International Journal of Multiphysics
Volume 18, No. 3, 2024
ISSN: 1750-9548

The Construction of the Semantic Network of Characteristic Words in Overseas Chinese Dialects

Zihui Xiao*, Zhifeng Peng
Center for Chinese Dialect Research, School of Chinese Language and Literature, Jinan University, Guangzhou, Guangdong, China
*Corresponding Author.

Abstract

The characteristic words of overseas Chinese dialects are the most distinctive part of the vocabulary in overseas Chinese dialects. The semantic network of characteristic words in overseas Chinese dialects treats characteristic words as semantic symbols, drawing knowledge graphs through semantic relations to form a formal description of semantic knowledge, presenting a comprehensive picture of Chinese dialects and culture in overseas Chinese communities. The article proposes the ideas and methods for dividing the semantic classification system of characteristic words in overseas Chinese dialects and designs synonym sets and data attribute description rules based on the characteristics of characteristic words in overseas Chinese dialects. Taking nouns as an example, it explains common semantic relationships such as synonymy, hyponymy, holonymy, antonymy, and meronymy. The construction of the semantic network provides rich lexical semantics and relational information for overseas Chinese dialect research and various fields of natural language processing.

Keywords: Overseas Chinese dialects, characteristic words, lexical semantic network, knowledge graph.

1. Introduction

Overseas Chinese dialects is an academic concept proposed by scholars from mainland China, referring to Chinese dialects distributed outside China's borders. Chinese communities overseas, living in different natural geographic and socio-cultural environments, have dialects that both preserve traditions and undergo variations and innovations. One prominent characteristic is the presence of a set of words with overseas features in the lexical system of overseas Chinese dialects, which are either absent or rare in mainland Chinese dialects. This paper refers to these as characteristic words of overseas Chinese dialects. These characteristic words fall into three categories: First, Referring to objects or concepts that exist only overseas. For example, in the Hokkien, Cantonese, and Hakka dialects of Malaysia, terms such as "Baba" (a boy born from a Chinese-Malay mixed marriage), "Nyonya" (a girl born from a Chinese-Malay mixed marriage), and "Sultan" (state ruler) are used. Second, Using the same characters as in mainland China, but with different meanings (or senses) or grammatical functions. For instance, in the Kokang dialect of northern Myanmar, "dǎ qīng jiā" means forming a bond between two families because their children recognize each other as godfather/godmother, while in the Yunnan dialect in China, it means forming a marital relationship between two families due to their children's marriage. Though the characters are the same, the meanings differ; Third, different expressions for the same object or concept in overseas Chinese communities. For example, in the Hokkien, Cantonese, and Hakka dialects of Malaysia, terms like "shù rén" (rubber tree), "zhǐ dài" (paper bag/plastic bag), and "gē chē" (overtaking) are used.

Characteristic words of overseas Chinese dialects cover various aspects of the lives of Chinese people and are frequently used in daily life. Although some characteristic words have been collected in research reports and monographs on overseas Chinese dialects over the past two decades [1-4], they are scattered and few in number.

International Journal of Multiphysics
Volume 18, No. 3, 2024
ISSN: 1750-9548

Different Chinese communities exhibit different characteristics in their dialects, with varying characteristic words and possibly different semantic rationales. The distribution and semantic relationships of these characteristic words are also diverse. Therefore, how to establish and describe the characteristic words, and even the lexical system, of overseas Chinese dialects is a question that needs consideration in overseas Chinese dialect research. This paper discusses related issues and constructs a database and knowledge semantic network of characteristic words in overseas Chinese dialects.

2. The Construction Goals and Special Value of the Semantic Network of Characteristic Words in Overseas Chinese Dialects

In 2001, Berners-Lee T proposed the concept of the Semantic Web, defining it as a conceptual standard for describing the objective world through a unified set of metadata to provide detailed semantic annotations for internet content [5]. A lexical semantic network is essentially a semantic network knowledge base aimed at expressing the lexical semantic knowledge and semantic relationships of one or more languages. Since the mid-1980s, many research institutions domestically and internationally have begun constructing various large-scale semantic knowledge networks (bases) for machine use, with scales ranging from tens of thousands to hundreds of thousands, and even millions. The major achievements in semantic knowledge networks (bases) based on English include Princeton University's WordNet, the University of California, Berkeley's FrameNet, Microsoft's MindNet, and the LLD developed by the University of Cambridge and the University of Edinburgh. In terms of Chinese, there are HowNet (or Zhiwang) by Dong Zhendong of the Institute of Computational Linguistics at the Chinese Academy of Sciences, the Chinese Concept Dictionary (CCD) by Peking University, and the Sino-Trans-SemDict by ChinaSoft, etc. In terms of knowledge description content, both domestic and international semantic knowledge base research focuses on the semantic classification and semantic relationships of different types of semantic units, showing significant relativity in semantic knowledge categories. The development and application of lexical semantic knowledge networks (bases) have become quite mature, forming many complete theoretical frameworks and annotation systems. These networks provide various application services for natural language processing tasks such as word sense disambiguation [6-7], information retrieval [8-9], machine translation [10-11], and question-answering systems [12], and also provide an extremely important platform for semantic research on Chinese. However, in the field of Chinese dialect research, there has been no construction or application of lexical semantic knowledge network technology.

