A Corpus-based Study on the Phenomenon of Gender Marker Falsification in Southern Dialects

Cheng Xirong

Shangrao Preschool Education College, Jiangxi, Shangrao, 334000, China

Keywords: Vagueness; Dialects; Gender Markers; Corpus

Abstract: The phenomenon of grammaticalization is an important phenomenon in the development of Chinese grammar. Gender markers in Chinese dialects have also undergone a process of evolution from real to imaginary. Compared with the grammaticalization of gender markers in northern dialects, the grammaticalization of gender markers in southern dialects is more obvious and the process is more complete. There are great differences in terms of words, word order and grammatical nature when expressing gender markers in Chinese southern dialects. Gender markers vary in degree of grammaticalization and grammatical connotation in different stages of development. The construction of natural discourse corpus of Chinese dialects is an important part of national language resources construction. It has been widely used in language teaching research, multilingual communication and information services, language recognition and speaker identification, oral cultural heritage and protection, etc. Based on the Chinese corpus, this paper analyses the phenomenon of gender labeling in Southern dialect monographs, and explores the mechanism and evolution of gender labeling.

1. Introduction

Gender relationship has always been the most important relationship between biological individuals, and is also one of the foundations of human social relations and cultural forms [1]. Gender awareness is a common psychological phenomenon in human society. Different nations and societies have certain differences in reflecting this problem. There is a great difference between the northern and southern dialects in reflecting the sex of animals, especially domestic animals. Physiological sex is formed in the embryonic stage before birth, just like the identity characteristics of each person, such as blood type and fingerprints, remain unchanged for life [2]. Social gender is differentiated according to the characteristics of different roles that people assume in society and acquired in society. The semantic meaning of gender marker comes from the gender connotation attached to nouns referring to human beings or animals. The physiological natural gender is the semantic basis for the grammaticalization of Chinese gender marker [3]. The structural characteristics of such words are inseparable from their semantic development, and the ambiguity of gender-marking semantics is the main reason for the change of such lexical relationship. Chinese is a tonal language, and tones have a decisive role in words.

The corpus has a speech corpus and a text corpus. Spoken corpora can come from words spoken in the text, conceived fluent speech, and natural discourse [4]. The speech corpus has a wide range of applications in the fields of language teaching, language research, multilingual communication, information services, speech recognition, speaker recognition, endangered language and oral culture preservation and development. Gender markers are words or word-forming components that have a gender difference [5]. In Chinese, the term “sex” in gender terms such as negative nouns and masculine nouns refers to the natural gender represented by lexical meanings [6]. To construct a natural spoken language corpus in Chinese dialects, we should first clarify the service orientation and service objects of the corpus. The existence of dialect resources lies in the original ecological discourses that contain vivid expressions, rich geo-knowledge and cultural content, and are closely related to the masses' lives. [7] Gender markers differ in degree of ambiguity at each stage of development, and their grammatical connotations are also different. The development of the gender-marking process in different languages and dialects is not synchronized, but the basic

DOI: 10.25236/clla.2019.042

trajectory of the ambiguity is basically the same [8]. Based on the Chinese corpus, this paper analyzes the phenomenon of gender labeling in the dialects of southern China, and discusses its blurring mechanism and evolution law.

2. Sources of Verbal Markers in Southern Dialects

2.1 Corpus on Multi-Purpose Driving

Serving the language life of the masses is the basic aim and fundamental purpose of the construction of dialect spoken corpus. In modern southern dialects, gender markers mainly come from two categories: one is the noun that represents the appellation of a person, and the other is the noun that is the proper name of an animal. People's understanding of the categories and attributes of things is reflected in language, which is to summarize similar things, related things, similar behaviors or attributes into one word or category. The ancients compared the gender characteristics of human men and women to observe various things and phenomena, and subjectively classified them into different genders. Language is unique to human beings. Cognition is the extension and development of the negative meaning of the objective world in human subjective world. Objectively, it is the accumulation and development of human subjective cognition [9]. The key to the protection and promotion of human linguistic and cultural diversity lies in the continuous learning and inheritance of all languages and cultures. It is natural for human beings to extend their gender perception to animals and then determine it in the form of language. After the generalization of gender markers, not only the scope of use has changed, but also the category of use has changed. The spoken dialect corpus should be able to provide the dialect groups with all kinds of corpus they need to learn, disseminate and inherit their native dialects.

