Journal of Zhejiang University SCIENCE A
ISSN 1673-565X (Print); ISSN 1862-1775 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

An OWL-based WordNet lexical ontology*

HUANG Xiao-xi†[1,2], ZHOU Chang-le[2,3]

([1]*School of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China*)
([2]*Center for the Study of Language and Cognition, Zhejiang University, Hangzhou 310028, China*)
([3]*Institute of Artificial Intelligence, Xiamen University, Xiamen 361005, China*)
†E-mail: itshere@zju.edu.cn

Received Oct. 26, 2006; revision accepted Jan. 11, 2007

Abstract: This paper discribes a data representation for WordNet 2.1 based on Web Ontology Language (OWL). The main components of WordNet database are transformed as classes in OWL, and the relations between synsets or lexcial words are transformed as OWL properties. Our conversion is based on the data file of WordNet instead of the Prolog database. This work can be used to enrich the work in progress of standard conversion of WordNet to the RDF/OWL representation at W3C.

Key words: WordNet, Web Ontology Language (OWL), Semantic Web, Ontology, Natural Language Processing (NLP)
doi:10.1631/jzus.2007.A0864 **Document code:** A **CLC number:** TP18; H03

INTRODUCTION

The Semantic Web is a vision for the future of the Web in which information is given explicit meaning, making it easier for machines to automatically process and integrate information available on the Web (Berners-Lee *et al*., 2001). The primary goal is to enable intelligent services such as information brokers, search agents, information filters and queries for knowledge on the Web instead of the conventional string matching. The World Wide Web Consortium (W3C) released the Resource Description Framework (RDF) and the Web Ontology Language (OWL) as W3C Recommendations on Feb. 10, 2004, intending RDF to be used to represent information and to exchange knowledge on the Web, OWL to be used to publish and share sets of terms called ontology, supporting advanced Web search, software agents and knowledge management (Daly *et al*., 2004). WordNet (Fellbaum, 1998) is a machine-readable lexicon widely used in the Natural Language Processing (NLP) community, for tasks such as word sense disambiguation or information retrieval. Recently, Semantic Web researchers began to consider WordNet as an auxiliary tool for ontology annotation, ontology mapping, etc. Many efforts were made in conversion from WordNet's Prolog database to RDF/OWL, differing in design architecture and application scope (Ciorascu *et al*., 2003; Assem *et al*., 2006; Graves and Gutierrez, 2006). All the existing conversions are based on WordNet's Prolog format before Version 2.0. In this paper, we present a novel OWL representation from the original data files of the latest version of WordNet, Version 2.1. Differing from previous versions, WordNet 2.1 has distinguished hyponyms that are classes from hyponyms that are instances (Miller and Hristea, 2006).

The main motivations to our work on the development of OWL representation for WordNet lie in two aspects. Firstly, no reasoning procedure is provided by WordNet, which provides information about concepts that correspond to word senses only. To use WordNet as a generic ontology or semi-ontology, we should provide a mechanism to reason with the information provided by WordNet. If we can represent WordNet as an ontology with OWL language, we can use OWL formal semantics to specify

* Project supported by the National Natural Science Foundation of China (No. 60373080) and the 985 Project of Zhejiang University, China

how to derive various logical consequences, then WordNet will be enriched with reasoning capability. Secondly, an OWL representation of WordNet is of great use in many applications, such as computer understanding system (Aref and Zhou, 2005), Web services (Bansal *et al*., 2005), semantic annotation, etc.

There are other projects focusing on lexicon meta-models. Lexical Markup Framework (LMF) (Francopoulo *et al*., 2006) is a model in progress that provides a common standardized framework for lexicons of NLP systems. The main motivations of LMF are to provide a common model for construction, use and management of lexical resources which may be monolingual, bilingual or multilingual. There is also a project similar to LMF called LIRICS (http://lirics.loria.fr). As for Aisan languages, Takenobu *et al*. (2006) proposed a framework based on MILE (Multilingual ISLE Lexical Entry) which aimed for a standardized infrastructure to develop multilingual lexical resources for various NLP applications. The framework of (Takenobu *et al*., 2006) is also situated in the context of W3C standards.

