Extended-HowNet- A Representational Framework for Concepts

Keh-Jiann Chen, Shu-Ling Huang, Yueh-Yin Shih, Yi-Jun Chen

Institute of Information Science
Academia Sinica, Taipei, Taiwan
kchen@iis.sinica.edu.tw, {josieh, yuehyin, chenyijun}@hp.iis.sinica.edu.tw

Abstract

Natural languages are means to denote concepts. However word sense ambiguities make natural language processing and conceptual processing almost impossible. To bridge the gaps between natural language representations and conceptual representations, we propose a universal concept representational mechanism, called Extended-HowNet, which was evolved from HowNet. It extends the word sense definition mechanism of HowNet and uses WordNet synsets as vocabulary to describe concepts. Each word sense (or concept) is defined by some simpler concepts. The simple concepts used in the definitions can be further decomposed into even simpler concepts, until primitive or basic concepts are reached. Therefore the definition of a concept can be dynamically decomposed and unified into Extended-HowNet at different levels of representations. Extended-HowNet are language independent. Any word sense of any language can be defined and achieved near-canonical representation. For any two concepts, not only their semantic distances but also their sense similarity and difference are known by checking their definitions. In addition to taxonomy links, concepts are also associated by their shared conceptual features. Fine-grain differences among near-synonyms can be differentiated by adding new features.

1. Introduction

Ontology is a specification of a conceptualization (Gruber, 1993). We proposed a frame-based entity-relation knowledge representation model called Extended-HowNet, which was evolved from HowNet (Dong & Dong, `http://www.keenage.com/`), to encode concepts. Concepts are represented and understand by their definitions and association links to other concepts. In Extended-HowNet, we define each lexical sense by simple concepts which are not necessary to be primitive concepts. The vocabularies used for definitions are WordNet synsets (Fellbaum, 1998). The advantage of using WordNet synsets is that each synset has unique sense and sense similarity between two synsets can be measured through WordNet ontology. The following examples illustrate partially what we intended to achieve.

e.g. <science fiction> Def:= {<book>: content = {<imagination>: domain= {<science>}}}; which says that a science fiction is a book with imaginary content in science domain.

In section 2, background works regarding lexical knowledge representation are introduced. Section 3 describes the formal definition of Extended-HowNet. Advantages of the system is addressed at section 4. Summarization and conclusion are drawn in section 5.

2. Backgrounds

To achieve natural language understanding, computer systems should know the sense similarity and dissimilarity of two sentences or two words. To achieve above goals, it requires supports of ontologies. Ontology provides the following functions.

a) Identifies synonym concepts and measures similarity distance between two concepts.
b) Knows the shared semantic features and feature differences between two concepts.
c) Provides unique index to each concept, such that associated knowledge can be coded and accessed.
d) Language independent sense encoding.
e) Logical inferences through conceptual property inheritance system.
f) Dynamic concept decomposition and composition mechanisms.

None of the currently available ontology provides all of the above functions. We intend to propose a sense representation framework extended from HowNet, to achieve the above functions.

2.1 WordNet-like ontologies

WordNet (Fellbaum, 1998) contains information about nouns, verbs, adjectives and adverbs in English and is organized around the notion of a synset. A synset, roughly denoted a concept, is a set of words with the same part-of-speech that can be interchanged in a certain context. For example, {car; auto; automobile; machine; motorcar} form a synset because they can be used to refer to the same concept. Synsets can be related to each other by semantic relations, such as hyponymy, meronymy, cause, etc and a synset is often further described by a gloss: “4-wheeled; usually propelled by an internal combustion engine”.

The disadvantage of WordNet-like ontologies is that each concept class has limited linking to other concepts. The major links are hyponymy relations which limit inheritance and inference capability to the classes on the taxonomy. For those features without used as classification criterion will not be possible to encode their inherent properties. For instances, the set of round objects, eatable things will not be natural classes in the taxonomy. Therefore there will not be any general inference rules, such as (roll @round object), (digest @eatable things) can be encoded for such class of concepts.

2.2 HowNet

HowNet is an on-line common-sense knowledge base unveiling inter-conceptual relations and inter-attribute relations of concepts as connoting in lexicons of the Chinese and their English equivalents. Introduction of HowNet can be seen at http://www.keenage.com/zhiwang/e_zhiwang.html.

