# Multilingual Central Repository version 3.0

**Aitor Gonzalez-Agirre, Egoitz Laparra, German Rigau**

IXA group UPV/EHU, Donostia Spain, Donostia Spain
agonzalez278@ikasle.ehu.com
egoitz.laparra, german.rigau@ehu.com

**Abstract**

This paper describes the upgrading process of the Multilingual Central Repository (MCR). The new MCR uses WordNet 3.0 as Interlingual-Index (ILI). Now, the current version of the MCR integrates in the same EuroWordNet framework wordnets from five different languages: English, Spanish, Catalan, Basque and Galician. In order to provide ontological coherence to all the integrated wordnets, the MCR has also been enriched with a disparate set of ontologies: Base Concepts, Top Ontology, WordNet Domains and Suggested Upper Merged Ontology. The whole content of the MCR is freely available.

**Keywords:** EuroWordNet, Lexical Semantics, Knowledge Integration

## 1. Introduction

Building large and rich knowledge bases is a very costly effort which involves large research groups for long periods of development. For instance, the Multilingual Central Repository (MCR)[1] (Atserias et al., 2004b), which follows the model proposed by the EuroWordNet project (LE-2 4003) (Vossen, 1998), is the result of the MEANING project (IST-2001-34460) (Rigau et al., 2002), as well as projects KNOW (TIN2006-15049-C03)[2] (Agirre et al., 2009), KNOW2 (TIN2009-14715-C04)[3] and several complementary actions associated to the KNOW[2] project. The original MCR was aligned to the 1.6 version of WordNet. In the framework of the KNOW[2] project, we decided to upgrade the MCR to be aligned to a most recent version of WordNet.

The previous version of the MCR, aligned to the English 1.6 WordNet version, also integrated the eXtended WordNet project (Mihalcea and Moldovan, 2001), large collections of selectional preferences acquired from SemCor (Agirre and Martinez, 2001) and different sets of named entities (Alfonseca and Manandhar, 2002). It was also enriched with semantic and ontological properties as Top Ontology (Àlvez et al., 2008), SUMO (Pease et al., 2002) or WordNet Domains (Magnini and Cavaglià, 2000).

The new MCR integrates wordnets of five different languages, including English, Spanish, Catalan, Basque and Galician. This paper presents the work carried out to upgrade the MCR to new versions of these resources. By using technology to automatically align wordnets (Daudé et al., 2003), we have been able to transport knowledge from different WordNet versions. Thus, we can maintain the compatibility between all the knowledge bases that use a particular version of WordNet as a sense repository. However, most of the ontological knowledge have not been directly ported from the previous version of the MCR.

[1] http://adimen.si.ehu.es/web/MCR
[2] http://ixa.si.ehu.es/know
[3] http://ixa.si.ehu.es/know2

## 2. Multilingual Central Repository 3.0

The first version of the MCR was built following the model proposed by the EuroWordNet project. The EuroWordNet architecture includes the Inter-Lingual Index (ILI), a Domain Ontology and a Top Ontology (Vossen, 1998).

Initially most of the knowledge uploaded into the MCR was aligned to WordNet 1.6 and the Spanish, Catalan, Basque and Italian WordNet and the MultiWordNet Domains, were using WordNet 1.6 as ILI (Bentivogli et al., 2002; Magnini and Cavaglià, 2000). Thus, the original MCR used Princeton WordNet 1.6 as ILI. This option also minimized side efects with other European initiatives (Balkanet, EuroTerm, etc.) and wordnet developments around Global WordNet Association. Thus, the Spanish, Catalan and Basque wordnets as well as the EuroWordNet Top Ontology and the associated Base Concepts were transported from its original WordNet 1.5 to WordNet 1.6 (Atserias et al., 1997; Benítez et al., 1998; Atserias et al., 2004a).

The release of new free versions of Spanish and Galician wordnets aligned to Princeton WordNet 3.0 (Fernández-Montraveta et al., 2008; Xavier et al., 2011) brought with it the need to update the MCR and transport all its previous content to a new version using WordNet 3.0 as ILI. Thus, as a first step, we decided to transport Catalan and Basque wordnets and the ontological knowledge: Base Concepts, SUMO, WordNet Domains and Top Ontology.

