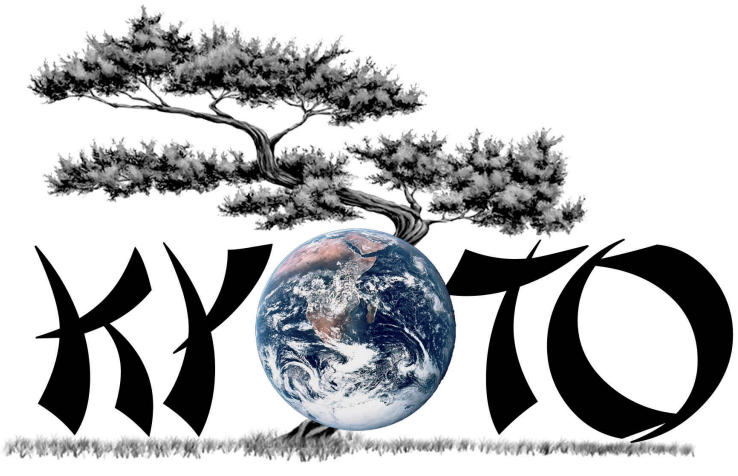
WORDNETS MAPPED TO CENTRAL ONTOLOGY – 1
DELIVERABLE 6.3

Version 2
Date March 2,2009

Axel Herold, Amanda Hicks – BBAW
Roxane Segers, Piek Vossen - VUA
German Rigau, Eneko Agirre, Egoitz Laparra – EHU
Monica Monachini, Antonio Toral, Claudia Soria - CNR-ILC



Knowledge Yielding Ontologies for Transition-based Organization

ICT 211423

Grant Agreement No.	**ICT 211423**
Project Acronym	**KYOTO**
Project full title	**Knowledge Yielding Ontologies for Transition-based Organization**
Technologies	
Funding Scheme	**FP7 – ICT**
Date latest version Annex I	**19-12-2007**
Project Coordinator	**Prof. Dr. Piek T.J.M. Vossen** **VU University Amsterdam** **Tel. + 31 (0) 20 5986466** **Fax. + 31 (0) 20 5986500** **Email: p.vossen@let.vu.nl**
Project website	**http://www.kyoto-project.eu/**
Deliverable Document Number	**D6.3**
Status	**Draft**
Security (distribution level)	**Public**
Contractual date of delivery	
Actual date of delivery	
Type	**report / prototype / software / ontology / wordnets / data / interface / demo / etc.**
WP contributing to the deliverable	**WP6**
WP responsible	
Authors	**Axel Herold, Amanda Hicks – BBAW** **Roxane Segers, Piek Vossen - VUA** **German Rigau, Eneko Agirre, Egoitz Laparra – EHU** **Monica Monachini, Antonio Toral, Claudia Soria - CNR-ILC**
EC project officer	**Werner Janusch**
Keywords	
Abstract	The deliverable describes the progress that has been made toward ontologizing the OntoWordNet with respect to the ontological meta properties of rigidity and non-rigidty, and the progress that has been made toward mapping the Dutch, English, and Italian wordnets onto OntoWordNet. We give a typology of wordnet to ontology mappings, and finally we provide a preliminary discussion of cross lingual rigidity validation within the context of the KYOTO framework.

Table of Content

1 Introduction

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 (5) 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. More specifically, the ontology provides the point of interface for various wordnets.

This deliverable describes the progress that has been made toward ontologizing the OntoWordNet with respect to the ontological meta properties of rigidity and non-rigidty, and the progress that has been made toward mapping the Dutch, English, Spanish, Basque, and Italian wordnets onto OntoWordNet. We give a typology of wordnet to ontology mappings, and finally we provide a preliminary discussion of cross lingual rigidity validation within the context of the KYOTO framework.

2 Description of Rudify

In this section we provide a discussion of rigidity and non-rigidity as ontological meta properties of concepts, the desirability of making these distinctions in an ontology, and describe the development of Rudify, a tool that semi-automatically evaluates rigidity and non rigidity of concepts, in addition to other meta properties.

2.1 *Rigid concepts in ontology*

We have developed Rudify, a collection of tools for ontology tagging used in WP6 for the semi-automatic determination of ontological meta properties, focusing on the meta property rigidity as defined in (Gauriano and Welty, 2002), with the aim of tagging nodes in OntoWordNet (OWN) in order to produce an ontology based on OWN that can (1) easily interface with a WordNet3.0 (since it is already derived from a WordNet1.6), (2) represent certain ontological relations amongst different kinds of concepts and ontological entities, and (3) improve reasoning with rigid and non-rigid terms.

Consider the following fragment taken from WordNet3.0:

Animal
    Chordate
        Cat
    Pet

The relation that structures the hierarchy is the *is-a* relation. For two terms x and y, x is-a y just in case every instance of x is also an instance of y. In this example; every pet is an animal; every chordate is an animal; and every cat is a chordate. Since the *is-a* relation is transitive, it also follows that every cat is an animal. From a linguistic point of view, the *is-a* relation is valuable for representing our use of the words *animal*, *pet*, *chordate*, and *cat*. From an ontological point of view, however, we notice that “cat” and “pet” are two different kinds of concepts; namely, “cat” is rigid and “pet” is non-rigid. Our goal is to begin the process of integrating the distinction between rigid and non-rigid concepts into OWN.

To illustrate the distinction between rigid and non-rigid concepts, let us consider an individual cat, Fluffy. Because we know that Fluffy is a cat, we can infer that he was a cat when he was born and will be a cat when he dies. On the other hand, if Fluffy is a pet, he may cease being a pet at some later date. The fact that Fluffy can cease being a pet but cannot cease being a cat reveals an ontological distinction between rigid and non-rigid concepts.

This distinction can be used to improve formal reasoning.

Consider the two propositions:

> Fluffy was not a pet on Monday.
> Fluffy was a pet on Tuesday.

Because "pet" is a non-rigid concept, there is no interesting conclusion that we can infer from these premises alone. This is not the case, however, with rigid concepts like cat.

> Fluffy was not a cat on Monday.
> Fluffy was a cat on Tuesday.

From these premises, we can actually infer that Fluffy was born on Tuesday.

2.2 *Development of Rudify*

The general idea behind Rudify is the assumption that a preferred set of linguistic expressions is used when talking about ontological meta properties. This idea has been developed and programmatically exploited, first by the AEON project (http://ontoware.org/projects/aeon/). Publications of the AEON approach include (Völker et al 2005) and (Völer et al 2008).

Rudify differs from the original AEON implementation in the follow respects.

- Rudify uses a hinting mechanism to deal with polysemeous lexical representations.
- Rudify employs Google's actively maintained AJAX based interface instead of the SOAP interface for which support was discontinued in 2007.
- Classification/Training is left to standalone Weka which is not incorporated into Rudify. This way it is much easier for non-programmers to make use of all of Weka's algorithms and their parameters. Weka can easily be substituted by any other program being capable of working with ARFF files.
- Rudify is written in Python instead of JAVA.

The general workflow of the Rudify toolchain is as follows:

1. conceptlist.py -- generate a list of concepts that are to be tagged with ontological meta properties.
2. rudify.py -- on the basis of a concept list generate an ARFF file containing numerical feature vectors for each concept's lexical representation
3. Training -- train a model for the classification of the numerical feature vectors or use a supplied model.
4. Classification -- create an ARFF file holding the classification for the cconcepts
5. ontotag.py -- on the basis of the ARFF file with class attribute(s) tag the ontology.

We chose to use Google as the primary source for evidence because of the huge database its query results are based on. We are not aware of any other data source comparable in size. Still, there are major drawbacks in relying on a closed source data repository like Google. Most notably for our application are:

- we do not/cannot control the data on which the query results are based.

- frequent and generally unreported updates of the data base render the query results unstable and irreproducible.

For a further discussion see Adam Kilgarriff: "Googleology is Bad Science" In: Computational Linguistics, vol. 33, no. 1 (March 2007), pp. 147-151. For later versions of the Rudify tools we hope to solve at least some of these problems by using an apropriately sized (web) corpus.

2.2.1 Chosing lexical representations and hinting

For a given synset, currently its first lexical representation is used as the target word form for constructing queries. This is a rather unsophisticated heuristic for deciding which lexical representations represent the synset "best" in actual language use. We are well aware of the fact that this needs to be improved.

Hint generation works on the basis of the synset's lexical representation (LR1) by traversing the synset's direct and indirect hypernyms. The lexical representation (LR2) of the hypernym occuring most often together with LR1 on Google is then added to the original query to "prime" the search engine for the desired sense of LR1.

2.2.2 Training

The handcrafted training set for Rudify consists of 100 prototypically rigid and non-rigid concepts. All of these concepts are monosemeous according to WordNet-3.0 and cover a wide variety of domains. For training, the Weka software suite was used. The list of training concepts is given in the appendix.

Four different algorithms have been used for classification:

- J48 (C4.5 decision tree)
- MLR (multinomial logistic regression, functional)
- NNge (nearest-neighbor-like, rule based)
- LWL (locally weighted learning, instance based)

2.2.3 Preliminary evaluation

We tested Rudify on different data sets (English language data):

- 50 region terms (handcrafted by WP1)
- 236 latin species names (handcrafted by WP1)
- 201 common species names (handcrafted by WP1)
- 297 basic level concepts (BLC-50, automatically created from WordNet-3.0)

Classifiers correctly classified all region terms and all Latin species names as rigid concepts. This holds also for the common English species names with three exceptions: "wildcat" was mis-classified as denoting a non-rigid concept by all four classifiers and "wolf" and "apollo" (a butterfly) were mis-classified by all classifiers except NNge. This mis-classification is due to the fact, that those lexical representations are not monosemeously denoting a single concept (a species) but are polysemous and also frequently used in figurative language (examples are taken from our log files):

- Mount Si High School teacher Kit McCormick is no longer a Wildcat. (generalization from a school mascot to a school member)
- Also the 400 CORBON is no longer a wildcat. (a handgun)
- He nearly gave in and became a Wildcat before finally deciding to honor his original commitment to the Ducks. (a football team's (nick)name)
- For example, the dog is no longer a wolf, and is now a whole seperate species. ()
- For four years, the space agency had been planning, defining, or defending some facet of what led up to and became Apollo. (a space mission's name)
- Others figuring prominently in the countyâ€™s history were Edward Warren, who established a trading post near what is now Apollo [...] (a geographical name)
- The patron of the city is now Apollo, god of light, [...] (a Greek deity)

We are currently testing and evaluating the approach on the ontological concepts of OntoWordNet-0.73. Due to its high number of concepts, not all of the OntoWordNet based data can be checked manually.

2.2.4 Lexical patterns for Spanish and Basque

UPV/EHU has started the first steps to run and test Rudify for Basque and Spanish. Basque is an agglutinative language and as such might require adapting Rudify for being able to work with suffixes. We have focused on rigidity.

