The Initial Research of Mongolian Literary Corpus-Take the Text of Da.Nachugdorji's Work for Instance

YinhuaHai
School of Mongolian Studies
Inner Mongolia University
Hohhot, China
haiyh2008@163.com

Abstract—**Today, the Mongolian corpus is gradually developed from the basic resource construction stage to an in-depth research covering multi-level processing or author-corpus-based quantitative analysis, and multi-functional electronic dictionary's development. However, there are still many shortcomings and deficiencies in the collection, development and processing of literary corpus. In this paper, the author will introduces the corpus of Da.Nachugdorji's Literature and will discusses its profound significance, and fulfill multi-level processing such as lexical, syntactic and semantic annotation, as well as dissertates the preliminary processing research of Mongolian literary corpus from the perspective of statistics on the POS, word and phrase frequency and computation of lexical richness.**

Keywords-Mongolian literary corpus; processing; statistical research

I. LITERARY CORPUS AND THE DEVELOPING SIGNIFICANCE

As everyone knows, Da.Nachugdorji is one of the founders of Mongolian contemporary literature and a famous writer and poet. He began his literary creations in the 1920s and produced many poems, essays, comedy works and novels in his short life. His exquisite writing skills, unique language descriptions, rich and refined vocabularies provide readers with an aesthetic enjoyment for reading and appreciating his literary works, affecting a large number of Mongolian literary researchers and enthusiasts throughout the history (mainly from 1920s to 1940s). As it's commented, "the greatest contribution of his literary works is that he created the classic framework of the Mongolian contemporary literature. 'Classic' means a framework that can profoundly affect the hearts of Mongolian people" [1]. Therefore, the development of his work corpus is not only an integral part of the construction of all Mongolian literary corpus, but also the best way to scientifically protect and efficiently make use of his works, and provide valuable resources for many research fields in Mongolian studies.

The development of the corpus of "Da.Nachugdorji's Work" began in January 2016. In fact, it is a balanced corpus of two different texts of traditional Mongolian and Cyrillic Mongolian. The former has been completed; its Cyrillic version is in the developing stage. In terms of source of corpora, as a raw material, "Da.Nachugdorji" [Mongol] (Inner Mongolian People's Publishing House, the first printing in April 1981, a total of 436 pages) was used to save 106 works for developing the corpora, including (the style, number of works and the proportion of corpus):

(1) Political and lyric poems: 31, 19%;
(2) Relevant health poems: 40, 14%;
(3) Comedy: 1, 19%;
(4) Fiction prose: 29, 46.7%;
(5) Translated poems: 5, 1.3%.

For the related information of above-mentioned works, this paper uses Access database2016 to save that information such as the style, catalog and writing time of the works, as shown in Figure 1. The corpus of "Da.Nachugdorji's Work" was completely accomplished manually. The Mongolian text (Microsoft Word) was saved in the notepad++2016 format, and transformed into Latin text format with "Mongolian Proofreading System Editor4.0"[1], and finally achieve two different formats of preservation (Mongolian and Latin). The scale of corpus has reached about 628,000 words at present, whose sample is shown in Figure 2. For the text type, there are three formats, including the traditional Mongolian, Latin and Cyrillic, among which, research group for the Cyrillic texts is cooperating with the Institute of Language and Literature of Mongolian Academy of Sciences.

In fact, although the original intention of constructing a large-scale corpus was mainly to study vocabulary. However, a large number of literary works that have been included in the corpus have invisibly formed a rich field of literary appreciation and literary criticism. Corpus Stylistics is an interdisciplinary subject that combines literary text analysis with literary criticism using corpus linguistics [2]. Chinese scholar Guanghui Ma dissertated that "the method of corpus in linguistics can provide a set of effective methods and tools for literary research, so that people can make a more detailed, deeper and more specific description on literary works" [3]. "The stylistic style of a literary work is of unique charm of the author in the exchange of words, and it is the reflection of the author's creative intentions between the lines" [4]. Therefore, this paper believes that the research of stylistic style must be originated from the text of the works, which refers to the using and expression of language. The corpus had cuts into the research field of literary appreciation and literary criticism from such perspective. Many scholars at home and abroad have made some successful experiments in this field. A variety of retrieval software has been used in corpus stylistics, including WordSmith Tool, AntConc, Readability Analyzer and other open source tools, to have quantitative and qualitative analysis of literary texts from the perspectives of

This research is sponsored by the Project of NSSF of China (18BYY193), the project of Mongolian Language and Character of Inner Mongolia (MW-YB-201701).