### 2.1. Upgrading from 1.6 to 3.0

This section describes the process carried out for adapting the MCR to ILI 3.0. Due to its size and complexity, all this process have been mainly automatic.

To perform the porting between the wordnets 1.6 and 3.0 we have followed a similar process to the one used to port the Spanish and Catalan versions from 1.5 to 1.6 (Atserias et al., 2004a).

**Upgrading ILI**: The algorithm to align wordnets (Daudé et al., 2000; Daudé et al., 2001; Daudé et al., 2003) produces two mappings for each POS, one in each direction (from 1.6 to 3.0, and from 3.0 to 1.6). To upgrade the ILI, different approaches were applied depending on the POS.

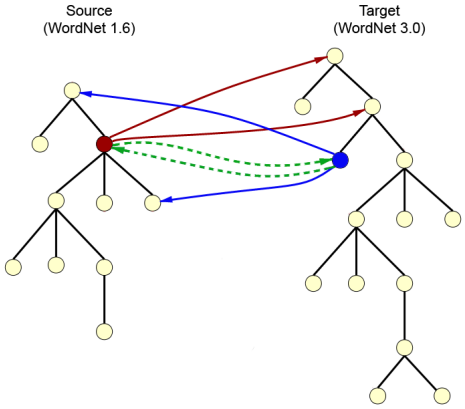


Figure 1: Example of a intersection when performing the mapping between two versions of WordNet.

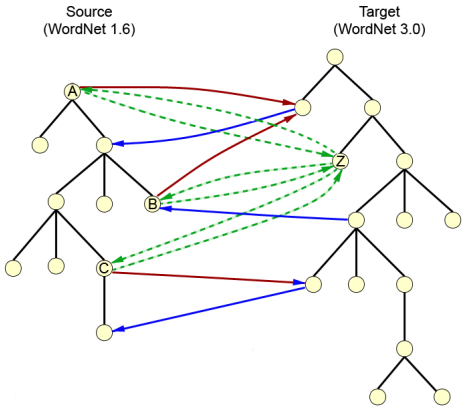


Figure 2: Example of a multiple intersection in the mapping between two versions of WordNet.

For nouns, those synsets having multiple mappings from 1.6 to 3.0 were checked manually (Pociello et al., 2008).

For verbs, adjectives and adverbs, for those synsets having multiple mappings, we took the intersection between the two mappings (from 1.6 to 3.0, and from 3.0 to 1.6). An example is shown in Figure 1, where the selected mapping between the two synsets is marked in green.

**Upgrading WordNets**: Finally, using the previous mapping, we transported from ILI 1.6 to ILI 3.0 the Basque (Pociello et al., 2008) and Catalan (Benítez et al., 1998) wordnets. The English WordNet was uploaded directly from the source files while the Spanish (Fernández-Montraveta et al., 2008) and Galician (Xavier et al., 2011) wordnets were directly uploaded from their database dumps.

It is possible to have multiple intersections for a source synset. When multiple intersections colapsed into the same target synset, we decided to join the set of variants from the source synsets to the target synset.

Figure 2 shows an example of this particular case (the intersections are displayed as dot lines). Therefore, the variants of the synsets A, B and C of WordNet 1.6 will be placed together in the synset Z of WordNet 3.0.

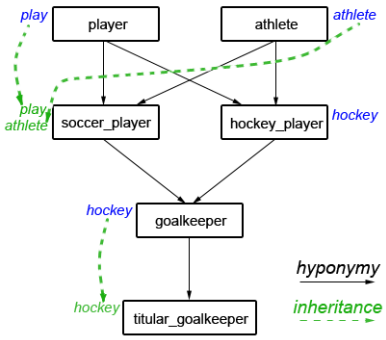


Figure 3: Example of a multiple intersection in the mapping between two versions of WordNet.