We took the lexical patterns for non-rigidity in English as inspiration, and created the following patterns for Basque (with approximate English translations between parenthesis):

X = NOUN

1. *X izateari utzi (stopped being X)*
2. *X bilakatu da (became X)*
3. *X bihurtu da (became X)*
4. *Xtzat hartu (is regarded as X)*
5. *Xtzat jo (is regarded as X)*
6. *X gisa (is regarded as X)*
7. *X izaten jarraitu (keep being X)*

The lexical pattern that would signal non-rigidity for Spanish are the following (with approximate English translations between parenthesis):

X = NOUN

1. *Dejar de ser X / dejar de ser un(a) X (stopped being X)*
2. *Convertirse en X / convertirse en un(a) X (became X)*
3. *Volverse X / volverse un(a) X (became X)*
4. *Tener por X (X != "objeto") / tener por un(a) X (is regarded as X)*
5. *Ahora ya es X / ahora ya es un(a) X (has become X)*

In order to evaluate whether these patterns are valid indicators for rigidity, we have directly queried google using some rigid and non-rigid terms. The following tables show, first the results for Basque, and then for Spanish.

Table 1: Google hits for three Basque rigid words (rows) and the seven lexical patterns for non-rigidity (columns).

	1	2	3	4	5	6	7
txakur (dog)	0	1	10	1	3	48	0
pertsona (person)	0	67	292	56	34	6.62	0
[illegible] (beach)	1	3	7	0	0	55	0

Table 2: Google hits for three Basque non-rigid words (rows) and the seven lexical patterns for non-rigidity (columns).

	1	2	3	4	5	6	7
irakasle (teacher)	41	78	401	22	1	10.600	4
ume (child)	183	104	1050	7	2	57	38
lagun	6	164	397	97	9	518	70

(friend)							

The tables show several issues with the Basque lexical patterns:

- when applied over *person*, the modifiers in front of person make it occur very often in the lexical patterns, e.g. *Gizalege gabeko pertsonatzat jo zuen Leninek* (Lenin regarded him as a rude person)
- pattern 6 does not seem to be very discriminative, as it produced many hits for the three rigid words
- pattern 1 and pattern 7 seem to be the best

Table 3: Google hits for three Spanish rigid words (rows) and the five lexical patterns for non-rigidity (columns). In parenthesis the results for the patterns including the determiner

	1	2	3	4	5
perro (dog)	874 (526)	1150 (3420)	213 (147)	0 (2)	2 (3)
persona (person)	6400 (29800)	27900 (209000)	122 (3340)	1670 (57)	1 (33)
playa (beach)	3 (1)	2090 (697)	4 (0)	1 (0)	0 (0)

Table 4: Google hits for three Spanish non-rigid words (rows) and the five lexical patterns for non-rigidity (columns). In parenthesis the results for the patterns including the determiner.

	1	2	3	4	5
profesor (teacher)	397 (6)	6180 (3990)	202 (5)	9 (0)	1 (1)
niño (child)	11100 (11800)	1010 (12900)	879 (386)	2 (6)	3 (4)
amigo (friend)	2200 (1460)	559 (500)	340 (102)	1720 (5)	6 (2)

The tables show that, contrary to Basque, the Spanish lexical patterns don not seem to be discriminative enough. It seems that just translating the English patterns does not work, and perhaps we will need to create new patterns for Spanish.

3 Quantitative Evaluation of Rudify Output on OntoWordNet

At the time of writing this draft devliverable, Rudify is in the process of evaluating the concepts in OntoWordNet. Once the data becomes available and has been evaluated, we will update this deliverable to include the evaluation of Rudify's performance on OWN.

4 Initial Wordnet to OntoWordNet Mappings

The intitial Wordnet to OntoWordNet Mappings are achieved using the BC concepts as a starting point. By creating mappings from each of the wordnets to the BC, each of these wordnets can also be straightforwardly mapped to the KYOTO-CORE-1 ontology.

4.1 *English*

Because OWN is based on WordNet1.6 and the BC were derived from WN3.0, most of the BC are already contained in OWN. Furthermore, all of the BC concepts are included in KYOTO-CORE-1 ontology. KYOTO-OWN-1 will be modified to reflect the KYOTO-CORE-1 ontology, in such a way that KYOTO-CORE-1 is a fragment of KYOTO-OWN-1. There are two stages to this process.

1. Those nodes that appear in KYOTO-CORE-1 but not in KYOTO-OWN will be added to KYOTO-OWN. More specifically, those BC terms that do not appear in WordNet1.6 and the Kyoto-1 species and regions will be added to KYOTO-OWN-1.

2. The hierarchy relations in KYOTO-OWN will be modified to reflect the hierarchy relations in KYOTO-CORE-1.

This work will be completed by the end of March.

4.2 *Dutch*

The Dutch wordnet is incorporated in the Cornetto database, a lexical semantic database for Dutch. We decided to fix the manual mappings of Dutch synsets to the Base Concepts (BC) in the Cornetto database and then export the result to the Dutch Wordnet-LMF database in KYOTO. The Cornetto database contains equivalence relations to English WordNet 2.0. These relations were derived automatically; for the most frequent and polysemous words in Dutch, the equivalence relations have been checked and edited by hand.

The Base Concepts were extracted from English WordNet 3.0 (EWN3.0), but the Dutch synsets point to English WordNet 2.0 (EWN2.0) and these relations definitely need revision. Therefore, we started a revision process that can be illustrated by figure 1:

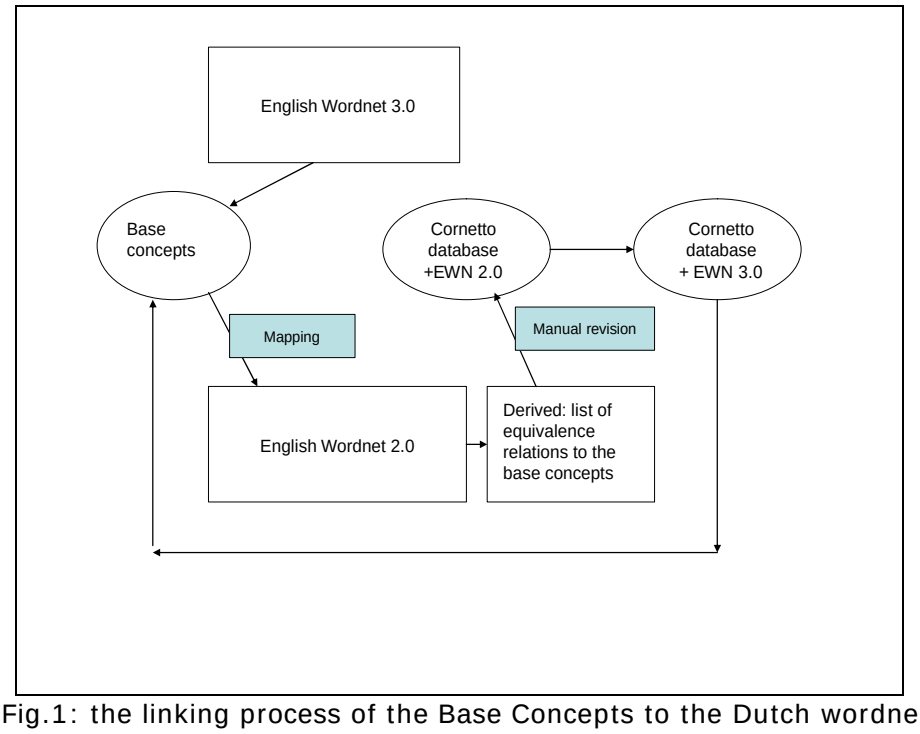


Fig.1: the linking process of the Base Concepts to the Dutch wordnet in Cornetto

We first extended the English WordNet 2.0 (EWN2.0) mappings in Cornetto with English WordNet 3.0 (EWN3.0) mappings. For this, we used the wordnet mapping tables from UPC NLP group in Barcelona: www.lsi.upc.es/~nlp/tools/**mapping**.html. We thus generated a new Cornetto database with both EWN2.0 and EWN3.0 mappings. From this database, we selected all Dutch synsets that have some equivalence relation to the EWN3.0 BCs. The 297 BCs thus yielded 680 Dutch synsets. This is 2.29 synsets per BC on average.

Next, we revised the EWN2.0 mappings in the Cornetto editor to improve the relation between the Dutch synsets and the BCs. When the editing is completed, we then can generate a new version of the Cornetto database by adding EWN3.0 mappings again using the improved EWN2.0 mappings. This new version of the Cornetto database is then used to generate the

Wordnet-LMF database for Dutch that is used in KYOTO. This database will have a very precise mapping to the EWN3.0 BCs which are the basis for the KYOTO core-ontology version 1.

The mapping of the English Base Concepts to the Dutch equivalents is not always a straightforward one to one mapping due to differences between both languages and the structure of the Wordnets. One difference is that the English wordnet is more extensive than the Dutch Wordnet; this means that some of the concepts in the BC list do exist in the Dutch language but are not incorporated in the Dutch Wordnet. This is, for example, the case for all kinds of taxonomic families and genera.

More complicated are the cases where a full equal synonymy relation isn't possible because the English concept is broader or narrower than the closest concept in Dutch. In a few cases, the English concept does not exist in Dutch at all, which is the case for a concept like *containerful*.

Another point is that the current equivalence relations in Cornetto happen to have a lot of many to one relations: many Dutch synsets point to the same English synset. This means that one Base Concept can point to up to ten or even more related synsets in Dutch. This is especially the case for concepts in Dutch that have no equivalence relation in English for example all kinds of words that relate to water management, such as words for areas close to dikes. In this case the Cornetto editors have chosen to make an equal-hypernym relation to the closest synset 'area' in the English wordnet. Now that the synset 'area' is a Base Concept, we will have many Dutch synsets that point to the Base Concept. Some of these were created by hand, others were derived automatically and need checking and revision. Finally, a not so frequent problem is that the Base Concept can be related with full equivalence to a Dutch synset, but the relation was never made.

Due to these complicating factors, the derived list of Dutch synsets from the original set of 297 BCs is much larger: 680 Dutch synsets. To get a better mapping between the Dutch wordnet and the BCs, we checked and revised all the relations by hand. The editing of the equivalence relations is concentrated on minimizing the amount of synsets that relate to a Base Concept. Furthermore, we make the type of relation as specific as possible, preferably an equal synonym. If the Base Concept is broader or narrower then the closest Dutch equivalent, we use a equal hyperonym or an equal hyponym relation. If the Base Concept isn't a concept in Dutch, we also use one or sometimes more equal hyperonym or hyponym relations. This applies to Base Concepts like taxonomic families that are not in Cornetto.

At this point about 50 percent of all the base concepts and their relations to the Cornetto database have been manually checked and revised. 1 shows that the average number of BCs per Dutch synset is now reduced from 2.29 to 0.88 BC3 synsets per Dutch synset.

