PAN LOCALIZATION PROJECT

FREQUENCY ANALYSIS OF MONGOLIAN CORPUS

Phase 2.1

October 24, 2008

CENTER FOR RESEARCH ON LANGUAGE PROCESSING
NATIONAL UNIVERSITY OF MONGOLIA, ULAANBAATAR, MONGOLIA

Table of Contents

List of Figures and Tables

Abstract

This document provides an overview of the frequency analysis of a corpus for Mongolian. The frequency analysis is discussed focusing on the corpus and its text categories. Various statistical information are extracted from the corpus such as average word type frequencies, average word length, distributions of the word lengths, used foreign words etc.

The process of selecting lexicon and POS tagging the lexicon are introduced in this report too. Additional frequency analysis is done on the tag sets. For example, the average count of tags that a word has is about 1.5 times. It also shows that nouns, verbs and adjectives are the main part of the lexicon.

As a result, it seems appropriate to tag all 5 million word corpus with this tag set without any critical problems. The tag set is expandable (in the boundary of tag creation rules and main tag sets) when it is required.

1. Introduction

In this phase of the project, we have created 10k word lexicon from the 5 million word texts, which has been collected from online news, articles, printed newspapers and other literature and law texts, for a Mongolian POS tagger and other uses. Analyzing corpus frequency for each text type and the entire corpus texts, the words in the lexicon are selected as common usage words based on the analysis. And also, we have analyzed the corpus frequency for other useful linguistic information such as foreign word usage, word length features, closed class words, etc.

For tagging the lexicon word entities, we have used the words, which were tagged in the previous phase of the project, at first, and the rest of the word entities have been tagged by hand according to the POS tagset.

2. Related Works

This work is the first attempt to collect large amount of Mongolian texts and we have collected 5 million word texts form online news, articles, printed newspapers and other literature and law texts.

This corpus will be very useful resource for Mongolian language and its usage. Further, we are planning to enrich our corpus with external knowledge such as POS tag, semantic information etc.

3. 5 Million Words Corpus

Five million word corpus is collected form online news, articles, printed newspapers and other literature and law texts. After collecting these texts are cleaned and corrected some errors in them.

There are still some minor errors to correct in the future such as misrecognized and mistyped words, and uncleaned punctuations etc.

4. Word Frequency Analysis

In this section total word frequency analysis on 5 million word corpus is presented. The purpose of the analysis is to derive a word list from the corpus for creating a lexicon for a Mongolian POS tagger that will be developed in the next phase of the project.

The creating word list is based on word frequencies in the corpus. The main word frequency information is shown in Table 1. AntConc[1][], a free concordancing tool, has been used for frequencing the corpus words. Before frequenting the words in the corpus, all files are cleaned and converted .txt and .rtf, Unicode files into UTF8 Unicode text files, since AntConc tool supports UTF8 encoding.

Table 1. Statistics of the Corpus

Text Types	Words	Texts	Words per file	Word Types	Foreign Words	Word Length	Average Word Type Freq
Literature	1,012,779	288	3517	78,972	28	7.9	12.8
Law	577,708	144	4012	15,235	40	8.5	37.9
Publish	2,460,225	2,518	977	118,601	1,153	8.0	20.7
Unen Sonin	949,558	1,260	754	61,125	0	7.8	15.5
Total	**5,000,270**	**4,210**	**1188**	**192,061**	**1,221**	**8.2**	**26.0**

As showin in Table 1, there are total 192,061 word types in the corpus and these words form the entire corpus. Considering that, each word type occurs approximately 26 times. Literature text uses one word type 13 times which is relatively less than others, thus we can say that it has richer vocabulary. On the other hand, Law text uses one word type more often which is 38 times.

Law and Literature type have relatively higher number (average 4000) of words per file, according to other two types Publish and Unen. Among these two types, Literature text has more (5 times) word types than Law text. Since Literature text uses rich vocalbulary, whereas Law text uses limited terms in its definition.

Similarly, for low density texts Publish and Unen, the former has more (3 times) word types than the latter. Actually these two types have similar origin, texts from newspapers, but their time of creation is different. Texts in Unen are from newspapers during 1980-1990, whereas Publish type is collected from modern online newschannels. We suspect that vocabulary used in modern news channels have variety of word types, such as foreign words etc. Also there are many spelling and typing error in these texts and these errors mutiplicate one word type to many word types. But texts from newspapers of 1980s have limited vocabulary size, text is well edited, reviewed, error-corrected.

Total word frequency in the corpus

In the corpus, there are totally 5,000,270 word tokens and they are categorized into 192,061 word types, a word list. The highest frequency 10 thousand words in the word list are selected. Their rankings are above 25 in the list and among those “нь” (personal possessive postposition of the third person) has the highest frequency of 98,319. In Table 4, some of the highest frequency words are shown and 10 thousand words (exactly 10,117) are shown in Appendix A.For the entire corpus, the highly ranked words have frequencies over 50 and they occur about 4,217,571 times (which makes 84% of the corpus).

[1] AntConc 3.2.1w (Windows), concordance tool
Developed by Laurence Anthony

Table 2. Most frequent 50 words

Rank	Word	Freq.	Rank	Word	Freq.	Rank	Word	Freq.	Rank	Word	Freq.	Rank	Word	Freq.
1	нь	98319	11	их	18380	21	би	11635	31	терийн	9236	41	гээд	7669
2	юм	38986	12	хүн	18236	22	болон	11488	32	шиг	9029	42	байдаг	7630
3	байна	36444	13	тэр	17033	23	гэсэн	11164	33	зүйл	8823	43	эрх	7597
4	ч	36332	14	байсан	16406	24	дээр	11050	34	тухай	8722	44	ын	7573
5	энэ	35356	15	аж	15410	25	дээ	10114	35	өөр	8645	45	сарын	7524
6	гэж	35260	16	олон	14905	26	улс	9986	36	болж	8637	46	уу	7444
7	л	28729	17	улсын	13962	27	монгол	9847	37	болсон	8489	47	гэдэг	7371
8	байгаа	25886	18	мөн	13093	28	ажил	9609	38	оны	8386	48	д	7333
9	нэг	21624	19	байх	12407	29	манай	9569	39	ахуйн	7954	49	үйл	7275
10	бол	19021	20	хоёр	12195	30	газар	9374	40	арга	7899	50	ямар	7181

Since there is a space limitation we can't show all the 10,117 words in this report, instead we illustrated these frequencies by a graph in Figure 1.

