Consistent annotation of EuroWordNet with the Top Concept Ontology

Javier Álvez[*], Jordi Atserias[**], Jordi Carrera[***], Salvador Climent[***], Antoni Oliver[***], and German Rigau[*]

[*] Basque Country University. [**] Web Research Group - Universitat Pompeu Fabra [***] Open University of Catalonia.

jibalgij @ si.ehu.es, jordi.atserias@upf.edu, jcarrerav@uoc.edu, scliment@uoc.edu, aoliverg@uoc.edu, german.rigau@ehu.es

Abstract. This paper presents the complete and consistent annotation of the nominal part of the EuroWordNet (EWN). The annotation has been carried out using the semantic features defined in the EWN Top Concept Ontology. Up to now only an initial core set of 1024 synsets, the so-called Base Concepts, were ontologized in such a way.

1. Introduction

Componential semantics has a long tradition in Linguistics since the work of post-structuralists as Hjelmslev in the thirties [cf. Simone 1990] or [Katz and Fodor 1963] among generativists. There is common agreement that this kind of lexical-semantic information can be extremely valuable for making complex linguistic decisions. Nevertheless, according to [Simone 1990], componential analysis cannot be actually achieved due to three main reasons (being the first the most important): (1) the vocabulary of a language is too large, (2) each word needs several features for its semantics to be adequately represented and (3) semantic features should be organized in several levels.

Our work provides a good solution to these problems, since 65.989 noun concepts from WordNet 1.6 (WN16) [Fellbaum 1998] corresponding to 116.364 noun lexemes (variants) have been consistently annotated with an average of 6.47 features per synset, being those features organized in a multilevel hierarchy. Therefore, it might allow componential semantics to be tested and applied in real world situations probably for the first time, thus contributing to a wide number of NLP tasks involving

semantic processing: Word Sense Disambiguation, Syntactic Parsing using selectional restrictions, Semantic Parsing or Reasoning.

Despite its wide scope, the work presented here is envisaged to be the first stage of an incremental and iterative process, as we do not assume that the current version of the EWN Top Concept Ontology (TCO) covers the optimal set of features for the aforementioned tasks. Currently, a second phase has started within the framework of the KNOW Project[1] in which the first version of the enriched lexicon is being used to label a corpus. We plan to use later this annotation for abstracting the semantic properties of verbs occurring in the corpus. This will lead, presumably, to a reformulation of the TCO, through addition, deletion or reorganisation of features.

In this paper, is organized as follows. After a brief summary of the state of the art (section §2), we present our methodology for annotating the nominal part of EWN (section §3). Then, we provide a qualitative analysis by providing some relevant examples (section §4). Section §5 summarizes a quantitative analysis and finally, section §6 provides some concluding remarks.

2. Previous Work and State of the Art

2.1 The EuroWordNet Top Ontology

The EWN TCO was not primarily designed to be used as a repository of lexical semantic information, but for clustering, comparing and exchanging concepts across languages in the EWN Project. Nevertheless, most of its semantic features (e.g. Human, Object, Instrument, etc.) have a long tradition in theoretical lexical semantics and have been postulated as semantic components of meanings. We will only describe here some of its major characteristics (see [Alonge et al. 1998] for further details).

The EWN TCO (Fig. 1) consists of 63 features and it is primarily organized following [Lyons 1977]. Correspondingly, its root level is structured in three disjoint types of entities:

- *1stOrderEntity* (physical things, e.g.: vehicle, animal, substance, object)
- *2ndOrderEntity* (situations, e.g.: happen, be, begin, cause, continue, occur)
- *3rdOrderEntity* (unobservable entities e.g.: idea, information, theory, plan)

*1stOrderEntitie*s are further distinguished in terms of four main ways of conceptualizing or classifying concrete entities:

[1] *KNOW*. Developing large-scale multilingual technologies for language understanding . Ministerio de Educación y Ciencia. TIN2006-15049-C03-02.

- *Form*: as an amorphous substance or as an object with a fixed shape (Substance or Object)
- *Composition*: as a group of self-contained wholes or as a necessary part of a whole, hence the subdivisions Group and Part.
- *Origin*: the way in which an entity has come about (Artifact or Natural).
- *Function*: the typical activity or action that is associated with an entity (Comestible, Furniture, Instrument, etc.)

These main features are then further subdivided. These classes are comparable to the Qualia roles as described in [Pustejovsky, 1995] and are based on empirical findings raised during the development of the EWN project, when the classification of the Base Concepts (BCs) was undertaken. Concepts can be classified in terms of any combination of these four roles. As such, these top concepts function more as features than as ontological classes.

Although the main-classes are intended for cross-classification, most of the subdivisions are disjoint classes: a concept cannot be both an *Object* and a *Substance*, or *Natural* and *Artifact*. As explained below, feature disjunction will play an important role in our methodology.

*2ndOrderEntitie*s can lexicalize both nouns and verbs (as well as adjectives and adverbs) denoting static or dynamic situations, such as birth, live, life, love, die and death. All 2ndOrderEntities are classified using two different classification schemes:

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

SituationType represents a basic classification in terms of the event-structure (in the formal tradition) or the predicate-inherent Aktionsart properties of nouns and verbs, as described for instance in [Vendler 1967]. *SituationType*s can be *Static* or *Dynamic*, further subdivided in *Property* and *Relation* on the one side and *UnboundedEvent* and *BoundedEvent* on the other.

*SituationComponent*s (e.g. *Location*, *Existence*, *Cause*, *Mental*, *Purpose*) emerged empirically when selecting verbal and deverbal Base Concepts in EWN. They resemble the cognitive components that play a role in the conceptual structure of events, as described in [Talmy 1985] and others. In fact, much in the same way as *Function* did for *1stOrderEntitie*s, they are good candidates for encoding important semantic properties of words denoting situations.