The unique knowledge and culture of Chinese communities are mainly inherited and disseminated through characteristic words in overseas Chinese dialects. The knowledge units of these characteristic words are interrelated, and there are certain logical relationships between the words, forming a relatively systematic knowledge system. Therefore, the overall construction goal of the semantic network of characteristic words in overseas Chinese dialects is to use the database of characteristic words in overseas Chinese dialects as the basic data resource and establish a structured database for the semantic network of characteristic words in overseas Chinese dialects through automatic matching and manual intervention. This aims to establish various semantic relationships such as synonym sets, antonymy, hyponymy, and meronymy, forming a rich structure of knowledge units. The resulting semantic network of characteristic words is equivalent to a dictionary of characteristic words in overseas Chinese dialects, but it differs from traditional dialect dictionaries as it not only deals with semantic or grammatical information related to the explained words but also focuses more on the natural semantic connections between the explained words and related words. The semantic network organizes words based on the psychological linguistic facts of similarity and meaning aggregation, gathering words with related meanings that would otherwise be scattered throughout a traditional dictionary. It describes the various semantic relationships such as synonymy, antonymy, hyponymy, meronymy, and holonymy between the explained words and related words, forming a semantic network centered around each explained word, almost describing all the basic semantic relationships of the explained word. For example, in the Kokang dialect of northern Myanmar, characteristic words related to wedding customs such as "dīu pìn" (dowry), "jiē qīn" (welcoming the bride), "xíng bào lǐ" (performing rituals), "zuò cuò lǐ" (performing wrong rituals), "jiǎng lǐ qián" (discussing bride price), "qiáo rì zi" (choosing a date), "huì qīn" (meeting relatives), "huàn kǒu" (exchanging words), etc., form a semantic network reflecting the local culture of "stealing the bride" (also called "bride kidnapping").

The special value of constructing a semantic network of characteristic words in overseas Chinese dialects is mainly reflected in two aspects:

Firstly, the semantic network has special value for the study of overseas Chinese dialects. Characteristic words are the most valuable part of the vocabulary system of overseas Chinese dialects, but they do not exist in isolation. They are connected by nodes with complex relationships, forming a system with a certain hierarchy. Without focusing on the overall and associative nature of characteristic words, even if some scattered data is accumulated, it will be found to be too fragmented or insufficient when it comes to cross-linguistic and cross-dialect comparisons, interpretations, measurements, or experimental analyses that combine linguistic elements and external elements. The semantic network transforms the characteristic words and knowledge in Chinese dialects into structured data, systematically classifying, integrating, and expressing vocabulary and knowledge, and more normatively describing the characteristic words and knowledge system of Chinese dialects. It explores the logical relationships between characteristic words and the objective events they reflect, providing valuable data for in-depth research in the field of overseas Chinese dialects and the innovation of linguistic theories and methods. Moreover, a comprehensive collection and recording of characteristic words in the Chinese dialects of overseas Chinese communities can help Chinese researchers examine the development and current state of Chinese vocabulary as a whole, providing unprecedented convenience for the study of Chinese lexical typology and semantic typology.

Secondly, the semantic network also has special value for Chinese natural language processing in the era of artificial intelligence. Natural Language Processing (NLP) is an important research field of artificial intelligence. Various NLP systems, such as machine translation systems and information retrieval systems, usually require an electronic dictionary with semantic knowledge to provide more comprehensive and in-depth semantic information for automatic computer analysis. Although a large amount of vocabulary and semantic knowledge has been accumulated for mainland Chinese, the lack of information on characteristic words in overseas Chinese dialects will still directly affect the effectiveness of the application of the global Chinese semantic knowledge base in NLP tasks. The semantic network of characteristic words in overseas Chinese dialects provides comprehensive and detailed lexical data for Chinese NLP, becoming an important source of semantic data for Chinese. These lexical data can be used for computational analysis in global Chinese machine learning, facilitating the construction of the Chinese system's data and addressing issues in the digital linguistic service industry.

3. The Semantic Hierarchy System of Characteristic Words in Overseas Chinese Dialects

The myriad complexities of the world are interconnected, and the meanings of words that reflect this world form a unity of order and disorder. Therefore, "it is not enough to only consider the meanings of words in semantic research; we must also consider their positions in the semantic system and the structural relationships between them" [13]. This implies that characteristic words require semantic classification and layering to establish a reasonably structured system, allowing semantic information to be reflected concisely and clearly. The semantic hierarchy system describes the cognitive understanding background of words, grouping words with the same cognitive understanding background into the same type for description, resulting in a lexical classification compilation based on semantic categories.

There have already been significant achievements in the semantic classification of vocabulary. Among these, the rationality of the WordNet semantic classification framework is widely recognized in the fields of lexical semantics and computational lexicography. WordNet classifies vocabulary by first dividing it into nouns, verbs, adjectives, and adverbs according to grammatical categories, and then organizing these into independent hierarchical structures through hypernym-hyponym relationships, forming a network of synonyms. However, due to different traditional habits and cultural backgrounds, the semantic layering of Chinese and English differs. For instance, in WordNet, time and place are not parallel categories; time is treated as a subordinate concept of abstract entities, while place is treated as a major category. This discrepancy makes it impossible to directly apply WordNet's semantic classification to Chinese vocabulary. Therefore, Chinese linguists consciously use "semantic features" or "semantic elements" to analyze word meanings, attempting to conduct a detailed semantic classification of Chinese vocabulary based on semantic features, especially categorical semantic features [14-16].

