International Journal of Knowledge and Language Processing
Volume 8, Number 2, **2017**

www.ijklp.org
KLP International ©2017 ISSN 2191-2734
pp.16—26

Review of Education-oriented Knowledge Base

Yi Huang

Institute of Scientific and Technical Information of China
No. 15 Fuxing Road, Haidian District,
100038, Beijing, China
huangyi@istic.ac.cn

Received June 2017; revised June 2017

ABSTRACT. *With the rapid development of international exchange and cooperation, English has become more and more prominent in China's middle school teaching. The current sharing of English education resources is mainly based on web pages, databases, there is a problem of duplication of resources, poor correlation, lack of a unified terminology standard, and cannot effectively communicate and share. How to effectively solve problems such as knowledge sharing, knowledge representation, and knowledge reasoning, and make resources serve English teaching, which determines the quality of teaching resources and the efficiency of student learning. The linguistic knowledge base is applied to English teaching in middle schools and has significant scientific value and application value for realizing information-based teaching. This paper reviews the current domestic research literature and summarizes it from multiple dimension, and propose a framework to construct an education-oriented knowledge base.*
Keywords: Semantic Orientation, Semantic Identification, Rule-based

1. **Introduction.** The primary way for middle school students in China to acquire English knowledge is classroom teaching. In traditional classrooms, on the one hand, teachers spend most of their time in explaining the sound and meaning of words, and urge students to strengthen vocabulary memory. However, the vocabulary meaning and usage learned by students are often too monotonous and rigid to use. On the other hand, many teachers tend to focus on vocabulary teaching or grammar teaching and do not combine the two well. Vocabulary and grammar are the two pillars of English learning. They complement each

other, and neither aspect can be neglected. The current state of English teaching in middle schools leads to the blindness and disorder of English learning, which results in a poor teaching effect. The reasons for this phenomenon are summarized as follows: 1) The teaching methods are single, and the students lack interest; 2) The teachers teach vocabulary or grammar in isolation; 3) The focus is unclear, and there is no distinction between primary and secondary knowledge; 4) The mechanized memorization method that emphasizes memorization; 5) It lacks recurrence and systematicity.

Given the students' problems in vocabulary and grammar acquisition, since the beginning of the new middle school English education syllabus and new textbooks in 1993, domestic scholars have conducted extensive and in-depth discussions and research on the topic of English teaching. Through the introduction of different advanced linguistic theories and teaching concepts around the globe, using advanced computer and information network technologies, new understandings and attempts have been made in English teaching methods. Although some research has achieved specific results, these resources are fundamentally fragmented and do not form a unified and interconnected system, which is not conducive to the exchange and sharing of resources. How to make practical use of existing resources and integrate them into a unified, functional and complete language knowledge base becomes an urgent problem to be solved. This paper reviews the current domestic research literature and summarizes it from multiple dimension, and propose a framework to construct an education-oriented knowledge base.

The rest of this paper is organized as follows. Section 2 reviews the corpus research. Section 3 reviews the ontology research. Section 4 reviews the education-orientated knowledge base. Section 5 describes the proposed framework. Finally, a conclusion is given in Section 6.

2. **Corpus research.**

2.1 **Definition of corpus.** The term corpus is derived from Latin, which means "body". With the passage of time and the evolution of language, it had been injected with many meanings such as "subject, mass, group, and collection". However, as a corpus, different scholars give different definitions:

Crystal [1] believes that "Corpus is a collection of linguistic materials. Written language, spoken language, or its recorded scripts are all sources of data and are often used as a starting point for language descriptions or as a way to verify hypotheses about language."

According to Leech [2], "Corpus is not a collection of simple text data. The collection of corpus should have a certain purpose. The collected corpus should be able to represent a specific language or text of integrated language information with real-time and practical use, for research purposes."

According to Sinclair [3], the corpus is a collection of linguistic materials selected and ranked according to linguistic criteria, and its significance lies in its use as a reference sample of language.

McEnery & Wilson [4] decomposes the definition of “corpus” into four specific levels, namely: 1) sampling and representativeness; 2) finite size; 3) machine-readable format; 4)

and a standard reference.