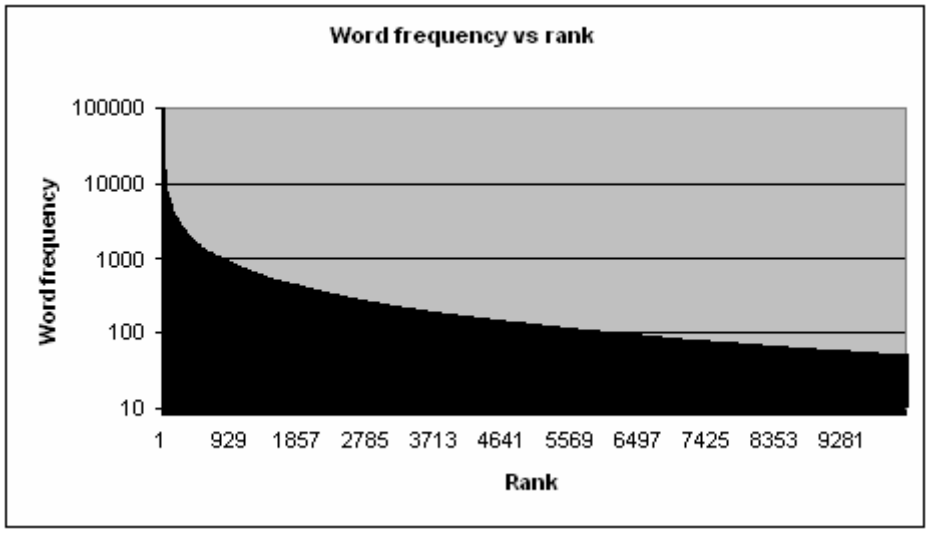

Figure 1. Word frequencies versus their ranks in the corpus (logarithmic scale)

Figure 1shows the word frequencies of the above highly ranked 10,117 words versus their ranks in logarithmic scale. Some other analyses on the word frequency are shown in Table 5.

Table 3. Highly ranked word statistics

		Highly ranked 10000 word selection						
Text Types	**Abbr.**	**Filter**	**Records**	**Sum**	**Percent**	**Common with ALL**	**Common with ALL**	**Average word length**
Literature	LT	>10	9,981	860,265	85%	5,928	59%	6.3
Law	LW	>2	8,284	568,775	98%	4,512	45%	8.1
Publish	PR	>25	9,959	2,087,780	85%	8,226	81%	6.2
Unen Sonin	RTF	>6	10,007	867,321	91%	6,070	60%	7.3
Total	**ALL**	>50	10,117	4,217,571	84%	10,117	100%	6.5

The most of the words forming highly ranked 10,000 words are from the Publish type texts and this is because it forms the most of the corpus. Table 3 shows statistics about highly ranked 10,000 words for the entire corpus and for each text types. The column "Common with ALL" shows the percentage of the common words between the main list of highly ranked 100,000

words and a list for the corresponding text types, i.e. most frequent 10,007 words of the Publish type text forms 81% of the total 10,117 most frequent words of the entire corpus.

The column “Percent” shows the ratio of the highly ranked words’ occurrences to the entire corpus. For example, highly ranked 10,117 words occur 4,217,571 times in the corpus and it makes 84% of the corpus. If we can manually tag (POS tag) these highly ranked 10,117 words, it means that we can get the 84% of the tagged corpus.

Highly ranked words from law, literature and press texts

In Table 4, the most frequent words of each text type are shown. In literature texts, most frequent words are mainly particle words such as “нь”, “ч”, “л” and pronouns such as “энэ” (this), “тэр” (that), “би” (I) etc (closed class words). In law texts, frequent words are mainly juridical terms, date and ordinals. In press texts “бай-” (is, exist, stay) words have high frequencies, which reflect daily conversation pattern.

Table 4. Most frequent 20 words of each text type

	Literature		Law		Publish		Unen	
Rank	Word	Freq.	Word	Freq.	Word	Freq.	Word	Freq.
1	нь	25954	энэ	7506	нь	52919	нь	13213
2	гэж	15392	улсын	6727	юм	23018	байна	9786
3	юм	12439	дугаар	6307	ч	21618	аж	9064
4	ч	11391	нь	6233	байна	20437	ахуйн	5486
5	л	8608	зүйл	5626	энэ	19802	олон	4626
6	нэг	6699	болон	4952	л	18865	намын	4539
7	тэр	6688	хуулийн	4776	гэж	17065	байгаа	4433
8	энэ	6156	хууль	3974	байгаа	15448	арга	4268
9	би	6097	сарын	3818	байсан	12444	улс	4232
10	хүн	5954	дүгээр	3642	нэг	11675	үр	4186
11	байна	5169	заасан	3544	бол	10839	ажил	4184
12	хоёр	4966	төрийн	3510	тэр	9636	их	4124
13	шиг	4903	оны	3496	их	9529	юм	3529
14	байгаа	4042	хуулиар	3363	хүн	9404	ажлын	3213
15	бол	3940	эрх	3319	гэсэн	8206	ч	3192
16	дээ	3823	тухай	3143	олон	7655	манай	3095
17	минь	3438	байгууллага	3048	байх	7389	зохион	2975
18	дээр	3399	бол	3046	мөн	6848	ын	2941
19	чинь	3390	бусад	2997	дээ	6270	шинэ	2778
20	юу	3303	өдрийн	2963	монгол	6241	чухал	2702