978-1-7281-5014-7/19/$31.00 ©2019 IEEE

word frequency, key words, concordance, collocation, clusters, article plot and so on. At present, the retrieval of novels and translation works has become a hot topic especially. Take Xueqin Cao's *A Dream of Red Mansions* for example, in recent years, the study based on its corpus is very popular in studies of literature and translation. In the database of "CNKI" (www.cnki.net), over hundreds of articles can be retrieved with the topic of "Research on the corpus-based *A Dream of Red Mansions*", which can show that the corpus has an undoubtedly great potential for research of literature, translation, and foreign language teaching.

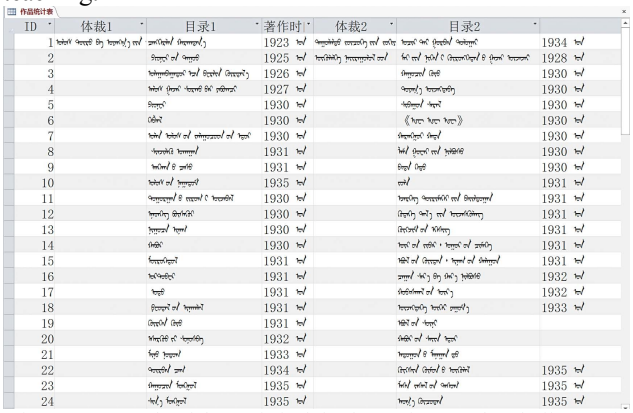

Figure 1. Sample of the statistical database of Da.Nachugdorji's Work

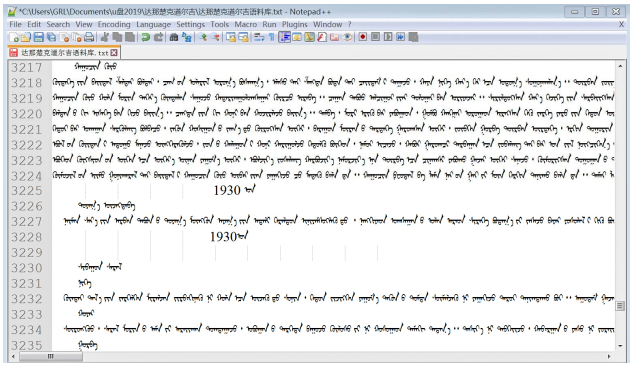

Figure 2. Sample of the corpus of " Da.Nachugdorji's Work" (Microsoft Mongolian version)