However, more and more scholars have agreed on the definition: machine-readable and authentic texts, that is, the use of computers by certain linguistic principles. Real corpus collected on a large scale and stored in a computer according to the purpose of a specific language study.

2.2 **Corpus linguistics.** Corpus linguistics is an interdisciplinary subject that emerged at the end of the 20th century to study the collection, storage, processing, and statistical analysis of natural language texts. The purpose is to engage in linguistic studies and guide natural language by providing objective and real linguistic evidence provided by the development of large-scale corpus and information processing system. For the definition of Corpus Linguistics, two Western linguists describe it as follows. Leech [5] believes that "the study of language in real-life examples of language use is called corpus linguistics". Crystal [6] thinks that "the use of corpus as a starting point for language descriptions or corpus as a method for validating hypotheses is called corpus linguistics".

2.3 **Corpus-based English Teaching Motivation.**

2.3.1 **Schema and schema theory**. The concept of a schema first appeared in the works of the German philosopher Kant. He believed that "in its own right, a schema is a product of imagination or a learner's previously acquired knowledge (i.e., background knowledge) Structure. In 1932, the British psychologist Bartlett first proposed the schema theory in his "Remembering" book. He thought that "schema is an active organization of previous experience (developing active pattern) is a process in which information stored in a learner's brain acts on new information and a process of absorbing new information." Gui [7] thinks that "Schema theory emphasizes that people have the decisive role of knowledge and knowledge structure in current cognitive activities, and that they are a way for people to use the existing structure to remember new information". Although the definitions of these schemas are not the same, the views expressed are the same, that is, schemas are all general knowledge of the world that is input and stored in mind.

In the process of learning English, the existing vocabulary and grammar knowledge in mind plays an essential role in the absorption and application of new knowledge. The corpus context can now provide corresponding language teaching materials for constructing and reconstructing schemata. In this way, teachers can use corpus or language knowledge base to set appropriate teaching tasks and goals, guide students to identify, select, classify, analyze, and synthesize new knowledge to continuously improve and improve their ability to acquire and use new knowledge.

2.3.2 **Lexical approach theory**. It was put forward by Lewis [8] in 1993. He believes that "Language is composed of grammatical vocabulary rather than lexicalized grammar" and proposes to use a large number of real chunks as a language for learning. The basic unit. Wang [9] pointed out that "word block is a combination of multiple words that are

pre-fabricated and frequently used. This kind of vocabulary combination has its structure and relatively stable meaning. It is stored as a whole and stored in real time. When the communication is extracted as a whole, there is no need to use grammar rules for processing analysis."

The corpus-based English teaching is consistent with Lewis's idea of inputting a large number of real chunks as the basic unit of language learning. By using tools, corpus can be automatically extracted and formed. With the help of tagged part-of-speech linguistic materials and reasonable retrieval, many syntactic structures can be analyzed and mined. Corpus presents a large number of examples to students in the form of data or contextual co-occurrence, which is conducive to attracting attention, strengthening memory, and helping them to use context to acquire semantics and summarize laws.

3. **Ontology research.**

3.1 **Definition of ontology**. Ontology was initially a philosophical concept, defined in philosophy as "a systematic explanation or explanation of objective existence, concerned with the abstract nature of objective facts" [10]. In the field of artificial intelligence, Neches et al. [11] first defined ontology, which was “giving out the basic terms and relationships that form the vocabulary of related fields, and the rules that govern these lexical extensions that are formed using these terms and relationships”. In 1993, Gruber [12] proposed that "ontology is a clear specification of a conceptual model." Studer [13] has conducted an in-depth study of the above definition, arguing that “ontologies are explicit formal specification descriptions of shared conceptual models”, including the meaning of conceptualization, explicit, formal and share.

From the above definition, it can be seen that ontology is a conceptual system used to describe knowledge in related fields, determine the basic vocabulary in the field, provide a shared understanding of the knowledge in the field, and give a precise definition of the relation of vocabularies from different levels.

3.2. **Ontology Model.** The logical structure of the ontology can be seen as a five-tuple [14]:

$$O := \{C, R, H^C, \mathrm{rel}, A^O\}$$

A finite set C. C is a collection of concepts.

A finite set R. R is a collection of relations.