There are about 97,145 words occurring only once in the entire corpus which makes 51% of the total word types and only 2% of the entire corpus. Among these, 40028 are in Literature, 4,969 are in Law, 54,752 are in Publish, and 35,941 are in Unen type. In the list of less frequent words, there are some words which are misrecognized, and mistyped. Beside this there are also some really rare words such as “аавынхантай” which means ‘with my fathers home’. Some of the less frequent word list is shown in Table 7 as an example.

Table 5. Less frequent words

Less frequent words		Literature		Law		Publish		Unen	
Word	Freq.	Word	Freq.	Word	Freq.	Word	Freq.	Word	Freq.
ахуйн	1	аавархуу	1	аваагүйг	1	данзангаас	1	бүтнээр	1
аавархуу	1	аавынхантай	1	авааариас	1	данида	1	бүтоэцнйп	1
аавынхантай	1	аагаараа	1	авааарийг	1	данид	1	бүтцнйг	1
аага	1	аагаараан	1	авааарь	1	данигайн	1	бүтцээрээ	1
аагаараа	1	аагархан	1	авагдаагүй	1	данзанчойдор	1	бүтцээс	1
аагаараан	1	ааггай	1	авагдсаны	1	данзанравжааг	1	бүтээмж	1
аагам	1	аагиж	1	авагчаар	1	данзандаржаагийн	1	бүтдэгийг	1
ааганд	1	аагий	1	авагчдын	1	дандовж	1	бүтээгдэхүүкий	1
аагархан	1	ааглаад	1	авагчтай	1	данзангийнх	1	бүтаэцийг	1
аагархах	1	ааглаж	1	авалт	1	данни	1	бүтвэг	1
аагархуу	1	аагчин	1	авалтыг	1	данзангаа	1	бүтвэгдэхүүннйг	1
ааггай	1	аадраас	1	авалцан	1	данжуурын	1	бүтвэя	1
аагиар	1	аадрас	1	авахуулахгүй	1	данжуураас	1	бүтднйг	1
аагиж	1	аажаагийн	1	авахуулж	1	данжанрав	1	бүтээгдэ	1
аагилан	1	аажий	1	авианы	1	данжаадууд	1	бүтдэггүй	1
аагилд	1	аажууруулавч	1	авилгатай	1	данжаад	1	бүтэамжнйг	1
аагимхан	1	аажууханаар	1	авлагагүй	1	данхалзац	1	бүтээгдэхүүн	1
аагинд	1	аалгүйтэж	1	авлагатай	1	данзанд	1	бүтээгдэхүүнцөцгийн	1

Word length and foreign words analyses

The average word length of the words in this corpus is 8.2. Following 2 shows the distribution of the words versus their length. The longest meaningful word in this corpus is “цахилгаанжуулахгүйгээс” (because of de-electrifying), which has 22 characters.

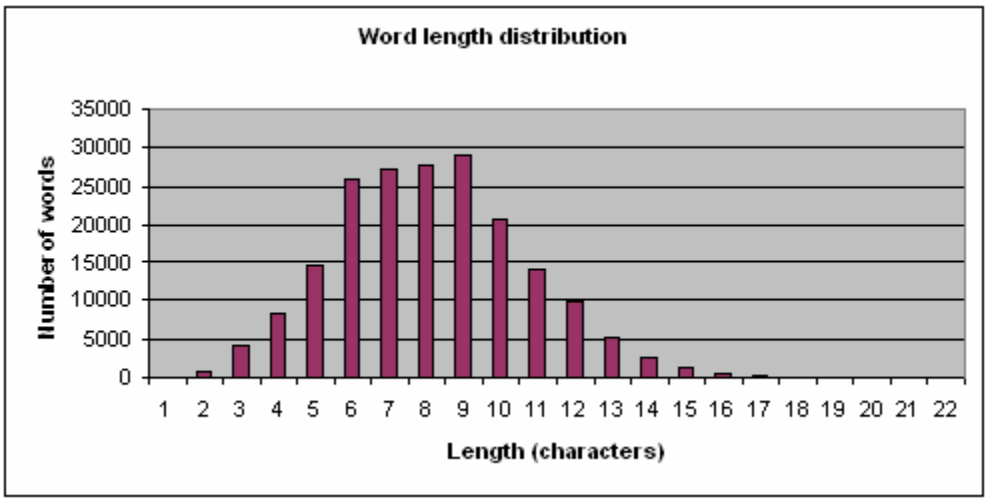

Figure 2. Word lengths in the corpus

There are 921 two character words and the shortest meaningful words are shown in the following Table 6.

Table 6. Two character words

	Word	Freq.		Word	Freq.		Word	Freq.		Word	Freq.
1	юм	38986	11	уг	3561	21	ус	2246	31	ил	1087
2	их	18380	12	ер	3500	22	ач	2015	32	ан	1028
3	аж	15410	13	чи	3430	23	ул	2014	33	ар	1012
4	би	11634	14	бэ	2973	24	ах	2012	34	ур	895
5	юу	6855	15	эд	2890	25	ёс	1701	35	эм	894
6	ур	6848	16	за	2788	26	яг	1679	36	ой	868
7	та	5871	17	ба	2593	27	га	1652	37	эс	785
8	эх	5088	18	эр	2468	28	ял	1213	38	нө	577
9	вэ	5004	19	ам	2453	29	ор	1128	39	ид	567
10	уг	3961	20	уе	2296	30	он	1122	40	ул	544

About text quality and Assurance methods

Texts in this corpus have some sort of errors such as mistyped words, misrecognized words (OCR) and words with different codes within a text. We corrected some possible errors in the text using special purpose tools developed by some of our project members.

