International Journal of Database Theory and Application
Vol.9, No.6 (2016), 247-256
http://dx.doi.org/10.14257/ijdta.2016.9.6.25

WordNet in Malay Language

Jer Lang Hong

School of Computing and IT, Taylor's University, Malaysia
jerlang.hong@taylors.edu.my

World Wide Web contains huge amount of data available in different languages across the world. Web browsers are tools used to display the data in graphical forms. With the evolution of Web 3.0, data has become an important part of human daily tasks, where it is used to process information, and formulate important decision rules for many organizations. Current tools used to conceptualize data are catered for some of the world well known languages such as English. However, these tools may not be able to support other languages as there are a wide range of languages with different syntax and representation. In this paper, we present a novel lexical semantic based database tool called MalayWordNet, specifically written for Malay language. Our tool is helpful for high end semantic based applications which use Malay language as part of their data presentation.

***Keywords**: Lexical Database, Linguistic Tools, Semantic Web*

1. Introduction

Languages are developed and used by different races and statistics in http://en.wikipedia.org/wiki/English _language showed that English language is by far the most commonly used in the world. Dictionaries presented in book form have been used for people to understand the meaning of individual words of a particular language. A good example is the Oxford English dictionary. However, these dictionaries are printed copy, presented in alphabetical order, and there are no relationships between each of the keywords and terms. With the introduction of Computer Technology and World Wide Web, researchers in computer and linguistic developed electronic version dictionaries, the first of which is the work of Sharp (Sharp's PWE560, see http://www.sharp.ca/products/ index.asp?cat=77&id=529). Electronic dictionaries provide a fast search technique, with some of them providing supports for stemming and lemmatization. Recently, electronic dictionaries are developed for a number of different languages for general use. The drawback of electronic dictionaries is that although they are able to work as ordinary dictionaries, they are not able to differentiate the relationships and taxonomies between each keyword (e.g. cat and dog are both mammal and warm blooded) [3].

In 1998, researchers from Princeton University developed a lexical database for English (WordNet) [4]. WordNet is an extended electronic dictionary of the previous generation of English dictionaries, with improved functionalities such as synonyms, hyponyms, and hypernyms. WordNet has since been improved further, with extra functionalities for word similarities matching, gloss overlap, and semantic based results matching. To date, WordNet consists of 155,287 words organized in 117,659 synsets for a total of 206,941 word-sense pairs, which is one of the largest lexical databases in the world.

Research has also been carried out to develop hybrid version of WordNet, some of the notable examples are SUMO [15], [23], DOLCE [22], extension of DOLCE [9], WonderNet [13], and BabelNet [18], WordNet 2 [11], MultiWordNet [16], lexical based WordNet [20]. These hybrids incorporate higher level ontology to that of WordNet, making them highly useable for more complex processing. There are also WordNet that is sentiment based [1], image based [5], and ontology based [8]. For more information related to WordNet, the readers are encouraged to refer to the paper in [7].

Current WordNet versions are able to support a number of languages, such as English, French, Chinese, and Russian. A french multilingual WordNet is developed by [2]. Since WordNet is developed in 1998, various extensions for multilingual support is adopted. However, not all the languages in this world are supported by WordNet. For example, WordNet is not able to support Malay language. Supporting Malay language is important as it is one of the widely used languages in the South East Asia region. For example, Indonesia is by far the 4th most populous country in the world, and its citizen has widely adopted Bahasa Indonesia as their main language, which is very similar to Malay Language. Furthermore, Malay language has old and traditional history back to that of Melaka Sultanate, where its language is full of culture and tradition. Not only the Malays speak and write Malay Language well, other non native citizens also speak and write this language fairly well.

Existing lexical database such as WordNet is shown to be very useful for many applications. Semantic based applications, such as the search engines and social commerce platforms, widely adopted semantic based tools such as WordNet due to its ability to recognize and detect the rich semantics of textual content. Not only that, data intensive applications require semantic tools in order to further process its highly irregular and diverse data. This data is available in huge quantity, making them difficult for processing by manual processing. Finally, semantic based tools are highly capable for poll prediction and business intelligence. For example, the results on current election of Barrack Obama and Hillary Clinton can be preanalyzed so that further strategy can be made to optimize the outcome of the election. On the other hand, the business and consumer feedbacks on the current state of the art smartphones (e.g. Apple, Samsung) can be used to further analyze and strategize a company business plan and operation.