Conventional sense representation tried to use semantic primitives to define and achieve canonical representation for concepts (Wizebicka, 1972), such as Conceptual Dependency representation (Schank, 1975) and HowNet. Using primitives to define concepts causes information degrading. It is almost impossible to understand a definition of a complex concept. Furthermore it is debatable that there exist a limited and fixed set of so-called primitives. In HowNet, word sense definition is restricted to a set of around two thousands primitive concepts, called sememes. A word sense is defined by its hypernymy sememe and additional semantic features. For instance, the HowNet definition of Warrior|戰士 is:

{human|人:belong={army|軍隊},
{fight|爭鬥:
agent={~},
domain={military|軍}}}

which says that a warrior is a human in army who plays the role of agent in the event of military fighting.

3. Extended-HowNet

We will use the notation <word> to denote the word sense. However a word may have ambiguous senses. Additional modifiers will be used for differentiation.[1] For instances, <money bank> and <river bank>, each denotes the concept of money bank and river bank. Therefore a word may have different senses and sense definitions. On the other hand, synonyms have same word senses and sense definitions. We adopt similar mechanism in HowNet to define word sense, except that a concept is defined by simpler or synonym concepts instead of semantic primitives only. For instance, <man> is a <human> of <male> gender, which is defined as <man>:={<human>: gender={<male>}}. It is similar to a conventional tree-like feature structures. Formally, the syntax for Extended-HowNet is shown in Table1.

The current version of HowNet uses about two thousand sememes. The set of sememes are also adopted at Extended-HowNet for the ground-level definitions. In Extended-HowNet, new concepts are defined by any well-defined concepts and a definition can be dynamically decomposed into lower level representations until ground-level definition is reached, in which all features in the definitions are sememes. For instance, the top level definition of <department of literature|文學系> is {<school department|學系>: predication= {<teach|教>: location={~}, theme={<literature|文>}}}. Since the concept <school department|學系> is not a primitive concept, the above definition can be further extended into the primitive level definition, {<InstitutePlace|場所>: domain = {<education|教育>}, predication= {<study|學習>: location={~}}, predication= {<teach|教>: location = {~}, theme={<literature|文>}}}.

[1] In order to denote the referred word sense, we use additional modifier or its hypernym for disambiguation. For instance, <request-ask> denotes the sense of request not the sense of questioning. For the notational simplicity, we will also use head word only, such as <teacher>, <volcano>,…, etc, to denote the referred concept, if without ambiguities.

Concept:={Hyper-concept : Feature,…, Feature} or {Concept} or {Sememe};	which means that a concept may be defined by 1) its hypernym concept and semantic features, or 2) a synonym concept, or 3) a primitive concept.
Features:=Relation(x)={Concept};	which says that a semantic feature is expressed by a (Relation, Concept) pair, which denotes the semantic relation (Relation) between semantic feature (Concept) and the argument x. Arguments are in the range of { ~, Speaker, Listener,…}, where ~ denotes the Hyper-concept in definition, such as <human> in the definition of <man>. Since arguments x used in the Relation(x) are mostly head hypernym concept in definitions, the feature representation of Relation(~)={Concept} will be abbreviated as Relation= {Concept}.
Relation:={property, content, host, location, agent, patient,….};	which is a set of semantic relations.
Sememe:={good, bad, do, die, alive, water, building,…,Speaker, Listener}	which is a set of semantic primitives. Each element of sememe will be replaced by its respective WordNet synsets for the purpose of universality. For readability, the notation for primitive concepts will be using <word| word's Chinese translation> hereafter.

Table 1. The syntax for Extended-HowNet

To describe precise definitions for concepts, several technical problems have to be solves. In order to achieve unambiguous definitions, each referred concept should be unambiguous. In Extended-HowNet, WordNet synsets were adopted as the vocabulary for conceptual indexing and representation. Second, what are major features of a concept which suffice to define the concept? The issue is discussed in the section 2.1. Third, the semantic composition and decomposition involve feature unification. During unification processing, feature values under the same relation type should be unified together. For instance, in the above example, the hypernym class <school department> of <department of literature|文學系> is not a primitive concept and was extend to the definition of {<InstitutePlace|場所>: domain={<education|教育>}, predication={<study|學習>: location={~}}, predication= {<teach|教>: location={~}}, and the reduplicated feature of predication={<teach|教>:location={~}, theme={<literature|文>}} is then unified.

