Towards a Filipino WordNet

Pierre Peter Paul L. Tan
De La Salle University
2401 Taft Avenue, Manila, Philippines
pierre.tan@yahoo.com

Nathalie Rose T. Lim
De La Salle University
2401 Taft Avenue, Manila, Philippines
limn@dlsu.edu.ph

ABSTRACT
A WordNet is a database of semantically linked lexicon. It had been used in several NLP applications like information retrieval and extraction, document structuring and categorization, audio and video retrieval, language teaching, translational applications, and parameterisable information systems. Currently, WordNets have already been developed using around forty languages around the world but despite its numerous applications, there is still no WordNet developed for the Filipino language. FilWordNet is a WordNet for Filipino that would be built using the top down approach in WordNet building as was used in building several other WordNets. It would explore on using the SUMO ontology as a structured inter-lingual index to link FilWordNet to other WordNets and explore on automating parts of the WordNet building process.

Keywords: wordnet, text analysis, knowledge acquisition

1. INTRODUCTION
A semantic network is a form of knowledge representation that was first introduced by Quillian [8]. It is used in artificial intelligence and machine translation and is composed of vertices and edges as a directed graph representing concepts and the relationships between them.

A WordNet is a database of semantically linked lexicon that was introduced by G.A. Miller in 1978. This project aimed to model the human lexicon and it started with three presuppositions. The first presupposition is the separability hypothesis which if applied to natural language processing tells us that a language's lexical component can be separated and studied independently. The second is the patterning hypothesis which tells us that people cannot master the lexical knowledge needed to use a natural language without using systematic patterns and relations among the meanings of words. The last presupposition is the comprehensiveness hypothesis which tells us that one has to store lexical data that is as comprehensive as that of a human to be able to process natural languages. Figure 1 illustrates a WordNet subset [3].

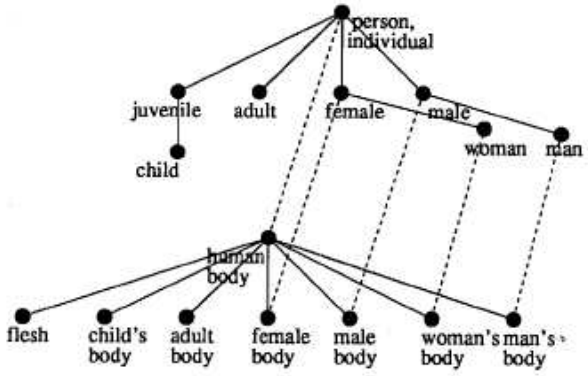

Figure 1. A WordNet Subset.

WordNets had been used in several natural language processing applications in the last decade. The study of Morato et al. [6] on the applications of WordNets identified several application areas: The first area is in information retrieval and extraction, particularly in concept identification in natural language and in query expansion. Applications in this area includes semantic disambiguation, semantic distance, and query expansion. The second area involves document structuring and categorization. Another area was in audio and video retrieval where WordNets were used to generate conceptual hierarchies for retrieval of multimedia data.

In addition to these application areas, WordNets were also used in language teaching, translational applications, and in parameterisable information systems which allowed personal searching of documents based on users' interests.

Today, WordNets had already been built using around forty languages around the world. However, there is still no WordNet for Filipino although related studies had already been done by Lat et al. [4] and Tiu [9]. These studies tried to automate the building of a bidirectional Filipino-English lexicon from corpora. FilWordNet, on the other hand, aims to build a lexical-semantic network that would model the Filipino language.

Building a WordNet for Filipino not only provides the possibility of usage in the said application areas but also provides a way of connecting the Filipino lexicon to other languages. In addition, this can also serve as a good reference for Filipino words as it includes not only lexical but also semantic information.