A concept hierarchy H^C . H^C is a directed relation $H^C \in C \times C$, which is called concept hierarchy or taxonomy.

A function rel: $R \rightarrow C \times C$. It relates concepts non-taxonomically.

A set of ontological axioms A^O . A^O is expressed in an appropriate logical language.

As the core structure of the ontology, this model is generally accepted and widely used for the description of ontology relationships.

3.3. **Knowledge Base and Ontology.** The logical structure of the knowledge base can be seen as a quad-tuple [15]:

$$KB := \{O, I, inst, instr\}$$

O is an ontology structure.

I is a collection of instances.

inst is a function. $C \rightarrow 2^I$ is called concept instantiation.

instr is a function. $R \rightarrow 2^{I \times I}$ is called relation instantiation

Regarding knowledge representation, ontology and knowledge base are all definitions, representations, and organizations of knowledge contained in a specific domain. Ontology as a concept-level description focuses on the description of terminology and terminology at the conceptual level, while the knowledge base focuses on the representation, organization, and storage of domain knowledge. Ontologies provide a standard set of description languages and rules for describing concepts and their relationships. The use of ontology description language to establish the concept and the relationship between the concepts, and associated with the inference rules to establish the knowledge base, can achieve the construction method and data sharing between different domains, different models.

3.4 **Language Knowledge Base Construction**. The construction of language knowledge base involves the arranging, discovering, formalization and standardization of language knowledge. The content of the language knowledge base and the manifestations of knowledge are various.

The construction of representative language knowledge bases at home and abroad is shown in Table 1:

TABLE 1. LANGUAGE KNOWLEDGE BASE CONSTRUCTION

PROJECT NAME	TIME	DEVELOPER	SIZE	FUNDAMENTAL SEMANTIC THEORY	CONSTRUCTION METHOD
WordNet	1985-	Princeton University	111,223 concepts; nouns, verbs, adjectives, adverbs; English	Relationship-based semantic description theory; a collection of synonyms, description of semantic relations	Manual build
FrameNet	1997-	University of California	458 frames, more than 4,000 words; English	Frame Semantics; Frame Elements, Valency, Semantic Relations	Manual build

MindNet	1993-	Microsoft Corporation	15.9 million words (nouns, verbs, adjectives); English	Semantic relationship description	Automatic build
VerbNet	2006-	University of Colorado	357 syntactic frames, more than 5,200 words	Beth Levin verb classification; description of semantic roles and semantic relations	Manual build
HowNet	1988-	Dong Zhendong et al.	116,533 records in Chinese-English bilingual	Semantic analysis, semantic roles, description of semantic relations	Manual build
Chinese Concept Dictionary (CCD)	2000-	Peking University	Nearly 70,000 concepts, bilingual in Chinese and English	WordNet semantic knowledge representation framework	Manual build

WordNet[16] was developed by the Cognitive Science Laboratory at Princeton University in 1985. WordNet describes objects that contain compound words, phrasal verbs, common collocation words, idioms, and words, where words are the most basic unit. WordNet classifies words into four categories: nouns, verbs, adjectives, and adverbs. WordNet is a semantic network of English language. It divides English words into synonym synonymously according to synonyms, and gives a concise definition of each synonym set, and the semantic relationship between it and other synonym sets. WordNet does not break down words into smaller meaningful units, nor does it contain larger organizational units than this, nor does it contain syntactic content of words.

FrameNet [17] is a computer lexicon compilation project supported by real corpus that was established in 1997 by the University of California, Berkeley under the leadership of Johnson and Fillmore. FrameNet consists of three parts: 1) The dictionary, which contains the traditional dictionary definition of the term, the rules of the record form syntax, the link to the annotation sample library, and links to the framework database and other machine-readable resources (e.g., WordNet, COMLEX); 2) The framework database, including the description of the basic concept structure of each framework, the framework elements and their descriptions, etc.; 3) The annotation sample library, which mainly contains marked example sentences, to illustrate the semantic attributes and syntax attributes of dictionary terms.