In this paper, we proposed a novel lexical semantic based database tool called MalayWordNet, specifically written for Malay language. Our initial study shows that it is feasible to develop Malay based lexical database as English and Malay languages share certain common representation. A careful and thorough investigation indicates that it is also possible to map English keywords to Malay keywords and constructs an entirely new lexical database based on the conventional English WordNet. In fact, new enhancements are proposed for our lexical database, which may help to further increase the accuracy of data intensive applications. Our tool is certainly helpful for high end semantic based applications which use Malay language as part of their data presentation.

This paper contains several sections. Section 2 describes the current work that is related to ours. Section 3 describes the problem in current lexical database tools while Section 4 provides the motivation for our research. Section 5 provides the methodological approach of our lexical database approach. In Section 6, we demonstrate experimental tests conducted on our method and novelty of our approach is provided in Section 7. Finally, Section 8 summarizes our work.

2.0 Related Work

2.1 Conventional WordNet

2.1.1 Overview

There are a number of ontological techniques available currently. Some of the common ones are CYC [10], WordNet [4], and SUMO [14]. However, we limit our discussion in this paper to the few state-of-the-art ontological tools. Details are presented in the next section.

2.1.2 WordNet

WordNet [4] was developed in 1998 as a light weight ontological technique, closer to a thesauri, and it is a lexical database for English for the semantic matching of words in Information Retrieval research [12]. WordNet contains a huge amount of information (150,000 words organized in over 115,000 synsets for a total of 207,000 word-sense pairs). WordNet represents nouns, adverbs, verbs and adjectives as a group of cognitive syno-nyms (synsets) with their own distinct concepts. Synsets are linked by means of conceptual semantic and lexical relations. A browser is used to manage and navigate the individual component in WordNet. It categorizes English words into several groups, such as hyper-nyms, synonyms, and antonyms.

2.1.3 CYC

CYC is developed by Lenat [10], [21] as part of his research work for MCC Corporation. Unlike WordNet, CYC covers a larger domain and provides more semantic information to the users. CYC provides more than hundreds thousands of terms, and millions of assertions related to the terms. The ontology in CYC knowledge has 47,000 concepts and 306,000 facts browsable by CYC web interface. CYC uses a mapping to define the con-cepts of each term. For example, CYC provides part of relationship between tree and leaves (leaves are part of a tree). Every concept mapped to the terms will return either a true or false statement. Based on this return value, users can then decide the appropriate actions for future processing. CYC has been successfully applied to Terrorism Knowledge Based application and has been used as part of Cyclopedia database (combining info taken from Wikipedia). However, studies indicate that CYC system and its underlying database is complicated, and it is also not scalable to large systems. An extension of CYC has also been developed [17].

2.1.4 BabelNet

BabelNet [18] is developed to overcome the drawback of WordNet. As stated in the literature of BabelNet, WordNet is a light weight ontological technique with limiting ontology domain and capability to provide sufficient information to the users. Using the combina-tion of WordNet and Wikipedia, BabelNet integrates the domain and knowledge base of these two systems, and could sufficiently provide the users with higher level ontology do-main. In addition, BabelNet is also able to distinguish word sense disambiguation accurate-ly using the information provided by Wikipedia domain knowledge.

2.1.5 YAGO

Yet Another Great Ontology (YAGO) is developed by Fabian and it is a lightweight on-tology with extensible functionalities for high data coverage and accuracy [6]. YAGO achieved an accuracy of 95% on its test cases. YAGO extracted data from Wikipedia and unified it with WordNet, and provides the users with 1 million entities and 5 million facts. YAGO also includes functionalities such as Is A as well as non taxonomical relations between entities.

2.1.6 WordNet++

WordNet++ is an extension of WordNet to solve word disambiguation problems. It extends the existing WordNet by providing extra high quality information from Wikipedia. WordNet++ could give high quality semantic information to the users, with support for word disambiguation using the interface of supervised tool Word Sense Disambiguation (WSD).

