# Towards the MEANING Top Ontology: Sources of Ontological Meaning

| | |
|---|---|
| **Document Number** | WP4.5 |
| **Project ref.** | IST-2001-34460 |
| **Project Acronym** | MEANING |
| **Project full title** | Developing Multilingual Web-scale Language Technologies |
| **Project URL** | http://www.lsi.upc.es/~nlp/meaning/meaning.html |
| **Availability** | Public |
| **Authors:** Jordi Atserias (UPC), Salvador Climent (UOC), German Rigau (UPV/EHU) | |

INFORMATION SOCIETY TECHNOLOGIES

| | |
|---|---|
| **Project ref.** | IST-2001-34460 |
| **Project Acronym** | MEANING |
| **Project full title** | Developing Multilingual Web-scale Language Technologies |
| **Security (Distribution level)** | Public |
| **Contractual date of delivery** | March 2004 |
| **Actual date of delivery** | April 2, 2004 |
| **Document Number** | WP4.5 |
| **Type** | Report |
| **Status & version** | v 1 Draft |
| **Number of pages** | 18 |
| **WP contributing to the deliberable** | WP 5 |
| **WPTask responsible** | German Rigau |
| **Authors** | Jordi Atserias (UPC), Salvador Climent (UOC), German Rigau (UPV/EHU) |
| **Other contributors** | |
| **Reviewer** | |
| **EC Project Officer** | Evangelia Markidou |

**Authors:** Jordi Atserias (UPC), Salvador Climent (UOC), German Rigau (UPV/EHU)

**Keywords:** Ontologies, WordNet, EuroWordNet

**Abstract:** This working paper describes the initial research steps towards the Top Ontology for the Multilingual Central Repository (Mcr) built in the Meaning project. The current version of the Mcr integrates five local wordnets plus four versions of Princeton's English WordNet, three ontologies and hundreds of thousands of new semantic relations and properties automatically acquired from corpora. In order to maintain compatibility among all these heterogeneous knowledge resources, it is fundamental to have a robust and advanced ontological support. This paper studies the mapping of main Sources of Ontological Meaning onto the wordnets and, in particular, the current work in mapping the EuroWordNet Top Concept Ontology.

# Contents

# 1 Introduction: the Meaning Project

meaning[1] [Rigau *et al.*, 2002] is a UE-funded project (IST-2001-34460) which has as one of its major goals the integration of several large-scale knowledge resources. meaning has designed a Multilingual Central Repository (Mcr) to act as a multilingual interface for integrating and distributing all the knowledge acquired in the project [Atserias *et al.*, 2004]. The Mcr follows the model proposed by the EuroWordNet project (Ewn): a multilingual lexical database with wordnets for several languages.

The Ewn architecture includes the Inter-Lingual-Index (Ili), a preliminary Domain Ontology (Do) and a Top Concept Ontology (Tco) [Vossen, 1998]. The Ili consists of a flat list of records that interconnect synsets across wordnets. During the Ewn Project around 1000 Ili-Records were selected as Base Concepts (Bc) and connected to the Tco.

Using the Ili, wordnets in the Mcr are interconnected so that it is possible to go from word meanings in one language or particular wordnet to their equivalents in other languages or wordnets.

In Ewn, the Ili was enhanced, enriched and structured by two separate ontologies:

- the **Top Concept ontology** (Tco), which is a hierarchy of language-independent concepts, reflecting important semantic distinctions, e.g. Object, Substance, Location, Dynamic;
- the **Domain ontology** (Do), which is a hierarchy of domain labels, which are knowledge structures grouping meanings in terms of topics or scripts, e.g. Transport, Sports, Medicine, Gastronomy;

The main purpose of the Tco is to provide a common framework for all the wordnets. It consists of 63 basic semantic distinctions that classify a set of Ili-records connected to Ewn Bc which represents the most important concepts in the different wordnets.

The current version of the Mcr uses the set of Princeton WordNet (Wn) 1.6 synsets as Ili. Initially most of the knowledge to be uploaded into the Mcr has been derived from Wn (automatic selectional preferences acquired from SemCor and BNC) and the Italian wordnet and the MultiWordNet Domains, both developed at IRST are using WordNet 1.6 as Ili [Bentivogli *et al.*, 2002; Magnini and Cavagli, 2000]. This option also minimises side effects with other European initiatives (Balkanet, EuroTerm, etc.) and wordnet developments around Global WordNet Association. However, the Ili for Spanish, Catalan and Basque wordnets, the Ewn Tco and the associated Bc were based on WordNet 1.5 [Atserias *et al.*, 1997; Benítez *et al.*, 1998].

After this short introduction, section 2 describes the main ontological resources used in Meaning. Section 3 illustrates with an example how we can detect errors and inconsistences with an integrated framework. Section 4 presents the inheritance mechanism used to expand the Tco properties. In section 5 we present the semi–automatic approach we

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

plan to follow to perform consistency checking of the Tco related to the diverse conceptual information used in Meaning. Section 6 illustrates the potential capabilities of an enhanced Tco with an example. Finally, section 7 provides some concluding remarks.

