

MONGOLIAN LEXICON DERIVED FROM CORPUS

J.Purev and Ch. Odbayar

CRLP

Center for Research on Language Processing
National University of Mongolia

(NUM)

OUTLINE

- Introduction to NUMTeam project
- Word Frequency Analysis: 10k words
- Lexicon
- Analysis on Lexicon
- Conclusion

NUM Team

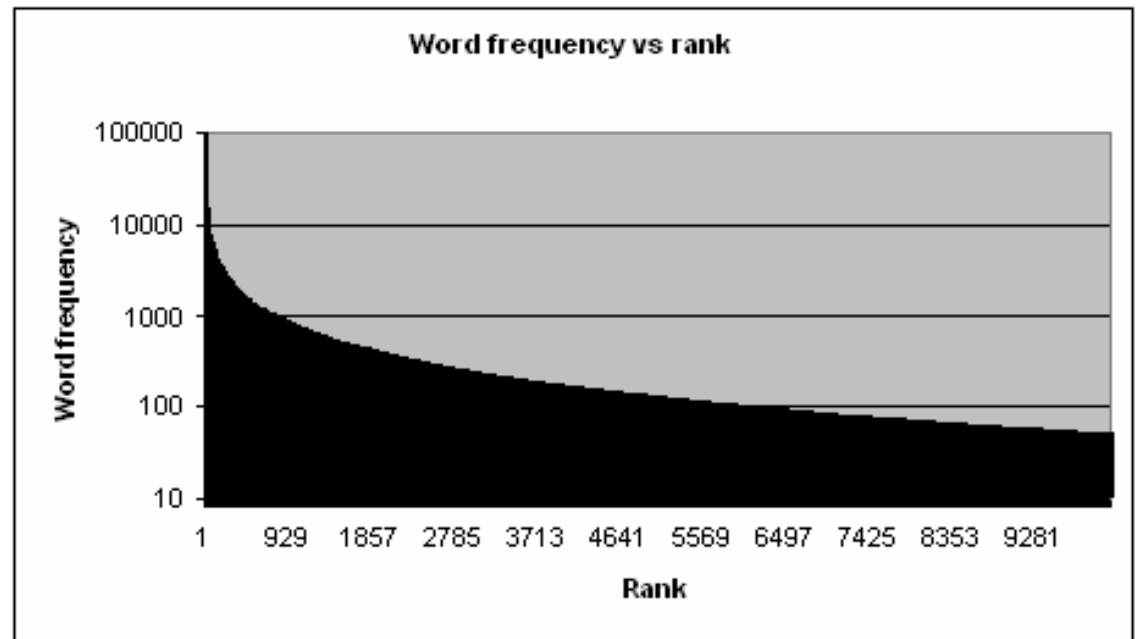
- NUMTeam, Mongolia
 - At Center for Research on Language Processing (CRLP)
 - First center in Mongolia
 - Established in May 2007
 - Supported by National University of Mongolia and PANLocalization
 - Currently 7 researchers and staffs including computer scientists and linguists
 - Goal: *Mongolian Localization and Processing*
 - Our team joint in PANL10n Project in 2006
- Project Objectives:
 - 5 million tagged words corpus
 - Cleaning Tools and Spell Checker
 - POS Tagset
 - POS Tagger

LEXICON CREATION

- After collecting the corpus and manually tagging 100k words
- Created a Mongolian lexicon for automatically tagging the corpus
 - Highly frequented words list from the corpus
 - Tag these words by using the manually tagged part of the corpus

WORD FREQUENCY

- Around 160 thousand distinct words in the corpus
 - *About the corpus has been presented in the previous day!*
- Highest frequency word
 - Almost 100 thousand occurrence
 - ‘нь’ possessive particle



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	төрийн	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