2.1.7 SUMO

Suggested Upper Merged Ontology (SUMO) provides the largest mapping of ontologies library where its library is used in search, linguistics, and reasoning. It is a formal ontology where all the keywords are mapped to WordNet lexicon and is written in SUO-KIF language.

2.1.8 Wikitology

Wikitology is an ontology tool developed based on the Wikipedia. It is useful as a tool for many language processing tasks. Each article is a concept in the ontology. The terms in the article are linked to each other and they may also interlink to other documents. Wikipedia ontology is created and maintained by diverse community. It has broad coverage, multilingual, and its content is very current. In fact, the quality of its content is very high, which is useful for many research works as it is maintained and created by trusted communities.

2.2 Multilingual Support

2.2.1 Overview

To the best of our knowledge, WordNet is the only ontological tool that has multilingual support for the users. Other ontological tools such as CYC and YAGO are in the development stages of providing support for various languages. Even though WordNet provides multilingual support for the users, it does not support the whole set of languages available in this world. Besides that, ontological tools such as WordNet++ extended from WordNet do not provide support for most of the languages supported by WordNet.

2.2.2 EuroWordNet

There are several different variant of WordNet developed for other languages, a notable example is EuroWordNet [19], [24], which is developed for European languages. EuroWordNet provides support for 5 types of European languages and is publicly available at http:// www.illc.uva.nl/EuroWordNet/. Similar to WordNet, users can use the publicly avail-able interfaces such as Java WordNet Interface (JWI) and Java WordNet Library (JWNL) to port their applications to EuroWordNet.

2.2.3 Other languages supported by WordNet

In addition to European Languages, WordNet also provides support for Chinese, Russian, Thai, Japanese, and Korean languages. Details of this work can be found at http://www.globalwordnet.org/gwa/wordnet_table.htm. It is noted however that all these software systems have different databases and mapping across their implementation due to the complexity and differences in their languages.

3. Problem Statement

With multilingual support, WordNet has been applied using a number of languages. As we are aware, WordNet has not been incorporated in Malay Language. Therefore, we aim to develop a Malay version of WordNet in our proposal. Works related to our proposal is the Asian WordNet (see http://asianwordnet.org/). In the early days, our country adopted English as part of our main language as we are a former British colony. Since we have achieved independence, Bahasa Malaysia has been developed and progressed further into a well-known language locally and internationally. In 2007, it is the 4th most spoken languages in the world (see http://www.ugmc.bizland.com/bmelayu.html). In the transition of our country's national language, most of the words used in Bahasa Malaysia are actually translation between English and Bahasa Malaysia. In fact, the terms used to define Bahasa Malaysia and English are actually very similar in many ways (verbs, adjectives, etc). Therefore, we are of the opinion that the idea of developing Malay WordNet is feasible and within the reach of our proposal. Several steps are given due considerations when conversion and translation are made. First of all, we conduct a random study where user study is carried out to test the correctness of mapping and translation on a preselected random samples. This step is repeated should the mapping and translation is incorrect for at least 5 times in the random samples. Then, a full and thorough checking is carried out for the en-tire library. Once the full checking is carried out, we then test the correctness of the mapping by applying sample test codes to the MalayWordNet library.

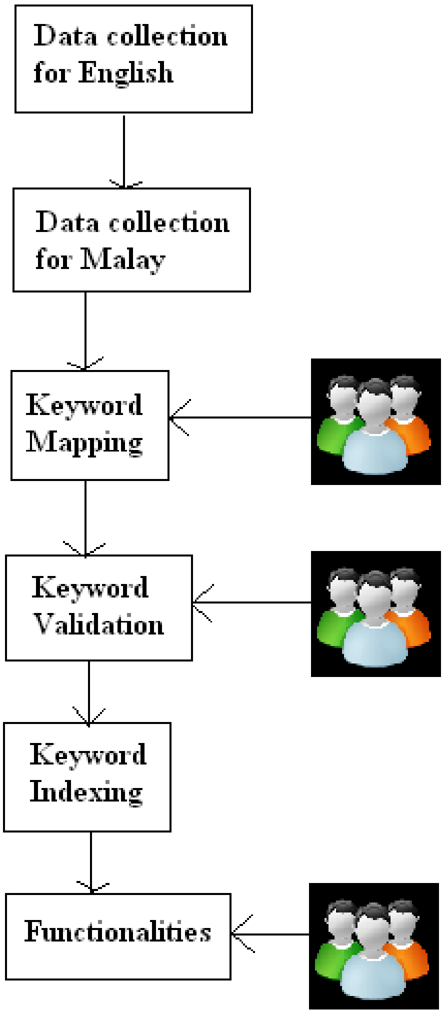

Figure 1. The Methodology of MalayWordNet

4. Motivation For Developing Malay Wordnet

The idea and motivation for developing Malay WordNet is a nontrivial task though it is feasible. This is because:

1. The word used in Malay language and English language is not a direct one to one mapping. For example, the word ‘faedah” could mean “interest in bank” or “your ad-vantage of being part of this team”. In fact, current translation engines are having great difficulties in translating and analyzing languages (see http://82.165.192.89/initial/ index.php?id=175). It is difficult and therefore incorrect to make an assumption that a word in the English language can be directly mapped to the Malay Language. However, a careful analysis and appropriate handling of the mapping will help in making the correct mapping.
2. The use of adjective in English language and Malay language is different. For English language, most of the root words are converted to adjectives by adding –ing (e.g. the word go will become going). In Malay, the root words are converted to adjectives by adding –me, -men, -meng (e.g. the word beli will become membeli). Furthermore, there is no current standard of applying the prefix and suffix to these words. To resolve this issue, we take extra precaution when applying the adjective between English and Ma-lay languages. A separate data structure is used to store the additional adjective for-mats in Malay language.
3. The same rules presented in Point 2 apply to other terms such as past tense, past participles, nouns, and so on. As the conversion of root words to other words is not similar in English and Malay languages, there are no currently acceptable standard for con-versions. Therefore, automated conversion of languages is highly unacceptable, one needs to manually analyze the words for English and Malay for conversion, as in the work of EuroWordnet [19].

The hierarchical structure and conceptual link presented in WordNet may be different in Malay WordNet, as these languages are highly ambiguous and there is no one to one mapping across each of the words. Therefore, the mapping of the word from English to Malay language has to be treated with proper care and detailed analysis with due consideration for the conversion to be carried out.

5. Proposed Methodology

The development of MalayWordNet can be divided into several stages (see Figure 1). These stages are as following:

1. Collection of data for English Language.
   We collect all the keywords of English Language and its related terms (e.g. verbs, adjectives etc) from the English dictionary in WordNet for our data collection. Root words are stored in a list, with their respective verbs, and adjectives. A data structure with its respective mapping of keywords is used to store these keywords.
2. Collection of data for Malay Language.
   We collect all the keywords of Malay Language and its related terms (e.g. kata kerja) from a well known electronic Malay dictionary Kamus Perdana, which is avail-able electronically. Words in this dictionary are tokenized and they are stored in a list, with their respective verbs, and adjectives. We choose Kamus Perdana as our source of information due to its highly comprehensive library available in electron-ic form.
3. Mapping of keywords between the two languages.

Once the dictionaries of both languages have been collected, we develop a system to map all the relevant keywords of these two languages. To achieve this, we use language translation engine such as Google Translate (see http://translate.google.com/#) to translate the meaning of these two keywords. Not all the words can be correctly mapped, as mapping of the words in Malay Language and English Language are not one to one. Words that are not correctly mapped are manually checked and mapped by the researchers. Otherwise, further treatment is required for the words that are incorrectly mapped. We either mapped them to other words, or create a multi mapped data structure where one word can be mapped to many different words. A special case may exist where multiple words are mapped to multiple other different words. We treat this case by providing multi graph data structure where many keywords are mapped to other different key-words.