It is believed that corpus stylistics has become an effective supplement to traditional stylistics and one of the indicators of organic integration with modern information processing technology. Like scholar Wikberg of Swedish University of Oslo (1997: 312-325), corpus-based text analysis can complement with traditional text analysis. How to make full use of the corpus, word index in Mongolian language to serve literary studies, foreign language teaching and learning (such as data-driven learning DDL) is lack of necessary development and deeper annotation. Since 1983, a number of Mongolian corpora including *Mongolian Secret History* (Chinese-Mongol version in 1983), HODOM, TOD, DURBELJIN Mongol etc. has been developed in China. There has been accumulation in corpora resources such as monolingual corpus including "*One million words corpus of modern Mongolian*" (that is simply referred to as "*One Million*"), "*10 million words corpus of Mongolian*"; some bilingual or multilingual parallel corpora such as Mongolian-Chinese, Chinese-Mongolian, traditional Mongolian—Cyrillic Mongolian, Chinese-Mongolian-English corpus (its scale is usually calculated as "sentence pairs"); as well as the spoken corpus. However, due to the slow process of basic resources of research and development, application and sharing, immature network technology and other relevant causes, there was a lag in resources construction progress and low efficiency in its utilization. It is not even expanded to more users for its applicability and practical effect through balanced large-scale development or finishing. It is still necessary to make up for the deficiencies. For the construction type of corpus, the above-mentioned Mongolian monolingual corpus belongs to the heterogeneous[2] type, which cannot be collected and preserved with the premise of predetermining relevant principles and proportions, resulting in the lack of balance and systematicness of corpus. The results of past research have shown that the types of *One million words corpus of modern Mongolian* include novels, languages, newspapers, and politics, which respectively account for 19.6%, 50.3%, 9.8% and 22.9% of the total corpus [5]. Based on the occupation ratio, it can be known that the novel corpus only accounts for 19.6%. Compared with news, politics, and newspapers and periodicals, the proportion of literary corpus is relatively low. Moreover, the "*One Million*" only contains two articles such as "Xagučin Xüü" and "Sibagun Saral" by Da.Nachugdorji. Therefore, in view of the fact that the "*10 million words corpus of Mongolian*" has not yet been applied, there should be more focus on the development and processing of literary corpus that can fully reflect the richness and diversity of Mongolian vocabulary. This will fill the gaps in literary corpus and can solve some unbalanced problems on the development of Mongolian corpus.

Besides, previous studies have developed special literary corpora such as "HOHE SVDVR", "NIGEN DABHVR ASAR" and "TVNGGALAG TAMIR" for specific applicative needs, with some research from the perspectives of corpus annotation processing, application and dictionary compilation (the basic information about the subject of paper, the name and scale of corpus are shown in Table Ⅰ). One of the common features of it is the development of applicative software or lexicographic software to provide a convenient tool for users, and the modern corpus linguistics method used has played a role of "taking the lead before and after" on the development and research of Mongolian literary corpus. Drawing on and referring to the experience of above-mentioned special corpus, the research group started to develop the corpus of *Da.Nachugdorji's Work*.

TABLE I. STATISTICS OF CORPUS INFORMATION IN PREVIOUS RESEARCH

subject of paper	name of corpus	scale(ten thousand)
Construction of common noun's section in the electronic dictionary of khokh svdar[6]	khokh svdar	44.3
Construction of function word's section in the electronic dictionary of khokh svdar[7]		
Establishing a corpus of Mongolia-based on Cyrillic Mongolian material[8]	TVNGGALAG TAMIR	16.8
Processing and application of corpus of NIGEN DABHVR ASAR[9]	NIGEN DABHVR ASAR	12.4

II. THE PROCESSING AND STATISTICAL ANALYSIS OF THE LITERARY CORPUS

Constructing Mongolian multi-level tagged corpus through lexical-grammatical-semantic-pragmatic level

annotation and processing of the text on *Da.Nachugdorji's Work*, provide knowledge support for the statistics on word frequency, POS distribution, computation on lexical richness, the context and its adjacent collocation research. Since the mid-1990s, the main feature of the development of corpus linguistics has been the widespread application of corpora at all levels of linguistics. New Zealand's famous corpus researcher G. Kennedy summarized contemporary corpus research into different levels of vocabulary, syntax, text structure and text type [10]. In fact, the corpus of *Da.Nachugdorji's Work* was originally a raw corpus of Mongolian language knowledge carried by computer. This corpus, like G.Kennedy classification, needs to be tagged, analyzed and processed through lexical, syntactical, semantic and other aspects to become a useful knowledge resource-processed corpus. The significance of this research at each level is as follows.

A. Processing of the Corpus

For corpus processing, lexical, syntactical and semantic level, based on the corpus of *Da.Nachugdorji's Work*, can complete tagging the POS, phrase structure, sentence structure and sentence component as well as semantic classification and semantic role information annotation. In this paper, Da.Nachugdorji's representative work *Xagučin Xüü* is discussed as an example to dissertate relevant contents. For the POS tagging, the word segmentation, lexical variation labeling and POS tagging of the text can be completed with the Mglex system developed by the research group.