Beside this, there are some minor errors like multi words connected without any delimiter (“байгальертөнцийнхамгийнухаантайамьтан”, the most clever animal in the nature), mixed code texts (ASCII codes in utf, text files).

Considering that the 84% of the entire corpus consists of the 10 thousand words, the word selection for the lexicon is well done.

About foreign words

In last decades Mongolian language has imported many foreign words without translation and used them within Mongolian texts. As an example, the following Table 7 shows the statistics of the foreign words appeared in the corpus.

Table 7. Statistics of foreign words

Text Types	Foreign Words	Some Example Words
Literature	28	holiday, beautiful, nescafe
Law	40	tatarica, saiga, lutra
Publish	1,153	bbc, mn, digest,bank, bloomberg
Unen Sonin	0	
Total	**1,221**	

Note that, foreign words column in the above table shows the number of words written only in Latin, and foreign words that are written in Cyrillic are not considered. Because counting foreign words in Cyrillic is a manual and time consuming work. These foreign words are not used in the corpus intentionally, they appear accidentally.

Conclusion

In this part, total word frequency analysis on 5 million word corpus is presented. The corpus is composed of 4 different types of texts such as Literature, Law, Publish and Unen. These types differ from each other on the source of their origin.

Each text type has its own characteristics, such as Literature texts have more rich vocabulary, whereas Law texts have limited number of terms. Law text has longer words (average word length is 8.5) and used one word approximately 38 times. On the other hand, in Literature text one word is used approximately 13 times, which is relatively less than other types. In the entire corpus one word type is used 26 times.

The average word length in the corpus is 8.2, and the longest meaningful word found in the corpus has 22 characters (“цахилгаанжуулахгүйгээс” because of de-electrifying). There are also plenty of 2 characters found.

For foreign word analysis, we considered only words that are in Latin. There is no such word in the Unen texts, which are the newspaper texts in the period of 1980-1990. But the most of the foreign words were found in Publish type texts, which are collected from modern internet news and articles.

As a result, the most frequent 10k words are obtained for each text types and also for entire corpus. The extracted highly ranked words makes 84% of the entire corpus. This means that if we can POS tag these words, we can get 84% of the tagged corpus.

Beside these, there are some minor errors originated from different encodings in the corpus, misrecognition in the OCR tools, and mistyping etc.

5. Lexicon

From Mongolian 5 million word corpus, approximately we have extracted the most frequent 10k words and selected them as a lexicon. After that, all possible POS tags (we call it “tag” hereafter) are attached to each word.

In order to attach tags, all the known word tags in the tagged corpus (from the Project Phase I) are used. This word-tag repository has about 1.5k word-tag pairs and covers about 42% (4,277 distinct words) of the entire lexicon.

The words in both word-tag repository and lexicon are identified and were attached to their tags. The remaining words in the lexicon are manually tagged. Some words’ tags are corrected. Thus all the words in the lexicon have corresponding tag(s).

Analysis on word-tag repository

In the word-tag repository, the most tag-rich word has 9 tags (word “ажил”, “work” in English), and the average tag count that a word has is 1.5. Following Figure 5 shows the distribution of the tag (for example, there are only 3,031 words which have only one tag).

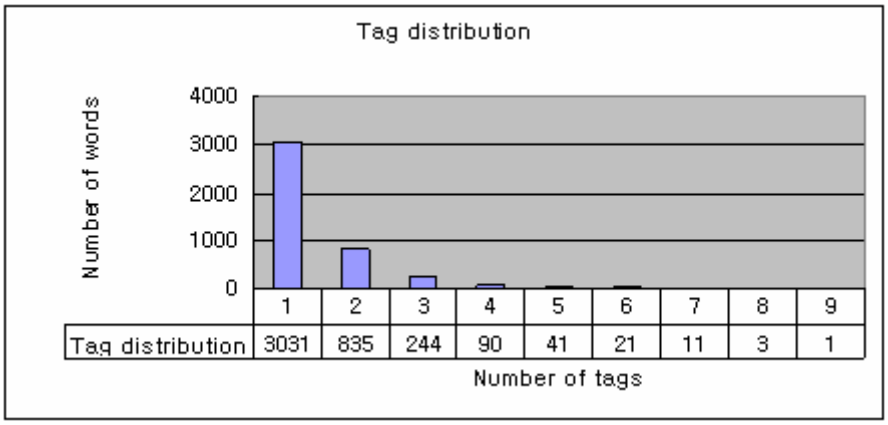

Figure 3. Word tags versus their counts in the corpus

There are total 250 distinct tags, among those tag “N” (noun) has the highest frequency of 2,744 and the second highest frequency tag is “JJ” (adjective) which is 2,375. See Table 8 for more detail.