MOST FREQUENT 50 WORDS

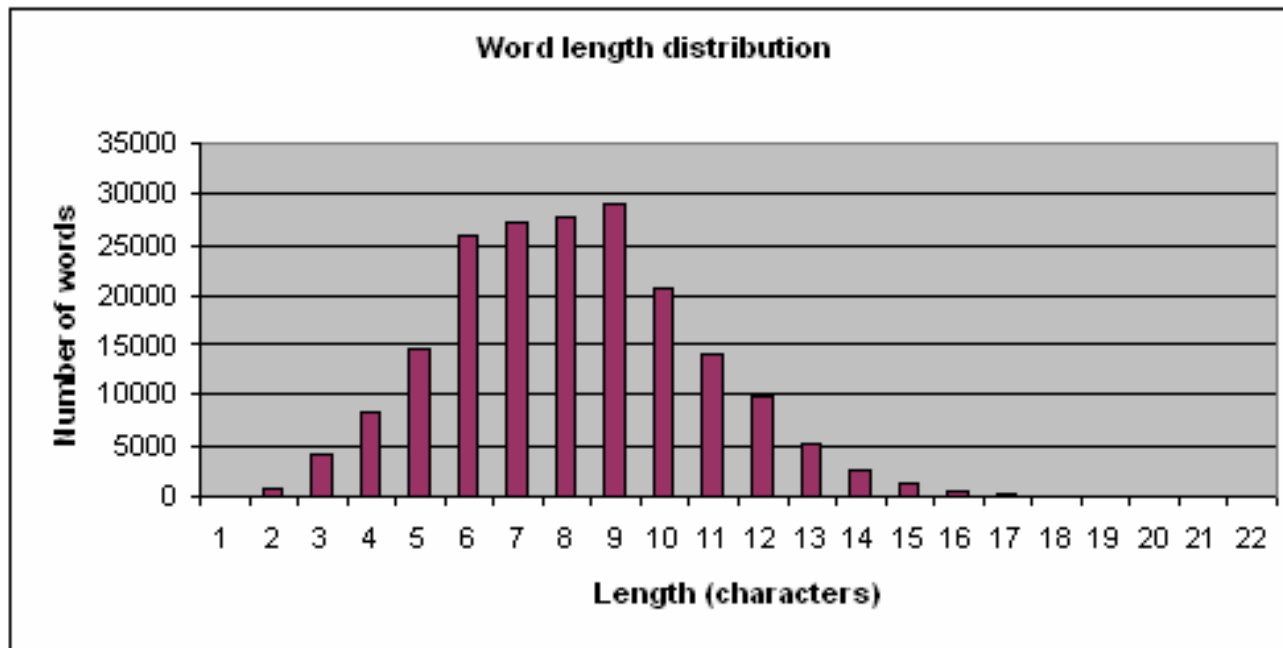
Rank	Word	Freq.	Rank	Word	Freq.	Rank	Word	Freq.	Rank	Word	Freq.	Rank	Word	Freq.
1	нь	98319	11	их	18380	21	би	11635	31	төрийн	9236	41	гээд	7669
2	юм	38986	12	хүн	18236	22	болон	11488	32	шиг	9029	42	байдаг	7630
3	байна	36444	13	тэр	17033	23	гэсэн	11164	33	зүйл	8823	43	эрх	7597
4	ч	36332	14	байсан	16406	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

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	төрийн	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

WORD FREQUENCY

- Some analyses on the word list
 - Average word length in Mongolian 6-9 char.
 - Many 2 and 3 character words are there



TWO CHARACTER WORDS

- There are 921 two-character words
- Most of them are content 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	үл	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	үр	6848		16	за	2788		26	яг	1679		36	ой	868
7	га	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	үг	3961		20	үе	2296		30	он	1122		40	ул	544

ANALYSIS ON CORPUS BALANCE

- Corpus balance is important
- For corpus balance analysis
 - Analysis on word frequency
 - Make frequencies of each text categories
 - Compare these frequencies
 - Generally, corpus balance is appropriate
 - But, there are some unbalanced parts

WORD FREQUENCY IN EACH TEXT STYLES

Rank	Literature		Law		Publish		Unen	
	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	хулийн	4776	гэж	17065	байгаа	4433
8	энэ	6156	хуль	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

CORPUS UNBALANCE

- Most of the basic words and their inflectional forms are included in the corpus
- But, 5 million words corpus is not enough to represent Mongolian
 - For example:
 - Word *хууль* (law) is occurred 6365 times in the corpus
 - Around 600 hundred words from the Law text are included in the corpus
 - This kind of word may have impact on statistical processing on the corpus
 - Word *хууль* is considered more common word than some words that are more common indeed

WORD FREQUENCY IN EACH TEXT STYLES

Rank	Literature		Law		Publish		Unen	
	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