1) The text is a lexically tagged corpus including the word segmentation, additional components, and POS, for example: HODEGE/Ne2-YIN/Fc11 BAYIDAL/Ne2 $ALTAR/Ac B0LTAR/ Ac /Wp1 CAG/Ne1-VN/Fc11 VLARIL/Ne2 0R0G_A/Ac B0SH_A/Ve1+G_A/Ft22/Wp1 ELESU/Ne2 TEI/Ai $ANGDA/Ne1+N/Fc8 BVTA/Ne1 TAI/Ai CAYIDAM/Ne1-I/Fc31 DAG_A/Ve1+JV/Fn1,/Wp1 HAYA/Dc NIGE/Mu HAR_A/Ac GER/Ne1-ECE/Fc41 VTVG_A/Ne2 SVGVNAGLA/Ve2+N_A/Fs21 ./Wp1

This is a fragment of *Xagučin Xüü* excerpted from the part of "*One Million*" (pre-processed by manual proofreading) that was a corpus of POS tagging. The original corpus is "

After each word "/" is a code for POS and additional components which consisting of a mixture of uppercase and lowercase letters and numbers. For example, Ne1, Ne2, Ac, Ve1, Ve2, Db, Mu, and Ai are marks of POS which means countable nouns, uncountable nouns, nature adjectives, transitive verbs, intransitive verbs, state adverbs, cardinal numbers, and relational adjectives respectively. Fc, Ft, Fn and Fs represent different configuration additional component tag codes such as grammatical case, shape verb, auxiliary verb, and verb statement of Mongolian language. These marks are derived from the national standard "Information Technology-Mongolian word and expression marks for information processing" (GBT26235-2010)[3]. So lexical information annotation provides a basic knowledge for lexical analysis and POS statistics.

2) The text is a syntactically parsed corpus, for example: {{{HAGVCIN/Ac/ HUU/Ne1}NP2d/sbj {{{{HVLA/Ac/spcM0RI/Ne1+N/Fc8}NP2dDEGER_E/Oa}OP/adb{HONDELEN/Ac/adbSAGV/Ve2+JV/Fn1}VP2b}VP5b/abd{HATARI/Ve2+GVL/Fe1 1+V/Zv1+GSAGAR/FnC/adb{HUR/Ve2+CU/Fn1/adbIR_E/Vz1+BE/Fs14} VP2s/prd}VP3b}VP8h}VP10u ./Wp1}S11.

This is a simple sentence based on the syntactic information of Mongolian Phrase Treebank. The original corpus is "

Where { }S denotes the beginning of a sentence, { A N}, {N N} and {V V} denote the start of the NP and VP phrases in a sentence respectively. Similar to the above, followed by each word "/" is a mark of POS, "." means the period, "Wp1" means the end of a sentence; sbj, prd, obj, spc, adb represent the five sentence components of Mongolian such as subject, predicate, object, attributive and adverbial. Such syntactically parsed corpus can be used to accomplish from the shallow-level study of statistics about the combinational patterns of various POS tags and the sentence number, sentence length or average sentence length to the in-depth study of quantitative analysis on the general sentences, compound sentences or various sentence patterns and so on.

3) The text is a semantically tagged corpus, for example: HODEGE（Nob21）-YIN BAYIDAL（NuB）\uij $ALTAR（Aboe2 ）B0LTAR（Aboe2） ,/Wp1 CAG（Nc） -VN VLARIL（Ncg3）\uij 0R0G_A （Abot2 ） B0SHAG_A （Abot2 ）,/Wp1 ELESUTEI $ANGDA（Nob1 ）\oro BVTATAI CAYIDAM（Nob1）\oro-I DAGAJV （Vh11）,/Wp1 HAYA NIGE HAR_A（Abbo2GER（Nob22）\irt -ECE VTVG_A（Nbbu14）\uij SVGVNAGLAN_A（Vsb3）./Wp1

Semantic annotations include semantic categories and semantic case. The former is marked according to the relevant mark of the semantic classification system of "Information Technology Mongolian Word's Semantic Classification System and Its Mark Set" (MGC/01-05)[4]. The latter is based on the semantic case system as a standard reference to completes the semantic notation that was developed by "The Frame Research on Mongolian Language" [11].