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

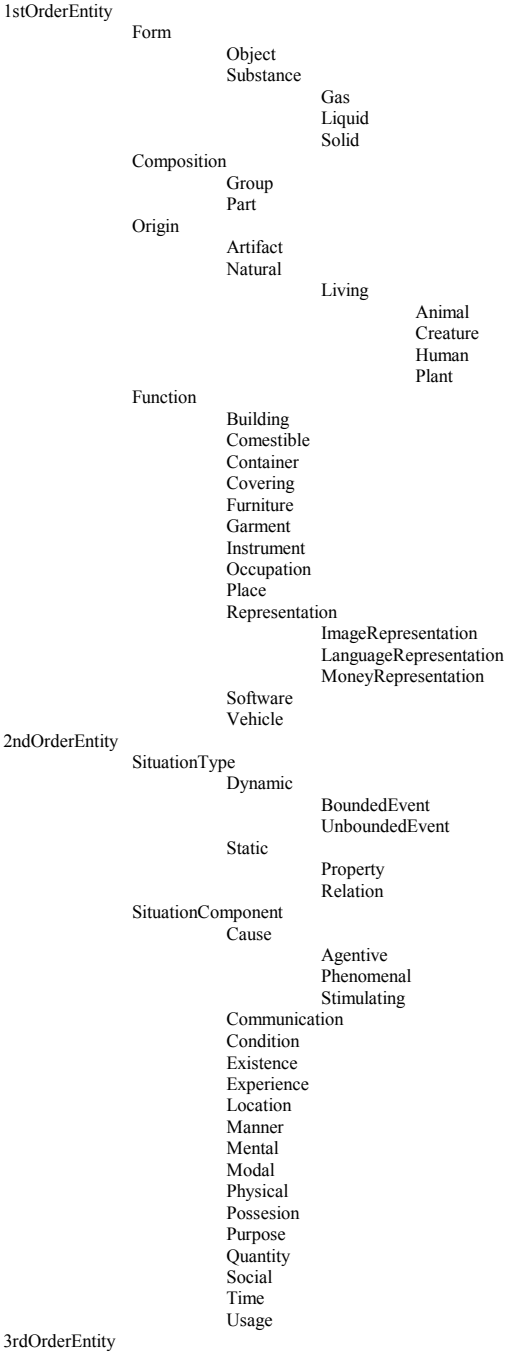

Fig. 1. The EWN Top Concept Ontology

Finally, *3rdOrderEntitie*s was not further subdivided, since there appeared to be a limited number of BCs of this kind in EWN.

The TCO has been redesigned twice, first by the EAGLES expert group [Sanfilippo et al. 1999] and then by [Vossen, 2001]. EAGLES expanded the original ontology by adding 74 concepts while the latter made it more flexible, allowing, for instance, to cross-classify features between the three orders of entities.

Moreover, the Global Wordnet Association [GWA 2007] recently distributed a taxonomy consisting of 71 so-called Base Types which can be seen as semantic primitives or taxonomic tops playing a key role in large-scale semantic networks. The Base Types have been derived by refining the original set of BCs. They are connected to both EWN synsets and TCO features, and represent an important synthesis effort in order to achieve a more elegant and economic modelling of the TCO.

2.2 Ontological information in the Multilingual Central Repository

In the framework of the UE-funded MEANING project [Rigau et al. 2002] a Multilingual Central Repository[2] (MCR) was designed and implemented in order to act as a multilingual interface for integrating and distributing lexical-semantic knowledge [Atserias et al. 2004a]. The MCR follows the model proposed by the EuroWordNet project (EWN) [Vossen 1998], i.e. a multilingual lexical database with wordnets for several languages. It includes wordnets for English, Spanish, Italian, Catalan and Basque.

The EWN architecture includes the Inter-Lingual-Index (ILI), which is a list of records that interconnect synsets across wordnets. Using the ILI, it is possible to go from word meanings in one language or particular wordnet to their equivalents in other languages or wordnets. The current version of the MCR uses the set of Princeton WordNet 1.6 synsets as ILI.

In the MCR, the ILI is connected to three separate ontologies: the EWN TCO (described above), the Domain Ontology (DO) [Magnini and Cavaglià, 2000] and the Suggested Upper Merged Ontology (SUMO) [Niles and Pease, 2001]. The DO is a hierarchy of 165 domain labels, which are knowledge structures grouping meanings in terms of topics or scripts, e.g. Transport, Sports, Medicine, Gastronomy. SUMO incorporates previous ontologies and insights by Sowa, Pierce, Russell and Norvig and others and, compared to EWN TCO, is much larger and deeper. The WN-SUMO mapping [Niles and Pease 2003] assigns only one SUMO category to every WN16 synset (being SUMO a large formal ontology), while the EWN TCO, as explained above, assigns a combination of a more reduced number categories. This makes the TCO much more suitable than that of SUMO for implementing componential

[2] http://adimen.si.ehu.es/cgi-bin/wei5/public/wei.consult.perl

semantics. While all the ILI is connected to the DO and to SUMO, only 1024 ILI-Records were connected to the TCO, i.e. those were selected as BCs in the EWN project.

2.3 Lexical Semantics for Robust NLP

Some NLP systems, such as knowledge-based Machine Translation systems usually include some kind of decision making (e.g. transfer module, PP-attachment) using lexical semantic features such as Human, Animate, Event, Path, Manner etc. [Hutchins 1995]. Its use, however, is restricted to demo systems, e.g. [Nasr et al. 1997] or, in real-world systems, to a limited number of lexical entries or/and to a very reduced number of semantic features, due to the difficulty of annotating a comprehensive lexicon with an exhaustive set of features.