The characteristic words of Chinese dialects in overseas Chinese communities have unique manifestations in terms of parts of speech and semantics. In terms of parts of speech, most characteristic words in overseas Chinese communities are generated to reflect new life and new things in their current places of residence, such as unique place names, food names, clothing names, festival names, plant and animal names, disease names, transportation names, farm tool names, and administrative names, leading to nouns overwhelmingly dominating in quantity. Other parts of speech include:(1) Measure words: e.g., "mǎ" (a piece of [money]), "bèng" (a bucket of [water]), "kān" (a room of [space]), "jīluó" (kilograms/kilometers in Tai Lue) in the Southwest Mandarin dialects of northern Thailand.(2)Numerals: e.g., "wàn" (ten thousand) is called "shí qiān" (ten thousand) in the Guangzhou dialect of Portland, USA.(3)Interjections: e.g., "wū" (expressing surprise or admiration), "wú" (reminding) ",xuěi" (expressing dissatisfaction or impatience) in the Southwest Mandarin dialects of northern Thailand.(4)Adverbs: e.g., "kè" (equivalent to "yě" in Mandarinin the Southwest Mandarin dialects of northern Thailand.(5)Verbs: e.g., "kāi zhà" (clearing land by burning trees and using the ashes as fertilizer), "kāi kǎn", starting to cut sugarcane), " tiào bǎi " (a dance event held in temple courtyards) in the Kokang dialect of northern Myanmar.(6)Adjectives: e.g., "suǒ" (sexy), "lěng" (cold), "dānghēi" (unlucky) in the Guangzhou dialect of Chicago, USA. Semantically, characteristic words cover various aspects of Chinese life, especially the following categories with numerous terms:(1) Terms for the unique natural ecological environment, natural phenomena, and plants and animals overseas.(2) Terms for various unique items and tools related to the clothing, food, housing, and transportation of overseas Chinese.(3) Terms reflecting the special political realities and social backgrounds of overseas regions, including unique systems, rituals, behavioral norms, and organizational forms used or created by indigenous Chinese in northern Myanmar.(4) Terms expressing the spiritual culture of overseas Chinese communities, including values, religious beliefs, literature, art, and aesthetic preferences.

Therefore, the design of the semantic hierarchy system for characteristic words in overseas Chinese dialects needs to consider the cultural and unique characteristics of the semantic system of characteristic words in overseas Chinese communities, referencing the semantic classification methods of WordNet and the Chinese Lexical Semantic Knowledge Base. Based on this, the principles of semantic classification and layering in this paper are:

First, using traditional grammatical categories as the outline and semantic categories as the goal. This principle ensures compatibility with classifications like WordNet and aligns with the characteristics of characteristic words, with different parts of speech having different semantic classifications.

Second, considering the social and cultural characteristics of overseas communities, with flexibility in the level of detail in classifications. Generally, nouns are numerous and finely classified, while other parts of speech are fewer and coarsely classified; concrete nouns are numerous and finely classified with hierarchical depths of 4-6 levels, while abstract nouns are fewer and coarsely classified to show general types.

Third, Classifying based on the daily perceptions of Chinese people, aiming for commonality and everyday use rather than academic specialization. Human beings classify things based on life experiences and knowledge to understand and utilize things, and this method can also be applied to the semantic system of vocabulary. For example, animal classification can be based on familiar categories like livestock, beasts, birds, and fish rather than biological academic classifications.

Based on these principles, characteristic words in overseas Chinese dialects are first divided into four categories: nouns, verbs, adjectives, and function words according to parts of speech. Then, semantic classification of different parts of speech is conducted based on life experience and knowledge. Finally, to accurately and efficiently describe the semantic categories of characteristic words, a semantic tag set is established according to the hierarchical classification above. Each semantic category at every level is marked with a unique code, such as nouns being marked with the initial letter "N," with N.1.1.1.2.1 being the semantic category code for livestock. This semantic classification tag set is imported into ontology modelling software to complete the construction of the semantic classification system for characteristic words.

4. The Collection of Characteristic Words and Knowledge in Overseas Chinese Communities

The data foundation for constructing a semantic network is the resource of characteristic word databases, which means that detailed and comprehensive characteristic words and their related knowledge need to be collected. The vocabulary and socio-cultural materials in existing works on overseas Chinese communities are extremely limited, making in-depth field investigations on vocabulary in these communities a crucial practical issue for constructing the semantic network.

The most commonly used method for traditional Chinese dialect vocabulary surveys is the translation method, where the investigator pronounces the dialect word according to a word list. Although this method is simple and quick, it has two major drawbacks for characteristic words: First, most characteristic words reflect unique knowledge and cultural backgrounds of overseas Chinese, with no corresponding concepts or expressions in Mandarin, making it difficult to comprehensively preset through a word list. Second, for characteristic words, the entry-based translation survey is a simple one-to-one translation, not reflecting the natural speech facts of the Chinese community, thus easily missing some vivid and important language information. This is especially true for some characteristic words with the same form as dialects in mainland China but with hidden special knowledge and cultural information. For example, the word "banana" exists in both the Zhenkang dialect of Yunnan and the Kokang dialect of Myanmar along the China-Myanmar border, with the same pronunciation and basic meaning. However, in the Kokang dialect, "banana" has a connotation of "not worrying about life" due to its homophony with the local dialect "pájiāo" and is thus a necessary offering in local sacrificial rituals (even using banana stems to make incense sticks). This cultural metaphor is absent in the Yunnan dialect's word for "banana", and such differences are difficult to discover through traditional vocabulary survey translation methods.