**Upgrading Base Concepts**: We used *Base Concepts* directly generated for WN 3.0[4] (Izquierdo et al., 2007).

**Upgrading SUMO**: SUMO has been directly ported from version 1.6 using the mapping. Those unlabelled synsets have been filled through inheritance. The ontology of the previous version is a modified version of SUMO, trimmed and polished, to allow the use of first-order theorem provers (like *E-prover* or *Vampire*) for formal reasoning, called AdimenSUMO[5]. The next step is to update AdimenSUMO using the latest version of SUMO for WordNet 3.0 (available on the website of SUMO[6]).

**Upgrading WordNet Domains**: As SUMO, what is currently in the MCR has been transported directly from version 1.6 using the mapping. Again, those unlabelled synsets have been filled through inheritance.

**Upgrading the Top Ontology**: Similar to SUMO and WordNet Domains, what is currently available in the MCR has been transported directly from version 1.6 using the mapping. Once more, those unlabelled synsets have been filled through inheritance. It remains to check the incompatibilities between labels following (Àlvez et al., 2008).

An example of how to perform the process of inheritance used for SUMO, WordNet Domains and Top Ontology is shown in Figure 3. The example is presented for domains, but it can be applied to the other two cases.

Figure 3 shows a sample hierarchy where each node represents a synset. The domains are displayed on the sides. The inherited domain labels are highlighted using dot lines. In this specific example synset ***soccer_player*** inherits labels *play* and *athlete* from its hypernyms ***player*** and ***athlete***, respectively. Note that synset ***hockey_player*** does not inherit any label form its hypernyms because of it owns a domain (*hockey*). Similarly, synset ***goalkeeper*** does not inherit domains coming from the synset ***soccer_player***. Finally, synset ***titular_goalkeeper*** inherits *hockey* domain (but neither *play* nor *athlete* domains).

Thus, some of the current content of the MCR will require a future revision. Fortunately, by cross-checking its ontological knowledge most of these errors can be easily detected.

[4]http://adimen.si.ehu.es/web/BLC
[5]http://adimen.si.ehu.es/web/adimenSUMO
[6]http://www.ontologyportal.org/

### 2.2. Web EuroWordNet Interface

WEI is a web application that allows consulting and editing the data contained in the MCR and navigating throw them. Consulting refers to exploring the content of the MCR by accessing words, a synsets, a variants or ILIs. The interface presents different searching parameters and displays the query results. The different searching parameters are:

- **Item**: a value to search for, it can be a Word, a Synset a Variant or an ILI.
- **Item type**: the type of item to search for: Word, a Synset a Variant or an ILI.
- **PoS**: the item' s gramatical cathegory or Part of Speech: Nouns, Verbs, Adjectives, Adverbs.
- **Search**: the type of search and subsearches (which are dinamically loaded from the database): Synonyms, Hyponyms, etc.
- **WordNet Source**: the WordNet from which navigate.
- **Navigation WordNet**: the WordNet to which navigate.
- **Gloss**: if selected it shows the glosses of the Synsets.
- **Score**: if selected, it shows the confidence factor.
- **Rels**: if selected, it shows information about the relations that each Synset has in all the target languages.
- **Full**: if selected, makes a recursive search.
- **Target WordNets**: the target WordNets of our search.

#### 2.2.1. Automatic translations

The new version of WEI is able to use *Automatic Translation Web Services* for translating automatically the glosses and examples from other wordnets. This new feature helps users to complete and/or improve the gloss or examples of a given WordNet more quicky. Both glosses and examples are taken from the original English WordNet and translated to the target language. Suggestions for glosses and/or examples will appear below the existing ones, and may choose the most appropriate. In the current version, the translations of the glosses and examples are translated only from English (despite the possibility of translating from any available source).