However, wordnets are large lexical resources freely-available and widely used by the NLP community. Currently, they serve a wide number of tasks involving some degree of semantic processing. In most of these tasks, wordnets are used to generalize or abstract a set of synsets to a subsuming one by following the WordNet hierarchy up. The main problem is finding the right level of generalization; that is, finding the concept which optimally subsumes a given set of concepts; but it could be the case that the class which would optimally capture the generalization is not lexical, but abstract –thus having to be represented through features. It can also be the case that wordnet simply is not the kind of taxonomy required, fact which can be due to several reasons: incompleteness, incorrect structuring, or perhaps that its structuring should be arranged differently for a particular NLP task.

Bearing these drawbacks in mind, some authors have turned to use the ontologies mapped onto WordNet to determine new sets of classes. For instance, [Atserias et al. 2005] and [Villarejo et al. 2005] have already used the MCR including SUMO, DO, WN16 Semantic (Lexicographer's) Files and a preliminary rough expansion of the TCO for Word Sense Disambiguation.

For many tasks, it seems that using a feature-annotated lexicon seems more appropriate than using the WordNet tree-structure, since (i) the WordNet hierarchy is not consistently structured [Guarino 1998] and (ii) a feature-annotated lexicon allows to make predictions based on measures of similarity even for words that, being sparsely distributed in WordNet, can only be generalized by reaching common hypernyms in levels too high in the hierarchy. Besides, a multiple-feature design allows to naturally depict semantically complex concepts, such as so-called *dot-objects* [Pustejovsky 1995], e.g., intrinsically polysemic words such as "letter", since a letter is something that can both be destroyed and carry information (as in "I burnt your love letter"). These aspects of meaning can be easily coded through using the EWN TCO, as shown in (1)

(1) "letter": FUNCTION: LanguageRepresentation
FORM: Object

In this direction, [Dzikovska et. Al. 2003] use a lexicon augmented with EWN TCO features both to implement selectional restrictions to limit the search space when parsing and to perform type-coercion in a dialogue system.

3. Methodology

Our methodology for annotating the ILI with the TCO[3] is based on the common assumption that hyponymy corresponds to feature set inclusion [Cruse 2002, p.8] and in the observation that, since wordnets are taken to be crucially structured by hyponymy " (…) by augmenting important hierarchy nodes with basic semantic features, it is possible to create a rich semantic lexicon in a consistent and cost-effective way after inheriting these features through the hyponymy relations" [Sanfilippo et al., 1999, pp. 204-205].

Nevertheless, performing such operation is not straightforward, as wordnets are not consistently structured by hyponymy [Guarino 1998]. Moreover, wordnets allow multiple inheritance. These are both drawbacks to overcome and situations to take advantage of by our methodology.

As told above, within the EWN project, a limited set of lexical base concepts[4] (the BCs) was annotated with TCO features. Despite being largely general in meaning, this set did not cover all of the upper level nodes in the wordnets. This was clearly a drawback for expanding features down all of WN1.6, thus the first step of our work consisted of annotating the gaps up the hierarchy, from the BCs to the Unique beginners. This was made semiautomatically: given that every synset in WN1.6 originally belongs to a so-called Semantic File (a flat list of 45 lexicographer files), those synsets were assigned a TCO feature via a table of expected equivalence between TCO nodes and Semantic Files.

This made the WN1.6 ready to be fully populated with at least one feature per synset. Nevertheless, in many cases, synsets got more than one feature, for one or more of the following reasons:

- They are BCs, so they were manually annotated with more than one feature

[3] We use WN16 since the ILI is drawn up on this version of WordNet

[4] Base Concepts (BCs) should not be confused with Basic Level Concepts (BLCs) as defined by [Rosch and Mervis, 1975] but in a future work BCs can be taken as a starting set to define that of BLCs. Since BLCs are supposed to be richer in distinctive features and the most psychologically salient lexical categories, they can also be relevant for advanced NLP tasks.

- In addition to their own manual annotation, they inherit features from one or more hypernyms
- They inherit features from different hypernyms, either located at different levels in a single line of hierarchy or by the effect of multiple inheritance

An initial rough expansion was the first ground for revision and inspection, following the strategy defined in [Atserias 2004b]. The a task has lasted for about three years and has involved several re-expansion cycles.

The manual work has been based on TCO feature incompatibilities. It consisted in automatically detecting co-occurrences in a synset of pairs of incompatible features. The axiomatic incompatibilities are the following:

- 1stOrderEntity - 2ndOrderEntity
- 1stOrderEntity - 3rdOrderEntity
- 3rdOrderEntity - 2ndOrderEntity [except for SituationComponent]
- 3rdOrderEntity - Mental[5]
- Object - Substance
- Gas - Liquid - Solid
- Artifact - Natural
- Animal - Creature - Human - Plant
- Dynamic - Static
- BoundedEvent - UnboundedEvent
- Property - Relation
- Physical - Mental
- Agentive - Phenomenal - Stimulating

The first rough expansion described above caused the following number of feature conflicts:

- 214 feature conflicts in 49 synsets caused by incompatible hand annotation
- 2247 feature conflicts in 743 synsets caused by hand annotation incompatible with inherited features
- 225.447 feature conflicts in 26.166 synsets caused by incompatibility between inherited features

The first type of conflicts usually indicates synsets causing ontological doubts to annotators within the EWN project (e.g. is “skin” an object or a substance?). The third type usually reveals errors in WordNet structure (i.e. *ISA overloading* [Guarino 1998]). The second type might be caused by either or both reasons.

