*International Journal of Knowledge and Language Processing* Volume 11, Number 1, **2020**  www.ijklp.org KLP International ©2020 ISSN 2191-2734 pp.24-31

## The Construction of the Verb Sub-library of VLAGAN EGULEN-U NILBVSVN

Dulan<sup>1</sup> and Dabhvrbayar<sup>2</sup> <sup>1</sup>School of Mongolian Studies Inner Mongolia University No.235 College Road Saihan District Hohhot, Inner Mongolia Autonomous Region, 010021, China 1026634236@ qq.com

<sup>2</sup> School of Mongolian Studies Inner Mongolia University No.235 College Road Saihan District Hohhot, Inner Mongolia Autonomous Region, 010021, China dabhvrbayar@163.com

## Selected paper from CLSW2019

**ABSTRACT.** VLAGAN EGULEN-U NILBVSVN is one of the greatest works of the famous Mongolian writer Injannasi, who plays a very important role in the history of Mongolian literature. In this paper, all the 1478 verbs in the corpus of VLAGAN EGULEN-U NILBVSVN are analyzed by Mongolian Information Processing. While analyzing the construction of the verb sub-library of VLAGAN EGULEN-U NILBVSVN, the paper introduces the way of compiling VLAGAN EGULEN-U NILBVSVN electronic dictionary and the process of recording its information contents such as entries, pronunciation, part of speech, syllable, paraphrases, word frequency, word frequency level, morphological form and example sentences.

Keywords: VLAGAN EGULEN-U NILBVSVN, the verb sub-library, fields, field's value

1. Introduction. Injannasi (1837-1892) is a well-known writer, poet and theoretical critic in modern Mongolian literary history. He has left behind a mass of works such as HOHE SVDVR, NIGEN DABHVR ASAR, VLAGAN\_A VHILAHV TINGHIM, VLAGAN EGULEN-U NILBVSVN, SARAN HOHOHE-YIN \$ASTIR, CILAGVN DER E-YIN

*\$IGUMJILEL, NOMCI-YIN HAGVRMAG, CAGAN EGULE* and *SARAN-V DAGVLAL*. Among them, *VLAGAN EGULEN-U NILBVSVN* is his first novel and has an obvious impact on his following creation of works.

Until now, based on the corpus of Injannasi's works, three electronic dictionaries have been compiled, i.e. HOHE SVDVR electronic dictionary, NIGEN DABHVR ASAR multi-functional electronic dictionary and VLAGAN A VHILAHV TINGHIM multi-functional electronic dictionary (which contain sub-libraries of parts of speech, work introduction, poetry, related papers and related books). Both the corpus processing and the constructing of the verb sub-library of VLAGAN EGULEN-U NILBVSVN contributes to the compilation of Injannasi's multi-functional electronic dictionary. Furthermore, it will enrich the Mongolian corpus materials and be used for Mongolian information retrieving and machine translation. Therefore, constructing the verb sub-library based on the corpus of VLAGAN EGULEN-U NILBVSVN has become significant.

To construct the verb sub-library, the first step is to process the corpus, including proofreading, tagging phrases and morphological features. Then, verbs are extracted from the corpus and recorded word by word. Finally, the word information such as pronunciation, part of speech, syllable, paraphrases, word frequency, word frequency level, morphological form and example sentences will be matched with corresponding verbs.

2. **The Source of the Corpus.** From 2005 to 2010, the modern Mongolian corpus has been expanded from 5 million words to 10 million words. *VLAGAN EGULEN-U NILBVSVN* is one of these additions. This corpus is based on the 1990 edition published by the Inner Mongolia Peoples Publishing House and following the Latin transcription rules of the Inner Mongolia University modern Mongolian corpus. The corpus sample is shown below:

□ VEN.txt - 记事本
文件(E)编辑(E)格式(Q) 查看(M) 帮助(H)
Page1 VLAGAN EGULEN-U NILBVSVN-V VG OROSIL INIYEDUMTU MVNGHAG-VD VYILAMATAGAI MERGED-I DEMEI BITEGEI ELEGLE, EMGENIL-UN SECEN-I VVLIHAI BALAMAD TAN YAGVNAI MEDEMUI. MINU BEY_E NILH_A-YIN UY_E-DUR BICHAN CECERLIG-TUR NAGADVHVY_A, TERE HU HURIYELENG-UN BARAGVN HOYITV ONCOG-TUR NOGOGAN NABCI BURHUJU NIGTA SIGVI SEGUDERLEN HAS-VN SALHIN OROHV BVI-J_A. HALAGVN NARAN NEBTEREHU UGEI DORBEN VLARIL CAG URGULJI CANG TOGTAJV, VSV SIMEDUGSEN TVLA HEREM-UN CILAGV, AGVLA-YIN ARV, MODON-V HALISV, GAJAR-VN HOROSO-DUR COM NOB NOGOGAN JAMAG VRGVJV NENGDE BASA NABCI-VIN GEREL-UN NOGOGAN TVYAG_A DUGURCU, MASI NIGEN HELEJU BOLVSI UGEI SONIN GAYIHALTAI SERCULENG SAYIHAN TVLA BI VRID HURIYELENG-DUR OROHV BURI ONCA TENDE ECIJU NIGEN VDAG_A TOGORIN, OBER_E

FIGURE 1. THE SAMPLE OF CORPUS VLAGAN EGULEN-U NILBVSVN.

## 3. Corpus Processing.

3.1. The Proofreading of the Corpus. In corpus processing, the accuracy of the corpus should be ensured first. In reality, however, since materials of the corpus are usually logged in manually, there unavoidably exists the possibility of misspellings and input errors due to different dialects or even mispronunciation. Therefore, the corpus is proofread with the 1990 edition *VLAGAN EGULEN-U NILBVSVN* and standardized by *Mongolian-Chinese* 

Dictionary<sup>1</sup>.

3.2. The Tagging of Fixed Phrases. On the basis of corpus proofreading, tagging phrases of corpus is accomplished with the Mongolian morphology analyzing software programmed by School of Mongolian studies of Inner Mongolia University. Then, the corpus is revised, based on professor Qinggeltu's *Detailed Interpretation of Modern Mongolian Fixed Phrase Grammatical Information Dictionary*<sup>2</sup>.

3.3. **The tagging of morphology.** Mglex, a software programmed by School of Mongolian studies, is applied to word segmentation and morphology tagging.

The tagging of the morphology is shown below:

文件①	编辑(E)	格式(0)	查看──	帮助(H)					
page1/	Ne1								
VLAGAN OROSI		GULE/Ne2	+N/Zx-I	U/Fc12	NILBV	SV/Ne2+N	I/Zx-V/Fc12	VG/Ne2	
INIYED	UMTU/AS	MVNGH	AG/Ac-	VD/Fp2	VYILA	MATAGAI/	Ac MERGED	/Ne2-	
I/Fc31	DEMEI	Ac BIT	EGEI/S	u ELEG	LE/Ve1	,/Wp1	EMGENIL/N	e2-	
							c TA/Rb+N		
YAGV/R	a+N/Zx+	AI/Fc13	MEDE.	/Ve1+MU	JI/Fs22	./Wp1	MINU/Rb		
BEY E/	Nel NI	LH A/Ac	-YIN/F	c11 UY	E/Ne1	-DUR/Fc2	2 BICIHAN	/Ac	
CECERL	IG/Ne1-	-TUR/Fc2	2 NAG	AD/Ve1+	-V/Zv1+	HVY A/Fr	A2 ./Wp1		
TERE=H	U/T HI	<b>IRIYELEN</b>	G/Ne1-	UN/Fc11	BARA	GVN=HOYI	TV/Yn ONC	OG/Ne1-	
TUR/Fc	22 NOC	GOGAN/Ac	NABC:	I/Ne1	BURHU/	Vel+JU/F	n1 NIGTA/	Ac	
SIGVI /	Ne2 SH	GUDERLE	/Ve1+N	/Fn3 H	AS/Ne2	-VN/Fc11	SALHI/Ne	2+N/Zx	
							NARA/Ne2+		
							CAG/Ne1	364050250500	
						./Wp1			4

FIGURE 2. THE SAMPLE OF MORPHOLOGY TAGGING OF CORPUS *VLAGAN EGULEN-U NILBVSVN*.