#### 2.2.2. Marks for synsets and variants

In the new version of WEI it is possible to assign a mark to a variant or synset to indicate special properties. We can also write a small note or comment to explain better the reason to assign that mark.
The available marks are the following:

- Variant marks:
  - DUBLEX: For those variant with dubious lexicalization.
  - INFL: Indicates that the variant is a inflected one.
  - RARE: Old fashioned or rarely used variant.
  - SUBCAT: Subcategorization.
  - VULG: For those variants that are vulgar, rude, or offensive.
- Synset marks:
  - GENLEX: Non-lexicalized general concepts that are introduced to better organize the hierarchy.
  - HYPLEX: Indicates that the hypernym has identical lexicalization.
  - SPECLEX: Domain specific terms that should be checked.

## 3. Current state of the MCR

In this section provide some information about the current state of the MCR, including the progress over the English WordNet.
Tables 1 and 2 present respectively the current number of synsets and variants, and the number of glosses of each wordnet per PoS.

## 4. Concluding Remarks and Future Directions

As a result of this work, the current version of the MCR consistently maintains new wordnet versions for five languages (English, Spanish, Catalan, Basque and Galician), and the ontological knowledge from WordNet Domains, Top Ontology and SUMO.
In particular, the main contributions of our work can be summarized as follows:
We have created a new version of the MCR using WordNet 3.0 as ILI.
We have uploaded into the new version of the MCR the English WordNet 3.0, the new Spanish WordNet 3.0 (Fernández-Montraveta et al., 2008) and a new Galician WordNet 3.0.
We have used a complete mapping from WordNet 1.6 to WordNet 3.0 (covering not only nouns, but verbs, adjectives and adverbs) to transport the Basque and Catalan wordnets and the ontological knowledge from the existing version of the MCR (using WordNet 1.6 as ILI) to the new MCR version (using WordNet 3.0 as ILI).
We have applied a very simple estrategy to complete the ontological information by exploiting basic inheritance mechanisms. This process has been applied to WordNet Domains, Top Ontology and SUMO.
The whole content of the MCR is freely available [7].

## 5. Acknowledgements

We thank the IXA NLP group from the Basque Country University. This work was been possible thanks to its support withing the framework of the KNOW2 (TIN2009-14715-C04-04) and PATHS (FP7-ICT-2009-6-270082) projects.
We also wish to thank the reviewers for their valuable comments.

[7] http://adimen.si.ehu.es/web/MCR

| WordNet | Nouns | Verbs | Adjectives | Adverbs | Synsets | WN % |
|---|---|---|---|---|---|---|
| EngWN3.0 | 147,360 | 25,051 | 30,004 | 5,580 | 118,431 | 100% |
| SpaWN3.0 | 40,009 | 11,107 | 7,005 | 1,106 | 59,227 | 50% |
| CatWN3.0 | 51,598 | 11,577 | 7,679 | 2 | 46,027 | 39% |
| EusWN3.0 | 41,071 | 9,472 | 148 | 0 | 30,615 | 26% |
| GalWN3.0 | 9,114 | 1,413 | 4,866 | 0 | 9,320 | 8% |

Table 1: Current number of synsets and variants of each WN.

| WordNet | Nouns | Verbs | Adjectives | Adverbs | Synsets | WN % |
|---|---|---|---|---|---|---|
| EngWN3.0 | 82,379 | 13,767 | 18,156 | 3,621 | 117,923 | 100% |
| SpaWN3.0 | 13,014 | 3,469 | 1,965 | 687 | 19,135 | 16% |
| CatWN3.0 | 6,289 | 44 | 840 | 1 | 7,174 | 6% |
| EusWN3.0 | 2,854 | 78 | 0 | 0 | 2,932 | 2% |
| GalWN3.0 | 4,997 | 2 | 3,111 | 0 | 8,111 | 7% |

Table 2: Current number of glosses of each WN.

## 6. References


Agirre, E. and Martinez, D. (2001). Knowledge sources for Word Sense Disambiguation. In *Proceedings of the Fourth International Conference TSD 2001, Plzen (Pilsen), Czech Republic. Published in the Springer Verlag Lecture Notes in Computer Science series. Václav Matousek, Pavel Mautner, Roman Moucek, Karel Tauser (eds.) Copyright Springer-Verlag. ISBN: 3-540-42557-8. "*.