In the next section, we briefly introduce the machine-readable dictionary WordNet, and some statistics on WordNet 2.1. Section 3 introduces the Web Ontology Language. Section 4 then describes the architecture of OWL representation of WordNet. In Section 5, we briefly introduce the application of this lexical ontology in our metaphor understanding system. Finally, Section 6 gives the conclusions and future work

WORDNET AND ITS STRUCTURE

WordNet is a widely used machine-readable lexicon in NLP. It is "a semantic dictionary that was designed as a network, partly because representing words and concepts as an interrelated system seems to be consistent with evidence for the way speakers organize their mental lexicons" (Fellbaum, 1998). A sense is organized as a synonym set (synset), namely, a concept. For example, in WordNet 2.1, the synset {person, individual, someone, somebody, mortal, soul} represents the concept with gloss of "a human being". Every synset consists of a list of synonymous word forms and semantic relations that describe relationships between the current synset and other synsets. A word form can be a single word or two or more words (referred to as collocations). WordNet includes four parts of speech (POS): nouns, verbs, adjectives and adverbs. There are 155327 words (147249 words ignoring POS), organized in 117597 synsets and 207016 word-sense pairs in WordNet 2.1. The detailed statistics about every POS of WordNet is listed in Table 1 (http://wordnet.princeton.edu).

Table 1 The number of words, synsets, and word-sense pairs in WordNet 2.1

POS	Unique strings	Synsets	Total word-sense pairs
Noun	117097	81426	145104
Verb	11488	13650	24890
Adjective	22141	18877	31302
Adverb	4601	3644	5720
Total	155327	117597	207016

There are several kinds of relations used to connect the different types of synsets. Table 2 shows the relations that are used to connect synsets along with their frequency counts, which are gained from the database with our programs. The semantic relations link the synsets to form a network. Some relations that hold between word forms have also been included in WordNet, such as derivational relatedness.

Table 2 shows that the portions of WordNet for each POS also have different properties, and may therefore require special treatments. For example, while the hypernymy/hyponymy and holonym/meronym relations are central to the organization of the nouns of WordNet, adjectives are organized primarily in terms of the antonymy and similarity relations. After investigating the WordNet database thoroughly, we concluded that the semantic relations can be divided into two kinds according to the element types of the relation, and we call them lexical linker and semantic linker respectively, where lexical linkers connect specific word senses while semantic linkers connect synsets. According to this distinction, we have grouped the total of 26 semantic pointers into 3 groups, as shown in Table 3. Lexical linker connects only between word senses, while semantic linker connects synsets including all the words in it.

Table 2 Relations and frequency counts by the linker type in WordNet 2.1

Relations	Noun	Verb	Adjective	Adverb
hypernym	75 134	13 124	–	–
hyponym	75 134	13 124	–	–
instance hypernym	8 515	–	–	–
instance hyponym	8 515	–	–	–
part holonym	8 874	–	–	–
part meronym	8 874	–	–	–
member holonym	12 262	–	–	–
member meronym	12 262	–	–	–
substance holonym	793	–	–	–
substance meronym	793	–	–	–
attribute	643	–	643	–
domain category	4 147	1 237	1 113	37
domain member category	6 534	–	–	–
domain region	1 247	2	76	2
domain member region	1 327	–	–	–
domain usage	942	16	227	73
domain member usage	1 258	–	–	–
entail	–	409	–	–
cause	–	219	–	–
also	–	589	2 683	–
verb group	–	1 748	–	–
similar	–	–	22 622	–
antonym	2 142	1 089	4 080	718
derivation	35 901	23 095	12 911	1
participle	–	–	124	–
pertainym	–	–	4 852	3 213
Total	265 297	54 652	49 331	4 044

ONTOLOGY LANGUAGES FOR SEMANTIC WEB (OWL)

The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation (Berners-Lee *et al*., 2001). A way to achieve this goal is to give the information on the Web a well-defined meaning. Several markup languages are developed for this purpose. Fig.1 shows the layer cake model for Web languages. The bottom two layers are the syntax basis of Semantic Web. The first semantic layer of Semantic Web is provided by RDF (Klyne and Carroll, 2004), which is a basis of Web metadata processing and is used for describing relationships between resources. The formal model for RDF can be represented as triples: <predicate, subject, object>. It has XML-based syntax and model semantics. RDFS stands for RDF Schema. It is a set of ontological modeling primitives on top of RDF model (Brickley and Guha, 2004). RDFS introduces classes, properties, subsumptions between classes, subsumptions between properties, and the domain and range of properties, on the top of RDF. Thus, RDFS can be viewed as an extensible, object-oriented type system based on RDF. The expressive ability of RDFS is finite, for example, it cannot define the transitivity or symmetry of properties. So it is necessary to introduce stronger ontology languages. OWL is such a language to enrich the expressibility of RDF(S).