Therefore, the investigation method for characteristic words in overseas Chinese dialects must be reformed. In actual overseas field investigations, it is found that although overseas Chinese are very familiar with their natural and social environments, they rarely describe vocabulary and the knowledge it contains scientifically like an encyclopedia. Instead, they mix these words and their knowledge into daily conversations, folk songs, myths, proverbs, and nursery rhymes. For example, when asking a Chinese person about characteristic words related to the rubber industry in the Tunai Hakka dialect of Malaysia, he was vague. However, he later provided a wealth of local rubber industry vocabulary information when recounting his experience in the rubber industry. Therefore, the investigation of characteristic words in overseas Chinese dialects should not be restricted to mechanical word lists and sentence lists but can adopt more flexible investigation methods. Early American descriptive linguistics did not emphasize independent word list surveys but extracted words from texts to form word lists. Some European linguists mainly used oral conversations and picture inquiries, rarely using word and sentence entry lists. Based on practical experience in overseas dialect surveys, the investigation of characteristic words in overseas Chinese dialects can adopt a long-text data collection model centered on "knowledge-communication" extracting words from long texts [17]. This model emphasizes knowledge-based activation, collection, and presentation of the pronunciations, vocabulary, and expressions of individuals and groups in the language community. The recording process is not a question-and-answer session with a word list but a collective exchange and communication within the language community, a process of knowledge activation, collection, co-construction, falsification, and retention, reflecting the natural state of language occurrence. It starts from knowledge-based words reflecting concepts and behavioral notions, leading to various long-text materials conveying related knowledge, such as monologues and dialogues. This collection method can comprehensively reflect the synchronic linguistic facts and practices, systematically represent the speech expressions of unique knowledge and cultural communication content in the target language, and extract a large amount of vivid, authentic vocabulary data, making it an ideal practice for recording overseas Chinese dialect vocabulary.

Of course, this method still requires detailed and standardized operational steps to avoid missing key points and inefficiency. Specifically, the method involves three basic steps:

First, Participant Observation: Fully engage in the social life of overseas Chinese communities to holistically perceive the operational processes of the material culture and social system from a Chinese perspective, accumulating related vocabulary and expressions through observation. As folk activities are cyclical, an ideal participant observation requires at least one year of residence to observe the local people's seasonal production activities, religious ceremonies, and festive events.

Second, Obtain Long Text Materials Centered on Local "Knowledge-Communication" Scenarios: Through preliminary participant observation, understand the historical background, natural environment, and socio-cultural environment of the investigation site, and create focused interview outlines around the local special natural and socio-cultural phenomena vocabulary. Then, conduct targeted interviews to obtain long text materials. These materials should be comprehensive, reflecting daily life and knowledge. The content must cover various fields of language community life, such as daily production, commercial activities, religious customs, cultural education and entertainment, and interactions with nature, covering various populations, extensive events, and behavioral practices. The interview methods can draw on the experiences of cultural anthropology, ethnobiology (ethnobotany and ethnozoology), etc., and three interview methods are recommended:

(1) Key Informant Interviews: Carefully select individuals born and raised overseas who are particularly knowledgeable about certain characteristic items or activities in the Chinese community, such as folk artists, religious figures, Chinese politicians, craftsmen, etc., and conduct direct interviews and truthful recordings.

(2) Focus Group Interviews: Gather several Chinese individuals (preferably 3-5) to chat about a local natural or socio-cultural characteristic topic, constructing the outline of the item/event based on their knowledge, memories, and experiences. Topics can be prepared by the investigator in advance and guided during the chat, or chosen by the Chinese individuals based on their interests and knowledge experiences.

(3) Audio-Visual Guided Interviews: Track the development process of an event or phenomenon (e.g., festivals, weddings) in the Chinese community, and, with permission, record and photograph physical objects reflecting the natural environment and socio-cultural life, such as videos of food preparation, handicraft making, tool manufacturing, festival scenes, and wedding and funeral ceremonies. Use these audio-visual materials to guide Chinese individuals in explaining and narrating cultural customs.

Third, Extract Characteristic Words to Form Word Lists and Conduct Linguistic and Encyclopedic Knowledge Collection of Characteristic Words: With a word list of characteristic words, different investigation parameters need to be set for different semantic fields of characteristic words to form a relatively unified investigation model. For example, the investigation parameters for characteristic food names should mainly include required raw materials, preparation processes, nutritional value, cultural value, and cultural significance of the food. For overseas place names, the investigation parameters should include topographical features (e.g., altitude, geographical location), resources (e.g., mineral resources, water resources), naming rationale, and origins. Setting parameters based on references can prevent disorganized and illogical investigation content while ensuring comprehensive information.

This way, the collected and annotated data describe the semantic properties of word-level language units but can ultimately form three sub-databases of corpus resources: the lexical database, the example sentence database, and the discourse database. The example sentence database annotates the usage of each lexeme in real materials, and the discourse database provides knowledge and cultural information contained in each lexeme. This data can collectively provide more comprehensive materials for language research and more resources for applied research, as they best reflect word sense descriptions, syntactic realizations, and semantic collocation patterns for natural language processing.

5. Attribute Table and Annotation of Metadata for Characteristic Words in Overseas Chinese Dialects

To establish a semantic network, it is necessary to create a standardized metadata set, assigning essential and appropriately descriptive data attributes to each entry. With scientific data attributes, all characteristic words can be comprehensively and systematically annotated, forming a resource database of characteristic words.

To meet the needs of Chinese information processing and semantic knowledge representation, the key to describing data attributes is to first determine an appropriate set of attribute names and clearly define the attribute information and attribute values for each name. Considering the characteristics of characteristic words in overseas Chinese dialects, the design of data attribute content for characteristic words can adopt a "linguistic information + encyclopedic information" integrated model. The attribute content of linguistic information includes word class, synonyms, origin, IPA, basic meaning, example sentences, *etc.*, while the attribute content of encyclopedic information includes knowledge experience, cultural connotation, multimedia materials, long oral narratives, etc. Specifically, the attributes are as follows:

International Journal of Multiphysics
Volume 18, No. 3, 2024
ISSN: 1750-9548

ID: A unique, non-repetitive identification number for each entry. Each sense of a polysemous word is recorded as a separate entry and assigned a unique ID to distinguish between different senses.