4. Validation of keywords.
   After mapping of the keywords, we need to conduct a user survey and evaluation to check and validate our mapping. A team of 5 researchers are chosen where they will conduct a user survey to validate the mapping. A sample size of 100 partici-pants is chosen to validate the mapping. Words in the Malay Language are divided into 20 categories, sorted by their acronyms and alphabets. A team of 5 is assigned to each category, validation is considered successful when all the members fully agree with the mapping done. Otherwise, the mapping is treated as incorrect and needs to be repeated. Two sets of user study are conducted, the first of which is random sampling, and the second is the full validation. The first test requires validation on a smaller set of random samples collected from the data while the second test requires a more comprehensive checking on the dictionary.
5. Indexing of keywords.
   We'll then index all the keywords in MalayWordNet according to the index implemented in English WordNet. The indexing in MalayWordNet needs proper and due considerations, as the mapping between Malay and English Language are not one to one either. Indexing is done using a N-Ary Tree structure, priority is given when inserting words into the tree, particularly the hierarchy and time to search a key-word in the tree. The tree must be balanced and sorted, to ensure that the searching time is minimal. We choose N-Ary tree as the data structure to store the keywords as this tree has a reasonable searching time.
6. Extra functionalities.
   Once indexing is done, we'll create all the functionalities provided by WordNet such as synsets, hypernyms, hyponyms etc. The synsets, hypernyms, and hyponyms are implemented according to that of WordNet. Validation is carried out in this stage, with due consideration given to the mapping.
7. Complete systems.
   A fully working system for the users is developed where it is equipped with Graphical User Interface, an intuitive and easy to use menus, highly portable across many platforms, and an online system is currently under development. The library of MalayWordNet will be provided for free for the users.

6. Experimental Test

Once the full system is completed, we conduct an experimental study to evaluate the performance of our system. A team of 10 researchers are chosen to evaluate our product. We benchmark our tool against the state of the art tool WordNet. We aim to achieve high performance processing, where our tool could robustly provide the synsets of the keywords entered, within minimal time frame for searching. Though our tool is not able to identically provide the features offered in WordNet due to the complicated mapping process, we aim to achieve at least 95% compatibility with WordNet. We conduct rigorous testing for some of the components of our system, notably the search of keywords, and their related synsets, hyponyms, and hypernyms. Then we validate the correctness of the keywords entered with respect to WordNet. Validation is considered correct if the keywords entered perfectly matched that of WordNet.

Table 1. WordNet and MalayWordNet

Terms	WordNet	MalayWordNet
Keywords	150,000	148,000
Synsets	115,000	112,000
Word-Sense Pair	207,000	195,000

Table 2. Accuracy of MalayWordNet

MalayWordNet	Term	Accuracy
Words Matched	146,000	98.65%
Synsets Matched	111,000	99.11%
Word-Sense Pair	190,000	97.44%

Table 1 shows the number of keywords, synsets, and word sense pair available in WordNet and MalayWordNet. Our tool MalayWordNet has comparable number of keywords, synsets, and word sense pair to that of WordNet. In Table 2, we demonstrate the performance of our tool with WordNet after the validation is carried out. Our tool has more than 97% accuracy for all the features implemented, which indicates that the mapping provided in our tool nearly matched that of WordNet. Therefore, a user who has used WordNet will be able to use our tool MalayWordNet with little modifications made to the code and minimal portability issues. Further test shows that the time taken to process a search query on the keywords only take 0.02s. The fast search required to process a keyword is certainly helpful for data intensive applications.

7. Novelty Of Our Approaches

Our novel lexical database has significant contribution to the research domain. First, our tool is certainly useful for other application domains such as Information Retrieval, Text Mining, Multimedia and even other industries like Healthcare and Airline. Our tool is also useful for many data intensive applications, particularly Search Engine applications. Search Engines such as Google, Yahoo, Bing, and Facebook can use our tool for language translation and semantic processing. Current search engines such as Google use statistical matching of text for language translation. Due to the fact that our tool has semantic matching capability, it will be an advantage for current search engines as it provides additional higher level information. Secondly, our tool can be commercialized and patented for industry use. Besides, other countries such as Indonesia, Singapore, and Brunei will also show great interests in using our products as their citizens also use similar languages as ours. Further extension for our tool to support similar languages such as Bahasa Indonesia is also possible.

8. Conclusion

We have developed a novel semantic based lexical database tool called MalayWordNet. We have shown that it is feasible to develop MalayWordNet as English language and Malay language exhibit certain similarities in their structure and presentation. Our tool is certainly helpful for many data intensive applications written in Malay language. Furthermore, our tool can be applied to many industries and domains, such as governmental organizations, health care and tourism industries. Further works include extending our tool to support Bahasa Indonesia, a language which is similar to Bahasa Malaysia.