# 2 Sources of Ontological Meaning

Although wordnets and ontologies are both graphs connecting concepts, they are different in nature: while wordnets build concepts upon lexical units of a particular language, nodes in ontologies are claimed to be language- independent concepts. Wordnets can be straightforwardly used for NLP tasks. On the contrary, ontologies, although being meaningful constructs, can not be straightforwardly used for NLP unless they are associated to linguistic units and structures.

Moreover, different ontologies usually are designed using different theoretical grounds; e.g. while Sumo incorporates previous ontologies and insights by Sowa, Pierce, Russell and Norvig and others, the Tco is based on more linguistic grounds: Lyons, Vendler, Verkuyl and Pustejovsky. Therefore, although different ontologies can be comparable, it would take a great theoretical effort to merge all of them in a unique standard and comprehensive construct to be consistently associated to Wn.

For this reason, in meaning we intend to adopt a hybrid and simple approach: to build the meaning To, different Sources of Ontological Meaning (Som) are assigned to language-independent Ili–records so that they can be mapped to Wn concepts and expanded throughout them using its internal semantic relations. The different Som do not need to be equivalent nor even compatible as they will stand as independent information. Besides, no claim of completion will be made.
Currently, Mcr integrates through Ili different Som:

1. An upgraded version of the Ewn Base Concepts (Bc)
2. An upgraded version of the Ewn Top Concept Ontology (Tco)
3. The WordNet Domains Ontology (Do) [Magnini and Cavagli, 2000], a hierarchy of 165 domain labels
4. The Suggested Upper Merged Ontology, Sumo [Niles and Pease, 2001]
5. WN Semantic Files (Sf), corresponding to lexicographical files from wordnet, e.g. noun.animal, verb.possession, etc.

## 2.1 The EuroWordNet Base Concepts

The EuroWordNet Base Concepts were selected manually to cover the most important concepts of the languages involved in the project [Vossen, 1998]. The main characteristic of the Base Concepts is their importance in the wordnets. According to our pragmatic point of view, a concept is important if it is widely used, either directly or as a reference for other widely used concepts. Importance is thus reflected in the ability of a concept to function as an anchor to attach other concepts or properties. This anchoring capability was defined in terms of three operational criteria that can be automatically applied to the available resources:

1. the number of relations (general or limited to hyponymy).

2. being widely used by several languages

3. high position of the concept in a hierarchy

The procedure of selecting the EuroWordNet Base Concepts and the Top Ontology is discussed in [Vossen, 1998]. The final set of common Base Concepts totalized 1030 WordNet 1.5 synsets.

In the first MEANING cycle, the Base Concepts from WN1.5 have been mapped to WN1.6. After a manual revision and expansion to all WN1.6 top beginners, the resulting BC for WN1.6 totalized 1,601 ILI-records. In that way, the new version of BC covers the complete hierarchy of ILI-records.

However, the definition of Base Concepts in EuroWordNet could not use the sense frecuency information currently available in the Princeton WordNet[2].

We suggest for future rounds to devise a simple and fully automatic method to derive the Base Concepts from the information contained into the MCR following the operational criteria defined above. However, we consider that the BCs should be general enough (being in the high part of the hierarchy) but also particular enough (being in the lower part of the hierarchy) to represent the main characteristics of each concept represented in the MCR.

Table 1 presents the hypernym chain for all the senses of the noun *church* in WN1.6. For each synset we show the result of summing up all the sense frequency counts appearing in SemCor (#occur.)[3] and the total number of direct relations (#rel.). Having calculated these two numbers for each synset (both representing the first two criteria defined above), a very simple arithmetic operations can be devised to obtain the BC for these particular synsets.

We suggest to study the following bottom-up approach to derive the whole set of BCs. Following the hypernym chain, we can obtain for both number of occurrences and number of relations, the local maxima of each synset. For instance, for **chain_1** the first local maximum for the #occur. corresponds to organization_2 (with 729 occurrences), and for the #rel. corresponds to faith_3 (with 12 relations). For **chain_2** the first local maximum for the #occur. corresponds to construction_3 (with 68 occurrences), and for the #rel. corresponds to building_1 (with 79 relations). Finally, for **chain_3** the first local maximum for the #occur. corresponds to service_3 (with 243 occurrences), and for the #rel. corresponds to religious_ceremony_1 (with 11 relations). Obviously, both criteria can also be combined. Furthermore, we suggest to collect all local maxima for each leaf of the WN hierarchies. All these local maxima will constitute the new Base Concepts of the MCR.

[2] WordNet started to contain sense frecuency information derived from SemCor in version 1.6

[3] For the rest of languages there is not available a sense tagged corpora for all words.

| #occur. | #rel. | offset | synset |
|---:|---:|---|---|
| 2338 | 18 | 00017954-n | group_1,grouping_1 |
| 0 | 19 | 05962976-n | social_group_1 |
| 729 | 37 | 05997592-n | organisation_2,organization_1 |
| 30 | 10 | 06002286-n | establishment_2,institution_1 |
| 15 | 12 | 06023733-n | faith_3,religion_2 |
| 62 | 5 | 06024357-n | Christianity_2,**church_1**,Christian_church_1 |
| 11 | 14 | 00001740-n | entity_1,something_1 |
| 51 | 29 | 00009457-n | object_1,physical_object_1 |
| 1 | 39 | 00011937-n | artifact_1,artefact_1 |
| 68 | 63 | 03431817-n | construction_3,structure_1 |
| 50 | 79 | 02347413-n | building_1,edifice_1 |
| 0 | 11 | 03135441-n | place_of_worship_1,house_of_prayer_1,house_of_God_1,house_of_worship_1 |
| 59 | 19 | 02438778-n | **church_2**,church_building_1 |
| 25 | 20 | 00017487-n | act_2,human_action_1,human_activity_1 |
| 611 | 69 | 00261466-n | activity_1 |
| 2 | 5 | 00662816-n | ceremony_3 |
| 0 | 11 | 00663517-n | religious_ceremony_1,religious_ritual_1 |
| 243 | 7 | 00666638-n | service_3,religious_service_1,divine_service_1 |
| 11 | 1 | 00666912-n | **church_3**,church_service_1 |