The task consisted on manual checking feature incompatibilities in order to (i) adding or deleting ontological features, and (ii) setting inheritance blockage points. A

[5] The incompatibility between 3rdOrderEntity and Mental and the compatibility between 3rdOrderEntity and Situation Components is explained below.

blockage point is an annotation in WN1.6 which breaks the ISA relation between two synsets, thus no information can be passed by inheritance through it.

When a case of feature incompatibility occurred, the synset involved, together with its structural surroundings (hypernyms, hyponyms), was analyzed. If the problem was due to a WN1.6 subsumption error, the corresponding link was *blocked* and synsets below the blockage point are annotated with new TCO features.

Changes in the annotation were made and blockage points were set until all conflicts were resolved. Then a second re-expansion of TCO features was launched which resulted in a new (smaller) number of conflicts. Following this iterative and incremental approach, inheritance was being re-calculated and the resulting data was re-examined several times. Although such hand-checking is extremely complex and laborious, and despite the large number of conflicts to solve, the task ended up being feasible because working on the topmost origin of one feature conflict results in fixing many levels of hyponyms. For instance, leaf_1, "the main organ of photosynthesis and transpiration in higher plants", is a synset that subcategorizes 66 kinds of leaves. It was originally categorized as *Substance*, but, being in that sense a bounded entity, it seemed clear that it cannot be assigned such TCO label. Therefore, fixing this case resulted in fixing as many as 66 conflicts downward with a single action.

The task has been carried out using application interfaces, which allowed access the synsets and their glosses in three languages at the same time: English, Spanish and Catalan. The information that was relied on in order to make decisions was of the following kinds:

- Relational information regarding every synset and neighboring ones; i.e. the WN1.6 structure
- The nature of the feature conflict (any of the three types of incompatibility aforementioned)
- Synsets' glosses as provided by EWN
- Glosses, descriptions and examples of the TCO features as provided in [Alonge et al. 1998]
- Usual word-substitution tests that acknowledge hyponymy, as in [Cruse 1986 pp. 88-92]

The task finished when finally a re-expansion of properties did not result in new conflicts. Then, two final steps were applied. First, as the TCO is itself a hierarchy, for every synset, its resulting annotation was expanded up-feature; e.g. if a synset beared the feature *Animal* it was also labelled *Living*, *Natural*, *Origin* and *1stOrderEntity*. Second, the whole noun hierarchy was been checked for consistency using several formal Theorem Provers like Vampire [Riazanov and Voronkov 2002] and E-prover [Schulz 2002]. This step resulted in a number of new conflicts which were finally fixed.

This methodology has led to detect many more inconsistencies in WordNet and much deeper into the hierarchy than previous approaches (e.g. [Martin 2003]).

This procedure can be seen as a shallow ontologization of WN1.6. That is, blocked links are reassigned to the TCO. This constitutes a pragmatic solution to the problem of the difficulty of complete wordnets ontologization. In this sense, our work will probably be the second one to ontologize the whole WordNet, after that with SUMO [Niles and Pease 2003]. However, our coding (i) is multiple (SUMO links every synset to only one label of the ontology) and (ii) it is more workable since it uses a more intuitive and simple TCO.

Regarding the completion of the work, the possibility that some areas in the WordNet hierarchy have remained unexamined cannot be completely excluded, although a very large number of changes have been introduced: (i.e. more than 13.000 manual interventions). Moreover, it should be noticed that, when removing links or features to fix errors, all hyponimy lines involved by the action have been reexamined and reannotated in order not to loss information.

4. Examples and qualitative discussion

In this section some examples of our methodology are presented at work. Hereinafter, noun synsets are represented by one of their variants enclosed in curly brackets and TCO features by its name in italics, capitalized and enclosed in square brackets. Inherited features are marked '+' while manually assigned features are marked '='. Indentations stand for ISA relations. The symbol 'x' as in '-x-' or '-x->' means that the relation has been blocked.

4.1 Bandung is not Java but a part of it

A simple but very typical case is the following, in which the conflict results from multiple inheritance and the incorrect use of hyponymy instead of meronomy in WN1.6:

{Bandung_1[6] [*Artifact*+ *Natural*+]}]
---> {Java_1 [*Natural*+]}
---> {island_1 [*Natural*+]}
---> {city_1 [*Artifact*=]}

Clearly, Bandung is a city, but it *is not a* Java (though it is *part of* Java). This case is revealed thanks to incompatibility between *Natural* and *Artifact*. It is fixed by blocking the subsumption link between Bandung_1 and Java_1:

{Bandung_1 [*Artifact*+]}]
-x-> {Java_1 [*Natural*+]}
---> {island_1 [*Natural*+]}
---> {city_1 [*Artifact*=]}

[6] A city in the island of Java.