Formally Extended-HowNet is a feature unification system of a quadruple of (Vocabulary, Grammar, Taxonomy of Concepts, Taxonomy of Relations), where

a) Vocabulary= WordNet Synsets;
b) Grammar= the above defined syntax for Extended-HowNet;
c) Taxonomy of concepts= the hierarchical structure of concepts and sememes formed by hyponym and part-whole relations;
d) Taxonomy of Relations= the hierarchical structure of the relations formed by hyponym relations.

In Extended-HowNet, we intend to unify WordNet and HowNet taxonomies as the taxonomy of concepts and combine the semantic relations of FrameNet, HowNet to form the taxonomy of relations.

3.1 Principles for Concept Definition

Meaning of a concept is supported by its associated concepts including its formal properties, constituents, purposes, relations to other concepts etc. To define a concept, it is not possible to encode all its associated relations. The principle for defining a concept is that first identify its immediate hypernym and then encode its most important features which suffice to differentiate this concept with other concepts. In principle, the qualia structure is the major features for a nominal-type concept (Pustejovsky, 1995) and event frame is for an event-type concept (Fillmore, *FrameNet*). The qualia of an object are (Pustejovsky, 1995):

a) Constitutive: the relation between an object and its constituents, such as material, parts, components etc..
b) Formal: which distinguishes the object within a larger domain, such as shape, magnitude, color etc..
c) Telic: purpose and function of the object.
d) Agentive: factors involved in the origin or "bringing about" of an object.

There are two different types of attribute features. One is simplex attribute type and another is complex relative clause type. The simplex attribute is a feature-value type and the value is expressed by some discrete elements. For the complex attribute the attribute relation is an eventive feature. The constitutive and formal properties can be represented by simple attribute-value pairs, i.e. Relation={Concept} pair, in Extended-HowNet. The telic and agentive properties are usually represented by eventive features which are event frames. For instances, the concepts of <teacher> and <student> may be defined and differentiated as <teacher>:={<human>:telic={<teach>:agent={~ }}} and <student>:= {<human>: telic= {<teach>: goal={~}}}. Event-type concepts are also defined by their hypernym event-type and brotherhood concepts are differentiated by their event frame elements which include participant roles and adjuncts as well as their semantic restrictions. For instances, according to FrameNet II, both <request-appeal> and <request-ask> have the sense of <communication-request>. They are differentiated by their manners:

<request-appeal> Def:=
{<commu-request>:
manner= <formal>}
<request-ask> Def:=
{<commu-request>:
manner= <informal>}

Noticed that the event frame and other features of <request-appeal> and <request-ask> are inherent from the event frame of <commu-request> which has participant roles of Speaker, Addressee, Message, and Topic.

3.2 Consistency and Integrity of Representations

The integrity of concept representation is supported by the dynamic conceptual associations within Extended-HowNet system. As we know, meaning of a word is expressed by its associations to other concepts. In Extended-HowNet system, a concept is associated to other concepts through

a) Taxonomies, such as SUMO (Niles & Pease, 2001), WordNet, HowNet, FrameNet, SIMPLE-CLIPS and EuroWordNet. The association relations include synonymy, hyponymy, antonymy, meronymy etc...

b) Dynamic definition extensions. High-level features (i.e. concepts) provide easy encoding for general knowledge. Usually important conceptual properties are associated with basic concepts not primitive concepts. For instances, Pluto is a dog and dog is a basic concept. To define a basic concept by primitive concepts is possible, but does not help too much to understand the basic concept. For instance, the associated properties of dogs, such as “dogs bark”, “dogs are pets”… are hardly associated with the primitive concept of ‘animal’.

For the representational consistency, it is obvious that inconsistent representation for synonyms may occur due to encoding ambiguities and complexities of representations. However the dynamic feature extension in Extended-HowNet provides a way to ensure that similar senses have similar sememe-level representations and multi-variant similarity measurements.