2. RELATED STUDIES
G.A. Miller's WordNet or the Princeton WordNet (PWN) [5] as it is now called is currently in version 2.1 (Windows) / 3.0 (UNIX)

and contains 155,327 concepts from the English language. It is comprised of nouns, verbs, adjectives, and adverbs. The PWN is organized using Synsets which are groups of words that refer to the same concept and has the same part-of-speech [12]. PWN currently has 117,597 synsets with their definition and/or example sentence called Gloss. PWN uses semantic relations to link these synsets to each other. Its semantic relations for nouns and verbs include:

For Nouns:

- Hyponymy – "kind of" relationship between two nouns or relationship between specific and more general concepts. (i.e. {bird} → {animal, animate being})
- Hyperonymy – opposite of Hyponymy. (i.e. {board} → {surfboard})
- Meronymy – "part of" relationship between two nouns or relationship between parts and wholes. Meronyms in WordNet has three types:
    - Stuff-Object Meronymy – Indicates that the first concept is a substance of the other. (i.e. {plastic} → {car, automobile})
    - Component-Object Meronymy – Indicates that the first concept is a part of the other. (i.e. {branch} → {tree})
    - Member-Collection Meronymy – Indicates that the first concept is a member of the other. (i.e. {fish} → {school})
- Holonymy – opposite of Meronymy. This relation also has three types:
    - Stuff-Object Holonymy
    - Component-Object Holonymy
    - Member-Collection Holonymy
- Coordinate Terms – synsets that share a hypernym
- Attribute – a noun for which adjectives express values. (i.e. {weight} → {light (vs heavy)})
- Sister Terms – matching strings that are both the immediate hyponyms of the same hypernym.

For Verbs:

- Entailment – a verb X entails Y if X cannot be done unless Y is or has been done. (i.e. {die, pass away, perish} → {leave, leave behind})
- Cause – opposite of entailment
- Troponymy – a relationship between two verbs where a verb is expressing a specific manner elaboration of another verb. X is a troponym of Y if to X is to Y in some manner. (i.e. {walk} → {spacewalk})
- Hyponymy & Hyperonymy
- Sister Terms

With a need to cover more languages aside from English, the idea of multilingual WordNets was concretized in 1999 by EuroWordNet (EWN). This project aimed to form a multilingual WordNet database while maintaining language-specific relations, achieving maximal compatibility across the different resources. In order to associate related words from different languages in a multilingual WordNet, EuroWordNet used an inter-lingual index (ILI). This intermediates multilingual synsets such that each synset would have at least one record connected to the ILI [12].

EWN's ILI is structured by the domain and top ontologies. The top ontology (or the top concepts) is a hierarchy of language independent concepts that reflect explicit opposition relations. The domain ontology, on the other hand, refers to the topic or event that relates to each ILI entry. EWN also extended PWN's semantic relations by adding features in relations, adding new relations, and broadening existing relations. Moreover, EWN also added relations that go between parts of speech removing PWN's restriction on the separation of parts of speech [12]. EWN refers to these as XPOS relations. Figure 2 illustrates the EWN data architecture [14].

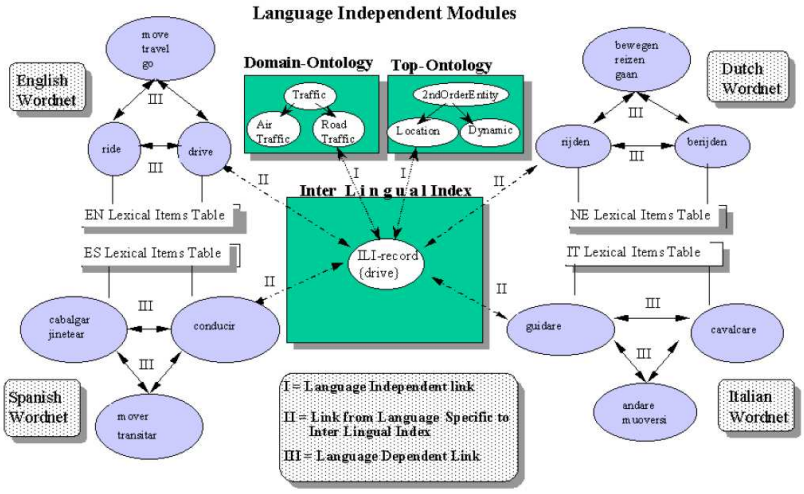

Figure 2. The EuroWordNet Data Architecture.

EWN is composed of 8 local WordNets including the English WordNet. Table 1 lists the local WordNets in EWN other than English.

Table 1. The Local WordNets of EuroWordNet.