2.2 Gender Marker Generalization

The gender marker in Chinese dialect words may only be applied to a specific object at first, then its applicable object becomes more and more broad, which we call referential generalization. If the study of Chinese dialectology cannot fully and deeply investigate and study the discourse, there can be no real theoretical and practical innovation and form its own unique subject charm [10]. Southern dialects have great similarities in expressing animal sex, and the sources of these marker words are different. The gender markers in Southern dialects can be further found in the names of some small animals, plants or some inanimate nouns. When the semantics of sex markers are metaphorically duplicated in some animal or plant names, it is possible to further map them into the names of inanimate objects with similar productive characteristics. At this time, the original meaningful gender markers no longer have the Semantic Connotation of natural gender, but only represent a certain kind of characteristic meanings, which is the expression of lexical meaning grammaticalization. Chinese dialect natural spoken language corpus should be able to meet the basic needs of speech application engineering.

3. Gender Marker Deficiency Mechanism

The study, inheritance and dissemination of dialects are the corpus uses that should be considered first. In traditional Chinese cultural concepts, men are generally the symbol of “strong” while women are generally the symbol of “weak”. Most females are weak and slender, so it's easy to associate strength with gender characteristics. From the perspective of the existing spoken corpus, the naturalness of the spoken corpus is insufficient, the coverage of genre and genre is relatively narrow, and the discourse content is predesigned. When sex markers completely replace the natural gender semantics with the nominal additions, it becomes logical that the name of small animals should be further transferred to the inanimate nouns with similar characteristics. As morphemes, they have real meaning, mainly the name of the person and the gender of the animal, while the general name of the animal, the name of the plant or the name of the inanimate object is no longer a mark of natural gender. Natural discourse is not necessarily a simple, repetitive, incoherent discourse, nor is it a casual discourse.

The nature of discourse is closely related to language and language behavior. If the utterance recorder is able to conform to the discourse situation, it is possible to return the utterance to the natural state by guiding the speech without perceiving the other party. In order to calculate the text similarity, the words in the text must first be modeled, so the method of establishing the word vector model including the input layer, the middle detection, and the output layer is first given. A sentence in a text can be represented as a vector of words:

$$HWt = \frac{\sum_{i=1}^{N} D_i(x)}{N} \tag{1}$$

The intermediate layer can be obtained by cumulatively summing the vectors in the input layer:

$$w(t) = w_2 + (w_1 - w_2)\frac{T-t}{T} \tag{2}$$

At this time, the output of the word vector can be calculated using the following formula:

$$T(x,y) = \frac{x \bullet y}{\|x\|^2 \times \|y\|^2} = \frac{\sum x_1 y_1}{\sqrt{\sum x_1^{\,2}}\sqrt{\sum y_1^{\,2}}} \tag{3}$$

Syntax tree is a very important feature in the process of calculating the semantic similarity of short texts. It reflects the structure of sentences and the relationship between each component. For the diagonal matrix with singular values, we remove the smaller value, leaving only the larger singular value. The row and column of the orthogonal matrix of the two vectors are correspondingly removed, so that a simplified new concept space can be obtained to simplify the calculation. Other basic sememes are restricted by the similarity between the first basic sememes, while relational sememes are restricted by the first basic sememes and other basic sememes. Feature selection is a very important step. Usually features can be extracted from the stack and input queue. These features mainly come from words, parts of speech and dependency relations. After analyzing the related problems of semantic distance calculation, we began to use CNN-based algorithm and LSTM algorithm based on depth learning to calculate the text similarity of pure text respectively. The calculation results are shown in Fig. 1.