3.3 Feature Inheritance and Conceptual Extension

The meaning of a concept is supported by its associated concepts. In Extended-HowNet, the defined concepts forms a hierarchical structure by is-a (hyponymy) relations. It is obvious that the associated property or knowledge regarding a particular concept can be accessed or encoded directly through its definition or indirectly inherent from its ancestors. Furthermore, the hierarchical taxonomy also provides a semantic distance between two concepts. However conventional taxonomies do not provide the exact semantic similarities and dissimilarities of two concepts. In Extended-HowNet, definitions of concepts not only provide semantic similarities but also encode semantic difference of two concepts. For instances, <teacher> and <student> are both <human> and inherit the properties of <human>. They also participant in the event of <teach>, but the semantic difference is that they act as different semantic roles and therefore inherit different property of their semantic relations.

Taxonomically unrelated but conceptually related concepts can also be computably associated through their Extended-HowNets. The following graphical relations, quoted from HowNet (Dong & Dong, http://www.keenage.com/), shows the concepts which may not associated with taxonomical relations but associated each other by other semantic relations.

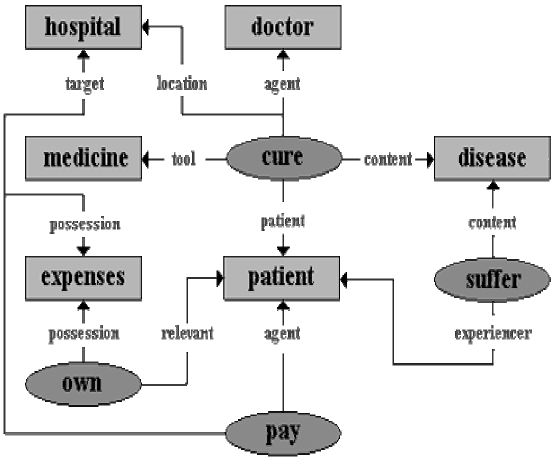

Figure 1. Concepts associated by semantic relations

3.4 Difficulties and Solutions

The above Extended-HowNet mechanism uses hypernym concepts as the type classifications of defined concepts and differentiates concepts of same hypernym class by their major features. However some types of concepts are hardly to have their natural hypernym concepts. For instances, parts of an object usually have no natural hypernym concepts. They linked to other concepts by part-whole relations in ontology. We extend Extended-HowNet a new notation % to denote part-of and use the feature relations of location or telic to differentiate different parts. For instance,

<foot|腳> Def:=
{ %<animal|獸>:
telic={<walk|行走> :
agent = {~}}

The definitions for relational-type concepts, such as kinship relations and directional relations, are different from definitions for entities. For instances, <grandfather> and <north-west> have to be expressed by composing primitive relations instead of feature attributes.

<grandfather>Def:= {father(father(human:x))}
<north-west> Def:={north(west(location:x))}

The detail discussions can be seen at (Chen et al., 2004).

Functional-type concepts, such as adverbs, prepositions, conjunctions, contain less content senses, but rich relation senses. Definitions of function words cannot just refer to their part-of-speeches, since part-of-speeches do not provide semantic information and cannot fit into the unification processes for semantic composition. Function words are defined by their relational senses and content senses (Chen et al., 2005). For instances, the adverb <in public|當眾> is defined as

Def:=manner={overt|公開}

and the preposition <by|被> is defined as

Def:=agent={}

It is necessary to make distinction between individual instances and generic concepts. For instance, proper names refer to individual not generic concepts. We use the notation of (<concept>) instead of {<concept>} to denote individual instance of <concept>. For instance,

<Tomas Edison|愛迪生 > Def:=
(< scientist| 科學家>:
name= 'Tomas Edison|愛迪生', …)
<Japan|日本> Def:=
(<country| 國家>:
name= 'Japan| 日本',…)

Some concepts are hard to be defined by common concepts. For instances, concepts belonging to certain special domains are hard to be defined in detail, such as <square root>, <prime number>, <gravity>, <palm tree>, which require the supports of domain ontology. For the moment, we propose that we do not provide detail definitions for domain specific concepts at Extended-HowNet and try to link them to domain ontology in the future.

3.5 Combine HowNet and WordNet to Form Extended-HowNet

In real implementation, we intend to integrate currently existing resources. We use the WordNet synsets as the vocabulary for the basic concepts of Extended-HowNet and translate the HowNet sense definitions into Extended-HowNet definitions as the first version. We buildup a mapping table between HowNet sememes and WordNet synsets first and then transfer sememes in the HowNet definitions into synset ids. The current version has more than 100,000 entries. The top-level ontology adopted in this system is a combined taxonomy by linking existing top-level ontologies of SUMO, HowNet, WordNet.