Table 8. Tag frequencies in the repository

No	Tag	#Word	No	Tag	#Word	No	Tag	#Word	No	Tag	#Word	No	Tag	#Word
1	n	2744	51	abrg	22	101	pnc	5	151	rnc	2	201	vpdc	1
2	jj	2375	52	nis	21	102	pti	5	152	l	2	202	auxdb	1
3	v	1519	53	vpbn	20	103	vabm	5	153	mdabi	2	203	auxdi	1
4	ng	944	54	vpbx	19	104	va1	5	154	mdapx	2	204	pnbs	1
5	nc	880	55	vadb	15	105	npgs	5	155	pnpm	2	205	np	1
6	nn	521	56	pr	15	106	vabxi	5	156	mdafx	2	206	qn	1
7	vac	469	57	vadn	15	107	vps	5	157	vapi	2	207	npghb	1
8	rb	378	58	vabi	14	108	vapn	5	158	nghc	2	208	npghl	1
9	nl	331	59	vabls	13	109	ptb	5	159	pjr	2	209	pvadis	1
10	vpc	241	60	npm	12	110	pni	4	160	vppxb	2	210	pvadi	1
11	ni	236	61	vpbi	12	111	vpfl	4	161	auxb	2	211	npls	1
12	vab	223	62	vadg	11	112	pnpc	4	162	pngs	2	212	pvabls	1
13	vad	196	63	npi	11	113	auxf	4	163	ptn	2	213	pun	1
14	cd	174	64	nbs	10	114	cdn	4	164	rn	2	214	vapls	1
15	vpb	165	65	vadc	10	115	vpdg	4	165	pts	2	215	ptm	1
16	nb	151	66	ptl	10	116	cdb	4	166	vppb	2	216	vaas	1
17	vag	119	67	prb	10	117	npcs	4	167	pvac	2	217	ptls	1
18	vpd	95	68	vpbls	10	118	vpbb	4	168	csb	1	218	ptis	1
19	ncs	93	69	vppx	10	119	vpbxi	4	169	auxdc	1	219	ptcs	1
20	vap	76	70	pnn	9	120	vapl	4	170	vpbxl	1	220	pnghn	1
21	npg	75	71	pnl	9	121	auxpc	3	171	cdis	1	221	ptbs	1
22	pt	73	72	vabb	9	122	auxpg	3	172	cdm	1	222	vpsi	1
23	rnn	67	73	jjr	9	123	pncs	3	173	auxbl	1	223	pnls	1
24	npn	67	74	ca	9	124	auxs	3	174	auxpl	1	224	pnm	1
25	vaf	64	75	npb	9	125	cdg	3	175	vppcs	1	225	pnp	1
26	nm	62	76	rnl	9	126	cdi	3	176	vpbix	1	226	ptpg	1
27	rng	58	77	ptg	9	127	mdab	3	177	cdgs	1	227	mdap	1
28	vpg	56	78	ptc	8	128	mdabx	3	178	csc	1	228	vp2	1
29	cc	53	79	auxp	8	129	pnpb	3	179	cscs	1	229	vasi	1
30	vpbl	52	80	rba	8	130	pc	3	180	auxbc	1	230	jjs	1
31	vabl	49	81	vapxb	8	131	vapg	3	181	auxpghn	1	231	m	1
32	npc	47	82	png	7	132	vpdn	3	182	vpfg	1	232	vapxi	1
33	ngs	46	83	pnb	7	133	vadm	3	183	vppxg	1	233	vapxc	1
34	vpf	46	84	rbr	7	134	vpdi	3	184	csn	1	234	mdabg	1
35	nls	46	85	nghn	7	135	vafl	3	185	auxbi	1	235	mdabxb	1
36	cs	41	86	vafx	7	136	vap1	3	186	auxpb	1	236	pnpbs	1
37	vpbg	36	87	auxd	6	137	vadcs	3	187	c	1	237	vaag	1
38	vpp	34	88	vpdb	6	138	vppc	3	188	auxg	1	238	mdadx	1
39	vabx	31	89	auxc	6	139	vp1	3	189	vpdl	1	239	vaap	1
40	abr	30	90	nghb	6	140	pnpl	3	190	mdabxl	1	240	mdapxb	1
41	npl	29	91	pn	6	141	csi	2	191	vpdis	1	241	vafmd	1
42	vas	29	92	pnpg	6	142	va3	2	192	cci	1	242	vads	1
43	vpbc	28	93	abrn	6	143	nms	2	193	auxbn	1	243	ng -_pun	1
44	vabg	28	94	vapc	6	144	auxpn	2	194	vpdcs	1	244	vadmd	1
45	pj	27	95	vapb	6	145	vabcs	2	195	vpp1	1	245	nghbs	1
46	vabc	27	96	pnpn	6	146	intj	2	196	cdcs	1	246	nghg	1
47	vabn	24	97	vadi	6	147	mdas	2	197	auxdg	1	247	vadls	1
48	jja	24	98	vpfn	6	148	csl	2	198	auxpbg	1	248	vabxc	1
49	vapx	23	99	mdaf	6	149	vpbm	2	199	vpfx	1	249	vpbcs	1
50	md	23	100	g	5	150	vafn	2	200	b	1	250	mdadi	1

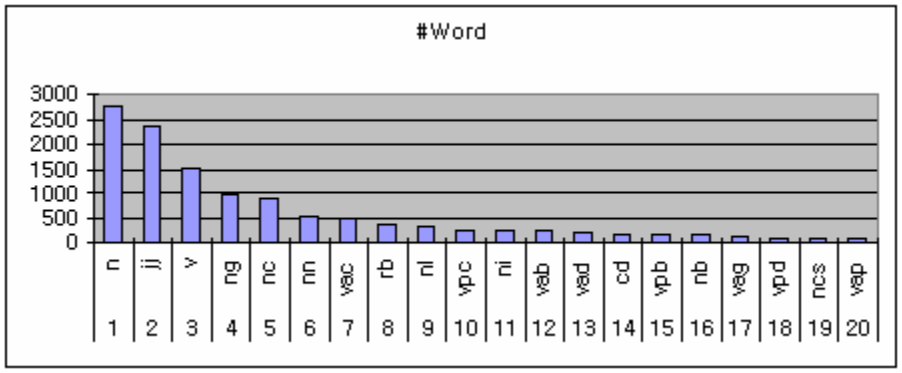

Figure 4. Tags versus their frequency

Analysis on the lexicon

The tagged lexicon contains 15,134 distinct word-tag pairs and 10k distinct words. The most tag-rich 2 words have 8 tags (word “мал”, “улс”: “animal”, “state/country” in English), and the average tag count that a word has is 1.6. Following Figure 5 shows the distribution of the tag (for example, there are only 5,993 words which have only one tag).