Table 5: Progress on editing the Dutch wordnet to BC mappings

	BC3.0	DWN2.0	Average nr of BC3 synsets per DWN synset
Total Synsets	297	680	2.29
Manual validated	138	121	0.88

When the work is completed, we expect that we will have about 250 Dutch synsets that relate to the 297 BCs.

Table 2 then shows the distribution of the equivalence relations, before and after the editing. Before editing there are 1630 equivalence relations between 297 BCs and 680 Dutch synsets. Most of these are EQ_NEAR_SYNONYM relations. After editing 50% of the BC mappings, we see that the total number of equivalence relations is reduced by 90% and the EQ_NEAR_SYNONYM mappings even by 96%. In the edited databasem EQ_SYNONYM and EQ_HAS_HYPONYM have become the most important relations.

Table 6: Equivelance relations to BCs before and after editing

Equivalence relation	Nr. before manual revision	Nr. after revision (138 out of 297 BCs covered)	Proportion
EQ_SYNONYM	186	101	54.30%
EQ_NEAR_SYNONYM	1380	52	3.77%
EQ_HAS_HYPONYM	27	13	48.15%
EQ_HAS_HYPERNYM	6	0	0.00%
EQ_HAS_HOLONYM	3	0	0.00%
EQ_HAS_MERONYM	27	0	0.00%
EQ_IS_CAUSED_BY	0	1	
EQ_IS_STATE_OF	1	1	100.00%
TOTAL	1630	168	10.31%

A further revision of the data and a new update of the database is scheduled for the end of March.

4.3 *Italian*

The Italian WordNet (IWN) is connected to WordNet 1.5 through the ILI mechanism of EuroWordNet. However, the BCs for English have been extracted from the last available version of WordNet, 3.0. Because of this, we have automatically added links from IWN to WN 3.0 by exploiting the mapping tables provided by UPC. These tables connect every pair of WordNet versions X and Y in both directions (from X to Y and from Y to X). From the tables that connect WN 1.5 to WN 3.0 and WN 3.0 to WN 1.5 we have derived a unique bidirectional table made of the intersection between the two directional mappings, in order to increase the quality of the obtained mappings.

Once IWN, through the ILI mechanism and the bidirectional table, is connected to WN 3.0, we have tackled the Base Concept issue from two points of view: (i) extracting them from IWN and obtaining the quivalent WN 3.0 synsets and (ii) departing from the BCs of WN 3.0 and deriving the correspondent BCs from IWN. Our motivation for (i) is to find BCs that because of differences related to culture, language and WordNet diversity of structure, might not be present in English but could be extracted from Italian. Concerning (ii), as the WN 3.0 will be part of the Kyoto-1 ontology, mapping them to IWN allows having a direct mapping from the Italian lexicon to the Kyoto-1 ontology.

From IWN to WN 3.0

We have transformed IWN to the input format required by the BLC extraction tool provided by EHU. Subsequently, we have applied this tool to IWN choosing the same threshold that has been used for English; i.e. in order to be extracted as a BC, a synset should represent at least 50 other synsets. This procedure extracts 72 synsets from IWN.

The next step consists in obtaining, from this set of synsets, the corresponding ones in WN 3.0. Each extracted synset from IWN has been enriched with the mapping to the SIMPLE ontology, whereas WN3.0 synsets come equipped with indication of the TCO. This knowledge helps us to check the correctness of the mapping between IWN BCs and WN3.0 BCs by comparing the two ontologies.

From WN 3.0 to IWN

The 297 BCs for WN 3.0 provided by EHU have been automatically mapped to WN 1.5 and from it to IWN. Considering any ILI relation this leads to a set of 4,275 synsets. Due to the high number of synsets and to the fact that connections with ILI relations other than "eq_synonymy" are not that meaningful for the current purpose, we have decided to consider only the IWN synsets that are connected through "eq_synonymy". This choice reduces dramatically the set of IWN synsets to 182.

Table 7: Progress on mapping Dutch synsets to BCs

	Initial number of BCs	Corresponding synsets
IWN ? WN 3.0	72	105
WN 3.0 ? IWN	297	182

At this point, these two different lists can be are compared. From the comparison, 42 synsets emerge that appear both in the Italian list and in the English list. These synsets can be considered as an intersection not cultural and language dependent.

The remaining 63 synsets require manual analysis. It can be the case that they can reveal erroneous mappings or they can represent language-dependent contribution to the Base Concepts and be candidate to inclusion to the KYOTO-1 ontology. Ultimately, they can also go to improve the list of the English BCs with new candidates: e.g., the IWN list of base concepts contains the synset for *mammifero_1* corresponding to *mammal_1* (any warm-blooded vertebrate having the skin more or less covered…) which is missing in the 297 BCs extracted from WordNet3.0 and which seems a good candidate for inclusion in the KYOTO-1 ontology.

Conversely, the path in the other direction, would allow after manual revision to enlarge the set of Italian BCs and having them linked to the ontology.

The results of the analysis of the two lists will be used to improve the mapping IWN-WN3.0 in the Wordnet-LMF version of IWN.

4.4 *Spanish and Basque*

As described in Deliverable D6.1, the current versions of the Spanish and Basque wordnets are the result products of more than ten years of combined effort of several research groups involved in different national and international projects.

The Spanish and Basque wordnets were built following the expand model (Vossen 98). That is, following an automatic method and exploiting several Spanish-English bilingual dictionaries, large sets of WordNet synsets were translated into equivalent synsets in Spanish (Farreres et al. 98). In that way, an aligned version of WordNet 1.5 was built. These preliminary versions were then corrected and augmented manually.

Both Spanish and Basque wordnets were enhanced in the MEANING project[1] (Rigau et al. 02). One of the major goals of this project was the integration of several large-scale knowledge resources. MEANING designed the Multilingual Central Repository (MCR) (Atserias et al. 04a) to act as a multilingual interface for integrating, distributing and ensuring the consistency and integrity of all the semantic knowledge produced by the project. The MCR design followed the model proposed by the EuroWordNet project, whose architecture includes the Inter-Lingual-Index (ILI), a Domain ontology and a Top Concept ontology (Vossen 98).

[1] http://www.lsi.upc.es/~nlp/meaning

During the MEANING project, the Spanish and Basque wordnets were ported to WN1.6 (Atserias et al. 04b). This process was performed thanks to a set of robust and accurate mappings which connects all English WNs maintaining its compatibility across WordNet versions (Daudé et al. 03). See also section 5 of D6.2 for further details. In fact, we are planning to port the whole Spanish and Basque WordNet to be aligned to WN3.0. We started the process by correcting manually those mappings for nouns that the automatic mapping process considers ambiguous. This process was carried out in the framework of the WNTERM project (Pociello et. 2008).

Following the model of the MEANING project, the KNOW project[2] maintains the last version of the MCR which integrates among other semantic resources the Spanish and Basque wordnets. The current version of the MCR can be consulted by using the Web EuroWordNet Interface (WEI)[3].

The current version of the MCR integrates into a common framework:

- The ILI based on WN1.6, including:
  - the EWN Base Concepts (2nd release) (Vossen 98)
  - the MEANING Top Concept ontology TCO 2.3 (Álvez et al. 08)
  - MultiWordNet Domains (Magnini and Cavaglià 00)
  - Suggested Upper Merged Ontology (SUMO) (Niles and Pease 03).
- Local WNs connected to the ILI:
  - English 1.5, 1.6, 1.7, 1.7.1, 2.0, 2.1, 3.0 (Fellbaum 98)
  - eXtended WN (Mihalcea and Moldovan 01)
  - Basque (Pociello et al. 08)
  - Italian (Pianta et al. 02)
  - Catalan and Spanish WN (Atserias et al. 04b).
- Large collections of semantic preferences acquired from SemCor (Agirre and Martinez 02).
- Instances
  - Named instances (Alfonseca and Manandhar, 02)
  - Named instances (Niles and Pease 03)
  - Named instances (Pianta et al. 02)

As the Spanish and Basque wordnets follow the expand model, both wordnets are tightly aligned to the English WordNet. Thus, thanks to the integration in the MCR of the Princeton WordNet 3.0, instead of applying the Basic Level Concept program on both the Spanish and Basque wordnets, we simply ported the Spanish and Basque wordnets to version 3.0.

After this process, we are able to find out which senses from KYOTO-OWN-1 have the corresponding translations to Spanish and Basque except for those BC concepts derived from WN3.0 without corresponding OWN label (that is, new concepts in WN3.0 which were not in WN1.6). These three concepts should be also translated to Spanish and Basque:

en30-00020090-n 03 264 substance | Substance

[2] http://ixa.si.ehu.es/know

[3] http://adimen.si.ehu.es/cgi-bin/wei/public/wei.consult.perl

en30-07555863-n 13 323 food solid_food | Comestible Solid

en30-08392137-n 14 108 terrorist_organization terrorist_group foreign_terrorist_organization FTO | Function Group Human

As a result of this process, only 12+3 concepts (around 5%) from KYOTO-OWN-1 have no corresponding Spanish translation:

en30-00661091-n 04 53 therapy | Agentive Condition Dynamic Physical Purpose Social UnboundedEvent | THER APY

en30-01525720-n 05 239 oscine oscine_bird | Animal Object | OSCINE__OSCINE_BIRD

en30-01727646-n 05 52 colubrid_snake colubrid | Animal Object | COLUBRID_SNAKE__COLUBRID

en30-05701944-n 09 281 basic_cognitive_process | Agentive Dynamic Mental | BASIC_COGNITIVE_PROCESS

en30-06043075-n 09 70 medicine medical_specialty | Agentive Condition Mental Purpose Social UnboundedEve nt | MEDICINE

en30-10284064-n 18 69 maker shaper | Function Human Object | MAKER__SHAPER

en30-12998815-n 20 98 agaric | Object Plant | AGARIC

en30-13086908-n 20 186 plant_part plant_structure | Object Part Plant | PLANT_PART

en30-13576355-n 23 143 indefinite_quantity | 3rdOrderEntity Quantity | INDEFINITE_QUANTITY

en30-13745420-n 23 54 large_integer | 3rdOrderEntity Quantity | LARGE_INTEGER

en30-13756125-n 23 59 containerful | 3rdOrderEntity Quantity | CONTAINERFUL

en30-15157225-n 28 164 day | 3rdOrderEntity Purpose Quantity Social Time | DAY_4

And only 51+3 concepts (around 18%) from KYOTO-OWN-1 have no corresponding Basque translation:

en30-00661091-n 04 53 therapy | Agentive Condition Dynamic Physical Purpose Social UnboundedEvent | THER APY

en30-01090446-n 04 177 commerce commercialism mercantilism | Agentive Dynamic Purpose Social | COMMERCE_