The mark in the brackets () after each word indicates the semantic classification information which it belongs to, such as Nob21 (natural place), NuB (condition), Aboe2 (disordered substance's attribute), Abot2 (dynamic attribute), Vh11 (moving action), Abbo2 (ordinary), Nbbu14 (natural phenomenon), Vsb3 (diffuse state), etc.; after each word "\" is followed by semantic case, such as uij (subject), oro (placement), irt (source), TVS (objective) and so on.

The core problem of semantic annotation is semantic disambiguation. In the above text, for example, HAR_A is a polysemous word with 11 different meanings such as (1) black ink, black; (2) dirty; (3) ordinary, rough; (4) pure, empty; ... [12]. It is labeled here as the third meaning according to the current context, meaning "ordinary, rough" rather than "color", which provides a basis for automatic disambiguation.

Through the above annotation processing, a multi-level finishing corpus with lexical-syntactical-semantic information (the sample corpus is shown in Figure 3) can be obtained, providing detailed knowledge resources for other applications such as machine translation, text retrieval, semantic automatic processing, and tree-bank development.

B. Statistical Analysis of the Corpus

Using the preliminary processed corpus with word segmentation and part-of-speech tagging, according to the

principle of statistical linguistics, the paper will discuss a research on word class, word frequency statistics and lexical richness calculation ant its importance by the text of "Xagučin Xüü".

1) For word class, a statistical analysis on all of text of "Xagučin Xüü" and its related word class information was conducted. According to the statistics, "Xagučin Xüü" has 405 words and 34 fixed phrases, which can be classified into 13 word classes. The corresponding number and proportion of each word class are shown in Table Ⅱ.

Table Ⅱ only shows the statistical results of the word class information of an article. According to the statistical data on Da.Nachugdorji's works in Figure 1, his literary works include political poetry, lyric poetry, healthy poetry, comedy, novels, prose and translated poetry, and the corpus also has a large scale. Among it, the 439 words of the prose "Xagučin Xüü" are involved in 13 word classes. Because *Xagučin Xüü* is the author's representative work, the study can use its 13 word classes as a basic standard to statistically analyze the percentage of Mongolian word classes in various stylistic articles in the corpus of *Da.Nachugdorji's Work*, which can reveal the distribution feature of each word class in different stylistic articles (of course, it is necessary to tag the POS to the corpus firstly).

POS	word												fixed phrase							
	noun						verb	function word					compound word					idiom	fixed word	term
	noun	quantifier	time and potion	adjective	Time word	pronoun		adverb	statement word	postposition	conjunction	modal word	compound noun	compound adjective	compound time and position	compound verb	compound adverb	verbal idiom		Personal name
number	142	2	17	40	16	10	17	133	3	14	4	6	14	5	1	4	1	4	3	2
ratio	32.35%	0.46%	3.87%	9.11%	3.64%	2.28%	3.87%	30.30%	0.68%	3.19%	0.91%	1.37%	3.19%	1.14%	0.23%	0.91%	0.23%	0.91%	0.68%	0.46%
	55.58%						30.30%	6.38%					5.7%					0.91%	0.68%	0.46%
total	92.25%												7.75%							

Figure 3. Sample of the processed corpus of Da.Nachugdorji

TABLE II. A STATISTICS ON OCCURRENCE NUMBER AND OCCUPATION RATIO OF POS ON *XAGUČIN XÜÜ*

2) The word frequency statistics and the lexical richness calculation can be accomplished furtherly. The purpose is to use a large amount of real language data to analyze the rule and pattern of Mongolian language in real using from researching the distribution frequency of the lexical information.