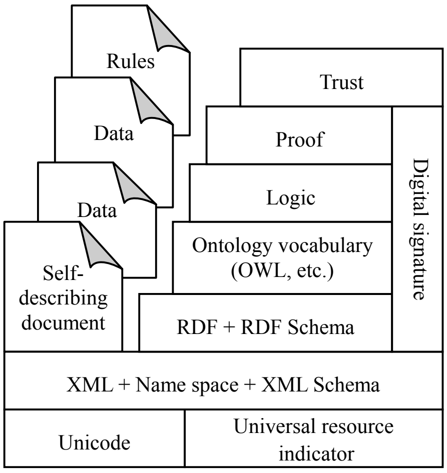

Fig.1 Layer of Web Language [adapted with changes from (Hendler, 2001)]

Table 3 Classification of semantic pointers in WordNet 2.1

Relation types	Classification standard	Semantic pointers
Lexical linker	Between senses	Antonym, participle, pertainym, derivation
Semantic linker	Between synsets	Hypernym, hyponym, instance hypernym, instance hyponym, part holonym, part meronym, member holonym, member meronym, substance holonym, substance meronym, entail, cause, similar, attribute, verb group
Both	Between senses or synsets	Also, domain category, domain member category, domain region, domain member region, domain usage, domain member usage

OWL is the W3C standard for representing ontologies on the Semantic Web. Ontologies are expected to play a key role in Semantic Web applications by providing a source of shared and precisely defined terms that can be used for describing Web resources. OWL is also designed to be compatible with the syntax of XML (Yergeau *et al*., 2004) and RDF (Hayes, 2004). Furthermore, OWL is also compatible with DAML+OIL (Horrocks, 2002), the most immediate predecessor of OWL, which was deeply influenced by description logics and frame systems. As trade-offs of the influences of description logics knowledge bases and frame systems, there are three sub-languages with increasing expressive power: OWL Lite, OWL DL and OWL Full, where the expressive ability of OWL DL is the same as Description Logic SHOIN(D) (Horrocks and Sattler, 2005). In OWL DL, the terms class, object property, datatype property, individual and datatype stand for concept, role, concrete role, object and concrete domain respectively in SHOIN(D), thus, it is convenient to use the reasoning theory in description logics in OWL. Our conversion of WordNet to OWL will be restricted in OWL DL.

ARCHITECTURE OF WORDNET OWL REPRESENTATION

We created our WordNet OWL model on the existing RDF/OWL model developed by the WordNet Task Force (Assem *et al*., 2006) and RDF model developed by Graves and Gutierrez (2006). Fig.2 shows the class hierarchy of the OWL Representation. The main classes are Synset, WordSense and Word, which are the same as those described in (Assem *et al*., 2006), where collocation was introduced as a separate class under Word. We think the collocations can be treated as single words through the property Lemma of Word (Table 5). With deep investigation on WordNet database, we found that there are about 40545 cased words, most of which are proper nouns or collocations. So we treat these words as a separate class named CasedWord as a subclass under Word. To include the morphology exception lists of noun, adjective, verb and adverb, we introduce a property exceptiveWord for Word.

Referred to the W3C Schema (Assem *et al*., 2006) and Graves' Schema (Graves and Gutierrez, 2006), we also divide the whole structure into three layers, namely, Word Layer, Sense Layer and Concept Layer as shown in Fig.3. Every WordSense represents one sense of a word, while a word can have several WordSense as polysemy. As the definition of synset in WordNet, every synset may have several WordSense to construct a concept, namely synset. According to Table 3, the lexical relations are located in the second layer, and semantic relations are located over the third layer, others are in both second and third layers. Tables 4 and 5 list the datatype properties and object properties defined in this schema respectively. The domain and range of each property are determined through Table 3. The schema was validated as OWL DL by the WonderWeb OWL Ontology Validator at http://phoebus.cs.man.ac.uk:9999/OWL/Validator.