Following samples are the first version of Extended-HowNet translated from the HowNet.

<exhibit as evidence|證物> Def:=
{<physical|物質>:
domain={<police|警>},
predication= {<prove|證明>:
instrument ={~}}}
(*Original HowNet definition* .)
<exhibit as evidence|證物>Def2:=

{[00010572N]:
domain={[06093563N]},
predication= {[00686544V+01816870V]:
instrument= {~}}}
(Definition is in terms of WordNet Synset id-number which reads as
{<substance>:
domain={<police>},
predication= {<testify+corroborate>:
instrument={~}}})

<motto|座右銘> Def:=
{<expression|詞語>:
predication= {<obey|遵循>:
content={~}}}
<motto|座右銘>Def2:=
{[05349662N+05059598N+04764807N]:
predication= {[01733968V]:
content={~}}}
(which reads as {<saying+term+construction>:
predication= {<obey>:
content={~}}})

In the future, we will redefine each complex concept by its immediate hypernym concept and major differentiation descriptions, to replace the conventional HowNet definition of using sememes only.

4. Advantages of Extended-HowNet

The following advantages of Extended-HowNet fulfill the purpose of bridging gaps between string processing and conceptual processing.

a) Feature representation is more precise and incremental. e.g.

<great dane|大丹狗>Def:=
{<dog|狗>:
place={<German|德國>},
telic={<hunt|狩獵>:
instrument={~},
size={<big|大型>},
evaluation={<gentle|溫和> },
color={<black white|黑白>}}}

A pure taxonomy approach, such as WordNet, does not provide detail description of a concept.

b) Features are criterion for classifying new types.For example, <great dane> is also classified as：

1) Hunting instruments according to its telicity feature: Other examples of the class are

<firearm> Def:=
{<gun|槍>:
telic={<hunt|狩獵>:
instrument={~}}}
<trap> Def:=
{<facility|設施>:
telic={<hunt|狩獵>:
instrument={~}}}.

2) Animals with black/white colors: Other examples of the class are

<panda> Def:=
{<beast|走獸>:
place={<China|中國>},
predication= {<eat|吃>:
patient={<bamboo|竹子>},
agent={~}},
color={<black white|黑白>}}
<zebra> Def:=
{<horse|馬>:
color={<black white|黑白>}}

c) Achieves near canonical semantic representation. Two sentences with different surface forms or in different languages may have similar Extended-HowNet representations. e.g.

a) 我 買了 一本 科幻小說。
b) I bought a science fiction.

Both sentences have the same representation of {<buy|買>: agent={<I|我>}, goal={<science fiction|科幻小說>: quantity={<one|一>}, time-before = {speaking time}}. Note that the above high level representation can be extended to lower level and WordNet synset representations.

d) Multi-level meaning decomposition. e.g.

<tailor store|裁縫店> Def:=
{<store|店>:
telic={<sew|裁縫>:
location={~}}}

which extend to

{<InstitutePlace|場所>:
{<produce|製造>:
PatientProduct=
{<clothing|衣物>},
location={~}}}

In contrast, in HowNet concepts are defined by primitive concepts; in the above example, the basic concept <InstitutePlace|場所> lost the information of “commerce” of <store|店>.

e) Extended-HowNet is universal and language independent, since it uses WordNet synsets as description language.

f) Extended-HowNet did not create a completely new ontology, but accommodates other ontologies, such as WordNet, HowNet, and FrameNet.

5. Summarization and Conclusion

To bridge gaps between natural language representations and conceptual representations, we proposed a universal concept representational mechanism, called Extended-HowNet, which uses the word sense definition mechanism of HowNet and the WordNet synsets as vocabulary to describe concepts. Fine-grain differences among near-synonyms can be differentiated by adding new features. The encoded features, including qualia structures, and ontological links provide the bases for manipulating intelligent semantic processing, such as type coercion, semantic composition, rule generalization, and logical inference. Near-canonical representations for concepts are achieved by following definition principles to define concepts. The semantic distance of two Extended-HowNet definitions can be computed by concept distance and representation distance. Better sense similarity measures can be derived through Extended-HowNet definitions. For instance, a weighted feature distance is more precise than the conventional hierarchical structure distances (Resnik, 1995).