Once the corpus processing is completed, the corpus is recorded in Microsoft Excel by word or by phrase. Then the verbs are listed in the verb sub-library.

4. Fields and Field's Value of Verb Sub-Library. *VLAGAN EGULEN-U NILBVSVN* electronic dictionary is one of Injannasi's multi-functional electronic dictionaries. With the requirements of research consistency, the fields of the verb sub-library in *VLAGAN EGULEN-U NILBVSVN* are the same as those in the sub-library of *NIGEN DABHVR ASAR* and *VLAGAN\_A VHILAHV TINGHIM* multi-functional electronic dictionary.

The fields of verb sub-library include Mongolian form(MU), Latin transcription (GU), pronunciation (DDG), Part of speech (UA), syllable (UB), paraphrase (MVT, GVT, HVT), word frequency (IT), word frequency level (DD), morphological form (HH) and example sentence (MJU, GJU) etc. The verb sub-library is shown below:

<sup>&</sup>lt;sup>1</sup> Mongolian Language Research Institute, Inner Mongolia University. 1999. Mongolian-Chinese dictionary (revised edition). Inner Mongolia University Press, Hohhot.

<sup>&</sup>lt;sup>2</sup> De, Qinggeltu. 2005. Detailed Interpretation of Modern Mongolian Fixed Phrase Grammatical Information Dictionary. Inner Mongolia Education Press, Hohhot.

2 2 2 Ariz	\$ · 文作 · 第 3 · 文作 · 第 3 · 20		戦入 (7000 6月 18日 - 184入 - 14日 - 186入 - 146 - 186 - 186	2189 X	☆ 数据 単編 ・ 工具 ・ 数据 - ふ 込 - ノ ワ ・ ○ ・ ● 第項 ・ ●	10日 开放工品 第日 - 和助 -   100 X - ム - 124 34   11 - 14 - 34 - 34 34   11 - 14 - 34 - 34 - 34 - 34 - 34 - 34 -	× 80° ⊒ ⊒• ●• 3•8• <b>2</b> • Δ• 20	J. 0					9 - 7
-	B113	• (*) B	fr 144	D	E.	R	c	в	I	I	ĸ	L	
1	GU	MU	DDG	UA	UB	MVT	GVT	HVT	IT	DD	HH	MJU	GJU
07	ALVSLAHV	Neddone	[alslăx]	Ve1	A.LVS.LA.HV	[1] 3444/8344/9 14449/1444/8344/ 344444/144/9 3444444/144/32	L1 AGVLA VSVN-V CAGAN_A ORDJV HOLA BOLHV, YAGVM_A TASIYALAJV	(及)隔,隅 远,隔开,间 隔;₩~~隔-	2	* <u>àààà</u> à	Fn3=1 Ft11+Fc 41=1	بدر اعذ استنقار یعا اسینتان لریشر ید	CI EDUGE HURIYELEN G-DUR SAGVGAD
	AMARAHV	للجيئة	[amrāx]	Ve2	A.MA.RA.HV		[1] AMVSHIGSEN- ECE BEY_E MAHABVD- VN JUDEREL TAYILVGDAHV :	(不及)①休 息,歌息:₩~ 歌一歌②舒服,	8	<b>*</b> \$\$\$\$\$	Fn1=2 Fn2=1 Fn3=1 Fs15=1	'ಬ್ಲಾ/ ನಿಟ್ಸ್ ' ಬೆಲ್ಲಪ್ 'ಸ್ಟಾ(್ ಪೆ ?	ENDE BAHAN AMARAHV UGEI VV ?
.09	AMIDVRAHV	-	[æmdrāx]	Ve2	A.MI.DV.RA.HV	[1] (4461 + 56491 57) 549968 al (349968 5666661 gal 531 5468 (349966 6 13997991 68	[1] HOMUN, AMIDAN BA VRGVMAL-VN TOROJU AMILAGSAN CAG-ACA	(不及)①生 活,生存,过 活,生息,谋	2	<b>★</b> ☆☆☆☆			CINU BEY_E- YI EL=GE=GE AMIDVRAGV LVN
10	AMISHVHV		[æmsgäx]	Ve2	A.MIS.HV.HV	[1] تينيد/ تم تعيد تين تخطيع المريد/ مراجع تين تعريد/ تعري ترجير تخطيعان تن	<b>L13</b> HOMUN BA AMIDV BVDAS AMISHVL-VN ERHETEN-IYER-IYEN	(不及)呼吸。	4	*****	Fh20=1 Fn1=3	ور بندر رام العند المحر رامه العربين م المحر رامه العربين	UJESE, ENE TU GUNGZI- YIN ECIGE \$AN=JIYUN-V
11	AMSAHV		[amsāx]	Vel	AM.SA.HV	[1]	[1] YOSON-V IDEGEN VMDAGAN-ACA AMA HURHU: CAGAN IDEGE~. [2]	(及)①尝味, 品味②(为了表 示领情)尝一	7	★☆☆☆☆	Fell+Fb 12=1 Fnl=1 Fs22=3	ریش میرد است میتورد معید و میتو	HERBE TEGUN-U HOYITV TORO-DUR

