Research on the Dependency Distance of Quantifiers in Modern Chinese Based on Big Data Text Taking the Dot Quantifier "粒(particle)" as an Example

Linlin Zou

linlangyuxia@foxmail.com

College of Chinese Language and Culture, Jinan University, Guangzhou, China

Abstract: In the era of information explosion, quantifiers were often ignored in the past natural language processing (NLP). However, with the deepening of NLP system, the importance of quantifiers in the field of NLP, information retrieval, Chinese teaching, and corpus construction has become increasingly prominent. By analyzing its Nominal structure semantic categories, we can find that: category b (specific affairs) has the highest proportion, accounting for 74.98%. b4 (materials) account for 26.38% of b catagory; catagory f (social activities) is the least, appearing only twice, and its proportion can be ignored. Therefore, material subclass of specific affairs class is often used in combination with "粒". In Liu's paper, he believes that the mean dependency distance of Chinese is 3.662, so the average number of words that Chinese users need to remember when processing sentences is about 3. By calculating the entries in the corpus, we find that the distance between "粒(particle)" and the subject or topic of the sentence is about 3.217.

Keywords: Chinese, Dependency Distance, Quantifiers, 粒(particle), Big Data.

1 INTRODUCTION

The widespread use of quantifiers is an important feature that distinguishes modern Chinese from other languages of the Sino-Tibetan language family and Indo-European language family. In the research process of Chinese grammar history, Chinese quantifiers have always been the research object of linguists. In the 1950s, a quantifier was named by "暂拟汉语教学语法系统简述" (A Brief Introduction to the Tentative Chinese Teaching Grammar System): "The word that represents the quantitative unit of things or actions."

With the development of linguistics and the deepening of research, the research on new perspectives and new theories of Chinese quantifiers is increasing. Chinese quantifiers, typology, cognitive linguistics and other fields of cross research are also emerging.

BDEDM 2023, January 06-08, Changsha, People's Republic of China

DOI 10.4108/eai.6-1-2023.2330293

According to the research of linguist Dryer, modern Chinese (broadly speaking, modern Chinese refers to the language used by the Han nationality after the May 4th Movement in 1919) is regarded as an atypical VO language [1]. The focus of typological research is mostly on word order and the correlation between word order and other features. Quantifiers are not the focus of typological research, so that the importance of quantifiers in word order is often ignored. Until recently, Comrie [2] and Haspelmath et al. [3] began to take quantifiers as an important parameter in typology. At the same time, with the continuous development of natural language processing (NLP) systems, it is found that quantifiers play a crucial role in word order types in the language system where quantifiers are forced to be used. The division and classification of Chinese quantifiers have an important impact on the field of NLP, information retrieval, Chinese teaching, corpus construction, etc. Clarifying the division and classification of Chinese quantifiers and measuring the dependency distance between modern Chinese quantifiers and subjects will play an important role in the field of NLP, especially in ambiguity processing and parallel translation.

2 LINGUISTIC BACKGROUND

2.1 Modern Chinese

The concept of modern Chinese has broad and narrow meanings. In a broad sense, modern Chinese refers to the language used by the Han nationality after the May 4th Movement. It not only includes modern standard Chinese (Putonghua), but also includes Chinese dialects. In a narrow sense, modern Chinese only refers to the common language of modern Han nationality -- modern standard Chinese Putonghua.

2.2 Quantifier

There are mainly two explanations about the causes of Chinese quantifiers: the first one comes from the functional school; The second interpretation comes from the typological school.

In many studies, people use “classifier” to translate Chinese “quantifiers”. However, some scholars believe that the category of classifiers is not equal to the meaning category of quantifiers. This is largely because the classification function of Chinese classifiers is not obvious, and the systematicness and motivation are not strong. Zhu pointed out in his analysis that sometimes there is a certain relationship in the meaning between nouns and the individual quantifiers that match them, such as something with extensibility is used for “张 (Zhang)” modification, which is a sort of classification [4]. However, it should be noted that Zhu added a sentence “This is only a few cases”, which shows that the motivation of Chinese classifiers is not very strong. For example, when weighing “鱼 (fish)”, “毛巾 (towel)” and “消息 (news)”, we use “条 (tiao)”, while “一只狗 (a dog)” and “一匹狼 (a wolf)” use different classifications.