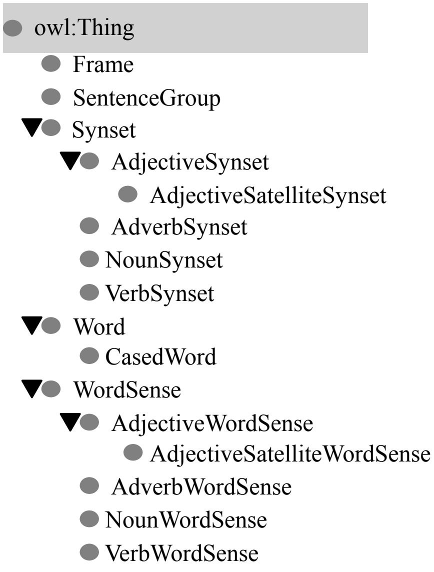

Fig.2 Class hierarchy of WordNet OWL representation

Table 4 The Datatype Properties in WordNet 2.1 OWL representation

Property name	Domain	Range
lemma	Word	xsd:string
exceptiveWord	Word	xsd:string
frameDescription	Frame	xsd:string
SentenceFrame	SentenceGroup	xsd:string
lexId	WordSense	xsd:string
senseNumber	WordSense	xsd:string
tagcount	WordSense	xsd:string
synsetId	Synset	xsd:string
gloss	Synset	xsd:string
sample	Synset	xsd:string

Table 5 The Object Properties in WordNet 2.1 OWL representation

Property name	Domain	Range	Inverse
hasSense	Word	WordSense	
antonymTo	WordSense	WordSense	antonymTo
casedWord	WordSense	CasedWord	
hasSynset	WordSense	Synset	
hasWord	WordSense	Word	
hasFrame	VerbWordSense	Frame	
hasSentenceGroup	VerbWordSense	SentenceGroup	
participle	AdjectiveWordSense	VerbWordSense	
pertainsTo	AdjectiveWordSense ⊔ AdverbWordSense	NounWordSense ⊔ AdjectiveWordSense	
hasAttribute	NounSynset	AdjectiveSynset	attributeOf
holonymOf	NounSynset	NounSynset	meronymOf
memberHolonymOf	NounSynset	NounSynset	memberMeronymOf
meronymOf	NounSynset	NounSynset	holonymOf
partHolonymOf	NounSynset	NounSynset	
substanceHolonymOf	NounSynset	NounSynset	
instanceHypern	NounSynset	NounSynset	instanceHyponym
instanceHyponym	NounSynset	NounSynset	instanceHypern
classifies	NounSynset ⊔ NounWordSense	Synset ⊔WordSense	classifiedBy
hypernym	NounSynset ⊔ VerbSynset	NounSynset ⊔ VerbSynset	hyponym
hyponym	NounSynset ⊔ VerbSynset	NounSynset ⊔ VerbSynset	hypernym
classifiedBy	Synset ⊔ WordSense	NounSynset⊔ NounWordSense	classifies
cause	VerbSynset	VerbSynset	causeBy
entails	VerbSynset	VerbSynset	entailedBy
sameVerbGroupAs	VerbSynset	VerbSynset	sameVerbGroupAs