According to the statistical findings, the word accounted for 92.25%, the fixed phrase accounted for 7.75% in 439 words of the whole text (punctuation excluded; the fixed phrase is counted as a word unit); from the POS distribution, there are 13 word classes. The word frequency of each word class has been counted separately, such as the noun word frequency statistics in Table Ⅲ and the fixed word frequency statistics in TableⅣ. The order of the proportion of each word class in mathematical order from high to low is expressed as follows:

a. Word class: noun > verb > adjective > time and position word = pronoun > numeral > statement word> time word > conjunction > postposition > adverb > quantifier > modal word;

b. Fixed phrase class: compound noun > compound adjective > compound verb > verbal idiom > fixed word > personal name > compound time and position word > compound adverb;

3) The total number of words and fixed phrases is calculated. The proportion of nouns with 66.6% in the whole article is relatively high, followed by verbs with 32.12%, and other words had 7.29%.

The proportions according to the number of occurrences of words are arranged in descending order as follows (the specific number of which is shown in Table Ⅴ)

1 occurrence of words > 2 occurrence words > 3 occurrence words > 4 occurrence words > 7 occurrence words > 5 occurrence words = 6 occurrence words = 8 occurrence words = 9 appearance words;

According to the principles of statistical linguistics and corpus stylistics, the high and low frequency of words have a certain relationship with the lexical richness of that article. In TableⅢ, the words with frequency of 0.23% in *Xagučin Xüü* occupy 79.2% of all texts (regardless of repeated vocabulary), and the words with frequency of 0.23%-0.91% occupy 98.08%. The statistical results show that *Xagučin Xüü* sentences had many low-frequency words and few

high-frequency words, and its vocabulary consists of many different low-frequency words.

TABLE III. FREQUENCY STATISTICS OF NOUN APPEAR ED IN THE TEXT OF "XAGUČIN XÜÜ"

Noun	ᠨᠢᠭᠡ᠂	ᠬᠡᠭᠡᠭᠡᠳ᠂	ᠬᠡᠯᠡᠨ᠂	ᠨᠠᠭᠠᠳ᠂	ᠮᠣᠩᠭᠣᠯ᠂	ᠬᠡᠪᠴᠢᠶᠡᠨ᠂	ᠪᠠᠭᠴᠠᠳ᠂	ᠬᠡᠪᠯᠡᠯ᠂	ᠪᠣᠯᠣᠨ᠂	ᠠᠭᠤᠷ᠂	ᠬᠣᠭᠣᠷᠣᠨ᠂	ᠬᠦᠮᠦᠨ᠂	ᠪᠠᠶᠢᠨ᠂	ᠪᠠᠶᠢᠭᠠᠯ᠂	ᠪᠠᠶᠢᠭᠤᠯᠯᠠᠭ᠂	ᠪᠠᠶᠢᠷ᠂	…
Number	3	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	…
Frequency	0.68%	0.23%	0.23%	0.23%	0.23%	0.23%	0.23%	0.23%	0.23%	0.46%	0.23%	0.23%	0.23%	0.23%	0.23%	0.23%	…
total	142noun																

It can be seen that the different words used in the author's sentence are very rich. According to Read (2000), lexical richness is the proportion of less common words or advanced vocabulary in the author's text. Therefore, there are 439 words totally except for repeated types. The total number of different words in the text is 431, which indicating that the lexical repetition rate is very low, the author's productive vocabulary is very high, and only about 2% of the words in this article belong to high frequency words. This result can fully indicate that the ratio of low-frequency words in the super 400 vocabularies to the total tokens of the text is more than 98%, mainly because the author's productive vocabulary is too high, unrepeated and advanced vocabulary appear more.

TABLE IV. FREQUENCY STATISTICS OF FIXED PHRASES APPEARED IN THE TEXT OF "XAGUČIN XÜÜ"