Acknowledgement

This work is carried out within the framework of a research grant funded by Taylor's University Research Grant Scheme (Project Code: TRGS/MFS/1/2013/SOCIT/03).

References

[1] S. Baccianella, A. Esuli and F. Sebastiani. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In Proceedings of the 7th Conference on Language Resources and Evaluation (LREC'10), Valletta, MT, 2010, pp. 2200–2204.

[2] S. Benoît, F. Darja. 2008. Building a free French wordnet from multilingual resources. In Proc. of Ontolex 2008, Marrakech, Maroc

[3] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, S. Hellmann, DBpedia – A crystallization point for the Web of Data. Web Semantics, 7(3), 2009, pp. 154–165

[4] Christiane Fellbaum, "WordNet: An Electronic Lexical Database," The MIT Press, Cambridge, MA, 1998.

[5] J. Deng, W. Dong, R. Socher, L. Li, K. Li, L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In Proc. of 2009 IEEE Conference on Computer Vision and Pattern Recognition

[6] Fabian M. Suchanek, GjergjiKasneci and Gerhard Weikum "Yago - A Core of Semantic Knowledge", 16th international World Wide Web conference, WWW 2007

[7] Francis Bond and Kyonghee Paik 2012a. A survey of wordnets and their licenses. In Proceedings of the 6th Global WordNet Conference (GWC 2012). Matsue. 64–71

[8] A. Gangemi, R. Navigli, P. Velardi. The OntoWordNet Project: Extension and Axiomatization of Conceptual Relations in WordNet, In Proc. of International Conference on Ontologies, Databases and Applications of SEmantics (ODBASE 2003), Catania, Sicily (Italy), 2003, pp. 820–838.

[9] Gangemi, A., Guarino, N., Masolo, C., Oltramari, A. 2003 Sweetening WordNet with DOLCE. In AI Magazine 24(3): Fall 2003, pp. 13–24

[10] Guha, R.V., Lenat, D.B., Building Large Knowledge Based Systems Reading, Massachusetts: Addison Wesley, 1990.

[11] S. M. Harabagiu, G. A. Miller, D. I. Moldovan. 1999. WordNet 2 – A Morphologically and Semantically Enhanced Resource. In Proc. of the ACL SIGLEX Workshop: Standardizing Lexical Re-sources, pp. 1–8

[12] Lassila, O. and McGuinness, D. (2001) The Role of Frame-Based Representation on the Se-mantic Web, Technical Report KSL-01-02, Knowledge Systems Laboratory, Stanford University, Stanford, California.

[13] Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A., Schneider, L.S. 2002. Won-derWeb Deliverable D17. The WonderWeb Library of Foundational Ontologies and the DOLCE on-tology

[14] Niles, I., and Pease, A. 2001. 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.

[15] A. Pease, I. Niles, J. Li. 2002. The suggested upper merged ontology: A large ontology for the Semantic Web and its applications. In Proc. of the AAAI-2002 Workshop on Ontologies and the Se-mantic Web, Edmonton, Canada.

[16] E. Pianta, L. Bentivogli, C. Girardi. 2002. MultiWordNet: Developing an aligned multilingual database. In Proc. of the 1st International Conference on Global WordNet, Mysore, India, pp. 21–25

[17] S. Reed and D. Lenat. 2002. Mapping Ontologies into Cyc. In Proc. of AAAI 2002 Confer-ence Workshop on Ontologies For The Semantic Web, Edmonton, Canada, 2002

[18] Roberto Navigli and Simone Paolo Ponzetto, BabelNet: Building a very large multilingual se-mantic network In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11-16 July 2010, pp. 216-225.

[19] Vossen, P., N. Calzolari, G. Adriaens, A. Sanfilippo, Y. Wilks (eds.) 1997 Proceedings of the ACL/EACL-97 workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Madrid, July 12th, 1997.

[20] Piek Vossen, Claudia Soria, Monica Monachini: Wordnet-LMF: a standard representation for multilingual wordnets, in LMF Lexical Markup Framework, edited by Gil Francopoulo ISTE / Wiley 2013

[21] http://www.cyc.com

[22] http://www.loa-cnr.it/DOLCE.html

[23] http://www.ontologyportal.org/

[24] http://www.illc.uva.nl/EuroWordNet/