In addition to the conventional taxonomic relation links, such as synonymy, hyponymy, antonymy, meronymy, Extended-HowNet also links concepts by their shared features. Multiple links mean multiple-inheritances. The shared properties of different concepts are associated with common ancestors in Extended-HowNet without redundant.

Extended-HowNet is language independent. It can bridge the gaps of translation equivalence between two languages. In EuroWordNet, each word sense of different language was intended to link to the synonymy WordNet synset. However many word sense cannot find a synonym synset, they have to create some interlingua-indices (ILI) to link translation equivalences among different languages (Vossen, 2000). The ILIs used in EuroWordNet do not carry semantic meanings and they are hard to form a completed ontological system. If WordNet synsets and the senses of all ILI are defined in Extended-HowNet, the result Extended-HowNet will become a shared ontology for all languages.

Extended-HowNet is universal and intends to achieve that any concept can be defined by Extended-HowNet. However there are still some problems needs further study. For instances, concepts belonging to certain special domains are complicated and hard to define, such as 'square root', 'prime number', 'gravity', 'palm tree',…,etc. The representation of them needs supports from related domain knowledge-bases instead of their fine-grain definitions.

The idea of using Extended-HowNet as intermediate language for machine translation needs further study, since differences between two words are not simply on semantics but also on their syntactic as well as discourse and social functions.

The semantic composition and decomposition mechanism at Extended-HowNet can be extended to encode deep semantics of phrases and sentences. The detail representations for references, quantifications, temporal relations will be our future works. The fine-grain features, in particular semantic/syntactic correlation features, will also be future refinement of Extended-HowNet, which is motivated by the syntactic differences within synonyms (Levin, 1993; Huang etc., 2000; Chen et al., 2005).

Acknowledgement: This research was supported in part by National Science Council under a Center Excellence Grant NSC 93-2752-E-001-001-PAE and Grant NSC93-2213-E-001-019.

References

Chen, Keh-Jiann, Shu-Ling Huang, Yueh-Yin Shih, Yi-Jun Chen，2004，*Multi-level Definitions and Complex Relations in Extended-HowNet*, Workshop on Chinese Lexical Semantics, Beijing University. (in Chinese)

Chen, Yi-Jun, Shu-Ling Huang, Yueh-Yin Shih, Keh-Jiann Chen，2005，*Semantic Representation and Definitions for Function Words in Extended-HowNet*, Workshop on Chinese Lexical Semantics, Xiamen University. (in Chinese)

Dong, Zhendong & Dong Qiang, *HowNet*, http://www.keenage.com/

Dowty, David R. 1991, " Semantic Proto-roles and Argument Selection", *Language*, Vol. 67(3), pp. 547-619.

Fellbaum, Christiane, 1998, *WORDNET-An Electronic Lexical Database*, the MIT Press.

Fillmore, Charles, *FrameNet*, http://www.icsi.berkeley.edu/~framenet/

Gruber, T.R. 1993, *Toward principles for the design of ontologies used for knowledge sharing*, Padua workshop on Formal Ontology.

Huang, Chu-ren, K. Ahrens, Li-li Chang, Keh-Jiann Chen, M. C. Liu, Mei-Chih Tsai, 2000, " The Module-Attribute Representation of Verbal semantics: From Semantics to Argument Structure" *International Journal of*

Computational Linguistics and Chinese Language Processing, Vol.5, #1, pp.19-46.

Levin, Beth. 1993, *English Verb Classes and Alternations: a Preliminary Investigation*, Chicago Press.

Niles, Ian and Adam Pease. 2001."Towards a Standard Upper Ontology," *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems* (FOIS-2001), Ogunquit, Maine, October 17-19.Pustejovsky, James 1995, *The Generative Lexicon*, the MIT press.

Resnik, Philip. 1995. "Using Information Content to Evaluate Semantic Similarity in a Taxonomy". *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, 448-453.

Schank, R., 1975, *Conceptual Information Processing*, Amsterdam, North-Holland.

SIMPLE-CLIPS Ontology, http://www.ilc.cnr.it/clips/Ontology.htm

Sowa, John, 2000, *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, Brooks Cole Publishing Co..

Vossen, Piek (ed.), 2000, *EuroWordNet General Document*, http://www.hum.uva.nl/~ewn.

Wiezbicka, A. 1972, *Semantic Primitives*, Athenaum, Frankfurt.