Language	Approach	Semi-automated	Number of Synsets		
			Noun	Verb	Adj./Adv.
Czech	Merge	Yes	9727	3097	0
Dutch	Merge	Yes	34455	9040	520
Estonian	Merge	No	5028	2650	0
French	Expand	Yes	17826	4919	0
German	Merge	Yes	9951	5166	15
Italian	Merge	Yes	30169	8796	1463

Spanish	Merge	Yes	17826	4919	0

In 2001, another project called BalkaNet [10] was started. This project aimed to extend EWN to five more Balkan languages. Aside from new language support, BalkaNet used more base concepts and placed more stress in the capture of differences between languages. This project also used a standardized XML format in data representation and developed a new editor and browser tool called VisDic. Table 2 lists the local WordNets in BalkaNet.

Table 2. The Local WordNets of BalkaNet

Language	Approach	Semi-automated	Number of Synsets		
			Noun	Verb	Adj./Adv.
Bulgarian	Merge	Yes	14174	4169	3097
Czech	Merge	Yes	21009	5155	2292
Greek	Merge	Yes	14426	3402	633
Serbian	Merge	Yes	5919	1803	337
Romanian	Merge	Yes	13345	4808	1686
Turkish	Merge	Yes	8691	2556	381

A recent WordNet project for Arabic (AWN) [1] that is currently being developed uses the Standard Upper Merged Ontology (SUMO) [7] as the ILI to other WordNets. This is possible since SUMO is currently mapped to PWN and the ILI used to link different WordNets in the previous multilingual WordNet projects is composed of synsets coming from PWN.

The Suggested Upper Merged Ontology by Niles and Pease is an ontology that was created by publicly merging several ontological contents. It attempts to capture the most general and reusable terms and definitions. It is currently the largest free, formal ontology available with 20,000 terms and 60,000 axioms.

Figure 3 illustrates the mapping of AWN and PWN through SUMO.

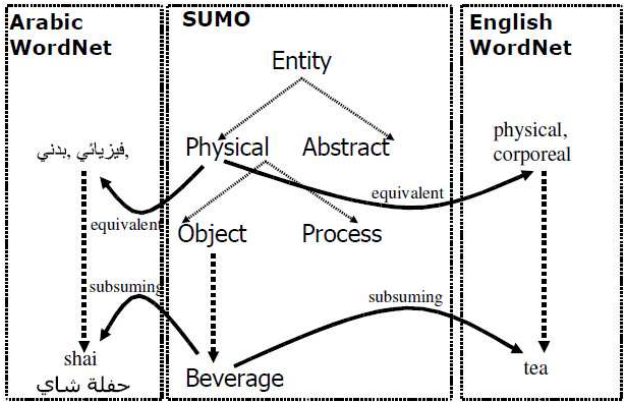

Figure 3. AWN and PWN Mapping through SUMO

The procedure in developing SUMO started with identifying all unlicensed high-level ontological content. They were then translated to the Suggested Upper Ontology Knowledge Interchange Format (SUO-KIF). This merging step is known as the syntactic merge. The next step, which is called the semantic merge, is involved with combining the existing ontologies to a single, consistent, and comprehensive framework.

The concepts in SUMO were mapped using three types of mappings and these are:

- Equivalent mapping (i.e. Mars and Earth)
- Subsuming mapping (i.e. Mars and Soil)
- Instance mapping (i.e. Mars and planet)

Vossen [13] introduced us to a top down approach in building WordNets. This approach had been used in several dozens of WordNet building projects including EWN and BalkaNet. Each EWN site first used an agreed list of base concepts which are composed of the most generalized and basic synsets. This served as an initial starting point and as the core of every local WordNets in EWN. In order to ensure the quality of each local WordNet, the internal relations between the synsets in the base concepts as well as the equivalence relations to the ILI were done manually. Each local EWN site also was allowed to add local base concepts. These synsets were composed of additional important concepts from the local language that were not found in the initial base concepts. After the core WordNets had been built. These were extended either manually or semi-automatically with more specific synsets coming from several language resources.

There are also two general approaches in building the semantic relations in each local WordNet. The first approach called the expand approach involves the translation of an existing WordNet to the local language. In doing so, semantic relations are also retained. Optionally, the builder might verify the correctness of the adapted relations. The second approach, on the other hand, which is called the merge approach, involves building of the local synsets as well as their semantic relations independently using existing resources. Both approaches could be done manually or be done semi-automatically.