4.2 A drug is a substance

This case is less straightforward but as well quite representative of malfunctions in the WN1.6 hierarchy. In WN1.6, {artifact_1} is both glossed as "a man-made object" and an hyponym of {physical_object_1}. Thus, in EWN it was annotated with the TCO feature [*Object*], which stands for bounded physical things. Nevertheless, its hyponym {drug_1} subsumes substances, therefore, it was annotated in EWN as [*Substance*]. It seems clear that the WN1.6 builders wanted to capture the fact that drugs are artificial compounds (although there indeed exist natural drugs[7]). But this fact, which is represented by the ISA relation between {drug_1} and {artifact_1} is not consistent with conceptualising {artifact_1} as a physical, bounded, object. In our work, feature expansion revealed the contradiction, since TCO features [*Object*] and [*Substance*] are incompatible:

{artifact_1 [*Object*=]}
--- {article_2 [*Object*+]}
--- {antiquity_3 [*Object*+]}
--- {... [*Object*+]}
--- {drug_1 [*Substance*= *Object*+]}
--- {aborticide_1 [*Substance*= *Object*+]}
--- {anesthetic_1 [*Substance*= *Object*+]}
--- {... [Substance=] [*Object*+]}

In this case, there were two possible solutions: either to underspecify {artifact_1} for *Object* and *Substance*, thus allowing it to subsume both kinds of entities, or blocking the subsumption relation between {drug_1} and {artifact_1}. We chosed the latter solution because {artifact_1} mainly subsumes hundreds of physical objects in WN1.6. Moreover, this solution is consistent with the glosses and respects the statement of {artifact_1} as hyponym of [physical_object_1}. Therefore, it seems better to treat {drug_1} as an exception than to change the whole structure:

{artifact_1 [*Object*=]}
--- {article_2 [*Object*+]}
--- {antiquity_3 [[*Object*+]}
--- {... [*Object*+]}
-x- {drug_1 [*Substance*=]}
--- {aborticide_1 [*Substance*+]}
--- {anesthetic_1 [*Substance*+]}
--- {... [*Substance*+]}

If we conceptualize the annotation with the TCO not just as simple feature labelling but as connecting WN1.6 to an upper flat abstract ontology, this solution is

[7] This fact prevents {drug_1} to be labelled [*Artifact*]. Only a number of its hyponyms can be done so.

equivalent to chopping off the {drug_1] subtree and link it to the [*Substance*] node of the TCO:

[*1stOrderEntity*]
--- [*Form*]
--- [*Object*]
--- {artifact_1}
--- {article_2}
--- {antiquity_3}
--- {...}
--- [*Substance*]
--- {drug_1}
--- {aborticide_1}
--- {anesthetic_1}
--- {...}

This vision was termed the shallow ontologization of WordNet in [Atserias 2004b].

4.3 The Statue of Liberty

In this section, a complete case is described showing how one single feature conflicting in the bottom of the hierarchy reveals a chain of inconsistencies up to the upper levels of the taxonomy, thus resulting in hundreds of wrongly classified synsets. We also show how our methodology is applied to solve these problems.

One conflict between first order and second order features originally taking place in {Statue_of_Liberty_1} climbs up to {creation_2} reveals the big confusion existing in WN1.6 regarding art, artistic genres, works of art and art performances (the last being events). Fixing this involved blockages and feature underspecification throughout the hierarchy. Finally, one synset {creation_2} should be underspecified as it would need disjunction of properties to be properly represented, as it can be either an object or an event.

In order to facilitate the explanation, synsets are represented by a single intuitive word, the *3rdOrderEntity* feature is more intuitively represented as [*Concept*] and only the more relevant synsets and features are shown.

As a starting point, there were four BCs manually annotated in EWN: {artifact} as [*Object*], {abstraction} as [*Concept*], {attribute} as [*Property*] and {sculpture} as [*ImageRepresentation*]. Figure 2 shows the clear-cut result of a direct expansion of properties by feature inheritance.

