An Na

RESEARCH ON TAGGING THE ENGLISH AND CHINESE IDIOM BASED ON CORPUS

1. The study of English and Chinese idioms

Terminology in this field has always been problematic, and extended discussions of the problem include those by many linguists[1]. There is no generally agreed common vocabulary. Different terms are sometimes used to describe identical or very similar kinds of unit; at the same time, a single term may be used to denote very different phenomena. It is therefore essential to clarify the kinds of unit and phenomenon which I will be discussing.

1.1. Terminology in English

Considering the structure and meaning of English idioms, we divided them into three kinds. The first kind is phraseme, for example *because of, according to, look into* and so on. These units include several words and play the role of simple words in sentences. The second kind is pure idiom. These units often include two or more words and express the literal and deep meanings. The third kind is proverbs, sayings and similes. They are always independent sentences.

[1] See *Everaert M., Van Der Linden E., Schenk A., Schreuder R.* (eds.) Idioms: Structural and Psychological Perspectives. Lawrence Erlbaum, 1995; *Moon R.* Fixed Expressions and Idioms in English. Oxford, 1998; *Nunberg G., Sag I.A., Wasow Th*. Idioms // Language. Vol. 70. 1994. P. 491–538 ; and so on.

1.2. Terminology in Chinese

In Chinese, researchers have same ideas in the use of terms; they all use «idiom». But their understanding of idiom's function and bound are not the same. No matter in and abroad China, there is no unified view on idiom. But in the course of corpus annotation, we cannot avoid idiom. So, for the research on corpus annotation, idiom in this paper mainly refer to the phraseme and pure idiom in English, and idioms, customary usages, two-part allegorical sayings, terms and also abbreviations which has the same character with idioms and customary usages in Chinese.

2. Idiom tagging in foreign corpus

Corpus annotation cannot avoid idiom tagging. Foreign corpus gave idioms part-of-speech (POS) tagging and semantic tagging.

2.1. Part-of-speech (POS) tagging

Part-of-speech (POS) tagging, also called grammatical tagging, is the commonest form of corpus annotation, and was the first form of annotation to be developed by UCREL at Lancaster. Their POS tagging software for English text, CLAWS (the Constituent Likelihood Automatic Word-tagging System), has been continuously developed since the early 1980s. The latest version of the tagger, CLAWS4, was used to POS tag c.100 million words of the British National Corpus (BNC)[1]. The CLAWS system uses the method of IDIOMTAG introduced by Blackwell and Mcenery for idiom tagging.

British National Corpus applied POS-tag different from LOB Corpus. Idioms are given Multiword tags which denote their grammatical function, for example, <*w AV0*>***of course*** (adverb), <*w PRP*>***according to*** (preposition), <*w CJS*>***except that*** (conjunction).

[1] *Garside R., Marshall I.* Claws4: The Tagging of the LOB Corpus. 1983.

From above analysis, we can see that idiom POS tagging in foreign corpus is complete.

3. The Tagging Method of Chinese Idiom and the Problems in it

On the tagging of Chinese idiom, we mostly study on the basic processing of contemporary Chinese Corpus at Peking University Specification made for tagging corpus of *People Daily* and *standardization for corpus processing made* by ministry of education institute of applied linguistics. These two POS tagging standardization are the mostly used in corpus annotation in China, and also in idiom tagging. By these standardization, when tagging idioms, we only give them one tag, and put lexical category tag and POS tag on the only one tag, for example, 'in', 'ip'. The standardization of Peking University and ministry of education institute of applied linguistics are different. The standardization of Peking University is more detail. The tagset of Peking University includes more tags than ministry of education institute of applied linguistics, for example 'n' (noun), 'v' (verb), 'a' (adjective), 'd' (adverb) and so on.

3.1. The problems in the tagging methods above

The first Specification of Peking University only gives idiom a lexical tag, for example 'i' (i indicates idioms), 'l' (l indicates customary usages), 'j' (j indicates abbreviations). In the new Specification, Peking University improved on the first Specification. Based on the lexical category tagging, they gave idiom POS tag. In the course of syntactic processing, we can process these tags, for example, 'in', 'jv', 'la' and so on. But this method increased the complicacy of rules when describing idioms.

3.2. Our tagging method

For better disposing idiom in our Broadcasting Corpus, as a start, I did an exclusive research on Chinese idioms after tagging them in

our corpus, focusing on their syntactic distributions. Then I constructed a database of idiom usages with data obtained for this quantitative analysis. In that way, I took a data-driven research paradigm from the descriptive methodology that most former linguists took. Since my goal is to set up a tagging manual for idioms so that their proper syntactic functions can be reflected, this paradigm ensures a genuine report on the diversified usage of idioms. The outcome of my study is a detailed manual on idiom tagging, and an annotated corpus of idioms with both syntactic functions and stylistic information.

In the course of tagging idioms, we applied the method of two tags. First, we gave idiom POS tag, such as 'v', 'a', 'n', 'd'. Second, we gave idiom lexical category tag, such as 'i' (idioms), 'j' (abbreviations), 'l' (customary usages), 'gy' (two part allegorical sayings).

4. Conclusion

Computer linguistics needs an annotation corpus as its knowledge in the course of natural language processing. When the researchers could get information from the corpus, this annotation corpus can be proved valuable.