Table 1: Hypernym chain for all senses of the noun church in Wn1.6

The resulting new Base Concepts can then be used to attach consistently new ontological properties defined into the Top Concept Ontology Tco.

## 2.2 The EuroWordNet Top Concept Ontology

The EuroWordNet Top Ontology consists of 63 higher-level concepts, excluding the top. Following [Lyons, 1977] EuroWordNet distinguish at the first level 3 types of entities:

- **1stOrderEntity** Any concrete entity (publicly) perceivable by the senses and located at any point in time, in a three-dimensional space, e.g.: vehicle, animal, substance, object.

- **2ndOrderEntity** Any Static Situation (property, relation) or Dynamic Situation, which cannot be grasped, heard, seen, felt as an independent physical thing. They can be located in time and occur or take place rather than exist, e.g.: happen, be, have, begin, end, cause, result, continue, occur..

- **3rdOrderEntity** Any unobservable proposition which exists independently of time and space. They can be true or false rather than real. They can be asserted or denied, remembered or forgotten, e.g.: idea, thought, information, theory, plan.

According to Lyons, 1stOrderEntities are publicly observable individual persons, animals and more or less discrete physical objects and physical substances. They can be located at any point in time and in, what is at least psychologically, a three-dimensional space. The 2ndOrderEntities are events, processes, states-of-affairs or situations which can be located in time. Whereas 1stOrderEntities exist in time and space 2ndOrderEntities occur or take place, rather than exist. The 3rdOrderEntities are propositions, such as ideas, thoughts, theories, hypotheses, that exist outside space and time and which are unobservable. They function as objects of propositional attitudes, and they cannot be said to occur or be located either in space or time. Furthermore, they can be predicated as true or false rather than real, they can be asserted or denied, remembered or forgotten, they may be reasons but not causes.

The first division of the ontology is disjoint: BCs cannot be classified as combinations of these Top Concepts. This distinction cuts across the different parts of speech in that:

- 1stOrderEntities are always (concrete) nouns (491 synsets)
- 2ndOrderEntities can be nouns, verbs or adjectives (535 synsets)
- 3rdOrderEntities are always (abstract) nouns (33 synsets).

The **1stOrderEntities** are distinguished in terms of four main ways of conceptualizing or classifying a concrete entity:

1. Origin: the way in which an entity has come about.
2. Form: as an a-morf substance or as an object with a fixed shape, hence the subdivisions Substance and Object.
3. Composition: as a group of self-contained wholes or as a part of such a whole, hence the subdivisions Part and Group.
4. Function: the typical activity or action that is associated with an entity.

These classes are comparable with Aristotle's Qualia roles as described in Pustejovsky's Generative lexicon, (the Agentive role, Formal role, Constitutional role and Telic Role respectively: [Pustejovsky, 1995] but are also based on our empirical findings to classify the BCs. BCs can be classified in terms of any combination of these four roles. As such the top-concepts function more as features than as ontological classes. Such a systematic cross-classification was necessary because the BCs represented such diverse combinations (e.g. it was not possible to limit Function or Living only to Object).

The main-classes are then further subdivided, where the subdivisions for Form and Composition are obvious given the above definition, except that Substance itself is further subdivided into Solid, Liquid and Gas. In the case of Function the subdivisions are based only on the frequency of BCs having such a function or role. In principle the number of roles is infinite but the above roles appear to occur more frequently in the set of common Base Concepts.

Finally, a more fine-grained subdivision has been made for Origin, first into Natural and Artifact. The category Natural covers both inanimate objects and substances, such as stones, sand, water, and all living things, among which animals, plants and humans. The latter are stored at a deeper level below Living. The intermediate level Living is necessary to create a separate cluster for natural objects and substances, which consist of Living material (e.g. skin, cell) but are not considered as animate beings. Non-living and Natural objects and substances, such as natural products like milk, seeds, fruit, are classified directly below Natural.

As suggested, each Bc that is a 1stOrderEntity is classified in terms of these main classes. However, whereas the main-classes are intended for cross-classifications, most of the subdivisions are disjoint classes: a concept cannot be an Object and a Substance, or both Natural and Artifact. This means that within a main-class only one subdivision can be assigned. Consequently, each Bc that is a 1stOrderEntity has at least one up to four classifications:

- fruit:
  - Comestible (Function)
  - Object (Form)
  - Part (Composition)
  - Plant (Natural, Origin)

As explained above, **2ndOrderEntities** can be referred to using nouns and verbs (and also adjectives or adverbs) denoting static or dynamic Situations, such as birth, live, life, love, die and death. All 2ndOrderEntities are classified using two different classification schemes, which represent the first division below 2ndOrderEntity:

- the SituationType: the event-structure in terms of which a situation can be characterized as a conceptual unit over time;
- the SituationComponent: the most salient semantic component(s) that characterize(s) a situation;

The SituationType reflects the way in which a situation can be quantified and distributed over time, and the dynamicity that is involved. It thus represents a basic classification in terms of the event-structure (in the formal tradition) or the predicate-inherent Aktionsart properties of nouns and verbs. Examples of SituationTypes are Static, Dynamic.