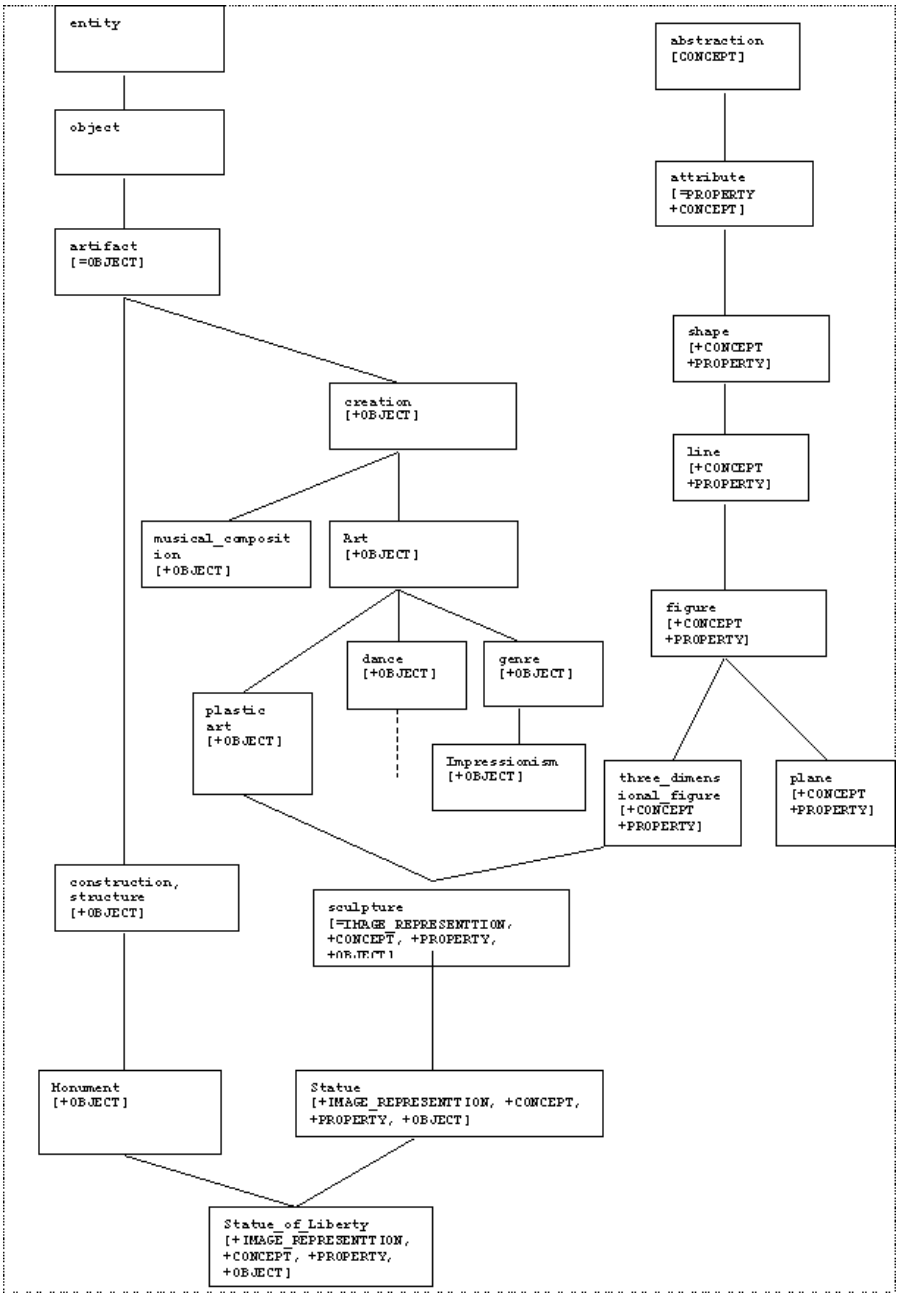

Fig. 2. The case of the Statue. Initial situation.

As a result of this process, several shocking annotations can be noticed at a first sight, for instance: (1) {musical composition}, {dance} and {impressionism} as [*Object*]; (2) {sculpture} as [*Property*], and (3) {Statue_of_Liberty} as [*Concept*].

Notice that we became aware of all this situation by inspecting the incompatibility of those TCO features inherited by {Statue of Liberty}. Due to multiple inheritance, the popular monument was taken to be an artifact, hence an object; but at the same time a kind of {art} —as e.g. {dance}, which is clearly an event, while {impressionism} is nothing but a concept. Moreover, {Statue of Liberty} appeared to be an abstraction, a [*Concept*], just as the geometric notion of a {plane}. Last, the statue also inherited [*Property*]. So, the result of applying full inheritance of ontological properties in WN1.6 resulted in multiple incompatible features eventually colliding at {Statue_of_Liberty}.

The analysis of the situation led to blockage of the following hierarchy paths, as it is shown in Figure 3:

- Between {artifact} and {creation}
- Between {art} and {dance} (but not between {art} and {genre})
- Between {plastic_art} and {sculpture}
- Between {three_dimensional_figure} and {sculpture}

Moreover, {creation} was underspecified by assigning the upmost neutral feature [*Top*] and [*Property*] was deleted in {attribute} since it is better represented by {attribute}'s hyponym {property} while the rest of hyponynms here considered (lines, planes, etc.) are, according to their glosses and relations, concepts.

The reasons behind these changes were the following:

(1) Although, intuitively, one might say that a creation is an artifact (for creations are made by men), according to the glosses and hyponyms one can realize that the synset {artifact} subsumes objects, while {creation} subsumes both objects and activities brought about by men (e.g. a "musical composition"). Therefore, {creation} can not inherit first order features, since they are incompatible with second order ones. Consequently, {creation} was here labeled as [*Top*] thus allowing its hyponyms to be further specified as entities or events since neither its gloss ("something that has been brought into existence by someone") nor the lack of homogeneity of its hyponyms allowed to make a choice. In a more flexible version of the TCO, as that proposed by [Vossen 2001], [*Origin*] features could be also attributed to second and third order entities. This will allow to assign [*Artifact*] to synsets like {Creation}. We intend to evolve to a TCO like this in the future.