_COMMERCIALISM__MERCANTILISM

en30-01342529-n 05 256 animal_order | Group | ANIMAL_ORDER

en30-01355326-n 05 70 eubacteria eubacterium true_bacteria | Living | EUBACTERIA__EUBACTERIUM__TRUE_BACT ERIA

en30-01428580-n 05 125 soft-finned_fish malacopterygian | Animal Object | SOFT-FINNED_FISH__MALACOPTERYG IAN

en30-01429349-n 05 171 fish_family | Group | FISH_FAMILY

en30-01432517-n 05 289 fish_genus | Group | FISH_GENUS

en30-01504437-n 05 143 bird_family | Group | BIRD_FAMILY

en30-01507175-n 05 399 bird_genus | Group | BIRD_GENUS

en30-01525720-n 05 239 oscine oscine_bird | Animal Object | OSCINE__OSCINE_BIRD

en30-01657723-n 05 162 reptile_genus | Group | REPTILE_GENUS

en30-01727646-n 05 52 colubrid_snake colubrid | Animal Object | COLUBRID_SNAKE__COLUBRID

en30-01759182-n 05 181 arthropod_family | Group | ARTHROPOD_FAMILY

en30-01762525-n 05 256 arthropod_genus | Group | ARTHROPOD_GENUS

en30-01862557-n 05 114 mammal_family | Group | MAMMAL_FAMILY

en30-02552171-n 05 100 spiny-finned_fish acanthopterygian | Animal Object | SPINY-FINNED_FISH__ACANTHOPT ERYGIAN

en30-02858304-n 06 64 boat | Artifact Instrument Object Vehicle | BOAT

en30-03419014-n 06 271 garment | Artifact Garment | GARMENT

en30-04194289-n 06 83 ship | Artifact Instrument Object Vehicle | SHIP

en30-04451818-n 06 167 tool | Artifact Instrument Object | TOOL

en30-04623612-n 07 99 disposition temperament | Experience Mental Property | DISPOSITION__TEMPERAMENT

en30-05267548-n 08 153 animal_tissue | Living Part Solid | ANIMAL_TISSUE

en30-06043075-n 09 70 medicine medical_specialty | Agentive Condition Mental Purpose Social UnboundedEve nt | MEDICINE

en30-06814870-n 10 88 musical_notation | 3rdOrderEntity Communication Purpose Usage | MUSICAL_NOTATION

en30-07283608-n 11 812 happening occurrence occurrent natural_event | BoundedEvent | HAPPENING__OCCURREN CE__NATURAL_EVENT

en30-07570720-n 13 126 nutriment nourishment nutrition sustenance aliment alimentation victuals | Comest

ible Substance | NUTRIMENT__NOURISHMENT__SUSTENANCE__ALIMENT__ALIMENTATION__VICTUALS

en30-07975026-n 14 270 gathering assemblage | Group Human | GATHERING__ASSEMBLAGE

en30-08552138-n 15 153 district territory territorial_dominion dominion | Part Place | DISTRICT__TERRITO RY

en30-08655464-n 15 53 American_state | Part Place | AMERICAN_STATE

en30-09360122-n 17 69 mountain_peak | Place | MOUNTAIN_PEAK

en30-09484664-n 18 129 mythical_being | Creature | MYTHICAL_BEING

en30-09522978-n 18 59 Hindu_deity | Creature Function | HINDU_DEITY

en30-09765278-n 18 143 actor histrion player thespian role_player | Human Object Occupation | ACTOR__HIS TRION__PLAYER__THESPIAN__ROLE_PLAYER

en30-09977660-n 18 105 criminal felon crook outlaw malefactor | Function Human Object | CRIMINAL__FELON_ _CROOK__OUTLAW__MALEFACTOR

en30-10284064-n 18 69 maker shaper | Function Human Object | MAKER__SHAPER

en30-10287213-n 18 65 man adult_male | Function Human Object | MAN__ADULT_MALE

en30-11534677-n 20 74 plant_order | Group | PLANT_ORDER

en30-11554175-n 20 64 gymnosperm_genus | Group | GYMNOSPERM_GENUS

en30-11556857-n 20 308 monocot_genus liliopsid_genus | Group | MONOCOT_GENUS__LILIOPSID_GENUS

en30-11562747-n 20 196 dicot_family magnoliopsid_family | Group | DICOT_FAMILY__MAGNOLIOPSID_FAMILY

en30-11579418-n 20 321 asterid_dicot_genus | Group | ASTERID_DICOT_GENUS

en30-11585340-n 20 298 rosid_dicot_genus | Group | ROSID_DICOT_GENUS

en30-11590783-n 20 74 fungus_family | Group | FUNGUS_FAMILY

en30-11592146-n 20 139 fungus_genus | Group | FUNGUS_GENUS

en30-12425281-n 20 125 liliaceous_plant | Object Plant | LILIACEOUS_PLANT

en30-13167078-n 20 114 fern_genus | Group | FERN_GENUS

en30-13576355-n 23 143 indefinite_quantity | 3rdOrderEntity Quantity | INDEFINITE_QUANTITY

en30-13756125-n 23 59 containerful | 3rdOrderEntity Quantity | CONTAINERFUL

en30-14070360-n 26 454 disease | Condition Physical Property | DISEASE

en30-14189204-n 26 60 blood_disease blood_disorder | Condition Physical Property | BLOOD_DISEASE__BLOOD_ DISORDER

en30-14253124-n 26 70 animal_disease | Condition Physical Property | ANIMAL_DISEASE

Obviously, for these concepts it will be necessary a manual inspection and treatment to assign the corresponding Spanish and Basque word senses. This process will be completed by the end of March.

5 Wordnet to wordnet mappings

5.1 *Connecting large-scale semantic structures*

There is an increasing need of broad-coverage and accurate lexical/semantic resources for the development of current Natural Language Processing (NLP) systems. Thus, one of the main issues in last years concerning NLP activities has been focused on the fast development and tuning of general semantic resources to particular applications and domains.

Using large-scale knowledge bases, such as WordNet (Fellbaum, 1998), has become a usual, often necessary, practice for most current Natural Language Processing (NLP) systems. Even now, building large and rich enough knowledge bases for broad–coverage semantic processing takes a great deal of expensive manual effort involving large research groups during long periods of development. In fact, hundreds of person-years have been invested in the development of wordnets for various languages (Vossen, 1998). Unfortunately, the outcomes of these projects are, usually, large and complex semantic structures, hardly compatible with resources developed in other projects and efforts. Obviously, this fact has severely hampered Human Language Technology (HLT) development.

To maintain compatibility between lexical knowledge bases of different languages and versions, past and new, it is fundamental to have available a robust, high accurate and fully automatic tool ready to use for enabling the mapping between lexical units in different knowledge bases.

For knowledge intensive NLP tasks, the availability of rich enough and wide coverage semantic resources is an issue of special interest. Most of these resources, such as DOLCE – OntoWordNet (Gangemi et al. 2002; Gangemi et al. 2003), SIMPLE (Lenci et al. 2000), Top Concept Ontology (Álvez et al. 2008), SUMO (Niles and Pease 01), Opencyc (*Matuszek et al. 2006*), EDR (Yokoy 1995) or WordNet (Fellbaum 1998) differ in great extent on several characteristics (e.g. broad coverage vs. domain specific, lexically oriented vs. conceptually oriented, granularity, kind of information stored, kind of relations, way of being built, etc.), and in most cases have been connected to a particular WordNet version.

To deal with the semantic mismatches between resources and versions, and to minimize side effects with respect to other initiatives world-wide, such as those exploiting Wikipedia[4] as DBpedia[5] (Auer et al. 2007), Geonames[6] or other wordnet developments around the Global WordNet Association, KYOTO requires a generic, powerful and robust mapping tool and a set of improved mappings between all involved resources.

Fortunately, most of these resources have been connected to WordNet. However, since these developments have been performed during a large period of time, different WordNet versions have been used. For instance, both the Italian wordnet and SIMPLE were connected to WN1.5. WN1.6 was used to map the Top Concept Ontology, OntoWordNet, some versions of SUMO and the Spanish and Basque wordnets, while WN2.0 was used to map Opencyc and the Dutch wordnet. Finally, WN3.0 is being used to connect Wikipedia articles in Dbpedia by means of the Yago[7] effort (Suchanek et al. 2007).

Thus, our main problem consists on selecting the best mapping technology for mapping wordnet versions, since by connecting the different WordNet versions, we will be able to semantically integrate all these large-scale knowledge resources.

5.2 *Available WordNet mappings*

To our knowledge, currently there are available the following wordnet mappings:

- Princeton mappings[8] provides Sensemap to help WordNet users to automatically convert noun and verb senses from an old version to their corresponding senses of the new version.
- Rada Mihalcea[9] provides a limited set of mappings corresponding to versions WN1.6, WN1.7, WN1.7.1 and WN2.0
- UPC mappings[10] provides a complete set of mappings for all WordNet versions from WN1.5 to WN3.0in both directions. The mappings include all POS (nouns, verbs, adjectives and adverbs).

5.3 *Sensemap from Princeton*

The Princeton web site provides the Sensemap files for the last WN version[11]. These mappings correspond to the last release, that is from WN2.1 to WN3.0.

---

4 http://*www.wikipedia.org/*
5 http://dbpedia.org/About
6 http://www.geonames.org/
7 http://www.mpi-inf.mpg.de/~suchanek/downloads/yago/
8 http://wordnet.princeton.edu/
9 http://www.cse.unt.edu/~rada/downloads.html
10 *http:www.lsi.upc.es/~nlp/tools/mapping.html*
11 http://wordnet.princeton.edu/3.0/WNsnsmap-3.0.tar.gz

This sense mapping was done as follows:

- Nouns and verbs unique to either database were ignored.
- Nouns and verbs that are monosemous in both databases were found and their sense_key s and synset_offset s were mapped. These sense mappings are in the files 2.1to3.0.{noun,verb}.mono.
- All senses of polysemous nouns and verb in version 2.1 were mapped to senses in version 3.0. Various heuristics were used to evaluate the similarity of 2.1 and 3.0 senses, and a score was assigned to each comparison. For each word, each 2.1 sense was compared to all of the 3.0 senses for the same word, and the 3.0 sense (or senses) with the highest score was deemed the best mapping. These sense mappings are in the file 2.1to3.0.{noun,verb}.poly. Heuristics include comparison of sense keys, similarity of synset terms, and relative tree location (comparision of hypernyms). Glosses are not used for comparisions, as they are often significantly modified.

Most WordNet users think that WordNet sensekeys are version-independent. For example, break#v#4 in WN2.1 is break#v#3 in WN3.0, but the sensekey is the same for both, "break%2:30:06::".

However, using Sensemap we can follow the sensekeys across WordNet versions. Surprisingly, according to Sensemap, it seems that for nouns, there are 22 monosemous and 121 polysemous mappings with different sensekeys. And for verbs, there are 2 monosemous and 135 polysemous mappings with different sensekeys.

The Sensemap files have been provided together with the official release for all WN versions since WN1.6. Table 3 reports the total amount of the different sensekeys across WordNet versions as provided by Sensemap.