FIGURE 3. THE SAMPLE OF VERB SUB-LIBRARY.

4.1. Entries. The creation of entries is the most essential step for the compilation of a dictionary. The entries in verb sub-libraries are written in the form of morphological ends "HVV/HUU".

4.2. **Mongolian Form of Entries.** In this field, the software Mongolian Editor 4.0 is used to transform the Latin transcription of entries to the Mongolian form of entries.

4.3. **Pronunciation.** Pronunciation is presented in the International Phonetic Alphabet (IPA) system according to the pronunciation in *Mongolian-Chinese Dictionary*.

4.4. **Part of Speech.** According to "information technology - Mongolian word and expression marks for information processing" (GB/T 26235-2010), Mongolian verbs are classified into four types: general verb, pro-verb, link verb and auxiliary verb. After a statistical analysis, the distribution of verbs in *VLAGAN EGULEN-U NILBVSVN* is shown below:

First-class	Second-class	Third-class	Entries' number	Total					
	1 1	Ve1	780						
verb	general verb	Ve2	678						
	pro-verb	Vt	4	1470					
	link verb	Vx	7	1478					
	auxiliary	Vz1	2						
	verb	Vz2	7						

 TABLE 1. STATISTICS OF VERB ENTRIES

As the table shows, there are 1478 entries listed in the sub-library of *VLAGAN EGULEN-U NILBVSVN*. Among them, general verbs account for 98.64% of all verbs, while pro-verbs for 0.27%, link verbs for 0.47% and auxiliary verbs for 0.61%. Further, transitive verbs and intransitive verbs account for 52.77% and 45.87% of all verbs, while first class auxiliary verbs and second-class auxiliary verbs account for 0.14% and 0.47% of

entries respectively. The classified verbs are distributed as follows: transitive verb > intransitive verb > link verb = second class auxiliary verb > pro-verb > first class auxiliary verb.

4.5. **Syllable.** In this field, verb entries are resolved into syllable, which is separated by Full Stop ".". For example, the entry "based" (go) is resolved into "O.CI.HV".

4.6. **Paraphrase.** The paraphrases of each entry are listed in the form of Latin transcription and Mongolian form respectively. Applying the paraphrasing software programmed by BAO Songbo, paraphrases from *Mongolian Language Dictionary* are recorded in this field. Then, the software Mongolian Editor 4.0 is used for the translation of paraphrases from Mongolian form into Latin transcription. For example:

۳۴۰۲۳۴ (take away):

 $\begin{bmatrix} 1 \end{bmatrix}_{\text{regred}, n'} \text{ trade in the last in the last intermed is intermed in the last intermed intermed in the last intermed intermed in the last intermed in$ 

ABACIHV:

【1】YAGVM\_A-YI ORON-ACA NI SILJIGULHU BVYV ABCV ODHV: MORI-BAN ~; NOM-IYAN ~. 【2】HEREG YABVDAL--DV UJELTE-YIN EGUDEN-ECE HANDVHV: MAGV--DV ~. 【3】HEREG YABVDAL--I SIIDHEN SIIDBURILEHU: HAVLI -BAR ~; AGVCILAN ~; DVR\_A JORIG-IYAR-IYAN ~.