Agirre, E., Rigau, G., Castellón, I., Alonso, L., Padró, L., Cuadros, M., Climent, S., and Coll-Florit, M. (2009). KNOW: Developing large-scale multilingual technologies for language understanding. *Procesamiento del Lenguaje Natural*, (43):377–378.

Alfonseca, E. and Manandhar, S. (2002). Distinguishing concepts and instances in WordNet. In *Proceedings of the first International Conference of Global WordNet Association*, Mysore, India.

Atserias, J., Climent, S., Farreres, J., Rigau, G., and Rodríguez, H. (1997). Combining Multiple Methods for the Automatic Construction of Multilingual WordNets. In *Proceedings of International Conference on Recent Advances in Natural Language Processing (RANLP'97)*, Tzigov Chark, Bulgaria.

Atserias, J., Rigau, G., and Villarejo, L. (2004a). Spanish WordNet 1.6: Porting the Spanish Wordnet across Princeton versions. In *LREC'04*.

Atserias, J., Villarejo, L., Rigau, G., Agirre, E., Carroll, J., Vossen, P., and Magnini, B. (2004b). The MEANING Multilingual Central Respository. In *Proceedings of the Second International Global WordNet Conference (GWC'04)*.

Bentivogli, L., Pianta, E., and Girardi, C. (2002). MultiWordNet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, Mysore, India.

Benítez, L., Cervell, S., Escudero, G., López, M., Rigau, G., and Taulé, M. (1998). Methods and Tools for Building the Catalan WordNet. In *Proceedings of ELRA Workshop on Language Resources for European Minority Languages*, Granada, Spain.

Daudé, J., Padró, L., and Rigau, G. (2000). Mapping WordNets Using Structural Information. In *38th Annual Meeting of the Association for Computational Linguistics (ACL'2000).*, Hong Kong.

Daudé, J., Padró, L., and Rigau, G. (2001). A Complete WN1.5 to WN1.6 Mapping. In *Proceedings of NAACL Workshop 'WordNet and Other Lexical Resources: Applications, Extensions and Customizations' (NAACL'2001).*, Pittsburg, PA, USA.

Daudé, J., Padró, L., and Rigau, G. (2003). Making Wordnet Mappings Robust. In *Proceedings of the 19th Congreso de la Sociedad Española para el Procesamiento del Lenguage Natural, SEPLN*, Universidad Universidad de Alcalá de Henares. Madrid, Spain.

Fernández-Montraveta, A., Vázquez, G., and Fellbaum, C. (2008). *Text Resources and Lexical Knowledge*, volume 33 of *Text, Translation, Computational Processing*, chapter The Spanish Version of WordNet 3.0, pages 175–182. Mouton de Gruyter.

Izquierdo, R., Suárez, A., and Rigau, G. (2007). Exploring the Automatic Selection of Basic Level Concepts. In *Proceedings of RANLP*.

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

Mihalcea, R. and Moldovan, D. (2001). eXtended WordNet: Progress Report. In *Proceedings of NAACL Workshop* WordNet and Other Lexical Resources: Applications, Extensions and Customizations, pages 95–100, Pittsburg, PA, USA.

Pease, A., Niles, I., and Li, J. (2002). The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications. In *AAAI-2002*.

Pociello, E., Gurrutxaga, A., Agirre, E., Aldezabal, I., and Rigau, G. (2008). WNTERM: Combining the Basque WordNet and a Terminological Dictionary. In *6th inter-*

*national conference on Language Resources and Evaluation, LREC'08*.

Rigau, G., Magnini, B., Agirre, E., Vossen, P., and Carroll, J. (2002). MEANING: A Roadmap to Knowledge Technologies. In *Proceedings of COLING Workshop A Roadmap for Computational Linguistics*, Taipei, Taiwan.

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

Xavier, G. G., Clemente, X. M. G., Pereira, A. G., and Lorenzo, V. T. (2011). Galnet: WordNet 3.0 do galego. *Linguamática*, 3(1):61–67.

Àlvez, 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. In *Proceedings of the the 6th Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech (Morocco).