The SituationComponents represent a more conceptual classification, resulting in intuitively coherent clusters of word meanings. The SituationComponents reflect the most salient semantic components that apply to our selection of Base Concepts. Examples of SituationComponents are: Location, Existence, Cause.

Typically, SituationType represents disjoint features that cannot be combined, whereas it is possible to assign any range or combination of SituationComponents to a word meaning. Each 2ndOrder meaning can thus be classified in terms of an obligatory but unique SituationType and any number of SituationComponents.

Since the number of **3rdOrderEntities** among the BCs was limited compared to the 1stOrder and 2ndOrder Entities we have not further subdivided them.

Base Concepts classified as 3rdOrderEntities: theory; idea; structure; evidence; procedure; doctrine; policy; data point; content; plan of action; concept; plan; communication; knowledge base; cognitive content; know-how; category; information; abstract; info;

In MEANING the TCO has been uploaded in for steps as explained in section 3.

The original set of Bc from Ewn based on Wn1.5 totalized 1,030 Ili–records. Now, the Bc from Wn1.5 have been mapped to Wn1.6. After a manual revision and expansion to all Wn1.6 top beginners, the resulting Bc for Wn1.6 totalized 1,601 Ili-records. In that way, the new version of Bc covers the complete hierarchy of Ili-records.

## 2.3 The MultiWordNet Domains

The initial EuroWordNet design included a Domain ontology. However, only the *Computer Domain* was included into the EuroWordNet database.

Information brought by Domain Labels is complementary to what is already in WordNet. First of all Domain Labels may include synsets of different syntactic categories: for instance MEDICINE groups together senses from nouns, such as doctor and hospital, and from Verbs such as operate.

Second, a Domain Label may also contain senses from different WordNet subhierarchies (i.e. deriving from different *unique beginners* or from different *lexicographer files*. For example, the SPORT contains senses such as athlete, deriving from life form, game equipment, from physical object, sport from act, and playing field, from location.

Meaning will use WordNet Domains [Magnini and Cavagli, 2000] which were partially derived from the Dewey Decimal Classification [4]. WordNet Domains is a hierarchy of 165 Domain Labels associated to WordNet 1.6 synsets (see Working Paper 4.1 for further details).

## 2.4 Suggested Upper Merged Ontology (Sumo)

Sumo[5] [Niles and Pease, 2001] is being created as part of the IEEE Standard Upper Ontology Working Group. The goal of this Working Group is to develop a standard upper ontology that will promote data interoperability, information search and retrieval, automated inference, and natural language processing. There is a complete set of mappings from WordNet 1.6 synsets to Sumo: nouns, verbs, adjectives, and adverbs.

[4] http://www.oclc.org/dewey

[5] http://ontology.teknowledge.com/

Sumo consists of a set of concepts, relations, and axioms that formalize an upper ontology. An upper ontology is limited to concepts that are meta, generic, abstract or philosophical, and hence are general enough to address (at a high level) a broad range of domain areas. Concepts specific to particular domains are not included in the upper ontology, but such an ontology does provide a structure upon which ontologies for specific domains (e.g. medicine, finance, engineering, etc.) can be constructed.

The current version of Sumo consists of 1,019 terms (all of them connected to WordNet 1.6 synsets, 4,181 axioms and 822 rules.

We think that further investigation is needed with respect comparing both Sumo and the EuroWordNet Top Ontology. For instance, the typology of processes in the Sumo was inspired by Beth Levin's well-received work entitled "Verb Classes and Alternations". Among other things, this work attempts to classify over 3,000 English verbs into 48 "semantically coherent verb classes". Some of the verb classes relate to static predicates in the ontology rather than to processes, and some classes are syntactically motivated, e.g. the class of verbs that take predicative complements.

Further, the ontology also defines formally its types and relations between them in the form of axioms.

### 2.5 Wn Semantic Files

During WordNet development synsets are organized into forty-five lexicographer files based on syntactic category and logical groupings. These lexicographer files can be also seen as a coarse–grained sense distinctions or subject codes [Rigau *et al.*, 1997]. Table 2 presents the synset distribution across Semantic Files in Wn1.6. From left to right, Semantic File number (SF), Frequency, Lexicographer File and Part–of–Speech (POS).

## 3 Integration of SOM

The integration of all these Som into a single platform both demands and allows for cross-checking.

For instance, we can improve Sumo labels and WordNet Domains mappings by merging and comparing them.

To illustrate how we can detect errors and inconsistences between different types of Som, we can see in the example in table 3 that the nouns corresponding to the Sumo process Breathing has been labeled with ANATOMY domain, some verbs with MEDICINE and some adjectives with FACTOTUM, when in fact, all these senses correspond to different Part-of-Speech of the same concept.