Sometimes, however, the paraphrasing software cannot find correct paraphrases for the entries because those paraphrases are not listed in *Mongolian Language Dictionary*. In this case, *H0RIN NAIMATV Dictionary*<sup>3</sup> may be of great help. If there are still no corresponding paraphrases, Consulting the Chinese version of *VLAGAN EGULEN-U NILBVSVN* translated by ZHAO Jingyang, along with *Chinese-Mongolian Dictionary*<sup>4</sup>, satisfactory answers can be available.

4.7. **Chinese Paraphrase.** The Chinese paraphrase is added according to the paraphrases in *Mongolian-Chinese Dictionary*. For example:

horzono (take away) :

(及)拿去,取去,领去,带去:ᠬᠭ~ 带到远处去; ᠳᡂᡣ ~ 叫去(take it: take it far away; call out) ②提到:ᠳᡵᠠ ᠳᡢ ᠳᠬ দলᠠ দলা দলা 提到原则上来分析(mention: analysis in principle) ③处理,处置,处分,处决,论处:ᠳᡂᠠ ~ 宽大处理; ᠳᠬ, ᠳ ~ 给予纪律处分; ㅋᠳ ↔ ~ 依法处决。(deal with, dispose, execute, punish: deal with leniently; disciplinary sanctions; execute according to law.)

<sup>&</sup>lt;sup>3</sup> Namjilma.2013. H0RIN NAIMATV Dictionary.Inner Mongolia Peoples Publishing House, Hohhot.

<sup>&</sup>lt;sup>4</sup> Institute of Mongolian Language and Writing, Inner Mongolia Academy Of Social Science. 2005. Chinese-Mongolian Dictionary (3<sup>rd</sup> Edition). The Ethnic Publishing House. Hohhot.

In case an entry does not have a paraphrase in *Mongolian-Chinese Dictionary*, the version of *VLAGAN EGULEN-U NILBVSVN* translated by ZHAO Jingyang is referred to.

4.8. Word Frequency. In this field, the frequency of occurrence of each entry in corpus *VLAGAN EGULEN-U NILBVSV*N is recorded. For example, if the word "CAYIHV" (whiten) appears 5 times, the frequency of this word is noted "5".

4.9. Word Frequency Level. The occurrences of each entry are graded according to their values. The level of occurrence is represented with a symbol  $\bigstar$ . To be specific, the level 1 noted  $<<\bigstar \bigstar \And \And \And \And \And \diamondsuit \end{Bmatrix}$  is the entry whose occurrence is between 1-10, the level 2 noted  $<<\bigstar \bigstar \And \And \And \diamondsuit \end{Bmatrix}$  is the entry whose occurrence is between 11-50, the level 3 noted  $<<\bigstar \bigstar \bigstar \And \diamondsuit \diamondsuit$  is the entry whose occurrence is between 51-100, the level 4 noted  $<<\bigstar \bigstar \bigstar \bigstar \diamondsuit$  is the entry whose occurrence is between 101-500 and the level 5 noted  $<<\bigstar \bigstar \bigstar \bigstar \diamondsuit$  is the entry whose occurrence is greater than 501. The entries categorized by the level of occurrence in corpus *VLAGAN EGULEN-U NILBVSVN* are shown below:

Word frequency level	Ve1	Ve2	Vt	Vx	Vz1	Vz2	Total
****	611	572	2	5	0	2	1192
★★☆☆☆	125	81	0	0	0	2	208
★★★☆☆	21	16	0	1	0	1	39
<b>★★★</b> ☆	20	7	2	0	1	1	31
****	3	2	0	1	0	1	8

 TABLE 2. STATISTICS OF WORD FREQUENCY LEVEL

4.10. **Morphological Form.** In this field, the morphological forms of entries appearing in corpus *VLAGAN EGULEN-U NILBVSVN* are recorded. The variations in term of voice, aspect, tense, modality, and evidentiality are noted with the morphological forms standardized in "Information technology-Mongolian word and expression marks for information processing".

The format is "morphological form tag = occurrence", and entries are usually transformed by multi morphological suffixes. If an entry is transformed by one morphological suffix, its morphological suffixes should be noted respectively in aforesaid format and separated from the next one by space. If its morphological suffixes are added one after another connectively, it should be noted in the format "morphological form1 + morphological form2 = occurrence". For example: ALVSLAHV (span):Fn3=1 Ft11+Fc41=1. "Fn3=1" means the morphological suffix "N" inflects the entry "ALVSLAHV" and appears one time in the corpus. "Ft11+Fc41=1" means the morphological suffixes "GSAN-ACA" inflects the entry "ALVSLAHV" and appears one time in the corpus.