phrase	ᠮᠣᠩᠭᠣᠯ ᠮᠣᠩᠭᠣᠯ᠂	ᠪᠠᠭᠴᠠᠳ ᠬᠡᠪᠯᠡᠯ᠂	ᠪᠠᠶᠢᠷ᠂	ᠪᠠᠶᠢᠷᠢᠨ ᠬᠦᠮᠦᠨ᠂	ᠪᠠᠶᠢᠭᠤᠯᠯᠠᠭ᠂	ᠪᠠᠶᠢᠭᠠᠯ᠂	ᠪᠣᠯᠣᠨ᠂	ᠪᠣᠯᠣᠭᠰᠠᠨ᠂	ᠪᠣᠯᠵᠤ᠂	ᠪᠣᠯᠵᠤ ᠪᠠᠶᠢᠭᠠᠯ᠂	ᠪᠣᠯᠵᠤ ᠪᠠᠶᠢᠭᠤᠯᠯᠠᠭ᠂	ᠪᠣᠯᠵᠤ ᠪᠠᠶᠢᠨ᠂	ᠪᠣᠯᠵᠤ ᠪᠣᠯᠣᠨ᠂	ᠪᠣᠯᠵᠤ ᠪᠣᠯᠣᠭᠰᠠᠨ᠂	ᠪᠣᠯᠵᠤ ᠪᠣᠯᠣᠨ᠂	ᠪᠣᠯᠵᠤ ᠪᠣᠶᠤ᠂	ᠪᠣᠯᠵᠤ ᠪᠠᠶᠢᠭᠠᠯ᠂	ᠪᠣᠯᠵᠤ ᠪᠠᠶᠢᠭᠤᠯᠯᠠᠭ᠂	ᠪᠣᠯᠵᠤ ᠪᠠᠶᠢᠨ᠂	ᠪᠣᠯᠵᠤ ᠪᠣᠯᠣᠨ᠂	ᠪᠣᠯᠵᠤ ᠪᠣᠯᠣᠭᠰᠠᠨ᠂	ᠪᠣᠯᠵᠤ ᠪᠣᠶᠤ᠂	ᠪᠣᠶᠤ᠂	ᠪᠣᠶᠤ ᠪᠠᠶᠢᠭᠠᠯ᠂	ᠪᠣᠶᠤ ᠪᠠᠶᠢᠭᠤᠯᠯᠠᠭ᠂	ᠪᠣᠶᠤ ᠪᠠᠶᠢᠨ᠂	ᠪᠣᠶᠤ ᠪᠣᠯᠣᠨ᠂	ᠪᠣᠶᠤ ᠪᠣᠯᠣᠭᠰᠠᠨ᠂	ᠪᠣᠶᠤ ᠪᠣᠯᠵᠤ᠂	ᠪᠣᠶᠤ ᠪᠣᠶᠤ᠂	ᠪᠣᠶᠤ ᠪᠣᠶᠤ ᠪᠣᠯᠵᠤ᠂	ᠪᠣᠶᠤ ᠪᠣᠶᠤ ᠪᠣᠯᠣᠭᠰᠠᠨ᠂	ᠪᠣᠶᠤ ᠪᠣᠶᠤ ᠪᠣᠯᠣᠨ᠂	ᠪᠣᠶᠤ ᠪᠣᠶᠤ ᠪᠣᠶᠤ᠂	ᠪᠣᠶᠤ ᠪᠣᠶᠤ ᠪᠣᠶᠤ ᠪᠣᠯᠵᠤ᠂	ᠪᠣᠶᠤ ᠪᠣᠶᠤ ᠪᠣᠶᠤ ᠪᠣᠯᠣᠭᠰᠠᠨ᠂	ᠪᠣᠶᠤ ᠪᠣᠶᠤ ᠪᠣᠶᠤ ᠪᠣᠯᠣᠨ᠂	ᠪᠣᠶᠤ ᠪᠣᠶᠤ ᠪᠣᠶᠤ ᠪᠣᠶᠤ᠂
number	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1
frequency	0.23%	0.23%	0.23%	0.23%	0.23%	0.23%	0.46%	0.23%	0.23%	0.23%	0.23%	0.23%	0.23%	0.23%	0.23%	0.23%	0.23%	0.23%	0.23%	0.23%	0.23%	0.23%	0.23%	0.46%	0.23%	0.23%	0.23%	0.23%	0.23%	0.23%	0.23%	0.23%
total	34 fixed phrase																															