This paper believes that the connotation category of “classifiers” cannot cover the connotation category of Chinese quantifiers, so this paper still uses the old name of “quantifiers”, and divides Chinese noun quantifiers into four categories: individual quantifiers, collective quantifiers, capacity quantifiers and body metaphor quantifiers according to the theory of cognitive linguistics. The most numerous, frequently used and representative of Chinese noun quantifiers are the individual quantifiers. The study of Chinese individual quantifiers is also the focus of

linguists. More research on quantifiers based on big data will help us deepen our research on quantifiers and contribute to the development of Chinese NLP technology. This paper based on big data corpus mainly studies the dependency distance between Chinese quantifiers and subjects by taking quantifier “粒 (particle)” as an example. Examples are as follows:

- Nǐ shì yī lì jīnzi，zài nǎlǐ dōu huì fāguāng.

You are a piece of gold that shines everywhere.

你是一粒金子，在哪里都会发光.

- Gèrén de tóngqíng xīn rútóng yī lì shāzi.

Personal compassion is like a grain of sand.

个人的同情心如同一粒沙子.

2.3 Dependency Distance

Dependency distance is an important concept in the field of dependency grammar research. It refers to the linear distance (inserted reference) between two words with syntactic relations in a sentence. The size of dependency distance reflects the constraint of human cognitive mechanism on syntactic structure (inserted reference). Dependency grammar holds that dependency distance is the linear distance between the dominant word and the subordinate word in a sentence.

With the help of CoreNLP developed by Stanford University, the version used in this paper is 4.4.0. The dependency analysis tree with “粒” is as follows:

- Tā yī kǒu dōngxī dōu méi chī, chī jǐ lì huāshēngmǐ dé la！

He didn’t eat a mouthful of food, and he just ate a few peanuts!

他一口东西都没吃，吃几粒花生米得啦！

The following figure 1 shows the dependency structure tree of the example sentence.

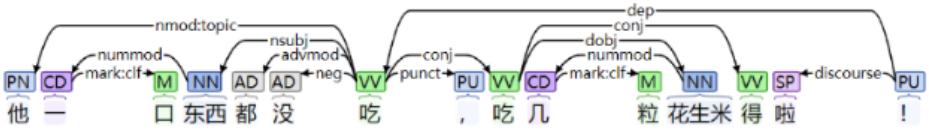

Figure 1. dependency structure tree of the example sentence.

In this sentence, the noun “东西 (food)” is the nominal subject of the sentence. The numeral “几 (a few)” and the quantifier “粒” forms a quantitative phrase to modify the noun “花生米 (peanut)”.

We want to use the analysis of dependency grammar to calculate the distance between quantifier “粒” and the subject or topic in the sentence, so as to show the closeness of the relationship between them.

3 A STUDY ON THE QUANTITATIVE PROPERTIES OF QUANTIFIERS

This paper mainly relies on the BCC corpus (http://bcc.blcu.edu.cn/) of Beijing Language and Culture University to conduct a diachronic retrieval of the quantifier "粒(particle)" in modern Chinese, and uses the dependency distance formula to calculate the dependency distance between the quantifier "粒(particle)" and the subject. Diachronic retrieval of it in BCC corpus is shown in the following figure 2:

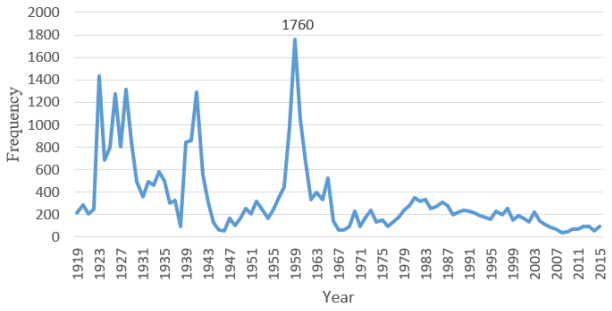

Figure 2. Frequency Statistics of Quantifiers by years.