Table 8: Amount of different sensekeys across WordNet versions

WN release	Mapping	#Different sensekeys
1.6	(WN1.5->WN1.6)	728
1.7.1	(WN1.6->WN1.7.1)	578
2.0	(WN1.7.1->WN2.0)	566
2.1	(WN2.0->WN2.1)	317
3.0	(WN2.1->WN3.0)	280

This means that using the sensekeys as unique identifiers for word senses across WordNet versions, a mapping error is introduced. Possibly, this error is augmented when using long distance WordNet versions (like from WN1.6 to WN3.0).

Moreover, sensekey mappings are only valid for word senses, not for synsets. That is, since sensekeys can be grouped in different synsets depending on the WordNet version, it is not possible to transport the knowledge associated to a particular synset (for instance, ontology labels or translated words in other languages) from one WordNet version to another.

Possibly, using these mappings (not directly the sensekeys) it is possible to follow the history of every nominal and verbal synset and every sensekey. But the sensemap also rely on a set of heuristics, which are not 100% correct. Furthermore, Sensemap only covers nouns and verbs. No Sensemap is provided for adjectives nor adverbs.

5.4 *Relaxation Labeling Algorithm*

The automatic mappings provided by UPC rely on the use of a robust robust approach for linking already existing lexical/semantic hierarchies (Daudé et al. 2000). This approach uses a constraint satisfaction algorithm (relaxation labeling) to select -among a set of candidates- the node in a target taxonomy that bests matches each node in a source taxonomy. The Relaxation Labeling algorithm exploits all information available in both semantic resources including the structural information (i.e. all synset relationships), and the words from the glosses and synset words in the cases where structure is not enough (adjectives and adverbs) (Daudé et al. 2001; Daudé et al. 2003).

In order to produce the mappings for all POS, the whole linking process is performed in three phases, following the relationship that establish dependencies among different PoS in WordNet, as described below:

The process starts by mapping the nominal part of both WordNet versions using only the hyper/hyponymy relationships plus the words of the synsets and the words from the glosses as additional constraints.

Similarly, verbs are mapped using hyper/hyponymy, antonymy and the "also-see" WordNet relationships plus some additional constraints coming from the words from synsets and glosses and also the verbal frames.

Then, adjectives are mapped using adj-to-adj relationships such as antonymy, "similar-to" and "also-see", as well as the adj-to-verb relationship "participle-of" and the adj-to-noun "pertains" and "attribute". Words from the synsets and glosses are also included as additional constraints.

Notice that the graph used to check the constraints imposed by adj-to-verb and adj-to-noun relationships was the result of the previous steps. That means that verbs and nouns are already mapped, reducing considerably the search space and accelerating relaxation labeling convergence.

5.5 *Comparison*

The figures presented in this section were computed by manually linking to WN1.6 a sample randomly chosen from WN1.5, and then use this sample mapping as a reference to evaluate system output (obviously, the system maps the entire taxonomies, though correctness is only evaluated on the sample list nodes). The validation sample consists of 1900 noun synsets, 1000 verb, 1000 adjective and 300 adverb synsets. See (Daudé et al. 2003) for further details.

Table 9: Performance comparison between Sensemap and Relax mappings.

	Sensemap		Relax	
	Precision	Recall	Precision	Recall
Noun	98,75%	99,65%	99,65%	99,70%
Verb	95,73%	97,88%	98,70%	98,90%

As shown in table 4, for both precision and recall for both nouns and verbs the Relax algorithm achieve better performances.

No comparison have been performed for the Rada Mihalcea mappings since the first mapping provided corresponds to WN1.6. So, there is no WN1.5 to WN1.6 mapping.

5.6 *Concluding remarks on wordnet-to-wordnet mappings*

WordNet sensekeys are not version-independent, that is, they are not stable across versions. Furthermore, as Sensemap does not provide mappings for adjectives and adverbs, the number of sensekey mismatches across WordNet versions is unknown. Moreover, the sensekey mappings are performed at a word sense level. That is, since sensekeys can be grouped in different synsets depending on the WordNet version, it is not possible (without manual checking) to transport the knowledge associated to a particular synset (for instance, ontology labels or translated words in other languages) from one WordNet version to another.

Conversely, the UPC mappings provide a robust and complete (for all nouns, verbs, adjectives and adverbs) set of mappings across all WordNet versions. Furthermore, the mapping is produced in both directions (from source to target and viceversa). Further research is required to measure the performance of the intersection of both mappings.

6 Typology of Wordnet to Ontology Mappings

In general there can be 3 ways to handle the relations between concepts in the lexicon and concept in the ontology:

1. all lexical concepts are also added as concepts to the ontology and there is a single and direct mapping from the lexical concept to the ontological concepts
2. there can be many to many relations between the lexicon concepts and the ontology concepts but all the axioms are in the ontology
3. there are axioms in the lexicon in addition to the axioms in the ontology

Obviously, there can be solutions that combine these three situations.

The first case puts an enormous burden on the ontology since it needs to include concepts for all lexicalized concepts in all languages. This means not only cultural specific concepts but also many roles and other contextuals that happen to be lexicalized in languages all over the world. Examples of the latter are e.g. female/male variants in non-English languages: *Lehrer* and *Lehrerin*. This proposal is now modeled in the Wordnet-LMF proposal in the KYOTO project.

In the second case, the ontology contains a subset of the lexicalized concepts. Axioms reside in the ontology and all inferencing as well. The lexicalized concepts that are not in the ontology need to related through multiple relations to more than one concept in the ontology. For example: Lehrer -> Teacher + Male, Lehrerin -> Teacher + Female. This proposal is represented in LexInfo. LexInfo e.g. proposes to link Schweineschnitzel to both Pig and Cutlet. The drawback is that the relation between the Pig and the Cutlet is not clear. LexInfo also has a proposal for linking slots in verb frames to Ontological fillers.

In the third case, complex mappings can be created. This can solve the Schweineschnitzel problem in that a little axiom is created that indicates that the Cutlet consists of the meat of a Pig. The ontology can remain minimal but axioms go in the lexicon. This proposal is modeled in the solution of the Cornetto project (Vossen et al 2008).

6.1 *Wordnet-LMF*

Currently in KYOTO, WordNets are described through the WordNet-LMF model, a common representation format, derived from the ISO Lexical Markup Framework which offers a metamodel as a standard framework for modeling and representing computational lexicons. LMF allows the definition of morpho-syntactic information (e.g. part-of-speech) as well as semantic representation of words in terms of senses. Since the model developed in KYOTO has been tailored to the representation of WordNet-like lexicons, senses are organized around synsets:

```
<LexicalEntry id="footprint">
  <Lemma writtenForm="footprint" partOfSpeech="n">
```

```
    </Lemma>
    <Sense id="footprint_1" synset="eng-30-06645039-n">
    </Sense>
</LexicalEntry>

<Synset id="eng-30-06645039-n" baseConcept="1">
      <Definition gloss="a trace suggesting that something was once present or felt or otherwise
important">
    <Statement example="the footprints of an earlier civilization" />
</Definition>
</Synset>
```

The conceptual units on their turn, are defined by lexical semantic relations and, informally, by semantic definitions. WordNet-LMF stops where lexical semantics of words stop.

The formal definition of the lexical entries in terms of an ontology, i.e. a set of semantic types in a formal system of representation of knowledge is left outside of the lexicon.

WordNet-LMF naturally lends itself to represent the two first ways of interfacing lexicons to ontologies; in any case, the formal knowledge resides in the ontology.

The WordNet LMF common representation format allows handling the mapping between synsets in the lexicon and concepts in the ontology at two different levels: the level of different monolingual WordNets and the level of the multilingual WordNet grid. Indeed, two different lexical objects are designed to represent the link to ontology(ies), the *MonolingualExternalRef* and the *InterlingualExternalRef*.

6.1.1 Mapping to an ontology at monolingual level

Every monolingual WordNet can enter in the KYOTO lexical grid equipped with the link to the ontology(ies) usually adopted for the semantic encoding of the monolingual lexical entries, i.e. the Dutch WordNet maintains the links to Terms and Axioms derived from SUMO and MILO (as in the Cornetto database); ItalWordNet carries the mapping to the SIMPLE-Owl ontology, the Spanish WordNet preserves the mapping of synsets to SUMO, etc...

In the WordNet-LMF model, at the level of language-specific WordNets , the lexicons can be anchored to the ontology either via the word senses, i.e. individual word meanings (*Sense* lexical object) or the synsets, concepts (*Synset* lexical object).

Any reference or correspondence from the WordNet monolingual lexicons to an external resource (an ontology as well as another lexical database) is made possible in WordNet-LMF by means of the object *MonolingualExternalRef*.

The type of information represented here varies according to the particular parent element in which it appears.

- when it is linked to the *Sense* element, it can be used to express mapping between a sense and its correspondent in another lexical resource. In the particular case of English WordNet, it can serve as a representational device to express the SenseKey value.
- when occurring inside the representation of the *Synset* element, then *MonolingualExternalRef* allows to encode reference to the domain and/or one or more links to an ontological system.

The *MonolingualExternalRef* expresses the name of the external resource and the particular identifier or node, by means of the two required attributes, ‘externalSystem’ and ‘externalReference’. The values of the ‘externalSystem’ are the names of the different ontologies. The ‘externalReference’ encodes the particular identifier of the node in the ontology,

whereas, the attribute 'relType' serves to specify relations with nodes in the ontology. Possible relations are: "at", the synset is an instance of the concept in the ontology; "plus", the synset is subsumed by the concept in the ontology; "equal", the synset is equivalent to the concept in the ontology.

An example of linking from concepts to a given ontology:

```
<Synset id="eng-30-06645039-n" baseConcept="1">
...
<MonolingualExternalRefs>
<MonolingualExternalRef externalSystem="SUMO" externalReference="superficialPart"
relType="at"/>
</MonolingualExternalRefs>
</Synset>
```

6.1.2 Mapping to the ontology at multilingual level

The WordNet-LMF format allows representing a multilingual grid of WordNets. Those synsets which are equivalent and correspond with each other in different monolingual WordNets are related at the multilingual level via the *SenseAxis* lexical object.

The object *InterlingualExternalRef* is used in WordNet-LMF to link a *SenseAxis* instance to an external system such as the KYOTO ontology and represents the means to anchor a multilingual group of synsets to an ontological node. Its intended use, thus, is to provide a representational device to link a group of synsets from different wordnets to the same ontological concept. The *InterlingualExternalRef* package should not be used to link a monolingual synset to an ontology.

The attribute *externalSystem* contains the name of the external resource and the *externalReference* encodes the particular identifier or node. The *relType* is intended for the representation of the type of relations with the ontology nodes. Possible relations are: "at", the synset is an instance of the concept in the ontology; "plus", the synset is subsumed by the concept in the ontology; "equal", the synset is equivalent to the concept in the ontology

The following example illustrates the case of Italian, Spanish, Chinese and English synsets for "fire", all related by an "equal_synonym" relation and pointing to the same SUMO ontological node "Combustion".