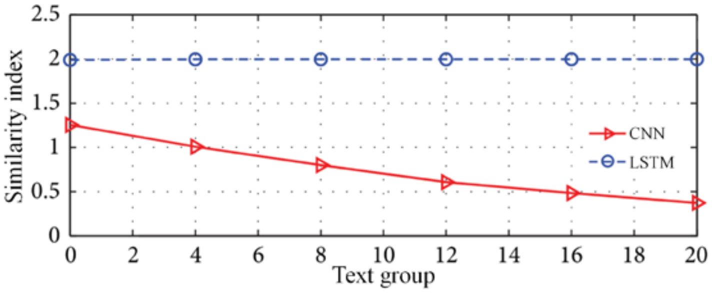

Fig. 1 Plain text similarity calculation results

As a corpus of spoken dialects, the corpus must fully reflect the linguistic facts and practices of dialects at the same time, and systematically reflect the unique knowledge and cultural content of the target dialect, that is, the traditional environmental knowledge based on the local natural community life of dialects. The gender markers in Chinese dialects are very complex, and there are many male and female markers. In natural sex, human males and females are similar to animals'males and females, so they can merge into positive, male or negative and female in language. As a complete system framework of spoken dialect corpus, it is impossible to embody comprehensiveness without the basic rules of magnitude, and it is also difficult to play a practical role in standardizing and citing the continuing expansion of the corpus in the future. Discourse in natural life is rich and colorful in both genre and expression style. Although there are obvious differences between the grammaticalized usage of gender markers in southern dialects and the grammatical categories of Russian and French, there are some important similarities between them that cannot be ignored. It is difficult for a single method to obtain effective discourses with large duration, multiple genres and rich content themes. The usage of grammaticalization is limited to the lexical level of Chinese and is realized by word formation. It is not systematic in grammatical rules and is not universal even in Chinese.

4. Conclusion

The grammaticalization of gender markers in southern Chinese dialects originates from nouns. Its natural gender is the semantic basis of grammaticalization. In a series of grammaticalization, gender markers gradually move from the lexical level to the grammatical level. Due to the large amount of data, corpora are mostly presented in the form of network databases except for electronic compact discs. The spoken Chinese dialect corpus not only provides corpus resources for language research and language engineering, but also serves the vast dialect areas and other people to learn dialects, spread regional culture and protect linguistic and cultural diversity. In some dialect words, the generalization of gender markers clearly reflects the idea that human beings are superior to animals. When the sex markers used for human beings are transferred to livestock, they will increase the esteem and affection. As a complete system framework of spoken dialect corpus, it is impossible to embody comprehensiveness without the basic rules of magnitude, and it is also difficult to play a practical role in standardizing and citing the continuing expansion of the corpus in the future. After the generalization of gender markers, not only the scope of use has changed, but also the category of use has changed. The spoken dialect corpus should be able to provide the dialect groups with all kinds of corpus they need to learn, disseminate and inherit their native dialects.

References

[1] Hualde, José Ignacio, Prieto P. Lenition of Intervocalic Alveolar Fricatives in Catalan and Spanish[J]. Phonetica, 2014, 71(2):109-127.

[2] Behravan H, Hautam?Ki V, Kinnunen T. Factors affecting i-vector based foreign accent recognition: A case study in spoken Finnish[J]. Speech Communication, 2015, 66:118-129.

[3] Lee, Clarissa. Historical Personalities: Tweeting Standard Narratives in the History of ScienceSocial Media Trends in Medical History[J]. Medical History, 2014, 58(04):627-628.

[4] Macmaster F P, Langevin L M, Jaworska N, et al. Corpus callosal morphology in youth with bipolar depression[J]. Bipolar Disorders, 2014, 16(8):889-893.

[5] Severin D. An electronic corpus of 15th-century Castilian cancionero manuscripts[J]. Health Policy, 2014, 70(2):792-795.

[6] Kuokkanen S, Polotsky A J, Chosich J, et al. Corpus luteum as a novel target of weight changes that contribute to impaired female reproductive physiology and function[J]. Systems Biology in Reproductive Medicine, 2016:1-16.

[7] Jedynak B M, Khudanpur S. Maximum Likelihood Set for Estimating a Probability Mass Function[J]. Neural Computation, 2005, 17(7):1508-1530.

[8] Morrill T, Dilley L, Forsythe H. Perceptual isochrony and prominence in spontaneous speech[J]. The Journal of the Acoustical Society of America, 2014, 136(4):2176-2176.

[9] Przemko K, Izabela H S, Anna L, et al. Relationship between Stereoscopic Vision, Visual Perception, and Microstructure Changes of Corpus Callosum and Occipital White Matter in the 4-Year-Old Very Low Birth Weight Children[J]. BioMed Research International, 2015, 2015:1-9.

[10] Jones Z E, Rourke C S. Speaker age effects on the voicing contrast of Tokyo Japanese stops[J]. Journal of the Acoustical Society of America, 2015, 137(4):2413-2413.