MindNet [18] was built by the Natural Language Processing team of Microsoft Research. MindNet is a lexical semantic knowledge base based on a wide-area syntax analyzer built on the existing dictionaries (e.g., LDOCE, AHD3) and Encarta. There are 24 different types

of semantic relationships in MindNet. The construction of MindNet is completed automatically, which embodies the idea of automatically acquiring, organizing, accessing, and extracting semantic information from natural language, and provides a credible approach and prospect for extracting semantic information on the model supporting common sense reasoning.

VerbNet [19] is a verb dictionary containing syntactic information and semantic information that was constructed in 2000 by Martha Palme and Kipper Karin of the University of Colorado. The verb classification standard of Beth Levin is the theoretical basis of VerbNet. The underlying assumption is that the syntactic frame is the most basic and direct reflection of the verb underlying semantics. VerbNet is classified according to shared word sense and syntactic behavior features. Each VerbNet class consists of three parts: MEMBERS, THEMROLES, and FRAMES. The contents of VerbNet are mainly expressed in two aspects: 1) static description and semantic feature description of verbs; 2) dynamic description, describing its syntactic frame and related semantic predicates and selection restrictions through the TAG tree.

HowNet [20] is a concept described by the Chinese and English words under the leadership of Dong Zhendong, executive director of the Chinese Information Society of China, to reveal concepts and concepts and between the attributes and attributes of concepts. The relationship is the essential content of the knowledge system. The definition of the relationship is achieved through HowNet Knowledge Dictionary Markup Language (KDML). HowNet focuses on the eight relationships between concepts, i.e., upper and lower levels, synonymy, antisense, righteousness, component-integrity, attribute-host, event-role, and material-finished product. In addition to the above relationships, HowNet also portrays a large number of features and dynamic roles.

Chinese Concept Dictionary (CCD) [21] is a Chinese-language semantic dictionary compatible with WordNet developed by the Institute of Computational Linguistics of Peking University. It inherits the main ideas of WordNet, uses synonym sets to describe concepts, and uses concepts to describe semantics, mainly including nouns, verbs, adjectives, and adverbs; the central relationships are synonymous, antisense, subordinate, and overall. Partial relations, etc.; At the same time, for the characteristics of Chinese, the relationship between concepts and concepts has been adjusted and improved.

4. **Education-orientated Knowledge Base Research**. The teaching domain knowledge base is a set of comprehensive, descriptive, procedural, and strategic knowledge in the subject area. The various types of knowledge in the collection are organized and represented by a specific representation method, and the relationship between the knowledge is established. It is a conceptualized representation of educational knowledge, namely the standardization of conceptual terminology in the field of education, as well as a clear description of its hierarchical relationships, and the expression of commonly recognized and shareable educational knowledge, which is the basis for achieving educational information sharing and exchange at the semantic level.

Knight et al. [22] linked the conceptual model with learning design and learning object

content by constructing an ontology-based framework (LOCO), and proved by examples that LOCO significantly improved knowledge retrieval at the automatic or semi-automatic processing and service level, as well as reuse efficiency. Garcia et al. [23] used ontologies to build the Weiner Lecture Archives. By establishing a hierarchical structure and semantic relationship network, the project effectively associates a knowledge point in a course with other knowledge points and notes, enabling students to quickly acquire notes and other knowledge associated with the knowledge point while watching a video.

Cui et al. [24] designed OntoEdu, a new type of teaching support platform. This platform takes the teaching ontology as the core and has functions such as expandability, freedom of choice and customization. It can continue to deepen its knowledge base as users use it, thereby further expanding the function of the teaching support platform itself, but the platform is still in the research stage. Liu [25] proposed a set of models to guide the construction of curriculum knowledge ontology in the E-Learning system. The model establishes the conceptual ontology model based on the curriculum knowledge points, extracts the core concepts of the curriculum knowledge points according to the teaching steps and teaching rules, and constructs the relationship between the concepts. The standard ontology language is used to define and describe the concepts. Course knowledge ontology model.

Although the studies mentioned above have achieved specific results, they are all focused on natural language processing and computer-related research. However, there is no universal significance for language knowledge bases, especially for the construction of English teaching knowledge bases. This paper mainly combines the correlative teaching theory, syntactic analysis and ontology theory of corpus linguistics, researchers and explores the methods and essential applications for the automatic construction of English teaching knowledge base in middle schools.