In this section, we would discuss nominal structure (NS) behind it and quantitatively study its dependency distance with the subject or topic of the sentence.

3.1 A Brief Study on the Semantic Categories of "粒"

We counted the use of the NSs behind "粒" in the corpus, and summarized the NSs with frequency ranking within 1000. The following Table I shows the NSs of the first five and the last five and their frequency.

TABLE I. TABLE OF NS FREQUENCY STATISTICS

NS	*FO*	*F*	NS	*FO*	*F*
沙 (sand)	1	296	阿斯匹林 (Aspirin)	996	1
米 (rice)	2	278	VC (Vitamin C)	997	1
种子 (seed)	3	203	排骨 (spareribs)	998	1
沙子 (sand)	4	170	牛肉干 (dried beef)	999	1
饭 (meal)	5	118	蝌蚪 (tadpole)	1000	1

a. Frequency order (FO), Frequency (F)

We plotted the frequency and order of the NS behind “粒”, the following figure 3 is obtained.

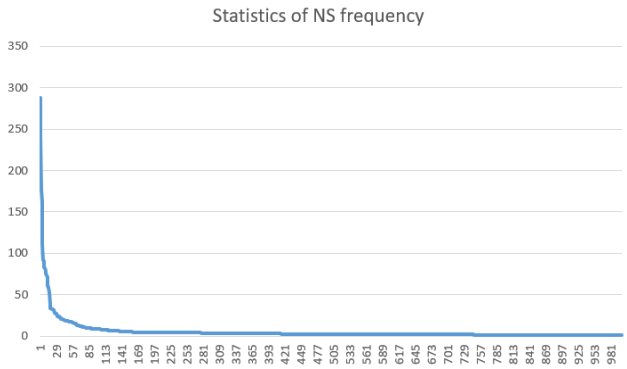

Figure 3. Statistics of NS frequency.

In Figure 3, the abscissa is the order and the ordinate is the frequency. The number of the first 100 NSs are 3268, accounting for 59% of the total. That is to say, the words behind “粒” have a lot of aggregation characteristics, often those words, such as “沙 (sand), 米 (rice), 饭 (rice), 石头 (stone), “珍珠 (pearl)”, etc.

Obviously, the frequency of NSs behind “粒” shows a power law distribution. The fitting function is as follows:

$$y = 619.91x^{-0.923} \quad (0 \le x \le 1000, x \in N^*) \tag{1}$$

Figure 3 and Formula 2 also show, to some extent, that it has a preference for NS, that is, it needs to match words of semantic categories within a certain category. Only those small words such as “沙子 (sand)” often appear, and can’t be collocated randomly, such as “一粒太阳 (a sun)”.

In order to explore the semantic category characteristics of NS, we need to use some semantic category tables in common domains. In the paper, “现代汉语分类词典” (A Thesaurus of Modern Chinese, TMC) is used as a reference semantic category table.

TMC is a dictionary that is classified and arranged according to the meaning of words. 82955 entries were collected in the TMC. It is developed based on the material warehouse and is also arranged according to the classification system of five semantic levels. There are 9 first level category, 62 second level category and 508 third level category. And there are 2057 in the fourth level category and 12659 in the fifth level category. The categories reflect the broad overview of the whole social life and Chinese vocabulary. In the meantime, TMC reflects the synonymous and antonymous relationship of words in detail. Therefore, TMC is a scientific semantic classification of modern Chinese [5-6]. So it was used to calculate the semantic category of quantifiers.

With the help of TMC, quantitative statistics of NSs semantic categories are shown in the figure 4 below:

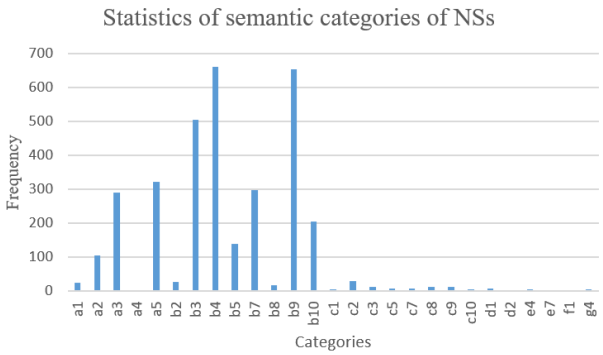

Figure 4. Statistics of semantic categories of NSs.