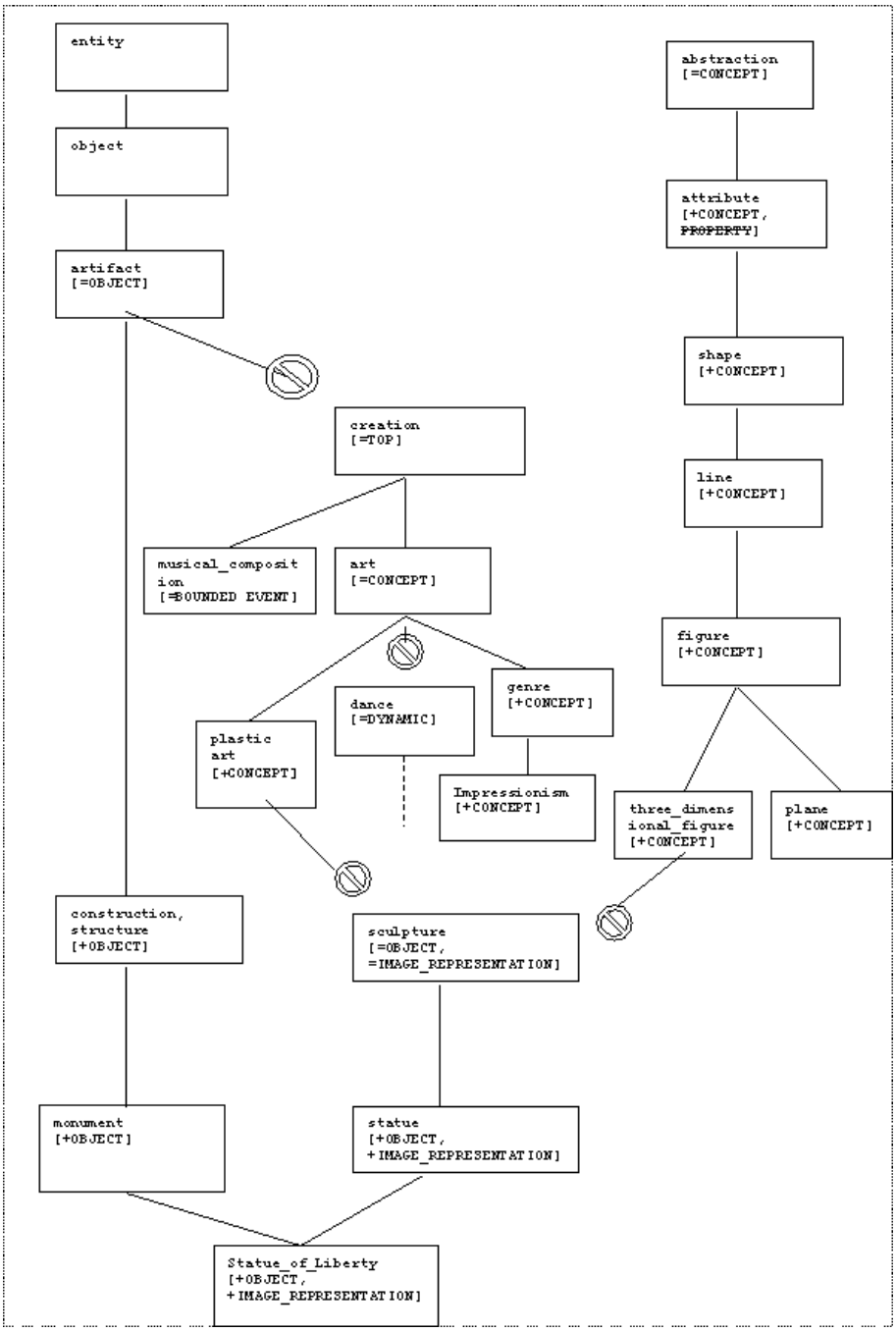

Fig. 3. The case of the Statue. Final result

(2) Although, intuitively, one might say that dance is a kind of art, according to the glosses and other hyponyms it is realized that {art} refers to the concept (like e.g. {impressionism}) while {dance} refers to an activity. Therefore, while "art" and "impressionism" are considered ideas, "dance", however, is an activity.

(3) Although, intuitively, one might say that sculpture is a plastic art, according to the glosses and other hyponyms one can realize that, as regards the senses given, {sculpture} refers to physical objects, while {plastic_art} refers to the abstract concept — a type of {art}, such as {impressionism}.

(4) Although, intuitively, one might say that a sculpture is a three dimensional figure, according to the glosses and other hyponyms it is realized that, {three_dimensional_figure} refers to the shape (the same as one-dimensional lines or two-dimensional planes, that is, abstract shapes). Therefore, in this sense, "sculptures" are objects while "figures" or "shapes" are geometrical abstractions.

The final result, as it can be seen in Figure 3, is a new quite reasonable labelling of the set of concepts, implicitly involving a reorganisation of the WN1.6 hierarchy. It is easy to realize how these limited set of decisions (four blockages, one feature deletion and few feature relabelling) subsequently affect hundreds of synsets. For instance, {creation} and {sculpture} relate to 713 and 28 hyponyms respectively.

4.4 Notes for further discussion

During the time devoted to carry on the work, a lot of interesting facts have been discovered about two objects of study: the structure of the noun hierarchy of WN1.6 and the nature of the EWN TCO features – as well as the mapping between both.

These facts are going to be further studied taken into consideration at least the following issues:

- To which extent those noun hierarchy problems correspond to those described in [Guarino 1998] or there are other kinds of facts distorting the WordNet structure

- Typical doubts or mistakes in the BCs annotation with the TCO carried on in the EWN Project

- Problems related to lack of clear definition of either synsets in WordNet or features in the TCO

For instance, a very common malpractice in EWN when annotating BCs with the TCO was that of the double coding of non-physical entities both as *3rdOrderEntity* and *Mental*. *Mental* is a subfeature for *2ndOrderEntity* and, as far as *2ndOrderEntity* and *3rdOrderEntity* are explicitly declared as incompatible, *Mental* and *3rdOrderEntity* can not coexist. Therefore, *Mental* has to be deleted, since what the encoder was intuitively doing in these cases was telling twice that the synset stands for a mental or conceptual entity. In a future enhanced TCO, following [Vossen 2001], it would be better to allow *Origin*, *Form*, *Composition* and *Function* features to be applied to situations and concepts, instead of the current classification based on *3rdOrderEntity* and *Mental* being disjunct. This will allow *Concept* to be cross-classified as for instance to classify "Neverland" as both *Concept* and *Place* in order to indicate that it is an imaginary location or to underspecify "creation" by classifying it simply as *Artifact*.

5. Quantitative analysis

Summarizing, the whole process provided a complete and consistent annotation of the nominal part of WN1.6 which consist of 65,989 synsets nominals with 116,364 variants or senses. All 227,908 initial incompatibilities were solved by manually adding or removing 13,613 TCO features and establishing 359 blockage points. Now, the final resource has 207,911 synset-feature pairs without expansion and 427,460 synset feature pairs with consistent feature inheritance.

6. Conclusions and further work

We have presented the full annotation of the nouns on the EuroWordNet (EWN) Interlingual Index (ILI) with those semantic features constituting the EWN Top Concept Ontology (TCO). This goal has been achieved by following a methodology based on an iterative and incremental expansion of the initial labelling through the hierarchy while setting the inheritance blockage points. Since this labelling has been set on the ILI, it can be also used to populate any other wordnet linked to it through a simple porting process.