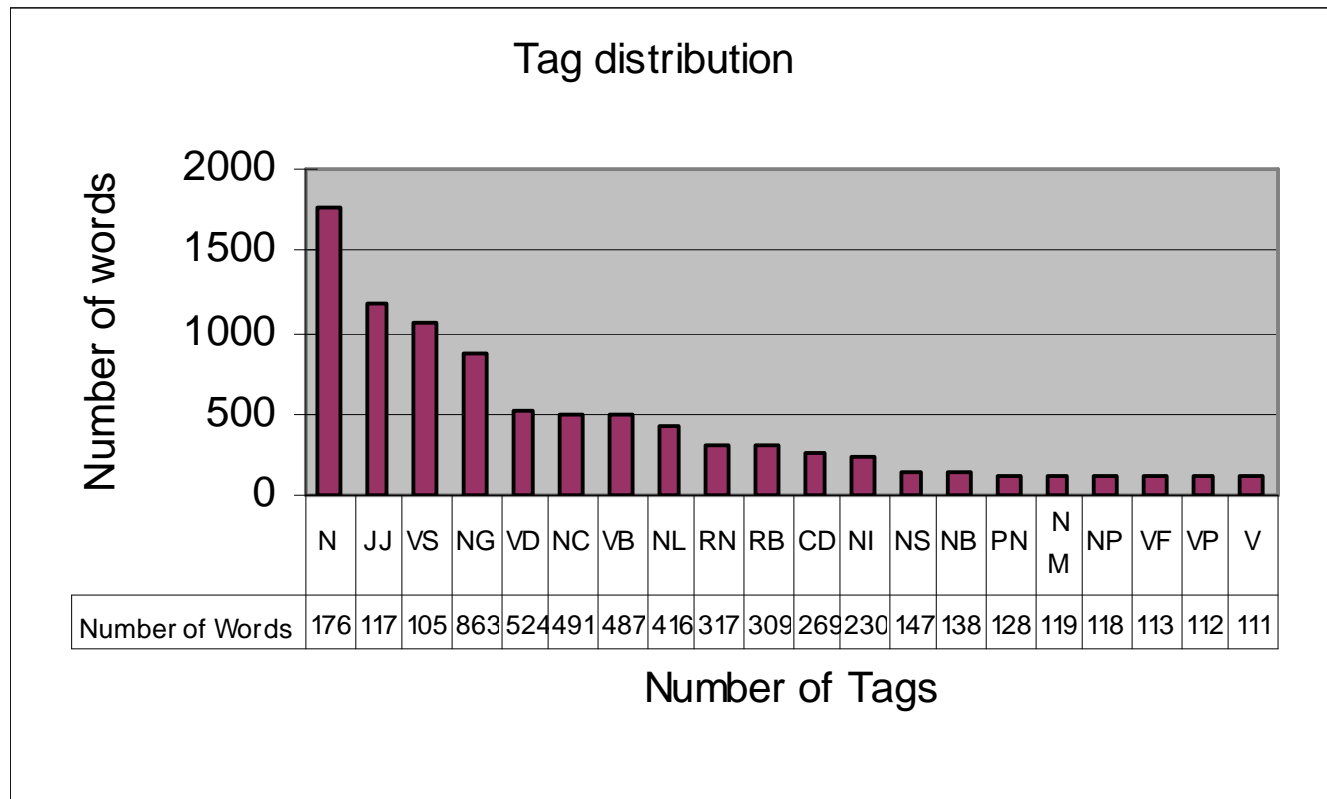
LEXICON

- After deriving the word list
 - Select the highest frequented words
 - Around the first 10k words are **85%** of the whole corpus
- To tag these 10k words
 - Using the manually tagged corpus