Example :

```
<SenseAxis id="sa_001" relType="eq_synonym">
<Target ID="ita-16-0001251-n"/>
<Target ID="spa-16-09686541-n"/>
<Target ID="zho-14-05231501-Na"/>
<Target ID="eng-30-13480848-n"/>
<InterlingualExternalRefs>
<InterlingualExternalRef externalSystem="SUMO" externalReference="Combustion"
relType="at"/>
<InterlingualExternalRefs>
</SenseAxis>
```

This architecture where corresponding synsets preserve at monolingual level their ontological typing and, at multilingual level, are referenced to the same KYOTO shared ontology allows obtaining an indirect mapping of pre-existent ontological type systems onto the KYOTO ontology.

6.2 *Wordnet to Sumo mappings in Cornetto*

Another way of handling the relation between concepts and the ontology was demonstrated in the Dutch Cornetto project (Vossen et al 2008), in which mappings between synsets and the SUMO ontology have been created using simple RDF like relation triplets.

The SUMO ontology terms were imported through the automatically derived equivalence relations to PWN2.0 (Niles and Pease 2001, Niles and Pease 2003). The mapping relations of the English synsets to SUMO and MILO were copied to the equivalent synsets in Dutch. These mappings were further manually adapted in an editing process.

In the SUMO to PWN mapping, the following relations are used:

= the synset is equivalent to the SUMO concept
+ the synset is subsumed by the SUMO concept
@ the synset is an instance of the SUMO concept
[the SUMO concept is subsumed by the synset

The mappings from PWN to SUMO are binary relations. In Cornetto, triplets are used to make more complex expressions. For the triplets, the above relations have been extended with all the relations that were defined in SUMO (April 2006) and are used in the axioms.

In the XML of the Cornetto database, the SUMO element 'ont_relation' contains a relation attribute (relation_name) with either the characters '+', '=', '[', '@' or one of the above relations as a value. Furthermore it has two attributes (arg1 and arg2) that represent the two arguments of the triplets. Other attributes are status 'true' or 'false', a name for whom created the mapping, and an attribute for 'negative'. Below is an example of the 'ont_relation' element:

```
<ont_relation status="false" name="dwn10_pwn16_pwn20_mapping" negative="false"
relation_name="+" arg1="" arg2="Artifact"/>
```

The relation name and the two arguments represent the triplet. The triplets are used as simplified representations of semantic implications. The arguments of the triplets follow the syntax of the relation names in SUMO. The fillers can be a SUMO term or a variable. Variables are integers, where the integer '0' is reserved to co-index with the referent of the synset that is being related. Empty argument slots are assumed to hold the value '0' as well. For example the following expressions are possible in the Cornetto database, where each a) and b) example is equivalent. The slight difference is due to whether the ontology mapping was automatically derived (a) or created by hand (b).

1. Equality:
   a. (=, 0, Circle)
   b. (=, , Circle)
2. Subsumption:
   a. (+, 0, Artifact)
   b. (+, , Artifact)

The two next types of triplets are used instead of the complex SUMO-KIF expressions in SUMO. These triplets are used to specify a complex mapping relation to the SUMO ontology, in case the basic mapping relations are not sufficient. This is especially the case for so-called non-rigid concepts that are not present in SUMO, e.g. 'theewater' (water for making tea) is not a type of water but water in a particular role.

3. Related:
   a. (part, 0, PlantBranch)
   b. (part, , PlantBranch)
4. Axiomatized:

a.(instance, 0, Water) (instance, 1, Making) (instance, 2, Tea) (resource, 0, 1) (result, 2,1)
b.(instance, , Water) (instance, 1, Making) (instance, 2, Tea) (resource, , 1) (result, 2,1)

By assuming default values for the KIF syntax, it is possible to generate more complex expressions that come close to the axioms in SUMO. Such an interpretation can be derived as follows: the default operator for a list of the triplets is AND, and we assume default existential quantification of any of the variables, specified as a value of the arguments. Furthermore, the convention was followed to use a zero symbol as the variable that corresponds to the denotation of the synset being defined and any other integer for other denotations. Finally, the symbol ⇔ was used in the explanation of the triplets for full equivalence (bidirectional subsumption). In the case of partial subsumption, the symbol ⇒ was used, meaning that the KIF expression is more general than the meaning of the synset. If no symbol is specified, an exhaustive definition by the KIF expression is assumed. The symbol ⇔ applies by default. A triplet for 'waakhond' should then be read as follows:

(instance, 0, Canine) (instance, 1, Guarding) (role, 0, 1)

The expression exhaustively defines the synset (Û), AND there exists an instance 0 of the type Canine (instance, 0, Canine), AND any referent of an expression with the synset {waakhond} as the head is also an instance of the type Canine (the special status of the zero variable), AND there exists an instance of the type Guarding 1 (instance, 1, Guarding), AND the entity 0 has a role relation with the entity 1 (role, 0 ,1).

The triplets can also be used to define new rigid types that are not in SUMO. In that case, we state that the synset has the names for these types. For names of rigid types, e.g. 'hond' as a Dutch name for Canine, the following expressions were proposed in Cornetto:

hond (=, 0, Canine); the synset {hond} is a Dutch name for the rigid type Canine
bokser (+, 0, Canine); the synset {bokser} is a Dutch name for a rigid concept which is a subclass of the type Canine

Naming relations were mostly imported from the SUMO mappings to the English Wordnet through the equivalence relation of the Dutch synset to the English synset. In the case of {bokser}, the mapping needs to be manually added because this dog race is not in the English Wordnet and not in SUMO. Possibly, SUMO could be extended with this type.

Within the wordnet hierarchy, we find many cases of mixtures of rigid and non-rigid concepts that have the same hyponym relation, i.e. both 'waakhond' (watchdog) and 'bokser' (boxer) are hyponyms of 'hond' (dog). In principle, the triplets can be used to differentiate between these through their mapping to the ontology. Note that the most frequent structure in the database now is a single triplet with the relation '+'. This means that the synset is simply labeled by subsumption to a SUMO concept and nothing is stated about the rigidity of the concept represented by the synset. Such more explicit mappings will be made in a future extension of the database.

6.3 *LexInfo*

LexInfo (Buitelaar et al 2008, 2009: http://lexonto.ontoware.org/lexinfo) is a rich lexicon model for associating linguistic information with ontologies. Like Wordnet-LMF, LexInfo model builds on the Lexical Markup Framework (LMF) as a core but extends it appropriately to accommodate the essential aspects of the LingInfo and LexOnto models. The LingInfo model (Buitelaar et al 2006a,b) provided a mechanism for modeling label internal linguistic structure, i.e., lexical inflection and morphological or syntactic decomposition of ontology labels, interpreted as terms. The LexOnto model (Cimiano et al 2007) on the other hand enabled the representation of label-external linguistic structure, i.e., predicate-argument structures that can be projected by lexical heads of ontology labels and their mapping to corresponding ontological structures.

The LexInfo model is an extension of LMF so that rich and complex lexical structures can have adequate mappings to elements in the ontological specification. All the linguistic information remains in the linguistic structure of LMF, which is clearly distinguished from the ontological specification. As such the LexInfo model is compatible and in fact complementary to the KYOTO representation of wordnets in Wordnet-LMF, as discussed above.

The LexInfo model is based on the following requirements for grounding linguistic information in ontologies:

1. capture morphological relations between terms, e.g., through inflection (cat, cats), separately from the domain ontology;
2. represent the morphological or syntactic decomposition of composite terms and the linking of the components to the ontology;
3. model complex linguistic patterns, such as subcategorization frames for specific verbs together with their mapping to arbitrary ontological structures;
4. specify the meaning of linguistic constructions with respect to an arbitrary (domain) ontology, and
5. clearly separate the linguistic and semantic (ontological) representation levels.

Morphological relations between terms and morphological and synactic composition are represented using the machinery from LMF. LexInfo further models the composition of complex morphemes in OWL by introducing an additional data type property order specifying the absolute order of a Component within a ListOfComponents. Components then in essence point to LexicalEntries which can again be composite, thus allowing for recursion. In order to capture how the parts of a compound are associated to the ontology, LexInfo uses the general mechanism of LMF to associate LexicalEntry objects with a Sense. For this subclasses owl:Property and owl:Class are defined. In this way, the model states that `Schweineschnitzel' is composed of two Lexical Entry objects where the first refers to the class pork and the second to the class cutlet. The crucial extension of LMF was here the fact that Entity (in the OWL 2 meta model) is specified as a subclass of lmf:Sense in the LexInfo ontology.

The properties of subcategorization frames are again modeled in LMF. The syntactic arguments are represented by introducing subclasses of lmf:SyntacticArgument, i.e., lexinfo:Subject, lexinfo:Object, lexinfo:PObject), etc. The semantic layer is represented in LMF by introducing subclasses of the lmf:PredicateRepresentation and lmf:SemanticPredicate classes, e.g., the classes lexinfo:ClassPredicativeRepresentation and lexinfo:ClassPredicate as well as lexinfo:PropertyPredicativeRepresentation and lexinfo:PropertyPredicate allowing to refer to a class or property (as predicate), respectively. Furthermore, LMF is extended with subclasses for the lmf:SemanticArgument class, i.e. lexinfo:Domain, lexinfo:Range etc., as well as appropriate subclasses allowing to specify the semantic arguments of a class (where properties are understood as slots of the frame represented by the class).

The crucial class establishing the connection between the syntax and semantic (ontological) levels is then the lmf:SynSemCorrespondence class, which is associated to various lmf:SynSemArgMaps, mapping a certain syntactic position, e.g. lexinfo:Subject and lexinfo:PObject of a particular verb, to semantic arguments of an ontological predicate. The next figure is taken from Buitelaar et al 2009 and shows the extension of LMF with subclasses for the representation of the frame for the verb “flow” and the SynSemCorresponde links to the semantic arguments of the process flowThrough in the ontology.