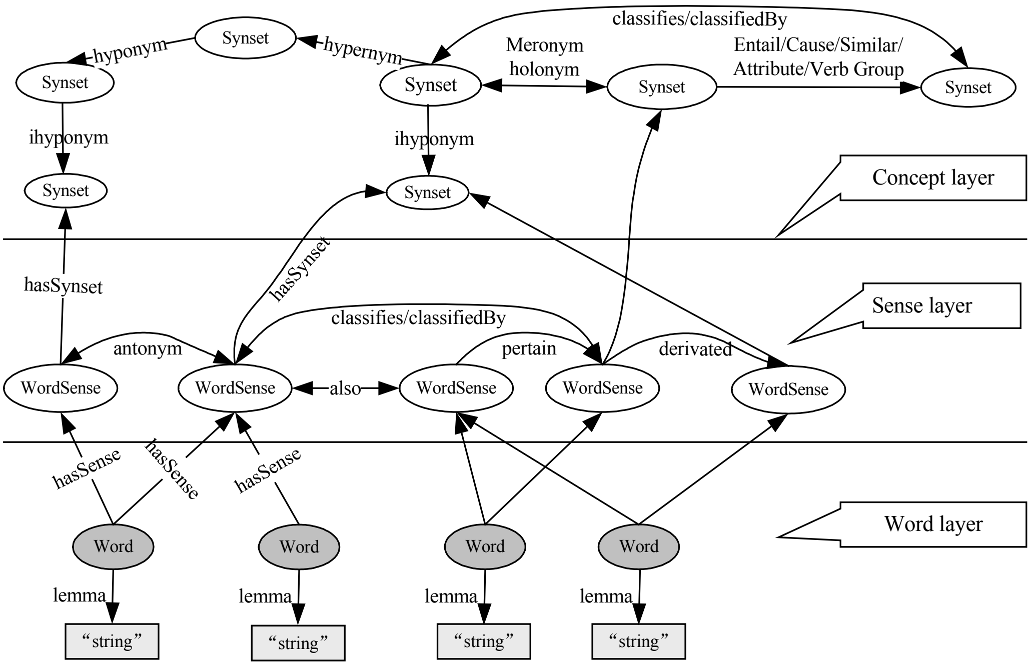

Fig.3 Schema of OWL representation for WordNet

Most of the properties defined in Table 5 are the same as in (Assem *et al*., 2006). The differences lie in the following: Firstly, the instance hyponym relation of WordNet 2.1 is defined as a sub-property of hyponym. And it is further restricted to NounSynset by owl:restriction and owl:allValuesFrom. Secondly, we introduce two new classes, Frame and SentenceGroup, to express the corresponding concept used in WordNet, instead of defined as xsd:string in (Assem *et al*., 2006). Thirdly, although the domain and range of some properties are defined in the form of union of two classes, such as pertainsTo, classifies, etc., as for specific class, it is also restricted to a definite class through owl:restriction and owl:allValuesFrom. For example, the property pertainsTo of AdverbWordSense is restricted to AdjectiveWordSense by the assertion "∀pertainsTo.AdjectiveWordSense", instead of being divided into two properties "adjectivePertainsTo" and "adverbPertainsTo" in (Assem *et al*., 2006).

APPLICATION

To illustrate the use of the model presented above, we are building a metaphor understanding system which uses the WordNet model as lexical knowledge source. We have proposed a logical approach to metaphor understanding (Huang and Zhou, 2005). Metaphor is a pervasive phenomenon in natural language. Metaphor understanding by computer has interesting applications in many NLP communities like machine translation, text summarization, information retrieval and question answering. The components of a metaphor are called target and source respectively. Fig.4 shows the system architecture in development.

A metaphorical sentence is firstly parsed by syntax and semantic parser. Here, we use MINIPAR, a broad-coverage parser available from (Lin, 1998). The system then uses lexical ontology and domain ontology to extract the components of the metaphor, namely source domain knowledge and target domain knowledge. The WordNet Lexical Ontology contributes to get the lexical knowledge. Then, the metaphorical meaning of the sentence is processed through the metaphor logic system (Huang and Zhou, 2005), based on the knowledge of source and target, where the WordNet Lexical ontology contributes to get the analogical or similar content between source and target.

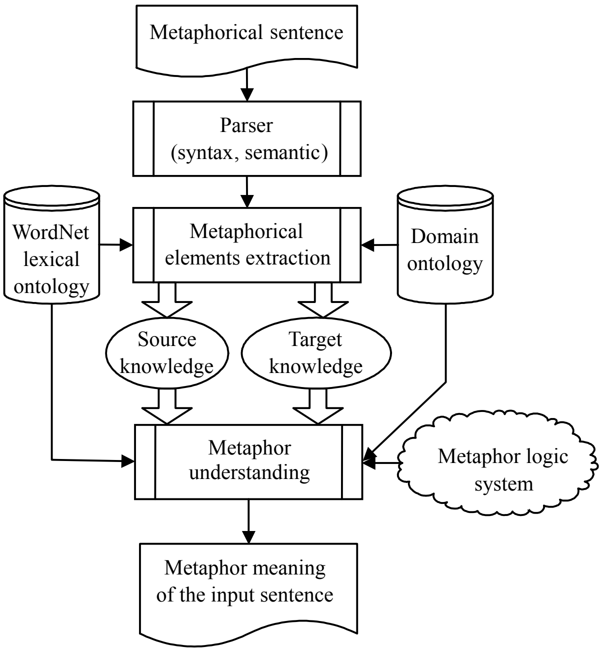

Fig.4 The system architecture of metaphor understanding

CONCLUSION AND FUTURE WORK

In this paper, we proposed a data representation for WordNet 2.1 based on OWL. The main components of WordNet database are transformed as classes in OWL, and the relations between synsets or lexcial words are transformed as OWL properties. Our conversion is based on the data file of WordNet instead of the prolog database. This work can be used to enrich the work in progress of standard conversion of WordNet to the RDF/OWL representation at W3C.