5. **Education-oriented Knowledge Base Construction**. Given the studies above, based on the theoretical idea and platform of automatic constructing ontology[26combined with and applying corpus linguistics teaching related theories and syntactic analysis theories, the paper proposes a framework to construct an Education-oriented Knowledge Base, aiming at the characteristics of middle school English teaching. We introduce ontology ideas into the process of building English teaching resources in middle schools, exploring a set of ideas and methods for automatically building a linguistic knowledge base organizes the knowledge reasonably and adequately so that students can build a good knowledge structure through the linguistic knowledge base and improve the efficiency of English learning. The overall framework is shown in figure 1.

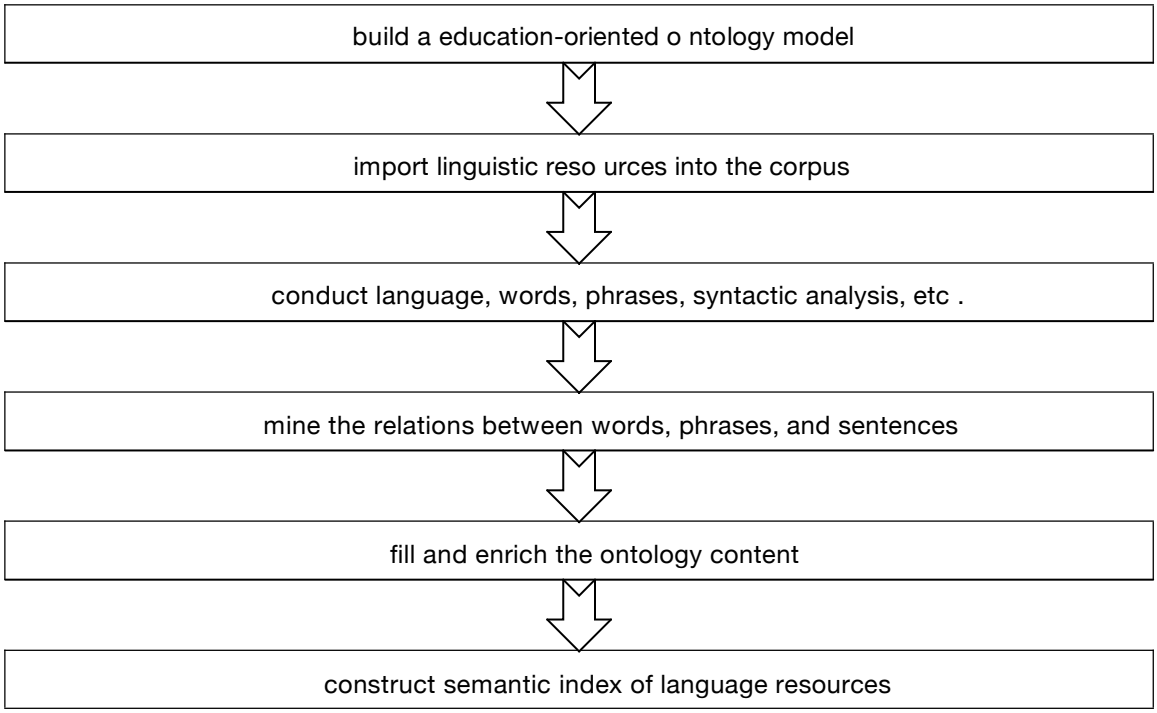

FIGURE 1. THE OVERALL FRAMEWORK OF EDUCATION-ORIENTED KNOWLEDGE BASE CONSTRUCTION

To be more specific, this article selects words, phrases, syntax components and grammar as the core concepts of language knowledge base for ontology modeling. By analyzing classification of words and phrases, characterizing the attributes and relations among them, a conceptual model of language resource is proposed. Introducing syntactic analysis to parse sentences, we can effectively associating sentences with words and phrases; and then with grammar rules, using corpus as the original material, a three-dimensional education-oriented knowledge base is constructed. By introducing ontology into the knowledge base construction process, and using a set of concepts and terminology provided by ontology to describe the language knowledge base, the concepts, grammar and relationships between words, phrases, and sentences are defined more precisely. The semantic retrieval function is easy to implement in the ontology-based knowledge base system, which helps the students to understand the structure and meaning of the profound language and improve the students' efficiency in learning English.