Word: The determination of words (including phrases) and non-words to be included according to the standard.

Synonyms: Words with the same meaning.

Word Class: The part of speech information corresponding to the word.

Origin: Mainly includes self-created words, loanwords (specifying the particular language), archaic words, etc.

IPA: The International Phonetic Alphabet description of the word.

Semantic Category: The category to which the word belongs in the semantic hierarchy classification system.

Basic Meaning: A simple description of the word's meaning, establishing the basic marker for meaning division.

Example Sentences: Typical application examples of the word in the given sense.

Knowledge Experience: Unique knowledge and cultural information contained within the word itself.

Cultural Connotation: Unique cultural information contained within the word itself.

Multimedia Resources: Related images/audio/video of the word.

Long Oral Narratives: Long oral dialect narratives by native speakers centered on the entry, mainly focusing on the knowledge experience and cultural connotations related to the word.

Remarks: Other content that needs to be explained.

Since linguistic information attributes are commonly present in most lexical corpora, this section will not elaborate on them but will focus on the attribute content of encyclopedic information and its annotation methods.

(1) "Knowledge Experience" and its annotation

The content of life experiences in characteristic words refers to the knowledge accumulated and passed down by overseas Chinese about the biological and ecological environment, including natural geographic knowledge, knowledge of flora and fauna, the interdependent relationship between humans and nature, production labor knowledge, knowledge of natural resource utilization, the relationship between social organization and social life with the natural ecosystem, and so on. For example, the Kokang people in northern Myanmar have a unique medicinal soup called "Yaogen Chicken" that is consumed during the Dragon Boat Festival. The materials, preparation methods, and effects of this soup represent the life experience knowledge embedded in it. Current Chinese dialect research works often provide simple semantic descriptions of vocabulary. Apart from a few Chinese dialect dictionaries, most of the vocabulary data in appendices of research works follow the "Mandarin-Dialect comparison word list" model. Many vocabulary lists often give perfunctory explanations like "a type of food" or "a type of tool" for words with cultural characteristics in the language community, without supplementing encyclopedic information, leading to a lack of distinction from other words and even potential misunderstandings.

Of course, in actual annotation, given the complexity of knowledge, it is difficult to cover all aspects for each word entry. If we try to be exhaustive, it can be challenging to focus on key points and might lead to information overload due to excessive detail. Ethnobotany often uses a cataloging method when recording plants, including the plant's local name, scientific name, uses, edible parts, processing methods, etc. Inspired by this, the investigation of characteristic word knowledge content can also be systematized and standardized. Similar types of characteristic words can have the same explanatory parameters set. The explanatory format is defined by the essential meaning parameters required for explanation.

(2) "Cultural Connotations" and its annotation

Many characteristic words contain rich cultural connotations, including the reasons behind their names, cultural metaphors, legends, and stories associated with the words. Without this information, non-Chinese speakers may not correctly understand the precise meaning and usage of the words. For example, "húnxiàn" refers to a red hand string commonly worn by Kokang Chinese. If we simply define it as a red hand string, readers might equate it with ordinary hand strings. This hand string is usually made by Buddhist or Taoist monks or from red threads offered to ancestors during the Zhongyuan Festival. In the old Kokang dialect, " hóng" (red) is homophonous with " hún " (soul), hence it is called " húnxiàn " believed to protect and keep the soul safe. Although similar items exist in mainland China, their cultural connotations in overseas Chinese dialects may differ. For example, " gǒuqǐ" (goji berries) in some Southeast Asian Chinese communities is homophonous with "gānjū," (living comfortably), symbolizing a prosperous and stable life. These cultural nuances should be noted in the annotation.

(3) "Multimedia Resources" and its annotation

Most characteristic words refer to things or concepts unique to overseas Chinese communities. Even with detailed textual annotations, it might still be difficult for domestic readers to imagine what these things look like. Therefore, it is best to use illustrations or even recordings and videos to provide materials that reflect language characteristics or knowledge content. These mainly include recordings of sample states, pictures, and videos of physical objects or production and living processes, and audio-visual materials of native speakers narrating ecological and cultural knowledge. Because characteristic cultural words in one language may be difficult to understand in another language due to a lack of similar environments and cultural backgrounds, accompanying images or videos can give readers a clear and vivid conceptual understanding.

(4) "Long Oral Narratives" and its annotation

The encyclopedic knowledge information mentioned earlier is organized and narrated from the perspective of an outsider through field investigations and interviews by language researchers. "Long Oral Narratives," however, belong to the narrative of knowledge from an insider's perspective, i.e., the native speaker's oral account of the ecological and cultural knowledge contained in the vocabulary. Native speakers' oral accounts, from the participant's perspective, help eliminate potential misunderstandings about the vocabulary due to language and cultural barriers faced by researchers and readers, objectively reflecting the knowledge content embedded in the vocabulary. This method encourages native speakers to express the connotations of words in their language, stimulating their expressive ability, especially aiding in language preservation. Moreover, the oral content itself is an important corpus that can be a crucial tool for language protection and revitalization.

In summary, the annotation of characteristic words should reorganize the linguistic information, encyclopedic knowledge (including life experience and cultural connotations) to comprehensively and exhaustively describe the various linguistic attributes and knowledge content of cultural characteristic words from multiple dimensions. The linguistic attributes, life experience content, and cultural connotations should interrelate to form a unified system, enabling readers to understand the meaning and knowledge connotations of words without prior knowledge of the language and culture.