TABLE V. STATISTICS ON THE NUMBER OF OCCURRENCES AND THE CORRESPONDING WORD'S NUMBER AND ITS PROPORTION

number of occurrences	1	2	3	4	5	6	7	8	9
word's number	248	355	17	7	1	1	2	1	1
proportion	79.23%	11.18%	5.43%	2.24%	0.23%	0.23%	0.46%	0.23%	0.23%

There are many measuring methods of lexical richness in general. Scholars of domestic and foreign had use different measuring methods for different research purposes. For example, Linnarud (1986) measures overall lexical abilities from four aspects: lexical uniqueness, lexical complexity, lexical diversity, and lexical density [13]. Based on this corpus, many different topics can be accomplished with employing modern technology and methods such as the calculation on the proportion of lexical/content words such as nouns, verbs and adjectives in the composition, the proportion of using different words, and the scope of vocabulary used by the author et all.

III. CONCLUSION

The corpus of *Da.Nachugdorji's Work* is only the starting point for digital research and development of the corpus of Mongolian literary works. Expanding and perfecting the scale and quality of this corpus is the core of the further study. Considering it as a premise, the research group will develop a lot of literary corpora such as "100 pieces of ancient literature of Mongolian" (1-4 volumes) (Mongol, C. DamdinSurung, 1979) and "the Classic Novels of Mongolian" (Mongol, 2009), "the Collected Work of R.Choinom" (Transcoded by B. DamdinRurung, 2014) , laying the foundation for the construction of large-scale literary corpus. On the other hand, one of the gratifying things is that "the Complete Works of Da.Nachugdorji" edited by Y.Temurjin(volumes 1-3) [14], was published by Yuanfang Publishing House in November 2016. In the near future, the research group can reference or use this version to regulate or expand the corpus, so as to achieve the perfect quality indicators of resources. At the same time, the research group will cooperate with the Institute of Language and Literature of the Mongolian Academy of Science to develop its Cyrillic edition. It is expected to achieve the sharing of different cultural resources and provide learners with a large number of real and natural corpus, which will become a teaching resources of learning and application.

REFERENCES

[1] Reprint from an interview with the starting ceremony of "the Complete Works of Da.Nachugdorji" and the symposium to commemorate the 110th anniversary of the birth of Da.Nachugdorji.Hohhot,China, December 16, 2016.

[2] MAHLBERG M : Corpus Stylistics and Dickens's fiction.New York:Routledge.2013, 5-14.

[3] GuanhuiMa, A Corpus-based Study of Novel Stylistics, Journal of Changshu Institute of Technology, No. 5, 2005, 4-6.

[4] [10] PinganHe, Corpus Linguistics and English Teaching, Foreign Language Teaching and Research Press, October 2004, 34, 9.

[5] HuashaBao,Badam-odsar : A status analysis and improvement strategy of Mongolian corpus, language calculation and content-

based text processing - Proceedings of the 7th National Conference on Computational Linguistics, Harbin, August 2003, 346-350.

[6] Hastuya_a, Construction of common noun's section in the electronic dictionary of khokh svdar, Master's degree thesis of Inner Mongolia University, May2008.

[7] GuirongLi, Construction of function word's section in the electronic dictionary of khokh svdar,, Master's degree thesis of Inner Mongolia University, May2010.

[8] Purubsurung, Establishing a corpus of Mongolia-based on Cyrillic Mongolian material, Doctor's degree thesis of Inner Mongolia University, May 2015.

[9] Xiaojuan, Processing and application of corpus of NIGEN DABHVR ASAR, Master's degree thesis of Inner Mongolia University, May 2015.

[11] D.Sarn_a, the Frame Research on Mongolian Language, Liaoning Nationalities Publishing House, March 2013.

[12] Mongolian Dictionary Compilation Group: "Mongolian Dictionary", Inner Mongolia People's Publishing House, November 1997, 1238

[13] JianlinChen, Research on the lexical richness of two genres of college English majors based on CEW corpus, Journal of Tianjin Foreign Studies University, No4, 2011.

[14] Tiejun edited, "the Complete Works of Da.Nachugdorji",Yuanfang Publishing House, November 2016.