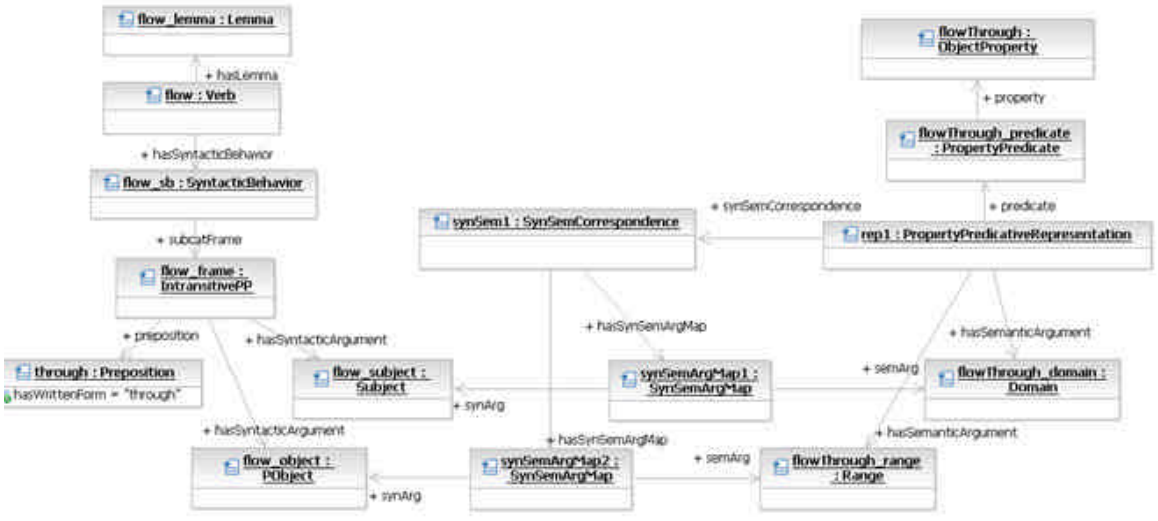


Figure 2: LexInfo representation for "flow", taken from Buitelaar et al 2009

The LexInfo extension of LMF thus can model FrameNet like information in combination with an ontology that defines processes and properties in terms of their semantic arguments.

7 Tentative cross-lingual rigidity validation

The KYOTO framework makes it possible to carry out a unique empirical validation of the ontological meta properties of the domain specific concepts. The basic methodology in KYOTO is that ontological properties are derived from linguistic observations, either through searches for non-rigidity patterns using the Rudify tool for each language or through user feedback on the potential ontology concepts in the Kyoto editing environment. In both cases, the terms in each language are the starting point, where the general schema is shown in the next figure:

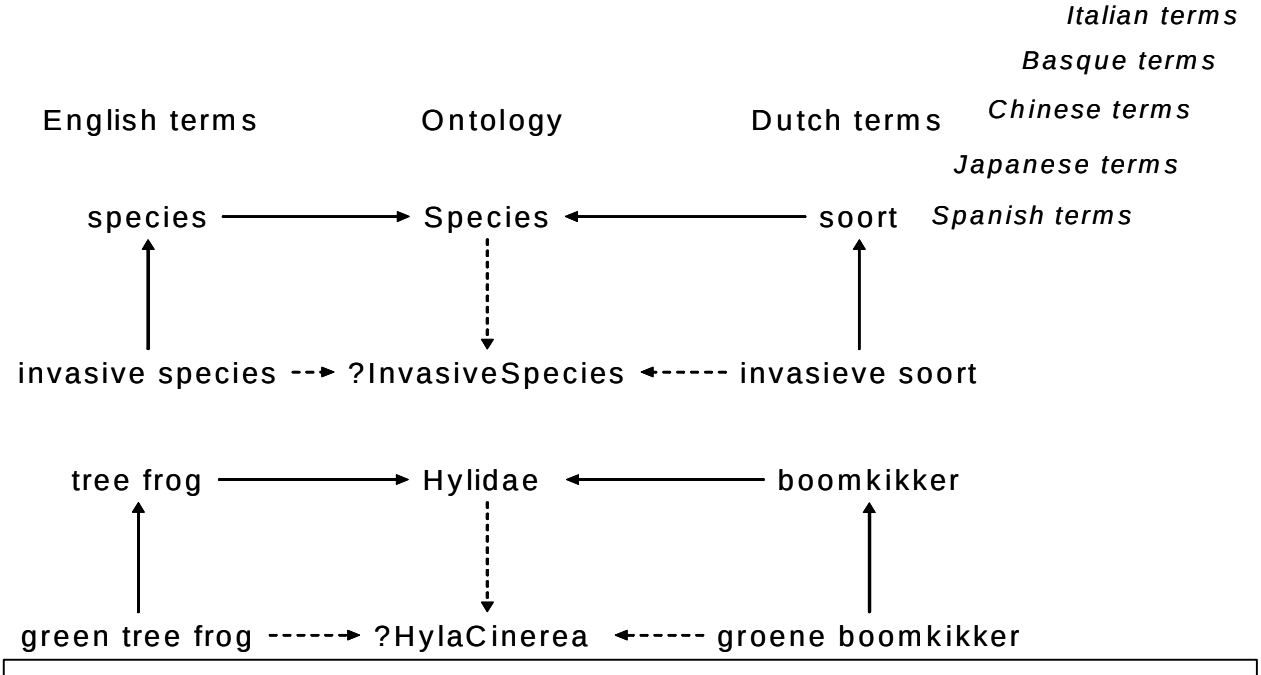


Figure 3: General schema for cross-lingual validation

In this figure, two cases for ontologization are shown:

1. a new concept InvasiveSpecies is proposed as a subclass of Species
2. a new concept HylaCinerea is proposed as a subclass of Hylidae

In this example, the concepts are proposed through terms in Dutch and English. Each term has an automatically extracted parent relation which is already present in the database. The concept represented by the term is linked to an existing ontology concept via this parent relation. These relations are represented as complete arrows.

Given the complete arrows, the system will create a proposal to extend the ontology hierarchy with a new subclass. However, the status of the subclass as a rigid or non-rigid concept needs to be validated through the linguistic use of the corresponding term in each language. The Rudify tool will yield many matches for non-Rigidity for the first case of InvasiveSpecies both in English and in Dutch:

English Google hits:

Garlic mustard *has become an invasive species* in temperate forests
the Colorado potato beetle started eating potato plants. It now *has become an invasive species*
Boa constrictor *has become an invasive species*

Dutch Google hits

hebben ontwikkeld tot een *invasieve soort (have developed to an invasive species)*
ontwikkelt tot een *invasieve soort (develops to an invasive species)*
een exoot soms gaat "woekeren" en zich ontwikkelt tot een *invasieve soort (an exotic sometimes starts to explore amd develops to an imvasive species)*

In addition, the domain experts in different language communities will be asked to confirm hypothesis regarding rigidity by answering questions in their respective languages:

English interview:

Is Garlic mustard always an invasive species?
Are all invasive species always a type of species?

Dutch interview:

Zijn exoten altijd invasieve soorten? (Are exotics always invasive species?)
Zijn invasieve soorten altijd een type soort? (Are invasive species always a type of species?)

Answering ‘no’ to the first question confirms the Rudify output that InvasiveSpecies is not a rigid concept and, therefore, should not be added as a subclass of species. Answering ‘yes’ to the second question, in this particular case, confirms that species are the kind of entity that can take on the role of an invasive species.

The case of HylaCinerea is different from invasive species insofar as, Rudify will generate no resulting matches for non-rigidity for HylaCinerea, both for the Dutch (groene boomkikker) and English (green tree frog) patterns. However, we expect that the interviews of the experts in each language are likely to generate consistent answers in both languages:

English interview:

Is a green tree frog always a type of tree frog?

Dutch interview

Is een groene boomkikker altijd een type boomkikker?

Likewise, we can conclude that HylaCinerea is therefore a rigid subtype of Hylidia.

Verifying that the proposals for ontological extensions in each language correspond to each other is a crucial part of this process. For example, we need to confirm that the English 'invasive species' and the Dutch 'invasieve soort' refer to the same concept. There are basically two ways to confirming this:

1. terms can be aligned though some other data source such as Latin species names or other multilingual thesauri;
2. ontological constraints derived from both terms can be matched across the languages;

The first case can, for instance, be applied to specific species. Existing multilingual thesauri and taxonomies can be used to propose mappings, for example, through their shared Latin names or through cross-links in DBpedia.

The second case can be applied to terms such as 'invasive species', where we need to learn what it means for a species to be an invasive species. This may require that other terms such as *invading* are added first to the database and their corresponding concepts to the ontology. We also expect that the term database will provide a rich set of properties and relations that apply to 'invasive species'. For example, the system can learn which specific species are called invasive in a particular language, e.g. certain types of ants or plants. The data collected for specific languages can then be compared through any available set of equivalence relations that are already present in the database: i.e. whenever these ants or plants are already aligned across these languages. On the basis of these mappings, the system can provide a list of concepts, ranked by the amount of overlap. These concepts would already exist in the ontology and are potential concepts to which the English term 'invasive species' may correspond.

By repeating this process for many different languages, it is possible to acquire a cross-validation of ontological meta properties for these new concepts. Likewise, we can differentiate portions of the ontology in terms of the cross-lingual and cross-cultural consistency.[12]

The above process can be applied to any state of the domain output that is generated in the community, with or without consulting the experts in the domain. For the current core-ontology, we are now acquiring the initial data to carry out experiments. The BCs described in section 4 and KYOTO deliverable D6.2 Central Ontology version 1, have been based on the concepts in a number of different wordnets. They cover all the nominal concepts in the English WordNet3.0, and we have constructed manual mappings from these concepts to synsets in the other languages. This data set provides a very precise equivalence mapping across these languages for the most important and basic concepts. The core-ontology itself has been derived by manually mapping English synsets onto a version of DOLCE that has been manually modified according to ontological principles , commitments, and distinctions made in DOLCE.

In a next phase, we will independently apply the Rudify method to the equivalent synsets in each of the related languages.

At the time of writing this draft deliverable, the data for this comparison are still being generated. Therefore, this comparison has not yet been carried out. We plan to update this deliverable soon, when the data becomes available.

[12] The formalism for representing ontological data (OWL-DL) needs to be extended to include the results of validating ontological.

8 References

Agirre E. and D. Martinez. (2002) *Integrating selectional preferences in wordnet*. In Proceedings of the first International WordNet Conference in Mysore, India.

Alfonseca E. and S. Manandhar (2002) An unsupervised method for general named entity recognition and automated concept discovery. In Proceedings of the 1st International Conference on General WordNet, Mysore, India.

Álvez J., Atserias J., Carrera J., Climent S., Laparra E., Oliver A. and Rigau G. (2008) *Complete and Consistent Annotation of WordNet using the Top Concept Ontology*. 6th international conference on Language Resources and Evaluation, LREC'08, Marrakesh, Morroco.

Atserias J., Villarejo L., Rigau G., Agirre E., Carroll J., Magnini B., Vossen P. (2004) *The MEANING Multilingual Central Repository*. In Proceedings of the Second International Global WordNet Conference (GWC'04). ISBN 80-210-3302-9. Brno, Czech Republic.

Atserias J., Rigau G., Villarejo L. (2004) *Spanish WordNet 1.6: Porting the Spanish Wordnet across Princeton versions*. LREC'04. ISBN 2-9517408-1-6. Lisboa.

Auer S., Bizer C., Kobilarov G., Lehmann J., Cyganiak R., Ives Z. 2007. Dbpedia: A Nucleus for a Web of Open Data. 6th International Conference on the Semantic Web. Busan, Korea.