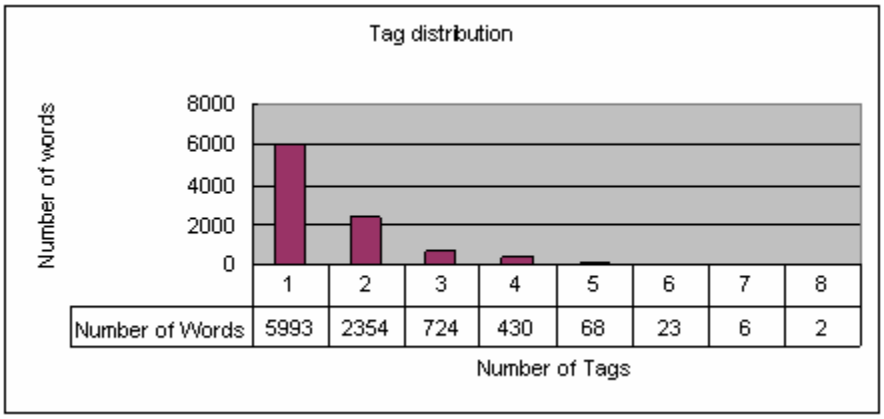

Figure 5. Word tags versus their counts in the lexicon

There are total 359 distinct tags, among those tags “JJ” (adjective) has the highest frequency of 1,524 and the second highest frequency tag is “N” (noun) which is 1,431. See Table 9 for more detail. Single tag words are 62% of the entire lexicon.

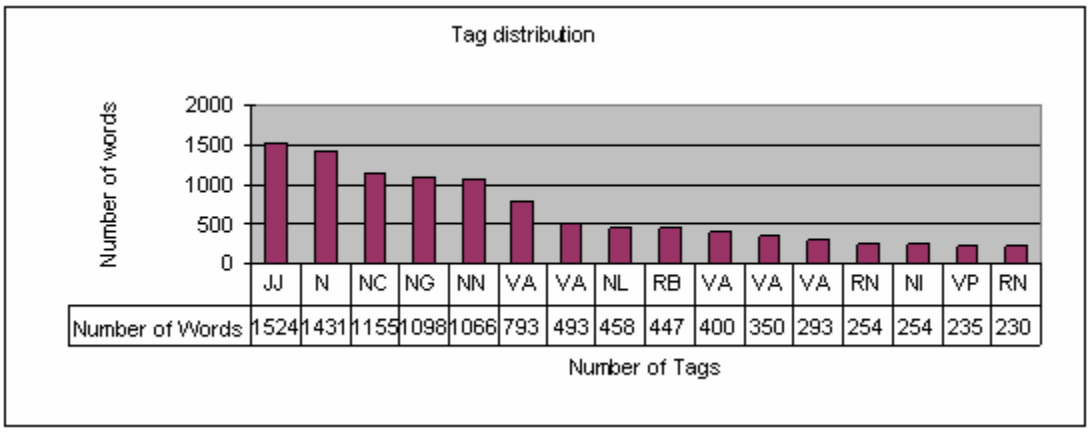

Figure 6. Tags versus their frequency in the lexicon

The count of tags that occurred only once is 118, which makes 2/3 of the total distinct tags. There are total 4,653 nouns (words with tag like “N*”) which is almost the half (48%) of the lexicon and 104 distinct abbreviations which makes 1% of the lexicon. 16% (1,581 words) of the lexicon is tagged as adjectives, which is relatively lower than nouns.