In Meaning the Tco has been uploaded in four steps (see [Atserias *et al.*, 2003] for further details):

1. Upgrading the Wn1.5 Bc to Wn1.6

| SF | Frequency | LF | POS |
|---|---|---|---|
| 00 | 14734 | adj.all | 3 |
| 01 | 3099 | adj.pert | 3 |
| 02 | 3575 | adv.all | 4 |
| 03 | 13 | noun.Tops | 1 |
| 04 | 5373 | noun.act | 1 |
| 05 | 7295 | noun.animal | 1 |
| 06 | 9811 | noun.artifact | 1 |
| 07 | 2634 | noun.attribute | 1 |
| 08 | 1592 | noun.body | 1 |
| 09 | 2261 | noun.cognition | 1 |
| 10 | 4548 | noun.communication | 1 |
| 11 | 851 | noun.event | 1 |
| 12 | 394 | noun.feeling | 1 |
| 13 | 2378 | noun.food | 1 |
| 14 | 1832 | noun.group | 1 |
| 15 | 2124 | noun.location | 1 |
| 16 | 41 | noun.motive | 1 |
| 17 | 1050 | noun.object | 1 |
| 18 | 6410 | noun.person | 1 |
| 19 | 524 | noun.phenomenon | 1 |
| 20 | 7873 | noun.plant | 1 |
| 21 | 908 | noun.possession | 1 |
| 22 | 521 | noun.process | 1 |
| 23 | 1104 | noun.quantity | 1 |
| 24 | 371 | noun.relation | 1 |
| 25 | 300 | noun.shape | 1 |
| 26 | 2550 | noun.state | 1 |
| 27 | 2392 | noun.substance | 1 |
| 28 | 875 | noun.time | 1 |
| 29 | 495 | verb.body | 2 |
| 30 | 2006 | verb.change | 2 |
| 31 | 635 | verb.cognition | 2 |
| 32 | 1388 | verb.communication | 2 |
| 33 | 411 | verb.competition | 2 |
| 34 | 229 | verb.consumption | 2 |
| 35 | 1953 | verb.contact | 2 |
| 36 | 606 | verb.creation | 2 |
| 37 | 303 | verb.emotion | 2 |
| 38 | 1247 | verb.motion | 2 |
| 39 | 410 | verb.perception | 2 |
| 40 | 688 | verb.possession | 2 |
| 41 | 1007 | verb.social | 2 |
| 42 | 671 | verb.stative | 2 |
| 43 | 78 | verb.weather | 2 |
| 44 | 82 | adj.ppl | 3 |
|  |  |  |  |

Table 2: Semantic File distribution in Wn1.6

2. Tco properties have been assigned to Wn1.6 synsets through the Wn 1.5 to 1.6 mapping [Daudé *et al.*, 2001].

3. For those Wn1.6 Tops (synsets without any parent) that do not have any assigned property through the mapping, we assigned to them the Tco properties via a table of equivalence between Tco and Sf.

4. The resulting properties were propagated top–down through the Wn hierarchy

The original set of Bc from Ewn based on Wn1.5 totalized 1,030 Ili–records. Now, the Bc from Wn1.5 have been mapped to Wn1.6. After a manual revision and expansion to all Wn1.6 top beginners, the resulting Bc for Wn1.6 totalized 1,601 Ili-records. In that way, the new version of Bc covers the complete hierarchy of Ili-records.

# 4 Expanding Tco properties

The Ewn project only performed a complete validation of the consistency of the Tco at the Bc level.

| **Synset** | **Word** | **SUMO** | **Domain** |
|---|---|---|---|
| 00003142v | exhale | Breathing | medicine |
| 00899001a | exhaled | Breathing | factotum |
| 00263355a | exhaling | Breathing | factotum |
| 00536039n | expiration | Breathing | anatomy |
| 02849508a | expiratory | Breathing | anatomy |
| 00003142v | expire | Breathing | medicine |

Table 3: Sumo vs. Domain labels

Assuming (as the builders of Sumo and Do have done) that the ontological properties have been correctly assigned to particular synsets and that Wn defines coherent subsumption chains, an automatic process can consistently inherit all the properties through the whole hierarchy of Wn - no matter the ontology they come from.

Meaning have performed an automatic expansion of the Tco properties assigned to the Bc. That is, we enriched the complete Ili structure with features coming from the Bc by inheriting the Top Concept features following the hyponymy relationship.

This way, once ontological properties are exported to the Ili and inherited through the whole Wn Hierarchy, all concepts in a Wn will result to be assigned with a set of semantic features as in the example shown in table 4.

| **lentil_1** | |
|---|---|
| *DOMAIN* | gastronomy |
| *SF* | food |
| *SUMO* | FruitOrVegetable |
| *TCO* | Comestible ; Plant |

Table 4: lentil_1

In order to provide consistency to the inheritance process we used the following basic incompatibilities among Tco properties (furtherly expanded to their daughter concepts) which were defined inside the Ewn project:

- substance - object
- plant - animal - human - creature
- natural - artifact
- solid - liquid - gas

As the classification of Wn is not always consistent with the Tco, these incompatibilities impeded the full automatic top–down propagation of the Tco properties. That semi-automatic process resulted in a number of synsets showing non–compatible information. Specifically:

- Sticking to TCO and according to the set of incompatibilities, some TCO properties assigned by hand appeared to be incompatible with either (a) inherited information, (b) information assigned via equivalence to SF or/and even (c) other TCO properties assigned by hand.