Both approaches have their own advantages and disadvantages. WordNets built using the merge approach can retain language-specific properties and structure as opposed to the other approach which strictly follows the source WordNet's structure. However, one major setback of the merge approach is that it is time consuming to do. Moreover, linking to an ILI is also harder in the merge approach because of the differences in structure. WordNets built using the expand approach, on the other hand, takes lesser time and effort than the other approach since it just copies and translates an existing WordNet. A major setback however is that the source language's WordNet structure is retained in the new WordNet structure of a different language.

A limitation of WordNets prior to version 2.1 in modeling the human lexicon was the distinction of classes and instances in nouns. Here, both classes and instances uses the hyponymy or the "kind of" relationship and are treated the same in the database (i.e. class → "An actor is a man" and instance → "Tom Hanks is a man"). Because of this, as of the latest WordNet version 2.1, instance tags were added to WordNet to differentiate instances in nouns. Noun instance include proper nouns and some of its challenges to a WordNet are its dependency to communities and languages. These nouns can denote several concepts. Certain proper nouns in a country might be different in another country. Several people can have identical names. The data model for

proper nouns should be designed in such a way to overcome these problems.

3. BUILDING FILWORDNET

3.1 Objectives, Limitations, and Challenges

The task of creating a Filipino WordNet would involve the collection, digitization, and preprocessing of these data. This would also include customized software tools that would automate the process of storing and preprocessing these information as well as semi-automating the building of the WordNet. A data architecture would also need to be developed in order to efficiently store these information and developing the data architecture would also require determining the different semantic relations that would be used. The design of FilWordNet should also follow the equivalence standards used by past projects in order for FilWordNet to be linkable to other WordNets that used Princeton WordNet or SUMO as their ILI.

Building the Filipino WordNet would also involve building of relations among synsets in the local WordNet as well as the linking to an Inter-Lingual Index (ILI) that would serve as the link between the Filipino WordNet and other WordNets. In our case, the ILI that would be used is the SUMO. Although, creating the relations among words can be done manually as was done in building the Princeton WordNet [3] and the Estonian WordNet [11], this would be a very tedious task. Hence, tools would also be developed to help semi-automate this stage of the development of the Filipino WordNet. FilWordNet would be linked first to the Princeton WordNet and then to SUMO using the existing Princeton WordNet to SUMO links.

Since Filipino not only includes Tagalog words but also borrowed words from other languages, its synset count is quite large. Because of this, the words that would comprise the Filipino WordNet would only be limited to the data sources that would be used. Data sources would include existing bilingual and monolingual dictionaries (electronic and non-electronic), parallel and non-parallel corpora, local thesauri, and other existing WordNets. There would also be an initial target of 3000 total synsets each containing word entries. In addition, only nouns and verbs which comprise the majority of words in other WordNets would be considered in this study.

Some of the unique features of the Filipino language that would be considered in the FilWordNet design are its complex system of affixes where each word can come in different forms (i.e. buhay [life], kabuhayan [livelihood], buhayin [give life], binuhay [gave life]. etc.). There are also words that come with different spellings.

3.2 Methodology

FilWordNet would follow the top down approach as was used in previous WordNet building projects like EuroWordNet and BalkaNet. This approach was chosen because it is able to leverage on the established practices and experiences of several WordNet builders in the past that used the approach. Its usage of common base concepts also allows easier multilinguality.

The semantic relations of FilWordNet are to be built using the merge approach mainly to retain the language-specific properties and structure.

Prior to the actual building of the FilWordNet, data sources like corpora, bilingual dictionaries, and lexical databases that are to be used should be preprocessed. This is to ease the extraction of information from these data sources. Preprocessing would include encoding and parsing of data as well as building the required software tools that would automate the extraction of information. Figure 4 shows a flowchart of the FilWordNet methodology that is based on the top down approach [13].

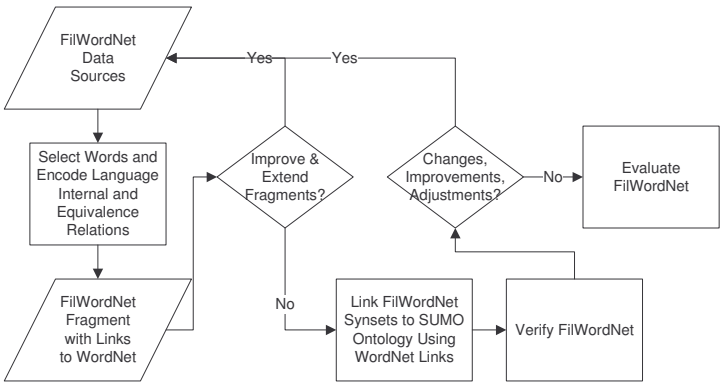

Figure 4. The FilWordNet Building Methodology

Data sources would particularly be useful during the initial phase of building the local Filipino WordNet. Here, synsets would come from existing lexical databases, thesauri, as well as corpora. Language internal relations, on the other hand, can also be extracted from corpora. And to determine the equivalence relations to a WordNet in a different language, bilingual dictionaries or aligned parallel corpora could be useful.