6. **Conclusion**. This paper introduces the current problems in English education resources and points out that the application of ontology to construct a language knowledge base can solve the problems of educational resource sharing and semantic retrieval. Then, it introduces the teaching arguments and ontology-related concepts based on corpus linguistics, listed relevant research results at home and abroad, and finally pointed out the necessity of building a language knowledge base. This paper reviews the current domestic research literature and summarizes it from multiple dimension, and propose a framework to

construct education-oriented knowledge base, by introducing syntactic analysis, concepts such as words, phrases, and sentences are effectively related.

REFERENCES

[1] Crystal D. A Dictionary of Linguistics and Phonetics (3rd Edition). Blackwell, 1991.

[2] Leech G, Garside R. "Running a Grammar Factory: The Production of Syntactically Analysed Corpora or Treebanks." Johansson and Stenström (1991): 15-32.

[3] Sinclair J. Lexical Grammar. BARBAI ir DIENOS, 2000 (24): 180-203.

[4] McEnery T, Wilson A. Corpus linguistics: An Introduction. Edinburgh University Press, 2001.

[5] Leech G. Introducing English Grammar. London: Penguin English, 1992.

[6] Crystal D. Stylish Profiling, in Aijmer & Alterberg, 1991, pp 221-238

[7] Gui SC. Memory and English Learning. Foreign Languages, 2003(3).

[8] Lewis M, Gough C. Implementing the lexical approach: Putting theory into practice. Vol. 3. No. 1. Hove: Language teaching publications, 1997.

[9] Wang LF. Learning from foreign countries, based on the local, innovative teaching. Chinese foreign language, 4.6 (2007).

[10] Deng HH. et al. Review of Ontology Research. Journal of Peking University (Natural Science), 2002, 38(5): 730-738.

[11] Neches R et a1. Enabling Technology for Knowledge Sharing. AI Magazine, 1991, 12(3): 36-56.

[12] Gruber TR. A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition, 1993, 5: 199-220.

[13] Fensel D et al. OIL: An Ontology Infrastructure for the Semantic Web. IEEE Intelligent Systems,2001(2).

[14] Maedche A, Staab S. Ontology learning for the Semantic Web. Kluwer Academic Pub, 2002.

[15] Maedche A, Staab S. Ontology learning for the Semantic Web. Kluwer Academic Pub, 2002.

[16] Miller GA et al. WordNet: An On-line Lexical Database. International Journal of Lexicography 3, 4 (Winter 1990), 235--312.

[17] Baker CF et al. The Berkeley FrameNet Project. Proceedings of the 17th international conference on Computational linguistics-Volume 1. Association for Computational Linguistics, 1998.

[18] Richardson SD et al. MindNet: Acquiring and Structuring Semantic Information From Text. Proceedings of the 17th international conference on Computational linguistics-Volume 2. Association for Computational Linguistics, 1998.

[19] Schuler KK. VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon. University of Pennsylvania, 2005.

[20] Dong ZD, Dong Q. HowNet and Chinese Studies. Contemporary Linguistics 3.1 (2001): 33-44.

[21] Yu JS et al. Chinese Concept Dictionary Specification. Journal of Chinese Language and Computing, 2003(06).

[22] Knight C et al. An Ontology-based Framework for Bridging Learning Design and Learning Content. JOURNAL OF EDUCATIONAL TECHNOLOGYAND SOCIETY 9.1 (2006): 23.

[23] Garcia DD et al. The Weiner lecture archives: an ontology-driven interface for viewing synchronized lectures and notes. Proceedings of the 41st ACM technical symposium on Computer science education.

ACM, 2010.

[24] Cui GZ et al. OntoEdu: Ontology-based teaching support platform. Annual Meeting of National Universities Educational Technology Cooperation Committee . (2003).

[25] Liu GR et al. Construction and implementation of curriculum knowledge ontology in E-Learning system. Journal of the China Society for Information and Technology 4 (2009): 006.

[26] Liu Y et al. "Field Ontology Automatic Construction Research." Journal of Beijing University of Posts and Telecommunications 29.Z2 (2006): 65-69.