- TCO properties, either original or inherited, are suspicious to be incompatible with other SOM.

By examining a subset of synsets, we realised that there are at least the following main sources of errors:

- Erroneous hand-made TCO mappings

- Erroneous statements of equivalence between TCO properties and SFS

- Erroneous ISA links in WN -which causes erroneous inheritance [Guarino and Welty, 2000]

- Multiple inheritance within WN can cause incompatibilities in inheritance of properties

The example shown in table 5 has incompatible information. 3rdOrderEntity can not coexist with properties only attributable to Events:

| **00660718 process_1** | |
|---|---|
| *DOMAIN* | factotum |
| *SF* | act |
| *SUMO* | IntentionalProcess |
| *TCO* | 3rdOrderEntity;Cause;Mental;Purpose |

Table 5: 00660718 process_1

# 5 Consistency checking

The procedure we will apply to solve the TCO incompatibilities is the following:

1. Hand-fixing TCO mappings where appearing incompatible properties

2. Setting inheritance–blocking–points and hand-fixing TCO mappings around these points (i.e. all involved hypernyms and hyponyms)

3. Recalculating the inheritance according to the information obtained in (1) and (2)

4. Reexamining the involved subtrees to check whether re–calculation of the inheritance produce new incompatibilities

5. Exporting the mappings and blocking–point information to the Ili.

It should be noticed that it is important to export also blocking–point information to the Ili in order to ease future correct exportation of Som's information to other wordnets, i.e. to prevent incorrect expansion of properties by inheritance. Inside a particular wordnet, when reaching a blocking point, a subsumption link can be considered as broken for ontological purposes –therefore, it will be assumed that the conceptual chain only proceeds upwards consistently to the Som (not to the hypernym synsets), via the Ili–records.

This process can be applied iteratively looking for suspicious synsets in Wn. In the first round we will check the list of 38 synsets which show incompatibility between hand-assigned Tco properties. In the second one we will check the set of Wn top beginners which only bear information mapped via the Tco–Sf table of equivalence. Third, we will check synsets showing incompatibility between information directly mapped via Tco and information mapped via the Tco–Sf table of equivalence. Last, we will check the remaining cases of incompatibility between Tco manual and inherited information.

Being more precise, for each synset in any of the subsets we will proceed as follows:

1. Fixing the properties of those synsets having contradictory Tco properties: Tco assignments are fixed in the synset and its immediate relatives (mainly hypernym and hyponyms). All these synsets will be marked as "hand–checked". The result will be correct Tco information assigned to several synsets as in the following example where, originally, non-agentive and non-intentional **00661612 stiffening_1** was inheriting all of the **00660718 process_1** properties as shown in table 6

| | |
|---|---|
| **00660718 process_1** | |
| *TCO* | Dynamic;Agentive;Purpose |
| **00661612 stiffening_1** | |
| *TCO* | Dynamic;Cause |

Table 6: 00660718 process_1 and 00661612 stiffening_1

2. For those synsets having false Wn subsumptions, we will introduce a blocking point between a pair of synsets. The result will be a list of blocking points, e.g.: between 00661612n and 00660718n.

3. We keep record of Tco–Sf erroneous equivalences, since they will be useful in the future to detect more synsets with erroneous mappings. The result will be a list of suspicious Tco–Sf equivalences, e.g.: [Tco:Agentive–Sf:ACT]

4. To study TCO–SUMO equivalences in such synsets. As in the previous step, they can be useful in the future to detect more synsets with mistaken mappings. The result will be a list of incompatible TCO–SUMO concepts, e.g.: [TCO:3rdOrderEntity–SUMO:Physical]

5. To inspect as well WN Domain assignments. The result will be a list of doubtful WN Domain assignments, e.g. 00364173n#play_3:ENTERPRISE

Following an iterative and incremental approach, the inheritance will be re-calculated, the resulting data will be re–examined, and the eventual correct information will be again uploaded into the MCR thus overwriting the pre-existent one

Although such hand–checking is extremely complex and delicate, we expect the task is affordable since critical conflicts seem to concentrate in a workable layer of synsets close to the higher part of the WN hierarchy.

# 6 TCO at work

In order to illustrate the potential capabilities of an enhanced TCO, consider the following example. Figure 1 presents a partial view of WN1.6 where solid lines represent direct connections between synsets and dotted lines represent indirect or inferred connections. Using the WN browser provided by Princeton we can ask for the direct and inherited PART-OF relations of a particular sense. A direct PART-OF relation occurs between plant_2 and plant_part_1, and and inherited PART-OF relation occurs between apple_2 and plant_part_1. However, while for succulent_1, the algorithm provides an inherited PART-OF relation to plant_part_1, for cactus_1 the algorithm do not provides any inherited relation at all. Consider now the following simple questions:

1. Does a cactus have leaves?

2. Does an orchad apple tree have leaves?

3. Does an orchad apple tree have fruits?

Obviously, this simple questions only can be answered applying a systematic inference mechanism on WN [Harabagiu and Moldovan, 1998]. What should be the correct behavior of the mechanism for inheriting correctly the PART-OF relation through the entire hierarchy of WN?