6. Issues in Determining Synonym Sets and Semantic Relationships for Characteristic Words in Overseas Chinese Dialects

The most critical aspect in the construction of a semantic network is determining the synonym sets and semantic mapping relationships for characteristic words.

6.1 Constructing Synonym Sets for Characteristic Words

A synonym relationship refers to the rational equivalence of meanings between different words in a language, where the semantic components are essentially consistent. In overseas Chinese dialects, it is common to have two or more expressions for the same concept, making synonym relationships prevalent. There are two types of synonym relationships for characteristic words in overseas Chinese dialects: (1) Completely Equivalent Synonyms: Different words can be used interchangeably, with fully equivalent semantic values. For example, in the Teochew dialect of Bangkok, Thailand, "diàndānchē" and "móduōsài" from English "motorcycle") both mean "motorcycle" and "chìpí" (and "pōlì" (from English "police") both mean "policeman." In these cases, the former are self-created words, while the latter are loanwords from English. (2) Non-equivalent Synonyms: Different words have similar rational meanings but cannot always be used interchangeably in certain contexts. Synonym relationships often differ in emotional connotation, degree of meaning, style of expression, and collocation, aiding in semantic differentiation [18]. For example, in the Guangzhou dialect of Chicago, USA, "báiguǐ", "lǎofān" and "guǐlǎo" all refer to Westerners, but "báiguǐ" carries a derogatory meaning, while the other two are neutral terms. Synonym relationships have been a focus of lexical semantics and computational semantics research, impacting the automatic semantic analysis and understanding by computers. Thus, describing the synonym relationships between characteristic words and constructing synonym sets is a prerequisite for establishing other semantic relationships. The synonym relationships discussed here include both complete synonyms and near synonyms, in line with the principles of constructing synonym sets.

After determining the Synsetsfor characteristic words, each synonym set should be assigned a unique SynsetID. Building semantic relationships with synonym sets as organizational units is also a necessary prerequisite to ensure compatibility of the characteristic word semantic network with lexical semantic knowledge bases of other languages, such as WordNet and CCD, facilitating the future merging and development of multilingual lexical semantic networks.

6.2 Determining Semantic Mapping Relationships between Characteristic Word Sets

Various cultural phenomena and knowledge in overseas Chinese communities are interconnected, and this association is part of lexical knowledge. In printed works and dictionaries, due to layout limitations, the original connections between cultural phenomena and knowledge are hard to reflect clearly through word lists or glossaries, with some lists even arranged alphabetically, completely disrupting the connections between vocabulary. The semantic network can reconstitute lexical semantic information into a complete, systematic cultural knowledge system, using "semantic mapping relationships" to annotate relational information and further explain the related semantic information of words, thus establishing connections between vocabulary and making it easier for readers to obtain all related lexical semantic information when querying an entry. The basic idea of WordNet is also to represent lexical semantics through relationships [19]. Specific words within and between different parts of speech have various semantic relationships. Here, we focus on the semantic mapping relationships within characteristic nouns.

I. Hypernym-Hyponym Relationship: This hierarchical semantic relationship, also known as semantic inclusion, involves one word being a superordinate (hypernym) of another, more specific word (hyponym) [20]. For example, in the Tunai Hakka dialect of Malaysia, "shùrén" is a hyponym of "shù" (tree). The hypernym-hyponym relationship is a main structure in the semantic network.

II. Synonym Relationship: Two nouns that can be interchanged in a certain context are considered synonyms. For instance, in the Teochew dialect of Bangkok, Thailand, "shífēngcuò" and "nǎitóucuò" both mean "luxury house" and can be used interchangeably.

III. Antonym Relationship: This oppositional relationship mainly includes three scenarios. (1) Polar Opposition: Words have completely opposite meanings, representing two extremes with intermediate expressions. (2) Complementary Opposition: Words are mutually exclusive, with no intermediate expression. For example, "bābā" (male offspring of a Chinese-Malay union) and "niángrě" (female offspring of a Chinese-Malay union) in the Min, Yue, and Hakka dialects of Malaysia. (3) Relational Opposition: Words share a common hypernym but have significant semantic differences, leading to a conventionalized opposition, such as "hánjì" (cold season) and "rèjì" (hot season) in the Teochew dialect of Bangkok, Thailand.

IV. Holonym-Meronym Relationship: This relationship involves one word (holonym) denoting a whole, while another word (meronym) denotes a part of it. It includes relationships between objects and their components, collectives and their members, and materials and the objects made from them. For example, "shùrén" (rubber tree) and "shùrénzi" (rubber seeds) in the Tunai Hakka dialect of Malaysia.

V. Troponymy Relationship: Words with this relationship semantically express the same thing and share a common hypernym. There are many such words in characteristic words, such as "yǎshēn" and "yǎcān" , both referring to a type of fruit in the Tunai Hakka dialect of Malaysia.

Once the five types of semantic relationships between characteristic words in the semantic network are identified, the collected characteristic words can be annotated for semantic relationships. In constructing synonym, holonym-meronym, troponymy, and antonym relationships, all words with semantic connections should be annotated, regardless of the semantic distance between them. For example, if words N1 and N2 have a holonym-meronym relationship, and words N2 and N3 have a holonym-meronym relationship, then words N1 and N3 also have a holonym-meronym relationship.