4.11. **Example Sentence.** In this field, example sentences of the entries appearing in corpus *VLAGAN EGULEN-U NILBVSVN* are listed in the verb sub-library. The example sentences which are written in Latin transcription are replicated from the corpus database directly. Then the example sentences in Mongolian form are transformed by the software Mongolian Editor 4.0.

5. Conclusions. In our study, corpus processing and constructing verb sub-library cost much manpower and time. The corpus processed in our work will contribute to relevant studies such as Corpus linguistics, Natural language processing and Computational linguistics. Some data about the study could be used in the computational linguistic research on Injannasi's work. In further research, a new research method could be studied for reducing labor intensity and saving manpower. Other Injannasi's works apart from *HOHE SVDVR, NIGEN DABHVR ASAR, VLAGAN\_A VHILAHV TINGHIM, VLAGAN EGULEN-U NILBVSVN* can be studied further to complete Injannasi's multi-functional dictionary and then to set up a cloud platform.

## REFERENCES

- [1] X. Bao, *Establishment of a Verb Sub-library in hohsudur Electronic Dictionary*, Un-published MA dissertation, Inner Mongolia University, 2008.
- [2] Dabhvrbayar, Research on Mongolian Phrase Structure Knowledge Base, Liaoning Ethnic Publishing House, Liaoning, 2014.
- [3] DuLan and Y. Hai, A Statistical Study of Type of The Word in D Nachugdorj's Novel Old Boy, Mongolian Studies of China, vol.45, no.3, pp.199-204, 2017.
- [4] Gaowa, Establishment of Proper Noun Sub-Library in *HOHE SVDVR* Electronic Dictionary, Unpublished MA dissertation, Inner Mongolia University, 2008.
- [5] Hasituya, Construction of Common Noun's Section in The Electronic Dictionary of *Khokh Sudar*, Unpublished MA dissertation, Inner Mongolia University, 2008.
- [6] S. Hua, Mongolian Author Special Corpora Building and the Lexicography of A New Dictionary, Research on Information Technology of Ethnic Languages and Characters: Papers Collection of the Eleventh National Symposium on Information of Ethnic Languages and Characters, pp.242-245, 2007.
- [7] G. Li, The Construction of Empty Word's Section in "Electronic Dictionary of *Khokh Sudar*", Unpublished MA dissertation, Inner Mongolia University, 2010.
- [8] Meijuan, Research in Mongolian Lexicography for Information Processing, Unpublished MA dissertation, Inner Mongolia University, 2008.
- [9] Mongolian Language Research Institute, Inner Mongolia University, *Modern Mongolian*, Inner Mongolia Peoples Publishing House, Hohhot, 1964.
- [10] Mongolian orthographic dictionary editorial team, *Orthographic Dictionary*, Inner Mongolia Peoples Publishing House, Hohhot, 1999.
- [11] H. Qian, Corpus Construction and Dictionary Compilation, *Lexicographical Studies*, vol.131 no.1, pp.58-68, 2002.

- [12] Wuyahan, The Chamber of *Red Tears* Corpus Processing and Application, Unpublished MA dissertation, Inner Mongolia University, 2017.
- [13] Xiaojuan, Processing and Application of Corpus of "NIGEN DABHVR ASAR", Unpublished MA dissertation, Inner Mongolia University, 2016.
- [14] Injannasi, VLAGAN EGULEN-U NILBVSVN, Inner Mongolia Peoples Publishing House, Hohhot, 1990.
- [15] J. Zhao, Hong Yun Lei, Inner Mongolia Peoples Publishing House, Hohhot, 2009.
- [16] Y. Zhao. The Construction of The Afjectival Section of the Electronic Dictionary of *Khoh Sodor*, Unpublished MA dissertation, Inner Mongolia University, 2009.
- [17] Y. Zhang, *Computational Lexicography and New Dictionaries*, Shanghai Lexicographical Publishing House, Shanghai, 2004.