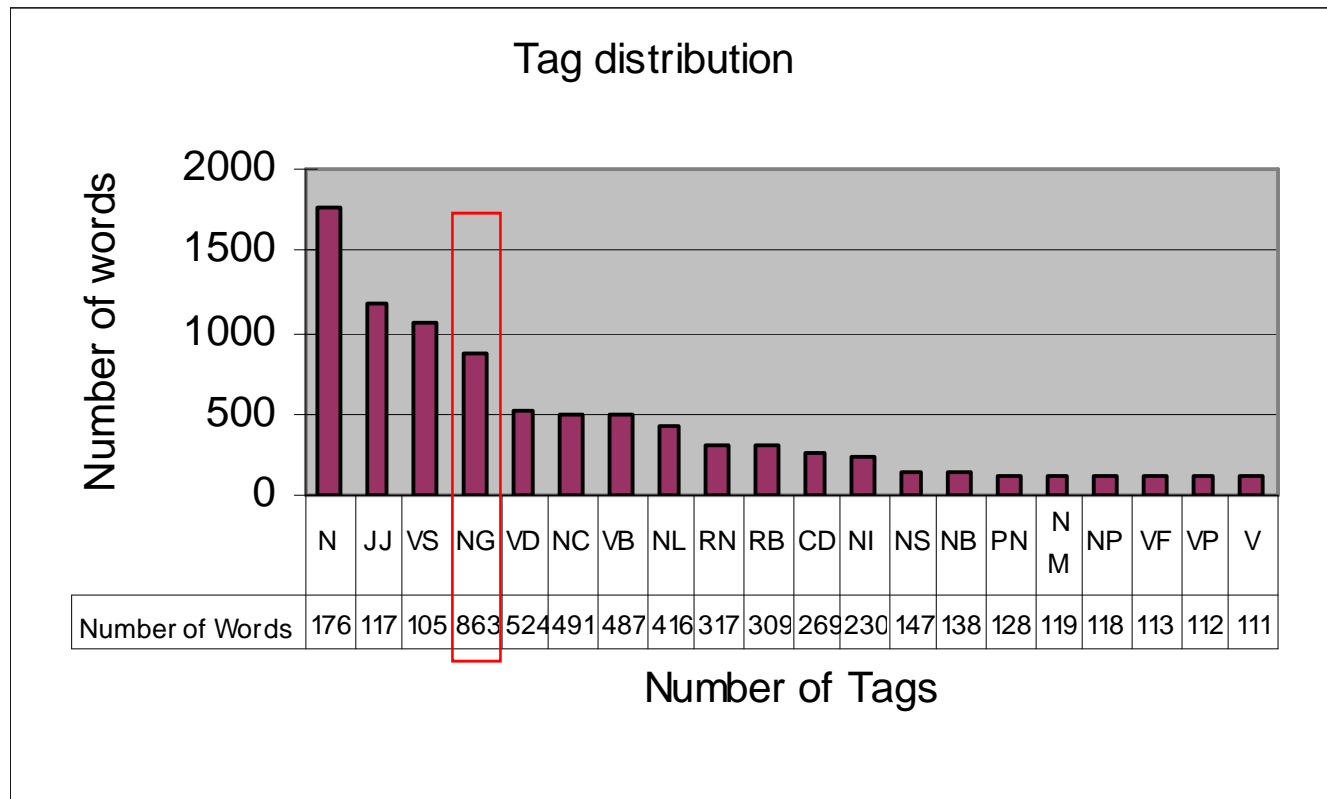
POS TAGSET

	Noun		Verb		Adword	Pronoun	Conjunction	Others
1	ABR	NMS	AUX	VBXL	DT	PJ	CC	CD
2	B	NP	GVB	VBS	JJ	PN	CS	PT
3	C	NPM	GVBB	VC	JJA			PR
4	G	NPMS	GVBC	VD	JJS			INF
5	I	NPB	GVBL	VDB	RB			INTJ
6	L	NPBS	GVBLS	VDC	RBA			MD
7	LS	NPC	GVBX	VDG				PUN
8	M	NPG	GVD	VDI				QN
9	N	NPGS	GVDS	VDL				NEG
10	NB	NPI	GVF	VDLS				
11	NBC	NPL	GVP	VDM				
12	NBS	NPLS	GVPC	VDS				
13	NC	NPM	GVPI	VDX				
14	ND	NPMS	GVPLS	VE				
15	NDS	NPS	GVPS	VF				
16	NCS	NS	GVS	VFX				
17	NG	NX	V	VG				
18	NGH	P	VB	VP				
19	NGS	PG	VBB	VPB				
20	NGHC	POS	VBBS	VPC				
21	NGHG	RN	VBC	VPG				
22	NGHB	S	VBG	VPI				
23	NGHLS		VBI	VPIS				
24	NGHM		VBIS	VPX				
25	NGHMS		VBL	VPS				
26	NI		VBLS	VPXB				
27	NIS		VBM	VPXG				
28	NL		VBX	VPXI				
29	NLS		VBXG	VS				
30	NM		VBXI	VSI				
31				VSS				
32				VSX				
33				VSI				

POS-TAG DISTERBUTION



POS-TAG DISTERBUTION



CONCLUSION

- In this phase
 - Analyzed the word frequency of the corpus
 - Determine mass of the words
 - Highly frequented 10k words
 - 85% of the whole corpus
 - Created a lexicon from the corpus for automatically tagging the corpus
 - Tagged the highly frequented 10k words
 - Using the manually tagged around 100k words