7. Issues of Data Storage and Querying in the Semantic Network of Characteristic Words

7.1 Data storage methods for the semantic network

There are three main ways to store semantic network data: traditional relational databases, graph databases, and RDF databases. A relational database uses a relational model to organize data and is the most common type of database model, storing data in rows and columns, and is particularly suitable for storing structured data. Most semantic networks, such as WordNet, are built using relational databases. RDF (Resource Description Framework) stores data in triples, consisting of two nodes and one edge, where the nodes can be two different entities. Each of these databases has its advantages, but they struggle with complex, unstructured data or long attribute descriptions. They also fail to effectively represent the hierarchical relationships and characteristics of semantic network data for characteristic words. Each synonym set in the characteristic words of overseas Chinese dialects includes two types of attributes: data properties (attributes of each word in the synonym set, such as word form, pinyin, part of speech, definition, semantic category, and semantic hierarchy) and object properties (attributes between synonym sets, such as semantic relationships between words). Given the multiple, complex data attributes of characteristic words, including unstructured data like images and videos, a graph database is more suitable due to its friendly attribute storage capabilities. A graph database uses a graph structure for semantic queries, representing and storing data with nodes, edges, and attributes. Semantic networks are a visual expression of linguistic knowledge, needing a specific form to present scattered, linear data, establish semantic associations between data, and display them through an intuitive visual interface. This makes it ideal for handling complex relational databases. With the help of Neo4j graph database, it is even possible to automatically generate knowledge graphs, facilitating the construction of knowledge graphs in the cultural domain of overseas Chinese dialects, and making it easier for researchers and users to query and understand.

To facilitate the use of the semantic network on a relational model platform, the primary data of the semantic network can first be converted into a relational model and stored using relational database management systems like MySQL. The semantic network can be stored as several data tables, including a main table of characteristic word data, a synonym set table, and a semantic relationship data table, linking various data tables through "ID, vocabulary" fields. If needed later, data migration techniques can be used to import data from the relational database into the graph database.

7.2 Online retrieval and browsing platform for the semantic network

The ultimate goal of the semantic network is to promote the use of characteristic word resources. To fully leverage its role in linguistic and cultural research and enhance its application value in natural language processing, an online advanced retrieval and browsing tool must be established. This tool should provide the corpus and statistical data resource information in the database to various users in a visualized manner. Therefore, a query function should be designed to enable the retrieval of various types of semantic information stored in the database of characteristic words in overseas Chinese dialects, thus truly reflecting the value of the lexical semantic network.

The characteristic word semantic network query system is designed to visually present the semantic information of characteristic words to users in various fields. Based on predefined search expressions, it allows users to search through vast amounts of data according to specific needs. In the lexical semantic network, each node represents a lexical meaning, with the basic content of a node being: {node code, word form, sense number, pronunciation, part of speech, basic definition, encyclopedic knowledge, word sense combination relations [word sense relationships]}. When a search term is entered into the semantic network's retrieval program, related words are linked together for a comprehensive semantic description. For example, entering a query for a specific word or linguistic structure quickly locates and extracts instances collected in the corpus, with the search results being data that meet the search criteria. The system functions of the characteristic word semantic network query platform can be broadly divided into two categories: database maintenance functions for management and maintenance personnel, and semantic information query functions for general users. For general users, the information that can be queried for each word includes metadata, linguistic information, and encyclopedic infomation (see Figure 1).