This resource[8] is intended to be useful for a large number of semantic NLP tasks and for testing for the first time componential analysis on real environments.

[8] http://lpg.uoc.edu/files/wei-topontology.2.2.rar

Moreover, those mistakes encountered in WordNet noun hierarchy (i.e. false ISA relations), which are signalled by more than 350 blocking annotations, provide an interesting resource which deserves future attention.

Further work will focus on the annotation of a corpus oriented to the acquisition of selectional preferences. This, compared to state-of-the-art synset-generalisation semantic annotation, will result in a qualitative evaluation of the resource and in gaining knowledge for designing an enhanced version of the Top Concept Ontology more suitable for semantically-based NLP.

References

[Alonge et al., 1998] A. Alonge, Bertagna F., Bloksma L., Climent S., Peters W., Rodríguez H., Roventini A. and Vossen P. (1998) The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology. In Piek Vossen (ed.) *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht.

[Atserias et al., 2004a] J. Atserias, L. Villarejo, G. Rigau, E. Agirre, J. Carroll, B. Magnini, and P. Vossen. (2004) 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.

[Atserias et al., 2004b] J. Atserias, S. Climent and G. Rigau (2004) Towards the MEANING Top Ontology: Sources of Ontological Meaning. *Proceedings of the LREC 2004*. Lisbon

[Atserias et al. 2005] J. Atserias, L. Padró, G. Rigau. (2005) An Integrated Approach to Word Sense Disambiguation. *Proceedings of the RANLP 2005*. Borovets, Bulgaria.

[Cruse 1986] D. A. Cruse (1986) *Lexical Semantics*. Cambridge University Press. NY

[Cruse 2002] D.A. Cruse (2002) Hyponymy and Its Varieties. In: R. Green, C.A. Bean, & S. H. Myaeng (eds.) *The Semantics of Relationships: An Interdisciplinary Perspective, Information Science and Knowledge Management*. Springer Verlag.

[Dzikovska et al. 2003] O. Dzikovska, Myroslava, Mary D. Swift i James F. Allen. (2003) Customizing meaning: building domain-specific semantic representations from a generic lexicon. Kluwer Academic Publishers.

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

[Guarino 1998] Guarino N. (1998) Some Ontological Principles for Designing Upper Level Lexical Resources. *Proceedings of the 1st International Conference on Language Resources and Evaluation*. Granada

[GWA 2007] *The Global WordNet Association web site.* Last accessed 04.06.2007-07-04 http://www.globalwordnet.org/gwa/gwa_base_concepts.htm

[Hutchins 1995] J. Hutchins (1995) A new era in machine translation research. *Aslib Proceedings 47 (1).*

[Katz and Fodor 1963] J.J. Katz and J.A. Fodor (1963) *The Structure of a Semantic Theory.* Language, 39: 170-210

[Lyons, 1977] J. Lyons (1977) *Semantics*. Cambridge University Press. Cambridge, UK

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

[Martin 2003] Martin Ph. (2003) Correction and Extension of WordNet 1.7. In *Proceedings of the 11th International Conference on Conceptual Structures.* Springer Verlag, LNAI 2746, pp. 160-173, Dresden, Germany,

[Nasr et al. 1997] A. Nasr, O. Rambow, M. Palmer, and J. Rosenzweig. (1997) Enriching Lexical Transfer With Cross-Linguistic Semantic Features (or How to Do Interlingua without Interlingua). In *Proceedings of the 2nd International Workshop on Interlingua,* San Diego, California, 1997.

[Niles and Pease, 2001] I. Niles and A. Pease. (2001) Towards a standard upper ontology. In *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems* (FOIS-2001)

[Niles and Pease 2003] I. Niles and A. Pease. (2003) Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Model Ontology. In *Proceedings of the 2003 International Conference on Information and Knowledge Engineering*. Las Vegas, USA.

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

[Riazanov and Voronkov 2002] Riazanov A. and Voronkov A. 2002 The Design and implementation of Vampire. *Journal of AI Communications.* 15(2). IOS Press.

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

[Rosch and Mervis, 1975] E. Rosch and C.B. Mervis. (1975) Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7, 573-605

[Sanfilippo et al. 1999] A. Sanfilippo, N. Calzolari, S. Ananiadou, et al. (1999) *Preliminary Recommendations on Lexical Semantic Encoding. Final Report*. EAGLES LE3-4244, 1999

[Schutz 2002] Schulz, S. 2002. A Brainiac Theorem Prover. *Journal of AI Communications 15(2/3)*: IOS Press.

[Simone 1990] R. Simone (1990) *Fondamenti di Linguistica*. Laterza & Figli. Bari-Roma. Trad. Esp.: Ariel, 1993

[Talmy 1985] L. Talmy (1985). Lexicalization patterns: Semantic structure in lexical forms. In Shopen 1985 ed. *Language typology and syntactic description: Grammatical categories and the lexicon*. Vol. 3. 57–149. Cambridge University Press. Cambridge, UK.

[Vendler 1967] Z. Vendler, (1967): *Linguistics in philosophy*. Ithaca, N.Y.: Cornell University Press.

[Villarejo et al. 2005] L. Villarejo, L. Màrquez and G. Rigau (2005) Exploring the construction of semantic class classifiers for WSD. In *Procesamiento del Lenguaje Natural,* nº35, pp. 195-202 Granada, Spain.

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

[Vossen, 2001] P. Vossen P. (2001) Tuning Document-Based Hierarchies with Generative Principles. In *GL'2001 First International Workshop on Generative Approaches to the Lexicon*. Geneva.