In order to test an inference mechanism on the PART-OF relation we implemented the following inference rules:

1. A has_hyperonym B and B has_hyperonym C => A has_hyperonym C

2. A has_hyponym B and B has_hyponym C => A has_hyponym C

3. A has_mero_part B and B has_mero_part C => A has_mero_part C

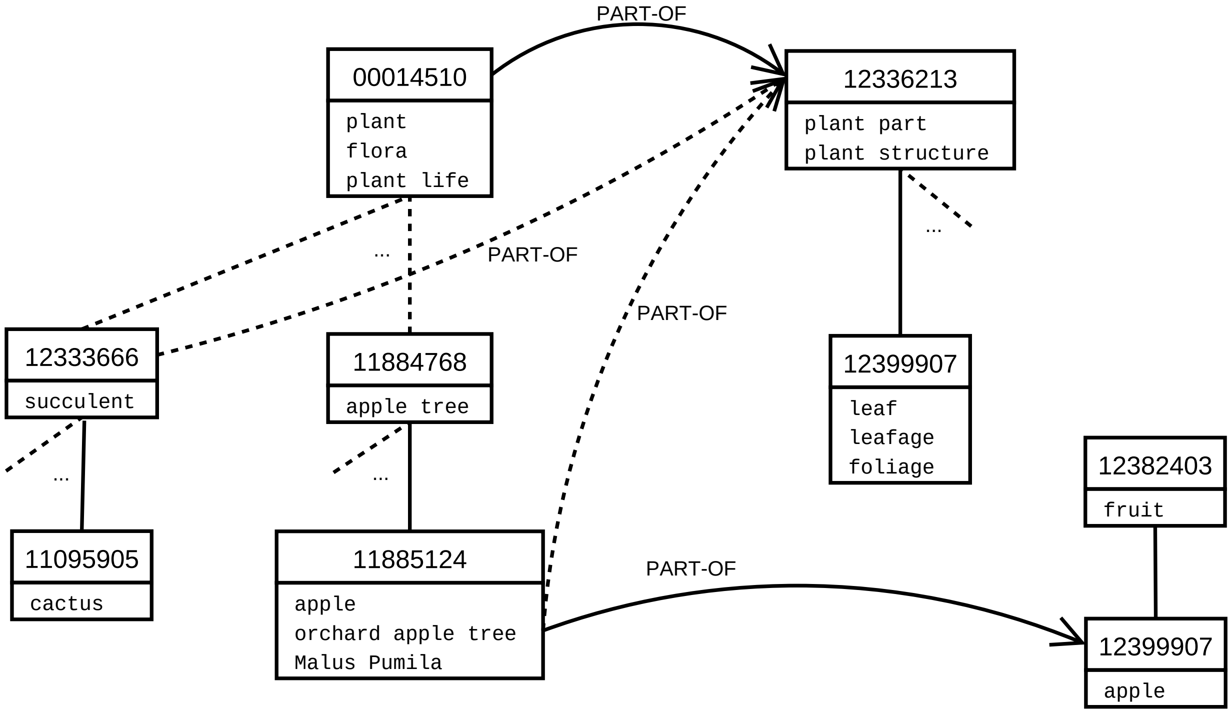

Figure 1: Partial view of Wn1.6

4. A has_hyperonym B and B has_mero_part C => A has_mero_part C

5. A has_mero_part B and B has_hyponym C => A has_mero_part C

The first two inference rules only represent the transitivity of the IS-A relation. The same holds for the third with respect PART-OF relation. The fourth inference rule will allow to inherit the PART-OF relation through an hypernym chain. For example, the relation cactus_1 has_mero_part plant_part_1. The last inference rule will allow to propagate a PART-OF relation through an hyponymy chain. For example, the relation plant_2 has_mero_part leaf_1. The resulting inferences derived by this rule are not precise. In a sense, these are abductions. For example, a tree_1 do not have as PART-OF all possible hyponyms of fruit_1. Moreover, the combination of the last two inference rules will allow to produce also relations such as cactus_1 has_mero_part leaf_1.

Notice that when applying these five simple inference rules we will obtain direct answers for the first three questions mentioned above.

When applying systematically these inference rules using an inference mechanism on a particular synset we obtain large collections of new explicit and inferred PART-OF relations. However, most of them are completely erroneous relations. For instance, for tree_1 (table 7 presents their main characteristics uploaded into the Mcr) we obtain 2,423 new PART-OF relations. However, most of them are erroneous because they are violating ontological properties. For instance, we are obtaining in that way PART-OF relations for

| | |
|---|---|
| *DOMAIN* | botany |
| *SF* | plant |
| *SUMO* | FloweringPlant+ |
| *TCO* | Group+ |
| *TCO* | Living+ |
| *TCO* | Object= |
| *TCO* | Plant= |

Table 7: tree_1 synset

tree_1 corresponding to all body_part_1 hierarchy (i.e. artery_1). This inference is produced because of the following inference chain:

tree_1 –ISA–> life_form_1 –PART-OF–> body_part_1 <–ISA– finger_1

The problem now is how to solve this unwanted phenomena produced by the arbitrary nature of the structure of Wn. Table 8 and 9 presents, respectively, the main characteristics uploaded into the Mcr of finger_1 and apple_1.

| | |
|---|---|
| *DOMAIN* | anatomy |
| *SF* | body |
| *SUMO* | BodyPart+ |
| *TCO* | Part+ |
| *TCO* | Living+ |

Table 8: finger_1 synset

While an apple_1 can be part of tree_1, a finger_1 can not. Thus, we suggest to use the Tco properties associated to a particular synset as a blocking marks to impede further inference propagation beyond this synset. Both tree_1 and apple_1 share Living and Plant Tco properties[6]. When applying the inference rules to propagate PART-OF relations we must also include the Tco Part property as a constraint.