The building of the core of the FilWordNet would include the definition of the base concepts using the 1024 common base concepts in EuroWordNet as well as manually encoding its internal and equivalence relations to the PWN. This stage would be done with the help of linguists. Afterwards, in extending the base concepts with more synsets, customized software tools as well as existing tools can now be utilized to automatically generate internal and equivalence relations. In addition, tree comparison algorithms can be used help increase the accuracy of determining the equivalence relations.

The next stage would involve linking the FilWordNet synsets to the SUMO. This can be done automatically by going through the equivalence relations of FilWordNet and PWN since PWN is already linked to the SUMO. With the linkage of both FilWordNet and PWN to SUMO, FilWordNet can be linked to other WordNets through the PWN synsets which were used as the ILI in several past projects. Linking other WordNets other than PWN however is beyond the scope of this study.

3.3 Architecture

The FilWordNet system architecture follows other WordNets closely. The main database is able to connect the editor as well as several application systems. The FilWordNet editor would be able to browse as well as update information from the FilWordNet database. Figure 5 illustrates the described system architecture. Figure 4 expounds the FilWordNet's top down building approach.

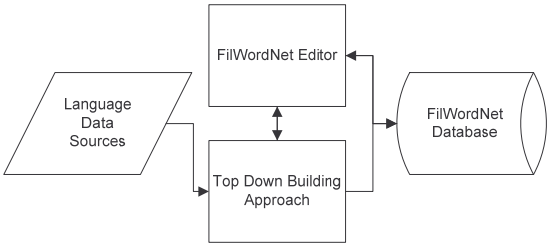

Figure 5. The FilWordNet System Architecture

The FilWordNet data architecture is designed to fully utilize the SUMO as the sole ILI to link FilWordNet to other WordNets. Since several synsets could share relational links to just a single concept in the SUMO. Finding the corresponding equivalent synsets from FilWordNet to PWN would require comparison of synset's relations. Figure 6 illustrates the data architecture of FilWordNet.

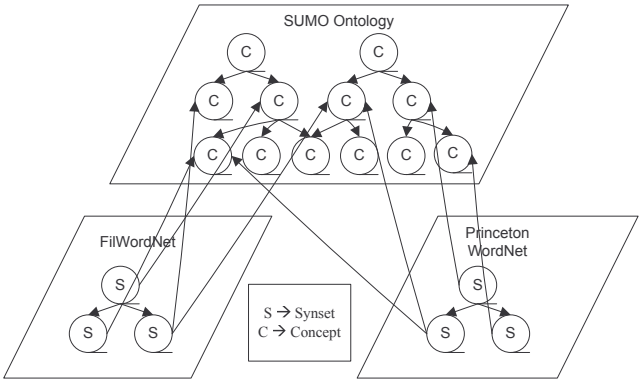

Figure 6. The FilWordNet Data Architecture

3.4 Verification

Upon completion of the FilWordNet, certain measures need to be fulfilled in order to assure completeness and consistency. This study would follow the measures for completeness and consistency that was used in the BalkaNet project [10].

To ensure completeness, the final WordNet should contain the most used words in the local vocabulary. This can be done by automatically generating a frequency list from existing corpora. Moreover, there should be no dangling relations or both ends of any relations should be present in the database. There should also be no gaps in the database. All synsets should be traceable to the top of the tree with their hyponyms. Any synset should also have at least one relation and at least one literal.

On the other hand, to ensure consistency, BalkaNet used three different steps. The first one involved the checking of the syntax of the XML files containing the WordNet data. In particular, automatic checking and correction were applied in these files. The next one checked for contradictions in the interpretation meanings of the synsets. And lastly, the consistency of the encoding of the semantic relations was checked. Here, synsets with different hypernyms (opposite of hyponyms) with their PWN equivalents and synsets without hypernyms are rechecked again. Semantic relationship loops were also checked and corrected. In addition, glosses were verified for errors like duplication.