Table 9. Tag frequencies in the lexicon

No	Tag	#Words
1	JJ	1525
2	N	1433
3	NC	1155
4	NG	1098
5	NN	1068
6	VAC	793
7	VAG	493
8	NL	458
9	RB	447
10	VAD	400
11	VAB	350
12	VAF	293
13	RNN	255
14	NI	254
15	VPC	235
16	RN	231
17	VPB	190
18	NCS	179
19	NM	169
20	VABN	161
21	NB	154
22	VADN	151
23	VPF	149
24	VAFN	138
25	VPD	131
26	VAP	123
27	NPN	119
28	VPG	118
29	VA2	98
30	ABR	92
31	NP	91
32	NLS	85
33	VPBN	80
34	NPG	76
35	PT	74
36	RNG	69
37	VPFN	63
38	VPDN	56
39	CC	54
40	VABL	52
41	ABRN	50
42	VAS	50
43	VABX	48
44	CS	46
45	CD	44
46	INTJ	43
47	JJA	40
48	RNC	40
49	VAFX	40
50	VAPX	40
51	MD	39
52	NGS	39
53	VPP	36
54	JJR	35
55	PJ	35
56	VPBL	34
57	ABRG	33
58	NPL	30
59	VAPN	29
60	RBR	27
61	VAFL	27
62	NIS	25
63	RNL	24
64	NPC	23
65	PRB	23
66	VABC	21
67	VABI	19
68	PNL	17
69	VABG	17
70	VA3	16
71	VABLS	16
72	NBS	15
73	PR	14
74	VA1	14
75	PNG	13
76	PTL	13
77	VADG	13
78	PNB	12
79	PNC	12
80	PNN	12
81	PTG	12
82	RBA	12
83	VABB	12
84	VAFI	12
85	PN	11
86	VAAD	11
87	VABXN	11
88	VADB	11
89	VADC	11
90	VAFC	11
91	VAFG	11
92	VAFXN	11
93	PVAS	10
94	VADCS	10
95	VAFLS	10
96	VAPXN	10
97	VPBC	10
98	VPBG	10
99	VPFL	10
100	CA	9
101	NGHN	9
102	VADX	9
103	PC	8
104	PNLS	8
105	PNPC	8
106	PNPG	8
107	PNPL	8
108	PVAD	8
109	V	8
110	VAFB	8
111	VP2	8
112	PTC	7
113	VADI	7
114	VPBX	7
115	VPFX	7
116	ARB	6
117	G	6
118	NGH	6
119	PNM	6
120	PNPN	6
121	PTB	6
122	PTI	6
123	PVAC	6
124	PVAF	6
125	VAADN	6
126	VADL	6
127	VPFC	6
128	AUXD	5
129	CDB	5
130	CDN	5
131	NPB	5
132	PNCS	5
133	PVAG	5
134	QW	5
135	RNM	5
136	VAAC	5
137	VPFG	5
138	ARBN	4
139	CDC	4
140	CDCS	4
141	CDG	4
142	CDL	4
143	CDM	4
144	NPCS	4
145	NPI	4
146	NPLS	4
147	PNI	4
148	PNPM	4
149	PTCS	4
150	PTLS	4
151	PVAB	4
152	VAA2	4
153	VAPC	4
154	VAPG	4
155	VFG	4
156	VPBI	4
157	AUXB	3
158	AUXBL	3
159	AUXDB	3
160	AUXF	3
161	AUXP	3
162	AUXPC	3
163	B	3
164	CDI	3
165	MDAD	3
166	MDAF	3
167	NGHG	3
168	NPM	3
169	PJR	3
170	PNGS	3
171	PNPB	3
172	PNPHN	3
173	PNPIS	3
174	PTM	3
175	PVABN	3
176	PVAFN	3
177	PVAP	3
178	VAAF	3
179	VAAG	3
180	VAAP	3
181	VAPB	3
182	VAPCS	3
183	VPBLS	3
184	VPBXN	3
185	VPFXN	3
186	VPPX	3
187	VPS	3
188	AUX3	2
189	AUXBB	2
190	AUXBC	2
191	AUXBG	2
192	AUXBI	2
193	AUXC	2
194	AUXDC	2
195	AUXDCS	2
196	AUXDL	2
197	AUXDX	2
198	AUXFC	2
199	AUXFG	2
200	AUXFX	2
201	AUXPB	2
202	AUXPCS	2
203	C	2
204	CSC	2
205	CSI	2
206	GH	2
207	MDAB	2
208	MDABX	2
209	MDAPX	2
210	MDAS	2
211	NGHB	2
212	NMS	2
213	NPD	2
214	PCN	2
215	PNGH	2
216	PNGHN	2
217	PRA	2
218	PRI	2
219	PTN	2
220	PTS	2
221	PVA1	2
222	PVA2	2
223	PVABLS	2
224	PVADN	2
225	RBG	2
226	RBN	2
227	RNB	2
228	RNI	2
229	RNLS	2
230	VAAS	2
231	VABCS	2
232	VACS	2
233	VADLS	2
234	VAFCS	2
235	VAPL	2
236	VAPM	2
237	VCS	2
238	VP1	2
239	VPDG	2
240	VPFI	2
241	VPPN	2
242	-?	1
243	ABRC	1
244	ABRL	1
245	ARBG	1
246	AUXBCS	1
247	AUXBIS	1
248	AUXBLS	1
249	AUXBX	1
250	AUXDG	1
251	AUXDI	1
252	AUXDLS	1
253	AUXFB	1
254	AUXFCS	1
255	AUXFI	1
256	AUXFIS	1
257	AUXFL	1
258	AUXG	1
259	AUXPCI	1
260	AUXPG	1
261	AUXPIS	1
262	AUXPX	1
263	AUXS	1
264	AUXSS	1
265	CCI	1
266	CDLS	1
267	CSB	1
268	CSCS	1
269	CSG	1
270	CSL	1
271	CSM	1
272	CSN	1
273	GHN	1
274	GHS	1
275	GS	1
276	I	1
277	JJS	1
278	JRB	1
279	L	1
280	M	1
281	MDABCS	1
282	MDABI	1
283	MDABXB	1
284	MDABXL	1
285	MDAC	1
286	MDADN	1
287	MDADX	1
288	MDAFCS	1
289	MDAFX	1
290	MDAG	1
291	MDAP	1
292	MDAPXB	1
293	NCPS	1
294	NPBI	1
295	NPF	1
296	NPGS	1
297	NPXI	1
298	NS	1
299	PCC	1
300	PNBS	1
301	PNGIS	1
302	PNIS	1
303	PNP	1
304	PNPCS	1
305	PNPGS	1
306	PNPHG	1
307	PNPHI	1
308	PNPLS	1
309	PNPNB	1
310	PRG	1
311	PTGS	1
312	PTIS	1
313	PTMS	1
314	PVA	1
315	PVABI	1
316	PVABIS	1
317	PVABX	1
318	PVABXN	1
319	PVADCS	1
320	PVADI	1
321	PVAFI	1
322	PVAFLS	1
323	PVAFX	1
324	PVAFXN	1
325	PVAPC	1
326	PVAPN	1
327	PVASA	1
328	RBRS	1
329	RG	1
330	RNCS	1
331	RNGS	1
332	VAAB	1
333	VAAFN	1
334	VABBS	1
335	VABIS	1
336	VABLD	1
337	VABXI	1
338	VADS	1
339	VADXN	1
340	VAFBS	1
341	VAFIS	1
342	VAP2	1
343	VAPCI	1
344	VAPI	1
345	VAPIS	1
346	VAPLS	1
347	VASS	1
348	VAXF	1
349	VFC	1
350	VP	1
351	VPBIX	1
352	VPBM	1
353	VPBPG	1
354	VPDI	1
355	VPDM	1
356	VPFIX	1
357	VPFLS	1
358	VPFM	1
359	VPGN	1

Verbs constitutes 3,005 (31%) of the entire lexicon. See following Table 10 for more detail.