Using this approach, from a total number of 2,423 new PART-OF relations, we finally obtain 583 possible new PART-OF relations (excluding the unwanted PART-OF relations).

Finally, as a validation methodology, we suggest also to perform a cycling process Tco revision/enrichment of the selected Bc by means of this powerful inference mechanism. Obviously, as a side effect we will obtain also an enriched version of the Mcr having thousands of new validated relations.

[6] in a corrected version of the Tco, tree_1 should have also the Natural property too

| | |
|---|---|
| *DOMAIN* | botany |
| *DOMAIN* | gastronomy |
| *SF* | food |
| *SUMO* | FruitOrVegetable+ |
| *TCO* | Part+ |
| *TCO* | Living+ |
| *TCO* | Comestible+ |
| *TCO* | Function+ |
| *TCO* | Natural+ |
| *TCO* | Object+ |
| *TCO* | Plant+ |

Table 9: apple_1 synset

# 7 Conclusions

In order to maintain compatibility among all the heterogeneous resources uploaded into the Mcr, it is fundamental to have a robust and advanced ontological support. This paper studied the mapping of the main Sources of Ontological Meaning onto the Mcr and, in particular, the current work with the Top Concept Ontology. We also presented a preliminary study on the utility of the Tco to support advanced ontological inference.

# References

[Atserias *et al.*, 1997] J. Atserias, S. Climent, X. Farreres, G. Rigau, and H. Rodríguez. Combining multiple methods for the automatic construction of multilingual wordnets. In *Procceeding of RANLP'97*, pages 143–149, Bulgaria, 1997.

[Atserias *et al.*, 2003] Jordi Atserias, Luís Villarejo, and German Rigau. Integrating and porting knowleges across languages. In *RANLP'03*, pages 31–37, Borovets, Bulgaria, 2003.

[Atserias *et al.*, 2004] Jordi Atserias, Luís Villarejo, German Rigau, Eneko Agirre, John Carroll, Bernardo Magnini, and Piek Vossen. The meaning multilingual central repository. In *Proceedings of the Second International Global WordNet Conference (GWC'04)*, Brno, Czech Republic, January 2004. ISBN 80-210-3302-9.

[Benítez *et al.*, 1998] L. Benítez, S. Cervell, G. Escudero, M. López, G. Rigau, and M. Taulé. Methods and tools for building the catalan wordnet. In *Proceedings of the ELRA Workshop on Language Resources for European Minority Languages, First International Conference on Language Resources & Evaluation*, Granada, Spain, 1998.

[Bentivogli *et al.*, 2002] L. Bentivogli, E. Pianta, and C. Girardi. Multiwordnet: developing an aligned multilingual database. In *First International Conference on Global WordNet*, Mysore, India, 2002.

[Daudé *et al.*, 2001] J. Daudé, L. Padró, and G. Rigau. A complete wn1.5 to wn1.6 mapping. In *Proceedings of NAACL Workshop "WordNet and Other Lexical Resources: Applications, Extensions and Customizations"*, Pittsburg, PA, United States, 2001.

[Guarino and Welty, 2000] Nicola Guarino and Christopher A. Welty. A formal ontology of properties. In *Proceedings of ECAI'2000 Workshop on Knowledge Acquisition, Modeling and Management*, pages 97–112, 2000.

[Harabagiu and Moldovan, 1998] S. Harabagiu and D. Moldovan. Knowledge processing on extended wordnet. In *WordNet: An Electronic Lexical Database and Some of its Applications, Editor C. Fellbaum*. MIT Press, 1998.

[Lyons, 1977] J. Lyons, editor. *Semantics 1*. Cambridge University Press, Cambridge, UK, 1977.

[Magnini and Cavagli, 2000] B. Magnini and G. Cavagli. Integrating subject field codes into wordnet. In *In Proceedings of the Second Internatgional Conference on Language Resources and Evaluation LREC'2000*, Athens. Greece, 2000.

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

[Pustejovsky, 1995] J. Pustejovsky, editor. *The Generative Lexicon*. MIT Press, Cambridge, MA, 1995.

[Rigau *et al.*, 1997] G. Rigau, J. Atserias, and E. Agirre. Combining unsupervised lexical knowledge methods for word sense disambiguation. In *Proceedings of joint 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics ACL/EACL'97*, Madrid, Spain, 1997.

[Rigau *et al.*, 2002] G. Rigau, B. Magnini, E. Agirre, P. Vossen, and J. Carroll. Meaning: A roadmap to knowledge technologies. In *Proceedings of COLING'2002 Workshop on A Roadmap for Computational Linguistics*, Taipei, Taiwan, 2002.

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