3.5 Evaluation

In this final stage of FilWordNet building, it would be evaluated by comparing translation words from two parallel corpora of Filipino and English. It would follow the computation and tabulation that were done in BalkaNet [2] which grouped results according to:

- Those that find a translation equivalent that has at least one ILI in common with the target word
- Those that find a translation equivalent that is semantically closely related with the target word
- Those that has the sense of the current occurrence of the target word was used in is not yet implemented in the source WordNet
- Those that has the sense of the current occurrence of the target word is defined in the source wordnet but the translation equivalent does not belong to that synset
- Those that has the sense of the current occurrence of the target word is defined in the source wordnet but the relevant synset (that contains the translation equivalent) is wrongly mapped on ILI
- Those with a translation equivalent that is not wrong but the translation itself is rather loose and does not justify adding the translation equivalent to the relevant synset
- Those with a translation equivalent that was wrongly chosen by the word alignment engine
- Those with a translation equivalent, although correctly chosen by the system, is wrong due to defective translation

The first two groups are considered to be good results for the WordNet. On the other hand, the next three groups denote some inconsistencies or incompleteness in the WordNet, whereas the last two groups represent an engine or human error.

4. CONCLUSION

Building a WordNet for the Filipino language would not only provide a valuable lexical-semantic resource for our local language but also a means of using our local language in numerous applications that utilize WordNets. In addition, there are still several improvements that can be done to the initial FilWordNet. One of which is to link FilWordNet to other WordNets aside from the Princeton WordNet. Maybe a group of WordNets for different Philippine dialects can also be done in the future. The initial FilWordNet would also lack support for adjectives as well as adverbs. Its initial synset count would also be not as huge as those in other WordNets but as soon as automated tools and processes had already been done for the FilWordNet, increasing the synset count would be easier and in line with this, FilWordNet would cover more concepts and become more useful.

5. REFERENCES

[1] Black, W., Elkateb, S., Rodriguez, H., Alkhalifa, M., Vossen, P., Pease, A., and Fellbaum, C. Introducing the Arabic WordNet Project. In *Proceedings of the Third International WordNet Conference (GWC-2006).* Jeju, South Korea, 2006.

[2] Christodoulakis, D. Assessment and Evaluation of the Project's Results. BalkaNet, 2004.

[3] Fellbaum, C. (Ed.). WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA, 1998.

[4] Lat, J., Ng, S., Sze, K., and Yu, G. AEFLex: Automatic English and Filipino Lexicon Builder. Unpublished undergraduate thesis, De La Salle University, Manila, 2006.

[5] Miller, G., Beckwith, R., Fellbaum, C., Gross, D. & Miller. K.J. Introduction to WordNet: an online lexical database. International Journal of Lexicography, 3, 4(1990), 35-244.

[6] Morato, J., Marzal, M.A., Llorens, J., & Moreiro, J. WordNet Applications. In *Proceedings of the Second Global WordNet Conference (GWC-2004).* Brno, Czech Republic, 2004.

[7] Niles, I., and Pease, A. Towards a Standard Upper Ontology. In *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001).* Chris Welty and Barry Smith, eds, Ogunquit, Maine, October 17-19, 2001.

[8] Quillian, M. Semantic memory. In M. Minsky (Ed.), Semantic information processing. MIT Press, Cambridge, MA, 1968.

[9] Tiu, E. Automatic Lexicon Extraction from Comparable, Non-Parallel Corpora. Unpublished master's thesis, De La Salle University, Manila, 2004.

[10] Tufis, D., Cristea, D., and Stamou, S. BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. Romanian Journal on Information Science and Technology, 7, 1-2(2004), 9-34.

[11] Vider, K., Paldre, L., Orav, H., Oim, H. The Estonian Wordnet. EuroWordNet, 1999.

[12] Vossen, P. WordNet, EuroWordNet and Global WordNet. Revue Française deRevue Française de Linguistique Appliquee. 2002.

[13] Vossen, P. EuroWordNet General Document. EuroWordNet, 2006.

[14] Vossen, P., Bloksma, L., Rodriguez, H., Climent, S., Calzolari, N., Roventini, A., Bertagna, F., Alonge, A., and Peters, W. The EuroWordNet Base Concepts and Top Ontology. EuroWordNet, 1998.