Table 10. Percentages of the some classes in the lexicon

Class	#Words	%Lexicon
Nouns	4,653	48%
Verbs	3,005	31%
Adjectives	1,581	16%
Abbr.	104	1%

Although “JJ”, adjective tag, shows the highest frequency, other types of this tag are not much. On the other hand, “N” tag has many other variations (it means inflected forms) like “NN”, “NG” etc. and these forms sum up to almost half of the lexicon.

The average length of the nouns in the lexicon is 6.7characters, 6.9 characters for verbs, 5.9 characters for adjectives and 2.6 characters for abbreviations. Obviously, abbreviations are shorter because of their nature of creation. Adjectives are shorter than verbs and nouns, because they are less inflectional. Also verbs are longer than nouns, for which we can say that verbs are more inflectional than nouns.

Additionally, tag set used in Phase I (about 2 hundred) is less than the tag set in the lexicon (about 3 hundred). This is mainly because of verbs in the lexicon, which have more inflectional forms, and we have added additional tags to the tag set for them. But these additional tags are created in the boundary of tag creation rules and these are possible tags. Newly created tags are composed of tags in the original tag sets, they just combined.

As a result, it seems appropriate to tag all 5 million word corpus with this tag set without any critical problems. The tag set is expandable (in the boundary of tag creation rules and main tag sets) when it is required.

Conclusion

The process of selecting lexicon and POS tagging the lexicon are introduced in this part of the report. Additional frequency analysis is done on the tag sets. The average count of tags that a word has is about 1.5 times. It shows that nouns, verbs and adjectives are the main part of the lexicon.

There are total 4,653 nouns which is almost the half (48%) of the lexicon and verbs constitutes 3,005 (31%) of the entire lexicon. The average length of the nouns in the lexicon is 6.7characters, 6.9 characters for verbs, 5.9 characters for adjectives and 2.6 characters for abbreviations.

Additionally, tag set used in Phase I (about 2 hundred) is less than the tag set in the lexicon (about 3 hundred). This is mainly because of verbs in the lexicon, which have more inflectional forms, and we have added additional tags to the tag set for them. But these additional tags are created in the boundary of tag creation rules and these are possible tags. Newly created tags are composed of tags in the original tag sets, they just combined.

As a result, it seems appropriate to tag all 5 million word corpus with this tag set without any critical problems. The tag set is expandable (in the boundary of tag creation rules and main tag sets) when it is required.

6. Conclusion

In this report, total word frequency analysis on 5 million word corpus is presented. The corpus is composed of 4 different types of texts such as Literature, Law, Publish and Unen. These types differ from each other on the source of their origin.

Each text type has its own characteristics, such as Literature texts have more rich vocabulary, whereas Law texts have limited number of terms. Law text has longer words (average word length is 8.5) and used one word approximately 38 times. On the other hand, in Literature text one word is used approximately 13 times, which is relatively less than other types. In the entire corpus one word type is used 26 times.

The average word length in the corpus is 8.2, and the longest meaningful word found in the corpus has 22 characters ("цахилгаанжуулахгүйгээс" because of de-electrifying). There are also plenty of 2 characters found.

For foreign word analysis, we considered only words that are in Latin. There is no such word in the Unen texts, which are the newspaper texts in the period of 1980-1990. But the most of the foreign words were found in Publish type texts, which are collected from modern internet news and articles.

As a result, the most frequent 10k words are obtained for each text types and also for entire corpus. The extracted highly ranked words makes 84% of the entire corpus. This means that if we can POS tag these words, we can get 84% of the tagged corpus.

Besides these, there are some minor errors originated from different encodings in the corpus, misrecognition in the OCR tools, and mistyping etc.

The process of selecting lexicon and POS tagging the lexicon are introduced in this part of the report. Additional frequency analysis is done on the tag sets. The average count of tags that a word has is about 1.5 times. It shows that nouns, verbs and adjectives are the main part of the lexicon.

There are total 4,653 nouns which is almost the half (48%) of the lexicon and verbs constitutes 3,005 (31%) of the entire lexicon. The average length of the nouns in the lexicon is 6.7characters, 6.9 characters for verbs, 5.9 characters for adjectives and 2.6 characters for abbreviations.

Additionally, tag set used in Phase I (about 2 hundred) is less than the tag set in the lexicon (about 3 hundred). This is mainly because of verbs in the lexicon, which have more inflectional forms, and we have added additional tags to the tag set for them. But these additional tags are created in the boundary of tag creation rules and these are possible tags. Newly created tags are composed of tags in the original tag sets, they just combined.

As a result, it seems appropriate to tag all 5 million word corpus with this tag set without any critical problems. The tag set is expandable (in the boundary of tag creation rules and main tag sets) when it is required.

Appendix A: 10 thousand word list

See the attached file named ”*mongolian_10k_lexicon.txt*”.

Reference

[1] Unen press

http://www.unen.mn/

[2] PennTreebank:

Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision, 2nd printing)
Beatrice Santorini, June 1990

Part of Speech Tagging Guidelines for the Penn Korean Treebank

Chung-hye Han and Na-Rae Han, 2001

The Part-Of-Speech Tagging Guidelines for the Penn Chinese Treebank (3.0)

Fei Xia, October 17, 2007

[3] Statistical Natural Language Processing:

Chris Manning and Hinrich Schütze, Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA: May 1999