In Figure 4, the letters a-g represent the largest semantic category names, respectively: biology, specific affairs, abstract things, space-time, biological activities, social activities, and auxiliary words. The number followed by each letter represents the sub category under this category.

Through the analysis of Figure 4, it is clear that category b has the highest proportion, accounting for 74.98%. b4 (materials) account for 26.38% of b catagory; catagory f is the least, appearing only twice, and its proportion can be ignored.

3.2 A Quantitative Study on the Dependency Distance of "粒"

In order to further explore the deep cognitive characteristics of "粒", we use dependency grammar to calculate the distance between quantifiers and sentence subjects or topics.

This paper refers to Liu's formula for calculating the mean dependency distance (MDD) of Chinese sentences [7]. The formula is as follows:

$$MDD = \frac{1}{n-1}\sum_{i=1}^{n-1}|DD_i| \tag{2}$$

Because quantifiers are often combined with nouns to form NSs, most quantifiers cannot be directly controlled by the subject or theme of the sentence. Therefore, if we follow Liu's formula, we cannot directly calculate the dependency distance between quantifiers and sentence subjects. But the NS of nouns is often dominated by the sentence subject, so we calculate the dependency distance between NS and the subject to represent the dependency distance between quantifiers and subjects.

In Liu's paper, he thinks that the MDD of Chinese is 3.662, so the average number of words that Chinese users need to remember when processing sentences is about three. By calculating the entries in the corpus, we find that the distance between "粒" and the subject or topic of the sentence is about 3.217.

4 CONCLUSIONS

In the era of information explosion, we can access all kinds of texts every day, and we can also collect more and more texts. Then, after having such text data, it is necessary to do in-depth data processing on them. Therefore, it is necessary to conduct an in-depth study of the syntactic and semantic structure of the text. This paper selects the so-called stop words "quantifiers" that are often ignored in Chinese NLP research, and takes "粒" as an example to discuss the necessity of mining potential information of text based on big data.

Quantifiers were often ignored in the past natural language processing. However, with the deepening of NLP system, the importance of quantifiers in the field of NLP, information retrieval, Chinese teaching, and corpus construction has become increasingly prominent.

By analyzing its NS semantic categories, we can find that: specific affairs have the highest proportion, accounting for 74.98%. Materials account for 26.38% of specific affairs; social activities are the least, appearing only twice, and its proportion can be ignored. Therefore material subclass of specific affairs class are often used in combination with "粒". In Liu's paper, he believes that the MDD of Chinese is 3.662, so the average number of words that Chinese users need to remember when processing sentences is about 3 [7]. By calculating the entries in the corpus, we find that the distance between "粒" and the subject or topic of the sentence is about 3.217.

REFERENCES

[1] Dryer, M. S. 1991. SVO Language and the OV/VO typology. Journal of Linguistics 27(2): 443-82

[2] Comrie, B. 2008. The areal typology of Chinese: Between north and southeast asia. In Djamouri, R. Meisterernst B. & R. Sybesma (eds.), Chinese Linguistics in Leipzig, CLE 2:1-21. Paris: Press of EHESS/CRLAO

[3] Haspelmath M., Dryer, M. S., Gil, D. , & Comrie B. 2005. The World Atlas of Language Structures. Oxford: Oxford University Press.

[4] D. Zhu, Lecture Notes on Grammar. Beijing: The Commercial Press, 2013, pp. 48. (in Chinese)

[5] X. Su, A Thesaurus of Modern Chinese. Beijing: The Commercial Press, 2013. (in Chinese)

[6] Y. Li, "An Empirical Study on the Readability of Chinese Text Based on Chinese Textbook Corpus," (Doctor Thesis) Jinan University, Guangzhou, China, 2020. (in Chinese)

[7] H. Liu, Dependency Relation & Language Networks, Beijing: Science, 2022, pp. 9. (in Chinese)