Buitelaar, Paul, Philipp Cimiano, Peter Haase, and Michael Sintek, 2009, Towards Linguistically Grounded Ontologies,

Buitelaar, Paul, Philipp Cimiano, Peter. Haase and Michael Sintek Towards, 2008, Linguistically Grounded Ontologies (Technical Report), Knowledge Management Dept. & Competence Center Semantic Web, DFKI, Germany.

Buitelaar, Paul, Thierry Declerck, Anette Frank, Stefania Racioppa, Malte Kiesel, Michael Sintek, Ralf Engel, Massimo Romanelli, Daniel Sonntag, Berenike Loos, Vanessa Micelli, Robert Porzel, and Philipp Cimiano, 2006a, Linginfo: Design and applications of a model for the integration of linguistic information in ontologies. In Proceedings of OntoLex06, a Workshop at LREC, 2006.

Buitelaar, Paul, Michael Sintek, and Malte Kiesel, 2006b, A lexicon model for multilingual/multimedia ontologies. In Proceedings of the 3rd European Semantic Web Conference (ESWC06), 2006.

Cimiano, Philipp, Peter Haase, Matthias Herold, Matthias Mantel, and Paul Buitelaar, 2007, Lexonto: A model for ontology lexicons for ontologybased nlp. In Proc. of the OntoLex (From Text to Knowledge: The Lexicon/Ontology Interface) workshop at ISWC07 (International Semantic Web Conference), 2007.

Daudé J., Padró L. and Rigau G., 2000. Mapping WordNets using Structural Information, Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics ACL'00. Hong Kong, China.

Daudé J., Padró L. and Rigau G., A Complete WN1.5 to WN1.6 Mapping, Proceedings of NAACL Workshop "WordNet and Other Lexical Resources: Applications, Extensions and Customizations". Pittsburg, PA, United States, 2001.

Daudé J., Padró L. and Rigau G., Validation and Tuning of Wordnet Mapping Techniques*. Proceedings of the International Conference on Recent Advances on Natural Language Processing (RANLP'03). ISBN 954-90906-6-3. Borovets, Bulgaria. 2003.*

Farreres X., Rigau G. and Rodríguez H., (1998) *Using WordNet for Building WordNets*. Proceedings of COLING-ACL Workshop "Usage of WordNet in Natural Language Processing Systems". Montreal, Canada.

Fellbaum, C. (ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA, The MIT Press.

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., Schneider, L. 2002. Sweetening Ontologies with DOLCE In A. Gómez-Pérez, V.R. Benjamins (eds.) Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web, 13th International Conference, EKAW 2002, Siguenza, Spain, October 1-4, 2002, Springer Verlag, pp. 166-181

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., 2003. Sweetening WORDNET with DOLCE. AI Magazine Volume 24(3). 13 – 24.

Guarino, Nicola and Chris Welty. 2002. Evaluating Ontological Decisions with OntoClean. *Communications of the ACM.* **45**(2):61-65. New York:ACM Press.

Kilgarriff, Adam, "Googleology is Bad Science" In: Computational Linguistics, vol. 33, no. 1 (March 2007), pp. 147-151.

A. Lenci, N. Bel, F. Busa, N. Calzolari, E. Gola, M. Monachini, A. Ogonowski, I. Peters, W. Peters, N. Ruimy, M. Villegas, and A. Zampolli. 2000. Simple: A general framework for the development of multilingual lexicons. International Journal of Lexicography, 2000.

Magnini, B., and Cavaglià, G. (2000). *Integrating subject field codes into WordNet*. Proceedings of the Second International Conference Language Resources and Evaluation Conference (LREC), Athens, Greece, 1413–1418.

Matuszek C, Cabral J., Witbrock M., DeOliveira J. 2006. An Introduction to the Syntax and Content of Cyc. In Proceedings of the 2006 AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering, Stanford, CA, March 2006.

Mihalcea R. and D. Moldovan. (2001) *Extended wordnet: Progress report*. In Proceedings of NAACL Workshop on WordNet and Other Lexical Resources, Pittsburgh, PA.

Niles, I. and Pease, A. (2003) *Mapping WordNet to the Suggested Upper Merged Ontology*. Proceedings of the International Conference on Information and Knowledge Engineering, Las Vegas, Nevada.

Suchanek F., Kasneci G. and Weikum G. 2007. YAGO: A Core of Semantic Knowledge. In proceedings of the 16th international World Wide Web conference (WWW'07).

Toshio Yokoi. 1995. The EDR electronic dictionary. Communications of the ACM archive. 38(11), 42-44.

Völker et al. (2005): Automatic Evaluation of Ontologies (AEON). In: Gil et al.: Proceedings of the 4th International Semantic Web Conference (ISWC2005), volume 3729 of LNCS, pp. 716-731, Berlin/Heidelberg. (http://www.aifb.uni-karlsruhe.de/WBS/ysu/publications/2005_iswc_aeon.pdf)

Völker et al. (2008): AEON: Â- An approach to the automatic evaluation of ontologies. In: Applied Ontology 3, pp. 41Â---62 (http://www.aifb.uni-karlsruhe.de/WBS/jvo/publications/aeon_jao_2008.pdf

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

Vossen, P., I.Maks, R. Segers and H. van der Vliet (2008). *Integrating Lexical Units, Synsets, and Ontology in the Cornetto Database*. In Proceedings of LREC-2008, Marrakech, Morocco, 2008.

9 Appendix: Rudify Training Data

Rigid Terms and Associated WordNet3.0 Synsets

algebra:::ALGEBRA
agate:::AGATE
beach:::BEACH
boyhood:::BOYHOOD
caffeine:::CAFFEINE__CAFFEIN
aluminum:::ALUMINUM__ALUMINIUM__AL__ATOMIC_NUMBER_13
antler:::ANTLER
evergreen:::EVERGREEN__EVERGREEN_PLANT
alkaloid:::ALKALOID
cerebrum:::CEREBRUM
freckle:::FRECKLE
gaggle:::GAGGLE
fungus:::FUNGUS
ankle:::ANKLE__ANKLE_JOINT__MORTISE_JOINT__ARTICULATIO_TALOCRURALIS
beanbag:::BEANBAG
conjunctiva:::CONJUNCTIVA
animal:::ANIMAL__ANIMATE_BEING__BEAST__BRUTE__CREATURE__FAUNA
carbohydrate:::CARBOHYDRATE__SACCHARIDE__SUGAR
nitrogen:::NITROGEN__N__ATOMIC_NUMBER_7
kangaroo:::KANGAROO
melanin:::MELANIN
acorn:::ACORN
beige:::BEIGE__ECRU
begonia:::BEGONIA
biceps:::BICEPS
graphite:::GRAPHITE__BLACK_LEAD__PLUMBAGO
binomial:::BINOMIAL
gorilla:::GORILLA__GORILLA_GORILLA
gene:::GENE
semicircle:::SEMICIRCLE__HEMICYCLE
biography:::BIOGRAPHY__LIFE__LIFE_STORY__LIFE_HISTORY
hug:::HUG__CLINCH
adrenalin:::ADRENALINE__EPINEPHRINE__EPINEPHRIN__ADRENALIN
cholesterol:::CHOLESTEROL
catalpa:::CATALPA__INDIAN_BEAN
chromosome:::CHROMOSOME
chordate:::CHORDATE
chloroform:::CHLOROFORM__TRICHLOROMETHANE
bicuspid:::PREMOLAR__BICUSPID
aphid:::APHID
corticoid:::CORTICOSTEROID__CORTICOID
candida:::CANDIDA
hormone:::HORMONE__INTERNAL_SECRETION
velvet:::VELVET
aphid:::APHID
adrenalectomy:::ADRENALECTOMY__SUPRARENALECTOMY
adenoid:::ADENOID__PHARYNGEAL_TONSIL
chamomile:::CHAMOMILE__CAMOMILE__CHAMAEMELUM_NOBILIS__ANTHEMIS_NOBILIS
leather:::LEATHER
walrus:::WALRUS__SEAHORSE__SEA_HORSE

Non-Rigid Terms and Associated WordNet3.0 Synset Labels

golfer:::GOLFER__GOLF_PLAYER__LINKSMAN
widower:::WIDOWER__WIDOWMAN
advisor:::ADVISER__ADVISOR__CONSULTANT
consumer:::CONSUMER
alarmist:::ALARMIST
coauthor:::COAUTHOR__JOINT_AUTHOR
classmate:::SCHOOLMATE__CLASSMATE__SCHOOLFELLOW__CLASS_FELLOW
legislator:::LEGISLATOR
chaperone:::CHAPERON__CHAPERONE
cyclist:::CYCLIST__BICYCLIST
niece:::NIECE
juror:::JUROR__JURYMAN__JURYWOMAN
boyfriend:::BOYFRIEND__FELLOW__BEAU__SWAIN__YOUNG_MAN
accomplice:::ACCESSORY__ACCOMPLICE__ACCESSARY
hostage:::HOSTAGE__SURETY
dishrag:::DISHRAG__DISHCLOTH
allergist:::ALLERGIST
amputee:::AMPUTEE
conspirator:::CONSPIRATOR__COCONSPIRATOR__PLOTTER__MACHINATOR
aunt:::AUNT__AUNTIE__AUNTY
doorman:::DOORKEEPER__DOORMAN__DOOR_GUARD__HALL_PORTER__PORTER__GATEKEEPER
lender:::LENDER__LOANER
ancestor:::ANCESTOR__ASCENDANT__ASCENDENT__ANTECEDENT
mediator:::MEDIATOR__GO-BETWEEN__INTERMEDIATOR__INTERMEDIARY
vegan:::VEGAN
counterfeit:::IMITATION__COUNTERFEIT__FORGERY
container:::CONTAINER
hangout:::HAUNT__HANGOUT__RESORT__REPAIR__STAMPING_GROUND
nonprofit:::NONPROFIT__NON-PROFIT-MAKING
habitat:::HABITAT
median:::MEDIAN
exemplar:::EXEMPLAR__EXAMPLE__MODEL__GOOD_EXAMPLE
workroom:::WORKROOM
drug:::DRUG
auditorium:::AUDITORIUM
gateway:::GATEWAY
caricature:::CARICATURE__IMITATION__IMPERSONATION
campsite:::CAMPSITE__CAMPGROUND__CAMPING_SITE__CAMPING_GROUND__BIVOUAC__ENCAMPMENT__CAMPING_AREA
hometown:::HOMETOWN
setback:::REVERSE__REVERSAL__SETBACK__BLOW
bronchodilator:::BRONCHODILATOR
birdseed:::BIRD_FOOD__BIRDSEED
platitude:::PLATITUDE__CLICHE__BANALITY__COMMONPLACE__BROMIDE bedtime:::BEDTIME
memento:::MEMENTO__SOUVENIR
euphemism:::EUPHEMISM
allergen:::ALLERGEN
victory:::VICTORY__TRIUMPH
fad:::FAD__CRAZE__FUROR__FURORE__CULT__RAGE
antidepressant:::ANTIDEPRESSANT__ANTIDEPRESSANT_DR