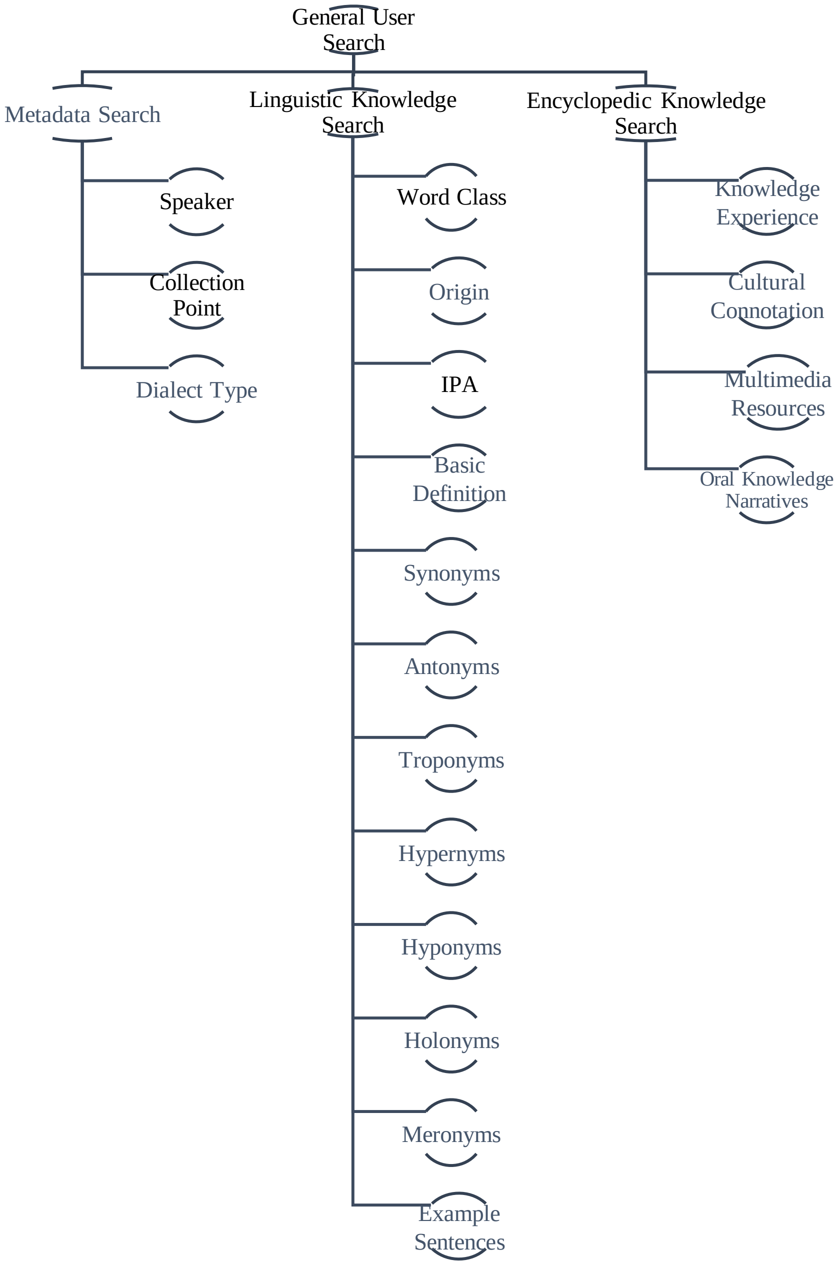

Figure 1 Structure diagram of general user query functions

Such a retrieval and browsing platform features convenient means for representing and retrieving semantic relationships, as well as descriptions of synonym sets, hypernym-hyponym relationships, and holonym-meronym relationships. In terms of content arrangement, resource presentation, management functions, and retrieval query methods, it facilitates management users in modifying and improving the database, and allows users to quickly access, query, and download the required corpus. It can provide ample, detailed, and visualized semantic resources for Chinese information processing, Chinese language research, and language teaching.

8. Conclusion

Overall, constructing a semantic network of characteristic words in the dialects of overseas Chinese communities involves linking numerous knowledge units to form a systematic knowledge organization system. This network can provide the latest and most comprehensive corpus for research on the vocabulary of Chinese dialects in overseas Chinese communities, serving as a valuable resource that can be effectively applied to various Chinese natural language processing tasks related to semantics. It can even offer an excellent platform for the preservation and dissemination of Chinese dialects and culture in overseas Chinese communities. However, we must also recognize that the construction of a characteristic word database for overseas Chinese dialects is significant but challenging. There are many issues in the construction of the semantic network that need further resolution in practice. For instance, due to the ambiguity and uncertainty of semantics, the setting of specific categories in the lexical semantic hierarchy may cause much debate. The process of annotating vocabulary is time-consuming and labor-intensive; hence, finding ways to expand and improve the scale and quality of the annotated corpus while minimizing manpower, material resources, and time costs is crucial. Additionally, we need to address how to continuously expand and improve the lexical semantic network to achieve the goal of dynamic updating and maintenance. Moreover, it is essential to explore what specific applications and service functions can be developed for the characteristic word semantic network platform.

References

[1] Chen X. J. An Overview of Chinese Dialects in Southeast Asian Chinese Communities Guangzhou: World Book Publishing Company, 2014.
[2] Li Y. M. Foreign Loanwords and Special Words in Teochew Dialect of Singapore, Dialects, 1991, Issue 1.
[3] Zhou C. J., Zhou Q. H. A Dictionary of Minnan Dialect in Singapore, Beijing: China Social Sciences Press, 2002.
[4] Lin T. A Dictionary of Dungan Language, Shanghai Dictionary Publishing House, 2023.
[5] Berners-Lee T., Hendler J., Lassila O. The Semantic Web-A New Form of Web Content that is Meaningful to Computers Will Unleash a Revolution of New Possibilities. Scientific American, 2001, 284(5): 34-43.
[6] Ren C., Cho S. A Framework for WordNet-based Word Sense Disambiguation, Journal of Korean Institute of Intelligent Systems, 2013, 23(4): 325-331.
[7] Roy A., Sarkar S., Purkayastha B. S. Knowledge Based Approaches to Nepali Word Sense Disambiguation, 2014, 3(3): 51-63.
[8] Srivastava A., Rehm G., Sasaki F. Improving Machine Translation through Linked Data. Prague Bulletin of Mathematical Linguistics, 2017, 108(1), 355-366.
[9] Nebhi K., Nerima L., Wehrli E. NERITS - A Machine Translation Mashup System Using Wikimeta and DBpedia, Lecture Notes in Computer Science, 2013, 7955: 312-318.
[10] Han B., Chen L., Tian X. Knowledge Based Collection Selection for Distributed Information Retrieval, Information Processing & Management, 2018, 54(1): 116-128.
[11] Chandurkar A., Bansal A. Information Retrieval from a Structured KnowledgeBase, International Conference on Semantic Computing. IEEE, 2017: 407-412.
[12] Saha A., Pahuja V., Khapra M M., et al. Complex Sequential Question Answering: Towards Learning to Converse Over Linked Question Answer Pairs with a Knowledge Graph, 2018.
[13] Zhang Z. Y, Zhang Q. Y. Lexical Semantics, Beijing, The Commercial Press, 2005, 4-8.
[14] Mei J. J, et al. Thesaurus of Synonyms, Shanghai Dictionary Publishing House, 1983.
[15] Yu J. S., Yu S. W. The Structure of the Chinese Concept Dictionary, Journal of Chinese Information Processing, 2002, 4, 13-21.
[16] Su X. C. Modern Chinese Classified Dictionary, The Commercial Press, 2013.
[17] Fan J. J. The Preservation and Protection of Endangered Languages in China, Jinan Journal (Philosophy & Social Sciences), 2018, 10, 1-18.
[18] Shi A. S. Semantics, Beijing: The Commercial Press, 2005.
[19] Fell baumC, Wordnet: An Electronic Lexical Database. Cambridge Massachusetees: MIT Press, 1999.
[20] Feng Z. W. Comprehensive Theory of Natural Language, Publishing House of Electronics Industry, 2018, p370.