In future work, we will attach VerbNet to this conversion. In this design, VerbNet and WordNet can share the words through the classes Word, WordSense and Synset. Furthermore, VerbNet's frame and role structure can enrich the architecture of the OWL respresentation of WordNet.

References

Aref, M.M., Zhou, Z., 2005. The Ontology Web Language (OWL) for a Multi-Agent Understating System. Proc. IEEE Integration of Knowledge Intensive Multi-Agent Systems. IEEE Computer Society, Waltham, USA, p.586-590. [doi:10.1109/KIMAS.2005.1427149]

Assem, M.V., Gangemi, A., Schreiber, G., 2006. Conversion of WordNet to a Standard RDF/OWL Representation. Proc. 5th International Conference on Language Resources and Evaluation. Genoa, Italy, p.237-242.

Bansal, A., Kona, S., Simon, L., Mallya, A., Gupta, G., Hite, T.D., 2005. A Universal Service-Semantics Description Language. Proc. 3rd European Conference on Web Services. IEEE Computer Society, p.1-12. [doi:10.1109/ECOWS.2005.4]

Berners-Lee, T., Hendler, J., Lassila, O., 2001. The semantic Web. *Scientific American*, **284**(5):34-43.

Brickley, D., Guha, R.V., 2004. RDF Vocabulary Description Language 1.0: RDF Schema. Http://www.w3.org/TR/rdf-schema, W3C.

Ciorascu, C., Ciorascu, L., Stoffel, K., 2003. knOWLer—Ontological Support for Information Retrieval Systems. Proc. 26th Annual International ACM SIGIR Conference, Workshop on Semantic Web. Toronto, Canada.

Daly, J.J., Forgue, M.C., Hirakawa, Y., 2004. World Wide Web Consortium Issues RDF and OWL Recommendations. Http://www.w3.org/2004/01/sws-pressrelease.html.en,W3C.

Fellbaum, C., 1998. WordNet: An Electronic Lexical Database. MIT Press, Cambridge.

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M., Soria, C., 2006. LMF for Multilingual Specialized Lexicons. LREC Workshop on Acquiring and Representing Multilingual, Specialized Lexicons: The Case of Biomedicine. Genova, Italy, p.27-32.

Graves, A., Gutierrez, C., 2006. Data Representation for WordNet: A Case for RDF. Proc. 3rd Global WordNet Association Conference. Jeju Island, Korea.

Hayes, P., 2004. Resource Description Framework (RDF) Semantics. Http://www.w3.org/TR/2004/REC-rdf-mt-20040210/, W3C.

Hendler, J., 2001. Agents and the semantic Web. *IEEE Intell. Syst.*, **16**(2):30-37. [doi:10.1109/5254.920597]

Horrocks, I., 2002. DAML+OIL: a description logic for the semantic Web. *IEEE Data Eng. Bull.*, **25**(1):4-9.

Horrocks, I., Sattler, U., 2005. A Tableaux Decision Procedure for SHIOQ. Proc. 19th International Conference on Artificial Intelligence. Morgan Kaufman, p.448-453.

Huang, X.X., Zhou, C.L., 2005. A Logical Approach for Metaphor Understanding. Proc. 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering. IEEE Computer Society, Wuhan, China, p.268-271. [doi:10.1109/NLPKE.2005.1598746]

Klyne, G., Carroll, J., 2004. Resource Description Framework (RDF): Concept and Abstract Syntax. Http://www.w3.org/TR/rdf-concepts/, W3C.

Lin, D., 1998. Dependency-based Evaluation of MINIPAR. Proc. Workshop on the Evaluation of Parsing Systems. Granada, Spain, p.298-312.

Miller, G.A., Hristea, F., 2006. WordNet nouns: classes and instances. *Computational Linguistics*, **32**(1):1-3. [doi:10.1162/coli.2006.32.1.1]

Takenobu, T., Sornlertlamvanich, V., Charoenporn, T., Calzolari, N., Monachini, M., Soria, C., Huang, C.R., Xia, Y.J., Yu, H., Prevot, L., *et al*., 2006. Infrastructure for Standardization of Asian Language Resources. Proc. COLING/ACL 2006 Main Conference Poster Sessions. Association for Computational Linguistics. Sydney, Australia, p.827-834.

Yergeau, F., Bray, T., Paoli, J., Sperberg-McQueen, C.M., Maler, E., 2004. eXtensible Markup Language (XML) 1.0. Http://www.w3.org/TR/REC-xml/, W3C.